Ergodic Convergence Rates of Markov Processes

—Eigenvalues, Inequalities and Ergodic Theory

(Volume II)

Mu-Fa Chen
(Beijing Normal University)

First Edition 2000
Second Edition 2009

Contents

Volume II

Volume I

Stoch. Proc. Appl. 87:2 (2000), 281–297

EQUIVALENCE OF EXPONENTIAL ERGODICITY AND L^2-EXPONENTIAL CONVERGENCE FOR MARKOV CHAINS

Mu-Fa Chen

Department of Mathematics, Beijing Normal University,
Beijing 100875, The People's Republic of China
(First version: September 10, 1998; Revised version: July 9, 1999);
Second revised version Sep. 27, 1999); Third revised version Nov. 4, 1999)

Abstract. This paper studies the equivalence of exponential ergodicity and L^2-exponential convergence mainly for continuous-time Markov chains. In the reversible case, we show that the known criteria for exponential ergodicity are also criteria for L^2-exponential convergence. Until now, no criterion for L^2-exponential convergence has appeared in the literature. Some estimates for the rate of convergence of exponentially ergodic Markov chains are presented. These estimates are practical once the stationary distribution is known. Finally, the reversible part of the main result is extended to the Markov processes with general state space.

1. Introduction

Let $Q = (q_{ij})$ be a regular, irreducible Q-matrix on a countable set E. Assume that the corresponding transition probability matrix (also called the Q-process to indicate the connection with the matrix Q) $P(t) = (p_{ij}(t) : i, j \in E))$ is stationary with distribution $\pi = (\pi_i)$. Refer to Anderson (1991) or Chen (1992) for general terminology and notations. Note that the Q-matrix and Q-process are replaced by q-matrix and q-function respectively in Anderson (1991). A traditional topic in the study of Markov chains is exponential ergodicity. The Q-process $P(t)$ is said to have exponentially ergodic convergence to its stationary distribution π, if there is an $\alpha > 0$ such that for all i, $j \in E$, there exists a constant C_{ij} so that

$$|p_{ij}(t) - \pi_j| \leqslant C_{ij} e^{-\alpha t} \qquad \text{for all } t \geqslant 0. \tag{1.1}$$

2000 *Mathematics Subject Classification.* 60J27.

Key words and phrases. Exponential ergodicity, spectral gap, quadratic form, Markov chains.

Research supported in part by NSFC (No. 19631060), Math. Tian Yuan Found., Qiu Shi Sci. & Tech. Found., RFDP, MCME and MCMCAS.

E-mail: mfchen@bnu.edu.cn

Received 27 July 1999, received in revised form 4 November 1999, accepted 11 November 1999

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TeX

The parameter α is called an exponentially ergodic convergence rate. It is well known that (1.1) is equivalent to exponential decay of $\|p_{i\cdot}(t) - \pi\|_{\mathrm{Var}}$ as $t \to \infty$ (cf. Chen (1992), Theorem 4.43 (2)), where $\|\cdot\|_{\mathrm{Var}}$ is the total variation $\big(\|\mu\|_{\mathrm{Var}} = \mu_+(E) + \mu_-(E) = \sup_{|f|\leqslant 1} \big|\int f \mathrm{d}\mu\big|\big)$. About the convergence in total variation, there is a great deal of publications, see for instance Down et al (1995) , Lund et al (1996), Meyn and Tweedie (1993), Nummelin (1984) and references within.

A transition probability matrix $P(t)$ defines in a natural way a strongly continuous, contractive semigroup, denoted by $\{P(t)\}_{t\geqslant 0}$, on the space $L^2(\pi)$. A recent topic in the study of Markov processes is L^2-exponential convergence. A Markov semigroup $\{P(t)\}_{t\geqslant 0}$ is said to have L^2-exponential convergence if there exists an $\varepsilon > 0$ such that

$$\|P(t)f - \pi(f)\| \leqslant \|f - \pi(f)\| e^{-\varepsilon t}, \qquad t \geqslant 0,\ f \in L^2(\pi), \tag{1.2}$$

where $\|\cdot\|$ denotes the usual L^2-norm and $\pi(f) = \int f \mathrm{d}\pi$. The parameter ε is called an L^2-exponential convergence rate.

The two convergences in (1.1) and (1.2) look like rather different, but they are proved in the paper to be nearly equivalent for continuous-time Markov chains.

Before moving on, let us review some notation (cf. Chen (1992), Corollary 6.62 and Chapter 9). Let $(\cdot,\cdot)$ denote the usual inner product on $L^2(\pi)$. Define two operators on $L^2(\pi)$:

$$D(f) = \lim_{t\downarrow 0} t^{-1}(f - P(t)f,\ f)$$

provided the limit exists and

$$D^*(f) = \frac{1}{2}\sum_{i,j} \pi_i q_{ij}(f_j - f_i)^2.$$

The domains of these operators are defined as the subsets of $L^2(\pi)$ on which the operators are finite:

$$\mathscr{D}(D) = \{f \in L^2(\pi) : D(f) < \infty\} \quad \text{and} \quad \mathscr{D}(D^*) = \{f \in L^2(\pi) :\ D^*(f) < \infty\}.$$

One can deduce a quadratic form (named Dirichlet form) on $\mathscr{D}(D)$ by the standard way:

$$D(f,g) = -(\Omega f, g)$$

for f, g in the L^2-domain of the generator of Ω of $P(t)$. Similarly, we have the quadratic form $(D^*, \mathscr{D}(D^*))$. Let $\mathscr{K}$ be the set of functions on E with finite support. Then, $\mathscr{K} \subset \mathscr{D}(D)$ and for all $f \in \mathscr{K}$, $D(f) = D^*(f)$. Since

$$t^{-1}(f - P(t)f, f) = (2t)^{-1}\sum_i \pi_i P(t)[f - f_i]^2(i),$$

by Fatou's lemma, we have $\mathscr{D}(D) \subset \mathscr{D}(D^*)$.

In the reversible case (i.e., $\pi_i q_{ij} = \pi_j q_{ji}$ for all i, j), the regularity assumption on $Q = (q_{ij})$ implies that $D(f) = D^*(f)$ and $\mathscr{D}(D) = \mathscr{D}(D^*)$(cf. Chen (1992), Corollary 6.62). In other words,

$$\mathscr{K} \text{ is dense in } \mathscr{D}(D^*) \text{ in the } \|\cdot\|_{D^*}\text{-norm } \big(\|f\|_{D^*}^2 := \|f\|^2 + D^*(f)\big). \tag{1.3}$$

This may also holds for irreversible Markov chains but it remains unproven.

Let $q_i = -q_{ii}$ for $i \in E$. In general, a simple sufficient (but not necessary) condition for (1.3) is that

$$\sum_i \pi_i q_i < \infty \tag{1.4}$$

(cf. Chen (1992), Lemma 9.7).

In the irreversible situation, one often adopts the following symmetrizing procedure. Let $\widehat{P}(t) = (\hat{p}_{ij}(t))$ be the dual of $P(t)$: $\hat{p}_{ij}(t) = \pi_j p_{ji}(t)/\pi_i$. It first deduces the dual Q-matrix $\widehat{Q} = (\hat{q}_{ij})$ and then leads to a reversible Q-matrix $\overline{Q} = (\bar{q}_{ij})$ as follows:

$$\hat{q}_{ij} = \pi_j q_{ji}/\pi_i, \qquad \bar{q}_{ij} = \big(q_{ij} + \hat{q}_{ij}\big)/2. \tag{1.5}$$

We now introduce the first main result of the paper. The further results including some estimates of convergence rates are presented in Sections 3 and 4. In the discrete-time case, the reversible part of the result below was proved in a recent paper Roberts and Rosenthal (1997). We believe that the result is more or less known, though it may not have previously been stated explicitly.

Theorem 1.1. Let $Q = (q_{ij})$ be a regular, irreducible Q-matrix on a countable set E and the corresponding Q-process is stationary. Then

(1) L^2-exponential convergence implies exponentially ergodic convergence.
(2) If the Q-process is reversible, then the two convergences are equivalent.
(3) Assume that the Q-process is not reversible but $\mathscr{K}$ is dense in $\mathscr{D}(D^*)$. If the $\overline{Q}$-process is exponentially ergodic, then the Q-process is not only exponentially ergodic but also L^2-exponentially convergent.

Note that part (3) of the theorem is somewhat different from the inverse statement of part (1). This is a technical point in our proof. However, as we will show in the next section, it is often true that exponential ergodicity of the Q-process implies that of the $\overline{Q}$-process. It that case, we do have the inverse implication.

In view of Theorem 1.1, the study of one type of convergence may benefit from the study of the other type of convergence. For instance, in the reversible case, the well known criteria for exponential ergodicity (cf. Anderson (1991), Chen (1992) or (2.1)) now become criteria for L^2-exponential convergence. Until now, no criterion for L^2-exponential convergence has appeared in the literature. Note that on the one hand, some nice progress has been made recently in the study on the spectral gap for Markov processes (refer to the survey article Chen (1997) for the present status of the study and for a comprehensive list of references). On the other hand, this paper presents some explicit comparisons between the drift constant δ used in Criterion (2.1) below, the spectral gap and the exponential convergence rate (cf. Theorems 3.1, 4.1, 4.3—4.5 given in Sections 3 and 4). Based on these facts, whenever the stationary distribution π is known, one may deduce immediately many new bounds for exponentially ergodic convergence rate. Certainly, when π is not known, the use of the Dirichlet forms has no advantage, and one must adopt different approach (the coupling methods for instance, cf. Chen (1992) and Chen (1997)).

Of course, Theorem 1.1 is meaningful for more general Markov processes. Here we consider only the reversible case (refer also to the last paragraph of Section 4). The discrete-time analog of the next result was presented in Roberts and Rosenthal (1997).

Theorem 1.2. Let $\{P(t)\}_{t\geqslant 0}$ be a Markov semigroup on a measurable state space $(E, \mathscr{E})$, reversible with respect to a probability measure π. Then L^2-exponential convergence (1.2) is equivalent to the following statement:

$$\begin{aligned}&\text{For each probability measure } \mu \ll \pi \text{ with } \mathrm{d}\mu/\mathrm{d}\pi \in L^2(\pi)\text{, there is } C_\mu < \infty\\ &\text{such that } \|\mu P(t) - \pi\|_{\mathrm{Var}} \leqslant C_\mu e^{-\varepsilon t}, \qquad t \geqslant 0.\end{aligned} \tag{1.6}$$

For discrete state space, by setting $\mu = \delta_i$ in (1.6), it follows that Theorem 1.2 generalizes the reversible part of Theorem 1.1. Next, by Theorem 1.2 again, the equivalence of the two convergences also holds once the transition probability $p(t, x, \cdot)$ satisfies that for some $h > 0$, $p(h, x, \cdot) \ll \pi$ and $\mathrm{d}p(h, x, \cdot)/\mathrm{d}\pi \in L^2(\pi)$ for all $x \in E$[1]. In view of this, it follows that the equivalence holds for a large class of reversible Markov processes. However, in the infinite-dimensional situation, the restriction on μ given in (1.6) can not be removed. For instance, when there exist several Gibbs states corresponding to the same semigroup $\{P(t)\}$, it can happen that for each Gibbs states π, (1.2) holds but there is no hope to remove the restriction on μ since the Gibbs states may be singular each other. In other words, assertion (1.6) does not necessarily imply ergodicity of the corresponding process in the infinite-dimensional situation.

The proof of Theorem 1.1 is delayed until Sections 3 and 4. In the next section, we recall some known results which will be used in the later proofs and explain some background which leads to Theorem 1.1. In Section 5, the application of the results obtained in the paper is illustrated by some examples. The proof of Theorem 1.2 is given in the last section.

2. Preliminaries and Background

In this section, we recall some known facts and some motivation for the present study. In particular, a formula of the L^2-exponential convergence rate is given. The complication of the relationship between the drift constant δ used in Criterion (2.1) below and the convergence rates are illustrated. Besides, part (3) of Theorem 1.1 is proved.

First, we make a remark about the relation of the dense condition (1.3) and the regularity of $\overline{Q}$. By Chen (1992), Theorem 4.69, the Q-matrix $\widehat{Q}$ is regular and has the same stationary distribution π. Clearly, the form $\big(\widehat{D}, \mathscr{D}\big(\widehat{D}\big)\big)$ coincides with $(D, \mathscr{D}(D))$ by definition. Next, set $\mathscr{D}\big(\overline{D}\big) = \mathscr{D}(D)$. Then $\big(\overline{D}, \mathscr{D}\big(\overline{D}\big)\big)$ also

[1]Note that in the reversible case,

$$\int p(h, x, y)^2 \pi(\mathrm{d}y) = \int p(h, x, y) p(h, y, x) \pi(\mathrm{d}y) = p(2h, x, x) < \infty.$$

coincides with $(D, \mathscr{D}(D))$. Thus, if (1.3) holds, then $\big(\overline{D}, \mathscr{D}\big(\overline{D}\big)\big)$ also satisfies (1.3) and hence $\overline{Q} = (\bar{q}_{ij})$ is regular having the same stationary distribution π (cf. Chen (1992), Corollary 6.62 and Theorem 9.9). Therefore, the forms $(D, \mathscr{D}(D))$, $\big(\widehat{D}, \mathscr{D}\big(\widehat{D}\big)\big)$ and $\big(\overline{D}, \mathscr{D}\big(\overline{D}\big)\big)$ all coincide with $(D^*, \mathscr{D}(D^*))$ under (1.3). Conversely, if $\overline{Q}$ is regular (refer to Chen (1992), Theorem 2.25 for a practical criterion), then (1.3) holds (first for $(\overline{D}, \mathscr{D}(\overline{D}))$ and then for $(D, \mathscr{D}(D))$) since $\overline{Q}$ is reversible. Thus, condition (1.3) is indeed equivalent to the regularity of $\overline{Q}$.

Recall that the largest L^2-convergence rate $\varepsilon_{\max}$ in (1.2), denoted by gap(Q) or gap(D) according to our convenience, is given by the following variational formula.

$$\begin{aligned}\operatorname{gap}(D) &= \inf\{D(f) : f \in \mathscr{D}(D),\ \pi(f) = 0 \text{ and } \|f\| = 1\} \\ &= \inf\big\{ -(\Omega f,\, f) : f \in \mathscr{D}(\Omega),\ \pi(f) = 0 \text{ and } \|f\| = 1\big\}\end{aligned}$$

here Ω is at the moment regarded as the generator of $P(t)$ with domain $\mathscr{D}(\Omega)$ in $L^2(\pi)$. Actually, this formula of $\varepsilon_{\max}$ holds for any reversible Markov semigroup $\{P(t)\}_{t\geqslant 0}$ if we use the notations $(D, \mathscr{D}(D))$ and $(\Omega, \mathscr{D}(\Omega))$ to denote the Dirichlet form and the generator of $\{P(t)\}_{t\geqslant 0}$ respectively (cf. Chen (1992), Theorem 9.1). Thus, the rate ε in (1.2) and (1.6) can be simultaneously replaced by gap(D). When E is finite and Q is reversible, gap(Q) is the smallest non-trivial eigenvalue of $-Q$, i.e., the gap between the first two eigenvalues of $-Q$. See also the remark at the end of this section.

Under the dense condition (1.3), the study on L^2-exponential convergence in the irreversible case can be completely reduced to the reversible one since $\operatorname{gap}(D) = \inf\{D^*(f) : \pi(f) = 0,\ \|f\| = 1\}$ and furthermore $\operatorname{gap}(D) = \operatorname{gap}\big(\widehat{D}\big) = \operatorname{gap}\big(\overline{D}\big)$.

We now show that part (3) of Theorem 1.1 is a simple consequence of the first two parts of the theorem. Since $\overline{Q}$ is reversible, by part (2) and the assumption, the $\overline{Q}$-process is L^2-exponentially convergent and so is the Q-process since $\operatorname{gap}(Q) = \operatorname{gap}(\overline{Q})$. Exponential ergodicity of the Q-process then follows from part (1) of the theorem.

Next, denote by $\hat{\alpha} = \hat{\alpha}(Q)$ the supremum of the possible exponentially ergodic convergence rate in (1.1). Unfortunately, there is no variational formula for $\hat{\alpha}$. We only know some criteria for the positivity of $\hat{\alpha}$. The most practical criterion is: $\hat{\alpha} > 0$ iff for some/every finite set A, there exists a function φ and constants $\delta > 0$, $C \geqslant 0$ such that

$$\varphi \geqslant 1 \qquad \text{and} \qquad \Omega\varphi \leqslant -\delta\varphi + C I_A \tag{2.1}$$

(cf. Anderson (1991), Section 6.6, Theorem 6.5 or Chen (1992), Theorem 4.45 (3) or Down et al (1995) ; see also the comment above Lemma 4.2 in Section 4). Here and in what follows, the operator Ω is defined on the set $\{f : \sum_{j\neq i} q_{ij}|f_j| < \infty \text{ for all } i\}$: $\Omega f(i) = \sum_j q_{ij}(f_j - f_i)$. Clearly, the operator Ω and the form $(D^*, \mathscr{D}(D^*))$ are both determined by the Q-matrix $Q = (q_{ij})$. The next two examples show that (2.1) is not enough to determine either $\hat{\alpha}$ or gap(D). Hence the equivalence of the convergences is not obvious.

Example 2.1. Consider the birth-death process on $\mathbb{Z}_+ = \{0, 1, \cdots\}$ with birth rates $b_i = i + 2$ for $i \geqslant 0$ and death rates $a_i = i^2$ for $i \geqslant 1$. Then condition (2.1) holds for every $\delta > 0$ whenever A is large enough.

Proof. Let $\varphi_i = i + 1 \geqslant 1$. Then

$$\Omega\varphi(i) = 2 + i - i^2 \leqslant -(i/2)\varphi_i + 3I_A(i),$$

where $A \supset \{0, 1, 2, 3\}$. Thus, for $A = \{0, 1, \cdots, m\}$ with $m \geqslant 3$, (2.1) holds with $\delta = m/2$ which can be as large as we want.

However, for this example, it is known that $\hat{\alpha} = \text{gap}(D) = 2$ (cf. Chen (1996), Section 1). □

Clearly, the large δ in the last example comes from the large size of A. The next example shows that the constant δ can be arbitrarily small if the set A is taken to be a singleton.

Example 2.2. Let $(\pi_i > 0)$ be an arbitrary distribution on a countable set E and let $q_{ij} = \pi_j$ for $j \neq i$. Then, $\hat{\alpha} \geqslant \text{gap}(D) = 1$. But, when $A = \{i\}$, (2.1) holds iff $\delta < \pi_i$ which can be arbitrarily small for infinite E.

Proof. It is rather straightforward to check that $\text{gap}(D) = 1$ and every non-constant function φ with $\pi(\varphi) = 0$ is an eigenfunction of $\lambda_1 = \text{gap}(D)$.

Fix a reference point, say $0 \in E$ to simplify the notation. Solving the equation

$$\varphi \geqslant 1 \qquad \text{and} \qquad \Omega\varphi \leqslant -\beta\varphi + CI_{\{0\}}, \tag{2.2}$$

we get

$$1 \leqslant \pi(\varphi) := \sum_i \pi_i\varphi_i \leqslant (1-\beta)\varphi_i, \qquad i \neq 0.$$

This implies that $\beta < 1$ and

$$\varphi_i \geqslant \pi(\varphi)/(1-\beta), \qquad i \neq 0.$$

Then

$$\pi(\varphi) = \pi_0\varphi_0 + \sum_{i\neq 0} \pi_i\varphi_i \geqslant \pi_0 + (1-\pi_0)\pi(\varphi)/(1-\beta).$$

Or

$$(\pi_0 - \beta)\pi(\varphi) \geqslant \pi_0(1-\beta) > 0.$$

Thus, we must have $\beta < \pi_0$ and $\pi(\varphi) \geqslant \pi_0(1-\beta)/(\pi_0-\beta)$.

Let $\beta < \pi_0$ and $c \geqslant \pi_0(1-\beta)/(\pi_0-\beta)$. Define $\varphi_i = c/(1-\beta)$, $i \neq 0$ and $\varphi_0 = 1$. Then, (2.2) holds for these φ, β and every

$$C \geqslant \beta + \pi_0 - 1 + c(1-\pi_0)/(1-\beta).$$

Finally, since the reference point 0 is arbitrary, we obtain the required assertion. □

The above two examples are both reversible. Irreversible Markov chains are much more complicated and up to now there is still no effective tool to estimate the exponentially ergodic convergence rate $\hat{\alpha}$. A recent approach is studying the stronger L^2-exponential convergence (i.e., the spectral gap) instead of studying exponential ergodicity directly. However, it often happens that $\hat{\alpha} > \text{gap}(D)$ as illustrated by the following simple example.

Example 2.3. Let
$$Q_1 = \begin{pmatrix} -1/2 & 1/2 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix}.$$
Then the eigenvalues of Q_1 are 0, $-5/4 \pm \sqrt{7}\, i/4$ but we have $\hat{\alpha} = 5/4 > 1 = \text{gap}(Q_1)$.

A natural question arises: for infinite E, can $\hat{\alpha}$ be positive yet $\text{gap}(Q) = 0$? To answer this question, we need some preparation. Recall that the dual $\widehat{Q}$-process ($\hat{p}_{ij}(t) = \pi_j p_{ji}(t)/\pi_i$) has the same stationary distribution π. Thus, the Q-process is exponentially ergodic iff so is the $\widehat{Q}$-process and they have the same convergence rate $\hat{\alpha}$. These facts may be enough to conclude exponential ergodicity of the $\overline{Q}$-process but we are unable to prove it at the moment and there is still no counterexample either. The problem is that when we look at Criterion (2.1), the function φ and constant δ used there for Q and $\widehat{Q}$ may be different. The same problem appears in the opposite implication: exponential ergodicity of the $\overline{Q}$-process implies the one of the Q-process. But this is overcome in a rather technical way, stated as part (3) of Theorem 1.1. We now mention a simpler sufficient condition:

$$\varphi \geqslant 1, \quad \Omega\varphi \leqslant -\delta\varphi + CI_A \quad \text{and} \quad \widehat{\Omega}\varphi \leqslant \hat{\delta}\varphi + \widehat{C}I_A \quad \text{for some} \quad \hat{\delta} < \delta \tag{2.3}$$

Note that only a single function φ is used here and $\hat{\delta}$ is allowed to be positive! Then we have
$$\varphi \geqslant 1 \quad \text{and} \quad \overline{\Omega}\varphi \leqslant -\frac{\delta - \hat{\delta}}{2}\varphi + \frac{C + \widehat{C}}{2} I_A \tag{2.4}$$
and so the $\overline{Q}$-process is exponentially ergodic. Condition (2.3), which will be further weakened in (4.4), often holds for Markov chains (see Example 5.3 for instance) and we have no counterexample of a Markov chain for which condition (2.3) does not hold. Thus, by Theorem 1.1, it is often true that $\hat{\alpha} > 0 \iff \text{gap}(D) > 0$, and so we are safe in using the above symmetrizing approach when (2.3) holds at least.

Another motivation of the study comes from Markov Chain Monte Carlo. In this context, we are given a distribution, say $\pi_0 = 1/2$, $\pi_1 = \pi_2 = 1/4$. The problem is to construct a Markov chain whose law converges rapidly to π. It is natural to construct a reversible one. For instance
$$Q_2 = \begin{pmatrix} -1/2 & 1/2 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{pmatrix}.$$
Then we have $\text{gap}(Q_2) = (7 - \sqrt{17})/4 \approx 0.72 < 1$. On the other hand, one may regard Q_1 as a perturbation of Q_2 with the same equilibrium distribution π. Then, the irreversible Q_1 has a faster exponentially ergodic convergence rate than the reversible Q_2. However, even for infinite E, any local perturbation does not change exponential ergodicity by (2.1). Thus, for every local perturbation Q_1

of a reversible Q_2 (which is the main interest in practice), whenever π is kept, condition (2.3) holds and hence $\hat{\alpha}(Q_1) > 0 \Longleftrightarrow \text{gap}(Q_2) > 0$ by Theorem 1.1. In view of this observation, we can just consider the class of reversible processes.

To conclude this section, we make a remark on the term "gap". In the irreversible case, the term is not necessarily closely related to the spectrum of Ω as illustrated by Example 2.3. Next, recall the reversible Q-matrix $\overline{Q}$ given by (1.5). Under (1.3), we have $\text{gap}(D) = \text{gap}(\overline{D}) = \inf \text{spec}(-\overline{Q})|_{1^\perp}$ (cf. Chen (1992), Theorem 9.9). In this sense, it has some spectral meaning.

3. Proof of the first part of Theorem 1.1

In this section, we prove that L^2-exponential convergence implies exponential ergodicity, without using the dense condition (1.3). The first proof below is the shortest one but its conclusion is weaker than the second proof, which is meaningful in a more general setup (cf. Chen (1998)).

The first proof. The proof is rather easy as shown in Chen (1992), Proposition 9.20. By (1.2), we have

$$e^{-2\text{gap}(D)t}\|f - \pi(f)\|^2 \geqslant \pi_{i_0}|p_{i_0 j_0}(t) - \pi_{j_0}|^2$$

for the function $f_j = \delta_{j j_0}$ and arbitrary i_0 and j_0. Hence

$$|p_{ij}(t) - \pi_j| \leqslant \sqrt{\pi_j(1-\pi_j)/\pi_i}\, e^{-\text{gap}(D)t}$$

for all i, j, which proves (1.1). In other words, the spectral gap always lower bounds exponentially ergodic convergence rate. $\square$

The second proof. As mentioned in the first section, (1.1) is equivalent to exponential decay of $\|p_{i\cdot}(t) - \pi\|_{\text{Var}}$ as $t \to \infty$. But the convergence rate in the total variation may be smaller than the one in (1.1). The next result shows that we still have the same lower bound.

Theorem 3.1. For every probability measure μ, whenever the function μ_i/π_i belongs to $L^2(\pi)$, we have

$$\|\mu P(t) - \pi\|_{\text{Var}} \leqslant \|\mu/\pi - 1\|\, e^{-\text{gap}(D)t}$$

for all $t \geqslant 0$, In particular,

$$\|p_{i\,\cdot}(t) - \pi\|_{\text{Var}} \leqslant \sqrt{\pi_i^{-1} - 1}\, e^{-\text{gap}(D)t}.$$

Proof. The proof is similar to Chen (1998), Theorem 1.1, where the assertion was proved in the reversible case. Recall that $\hat{p}_{ij}(t) = \pi_j p_{ji}(t)/\pi_i$ and $\text{gap}(D) =$

$\text{gap}(\widehat{D})$. Assume that $\|\mu/\pi - 1\| < \infty$. Then, we have

$$\begin{aligned}
\|\mu P(t) - \pi\|_{\text{Var}} &= \sum_j |\mu P(t)(j) - \pi_j| \\
&= \sum_j \bigg| \sum_i (\mu_i - \pi_i) p_{ij}(t) \bigg| \\
&= \sum_j \pi_j \bigg| \sum_i \hat{p}_{ji}(t)(\mu_i/\pi_i - 1) \bigg| \\
&= \|\widehat{P}(t)(\mu/\pi - 1)\|_{L^1(\pi)} \\
&\leqslant \|\widehat{P}(t)(\mu/\pi - 1)\|.
\end{aligned}$$

Because the function $\mu/\pi - 1 \in L^2(\pi)$ has mean zero, by (1.2) and definition of $\text{gap}(D)$ (cf. Chen (1992), Theorem 9.1), the right-hand side is governed by

$$\|\mu/\pi - 1\| e^{-\text{gap}(\widehat{D})t} = \|\mu/\pi - 1\| e^{-\text{gap}(D)t}. \quad \square$$

It was also proved in Chen (1998), that the convergence rate given in Theorem 3.1 is indeed sharp for birth-death processes.

4. Proof of the second part of Theorem 1.1

This section is devoted to proving that exponential ergodicity implies L^2-exponential convergence. The proofs given in this section are very technical. The organization goes as follows. First, we deal with the reversible case, for which two different proofs are presented (Theorem 4.1 and Theorem 4.3). Another different proof will be presented in Section 6. Then, we reduce the irreducible case to the reversible one (Theorem 4.4). Finally, a criterion for the positivity of $\text{gap}(D)$ is presented (Theorem 4.5).

Reversible case. Let (X_t) be a Markov chain with transition probability $P(t)$. Define $\tau_A = \inf\{t \geqslant 0 : X_t \in A\}$. When A is a singleton, say $0 \in E$ for simplicity, we write τ_0 instead of $\tau_{\{0\}}$. The first assertion in the next result is due to [14] in the discrete-time case.

Theorem 4.1. In the reversible case, if there exists a constant $\beta > 0$ such that $\mathbb{E}^i \exp[\beta\tau_0] < \infty$ for all $i \in E$, then $\text{gap}(D) \geqslant \beta$. Furthermore, the last condition holds iff there exists a function φ defined on E such that

$$\varphi \geqslant 1 \qquad \text{and} \qquad \Omega\varphi(i) \leqslant -\beta\varphi_i \qquad \text{for all } \ i \neq 0. \tag{4.1}$$

Proof. (a) To prove the first assertion, fix $t > 0$ and consider the discrete-time chain $(X_{nt})_{n\geqslant 0}$ with transition probability $P = P(t)$. For $n \geqslant 0$ and $i \neq 0$, define

$$P_D^n f(i) = \mathbb{E}^i[f(X_{nt}) : \tau_0 > nt]$$

and write $P_D = P_D^1$. Then we have

$$P_D^n 1(i) = \mathbb{P}^i[\tau_0 > nt] \leqslant e^{-\beta nt}\mathbb{E}^i e^{\beta\tau_0} \qquad i \neq 0.$$

Thus, following the proof of Sokal and Thomas (1988), Lemma 3.11 (roughly speaking, the lemma says that for discrete-time Markov chains,

$$\|P_D\|_{L^2(\pi;E\setminus\{0\})} \leqslant \bar{r}^{-1}$$

whenever $\mathbb{E}^i\bar{r}^{\tau_0} < \infty$ for some $\bar{r} > 1$ and all i), we know that the operator norm $\|P_D\|$ in $L^2(\pi; E \setminus \{0\})$ is bounded above by $e^{-\beta t}$. At this point, we need not only the reversibility but also (1.3). However, condition (1.3) is automatic in the reversible case as mentioned before. Since $P = P(t)$ is nonnegative definite on $L^2(\pi)$, following the proof of Sokal and Thomas (1988), Lemma 3.12, the operator norm of $P(t)$ on $L^2(\pi; E) \setminus \{\text{constants}\}$ is bounded above by $\|P_D\|$. Hence, for every f with $\pi(f) = 0$ and $\|f\| = 1$, we get

$$(f,\, P(t)f) \leqslant \|P_D\| \leqslant e^{-\beta t}.$$

Therefore,

$$D(f) = \lim_{t\downarrow 0}\frac{1}{t}(f - P(t)f,\, f) \geqslant \lim_{t\downarrow 0}\frac{1}{t}\big(1 - e^{-\beta t}\big) = \beta$$

and so $\mathrm{gap}(D) \geqslant \beta$.

(b) To prove the last assertion, note that if φ satisfies (4.1), then so does $\tilde{\varphi}_i := \varphi_i$ for $i \neq 0$ and $\tilde{\varphi}_0 := 1$. On the other hand, by Anderson (1991), Section 6.2, Lemma 1.5, $\mathbb{E}^i \exp[\beta\tau_0] < \infty$ iff there exists a function y on E such that $y_0 = 0$ and $\Omega y(i) \leqslant -\beta y_i - 1$ for all $i \neq 0$. Hence, the required assertion follows by using the transform $\tilde{\varphi}_i = \beta y_i + 1$, $i \in E$. □

It should be pointed out that the continuous-time version (i.e. the first assertion of Theorem 4.1) was mentioned in Landim et al (1996), Proposition 4.1 without proof. Moreover, condition (4.1) was replaced by a stronger one (Landim et al (1996), Proposition 4.2) which is usually less effective since it fails for the simplest chain with two states. An estimate of the exponential convergence rate for stochastically ordered jump processes with continuous state space $[0, \infty)$ was obtained in Lund et al (1996) .

To present an improved result with a simpler proof, we need some preparation. Consider an exponentially ergodic chain. First, we show that for every finite set A, there exists a function φ and a constant $\delta > 0$ such that (4.3) below holds.

By Anderson (1991), Section 6.6, Theorem 6.5 or Chen (1992), Theorem 4.45(2), a Markov chain is exponentially ergodic iff for every finite set A, there exists some $0 < \delta < q_i$ for all i and a finite non-negative sequence (y_i) such that

$$\begin{cases} \sum_{j\notin A\cup\{i\}} q_{ij}y_j \leqslant (q_i - \delta)y_i - 1, & i \notin A \\ y_i = 0, & i \in A \\ \sum_{j\notin A} q_{ij}y_j < \infty, & i \in A. \end{cases} \tag{4.2}$$

As we mentioned before, this well known criterion does not say anything about the convergence rate. By using the transform $\varphi_i = \delta y_i + 1$, (4.2) can be rewritten in the simpler form (2.1) (but (4.2) and (2.1) are indeed equivalent). By replacing φ with φI_{A^c} in (2.1), we obtain the following condition.

$$\varphi|_A = 0, \quad \varphi|_{A^c} > 0 \quad \text{and} \quad \Omega\varphi \leqslant -\delta\varphi \quad \text{on} \quad A^c. \tag{4.3}$$

Next, define

$$\lambda_0(A^c) = \inf\{D(f):\ f \in \mathscr{D}(D),\ f|_A = 0 \text{ and } \pi(f^2) = 1\}.$$

By the dense condition (1.3), we have

$$\lambda_0(A^c) = \inf\{D^*(f):\ f|_A = 0 \text{ and } \pi(f^2) = 1\}.$$

Lemma 4.2. For a reversible process, under (4.3), we have $\lambda_0(A^c) \geqslant \delta$.

Proof. (a) Choose finite sets E_n containing A such that $E_n \uparrow E$. Let

$$\tau_n = \inf\{t \geqslant 0:\ X_t \notin E_n \setminus A\}.$$

Note that for every function f with finite support,

$$\big\{e^{\delta t} f(X_t)\big\}_{t\geqslant 0} \quad \text{is a } \mathbb{P}^i\text{-martingale with respect to the operator } \partial/\partial t + \Omega.$$

Since (4.3) also holds for $\varphi_n := \varphi I_{E_n\setminus A}$ on $E_n \setminus A$ and

$$\mathbb{P}^i[\tau_n = t] = \sum_{m=1}^{\infty} \mathbb{P}^i[\tau_n = \tau^{(m)} = t] = 0$$

(where $\tau^{(m)}$ is the mth jump time of the chain), we have

$$\begin{aligned}\mathbb{E}^i\big[e^{\delta(t\wedge\tau_A\wedge\tau_n)}\varphi_n(X_{t\wedge\tau_A\wedge\tau_n})\big] &= \mathbb{E}^i\big[e^{\delta(t\wedge\tau_n)}\varphi_n(X_{t\wedge\tau_n})\big]\\ &= \varphi_n(i) + \mathbb{E}^i\int_0^{t\wedge\tau_n}(\partial/\partial s + \Omega)\big[e^{\delta\cdot}\varphi_n\big](s, X_s)\mathrm{d}s\\ &\leqslant \varphi_i \qquad \text{for all } i \in E_n \setminus A.\end{aligned}$$

Letting $n \uparrow \infty$ and using Fatou's Lemma, we get

$$\mathbb{E}^i\big[e^{\delta(t\wedge\tau_A)}\varphi(X_{t\wedge\tau_A})\big] \leqslant \varphi_i \quad \text{for all } i \notin A.$$

The restriction $\varphi|_A = 0$ implies that $\mathbb{E}^i\big[e^{\delta t}\varphi(X_{t\wedge\tau_A})\big] \leqslant \varphi_i$ and hence

$$\mathbb{E}^i\varphi(X_{t\wedge\tau_A}) \leqslant \varphi_i e^{-\delta t}, \qquad t \geqslant 0,\ i \notin A.$$

(b) Next, since $E_n \setminus A$ is finite, there exists a function u_n with unit norm and $u_n|_{(E_n\setminus A)^c} = 0$ satisfying

$$D(u_n) = \lambda_0(E_n \setminus A) = \inf\{D(f) : f|_{(E_n\setminus A)^c} = 0 \text{ and } \|f\| = 1\}.$$

Because $D(|f|) \leqslant D(f)$, u_n must be non-negative. Furthermore, u_n should be an eigenfunction of Ω on the finite set $E_n \setminus A$: $\Omega u_n = -\lambda_0(E_n \setminus A)u_n$ on $E_n \setminus A$. The reversibility of Q is required at this point. In the irreversible case, one obtains the equation $\overline{\Omega} u_n = -\lambda_0(E_n \setminus A)u_n$ rather than $\Omega u_n = -\lambda_0(E_n \setminus A)u_n$, which may have no solution at all (cf. Example 5.4). We now follow the proof given in Chen and Wang (1998), Proof of Theorem 3.2. Since $E_n \setminus A$ is finite, there exists a positive c_1 such that $u_n(X_{t\wedge\tau_n}) \leqslant c_1\varphi(X_{t\wedge\tau_A})$. Thus,

$$u_n(i)e^{-\lambda_0(E_n\setminus A)t} = \mathbb{E}^i u_n(X_{t\wedge\tau_n}) \leqslant c_1\mathbb{E}^i\varphi(X_{t\wedge\tau_A}) \leqslant c_1\varphi_i e^{-\delta t}, \qquad i \in E_n \setminus A.$$

This implies that $\lambda_0(E_n \setminus A) \geqslant \delta$. Finally, because (1.3) holds in the reversible case, it is easy to show that $\lambda_0(E_n \setminus A) \downarrow \lambda_0(A^c)$ as $n \to \infty$ and so the required assertion follows. $\square$

Theorem 4.3. In the reversible case, if (4.3) holds with $A = \{0\}$, then $\mathrm{gap}(D) \geqslant \lambda_0(\{0\}^c) \geqslant \delta$.

Proof. (a) Choose f such that $\pi(f) = 0$. Let $A = \{0\}$ and $c = f_0$. Then

$$D(f) = D(f-c) \geqslant \lambda_0(A^c)\|f-c\|^2 = \lambda_0(A^c)\big(\|f\|^2 + c^2\big) \geqslant \lambda_0(A^c)\|f\|^2.$$

This means that $\mathrm{gap}(D) \geqslant \lambda_0(A^c)$.

(b) The second inequality now follows from Lemma 4.2. $\square$

In view of Theorems 3.1 and 4.1, (4.1) and (4.3) are not only criteria for $\mathrm{gap}(D) > 0$ or $\hat{\alpha} > 0$ but also give us a useful estimate for exponentially ergodic convergence rate in the reversible situation. The above proofs work only in the reversible case and moreover, one can not replace τ_0 by τ_A if A is not a singleton. For irreversible counterexamples of Theorem 4.1 and Theorem 4.3, see proof (b) of Example 5.3.

Irreversible case. A condition parallel to, but weaker than (2.3) is as follows.

$$\varphi|_A = 0,\ \varphi|_{A^c} > 0,\ \Omega\varphi \leqslant -\delta\varphi \text{ and } \widehat{\Omega}\varphi \leqslant \hat{\delta}\varphi \text{ on } A^c \text{ for some } \hat{\delta} < \delta. \tag{4.4}$$

Then, we have $\overline{\Omega}\varphi \leqslant -\big(\delta - \hat{\delta}\big)\varphi/2$ on A^c. The following result is now a straightforward consequence of Theorem 4.3.

Theorem 4.4. Assume that (1.3) holds in the irreversible case. If (4.3) holds with Ω replaced by $\overline{\Omega}$ and $A = \{0\}$, then $\mathrm{gap}(Q) = \mathrm{gap}(\overline{Q}) \geqslant \delta$. In particular, the last condition holds if (4.4) is satisfied with $A = \{0\}$, in which case $\mathrm{gap}(Q) \geqslant (\delta - \hat{\delta})/2$.

General estimate. As we have seen from Example 2.2 the lower bound given by Theorems 4.1, 4.3 and 4.4 may be very small. On the other hand, as we have seen from Example 2.1, Criterion (2.1) is much more practical than (2.2). Thus, it is natural to use (2.1) instead of (2.2). That is the goal of this subsection. For any subset B of E, define a restricted form D_B^* on B and the associated measure π_B as:

$$D_B^*(f) = \frac{1}{2}\sum_{i,j\in B}\pi_i q_{ij}(f_j - f_i)^2, \qquad \pi_B(i) = \pi_i \Big/ \sum_{j\in B}\pi_j.$$

The spectral gap of D_B^* is

$$\operatorname{gap}(D_B^*) = \inf\{D_B^*(f) :\ \pi_B(f) = 0,\ \pi_B(f^2) = 1\}.$$

For finite B, since D_B^* coincides with the form generated by the symmetrized Q-matrix $(\bar{q}_{ij})$, it follows that $\operatorname{gap}(D_B^*) > 0$ even though the new Q-matrix can be reducible when restricted to B (cf. Chen (1992), Theorem 9.9 and Chen and Wang (1998)). Let $\pi(C) = \sum_{j\in C} \pi_j$ and $M_A = \max_{i\in A}\big(q_i + \sum_{j\notin A} q_{ij}\big)$. In the reversible case, M_A can be replaced by $2\max_{i\in A}\sum_{j\notin A} q_{ij}$. For finite A, $M_A \leqslant 2\max_{i\in A} q_i < \infty$. Thus, whenever $\lambda_0(A^c) > 0$ for some finite A, we can make B large enough so that the right-hand side of (4.5) below is positive.

Theorem 4.5. Assume that (1.3) holds in the irreversible case. Then, for any $A \subset B$ with $0 < \pi(A), \pi(B) < 1$, we have

$$\frac{\lambda_0(A^c)}{\pi(A)} \geqslant \operatorname{gap}(D) \geqslant \frac{\operatorname{gap}(D_B^*)[\lambda_0(A^c)\pi(B) - M_A\pi(B^c)]}{2\operatorname{gap}(D_B^*) + \pi(B)^2[\lambda_0(A^c) + M_A]}. \tag{4.5}$$

In particular, $\operatorname{gap}(D) > 0$ iff $\lambda_0(A^c) > 0$ for some finite A.

Proof. The proof is a slight modification of a much more general result Chen and Wang (1998), Theorem 3.1 applied to $\overline{Q} = (\bar{q}_{ij})$. We sketch the proof here. We must keep in mind that $\pi_i q_{ij}$ may not be symmetric. For the upper bound, noticing that for every f with norm one and $f|_A = 0$, we have

$$\pi(f^2) - \pi(f)^2 = 1 - \pi\big(fI_{A^c}\big)^2 \geqslant 1 - \pi\big(A^c\big) = \pi(A)$$

so $\operatorname{gap}(D) \leqslant D(f)/\pi(A)$. Then, it follows that $\operatorname{gap}(D) \leqslant \lambda_0\big(A^c\big)/\pi(A)$ as required.

For the lower bound, the original proof is based on Cheeger's splitting technique and consists of two estimates:

$$D(f) \geqslant \operatorname{gap}(D_B^*)\pi(B)^{-2}\big[\gamma - \pi\big(B^c\big)\big], \tag{4.6}$$

$$D(fI_{A^c}) \leqslant 2D(f) + M_A\gamma \tag{4.7}$$

for every f with $\pi(f) = 0$ and $\pi(f^2) = 1$, here $\gamma = \pi\big(f^2 I_B\big)$. Notice that

$$D\big(fI_{A^c}\big) \geqslant \lambda_0\big(A^c\big)\pi\big(f^2 I_{A^c}\big) \geqslant \lambda_0\big(A^c\big)(1-\gamma).$$

Once (4.6) and (4.7) have been proved, we obtain two lower bounds of $D(f)$, say $g_1(\gamma)$ and $g_2(\gamma)$. Then $D(f) \geqslant \inf_{\gamma\in[0,1]}\max\{g_1(\gamma), g_2(\gamma)\}$. Optimization of this lower bound with respect to $\gamma \in [0,1]$ produces the lower bound in (4.5).

We now prove inequality (4.6). Because

$$D(f) = D^*(f) \geqslant D_B^*(fI_B),$$

one needs to show that

$$\pi\big(f^2 I_B\big) - \pi(B)^{-1}\pi\big(fI_B\big)^2 \geqslant \pi(B)^{-1}\big[\gamma - \pi\big(B^c\big)\big].$$

This can be done by using the inequality:

$$\pi(fI_B)^2 = \pi(fI_{B^c})^2 \leqslant \pi\big(f^2 I_{B^c}\big)\pi\big(B^c\big) = (1-\gamma)\pi\big(B^c\big).$$

The inequality (4.7) is based on

$$|(fI_{A^c})_j - (fI_{A^c})_i| \leqslant |f_j - f_i| + I_{\{(i,j)\in B\times A^c\cup A^c\times B\}}|(fI_A)_j - (fI_A)_i|$$

and $(a+b)^2 \leqslant 2(a^2+b^2)$. The coefficient M_A comes from the following calculation:

$$\begin{aligned}\sum_{(i,j)\in B\times A^c\cup A^c\times B} \pi_i q_{ij}[(fI_A)_j - (fI_A)_i]^2 &= \sum_{i\in A}\pi_i f_i^2 \sum_{j\notin A} q_{ij} + \sum_{i\notin A}\pi_i \sum_{j\in A} q_{ij} f_j^2\\
&\leqslant \sum_{i\in A}\pi_i f_i^2 \sum_{j\notin A} q_{ij} + \sum_{i}\pi_i \sum_{j\neq i} q_{ij} (fI_A)_j^2\\
&= \sum_{i\in A}\pi_i f_i^2 \sum_{j\notin A} q_{ij} + \sum_{i\in A}\pi_i q_i f_i^2\\
&\leqslant \max_{i\in A}\Big\{\sum_{j\notin A} q_{ij} + q_i\Big\}\pi\big(f^2 I_A\big)\\
&\leqslant M_A\gamma.\end{aligned}$$

In the second equality, we used the stationary property of the process (cf. Chen (1992), Theorem 4.17). □

To conclude this section, we mention that the irreversible part of Theorem 1.1 can also be extended to a more general setting, because the analogs of Theorem 3.1, Criterion (2.1) and Theorem 4.5 have been obtained in Chen (1998), Down et al (1995) and Chen and Wang (1998) respectively.

5. Examples

In this section, we discuss four examples. the first two examples are reversible, they illustrate the application of Theorem 1.1 and Theorem 4.5. The last two examples are irreversible, they illustrate the application of Theorem 4.4; they also show the effectiveness of condition (4.4) and the independence of the convergence rates and the eigenvalues of the operator Ω.

Example 5.1. Let $E = \mathbb{Z}_+$. Consider the birth-death process with death rates a_i and birth rates b_i: $a_i = b_i = i^\gamma$ $(i \geqslant 1)$ for some $\gamma > 0$, $a_0 = 0$ and $b_0 = 1$. The process is exponentially ergodic iff $\gamma \geqslant 2$.

Proof. It is well-known that the chain is ergodic iff $\gamma > 1$. Moreover $\text{gap}(D) > 0$ iff $\gamma \geqslant 2$ (cf. Chen (1996) or Chen and Wang (1998), Example 4.5). Thus, by part (2) of Theorem 1.1, the process is exponentially ergodic iff $\gamma \geqslant 2$. The assertion is well-known when $\gamma > 2$ but is new for $\gamma \in (1, 2]$ (cf. Anderson (1991), Proposition 6.6 or Chen (1992), Corollary 4.51). We remark that $\sum_i \pi_i q_i = \infty$ for this example. □

Example 5.2. This is a continuation of Example 2.1. By Lemma 4.2, it follows that $\lambda_0(A^c) \geqslant m/2$ for $A = \{0, 1, \cdots, m\}$ with $m \geqslant 3$. Next, fix A and choose N large enough so that

$$\pi(B) \geqslant \frac{M_A}{M_A + \lambda_0(A^c)}, \qquad B := \{0, 1, \cdots, N\}.$$

We have the following dual variational formula Chen (1999), Theorem 3.2 for $\text{gap}(D_B^*)$:

$$\text{gap}(D_B^*) = \sup_{w \in \mathscr{W}} \inf_{0 \leqslant i \leqslant N-1} b_i \mu_i (w_{i+1} - w_i) \Big/ \sum_{j=i+1}^{N} \mu_j w_j$$

where $\mu_0 = 1$, $\mu_n = b_0 \cdots b_{n-1}/a_1 \cdots a_n$, $1 \leqslant n \leqslant N$ and $\mathscr{W}$ is the set of all strictly increasing sequences (w_i) with $\sum_{i=0}^{N} \mu_i w_i \geqslant 0$. The result is valid even for infinite N (cf. Chen (1996) or Chen (1999)). Note that (when $N < \infty$) each $w \in \mathscr{W}$ gives us a non-trivial lower bound of $\text{gap}(D_B^*)$ and then we obtain a non-trivial lower bound of $\text{gap}(D)$ (and furthermore of $\hat{\alpha}$) by Theorem 4.5 (and Theorem 3.1). $\square$

For general Markov chains with Q-matrix $Q = (q_{ij})$ and stationary distribution π, once the elements $\bar{q}_{i,i+1}$ and $\bar{q}_{i,i-1}$ of the symmetrizing matrix $\overline{Q} = (\bar{q}_{ij})$ are positive, the Dirichlet form $D(f)$ is bounded below by a form of birth-death process with birth rates $b_i = \bar{q}_{i,i+1}$ and death rates $a_i = \bar{q}_{i,i-1}$. Thus, the procedure used in Example 5.1 is still applicable. In other words, it is now often easy to obtain a non-trivial lower bound of $\hat{\alpha}$ once π is explicit.

Next, we consider the irreversible case. A Markov chain on $\mathbb{Z}_+$ is called a *single death process* if $q_{i,i-1} > 0$ for all $i \geqslant 1$ but $q_{ij} = 0$ for all $i \geqslant 2$ and $0 \leqslant j \leqslant i-2$. There is no restriction on the rates q_{ij} for $j > i$. Such a process has an advantage: its stationary distribution is computable by an iterative procedure:

$$\pi_1 = \frac{\pi_0 q_0}{q_{10}}, \qquad \pi_{n+1} = \frac{\pi_n q_n}{q_{n+1,n}} - \sum_{k=0}^{n-1} \frac{\pi_k q_{kn}}{q_{n+1,n}}, \qquad n \geqslant 1 \tag{5.1}$$

This provides us a chance to apply the estimate from Theorem 4.4. However, in order to illustrate some idea and make the computation possible by hand, we consider here two very particular examples only.

Example 5.3. Let $q_{0k} > 0$, $q_k = q_{k,k-1} > 0$ for all $k \geqslant 1$, $q_0 = \sum_{k \geqslant 1} q_{0k}$ and $q_{ij} = 0$ for all other $j \neq i$. Suppose that $0 < c_1 = \inf_{i \geqslant 1} q_i \leqslant \sup_{i \geqslant 1} q_i = c_2 < \infty$. Then the process is exponentially ergodic iff $\{q_{0k}\}$ has geometric decay: $q_{0k} \leqslant c\theta^k$ for some constants c and $\theta < 1$.

Proof. (a) First, we compute the stationary distribution. From the iterative procedure (5.1) plus some computations, it follows that the Q-matrix $Q = (q_{ij})$ has a stationary distribution π iff $\sum_{k=2}^{\infty} (1/q_n) \big(q_0 - \sum_{k=1}^{n-1} q_{0k}\big) < \infty$. If this holds,

then

$$\pi_0 = \left\{1 + \frac{q_0}{q_1} + \sum_{k=2}^{\infty} \frac{1}{q_n}\left(q_0 - \sum_{k=1}^{n-1} q_{0k}\right)\right\}^{-1} \quad \text{and}$$
$$\pi_n = \frac{\pi_0}{q_n}\left(q_0 - \sum_{k=1}^{n-1} q_{0k}\right) \qquad \text{for all } n \geqslant 1. \tag{5.2}$$

(b) We now prove the conclusion. By solving the inequality $\Omega\varphi(i) \leqslant -\beta\varphi_i$ for $i \geqslant 1$, we get

$$\varphi_i \geqslant \prod_{j=1}^{i} \frac{q_j}{q_j - \beta}\varphi_0 \qquad \text{for } i \geqslant 1$$

whenever $\beta < c_1$. Let $\varphi_0 = 1$. Then, condition $\Omega\varphi(0) \leqslant C - \beta\varphi_0$ gives us

$$\infty > C \geqslant \beta - q_0 + \sum_{i=1}^{\infty} q_{0i} \prod_{j=1}^{i} \frac{q_j}{q_j - \beta}. \tag{5.3}$$

Because $c_2 = \sup_{k\geqslant 1} q_k < \infty$, we have

$$\sum_{i=1}^{\infty} q_{0i} \prod_{j=1}^{i} q_j/(q_j - \beta) \geqslant \sum_{i=1}^{\infty} q_{0i}(1 - \beta/c_2)^{-i}.$$

Thus, (5.3) holds only if $\{q_{0k}\}$ has geometric decay. From this, one can easily construct some examples ($\{q_{0k}\}$ has only polynomial decay for instance) for which $\hat{\alpha} = 0$ and so $\text{gap}(Q) = 0$ by Theorem 1.1. Thus, Theorems 4.1 and 4.3 fail since conditions (4.1) and (4.3) (with $A = \{0\}$) do not use the sequence $\{q_{0k}\}$.

Conversely, if $\{q_{0k}\}$ has geometric decay, then

$$\sum_{i=1}^{\infty} q_{0i} \prod_{j=1}^{i} q_j/(q_j - \beta) \leqslant \sum_{i=1}^{\infty} q_{0i}(1 - \beta/c_1)^{-i} < \infty$$

for sufficient small $\beta > 0$. Hence (5.3) holds and so we have thus proved the required conclusion. □

Example 5.4. Everything is the same as in Example 5.3 but $q_k = 1$ and $q_{0k} = \theta^k$ $(k \geqslant 1)$ for some $\theta < 1$. Then (2.3) and (4.4) hold. Moreover, when $\theta \leqslant 1/2$, $\text{gap}(Q) \geqslant 1 - \sqrt{\theta}$. However, the operator Ω has no non-zero real eigenvalues λ in the ordinary sense: $\Omega f(i) = -\lambda f_i$ for some real $f \neq 0$ and all $i \in E$.

Proof. (a) First, we prove (2.3). By (5.1), we have

$$q_0 = \frac{\theta}{1-\theta}, \qquad \pi_0 = \left\{1 + \frac{\theta}{(1-\theta)^2}\right\}^{-1} \quad \text{and} \quad \pi_n = \frac{\pi_0\theta^n}{1-\theta} \qquad \text{for all } n \geqslant 1.$$

Moreover, $\varphi_i := (1-\beta)^{-i}\,(i \geqslant 0)$ satisfies $\Omega\varphi(i) = -\beta\varphi_i\,(i \geqslant 1)$ and (5.3) for all $\beta < 1-\theta$. On the other hand, for the dual matrix $\widehat{Q} = (\hat{q}_{ij})$, we have $\hat{q}_{i0} = 1-\theta$ and $\hat{q}_{i,i+1} = \theta\,(i \geqslant 1)$. Thus,

$$\sum_j \hat{q}_{ij}(\varphi_j-\varphi_i) = \theta(\varphi_{i+1}-\varphi_i)+(1-\theta)(\varphi_0-\varphi_i) = \left[\frac{\theta}{1-\beta}-1+(1-\theta)(1-\beta)^i\right]\varphi_i$$

for all $i \geqslant 1$. We now prove that

$$\hat{\beta} := \theta(1-\beta)^{-1} - 1 + (1-\theta)(1-\beta) < \beta.$$

These two facts imply (2.3). Let $\tilde{\beta} = 1-\beta$. Then, $\tilde{\beta} \in (\theta, 1)$. To prove the last inequality, it suffices to show that $(2-\theta)\tilde{\beta}^2 - 2\tilde{\beta} + \theta < 0$. Solving this equation, we get two roots: $\tilde{\beta}_1 = \theta/(2-\theta)$ and $\tilde{\beta}_2 = 1$. The inequality now follows by confirming that $\tilde{\beta}_1 < \theta$.

(b) We show that the operator Ω has no non-zero eigenvalues. That is, $\Omega f = -\lambda f$ has no non-trivial solution ($\lambda \neq 0$ and $f \neq 0$). Thus, it is no hope to estimate $\hat{\alpha}$ by using the eigenvalues of Ω. Solving the equation $\Omega f(i) = -\lambda f_i\ (i \geqslant 1)$, one gets $(1-\lambda)f_i = f_{i-1}\,(i \geqslant 1)$. From this, it follows that $f_i \equiv 0$ once $\lambda = 1$. Otherwise, $f_i = (1-\lambda)^{-i} f_0$ for all $i \geqslant 1$ and $f_0 \neq 0$. From $\Omega f(0) = -\lambda f_0$, it follows that we must have $\theta < |1-\lambda|$ and

$$\sum_{k=1}^{\infty} \theta^k (1-\lambda)^{-k} - \theta(1-\theta)^{-1} = -\lambda.$$

But when $\theta \leqslant 1/2$, the last equation holds iff $\lambda = 0$.

(c) The rest of the proof estimates $\hat{\alpha}$. Let $\varphi_0 = 0$ and $\varphi_i = (1-\beta)^{-i+1}\,(i \geqslant 1)$. Then (4.3) holds with $A = \{0\}$ and $\delta = \beta < 1$. Moreover, (4.4) holds with $A = \{0\}$ and $\hat{\delta} = \theta/(1-\beta) - 1$ whenever $\beta < \sqrt{1-\theta}$. Thus,

$$\overline{\Omega}\varphi \leqslant -\frac{1}{2}\left\{\beta - \frac{\theta}{1-\beta} + 1\right\}\varphi =: -\bar{\beta}\varphi \qquad \text{on} \quad \{0\}^c. \tag{5.4}$$

Maximizing $\bar{\beta}$ with respect to $\beta < \sqrt{1-\theta}$, we get $\bar{\beta} = 1 - \sqrt{\theta}$. Thus, by Theorem 4.4, we obtain

$$\text{gap}(D) \geqslant \lambda_0(\{0\}^c) \geqslant \bar{\beta}$$

and hence $\hat{\alpha} \geqslant \bar{\beta} > 0$ by part (1) of Theorem 1.1. $\square$

We remark that the lower bound produced by Theorem 4.1 is the same for this example. If one uses (2.3) instead of (4.4), then the resulting lower bound is $1 - \sqrt{\theta(2-\theta)}$ which is smaller than $1 - \sqrt{\theta}$.

As we mentioned before, the lower bound provided by Theorems 4.1, 4.3 and 4.4 may be rather rough. A possible way to improve the estimate of gap(D) is by directly using the methods developed in the study on the spectral gap for reversible processes (cf. Chen (1997) and Chen (1999)). We now illustrate one of the methods.

Let f satisfy $\pi(f) = 0$ and $\|f\| = 1$. Denote $\sum_{i\in[a,b]} \pi_i$ by $\pi[a,b]$. Then, we have

$$\begin{aligned}
1 &= \frac{1}{2}\sum_{i,j} \pi_i\pi_j(f_j - f_i)^2 \\
&= \sum_{i<j} \pi_i\pi_j(f_j - f_0 + f_0 - f_i)^2 \\
&\leqslant 2\sum_{i<j} \pi_i\pi_j(f_j - f_0)^2 + 2\sum_{i<j} \pi_i\pi_j(f_i - f_0)^2 \\
&\leqslant 2\sum_{j\geqslant 1} \pi_j\pi[0, j-1]\sum_{k=1}^{j}(f_k - f_{k-1})^2\gamma^k\sum_{\ell=1}^{j}\gamma^{-\ell} \\
&\quad + 2\sum_i \pi_i\theta^i(f_i - f_0)^2\frac{\pi[i+1,\infty)}{\theta^i} \\
&=: 2I_1 + 2I_2,
\end{aligned}$$

where $\gamma \in (0,1)$ is a constant to be determined later. Next,

$$\begin{aligned}
I_1 &\leqslant \sum_{k\geqslant 1}(f_k - f_{k-1})^2\sum_{j\geqslant k}\pi_j\sum_{\ell=1}^{j}\gamma^{-\ell+k} \leqslant \frac{\gamma}{(1-\gamma)(\gamma-\theta)}\sum_{k\geqslant 1}\pi_k(f_k - f_{k-1})^2, \\
&\quad \gamma \in (\theta, 1). \\
I_2 &\leqslant \frac{\pi_0\theta}{(1-\theta)^2}\sum_i \pi_i\theta^i(f_i - f_0)^2.
\end{aligned}$$

Minimizing $\gamma/(1-\gamma)(\gamma-\theta)$ with respect to γ, we get the minimum $\big(1-\sqrt{\theta}\,\big)^{-2}$. Thus,

$$\begin{aligned}
1 &\leqslant \frac{2}{\big(1-\sqrt{\theta}\,\big)^2}\sum_{k\geqslant 1}\pi_k q_{k,k-1}(f_k - f_{k-1})^2 + \frac{2\pi_0\theta}{(1-\theta)^2}\sum_i \pi_i q_{0i}(f_i - f_0)^2 \\
&\leqslant 2\max\left\{\frac{1}{\big(1-\sqrt{\theta}\,\big)^2},\ \frac{\pi_0\theta}{(1-\theta)^2}\right\}D(f).
\end{aligned}$$

Comparing this with the estimate given in Example 5.4, since $1-\sqrt{\theta} \geqslant \big(1-\sqrt{\theta}\,\big)^2/2$, we see that this usually quite effective method does not make any improvement.

6. Proof of Theorem 1.2

To prove Theorem 1.2, we need some preparation. As an analog of the usual L^p-space of functions, we define some space of finite signed measures as follows [Roberts and Rosenthal (1997)]:

$$\mathscr{L}^p(\pi) = \{\mu : \mu \text{ is a (finite) signed measure, } \mu \ll \pi \text{ and } \mathrm{d}\mu/\mathrm{d}\pi \in L^p(\pi)\},$$

where $1 \leqslant p < \infty$. For $\mu \in \mathscr{L}^p(\pi)$, set

$$\|\mu\|_{\mathscr{L}^p(\pi)}^p = \int |\mathrm{d}\mu/\mathrm{d}\pi|^p \mathrm{d}\pi.$$

When $p = 2$, $\mathscr{L}^2(\pi)$ is a Hilbert space with inner product

$$\langle \mu, \nu \rangle = \int (\mathrm{d}\mu/\mathrm{d}\pi)(\mathrm{d}\nu/\mathrm{d}\pi)\mathrm{d}\pi.$$

Due to the reversibility of $P(t)$, it is easy to check that the action of $P(t)$ on $\mu \in \mathscr{L}^2(\pi)$ is equivalent to the action of $P(t)$ on $f \in L^2(\pi)$. Moreover,

$$\mu \in \mathscr{L}^p(\pi) \Longrightarrow \mu P(t) \in \mathscr{L}^p(\pi)$$

for all $t \geqslant 0$. In particular, the L^2-exponential convergence can be restated as follows:

$$\begin{aligned} &\text{For every signed measure } \mu \in \mathscr{L}^2(\pi) \text{ with } \mu(E) = 0, \\ &\|\mu P(t)\|_{\mathscr{L}^2(\pi)} \leqslant \|\mu\|_{\mathscr{L}^2(\pi)} e^{-\varepsilon t}, \qquad t \geqslant 0. \end{aligned} \tag{6.1}$$

Proof of Theorem 1.2. Because of the above discussions, we need only to show that (1.6) $\Longleftrightarrow$ (6.1).

(6.1) $\Longrightarrow$ (1.6). The proof is very much the same as the proof of Theorem 3.1. Note that

$$\|\mu\|_{\mathrm{Var}} \leqslant \|\mu\|_{\mathscr{L}^1(\pi)} \leqslant \|\mu\|_{\mathscr{L}^2(\pi)}.$$

By (6.1), we have for every probability measure $\mu \in \mathscr{L}^2(\pi)$,

$$\|\mu P(t) - \pi\|_{\mathrm{Var}} = \|(\mu - \pi)P(t)\|_{\mathrm{Var}} \leqslant \|\mu - \pi\|_{\mathscr{L}^2(\pi)} e^{-\varepsilon t}.$$

This gives (1.6) with

$$C_\mu = \|\mu - \pi\|_{\mathscr{L}^2(\pi)} = \left(\|\mu\|_{\mathscr{L}^2(\pi)}^2 - 1\right)^{1/2}.$$

(1.6) $\Longrightarrow$ (6.1). By using spectral theory of bounded self-adjoint operators, it was proved in Roberts and Rosenthal (1997), Theorem 2.1 that for a reversible Markov operator P, the following statements are equivalent.

(i) There is $\rho < 1$ such that for every signed measure $\mu \in \mathscr{L}^2(\pi)$ with $\mu(E) = 0$, $\|\mu P\|_{\mathscr{L}^2(\pi)} \leqslant \rho \|\mu\|_{\mathscr{L}^2(\pi)}$.

(ii) There is $\rho < 1$ such that for every probability measure $\mu \in \mathscr{L}^2(\pi)$, there is $C_\mu < \infty$ such that $\|\mu P^n - \pi\|_{\mathrm{Var}} \leqslant C_\mu \rho^n$ for all $n \geqslant 1$.

Now, we fix $t > 0$. From (1.6), it follows that assertion (ii) holds with $P = P(t)$ (and then $P^n = P(nt)$ for all n) and $\rho = e^{-\varepsilon t}$. Therefore, assertion (i) holds with the same P and ρ. That is (6.1). $\square$

Acknowledgements. The author is greatly indebted to Jennifer Lynn Steichen and the referees, whose very careful comments on the earlier versions of the paper and whose introduction of Roberts and Rosenthal (1997) greatly improved the quality of the paper.

References

[1] Anderson, W. J., 1991, *Continuous-Time Markov Chains*, Springer Series in Statistics, Springer, Berlin.

[2] Chen, M. F., 1992, *From Markov Chains to Non-Equilibrium Particle Systems*, World Scientific, Singapore.

[3] Chen, M. F., 1996, *Estimation of spectral gap for Markov chains*, Acta Math. Sinica. New Ser. 12(4), 337–360.

[4] Chen, M. F., 1997, *Coupling, spectral gap and related topics*, (I): Chinese Sci. Bulletin, 42(14), 1472–1477 (Chinese Edition); 42(16), 1321–1327 (English Edition). (Ⅱ): 42(15), 1585–1591 (Chinese Edition); 42(17), 1409–1416 (English Edition). (Ⅲ): 42(16), 1696–1703 (Chinese Edition); 1997, 42(18), 1497–1505 (English Edition).

[5] Chen, M. F., 1998, *Estimate of exponential convergence rate in total variation by spectral gap*, Acta Math. Sinica. New Ser. 14(1), 9–16.

[6] Chen, M. F., 1999, *Analytic proof of dual variational formula for the first eigenvalue in dimension one*, Sci. China (A) 29(4), 327–336 (Chinese Edition); 42(8), 857–866 (English Edition).

[7] Chen, M. F. and Wang, F. Y., 1998, *Cheeger's inequalities for general symmetric forms and existence criteria for spectral gap*, (Ann. Probab., to appear, **Abstract**). Chinese Sci. Bull. 1998, 43(18), 1516–1519.

[8] Down, D., Meyn, S. P. and Tweedie, R. L., 1995, *Exponential and uniform ergodicity of Markov processes*, Ann. Probab. 23, 1671–1691.

[9] Landim, C., Sethuraman, S. and Varadhan, S., 1996, *Spectral gap for zero-range dynamics*, Ann. Probab. 24(4), 1871–1902.

[10] Lund, R. B, Meyn, S. P. and Tweedie, R. L., 1996, *Computable exponential convergence rates for stochastically ordered Markov processes*, Ann. Appl, Probab. 6(1), 218–237.

[11] Meyn, S. P. and Tweedie, R. L., 1993, *Markov Chains and Stochastic Stability*, Springer, London.

[12] Nummelin, E., 1984, *General Irreducible Markov Chains and Non-negative Operators*, Cambridge Univ., Press, Cambridge.

[13] Roberts, G. O. and Rosenthal, J. S., 1997, *Geometric ergodicity and hybrid Markov chains*, Electron. Comm. Probab. 2, 13–25.

[14] Sokal, A. D. and Thomas, L. E., 1989, *Exponential convergence to equilibrium for a class of random-walk models*, J. Statis. Phys. 54, Nos.3/4, 797–828.

Appendix (Unpublished). Proofs of the Lemmas used in the proof of Theorem 4.1

Proposition A.1. Let X be a complex Banach space, A be a bounded linear operator on X with spectrum $\sigma(A)$. Then

$$\begin{aligned} r(A) &:= \sup\{|\lambda| :\ \lambda \in \sigma(A)\}\\ &= \lim_{n\to\infty} \|A^n\|^{1/n} = \inf \|A^n\|^{1/n}\\ &= \sup_{x\in X} \varlimsup_{n\to\infty} \|A^n x\|^{1/n}\\ &= \sup_{x\in X,\, \ell\in X^*} \varlimsup_{n\to\infty} |\langle \ell,\, A^n x\rangle|^{1/n}. \end{aligned}$$

When X is a Hilbert space, we have $r(A) = \sup_{x\in X} \varlimsup_{n\to\infty} (x,\, A^n x)^{1/n}$.

Proof. (a) The second line is the well-known Gelfand Theorem, refer to [Xia D. X. et al (1979), Real Analysis and Functional Analysis (2nd Edition), Vol. 2, Press of Higher Education, p. 147, Theorem 10 and p. 289, Theorem 3].

(b) Because

$$\varlimsup_{n\to\infty} \|A^n x\|^{1/n} \leqslant \lim_{n\to\infty} \|A^n\|^{1/n} \lim_{n\to\infty} \|x\|^{1/n} = \lim_{n\to\infty} \|A^n\|^{1/n},$$

we have $\sup_{x\in X} \varlimsup_{n\to\infty} \|A^n x\|^{1/n} \leqslant r(A)$.

Conversely, fix λ such that

$$\lambda > c(A) := \sup_{x\in X} \varlimsup_{n\to\infty} \|A^n x\|^{1/n}.$$

Set $B = A/\lambda$. Then, for every fixed $x\in X$, $\{B^n x : n \geqslant 1\}$ is bounded. By the Uniformly Boundedness Theorem (cf. the quoted book, p. 201), there is $M<\infty$ such that $\sup_n \|B^n\| \leqslant M$. Hence $\varlimsup_{n\to\infty} \|A^n\|^{1/n} \leqslant \lambda$. Because, $\lambda (> c(A))$ is arbitrary, we get $r(A) = \varlimsup_{n\to\infty} \|A^n\|^{1/n} \leqslant c(A)$.

(c) Note that

$$\begin{aligned} \varlimsup_{n\to\infty} |\langle \ell,\, A^n x\rangle|^{1/n} &\leqslant \varlimsup_{n\to\infty} [\|\ell\|^{1/n} \|A^n x\|^{1/n}]\\ &\leqslant \varlimsup_{n\to\infty} [(\|\ell\|\, \|x\|)^{1/n}\, \|A^n\|^{1/n}]\\ &= \varlimsup_{n\to\infty} \|A^n\|^{1/n}\\ &= r(A). \end{aligned}$$

On the other hand, let

$$\begin{aligned} \lambda > c(A) &:= \sup_{x\in X,\, \ell\in X^*} \varlimsup_{n\to\infty} |\langle \ell,\, A^n x\rangle|^{1/n}\\ &= \sup_{x\in X,\, \ell\in X^*,\, \|x\|=1,\, \|\ell\|=1} \varlimsup_{n\to\infty} |\langle \ell,\, A^n x\rangle|^{1/n}. \end{aligned}$$

Put $B = A/\lambda$. Then, for fixed $x \in X$ and $\ell \in X^*$, the set $\{\langle \ell, B^n x\rangle : n \geqslant 1\}$ is bounded. This means that for fixed $x \in X$, $\{B^n x : n \geqslant 1\}$ is weakly bounded and so is strongly bounded. Applying the Uniformly Boundedness Theorem again, it follows that $\sup_{n\geqslant 1} \|B^n\| < \infty$. But $\lambda > c(A)$ can be arbitrary, we obtain $\overline{\lim}_{n\to\infty} \|A^n\|^{1/n} \leqslant c(A)$.

(d) For Hilbert space, use the fact that

$$\begin{aligned}(x, A^n y) =&\frac{1}{4}\Big[\big(x+y, A^n(x+y)\big) - \big(x-y, A^n(x-y)\big)\Big]\\ &-\frac{i}{4}\Big[\big(x+iy, A^n(x+iy)\big) - \big(x-iy, A^n(x-iy)\big)\Big]\end{aligned}$$

and that $a+b+c+d \leqslant 4\max\{a,b,c,d\}$. It follows that

$$r(A) = \sup_{x,\,y\in X}\varlimsup_{n\to\infty} |(x,\, A^n y)|^{1/n} \leqslant \sup_{x\in X}\varlimsup_{n\to\infty} |(x,\, A^n x)|^{1/n}.$$

The inverse inequality is obvious. $\square$

Before moving further, we remark that for non-self-adjoint operator, one can not replace $\sup_{x\in X}$ by $\sup_{x\in D}$, where D is a dense set in X. An easy counterexample is as follows. Take $X = \ell^2(\mathbb{Z})$ and A the shift operator. However, the conclusion is true for self-adjoint operator.

Proposition A.2. Let X be a complex Hilbert space, A be a bounded, self-adjoint linear operator. Then for every dense set D, we have

$$\|A\| = r(A) = \sup_{x\in D}\varlimsup_{n\to\infty} (x,\, A^n x)^{1/n}.$$

Next, take $X = L^2_{\mathbb{C}}(\pi)$ and assume that $A\big(L^2_{\mathbb{R}}(\pi)\big) \subset L^2_{\mathbb{R}}(\pi)$. Then for every dense set D in $L^2_{\mathbb{R}}(\pi)$, the above formula still holds.

Proof. (a) It is well known that $\|A\| = r(A)$. Let E_A be the spectral projection of A. Given $\varepsilon > 0$, let

$$S_\varepsilon = [-r(A),\, -r(A)+\varepsilon) \cup (r(A)-\varepsilon,\, r(A)].$$

Since S_ε^c is closed, $E_A(S_\varepsilon) \neq 0$. Choose $x \in D$ such that $E_A(S_\varepsilon)(x) \neq 0$. Then for even n, we get

$$\begin{aligned}(x,\, A^n x) &= \int \lambda^n (x,\, E_A(\mathrm{d}\lambda)x)\\ &\geqslant (r(A)-\varepsilon)^n (x,\, E_A(S_\varepsilon)x)\\ &= (r(A)-\varepsilon)^n \|E_A(S_\varepsilon)x\|^2 > 0.\end{aligned}$$

Thus,

$$\varlimsup_{n\to\infty} |(x,\, A^n x)|^{1/n} \geqslant (r(A)-\varepsilon)\varlimsup_{n\to\infty} \|E_A(S_\varepsilon)x\|^{2/n} = r(A)-\varepsilon.$$

But ε can be arbitrary, this gives us $\overline{\lim}_{n\to\infty} |(x,\, A^n x)|^{1/n} \geqslant r(A)$.

(b) To prove the last assertion, note that every $x \in L^2_{\mathbb{C}}(\pi)$ can be expressed as $x = y + iz$ with $y, z \in L^2_{\mathbb{R}}(\pi)$ and $D + iD$ is dense in $L^2_{\mathbb{C}}(\pi)$. Moreover, $(x, A^n x) = (y, A^n y) + (z, A^n z)$. So,

$$|(x, A^n x)|^{1/n} \leqslant 2^{1/n}\big[|(y, A^n y)|^{1/n} \vee |(z, A^n z)|^{1/n}\big].$$

Thus, by (a), we obtain

$$r(A) = \sup_{x\in D+iD} \overline{\lim_{n\to\infty}} |(x, A^n x)|^{1/n} \leqslant \sup_{y\in D} \overline{\lim_{n\to\infty}} |(y, A^n y)|^{1/n}. \quad \square$$

Lemma A.3(Lemma 3.12 of [14]). Let P be a transition probability matrix on a countable set with reversible measure π. Suppose that P is nonnegative definite on $L^2(\pi)$.[2] Denote by P_D the matrix obtained by deleting one (say 0)-row and 0-column. Then the operator norm of P on $L^2(\pi) \setminus \{\text{constants}\}$ is less or equal to the norm of P_D on $L^2(\pi; E \setminus \{0\})$.

Proof. Given $\varphi \in L^2(\pi)$ with $\pi(\varphi) = 0$. Set $c = \varphi(0)$. Then

$$\begin{aligned}
0 &\leqslant (\varphi,\, P\varphi)_{L^2(\pi)}\\
&= \big(\varphi - c,\, P(\varphi - c)\big)_{L^2(\pi)} - c^2\\
&= \big(\varphi - c,\, P_D(\varphi - c)\big)_{L^2(\pi; E\setminus\{0\})} - c^2\\
&\leqslant \|P_D\|_{L^2(\pi;E\setminus\{0\})} \|\varphi - c\|^2_{L^2(\pi)} - c^2\\
&= \|P_D\|_{L^2(\pi;E\setminus\{0\})} \big(\|\varphi\|^2_{L^2(\pi)} + c^2\big) - c^2\\
&\leqslant \|P_D\|_{L^2(\pi;E\setminus\{0\})} \|\varphi\|^2_{L^2(\pi)}.
\end{aligned}$$

Note that the second step works only if a single point is deleted from E, even though the equality can be replaced by inequality.

Lemma A.4(Lemma 3.11 of [14]). The assumption is the same as in Lemma $A.3$. Denote by τ_0 the hitting time at 0. If there is $\bar{r} > 1$ such that $\mathbb{E}_x \bar{r}^{\tau_0} < \infty$ for all $x \in E$, then $\|P_D\|_{L^2(\pi;E\setminus\{0\})} \leqslant \bar{r}^{-1}$.

Proof. Note that $\big(P_D^n 1\big)(x) = \mathbb{P}_x[\tau_0 > n] \leqslant \bar{r}^{-n-1} \mathbb{E}_x \bar{r}^{\tau_0}$. For every ψ with compact support, we have

$$\begin{aligned}
\big|\big(\psi,\, P_D^n \psi\big)_{L^2(\pi;E\setminus\{0\})}\big| &\leqslant \big(|\psi|,\, P_D^n |\psi|\big)_{L^2(\pi;E\setminus\{0\})}\\
&\leqslant |\psi|^2_\infty \sum_{x\in\operatorname{supp}\psi} \pi(x) \big(P_D^n 1\big)(x)\\
&\leqslant \bar{r}^{-n-1} |\psi|^2_\infty \sum_{x\in\operatorname{supp}\psi} \pi(x) \mathbb{E}_x \bar{r}^{\tau_0}.
\end{aligned}$$

Thus, by Proposition A.2, we get

$$\|P_D\|_{L^2(\pi;E\setminus\{0\})} = \sup_{\psi:\,\psi \text{ has compact support}} \overline{\lim_{n\to\infty}} \big|\big(\psi,\, P_D^n \psi\big)_{L^2(\pi;E\setminus\{0\})}\big|^{1/n} \leqslant \bar{r}^{-1}. \quad \square$$

[2] Without this additional condition, $(\varphi,\, P\varphi)_{L^2(\pi)}$ can be negative. However, one may avoid this by replacing P with P^2.

Sciences in China (B), 2000, 30:3, 203–209 (Chinese Edition); (A)43:6 (English Edition), 601–608

LOGARITHMIC SOBOLEV INEQUALITY FOR SYMMETRIC FORMS

Mu-Fa Chen

(Dept. of Math., Beijing Normal University, Beijing 100875)
Received August 2, 1999.

Abstract Some estimates of logarithmic Sobolev constant for general symmetric forms are obtained in terms of new Cheeger's constants. The estimates can be sharp in some sense.

Keywords: Logarithmic Sobolev inequality, symmetric form, birth-death process

Let $(E, \mathscr{E}, \pi)$ be a measurable probability space satisfying $\{(x,x) : x \in E\} \in \mathscr{E} \times \mathscr{E}$ and denote by $L^2(\pi)$ the usual real L^2-space with norm $\|\cdot\|$. For a symmetric form $D(f,g)$ with domain $\mathscr{D}(D)$ on $L^2(\pi)$, logarithmic Sobolev inequality means that

$$\pi\big(f^2 \log f^2\big) \leqslant 2c^{-1} D(f,f), \qquad f \in \mathscr{D}(D), \quad \|f\| = 1 \tag{0.1}$$

for some constants c, where $\pi(f) = \int f \mathrm{d}\pi$. In what follows, the largest possible c is denoted by σ and is called the *logarithmic Sobolev constant*.

The inequality goes back to L. Gross (1976)[1]. It has attracted a great deal of research in the past two decades. The reader may refer to the survey articles[2],[3] for the present status of the study and also for references. Note that the most of the publications deal with diffusions and the inequality for (unbounded) Markov chains have been open for a long time. Very recently, the inequality has been studied in ref. [4] for finite Markov chains and in ref. [5] for general symmetric forms. The method used in the last paper is different from the previous one, that is the Cheeger's technique for unbounded operators developed in ref. [6]. We mention that the inequality does not hold for bounded operators in infinite spaces. The purpose of the note is to present some new or improved explicit lower bounds for the constant σ by using different proofs. See also the comment after the proof of Theorem 1.1 given in sec. 3 for a detailed comparison with ref. [5].

1 Main results. The symmetric form $(D, \mathscr{D}(D))$ considered here is as follows:

$$\begin{aligned} D(f,g) &= \frac{1}{2}\int J(\mathrm{d}x, \mathrm{d}y)[f(x)-f(y)]\,[g(x)-g(y)], \\ &\qquad f, g \in \mathscr{D}(D) := \{f \in L^2(\pi) : D(f,f) < \infty\}, \end{aligned} \tag{1.1}$$

where J is non-negative and symmetric: $J(\mathrm{d}x, \mathrm{d}y) = J(\mathrm{d}y, \mathrm{d}x)$. Without loss of generality, assume that $J(\{(x,x) : x \in E\}) = 0$.

Take and fix a non-negative, symmetric function $r \in \mathscr{E} \times \mathscr{E}$ such that

$$J^{(1)}(\mathrm{d}x, E)/\pi(\mathrm{d}x) \leqslant 1, \qquad \pi\text{-a.s.}, \tag{1.2}$$

where $J^{(\alpha)}(\mathrm{d}x, \mathrm{d}y) = I_{\{r(x,y)^\alpha > 0\}} J(\mathrm{d}x, \mathrm{d}y)/r(x,y)^\alpha$, $\alpha \geqslant 0$. Throughout the paper, we adopt the convention that $r^0 = 1$ for all $r \geqslant 0$. Correspondingly, we have symmetric forms $\big(D^{(\alpha)}, \mathscr{D}\big(D^{(\alpha)}\big)\big)$ generated by $J^{(\alpha)}$. Define

$$\begin{aligned}
\lambda_1^{(\alpha)} &= \inf\{D^{(\alpha)}(f,f) : \ \pi(f) = 0, \ \|f\| = 1\},\\
\kappa^{(\alpha)} &= \inf_{\pi(A)\in(0,1)} \frac{J^{(\alpha)}(A \times A^c)}{-\pi(A)\log \pi(A)},\\
\xi^\delta &= \inf_{\pi(A)>0} \frac{J^{(1/2)}(A \times A^c) + \delta\pi(A)}{\pi(A)\sqrt{1 - \log \pi(A)}}, \quad \delta > 0,\\
\xi_r &= \inf_{\pi(A)\in(0,r]} \frac{J^{(1/2)}(A \times A^c)}{\pi(A)\sqrt{1 - \log \pi(A)}}, \quad r \in (0,1)\\
\xi^\infty &= \lim_{\delta\to\infty} \xi^\delta = \sup_{\delta>0} \xi^\delta,\\
\xi_0 &= \lim_{r\to 0} \xi_r = \sup_{r>0} \xi_r.
\end{aligned}$$

When $\alpha = 0$, we return to the original form and so the superscript "(α)" is omitted everywhere from our notations. Noting that

$$-t\log t \leqslant -(1-t)\log(1-t)$$

on $[1/2, 1)$, one may replace "$\pi(A) \in (0,1)$" by "$\pi(A) \in (0, 1/2]$" in the definition of $\kappa^{(\alpha)}$. Next, since

$$-(\log t)/\sqrt{1 - \log t} \geqslant (\log 2)/\sqrt{1 + \log 2}$$

on $(0, 1/2]$, we have

$$\kappa^{(1/2)}(\log 2)/\sqrt{1 + \log 2} \leqslant \xi_{1/2} \leqslant \xi_0.$$

Now, the main results of the paper can be stated as follows.

Theorem 1.1. We have

$$\begin{aligned}
2\kappa &\geqslant \sigma\\
&\geqslant \frac{2\lambda_1}{1{+}16\inf_{\delta>0} A(\delta)}\\
&\geqslant \frac{2\lambda_1}{1{+}16\inf_{r\in(0,1)} B(r)}\\
&\geqslant \max\left\{\frac{2\lambda_1\chi_{(0,\infty)}(\xi_0)}{65 - 64\log r_0}, \ \frac{\xi_{1/2}^3}{25 + 33\xi_{1/2} + 11\xi_{1/2}^2}\right\},
\end{aligned}$$

where

$$A(\delta) = \frac{(2+\delta)(\lambda_1+\delta)}{(\xi^\delta)^2},$$
$$B(r) = \left[\frac{\lambda_1}{\xi_r} + \sqrt{1-\log r}\right]\left[\frac{2}{\xi_r} + \sqrt{1-\log r}\right]$$

and when $\xi_0 > 0$, r_0 is the unique solution to the equation:

$$\sqrt{1-\log r} = \frac{\lambda_1+2}{2\xi_r}, \qquad r \in (0,1).$$

Especially, we have $\sigma > 0$ whenever $\lambda_1 > 0$ and $\xi^\infty > 0$. Moreover,

$$\xi_0 > 0 \Longrightarrow \xi^\infty > 0,$$

and the inverse implication holds except $\mathscr{E}$ contains only a finite number of π-atoms.

Theorem 1.2. We have

$$2\kappa \geqslant \sigma \geqslant \frac{2\lambda_1\kappa^{(1/2)}}{\sqrt{\lambda_1(2-\lambda_1^{(1)}) + 3\kappa^{(1/2)}}} \geqslant \frac{1}{8}{\kappa^{(1/2)}}^2.$$

It is reasonable to keep $\lambda_1^{(\alpha)}$ in the above formulas since there are several ways to estimate the lower bound of $\lambda_1^{(\alpha)}$. For instance, by using the estimates of $\lambda_1^{(\alpha)}$ given in ref. [6], we obtain the last lower bound in each of the above theorems (cf. sec. 3 below). Because the lower bounds in Theorem 1.2 are meaningful iff $\kappa^{(1/2)} > 0$ which implies $\xi_{1/2} > 0$, it follows that Theorem 1.1 is better than Theorem 1.2 qualitatively. However, the lower bounds given by these two theorems are not comparable quantitatively.

To illustrate the application of the above results, we now consider the ergodic birth-death process with birth rates $b_i > 0\,(i \geqslant 0)$ and death rates $a_i > 0\,(i \geqslant 1)$. Then

$$\pi_0 = \frac{1}{\mu}, \quad \pi_i = \frac{b_0\cdots b_{i-1}}{a_1\cdots a_i\mu}, \quad \mu = 1 + \sum_{m=1}^\infty \frac{b_0\cdots b_{m-1}}{a_1\cdots a_m};$$

$J_{ij} = \pi_i b_i$ if $j = i+1$, $J_{ij} = \pi_i a_i$ if $j = i-1$ and $J_{ij} = 0$ for all other j. Take

$$r_{ij} = (a_i+b_i)\vee(a_j+b_j), \qquad i \neq j.$$

The two results in the next corollary are essentially due to ref. [5]. The example is not difficult to check by using the corollary and shows that Theorem 1.1 is truly stronger than Theorem 1.2.

Corollary 1.1. For birth-death process, the following assertions hold.

(1) $\xi_0 > 0$ iff

$$c := \inf_{i\geqslant 1} \frac{\pi_i a_i}{\sqrt{r_{i,i-1}}} \Big/ \Big(\sum_{j\geqslant i}\pi_j\Big)\sqrt{1-\log\sum_{j\geqslant i}\pi_j} > 0.$$

Moreover, we have $\xi_r \geqslant c$ for all $r < \pi_0$.

(2) $\kappa^{(\alpha)} > 0$ iff

$$\inf_{i \geqslant 1} \frac{\pi_i a_i}{r_{i,i-1}^{\alpha}} \Big/ \Big(-\sum_{j \geqslant i} \pi_j \Big) \log \sum_{j \geqslant i} \pi_j > 0.$$

Example 1.1. Take $a_i = b_i = i^2 \log^{\gamma}(i+1)$. Then $\kappa^{(1/2)} > 0$ iff $\gamma \geqslant 2$ and $\xi_0 > 0$ iff $\gamma \geqslant 1$. Moreover, logarithmic Sobolev inequality holds iff $\gamma \geqslant 1$.

Example 1.1 shows that Theorem 1.1 can be sharp qualitatively and can be better than Theorem 1.2. The next example shows that our estimates can also be good enough quantitatively. Moreover, the first lower bound of Theorem 1.2 can be better than Theorem 1.1.

Example 1.2. Let E be finite and $(\pi_i > 0)$ be an arbitrary distribution on E. Take

$$J_{ij} = \pi_i \pi_j \, (i \neq j), \qquad r_{ij} = (1-\pi_i) \vee (1-\pi_j) \, (i \neq j)$$

and put $\pi_* = \min \pi_i$. Then, the main estimates given by Theorem 1.1 and Theorem 1.2 have the same leading order $(-\log \pi_*)^{-1}$ (as $\pi_* \to 0$), which is exact.

Proof. It is simple to prove that $\lambda_1 = 1$.

a) Because

$$\begin{aligned} \xi^{\delta} &= \inf_{\pi(A)>0} \frac{J^{(1/2)}(A \times A^c) + \delta \pi(A)}{\pi(A)\sqrt{1 - \log \pi(A)}} \\ &\geqslant \inf_{\pi(A)>0} \frac{\delta}{\sqrt{1-\log \pi(A)}} \\ &\geqslant \frac{\delta}{\sqrt{1 - \log \pi_*}} \end{aligned}$$

and

$$\xi^{\delta} \leqslant \Big(\frac{\pi_*}{\sqrt{1-\pi_*}} \sum_{j \neq i_*} \pi_j + \delta \pi_* \Big) \Big/ \pi_* \sqrt{1 - \log \pi_*} = \frac{\delta + \sqrt{1-\pi_*}}{\sqrt{1-\log \pi_*}},$$

we have

$$\inf_{\delta > 0} \frac{(2+\delta)(\lambda_1 + \delta)}{(\xi^{\delta})^2} \leqslant (1 - \log \pi_*) \inf_{\delta>0} \frac{(2+\delta)(1+\delta)}{\delta^2} = 1 - \log \pi_*$$

and

$$\inf_{\delta > 0} \frac{(2+\delta)(\lambda_1 + \delta)}{(\xi^{\delta})^2} \geqslant (1 - \log \pi_*) \inf_{\delta>0} \frac{(2+\delta)(1+\delta)}{(\delta + \sqrt{1-\pi_*})^2} = 1 - \log \pi_*.$$

Hence, the best lower bound we can get from Theorem 1.1 is

$$\sigma \geqslant \frac{2}{1 + 16(1 - \log \pi_*)}.$$

b) Next, take $r = \pi_*$. Then for each i with $\pi_i = \pi_*$, we have $r_{ij} = 1 - \pi_i$ $(j \neq i)$. Moreover,

$$\xi_r = \frac{J^{(1/2)}(\{i\} \times \{i\}^c)}{\pi_i \sqrt{1 - \log \pi_i}} = \sqrt{\frac{1 - \pi_*}{1 - \log \pi_*}}.$$

Therefore, by using the second lower bound of Theorem 1.1, we obtain

$$\sigma \geqslant \frac{2(1 - \pi_*)}{1 - \pi_* + 16(1 - \log \pi_*)(1 + \sqrt{1 - \pi_*}\,)(2 + \sqrt{1 - \pi_*}\,)}.$$

c) Because $r_{ij} \leqslant 1 - \pi_*$,

$$\pi(A) > 0 \Longrightarrow \pi(A) \geqslant \pi_*$$

and $-(1 - t)/\log t$ is increasing on $(0, 1)$, we have

$$\kappa^{(1/2)} \geqslant \inf_{\pi(A) \in (0,1)} \frac{1 - \pi(A)}{-\sqrt{1 - \pi_*}\, \log \pi(A)} \geqslant \frac{\sqrt{1 - \pi_*}}{-\log \pi_*}.$$

Thus, applying to the first lower bound of Theorem 1.2, we obtain

$$\sigma \geqslant \frac{2\sqrt{1 - \pi_*}}{3\sqrt{1 - \pi_*} - \sqrt{2}\, \log \pi_*}.$$

Among these estimates, c) is better than a), which is better than b). This means that the first lower bound of Theorem 1.2 can be better than Theorem 1.1.

d) Following c), one deduces that $\kappa = -(1 - \pi_*)/\log \pi_*$.

Finally, comparing the estimates given in a)—d) with the precise result, Theorem A.1 in ref. [4]:

$$\sigma = \frac{2(1 - 2\pi_*)}{\log(\pi_*^{-1} - 1)},$$

one sees that each of the estimates has the exact leading order. However, the last estimate of Theorem 1.2 yields the order: $(-\log \pi_*)^{-2}$, which is not exact. $\square$

The remainder of the paper is organized as follows. In the next section, we introduce a more general result and complete its proof. The proofs of the theorems and the corollary are delayed to sec. 3.

2 A general result and its proof.

In this section, we consider the general symmetric form

$$\overline{D}(f, g) = \frac{1}{2} \int J(\mathrm{d}x, \mathrm{d}y)[f(x) - f(y)]\,[g(x) - g(y)] + \int K(\mathrm{d}x) f(x) g(x),$$
$$f, g \in \mathscr{D}(\overline{D}) := \{f \in L^2(\pi) : \overline{D}(f, f) < \infty\}, \tag{2.1}$$

where J is the same as in the last section and K is a non-negative measure. Again, choose a non-negative, symmetric function $\bar{r} \in \mathscr{E} \times \mathscr{E}$ and a non-negative function $s \in \mathscr{E}$ such that

$$[\bar{J}^{(1)}(\mathrm{d}x, E) + K^{(1)}(\mathrm{d}x)]/\pi(\mathrm{d}x) \leqslant 1, \qquad \pi\text{-a.s.}, \tag{2.2}$$

where $\bar{J}^{(\alpha)}(\mathrm{d}x, \mathrm{d}y)$ is defined in the same way as before but replacing r with $\bar{r}$ and

$$K^{(\alpha)}(\mathrm{d}x) = I_{\{s(x)^\alpha > 0\}} \frac{K(\mathrm{d}x)}{s(x)^\alpha},$$

$\alpha \geqslant 0$. Then, we have the form $(\overline{D}^{(\alpha)}, \mathscr{D}(\overline{D}^{(\alpha)}))$ generated by $(\bar{J}^{(\alpha)}, K^{(\alpha)})$ as in (2.1). Define

$$\lambda_0^{*(\alpha)} = \inf\{\overline{D}^{(\alpha)}(f, f) : \ \|f\| = 1\}.$$

Fix a continuous increasing function U on $[0, \infty)$ with $U(0) = 1$ such that

$$U' \text{ is piecewise continuous and } \ c_1 := \sup_{t \geqslant 0} \frac{tU'_\pm(t)}{U(t)} < \infty \tag{2.3}$$

where $U'_\pm$ denote the right- and left-derivatives of U. Next, define

$$\xi^* = \inf_{\pi(A) > 0} \frac{\bar{J}^{(1/2)}(A \times A^c) + K^{(1/2)}(A)}{\pi(A)\sqrt{U(\pi(A)^{-1})}}.$$

We can now state the last main result of the paper.

Theorem 2.1. Set

$$\sigma(U) = \inf\big\{\overline{D}(f, f)/\pi\big(f^2 U\big(f^2\big)\big) : \ \|f\| = 1\big\}.$$

Then, we have

$$\inf_{\pi(A) > 0} \frac{J(A \times A^c) + K(A)}{\pi(A)U(\pi(A)^{-1})} \geqslant \sigma(U) \geqslant \frac{\xi^{*2}}{4(1 + c_1)^2\big(2 - \lambda_0^{*(1)}\big)}.$$

Proof. Put $E_U(f) = \pi\big(f^2 U\big(f^2\big)\big)$. The first inequality is easy. Simply compute $\overline{D}(f, f)/E_U(f)$ for $f = I_A/\sqrt{\pi(A)}$ with $\pi(A) > 0$.

a) To prove the second inequality, we need more notations.

When $K(\mathrm{d}x) \neq 0$, it is convenient to enlarge the space E by letting $E^* = E \cup \{\infty\}$. For any $f \in \mathscr{E}$, define f^* on E^* by setting f: $f^* = f I_E$. Next, define $J^{*(\alpha)}$ on $E^* \times E^*$ by

$$J^{*(\alpha)}(C) = \begin{cases} \bar{J}^{(\alpha)}(C), & C \in \mathscr{E} \times \mathscr{E}, \\ K^{(\alpha)}(A), & C = A \times \{\infty\} \text{ or } \{\infty\} \times A, \ A \in \mathscr{E}, \\ 0, & C = \{\infty\} \times \{\infty\}. \end{cases}$$

Then, we have $J^{*(\alpha)}(\mathrm{d}x, \mathrm{d}y) = J^{*(\alpha)}(\mathrm{d}y, \mathrm{d}x)$ and

$$\begin{aligned}
&\int_E \bar{J}^{(\alpha)}(\mathrm{d}x, E) f(x)^2 + K^{(\alpha)}(f^2) = \int_{E^*} J^{*(\alpha)}(\mathrm{d}x, E^*) f^*(x)^2, \\
&\overline{D}^{(\alpha)}(f, f) = \frac{1}{2}\int_{E^* \times E^*} J^{*(\alpha)}(\mathrm{d}x, \mathrm{d}y)[f^*(y) - f^*(x)]^2, \\
&\frac{1}{2}\int_{E \times E} \bar{J}^{(\alpha)}(\mathrm{d}x, \mathrm{d}y)|f(y) - f(x)| + \int_E K^{(\alpha)}(\mathrm{d}x)|f(x)| \\
&\qquad = \frac{1}{2}\int_{E^* \times E^*} J^{*(\alpha)}(\mathrm{d}x, \mathrm{d}y)|f^*(y) - f^*(x)|.
\end{aligned}$$

Note that if we set $r^*(x,y)=\bar r(x,y)$, $r^*(x,\infty)=r^*(\infty,x)=s(x)$ for all x, $y\in E$ and $r^*(\infty,\infty)=0$, then $J^{*(\alpha)}$ can be also expressed by

$$J^{*(\alpha)}(\mathrm{d}x,\mathrm{d}y)=I_{\{r^*(x,y)^\alpha>0\}}J^*(\mathrm{d}x,\mathrm{d}y)/r^*(x,y)^\alpha.$$

b) Following ref. [5], take $\varphi(t)=t\sqrt{U(t)}$ and $\eta(t)=\varphi(t^2)$. Note that φ is a strictly increasing function and so is η. Moreover,

$$\varphi'(t)=\sqrt{U(t)}\left[1+\frac{tU'(t)}{2U(t)}\right]$$

and

$$\eta'(t)=\frac{2\eta(t)}{t}\left[1+\frac{t^2U'(t^2)}{2U(t^2)}\right]$$

except a finite number of points on each finite interval. By (2.3), we have

$$\eta'(t)\leqslant c_2\eta(t)/t=c_2t\sqrt{U(t^2)},\qquad c_2:=2+c_1.$$

Given $s<t$, label the discontinuous points in $[s,t]$ by $s=t_1<\cdots<t_m=t$. Then, by Mean Value Theorem, there exist $\theta_i\in(t_i,t_{i+1})$ such that

$$\begin{aligned}0&\leqslant\eta(t)-\eta(s)\\&=\sum_{i=1}^{m-1}[\eta(t_{i+1})-\eta(t_i)]\\&=\sum_{i=1}^{m-1}\eta'(\theta_i)(t_{i+1}-t_i)\leqslant\\&\leqslant c_2\sum_{i=1}^{m-1}\frac{\eta(\theta_i)}{\theta_i}(t_{i+1}-t_i)\\&\leqslant c_2\frac{\eta(t)}{t}\sum_{i=1}^{m-1}(t_{i+1}-t_i)\\&=c_2\frac{\eta(t)}{t}(t-s)\end{aligned}$$

since $\eta(t)/t$ is increasing.

c) Let $f\geqslant 0$, $\|f\|=1$ and set $g^*=\varphi\big(f^{*2}\big)$. Then by b), we have

$$|g^*(y)-g^*(x)|\leqslant c_2|f^*(y)-f^*(x)|\frac{\eta(f^*(x)\vee f^*(y))}{f^*(x)\vee f^*(y)}.$$

Thus, by Cauchy-Schwarz inequality and (2.2), we have

$$\begin{aligned}
I^* &:= \frac{1}{2}\int J^{*(1/2)}(\mathrm{d}x,\mathrm{d}y)|g^*(y)-g^*(x)| \\
&\leqslant \frac{c_2}{2}\bigg[\int J^*(\mathrm{d}x,\mathrm{d}y)|f^*(y)-f^*(x)|^2\bigg]^{1/2}\bigg[\int J^{*(1)}(\mathrm{d}x,\mathrm{d}y)\Big[f^*(y)\sqrt{U\big(f^*(y)^2\big)} \\
&\qquad\qquad + f^*(x)\sqrt{U\big(f^*(x)^2\big)}\,\Big]^2\bigg]^{1/2} \\
&= \frac{c_2}{\sqrt{2}}\sqrt{\overline{D}(f,f)}\bigg[\int J^{*(1)}(\mathrm{d}x,\mathrm{d}y)\Big[2f^*(y)^2U\big(f^*(y)^2\big)+2f^*(x)^2U\big(f^*(x)^2\big)\Big] \\
&\qquad\qquad - 2\overline{D}^{(1)}\Big(f\sqrt{U\big(f^2\big)},f\sqrt{U\big(f^2\big)}\,\Big)\bigg]^{1/2} \\
&\leqslant \frac{c_2}{\sqrt{2}}\sqrt{\overline{D}(f,f)}\Big[4E_U(f)-2\lambda_0^{*(1)}E_U(f)\Big]^{1/2} \\
&\leqslant c_2\sqrt{\big(2-\lambda_0^{*(1)}\big)\overline{D}(f,f)E_U(f)}.
\end{aligned}\tag{2.4}$$

d) Define $h(t)=\pi\big(f^{*2}>t\big)$,

$$A_t=\{g^*>t\}=\{\varphi(f^{*2})>t\}.$$

Then $h(t)\leqslant 1\wedge t^{-1}$,

$$\pi(A_t)=\pi\big(f^{*2}>\varphi^{-1}(t)\big)=h\circ\varphi^{-1}(t),$$

where φ^{-1} denotes the inverse function of φ. By definition of ξ^*, we have

$$\begin{aligned}
I^* &\geqslant \xi^*\int_0^\infty \pi(A_t)\sqrt{U\big(\pi(A_t)^{-1}\big)}\,\mathrm{d}t \\
&= \xi^*\int_0^\infty h(s)\sqrt{U\big(h(s)^{-1}\big)}\,\varphi'(s)\mathrm{d}s \\
&\geqslant \xi^*\int_0^\infty h(t)\sqrt{U(t)}\,\varphi'(t)\mathrm{d}t \\
&= \xi^*\int \mathrm{d}\pi\int_0^{f^{*2}}\sqrt{U(t)}\,\varphi'(t)\mathrm{d}t.
\end{aligned}$$

Next, by (2.3) and the absolute continuity of U, we have

$$\int_0^r\sqrt{U}\,\varphi'\geqslant c_3rU(r)\quad\text{for all } r\geqslant 0,$$

where

$$c_3=\frac{2+c_1}{2(1+c_1)}>\frac{1}{2}.$$

Hence,

$$I^* \geqslant c_3\xi^* \int \mathrm{d}\pi f^{*2} U\big(f^{*2}\big) = c_3\xi^* E_U(f). \tag{2.5}$$

e) Combining (2.4) with (2.5), we get

$$c_3\xi^* E_U(f) \leqslant c_2\sqrt{\big(2-\lambda_0^{*(1)}\big)\overline{D}(f,f)E_U(f)}.$$

That is,

$$\frac{\overline{D}(f,f)}{E_U(f)} \geqslant \frac{c_3^2\xi^{*2}}{c_2^2\big(2-\lambda_0^{*(1)}\big)} = \frac{\xi^{*2}}{4(1+c_1)^2\big(2-\lambda_0^{*(1)}\big)}. \qquad \square$$

3 Proofs of Theorems 1.1, 1.2 and Corollary 1.1.

Proof of Theorem 1.1. The upper bound follows from (0.1) by setting

$$f = I_A/\sqrt{\pi(A)}$$

with $\pi(A) \in (0,1)$.

a) Let (1.2) hold for some symmetric function r. Fix $\delta > 0$ and take $K(\mathrm{d}x) = \delta\pi(\mathrm{d}x)$. Next, take $\bar{r} = (1+\delta)r$ and $s = 1+\delta$ so that (2.2) holds. Finally, take

$$U(t) = 1 + \log^+ t = 1 + \max\{0, \log t\}.$$

Then we have

$$\xi^* = \frac{1}{\sqrt{1+\delta}}\,\xi^\delta, \qquad \lambda_0^{*(\alpha)} = \frac{\delta}{(1+\delta)^\alpha}, \qquad c_1 = \sup_{t>0}\frac{tU'_\pm(t)}{U(t)} = 1.$$

Therefore, for the symmetric form $(D, \mathscr{D}(D))$ given in (1.1), by Theorem 2.1, we obtain

$$\sigma(U) = \inf_{\|f\|=1}\frac{D(f,f)+\delta}{E_U(f)} \geqslant \frac{(\xi^\delta)^2/(1+\delta)}{16\big(2-\delta/(1+\delta)\big)} = \frac{(\xi^\delta)^2}{c_\delta},$$

where $c_\delta = 16(2+\delta)$. Thus,

$$\pi\big(f^2\log^+ f^2\big) \leqslant c_\delta\big(\xi^\delta\big)^{-2}[D(f,f)+\delta] - 1, \qquad \|f\| = 1. \tag{3.1}$$

By Proposition 3.10 in ref. [2], the following inequality

$$\pi\big(f^2\log f^2\big) \leqslant C_1 D(f,f) + C_2, \qquad \pi(f) = 0, \quad \|f\| = 1 \tag{3.2}$$

implies that

$$\sigma \geqslant 2/[C_1 + (C_2+2)\lambda_1^{-1}]. \tag{3.3}$$

Applying this to (3.1), it follows that

$$\sigma \geqslant \frac{2\lambda_1}{1 + c_\delta(\lambda_1+\delta)/(\xi^\delta)^2}. \tag{3.4}$$

This proves the first lower bound of σ since $\delta > 0$ can be arbitrary.

b) Fix $r \in (0,1)$. Note that

$$\xi^\delta \geqslant \xi_r \wedge \inf_{\pi(A)>r} \big[\delta/\sqrt{1-\log \pi(A)}\,\big] \geqslant \xi_r \wedge \big[\delta/\sqrt{1-\log r}\,\big].$$

Take $\delta_r = \xi_r\sqrt{1-\log r}$. We have $\xi^{\delta_r} \geqslant \xi_r$. Inserting this into (3.4), we get

$$\sigma \geqslant \frac{2\lambda_1}{1+16\big[(\lambda_1/\xi_r)+\sqrt{1-\log r}\,\big]\,\big[(2/\xi_r)+\sqrt{1-\log r}\,\big]}. \tag{3.5}$$

Since $r \in (0,1)$ is arbitrary, we obtain the second lower bound of σ. By using the inequality $(a+c)(b+c) \leqslant [c+(a+b)/2]^2$, one deduces the third lower bound of σ.

c) Define

$$k^{(\alpha)'} = \inf_{\pi(A)\in(0,1/2]} J^{(\alpha)}(A\times A^c)/\pi(A).$$

By ref. [6], we have

$$\lambda_1 \geqslant \frac{k^{(1/2)'^{\,2}}}{1+\sqrt{1-k^{(1)'^{\,2}}}}. \tag{3.6}$$

Note that $1-\log\pi(A) \geqslant 1+\log 2$ when $\pi(A)\leqslant 1/2$. We have

$$k^{(1/2)'} \geqslant \sqrt{1+\log 2}\,\xi_{1/2}.$$

Hence

$$\lambda_1 \geqslant (1+\log 2)\xi_{1/2}^2/2$$

by (3.6). Therefore, by (3.5), we obtain

$$\sigma \geqslant \frac{c^2\xi_{1/2}^3}{\xi_{1/2}+8c[2+c\xi_{1/2}]^2} \geqslant \frac{\xi_{1/2}^3}{25+33\xi_{1/2}+11\xi_{1/2}^2},$$

where $c=\sqrt{1+\log 2}$.

d) We have seen in b) that $\xi_r > 0 \Longrightarrow \xi^\delta > 0$. Conversely, by assumption, there is a sequence $\{A_n\}\subset \mathscr{E}$ such that $\pi(A_n)\to 0$. Hence

$$\begin{aligned}
\xi^\delta &\leqslant \inf_{\pi(A)\in(0,r]} \frac{J^{(1/2)}(A\times A^c)+\delta\pi(A)}{\pi(A)\sqrt{1-\log\pi(A)}}\\
&\leqslant \varliminf_{\pi(A)\to 0} \frac{J^{(1/2)}(A\times A^c)+\delta\pi(A)}{\pi(A)\sqrt{1-\log\pi(A)}}\\
&\leqslant \varliminf_{\pi(A)\to 0} \frac{J^{(1/2)}(A\times A^c)}{\pi(A)\sqrt{1-\log\pi(A)}} + \varlimsup_{\pi(A)\to 0} \frac{\delta}{\sqrt{1-\log\pi(A)}}\\
&= \lim_{r\to 0}\inf_{\pi(A)\in(0,r]} \frac{J^{(1/2)}(A\times A^c)}{\pi(A)\sqrt{1-\log\pi(A)}}\\
&= \xi_0.
\end{aligned}$$

Note that condition $\pi(A) \to 0$ is necessary in the first equality. Therefore, $\xi^\delta > 0 \Longrightarrow \xi_0 > 0$. The conlusion also holds when $\mathscr{E}$ contains only a finite number of π-atoms but still $\inf_{\pi(A)\in(0,1)} J^{(1/2)}(A \times A^c) > 0$ (which is somehow the irreducibility). $\square$

It is the position to mention the difference between this paper and ref. [5]. First, since $\inf U > 0$, for continuous space E, one may have $J^{(1/2)}(A \times A^c) \to 0$ as $\pi(A) \to 1$ and so $\lim_{r\to 1} \xi_r = 0$. This is avoided by the use of ξ^δ. Next, the goal of the proof is essentially (3.2) rather than (0.1). These two facts lead us to consider the form (2.1) with $K(\mathrm{d}x) \neq 0$ instead of (1.1). It then simplifies the original proof given in ref. [5] and enables us to work out the explicit lower bounds, and furthermore to make the comparison of Theorems 1.1 and 1.2. Note that one can even allow $\delta \to \infty$ at the final step to optimizing the resulting bound as illustrated by Example 1.2. Besides, the continuity of U' is relaxed to be piecewise continuous due to the use of $U(t) = 1 + \log^+ t$. Another useful function is $U_n = 1 + \log_n^+ t$, where $\log_1^+ t = 0 \vee \log t\, (t \geqslant 0)$ and $\log_m^+ = \log_1^+ \log_{m-1}^+$. Then $c_1 = 1/\prod_{k=0}^{n-1} e^{(k)}$, where $e^{(0)} = 1$ and $e^{(m)} = \exp[e^{(m-1)}]$. Moreover, Theorem 2.1 can also be extended to the more general setup studied in ref. [5].

Proof of Theorem 1.2. The proof is partially due to F. Y. Wang. Let $\pi(f) = 0$ and $\|f\| = 1$.

a) Set

$$\varepsilon = \sqrt{2 - \lambda_1^{(1)}}/[2\kappa^{(1/2)}]$$

and

$$E(f) = \pi\big(f^2 \log f^2\big).$$

Then, it can be proved that

$$E(f) \leqslant 2\varepsilon\sqrt{D(f,f)} + 1. \tag{3.7}$$

This is much easier to prove than Theorem 2.1. Actually, here we adopt $g = f^2$ instead of $g = \varphi(f^2)$ used there. First, one shows that

$$I := \frac{1}{2}\int J^{(1/2)}(\mathrm{d}x,\mathrm{d}y)\big|f(y)^2 - f(x)^2\big| \leqslant \sqrt{\big(2 - \lambda_1^{(1)}\big)D(f,f)}. \tag{3.8}$$

The proof is standard as used several times before (cf. ref. [6]). Next, set $A_t = \big\{f^2 > t\big\}$ and prove that

$$I \geqslant \kappa^{(1/2)}[E(f) - 1]. \tag{3.9}$$

This is also not hard (cf. Proof d) of Theorem 2.1). Combining (3.8) with (3.9), we get (3.7).

b) By (3.7), we have

$$E(f) \leqslant 2\varepsilon\sqrt{D(f,f)} + 1 \leqslant \gamma\varepsilon D(f,f) + \varepsilon/\gamma + 1,$$

where $\gamma > 0$ is a constant to be specified below. Combining this with (3.2) and (3.3), it follows that

$$\sigma \geqslant \frac{2}{\varepsilon\gamma + [\varepsilon/\gamma + 3]/\lambda_1}.$$

Maximizing the right-hand side with respect to γ, we get

$$\sigma \geqslant \frac{2\lambda_1 \kappa^{(1/2)}}{\sqrt{(2-\lambda_1^{(1)})\lambda_1} + 3\kappa^{(1/2)}}. \tag{3.10}$$

On the other hand, it was proved in ref. [6] that

$$\lambda_1^{(1)} \geqslant 1 - \sqrt{1 - {k^{(1)'}}^2}.$$

Combining this with (3.6) and noting that $k^{(1/2)'} \geqslant (\log 2)\kappa^{(1/2)}$, it follows that the right-hand side of (3.10) is bounded below by

$$\frac{2(\log 2)^2 {\kappa^{(1/2)}}^2}{(\log 2 + 3)\big[1 + \sqrt{1 - {k^{(1)'}}^2}\,\big]} \geqslant \frac{1}{8}{\kappa^{(1/2)}}^2. \quad \square$$

We remark that one may use $D(f,f) + \delta$ instead of $D(f,f)$ in the last proof as we did in the proof of Theorem 1.1. However, when optimizing the resulting estimate with respect to δ, one gets $\delta = 0$ and so the use of δ makes no improvement. The reason is that the constant used in Theorem 1.2 is $\kappa^{(1/2)}$ but not ξ^{δ}.

Proof of Corollary 1.1. Since $t\sqrt{1-\log t}$ is increasing on $(0,1)$ and

$$\pi(A) \leqslant r < \pi_0 \Longrightarrow 0 \notin A,$$

it follows that $i_0 := \inf A \geqslant 1$ and furthermore

$$\xi_r \geqslant \frac{\pi_{i_0} a_{i_0}}{\sqrt{r_{i_0, i_0 - 1}}} \Bigg/ \bigg(\sum_{j \geqslant i_0} \pi_j\bigg) \sqrt{1 - \log \sum_{j \geqslant i_0} \pi_j} \geqslant c.$$

For the proofs of the other assertions, refer to ref. [5]. $\square$

Acknowledegements This work was supported in part by National Natural Science Foundation of China (No. 19631060), Mathematical Tian Yuan Foundation, Qiu Shi Science & Technology Foundation, RFDP and MCEC.

References

[1]. Gross, L., *Logarithmic Sobolev inequalities*, Amer. J. Math. 1976, 97: 1061.
[2]. Bakry, D., *L'hypercontractivité et son utilisation en théorie des semigroupes*, LNM, **1581**, 1994, Springer, Berlin.
[3]. Gross, L., *Logarithmic Sobolev inequalities and contractivity properties of semigroups*, LNM **1563**, 1993, Springer-Verlag.
[4]. Diaconis, P. and Saloff-Coste, L., *Logarithmic Sobolev inequalities for finite Markov chains*, Ann. Appl. Prob. 1996, 6(3): 695.
[5]. Wang, F. Y., *Sobolev type inequalities for general symmetric forms*, to appear in Proc. Amer. Math. Soc. 2000.
[6]. Chen, M. F. Wang, F. Y., *Cheeger's inequalities for general symmetric forms and existence criteria for spectral gap*, to appear in Ann Prob, Abstract. Chin. Sci. Bulletin 1998, 43(18): 1516.

In “Applied Probability”, 25-34, eds. R. Chan et al.,
AMS/IP Studies in Advanced Mathematics, vol. 26, 2002

A NEW STORY OF ERGODIC THEORY

Mu-Fa Chen

(Beijing Normal University)

May 29, 1999

Abstract. In the recent years, a great effort has been made to develop a new ergodic theory for Markov processes. It is mainly concerned with the study on several different inequalities. Some of them are very classical but some of them are rather new. The Liggett-Stroock form of Nash-type inequalities, the related ones and their comparison are discussed. Based on some new isoperimetric or Cheeger's constants, a simple sufficient condition for the inequalities is reported. The resulting condition can be sharp qualitatively. Finally, a diagram of the inequalities and the traditional three types of ergodicity is presented.

The paper is divided into twelve short sections. In the first eleven sections, various currently interested inequalities are discussed. The relationship of the inequalities and the three types of traditional ergodicity is exhibited in the last section, which may be glanced over before reading the details of the paper.

1. Notations. Let $(E, \mathscr{E}, \pi)$ be a measure space with σ-finite non-negative measure π. Denote by $L^p(\pi)$ the usual real L^p-space with norm $\|\cdot\|_p$. Consider a symmetric form D on $L^2(\pi)$ with domain $\mathscr{D}(D)$. Two typical forms are the following:

$$\begin{aligned}&D(f) := D(f,f) = \frac{1}{2}\int_{\mathbb{R}^d} \langle a(x)\nabla f(x), \nabla f(x)\rangle \pi(\mathrm{d}x) + \int_{\mathbb{R}^d} c(x) f(x)^2 \pi(\mathrm{d}x),\\ &\mathscr{D}(D) \supset C_0^\infty(\mathbb{R}^d),\end{aligned}$$

where $\langle\cdot,\cdot\rangle$ is the standard inner product in $\mathbb{R}^d$, a is positive definite and $c \geqslant 0$;

$$\begin{aligned}&D(f) = \frac{1}{2}\int_{E\times E} J(\mathrm{d}x,\mathrm{d}y)[f(y)-f(x)]^2 + \int_E K(\mathrm{d}x) f(x)^2,\\ &\mathscr{D}(D) = \{f \in L^2(\pi) : D(f) < \infty\},\end{aligned} \tag{1}$$

Research supported in part by NSFC (No. 19631060), Math. Tian Yuan Found., Qiu Shi Sci. & Tech. Found., RFDP and MCME.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

where J is a non-negative symmetric measure having no charge on the diagonal $\{(x,x): x\in E\}$ and $K(\mathrm{d}x)$ is a non-negative measure. As usual,

$$D(f,g):=[D(f+g)-D(f-g)]/4.$$

A particular example in our mind is the symmetrizable jump process with q-pair $(q(x), q(x,\mathrm{d}y))$ and symmetrizing measure π, for which we have $J(\mathrm{d}x,\mathrm{d}y)=\pi(\mathrm{d}x)q(x,\mathrm{d}y)$. More especially, for a Q-matrix $Q=(q_{ij})$ with symmetrizing measure $(\pi_i>0)$, we have density $J_{ij}=\pi_i q_{ij}\ (j\neq i)$ and

$$K_i=-\pi_i\sum_j q_{ij}\geqslant 0$$

with respect to the counting measure.

2. Liggett-Stroock form of Nash-type inequalities[17]**.** The main inequality we are interested in is the following:

$$\|f\|^2\leqslant CD(f)^{1/p}V(f)^{1/q},\qquad f\in L^2(\pi), \tag{2}$$

where $\|f\|=\|f\|_2$, $C=C(p)$ is a constant and $1/p+1/q=1$ with $1\leqslant p<\infty$. Now, only the functional $V\geqslant 0$ has to be specified. The simplest case is that $p=1$ and hence $1/q=0$, then there is nothing to do about V and (2) is reduced to *Poincaré inequality* (1890):

$$\|f\|^2\leqslant CD(f),\qquad f\in L^2(\pi). \tag{3}$$

Thus, we may assume in what follows that $1<p<\infty$. Next, since both $\|f\|^2$ and $D(f)$ have degree two of homogeneous, it is natural to assume that $V(cf)=c^2V(f)$ for all constant c. However, if we take $V(f)=D(f)$ or $\|f\|^2$, then (2) is again reduced to (3) for all $p\in(1,\infty)$.

Next, one may look at the other L_r-norm: $V(f)=\|f\|_r^2$, $r\neq 2$. We may assume that $r\in[1,2)$ since the case of $r\in(2,\infty)$ can be reduced to the one of $r\in[1,2)$ by the symmetry of the form. When $r=1$, it is called *Nash inequality* (1958):

$$\|f\|^2\leqslant CD(f)^{1/p}\|f\|_1^{2/q},\qquad f\in L^2(\pi). \tag{4}$$

By setting $p=1+2/\nu\ (\nu>0)$, one gets the more familiar form of Nash inequality

$$\|f\|^{2+4/\nu}\leqslant CD(f)\|f\|_1^{4/\nu},\qquad f\in L^2(\pi).$$

Sometimes, the form $D(f)$ on the right-hand side of (4) is replaced by a new form $\widetilde{D}(f)=D(f)+\delta\|f\|^2$ for some $\delta\geqslant 0$. It is surprising but actually proved in [7] that for all $r\in[1,2)$, inequality (2) with $V(f)=\|f\|_r^2\ (r\in[1,2))$ is equivalent to (4) and hence we need only to consider (4).

Of course, there are many other choices of V:

$$V(f)=\sup_x|f(x)|^2,\qquad \sup_{x\neq x_0}\left|\frac{f(x)-f(x_0)}{\rho(x,x_0)}\right|^2,\qquad \sup_{x\neq y}\left|\frac{f(y)-f(x)}{\rho(x,y)}\right|^2=\mathrm{Lip}(f)^2,$$

where ρ is a distance and x_0 is a reference point in E. The last one was used by Liggett (1991) and we may call the corresponding inequality *Liggett inequality.*

3. Alternative form of (2). From now on, we often restrict ourselves to the case that π is a probability measure and the form $(D, \mathscr{D}(D))$ satisfies $D(1) = 0$. Then, the right-hand side of (2) becomes zero for constant function $f = 1$. Thus, it is necessary to make a change of the left-hand side of (2). For this, one simply uses the variation of f: $\text{Var}(f) = \pi(f^2) - \pi(f)^2$ instead of $\|f\|^2$, where $\pi(f) = \int f \text{d}\pi$. Then we obtain the alternative form of (2) as follows.

$$\text{L.S. inequality :} \qquad \text{Var}(f) \leqslant CD(f)^{1/p}V(f)^{1/q}, \qquad f \in L^2(\pi). \tag{5}$$

Certainly, this contains the alternative forms of the particular inequalities.

$$\text{Nash inequality :} \qquad \text{Var}(f) \leqslant CD(f)^{1/p}\|f\|_1^{2/q}.$$
$$\text{Poincaré inequality :} \qquad \text{Var}(f) \leqslant CD(f).$$
$$\text{Liggett inequality :} \qquad \text{Var}(f) \leqslant CD(f)^{1/p}\text{Lip}(f)^{2/q}.$$

In what follows, unless otherwise stated, we mainly deal with these inequalities.

4. The second class of inequalities. Keeping the right-hand side of (3) but making a change of the left-hand side, one gets the following inequality[24]

$$\left\{\int |f|^{2p/(p-1)} U\big(f^2/\|f\|^2\big)\text{d}\pi\right\}^{(p-1)/p} \leqslant CD(f), \qquad f \in L^2(\pi). \tag{6}$$

When $U = 1$, it is just *Sobolev inequality* (1936):

$$\|f\|^2_{2p/(p-1)} \leqslant CD(f), \qquad f \in L^2(\pi).$$

Since $2p/(p-1) \geqslant 2$, the inequality is stronger than the Poincaré one unless $p = \infty$. When $U = \log$ and $p = \infty$, (6) is *logarithmic Sobolev inequality* (L. Gross, 1975):

$$\text{LogS :} \qquad \int f^2 \log\big(f^2/\|f\|^2\big)\text{d}\pi \leqslant CD(f), \qquad f \in L^2(\pi). \tag{7}$$

The advantage of the last inequality is that it is a powerful tool in the study of infinite-dimensional analysis but not the Sobolev one.

5. Importance of the inequalities. Denote by $(P_t)_{t\geqslant 0}$ the semigroup generated by the form $(D, \mathscr{D}(D))$. Then we have the following fundamental result.

Theorem 1[17]**.** Let $V : L^2(\pi) \to [0, \infty]$ satisfy $V(c_1 f + c_2) = c_1^2 V(f)$ for all constants $c_1, c_2 \in \mathbb{R}$.

(1) Assume additionally that $V(P_t f) \leqslant V(f)$ for all $t \geqslant 0$ and $f \in L^2(\pi)$ (it is automatic when $V(f) = \|f\|_r^2$). If (5) holds, then

$$\text{Var}(P_t f)\big(= \|P_t f - \pi(f)\|^2\big) \leqslant CV(f)/t^{q-1}, \qquad t > 0. \tag{8}$$

(2) Conversely, (8) $\Longrightarrow$ (5).

Thus, inequality (5) describes L^2-algebraic convergence of the semigroup to its equilibrium state π. For Poincaré inequality, one indeed has L^2-exponential convergence:

$$\mathrm{Var}(f) \leqslant CD(f) \Longleftrightarrow \mathrm{Var}(P_t f) \leqslant \mathrm{Var}(f) e^{-2t/C}. \tag{9}$$

The smallest constant $C = \lambda_1^{-1}$,

$$\lambda_1 := \inf\{D(f) : \pi(f) = 0, \ \|f\| = 1\}$$

is called the *spectral gap* of the form $(D, \mathscr{D}(D))$. The similar results hold for (2) and (3).

Based on (8) and (9), inequalities (5) and (7) now consist of the main tools in the study of phase transitions and the effectiveness of random algorithms.

6. Relation of the above inequalities. There is a simple comparison between the above inequalities:

$$\text{Nash ineq.} \Longrightarrow \text{LogS ineq.} \Longrightarrow \text{Poincaré ineq.} \Longrightarrow \text{Liggett ineq.} \tag{10}$$

Here "$\Longrightarrow$" means "implies" as usual but the last implication needs a mild condition. We will come back to this comparison in the last section. We remark that for (5) with $V(f) = \|f\|_r^2$, there is a jump at $r = 2$: For each $r < 2$, it is stronger than logarithmic Sobolev inequality but at $r = 2$, it becomes suddenly weaker than logarithmic Sobolev inequality.

7. Methods. As far as we know, there are two general and powerful methods to handle these inequalities.

a) The probabilistic method—coupling method. It has been successfully applied to the Riemannian geometry, elliptic operators and Markov chains. Refer to the survey articles [5] for the present status of the study and also for a comprehensive list of publications.

b) The second powerful method comes from Riemannian geometry, which is the one we are going to discuss here.

8. Isoperimetry. A very ancient geometric result says that among the different regions with fixed length of boundary, the circle has the largest area. That is, for a region A with smooth boundary ∂A, we have

$$\frac{|\partial A|}{|A|^{1/2}} \geqslant \frac{2\pi r}{\sqrt{\pi r^2}} = 2\sqrt{\pi}, \tag{11}$$

where $|A|$ denotes the volume (length, area) of A. The higher-dimensional analog is also true. That is the following *isoperimetric inequality*:

$$\frac{|\partial A|}{|A|^{(d-1)/d}} \geqslant \frac{|S_{d-1}|}{|B_d|^{(d-1)/d}}, \tag{12}$$

where B_d is the d-dimensional unit ball and S_{d-1} is its surface. The right-hand side is called the *isoperimetric constant*. Refer to [2] for more details.

9. Cheeger's constants. It is well known that isoperimetric inequality plays a critical role in the study of Sobolev-type inequality. In 1970, Cheeger observed that the same idea can also be used to study Poincaré inequality. To do so, Cheeger introduced the so-called *Cheeger's constants*:

$$h = \inf_{A \subset M} \frac{|\partial A|}{|A|}; \qquad k = \inf_{A \cup B = M} \frac{|\partial (A \cap B)|}{|A| \wedge |B|} \tag{13}$$

for a compact manifold M. Comparing (13) with (12), it follows that the power $(d-1)/d$ disappears here. Cheeger established the following *Cheeger inequalities*:

$$D(f) \geqslant \frac{h^2}{4}\|f\|^2; \qquad D(f) \geqslant \frac{k^2}{4}\mathrm{Var}(f). \tag{14}$$

For the first one, the Dirichlet boundary is imposed[3], [16], [25].

The Cheeger's technique was used to study the estimate of spectral gap for jump processes. For instance, it was proved by Lawler and Sokal (1988)[15] that

$$\lambda_1 \geqslant \frac{k^2}{2M},$$

where $M = \sup_x q(x)$ and

$$k = \inf_{\pi(A) \in (0,1)} \frac{\int_A \pi(\mathrm{d}x) q(x, A^c)}{\pi(A) \wedge \pi(A^c)}.$$

As far as we know, in the past ten years or so, this result has been collected into six books [1], [4], [10], [11], [21] and [22]. From the titles of the books, one sees a wider range of applications of the study on spectral gap. The main problem is that the lower estimate vanishes when one passes to the unbounded operators. The problem has been open for more than ten years.

For logarithmic Sobolev inequality, there is a large number of publications in the context of diffusion processes in the past two decades or more. However, there was no result in the context of jump processes until Diaconis and Saloff-Coste's paper appeared in 1996[12]. They proved that for finite Markov chains, the logarithmic constant

$$\sigma := \inf \left\{ D(f) \Big/ \int f^2 \log[|f|/\|f\|] : \|f\| = 1 \right\}$$

satisfies

$$\sigma \geqslant \frac{2(1 - 2\pi_*)\lambda_1}{\log[1/\pi_* - 1]}$$

under the condition that $\sum_j |q_{ij}| = 1$ for all i, where $\pi_* = \min_i \pi_i$. Clearly, for infinite E, $\pi_* = 0$ and so the result is meaningless. This is also a challenge open problem ([5]; Problem 13). For Nash inequality, we are in the same situation (cf. [21]).

10. Main theorem. We are now glad to be able to report some answers to the above open problems. To do so, we need some new types of isoperimetric or Cheeger's constants.

For Poincaré ineq.[9] $$k^{(\alpha)} = \inf_{\pi(A)\in(0,1)} \frac{J^{(\alpha)}(A\times A^c)}{\pi(A)\wedge\pi(A^c)}$$

For Nash ineq.[7] $$k^{(\alpha)} = \inf_{\pi(A)\in(0,1)} \frac{J^{(\alpha)}(A\times A^c)}{[\pi(A)\wedge\pi(A^c)]^{(\nu-1)/\nu}}, \quad \nu = 2(q-1)$$

For LogS ineq.[24], [8] $$k^{(\alpha)} = \lim_{r\to 0}\inf_{0<\pi(A)\leqslant r} \frac{J^{(\alpha)}(A\times A^c)}{\pi(A)\sqrt{\log[e+\pi(A)^{-1}]}}$$

$$k^{(\alpha)} = \lim_{\delta\to\infty}\inf_{\pi(A)>0} \frac{J^{(\alpha)}(A\times A^c)+\delta\pi(A)}{\pi(A)\sqrt{1-\log\pi(A)}}$$

Here only $J^{(\alpha)}$ has not defined yet. Note that the original kernel J can be very unbounded. To avoid this, choose a symmetric function $r(x,y)$ so that

$$J^{(1)}(\mathrm{d}x, E)/\pi(\mathrm{d}x) \leqslant 1, \qquad \pi\text{-a.e.}, \tag{15}$$

where

$$J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y) = I_{\{r(x,y)>0\}}\frac{J(\mathrm{d}x,\mathrm{d}y)}{r(x,y)^\alpha}$$

for $\alpha\in(0,1]$ and $J^{(0)} = J$. For jump processes, one simply chooses $r(x,y) = q(x)\vee q(y)$. This key idea comes from [9]. Note that the second one is close to (12) and the first one is close to (13) since for manifolds, $|A| = \pi(A)$.

Having the constants at hand, it is a simple matter to state our main result.

Theorem 2. Let π be a probability measure. For the form given by (1) with $K(\mathrm{d}x) = 0$, if $k^{(1/2)} > 0$, then the corresponding inequality in (5) or (7) holds.

The theorem is proved in four papers [9], [7], [24] and [8] in which some explicit lower bounds in terms of $k^{(\alpha)}$ are also presented. We remark that even though the above condition $k^{(1/2)} > 0$ is in general not necessary but it can still be sharp qualitatively.

To give some impression about how the Cheeger's constants are related to the inequalities, we now sketch of the proof for the new type of Cheeger's inequalities which imply Poincaré ones.

11. Sketch of the proof. a) The first step is a simple observation about the set form and the functional form of the Cheeger's constants. Fix $B\in\mathscr{E}$ with $\pi(B)\in(0,1)$. We have

$$\begin{aligned} h_B^{(\alpha)} &= \inf_{A\subset B}\frac{J^{(\alpha)}(A\times A^c)}{\pi(A)} \quad \text{(set form)}\\ &= \inf\left\{\frac{1}{2}\int J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y)|f(y)-f(x)| : f\geqslant 0,\ f|_{B^c}=0,\ \|f\|_1 = 1\right\}\\ &\qquad \text{(functional form).}\end{aligned}$$

The proof is not hard. By taking $f = I_A\,(A \subset B)$, the previous one follows from the latter one. The other implication uses a co-area formula from geometry.

b) The next step is using Cauchy-Schwarz inequality. Let f satisfy $f|_{B^c} = 0$ and $\|f\| = 1$. From a) and condition (15), it follows that

$$\begin{aligned}
h_B^{(1)^2} &\leqslant \left\{\frac{1}{2}\int J^{(1)}(\mathrm{d}x, \mathrm{d}y)|f(y)^2 - f(x)^2|\right\}^2 \\
&\leqslant \frac{1}{2}D^{(1)}(f)\int J^{(1)}(\mathrm{d}x, \mathrm{d}y)[f(y) + f(x)]^2 \\
&= \frac{1}{2}D^{(1)}(f)\int J^{(1)}(\mathrm{d}x, \mathrm{d}y)\big[2(f(y)^2 + f(x)^2) - (f(y) - f(x))^2\big] \\
&\leqslant D^{(1)}(f)\big[2 - D^{(1)}(f)\big].
\end{aligned}$$

This gives us

$$D^{(1)}(f) \geqslant 1 - \sqrt{1 - h_B^{(1)^2}}.$$

c) By another use of Cauchy-Schwarz inequality and (15), we get

$$h_B^{(1/2)^2} \leqslant D(f)\big[2 - D^{(1)}(f)\big].$$

Combining the above two estimates, we finally obtain a new type of the first Cheeger's inequality:

$$D(f) \geqslant \frac{h_B^{(1/2)^2}}{1 + \sqrt{1 - h_B^{(1)^2}}} = \frac{h_B^{(1/2)^2}}{1 + \sqrt{1 - h_B^{(1)^2}}}\|f\|^2, \qquad f|_{B^c} = 0.$$

From the proof, the relation between the constants $h_B^{(1/2)}$, $h_B^{(1)}$ and inequality (3) should be clear. Define

$$\lambda_0(B) = \inf\{D(f) : f|_{B^c} = 0,\ \|f\| = 1\}.$$

Then we have proved that

$$\lambda_0(B) \geqslant \frac{h_B^{(1/2)^2}}{1 + \sqrt{1 - h_B^{(1)^2}}}. \tag{16}$$

d) Finally, we adopt another idea due to Cheeger: the splitting technique. That is

$$\lambda_1 \geqslant \inf_{0 \leqslant \pi(B) \leqslant 1/2} \lambda_0(B)$$

which is a key of the proof. Noting that

$$k^{(\alpha)} = \inf_{\pi(A) \in (0, 1/2]} h_B^{(\alpha)},$$

one can deduce from (16) another new type of the second Cheeger's inequality (compare with (14)):

$$D(f) \geqslant \frac{{k^{(1/2)}}^2}{1+\sqrt{1-{k^{(1)}}^2}}\mathrm{Var}(f).$$

That is,

$$\lambda_1 \geqslant \frac{{k^{(1/2)}}^2}{1+\sqrt{1-{k^{(1)}}^2}}.$$

Refer to [9], [7], [24] and [8] for details and further references.

Finally, we discuss the relationship between the above inequalities and the traditional ergodic theory.

12. New story of ergodic theory. For simplicity, we restrict ourselves to continuous-time, irreducible Markov chains with transition probability matrix $P(t) = (p_{ij}(t))$ on a countable state space E. The chain is called *ergodic* if there exists a distribution $(\pi_i > 0 : i \in E)$ such that

$$\lim_{t\to\infty} p_{ij}(t) = \pi_j, \qquad i, j \in E. \tag{17}$$

It is called *exponentially ergodic* if there is a constant $\varepsilon > 0$ such that for all $i, j \in E$, there exists a constant C_{ij} so that

$$|p_{ij}(t) - \pi_j| \leqslant C_{ij}e^{-\varepsilon t}, \qquad t > 0. \tag{18}$$

The chain is called *strongly ergodic* if

$$\lim_{t\to\infty} \sup_i |p_{ij}(t) - \pi_j| = 0. \tag{19}$$

These three types of ergodicity consist of the most common topics in the study of ergodic theory for Markov processes. It is known that strong ergodicity implies the uniformly exponential decay $\sup_i |p_{ij}(t) - \pi_j| \leqslant C_j e^{-\varepsilon t}$ for some constants C_j and $\varepsilon > 0$. Hence

$$\text{Strong ergodicity} \Longrightarrow \text{Exponential ergodicity} \Longrightarrow \text{Ordinary ergodicity}. \tag{20}$$

A question arises naturally: Are there any relationship between the three types of ergodicity and the inequalities discussed above? The answer is affirmative, even though these two classes of objects look like very different.

Now, the new story of ergodic theory can be summarized as the following diagram, which is the main new result of this paper.

Theorem 3. For reversible Markov chains, the following implications hold.

$$\begin{array}{ccccc}
 & & \text{Nash inequality} & & \\
 & \swarrow\!\!\swarrow & & \searrow\!\!\searrow & \\
\text{Logarithmic Sobolev inequality} & & & & \text{Strong ergodicity} \\
\Downarrow & & & & \Downarrow \\
\text{Poincaré inequality} & & \Longleftrightarrow & & \text{Exponential ergodicity} \\
 & & \Downarrow & & \\
 & & L^2\text{-algebraic ergodicity} & & \\
 & & \Downarrow & & \\
 & & \text{Ordinary ergodicity} & &
\end{array}$$

Here L^2-algebraic ergodicity means that (8) holds for some V satisfying the first assumption of Theorem 1 and $V(f) < \infty$ for all functions f with finite support.

Before moving on, let us make some remarks about the theorem. First, the most parts of the theorem also holds for general Markov processes under some mild assumption (see the proof given below for details). Next, the theorem is complete in the sense that all of the one-side implications "$\Longrightarrow$" can not be replaced by "$\Longleftrightarrow$". Besides, strong ergodicity and logarithmic Sobolev inequality are not comparable as will be shown soon. Thirdly, there are well known criteria for the three types of ergodicity in terms of Q-matrix, but none of them provides us any explicit "convergence rate". On the other hand, the study on the inequalities are often devoted to estimate the rates but up to now there is still no criterion for the inequalities in the publications. Thus, due to the equivalence given in the above theorem, the study on one side can benefit from the other. Finally, the inequalities are now powerful tools in infinite-dimensional situation, but the three types of ergodicity are more or less finite-dimensional objects. In any case, it is hoped that the diagram has made a meaningful change to the picture of ergodic theory.

The original purpose of the study on Liggett inequality is for L^2-algebraic convergence. However, the inequality depends heavily on the choice of the distance ρ. If

$$C_\rho := \int \pi(\mathrm{d}x)\pi(\mathrm{d}y)\rho(x,y)^2 < \infty,$$

then for every f with $\mathrm{Var}(f) = 1$, we have

$$1 = \frac{1}{2}\int \pi(\mathrm{d}x)\pi(\mathrm{d}y)[f(x) - f(y)]^2 \leqslant \frac{1}{2}C_\rho \mathrm{Lip}(f)^2.$$

Hence, replacing $V(f)$ with $\mathrm{Var}(f)$ on the right-hand side of (5), it follows that Poincaré inequality $\Longrightarrow$ Liggett inequality. This fact plus the assumptions of part (1) of Theorem 1 implies L^2-algebraic convergence with $V(f) = \mathrm{Lip}(f)^2$. Next, because $\mathrm{Lip}(I_{\{k\}}) = \sup_{j\neq k}\rho(k,j)^{-1}$, the last condition of Theorem 3 becomes $\inf\limits_{j:\, j\neq k}\rho(k,j) > 0$ for all $k \in E$. Certainly, if one uses

$$V(f) = \sup_{x\neq x_0}\left|\frac{f(x) - f(x_0)}{\rho(x,x_0)}\right|^2$$

instead of $V(f) = \mathrm{Lip}(f)^2$, the resulting conditions are different.

Proof of Theorem 3. First, we prove the implication: "Nash inequality $\Longrightarrow$ Strong ergodicity". Assume that Nash inequality holds. Then by (8), we have

$$\|P_t f - \pi(f)\| \leqslant C\|f\|_1/t^{(q-1)/2}$$

and so

$$\|(P_t - \pi)f\| \leqslant C\|f - \pi(f)\|_1/t^{(q-1)/2}.$$

This means that the operator norm $\|P_t - \pi\|_{1\to 2}$ as a mapping from $L^1(\pi)$ to $L^2(\pi)$ is bounded above by $C/t^{(q-1)/2}$. Because of the symmetry of $P_t - \pi$, we get

$$\|P_t - \pi\|_{1\to\infty} \leqslant \|P_t - \pi\|_{1\to 2}\|P_t - \pi\|_{2\to\infty} = \|P_t - \pi\|_{1\to 2}^2.$$

Hence

$$\begin{aligned}\sup_x \|P_t(x,\cdot) - \pi\|_{\mathrm{Var}} &= \sup_x \sup_{|f|\leqslant 1} |(P_t(x,\cdot) - \pi)f| \\ &\leqslant \sup_x \sup_{\|f\|_1\leqslant 1} |(P_t(x,\cdot) - \pi)f| \\ &= \|P_t - \pi\|_{1\to\infty} \\ &\leqslant C/t^{q-1} \to 0\end{aligned}$$

as $t \to \infty$. This proves strong ergodicity of $(P_t(x, \mathrm{d}y))$.

The implications "Nash inequality $\Longrightarrow$ Logarithmic Sobolev inequality" and "Logarithmic Sobolev inequality $\Longrightarrow$ Poincaré inequality" are proved in [7] and [14] respectively. The implications "Poincaré inequality $\Longleftrightarrow$ L^2-exponential convergence $\Longrightarrow$ L^2-algebraic convergence" are obvious by (9). All the above proofs work for general reversible Markov processes. The implication "Exponential ergodicity $\Longleftrightarrow$ Poincaré inequality" is proved in [6] for Markov chains but it was mentioned there the result should work in more general setup. However, it is known that the equivalence fails in the infinite-dimensional situation, for instance when there exist phase transitions. To see that "L^2-algebraic convergence $\Longrightarrow$ Ordinary ergodicity", simply note that

$$\pi_i|p_{ik}(t) - \pi_k|^2 \leqslant \sum_j \pi_j|p_{jk}(t) - \pi_k|^2 \leqslant CV(I_{\{k\}})/t^{q-1} \to 0$$

as $t \to \infty$.

Finally, one may replace "pointwise" by "total variation" in definition of the three types of ergodicity (these definitions are indeed equivalent to the ones given by (17)–(19) in the discrete case):

$$\begin{aligned}&\text{Ordinary ergodicity}: && \lim_{t\to\infty} \|p_t(x,\cdot) - \pi\|_{\mathrm{Var}} = 0\\ &\text{Exponential ergodicity}: && \|p_t(x,\cdot) - \pi\|_{\mathrm{Var}} \leqslant C(x)e^{-\varepsilon t}\\ &\text{Strong ergodicity}: && \lim_{t\to\infty} \sup_x \|p_t(x,\cdot) - \pi\|_{\mathrm{Var}} = 0.\end{aligned}$$

Then, the implications in (20) hold for general Markov processes when $\mathscr{E}$ is countably generated. In other words, if the Markov process corresponding to the semigroup (P_t) is irreducible and aperiodic in the Harris sense, then (20) holds. To see this, noting that by [4; §4.4] and [13], the continuous-time case can be reduced to the discrete-time one and then the conclusion follows from [18; Chapter 16]. □

We now show by some examples that strong ergodicity and logarithmic Sobolev inequality are not comparable. To do so, we need the following result taken from [26] and [27].

Theorem 4. For regular birth-death process with birth rates $b_i\,(i \geqslant 0)$ and death rates $a_i\,(i \geqslant 1)$, the process is strong ergodic iff

$$S := \sum_{n=1}^{\infty} a_{n+1}^{-1}\Big\{1 + \sum_{k=1}^{n} b_k \cdots b_n / a_k \cdots a_n\Big\} < \infty.$$

Example 5. Consider the birth-death process with either

(1) $b_i = a_i = i^2 \log^{\gamma} i\ (i \geqslant 1)$, $\gamma \in \mathbb{R}$ or

(2) $b_i = i^{\gamma}/2$ and $a_i = i^{\gamma}\ (i \geqslant 1)$, $\gamma \geqslant 0$ and $b_0 = 1$.

Then logarithmic Sobolev inequality holds iff $\gamma \geqslant 1$ and the process is strongly ergodic iff $\gamma > 1$.

Proof. The second assertion follows from Theorem 4. The first assertion is proved in [24] and [7; Examples 2.6 and 2.9]. The key idea is that for logarithmic Sobolev inequality of birth-death process, the constant $k^{(1/2)}$ used in Theorem 2 is equal to

$$\xi := \inf_{i \geqslant 1} \pi_i a_i \Big/ \bigg(\sum_{j \geqslant i} \pi_j\bigg) \sqrt{r_{i,i-1}\bigg(1 - \log \sum_{j \geqslant i} \pi_j\bigg)} > 0,$$

where $r_{ij} = (a_i + b_i) \vee (a_j + b_j)$. Noting that $\pi_i a_i = \pi_{i-1} b_{i-1}$, we have $\xi > 0$ iff

$$\sup_{k \geqslant 1} \frac{1}{\pi_k b_k} \bigg(\sum_{j \geqslant k+1} \pi_j\bigg) \bigg[r_{k+1,k}\bigg(1 - \log \sum_{j \geqslant k+1} \pi_j\bigg)\bigg]^{1/2} < \infty.$$

On the other hand, it is easy to confirm that $S < \infty$ iff

$$\sum_{k \geqslant 0} \frac{1}{\pi_k b_k} \sum_{j \geqslant k+1} \pi_j < \infty.$$

These facts indicate the relationship between "$S < \infty$" and "$\xi > 0$". $\square$

Example 6. Given a distribution $(\pi_i > 0)$ on a countable set E. Let $q_{ij} = \pi_j$ for all $j \neq i$ and $q_i = -q_{ii} = 1 - \pi_i$. Then the process is strongly ergodic but logarithmic Sobolev inequality fails.

Proof. It is proved in [24] that logarithmic Sobolev inequality does not hold for bounded operators in infinite space and so the second assertion follows. The proof is simply applying (7) to $f = I_A/\sqrt{\pi(A)}$ with $\pi(A) \in (0,1)$ and then letting $\pi(A) \to 0$.

To prove the first assertion, for simplicity, let $0 \in E$. Next, let $f_0 = 0$ and $f_i = 1/\pi_0\,(i \neq 0)$. Then

$$\pi(f) := \sum_j \pi_j f_j = \pi_0^{-1} - 1.$$

Hence

$$\sum_{j \neq i} q_{ij}(f_j - f_i) = \pi(f) - f_i = -1$$

for all $i \neq 0$ and

$$\sum_{j \neq 0} q_{0j}(f_j - f_0) = \pi(f) < \infty.$$

Because $0 \leqslant f \leqslant \pi_0^{-1}$, the first assertion now follows from the well know criterion for strong ergodicity (cf. [4; Part (4) of Theorem 4.45]). $\square$

References

1. Aldous, D. G. & Fill, J. A. (1994–), *Reversible Markov Chains and Random Walks on Graphs*, URL **www. stat.Berkeley.edu/users/aldous/book.html**.
2. Berger, E. and Gostiaus, B. (1988), *Differential Geometry: Manifolds, Curves, and Surfaces*, Springer-Verlag, New York.
3. Cheeger, J. (1970), *A lower bound for the smallest eigenvalue of the Laplacian*, Problems in analysis, a symposium in honor of S. Bochner 195–199, Princeton U. Press, Princeton.
4. Chen, M. F. (1992), *From Markov Chains to Non-Equilibrium Particle Systems*, World Scientific.
5. Chen, M. F. (1997), *Coupling, spectral gap and related topics*, (I): Chin. Sci. Bulletin, 42:14, 1472–1477 (Chinese Edition); 42:16, 1321–1327 (English Edition). (Ⅱ): 42:15, 1585–1591 (Chinese Edition); 42:17, 1409–1416 (English Edition). (Ⅲ): 42:16, 1696–1703 (Chinese Edition); 1997, 42:18, 1497–1505 (English Edition).
6. Chen, M. F. (2000), *Equivalence of exponential ergodicity and L^2-exponential convergence for Markov chains*, Stoch. Proc. Appl. 87, 281–297.
7. Chen, M. F. (1999), *Nash inequalities for general symmetric forms*, Acta Math. Sin. Eng. Ser. 15:3, 353–370.
8. Chen, M. F. (2000), *Logarithmic Sobolev inequality for symmetric forms*, Sci. in China (A) 30:3, 203–209 (Chinese Edition); 43:6 (English Edition), 601–608..
9. Chen, M. F. and Wang, F. Y. (1998), *Cheeger's inequalities for general symmetric forms and existence criteria for spectral gap*, Preprint. Abstract. Chin. Sci. Bulletin 43:14, 1475–1477 (Chinese Edition); 43:18, 1516–1519 (English Edition).
10. Chung, F. R. K. (1997), *Spectral Graph Theory*, CBMS, **92**, AMS, Providence, Rhode Island.
11. Colin de Verdière, Y. (1998), *Spectres de Graphes*, Soc. Math. France.
12. Diaconis, P. and Saloff-Coste, L., *Logarithmic Sobolev inequalities for finite Markov chains*, Ann. Appl. Prob. 1996, 695–750.
13. Down, D., Meyn, S. P. and Tweedie, R. L. (1995), *Exponential and uniform ergodicity of Markov processes*, Ann. Prob. 23, 1671–1691.
14. Gross, L. (1976), *Logarithmic Sobolev inequalities*, Amer. J. Math. 97, 1061–1083.
15. Lawler, G. F. and Sokal, A. D. (1988), *Bounds on the L^2 spectrum for Markov chain and Markov processes: a generalization of Cheeger's inequality*, Trans. Amer. Math. Soc. **309**, 557–580.
16. Li, P. (1993), *Lecture Notes on Geometric Analysis*, Seoul National Univ., Korea.
17. Liggett, T. M. (1991), *L_2 rates of convergence for attractive reversible nearest particle systems: the critical case*, Ann. of Prob. 19:3, 935–959.
18. Meyn, S. P. & Tweedie, R. L. (1993), *Markov chains and Stochastic Stability*, Springer-Verlag.
19. Nash, J. (1958), *Continuity of solutions of parabolic and elliptic equations*, Amer. J. Math. 80, 931–954.
20. Poincaré, H. (1890), *Sur les équations aux dérivées partielles de la physique mathématique*, Amer. J. Math. 12, 211–294.
21. Saloff-Coste, L. (1997), *Lectures on finite Markov chains*, LNM **1665**, 301–413, Springer-Verlag.
22. Sinclair, A. (1993), *Algorithms for Random Generation and Counting: A Markov Chain Approach*, Birkhäuser.

23. Sobolev, S. (1936), *On a theorem in functional analysis*, Mat. Sbornik 4 (1938), 471–497 (In Russian); AMS Translations, Ser. 2, 34 (1963), 39–68.
24. Wang, F. Y. (1998), *Sobolev type inequalities for general symmetric forms*, to appear in Proc. of Amer. Math. Soc.
25. Yau, S. T., Schoen, R. (1988), Differential Geometry, Science Press (In Chinese), Beijing, China.
26. Zhang, Y. H. (2001), *Strong ergodicity for continuous-time Markov chains*, to appear in J. Appl. Prob..
27. Zhang, H. J., Lin, X. and Hou, Z. T. (1998), *Uniformly polynomial convergence for standard transition functions*, preprint.

Received 27 July 1999, received in revised form from 4 November 1999, accepted 11 November 1999.

Department of Mathematics, Beijing Normal University, Beijing 100875, The People's Republic of China. E-mail: mfchen@bnu.edu.cn

Chin. Sci. Bulletin, 1999, 44(23), 2465–2470 (Chinese Ed.); 2000, 45:9 (English Ed.), 769–774

Eigenvalues, inequalities and ergodic theory

Mu-Fa Chen
(Beijing Normal University, Beijing 100875)

Abstract This paper surveys the main results obtained during the period 1992–1999 on three aspects mentioned at the title. The first result is a new and general variational formula for the lower bound of spectral gap (i.e., the first non-trivial eigenvalue) of elliptic operators in Euclidean space, Laplacian on Riemannian manifolds or Markov chains (§1). Here, a probabilistic method—coupling method is adopted. The new formula is a dual of the classical variational formula. The last formula is actually equivalent to Poincaré inequality. To which, there are closely related logarithmic Sobolev inequality, Nash inequality, Liggett inequality and so on. These inequalities are treated in a unified way by using Cheeger's method which comes from Riemannian geometry. This consists of §2. The results on these two aspects are mainly completed by the author joint with F. Y. Wang. Furthermore, a diagram of the inequalities and the traditional three types of ergodicity is presented (§3). The diagram extends the ergodic theory of Markov processes. The details of the methods used in the paper will be explained in a subsequent paper under the same title.

Keywords Eigenvalue inequality ergodic theory Markov process

1 New variational formula for the lower bound of spectral gap

1.1 Story of estimating λ_1 in geometry

We recall the study on λ_1 in geometry. From the story below, one should have some feeling about the difficulty of the hard mathematical topic.

Consider Laplacian Δ on a compact Riemannian manifold (M, g), where g is the Riemannian metric. The spectrum of Δ is discrete: $\cdots \leqslant -\lambda_2 \leqslant -\lambda_1 < -\lambda_0 = 0$ (may be repeated). Estimating these eigenvalues λ_k (especially λ_1) consists an important section and chapter of the modern geometry. As far as we know, until now, five books have been devoted to this topic. Here we list only the geometric books but ignore the ones on general spectral theory[1]—[5]. Denote by d, D and K respectively the dimension, the diameter and the lower bound of Ricci curvature ($\text{Ricci}_M \geqslant Kg$) of the manifold M. We are interested in estimating λ_1 in terms of these three geometric quantities. For an upper bound, it is relatively easy. Applying a test function $f \in C^1(M)$ to the classical variational formula

$$\lambda_1 = \inf\{\textstyle\int_M \|\nabla f\|^2 : \ f \in C^1(M), \ \int f\text{d}x = 0, \ \int f^2\text{d}x = 1\},$$

where "dx" is the Riemannian volume element, one gets an upper bound. However, the lower bound is much harder. The previous works have studied the lower estimates case by case by using different elegant methods. Eight of the most

Typeset by $\mathcal{A}_{\mathcal{M}}\mathcal{S}$-TEX

beautiful lower bounds are listed in the following table.

A. Lichnerowicz (1958)	$\dfrac{d}{d-1}\,K, \quad K \geqslant 0.$	(1)
P. H. Bérard, G. Besson & S. Gallot (1985)	$d\left\{\dfrac{\int_0^{\pi/2}\cos^{d-1}t\mathrm{d}t}{\int_0^{D/2}\cos^{d-1}t\mathrm{d}t}\right\}^{2/d}, \quad K = d-1 > 0.$	(2)
P. Li & S. T. Yau (1980)	$\dfrac{\pi^2}{2\,D^2}, \quad K \geqslant 0.$	(3)
J. Q. Zhong & H. C. Yang (1984)	$\dfrac{\pi^2}{D^2}, \quad K \geqslant 0.$	(4)
P. Li & S. T. Yau (1980)	$\dfrac{1}{D^2(d-1)\exp\left[1+\sqrt{1+16\alpha^2}\right]}, \quad K \leqslant 0.$	(5)
K. R. Cai (1991)	$\dfrac{\pi^2}{D^2} + K, \quad K \leqslant 0.$	(6)
H. C. Yang (1989) & F. Jia (1991)	$\dfrac{\pi^2}{D^2}e^{-\alpha}, \quad \text{if } d \geqslant 5, \quad K \leqslant 0.$	(7)
H. C. Yang (1989) & F. Jia (1991)	$\dfrac{\pi^2}{2\,D^2}e^{-\alpha'}, \quad \text{if } 2 \leqslant d \leqslant 4, \quad K \leqslant 0,$	(8)

where $\alpha = D\sqrt{|K|(d-1)}/2$, $\alpha' = D\sqrt{|K|((d-1)\vee 2)}/2$. All together, there are five sharp estimates ((1), (2), (4), (6) and (7)). The first two are sharp for the unit sphere in two- or higher-dimension but it fails for the unit circle; the fourth, the sixth and the seventh estimates are all sharp for the unit circle. The above authors include several famous geometers and the estimates were awarded several times. From the table, it follows that the picture is now very complete, due to the effort by the geometers in the past 40 years. For such a well-developed field, what can we do now? Our original starting point is to learn from the geometers, study their methods, especially the recent new developments. It is surprising that we actually went to the opposite direction, that is, studying the first eigenvalue by using a probabilistic method. It was indeed not dreamed that we could finally find a general formula.

1.2 New variational formula

To state the result, we need two notations

$$C(r) = \cosh^{d-1}\left[\frac{r}{2}\sqrt{\frac{-K}{d-1}}\right], \qquad r \in (0, D).$$
$$\mathscr{F} = \{f \in C[0, D] : f > 0 \text{ on } (0, D)\}.$$

Here the dimension d, the diameter D and the lower bound of Ricci curvature K have all been used.

Theorem [General formula] (Chen & Wang[6]).

$$\lambda_1 \geqslant \sup_{f\in\mathscr{F}}\inf_{r\in(0,D)}\frac{4f(r)}{\int_0^r C(s)^{-1}\mathrm{d}s\int_s^D C(u)f(u)\mathrm{d}u}.$$

The new variational formula has its essential value in estimating the lower bound. It is a dual of the classical variational formula in the sense that "inf" is replaced by "sup". The last formula goes back to Lord S. J. W. Rayleigh(1877) or E. Fischer (1905). Noticing that there are no common points in these two formulas, this explains the reason why such a formula never appeared before. Certainly, the new formula can produce a lot of new lower bounds. For instance, the one corresponding to the trivial function $f\mathscr{E}uiv1$ is still non-trivial in geometry. Next, let α be the same as above and let $\beta = \dfrac{\pi}{2D}$. Applying the formula to the test functions $\sin(\beta r)$, $\sin(\alpha r)$, $\sin(\beta r)$ and $\cosh^{d-1}(\alpha r)\sin(\beta r)$ successively, we obtain the following:

Corollary (Chen & Wang[6]).

$$\lambda_1 \geqslant \frac{\pi^2}{D^2} + \max\left\{\frac{\pi}{4d}, 1-\frac{2}{\pi}\right\}K, \qquad K \geqslant 0 \tag{9}$$

$$\lambda_1 \geqslant \frac{dK}{d-1}\left\{1-\cos^d\left[\frac{D}{2}\sqrt{\frac{K}{d-1}}\right]\right\}^{-1}, \qquad d>1, \quad K \geqslant 0 \tag{10}$$

$$\lambda_1 \geqslant \frac{\pi^2}{D^2} + \left(\frac{\pi}{2}-1\right)K, \qquad K \leqslant 0 \tag{11}$$

$$\lambda_1 \geqslant \frac{\pi^2}{D^2}\sqrt{1-\frac{2D^2K}{\pi^4}}\cosh^{1-d}\left[\frac{D}{2}\sqrt{\frac{-K}{d-1}}\right] \qquad d>1, \quad K \leqslant 0. \tag{12}$$

Comments.

(1) The corollary improves all the estimates (1)—(8). (9) improves (4); (10) improves (1) and (2); (11) improves (6); (12) improves (7) and (8).

(2) The theorem and corollary valid also for the manifolds with convex boundary with Neumann boundary condition. In this case, the estimates (1)—(8) are believed by geometers to be true. However, only the Lichnerowicz's estimate (1) was proved by J. F. Escobar until 1990. Except this, the others in (2)—(8) (and furthermore (9)—(12)) are all new in geometry[6].

(3) For more general non-compact manifolds, elliptic operators or Markov chains, we also have the corresponding dual variational formula[7],[8]. The point is that only three parameters d , D and K are used in the geometric case, but there are infinite parameters in the case of elliptic operators or Markov chains. Thus, the latter cases are more complicated. Actually, the above formula is a particular example of our general formula for elliptic operators. In dimensional one, our formula is complete.

(4) The probabilistic method—coupling method was developed by the present author before this work for more than ten years. The above study was the first time for applying the method to estimating the eigenvalues. For a long time, almost nobody believes that the method can achieve sharp estimate. From these facts, the influence of the above results to probability theory and spectral theory should be clear[8].

2 Basic inequalities and new forms of Cheeger's constants

2.1 Basic inequalities

Let $(E, \mathscr{E}, \pi)$ be a probability space satisfying $\{(x,x) : x \in E\} \in \mathscr{E} \times \mathscr{E}$. Denote by $L^p(\pi)$ the usual real L^p-space with norm $\|\cdot\|_p$. Write $\|\cdot\| = \|\cdot\|_2$. Our main object is a symmetric form $(D, \mathscr{D}(D))$ on $L^2(\pi)$. For Laplacian on manifold, the form used in the last part is the following

$$D(f) := D(f,f) = \int_M \|\nabla f\|^2 \mathrm{d}x, \qquad \mathscr{D}(D) \supset C^\infty(M).$$

Here, only the diagonal elements $D(f)$ is written, but the non-diagonal elements can be then deduced from the diagonal ones by using the quadrilateral role. The classical variational formula for spectral gap now can be rewritten into the following form.

$$\textit{Poincaré inequality}: \qquad \mathrm{Var}(f) \leqslant CD(f), \qquad f \in L^2(\pi)$$

where $\mathrm{Var}(f) = \pi(f^2) - \pi(f)^2$, $\pi(f) = \int f \mathrm{d}\pi$ and $C(= \lambda_1^{-1})$ is a constant. Thus, the study on the spectral gap is the same as the one on Poincaré inequality of the form $(D, \mathscr{D}(D))$. Nevertheless, we have more symmetric forms. For an elliptic operator in $\mathbb{R}^d$, the corresponding form is as follows.

$$D(f) = \frac{1}{2} \int_{\mathbb{R}^d} \langle a(x) \nabla f(x), \nabla f(x) \rangle \pi(\mathrm{d}x), \qquad \mathscr{D}(D) \supset C_0^\infty(\mathbb{R}^d),$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^d$ and $a(x)$ is positive definite. Corresponding to an integral operator (or symmetric kernel) on $(E, \mathscr{E})$, we have the symmetric form

$$\begin{gathered} D(f) = \frac{1}{2} \int_{E \times E} J(\mathrm{d}x, \mathrm{d}y)[f(y) - f(x)]^2, \\ \mathscr{D}(D) = \{f \in L^2(\pi) : D(f) < \infty\}, \end{gathered} \tag{13}$$

where J is a non-negative, symmetric measure having no charge on the diagonal set $\{(x,x) : x \in E\}$. A typical example in our mind is the reversible jump process with q-pair $(q(x),\, q(x, \mathrm{d}y))$ and reversible measure π. Then $J(\mathrm{d}x, \mathrm{d}y) = \pi(\mathrm{d}x) q(x, \mathrm{d}y)$. More especially, for a reversible Q-matrix $Q = (q_{ij})$ with reversible measure $(\pi_i > 0)$, we have density $J_{ij} = \pi_i q_{ij}$ $(j \neq i)$ with respect to the counting measure.

For a given symmetric form $(D, \mathscr{D}(D))$, except Poincaré inequality, there are also other basic inequalities.

$$\begin{aligned} &\textit{Nash inequality}: && \mathrm{Var}(f) \leqslant CD(f)^{1/p} \|f\|_1^{2/q}, && f \in L^2(\pi) \\ &\textit{Liggett inequality}: && \mathrm{Var}(f) \leqslant CD(f)^{1/p} \mathrm{Lip}(f)^{2/q}, && f \in L^2(\pi) \end{aligned}$$

where C is a constant and $\mathrm{Lip}(f)$ is the Lipschitz constant of f with respect to some distance ρ. The above three inequalities are actually particular cases of the following one

$$\textit{Liggett-Stroock inequality}: \qquad \mathrm{Var}(f) \leqslant CD(f)^{1/p} V(f)^{1/q}, \qquad f \in L^2(\pi)$$

where $V : L^2(\pi) \to [0, \infty]$ is homogeneous of degree two: $V(c_1 f + c_2) = c_1^2 V(F)$, c_1, $c_2 \in \mathbb{R}$. Another closely related one is
Logarithmic Sobolev inequality:

$$\int f^2 \log\left(f^2/\ |f\|^2\right) \mathrm{d}\pi \leqslant CD(f), \qquad f \in L^2(\pi).$$

2.2 Statue of the research

From now on, we restrict ourselves to the symmetric form (13) corresponding to integral operators. The question is under what condition on the symmetric measure J, the above inequalities hold. In contrast with the probabilistic method used in the last part, here we adopt Cheeger's method (1970) which comes from Riemannian geometry.

We call

$$\lambda_1 := \inf\{D(f) : \pi(f) = 0,\ \|f\| = 1\}$$

the *spectral gap* of the form $(D, \mathscr{D}(D))$. For bounded jump processes, the main known result is the following.

Theorem (Lawler & Sokal (1988)).

$$\lambda_1 \geqslant \frac{k^2}{2M},$$

where

$$k = \inf_{\pi(A) \in (0,1)} \frac{\int_A \pi(\mathrm{d}x) q(x, A^c)}{\pi(A) \wedge \pi(A^c)}, \qquad M = \sup_{x \in E} q(x).$$

In the past seven years, the theorem has been collected into six books[9]—[14]. From the titles of the books, one sees the wider range of the applications of the study. The problem is: the result fails for unbounded operator. Thus, it has been a challenge open problem in the past ten years or more to handle the unbounded situation.

As for logarithmic Sobolev inequality, there is a large number of publications in the past twenty years or more for differential operators. However, there was almost no result for integral operators until the next result appeared.

Theorem (Diaconis & Saloff-Coste (1996)). Let E be a finite set and

$$\sum_j |q_{ij}| = 1$$

holds for all i. Then the *logarithmic Sobolev constant*

$$\sigma := \inf\left\{D(f) \bigg/ \int f^2 \log[|f|/\|f\|] : \|f\| = 1\right\}$$

satisfies

$$\sigma \geqslant \frac{2(1 - 2\pi_*)\lambda_1}{\log[1/\pi_* - 1]},$$

where $\pi_* = \min_i \pi_i$.

Obviously, the result fails again for infinite E. The problem is due to the limitation of the method used in the proof.

2.3 New result

Corresponding to three inequalities, we introduce respectively the following new forms of Cheeger's constants.

Inequality	Constant $k^{(\alpha)}$	
Poincaré	$\inf_{\pi(A)\in(0,1)} \frac{J^{(\alpha)}(A\times A^c)}{\pi(A)\wedge\pi(A^c)}$	(Chen & Wang[15])
Nash	$\inf_{\pi(A)\in(0,1)} \frac{J^{(\alpha)}(A\times A^c)}{[\pi(A)\wedge\pi(A^c)]^{(\nu-1)/\nu}}$ $\nu = 2(q-1)$	(Chen[16])
Log. Sobolev	$\lim_{r\to 0}\inf_{\pi(A)\in(0,r]} \frac{J^{(\alpha)}(A\times A^c)}{\pi(A)\sqrt{\log[e+\pi(A)^{-1}]}}$	(Wang[17])
	$\lim_{\delta\to\infty}\inf_{\pi(A)>0} \frac{J^{(\alpha)}(A\times A^c)+\delta\pi(A)}{\pi(A)\sqrt{1-\log\pi(A)}}$	(Chen[18])

where $r(x,y)$ is a symmetric, non-negative function such that

$$J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y) := I_{\{r(x,y)>0\}}\frac{J(\mathrm{d}x,\mathrm{d}y)}{r(x,y)^\alpha}\ (\alpha>0)$$

satisfies

$$\frac{J^{(1)}(\mathrm{d}x,E)}{\pi(\mathrm{d}x)} \leqslant 1, \qquad \pi\text{-a.s.}$$

For convenience, we use the convention $J^{(0)} = J$. Now, our main result can be easily stated as follows.

Theorem. $k^{(1/2)} > 0 \Longrightarrow$ the corresponding inequality holds.

The result is proved in four papers [15]—[18]. At the same time, some estimates for the upper or lower bounds are also presented. These estimates can be sharp or qualitatively sharp, which did not happen before in using Cheeger's technique.

3 New picture of ergodic theory

3.1 Importance of the inequalities

Let $(P_t)_{t\geqslant 0}$ be the semigroup determined by the symmetric form $(D, \mathscr{D}(D))$. Then, various applications of the inequalities are based on the following result.

Theorem.

(1) Let $V(P_t f) \leqslant V(f)$ for all $t \geqslant 0$ and $f \in L^2(\pi)$ (which is automatic when $V(f) = \|f\|_r^2$). Then Liggett-Stroock inequality implies that

$$\mathrm{Var}(P_t f) \leqslant CV(f)/t^{q-1}, \qquad t > 0. \tag{14}$$

(2) Conversely, (14) $\Longrightarrow$ Liggett-Stroock inequality.

(3) Poincaré inequality $\Longleftrightarrow$ $\mathrm{Var}(P_t f) \leqslant \mathrm{Var}(f)\exp[-2\lambda_1 t]$.

Note that $\mathrm{Var}(P_t f) = \|P_t f - \pi(f)\|^2$. Therefore, the above inequalities describe some type of L^2-ergodicity of the semigroup $(P_t)_{t\geqslant 0}$. In particular, we call (14) L^2-*algebraic convergence*. These inequalities have become powerful tools in the study on infinite-dimensional mathematics (phase transitions, for instance) and the effectiveness of random algorithms.

3.2 Three traditional types of ergodicity

In the study of Markov processes, the following three types of ergodicity are well known.

$$\begin{array}{ll} \textit{Ordinary ergodicity}: & \lim_{t\to\infty} \|p_t(x,\cdot) - \pi\|_{\mathrm{Var}} = 0 \\ \textit{Exponential ergodicity}: & \|p_t(x,\cdot) - \pi\|_{\mathrm{Var}} \leqslant C(x)e^{-\varepsilon t} \\ \textit{Strong ergodicity}: & \lim_{t\to\infty} \sup_x \|p_t(x,\cdot) - \pi\|_{\mathrm{Var}} = 0 \end{array}$$

where $p_t(x, \mathrm{d}y)$ is the transition function of the Markov process and $\|\cdot\|_{\mathrm{Var}}$ is the total variation norm. They obey the following relation: Strong ergodicity $\Longrightarrow$ Exponential ergodicity $\Longrightarrow$ Ordinary ergodicity. Now, it is natural to ask the following question. Does there exist any relation between the above inequalities and the traditional three types of ergodicity?

3.3 New picture of ergodic theory

Theorem[16],[19],[20]. For reversible Markov chains, we have the following diagram:

$$\begin{array}{ccc} & \text{Nash inequality} & \\ \swarrow & & \searrow \\ \text{Log. Sobolev inequality} & & \text{Strong ergodicity} \\ \Downarrow & & \Downarrow \\ \text{Poincaré inequality} & \Longleftrightarrow & \text{exponential ergodicity} \\ & \Downarrow & \\ & L^2\text{-algebraic ergodicity} & \\ & \Downarrow & \\ & \text{Ordinary ergodicity} & \end{array}$$

where L^2-*algebraic ergodicity* means that (14) holds for some V having the properties: V is homogeneous of degree two, $V(f) < \infty$ for all functions f with finite support.

Comments.

(1) The diagram is complete in the following sense. Each single-side implication can not be replaced by double-sides one. Moreover, strong ergodicity and logarithmic Sobolev inequality are not comparable.

(2) The application of the diagram is obvious. For instance, one obtains immediately some criteria (which are indeed new) for Poincaré inequality to be held from the well-known criteria for the exponential ergodicity. On the other hand, by using the estimates obtained from the study on Poincaré inequality, one may estimate exponentially ergodic convergence rate (for which, the knowledge is still very limited).

(3) Except the equivalence, all the implications in the diagram are suitable for more general Markov processes. The equivalence in the diagram should be also suitable for more Markov processes but it may be false in the infinite-dimensional situation.
(4) No doubt, the diagram extends the ergodic theory of Markov processes.

Acknowledgement Research supported in part by NSFC (No. 19631060), Math. Tian Yuan Found., Qiu Shi Sci. & Tech. Found., RFDP and MCME.

References

[1]. Chavel I., *Eigenvalues in Riemannian Geometry*, New York: Academic Press, 1984.
[2]. Bérard P H, *Spectral Geometry: Direct and Inverse Problem*, **LNM.** vol 1207, New York: Springer-Verlag, 1986.
[3]. Yau S T, Schoen R, *Differential Geometry* (*In Chinese*), Beijing: Science Press, 1988.
[4]. Li P, *Lecture Notes on Geometric Analysis*, Seoul National Univ, Korea, 1993.
[5]. Ma C Y, *The Spectrum of Riemannian Manifolds* (*In Chinese*), Nanjing: Press of Nanjing U, 1993.
[6]. Chen M F, Wang F Y, *General formula for lower bound of the first eigenvalue*, Sci Sin, 1997, 40:4, 384–394.
[7]. Chen M F, Wang F Y, *Estimation of spectral gap for elliptic operators*, Trans Amer Math Soc, 1997, 349: 1239–1267.
[8]. Chen M F, *Coupling, spectral gap and related topics*, Chin Sci Bulletin, 1997, (I): 42:16, 1321–1327; (Ⅱ): 42:17, 1409–1416; (Ⅲ): 42:18, 1497–1505.
[9]. Chen M F, *From Markov Chains to Non-Equilibrium Particle Systems*, Singapore: World Scientific, 1992.
[10]. Sinclair A, *Algorithms for Random Generation and Counting: A Markov Chain Approach*, Boston: Birkhäuser, 1993.
[11]. Saloff-Coste L, *Lectures on finite Markov chains*, LNM **1665**, 301–413, New York: Springer-Verlag, 1997.
[12]. Chung F R K, *Spectral Graph Theory*, CBMS, **92**, Rhode Island: AMS, Providence, 1997.
[13]. Colin de Verdière Y, *Spectres de Graphes*, Paris: Publ Soc Math France, 1998.
[14]. Aldous D G & Fill J A (1994–), *Reversible Markov Chains and Random Walks on Graphs*, URL **www.stat.Berkeley.edu/users/aldous/book.html**.
[15]. Chen M F, Wang F Y, *Cheeger's inequalities for general symmetric forms and existence criteria for spectral gap*, **Abstract**. Chin Sci Bulletin, 1998, 43:18, 1516–1519. Ann. Prob. 2000, 28:1, 235–257.
[16]. Chen M F, *Nash inequalities for general symmetric forms*, Acta Math Sin Eng Ser, 1999, 15:3, 353–370.
[17]. Wang F Y, *Sobolev type inequalities for general symmetric forms*, to appear in Proc Amer Math Soc, 1999.
[18]. Chen M F, *Logarithmic Sobolev inequality for symmetric forms*, Sci Chin, 2000, 43:6, 601–608.
[19]. Chen M F, *Equivalence of exponential ergodicity and L^2-exponential convergence for Markov chains*, Stoch Proc Appl, 2000,87, 281–297.
[20]. Chen M F, *A new story of ergodic theory*, to appear in Proceedings of IMS Workshop on Applied Probability, Hong Kong: Intern. Press.

(Received August 2, 1999)

Advances in Mathematics (China) 28:6 (1999), 481–505

EIGENVALUES, INEQUALITIES AND ERGODIC THEORY (II)

Mu-Fa Chen

(Beijing Normal University)
Received: 1999-09-14

Abstract. The previous paper (Chen (1999e)) surveys the main results obtained during 1992–1999 on three aspects mentioned in the title. The present paper explains the main methods used in the proofs of the mentioned results. Mainly, there are two methods. One is a probabilistic method—coupling method; the other one is the Cheeger's method which comes from Riemannian geometry. Finally, a new picture of the ergodic theory is exhibited. This paper is rather self-contained. We choose some typical results with complete proofs and adopt the language as elementary as possible. The aim is to give the new comers a quick overview of the new progress and new ideas so that the readers may be easier to get into this active research field. Because the topics are quite wider, it is regretted that many beautiful results and a lot of references are missed, due to the limitation of the length of the paper. Parts of the materials given below has appeared in Chen (1994, 1997, 1998a).

Part I. Background

1. Definition. Consider a birth-death process with state space $E = \{0, 1, 2, \cdots\}$ and Q-matrix

$$Q = (q_{ij}) = \begin{pmatrix} -b_0 & b_0 & 0 & 0 & \dots \\ a_1 & -(a_1+b_1) & b_1 & 0 & \dots \\ 0 & a_2 & -(a_2+b_2) & b_2 & \dots \\ \vdots & \ddots & \ddots & \ddots & \end{pmatrix}$$

where $a_k,\ b_k > 0$. Since the sum of each row equals 0, we have $Q1 = 0 = 0\cdot 1$. This means that the Q-matrix has an eigenvalue 0 with eigenvector 1. Next, consider the finite case, $E_n = \{0, 1, \cdots, n\}$. Then, the eigenvalues of $-Q$ are discrete:

$$0 = \lambda_0 < \lambda_1 \leqslant \cdots \leqslant \lambda_n.$$

2000 *Mathematics Subject Classification.* 35P15, 47A75, 49R05, 60J99.

Key words and phrases. Eigenvalue, inequality, ergodic theory, Markov process.

Research supported in part by NSFC (No. 19631060), Math. Tian Yuan Found., Qiu Shi Sci. & Tech. Found., RFDP and MCME.

Typeset by $\mathcal{AMS}$-TEX

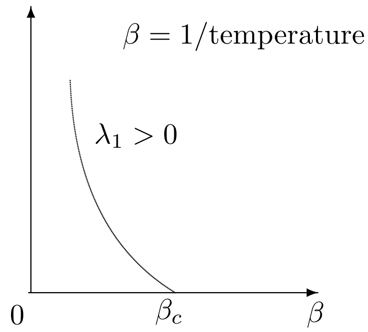
Hence, there is a gap between λ_0 and λ_1:

$$\operatorname{gap}(Q) := \lambda_1 - \lambda_0 = \lambda_1.$$

In the infinite case, the gap can be 0. Certainly, one can consider the self-adjoint elliptic operators in $\mathbb{R}^d$ or the Laplacian Δ on manifolds or an infinite-dimensional operator as in the study of interacting particle systems. In the last case, the operator depends on a parameter β. For different β, the system has completely different behavior.

2. Applications.

(1) Phase Transitions. In the study of interacting particle systems, a physical model is described by a Markov process with semigroup $\{P_t\}_{t\geqslant 0}$ (depending on temperature $1/\beta$) having stationary distribution π. Let $L^2(\pi)$ be the usual L^2-space with norm $\|\cdot\|$.

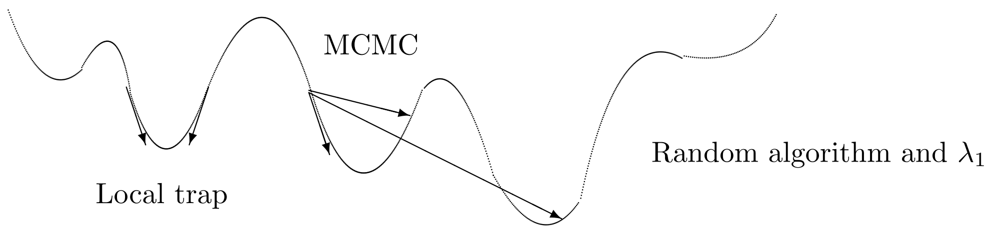


The picture means that in higher temperature (small β), the corresponding semigroup $\{P_t\}_{t\geqslant 0}$ is exponentially ergodic in the L^2-sense:

$$\|P_t f - \pi(f)\| \leqslant \|f - \pi(f)\| e^{-\lambda_1 t},$$

where $\pi(f) = \int f d\pi$, with the largest rate λ_1 and when the temperature goes to the critical value $1/\beta_c$, the rate will go to zero. This provides a way to describe the phase transitions and it is now an active research field. The next application we would like to mention is

(2) Markov chains Monte Carlo (MCMC). Consider a function with several local minimums. The usual algorithms go at each step to the place which decreases the value of the function. The problem is that one may pitfall into a local trap.



The MCMC algorithm avoids this by allowing some possibility to visit other places, not only towards to a local minimum. The random algorithm consists of two steps. a) Construct a distribution according to the local minimums in terms of Gibbs principle. b) Construct a Markov chain with the stationary distribution. The idea is great since it reduces some NP-problems to the P-problems in computer sciences. The effectiveness of a random algorithm is determined by λ_1 of the Markov chain. Refer to Sinclair (1993) for further information.

Since the spectral theory is a central part in each branch of mathematics and the first non-trivial eigenvalue is the leading term of the spectrum, it should not be surprising that the study of λ_1 has a very wider range of applications.

3. Difficulty.

We have seen the importance of the topic but it is extremely difficult. To get some concrete feeling, let us look at the following examples.
a) Consider birth-death processes with finite state space. It is trivial when $E = \{0, 1\}$, $\lambda_1 = a_1 + b_0$. The result is very nice since the increase of either a_1 or b_0 will increase λ_1. If we go one more step, $E = \{0, 1, 2\}$, then we have four parameters b_0, b_1 and a_1, a_2 only and

$$\lambda_1 = 2^{-1}\big[a_1 + a_2 + b_0 + b_1 - \sqrt{(a_1 - a_2 + b_0 - b_1)^2 + 4a_1 b_1}\big].$$

Now, the role for λ_1 played by the parameters becomes ambiguous.

Next, consider the infinite state space. Denote by g and $D(g)$ respectively the eigenfunction of λ_1 and the degree of g.

b_i	a_i	λ_1	$D(g)$
$i+1$	$2i$	1	1
$i+1$	$2i+3$	2	2

The change of the death rate from $2i$ to $2i + 3$ leads to the change of λ_1 from one to two. More surprisingly, the eigenfunction g is changed from linear to quadratic.
b) Consider diffusions with operator

$$L = a(x)\frac{d^2}{dx^2} + b(x)\frac{d}{dx}.$$

The state space for the first row below is the full line and for the last two rows is the half line $[0, \infty)$ with reflection boundary.

$a(x)$	$b(x)$	λ_1	$D(g)$
1	$-x$	1	1
1	$-x$	2	2
1	$-(x+1)$	3	3

From these, one sees that the eigenvalue λ_1 is very sensitive and the relation between λ_1 and the coefficients (a_i, b_i) or $(a(x), b(x))$ can not be very simple.

Part II. New Variational Formula, Coupling Method

In the first two sections below, we introduce some variational formulas of λ_1 for diffusion on half line or for birth-death processes. The mathematical tool used to derive the formulas is the couplings which are explained in Sections 3, 5 and 6. The key proof is sketched in Section 4.

1. Diffusion in half line. Consider the differential operator $L = a(x)\,\mathrm{d}^2/\mathrm{d}x^2 + b(x)\,\mathrm{d}/\mathrm{d}x$ on the interval $[0, D)$ ($D \leqslant \infty$) with Neumann boundary condition. Suppose that $a(x) > 0$ everywhere and $Z := \int_0^D \frac{\mathrm{d}x}{a(x)} \exp[C(x)] < \infty$, where $C(x) = \int_0^x b/a$. Set $\pi(\mathrm{d}x) = \frac{1}{Za(x)} \exp[C(x)]\mathrm{d}x$. On $L^2(\pi)$, the operator L has again a trivial eigenvalue $\lambda_0 = 0$, we are now interested in the nearest eigenvalue λ_1; that is, the smallest λ such that $Lf = -\lambda f$ for some non-constant f. A classical variational characterization (the Min-Max theorem) is as follows.

$$\lambda_1 = \inf\{D(f) : f \in C^1[0, D],\ \pi(f) = 0 \text{ and } \pi(f^2) = 1\}, \tag{2.1}$$

where

$$D(f) = \int_0^D a(x) f'(x)^2 \pi(\mathrm{d}x)$$

and $\pi(f) = \int f\mathrm{d}\pi$. Actually, this formula is valid in completely general situation (refer to [Chen (1992); Chapter 9] for instance). The formula is especially powerful for an upper estimate of λ_1 since every function f with $\pi(f) = 0$ and $\pi(f^2) = 1$ gives us an upper bound.

Theorem 2.1[Chen & Wang (1997b)]. Let

$$\mathscr{F} = \{f : f' > 0 \text{ on } (0, D) \text{ and } \pi(f) \geqslant 0\}.$$

Then, we have

$$\lambda_1 \geqslant \sup_{f\in\mathscr{F}} \inf_{x\in(0,D)} \left\{ \frac{e^{-C(x)}}{f'(x)} \int_x^D \frac{f(u)c^{C(u)}}{a(u)} \mathrm{d}u \right\}^{-1}. \tag{2.2}$$

Moreover, the equality holds once the equation $af'' + bf' = -\lambda_1 f$ has a non-constant solution $f \in C^2[0, D]$ with $f'(0) = 0$ and $f'(D) = 0$ when $D < \infty$.

The formula (2.2) is a dual of (2.1) in the sense that the "inf" in (2.1) is replaced with "sup" in (2.2). It is now quite easy to get a meaningful lower bound of λ_1 by applying (2.2) to a suitable test function $f \in \mathscr{F}$. Note that there is no common point between (2.1) and (2.2). This explains the reason why such a simple result has not appeared before even though the topic is treated in almost every textbook on differential equations.

2. Markov chains. Here we consider only a special case.

Theorem 2.2[Chen (1996, 1999)]. Let $E=\{0,1,2,\cdots,N\}, N\leqslant\infty$,

$$q_{i,i+1}=b_i>0\ (0\leqslant i\leqslant N-1),\qquad q_{i,i-1}=a_i>0\ (1\leqslant i\leqslant N)$$

and $q_{ij}=0$ for other $i\neq j$. Denote by $\mathscr{W}$ the set of all strictly increasing sequence (w_i) with $\sum_{i=0}^N \mu_i w_i\geqslant 0$ and define

$$I_i(w)=\frac{1}{b_i\mu_i(w_{i+1}-w_i)}\sum_{j=i+1}^N \mu_j w_j,\qquad 0\leqslant i\leqslant N-1,$$

where

$$\mu_0=1,\quad \mu_n=b_0\cdots b_{n-1}/a_1\cdots a_n,\quad 1\leqslant n\leqslant N.$$

Then, we have

$$\lambda_1=\sup_{w\in\mathscr{W}}\ \inf_{0\leqslant i\leqslant N-1} I_i(w)^{-1}.$$

We now turn to discuss the trilogy of couplings: The Markovian coupling, the optimal Markovian coupling and the construction of distances for couplings.

3. Markovian couplings.

We concentrate on diffusions since the story for Markov chains is similar. Given an elliptic operator in $\mathbb{R}^d$

$$L=\sum_{i,\,j=1}^d a_{ij}(x)\frac{\partial^2}{\partial x_i\partial x_j}+\sum_{i=1}^d b_i(x)\frac{\partial}{\partial x_i}.$$

An elliptic (may be degenerated) operator $\widetilde{L}$ on the product space $\mathbb{R}^d\times\mathbb{R}^d$ is called a *coupling of* L if it satisfies the following marginality:

$$\widetilde{L}f(x,y)=Lf(x)\ (\text{resp. } \widetilde{L}f(x,y)=Lf(y)),\quad f\in C_b^2(\mathbb{R}^d),\ x\neq y,$$

where on the left-hand side, f is regarded as a bivariate function. From this, it is clear that the coefficients of any coupling operator $\widetilde{L}$ should be of the form

$$a(x,y)=\begin{pmatrix} a(x) & c(x,y)\\ c(x,y)^* & a(y)\end{pmatrix},\qquad b(x,y)=\begin{pmatrix} b(x)\\ b(y)\end{pmatrix}.$$

This condition and the non-negative definite property of $a(x,y)$ consist of the *marginality* of $\widetilde{L}$ in the context of diffusions. Obviously, the only freedom is the choice of $c(x,y)$.

Three examples:

(1) *Classical coupling.* $c(x,y)\equiv 0, x\neq y$.
(2) *March coupling* (Chen and Li(1989)). Let $a(x)=\sigma(x)^2$. Take $c(x,y)=\sigma(x)\sigma(y)$.
(3) *Coupling by reflection.* Set $\bar{u}=(x-y)/|x-y|$ and take

$$c(x,y)=\sigma(x)\left[\sigma(y)-2\frac{\sigma(y)^{-1}\bar{u}\bar{u}^*}{|\sigma(y)^{-1}\bar{u}|^2}\right],\det\sigma(y)\neq 0,\ x\neq y\ \text{[Lindvall \& Rogers (1986)]}$$

$$c(x,y)=\sigma(x)\big[I-2\bar{u}\bar{u}^*\big]\sigma(y),\quad x\neq y\quad \text{[Chen \& Li (1989)]}.$$

In the case that $x = y$, the first and the third ones are defined to be the same as the second one. Each coupling has its own character. A nice way to interpret the first coupling is to use a Chinese idiom: fall in love at first sight. The word "march" is a Chinese command to soldiers to start marching. We are now ready to talk about

4. Sketch of the main proof[Chen & Wang (1993)]**.**
Here we adopt the analytic language. Given a self-adjoint elliptic operator L, denote by $\{P_t\}_{t\geqslant 0}$ the semigroup determined by L: $P_t = e^{tL}$. Corresponding to $\widetilde{L}$, we have $\{\widetilde{P}_t\}_{t\geqslant 0}$. The coupling simply means that

$$\widetilde{P}_t f(x, y) = P_t f(x) \text{ (resp. } \widetilde{P}_t f(x, y) = P_t f(y)) \tag{2.3}$$

for all $f \in C_b^2(\mathbb{R}^d)$ and all (x, y) $(x \neq y)$, where on the left-hand side, f is again regarded as a bivariate function.
Step 1. Let g be an eigenfunction of $-L$ corresponding to λ_1. By the standard differential equation of the semigroup, we have $\frac{d}{dt} P_t g(x) = P_t L g(x) = -\lambda_1 P_t g(x)$. Solving this ordinary differential equation in $P_t g(x)$ for fixed g and x, we obtain

$$P_t g(x) = g(x) e^{-\lambda_1 t}. \tag{2.4}$$

Step 2. Consider compact space. Then g is Lipschitz with respect to the distance ρ. Denote by c_g the Lipschitz constant. Now, the main condition we need is the following:

$$\widetilde{P}_t \rho(x, y) \leqslant \rho(x, y) e^{-\alpha t}. \tag{2.5}$$

This condition is implied by

$$\widetilde{L} \rho(x, y) \leqslant -\alpha \rho(x, y), \quad x \neq y \tag{2.6}$$

(cf. Lemma A6). Setting $g_1(x, y) = g(x)$ and $g_2(x, y) = g(y)$, we obtain

$$\begin{aligned} e^{-\lambda_1 t} |g(x) - g(y)| &= \big|P_t g(x) - P_t g(y)\big| \quad \text{(by (2.4))} \\ &= \big|\widetilde{P}_t g_1(x, y) - \widetilde{P}_t g_2(x, y)\big| \quad \text{(by (2.3))} \\ &\leqslant \widetilde{P}_t |g_1 - g_2|(x, y) \\ &\leqslant c_g \widetilde{P}_t \rho(x, y) \quad \text{(Lipschitz property)} \\ &\leqslant c_g \rho(x, y) e^{-\alpha t} \quad \text{(by (2.5))}. \end{aligned}$$

Letting $t \to \infty$, we must have $\lambda_1 \geqslant \alpha$. $\square$

The proof is unbelievably straightforward. It is universal in the sense that it works for general Markov processes. A good point in the proof is the use of eigenfunction so that we can achieve the sharp estimates. On the other hand, it is crucial that we do not need too much knowledge about the eigenfunction, otherwise, there is no hope to work out since the eigenvalue and its eigenfunction are either known or unknown simultaneously. Except the Lipschitz property of g with respect to the distance, which can be avoided by using a localizing procedure for the non-compact case, the key of the proof is clearly the condition (2.6). For this, one needs not only a good coupling but also a good choice of the distance.

5. Optimal Markovian Coupling.

Since there are infinitely many choices of coupling operators, it is natural to ask the following questions. Does there exist an optimal one? Then, in what sense of optimality we are talking about?

Definition 2.3[Chen (1994)]**.** Let $(E, \rho, \mathscr{E})$ be a metric space. A coupling operator $\overline{L}$ is called *ρ-optimal* if

$$\overline{L}\,\rho(x_1, x_2) = \inf_{\widetilde{L}} \widetilde{L}\rho(x_1, x_2) \quad \text{for all } x_1 \neq x_2,$$

where $\widetilde{L}$ varies over all coupling operators.

To construct an optimal Markovian coupling is not an easy job even though there is often no problem for the existence. Here, we mention a special case only.

Theorem 2.4[Chen(1994)]**.** Let $f \in C^2(\mathbb{R}_+; \mathbb{R}_+)$ with $f(0) = 0$ and $f' > 0$. Suppose that $a(x) = \varphi(x)\sigma^2$ for some positive function φ, where σ is constant matrix with $\det\sigma > 0$.

(1) If $\rho(x, y) = f(|\sigma^{-1}(x - y)|)$ with $f'' \leqslant 0$, then the coupling by reflection is ρ-optimal. That is, $c(x, y) = \sqrt{\varphi(x)}\big[\sigma^2 - 2\bar{u}\bar{u}^*/|\sigma^{-1}\bar{u}|^2\big]\sqrt{\varphi(y)}$.

(2) If $\rho(x, y) = f(|\sigma^{-1}(x - y)|)$ with $f'' \geqslant 0$, then the march coupling is ρ-optimal. That is, $c(x, y) = \sqrt{\varphi(x)}\sigma^2\sqrt{\varphi(y)}$.

(3) If $d = 1$ and $\rho(x, y) = |x - y|$, then all the three couplings mentioned above are ρ-optimal.

Note that in case (2), ρ may not be a distance but the definition of ρ-optimal coupling is still meaningful.

6. Construction of Distances.

In view of the above theorem, one sees that the optimal coupling depends heavily on ρ and furthermore, even for a fixed optimal coupling, there is still a large class of ρ can be chosen, for which, the resulting estimate of α given in (2.6) may be completely different. For instance, the sharp estimates for the Laplacian on manifolds can not be achieved if one restricted to the Riemannian distance only. Thus, the construction of the distances plays a key role in the application of our coupling approach. However, we now have a unified construction for the distance ρ used for three classes of processes. Here, we write down the answer for diffusions in half line only.

$$g(r) = \int_0^r e^{-C(s)} ds \int_s^\infty \frac{f(u)e^{C(u)}}{a(u)} du, \quad f \in \mathscr{F}, \qquad \rho(x, y) = |g(x) - g(y)|.$$

The idea of finding these distances was explained step by step in the survey article [Chen, 1997] which contains much more materials. For instance, the geometric aspect of the study was discussed there but is not touched here. This construction of distances consists of the last part of the trilogy of couplings.

To conclude this part, we introduce a direct, analytic proof of Theorem 2.1. However, there is still no at the moment such a simple proof in the higher-dimensional situation.

7. Analytic proof of Theorem 2.1[Chen (1999a)]**.** a) Set

$$I(f)(x) = \frac{e^{-C(x)}}{f'(x)} \int_x^D \frac{f(u)e^{C(u)}}{a(u)} \mathrm{d}u.$$

Let $g \in C^1[0, D]$ with $\pi(g) = 0$ and $\pi(g^2) = 1$. Then for every $f \in \mathscr{F}$, we have

$$\begin{aligned}
1 &= \frac{1}{2} \int_0^D \pi(\mathrm{d}x)\pi(\mathrm{d}y)[g(y) - g(x)]^2 \\
&= \int_{\{x \leqslant y\}} \pi(\mathrm{d}x)\pi(\mathrm{d}y) \bigg(\int_x^y g'(u)\sqrt{f'(u)}/\sqrt{f'(u)}\, \mathrm{d}u \bigg)^2 \\
&\leqslant \int_{\{x \leqslant y\}} \pi(\mathrm{d}x)\pi(\mathrm{d}y) \int_x^y g'(u)^2 f'(u)^{-1} \mathrm{d}u \int_x^y f'(\xi)\mathrm{d}\xi \\
&\qquad\qquad \text{(by Cauchy-Schwarz inequality)} \\
&= \int_{\{x \leqslant y\}} \pi(\mathrm{d}x)\pi(\mathrm{d}y) \int_x^y a(u)g'(u)^2 e^{C(u)} \frac{e^{-C(u)}}{a(u)f'(u)} \mathrm{d}u\, \big[f(y) - f(x)\big] \\
&= \int_0^D a(u)g'(u)^2 \pi(\mathrm{d}u) \frac{Ze^{-C(u)}}{f'(u)} \int_0^u \pi(\mathrm{d}x) \int_u^D \pi(\mathrm{d}y)\, \big[f(y) - f(x)\big].
\end{aligned} \tag{2.7}$$

But

$$\begin{aligned}
&\int_0^u \pi(\mathrm{d}x) \int_u^D \pi(\mathrm{d}y)\, \big[f(y) - f(x)\big] \\
&\quad= \int_0^u \pi(\mathrm{d}x) \int_u^D f(y)\pi(\mathrm{d}y) - \int_0^u f(x)\pi(\mathrm{d}x) \int_u^D \pi(\mathrm{d}y) \\
&\quad= \int_u^D f(y)\pi(\mathrm{d}y) - \int_u^D \pi(\mathrm{d}x) \int_u^D f(y)\pi(\mathrm{d}y) - \int_0^u f(x)\pi(\mathrm{d}x) \int_u^D \pi(\mathrm{d}y) \\
&\quad= \int_u^D f(y)\pi(\mathrm{d}y) - \bigg[\int_u^D \pi(\mathrm{d}x)\bigg] \int_0^D f(y)\pi(\mathrm{d}y) \\
&\quad\leqslant \int_u^D f(y)\pi(\mathrm{d}y) \quad (\text{since } \pi(f) \geqslant 0) \\
&\quad= \frac{1}{Z} \int_u^D \frac{f(y)e^{C(y)}}{a(y)}\, \mathrm{d}y.
\end{aligned}$$

Combining this with (2.7), we obtain

$$\int_0^D a(x)g'(x)^2\pi(\mathrm{d}x) \geqslant \inf_{x \in (0,D)} I(f)(x)^{-1}.$$

Then (2.2) follows by making infimum over g and then supremum over $f \in \mathscr{F}$.

b) The proof of the last assertion of Theorem 2.1 is more technical. Refer to [Chen & Wang (1997b)] and [Chen (1999a)]. □

Part III. Poincaré Inequality, Cheeger's Method

Let $(E, \mathscr{E}, \pi)$ be a probability space. Suppose that $\{(x,x) : x \in E\} \in \mathscr{E}$. Denote by $L^2(\pi)$ the usual real L^2-space with norm $\|\cdot\|$. Consider the symmetric form $(D, \mathscr{D}(D))$: its diagonal element is

$$D(f) := D(f,f) = \frac{1}{2}\int_{E\times E} J(\mathrm{d}x, \mathrm{d}y)[f(y) - f(x)]^2 \tag{3.1}$$

where J is a non-negative, symmetric measure having no charge on the diagonal set $\{(x,x) : x \in E\}$. A typical example in our mind is the reversible jump process with q-pair $(q(x), q(x, \mathrm{d}y))$ and reversible measure π. Then $J(\mathrm{d}x, \mathrm{d}y) = \pi(\mathrm{d}x)q(x, \mathrm{d}y)$. More especially, for a reversible Q-matrix $Q = (q_{ij})$ with reversible measure $(\pi_i > 0)$, we have density $J_{ij} = \pi_i q_{ij}$ $(j \neq i)$ with respect to the counting measure. The domain of D is taken to be the maximal one:

$$\mathscr{D}(D) = \{f \in L^2(\pi) : D(f) < \infty\}.$$

Since $D(1) = 0$, we have the trivial eigenvalue $\lambda_0 := \inf\{D(f) : \|f\| = 1\} = 0$. We are now interested in the first non-trivial eigenvalue $\lambda_1 = \inf\{D(f) : \pi(f) = 0, \|f\| = 1\}$, where $\pi(f) = \int f \mathrm{d}\pi$. It is also called the *spectral gap* of the form $(D, \mathscr{D}(D))$. In other words, we are interested in

$$\textit{Poincaré inequality}: \qquad \mathrm{Var}(f)(= \|f - \pi(f)\|^2) \leqslant \lambda_1^{-1} D(f), \qquad f \in L^2(\pi).$$

The first key idea is bounding the symmetric measure J. For this, choose a non-negative, symmetric measurable function $r(x,y)$ such that

$$J^{(1)}(\mathrm{d}x, E)/\pi(\mathrm{d}x) \leqslant 1, \qquad \pi\text{-a.s}, \tag{3.2}$$

where

$$J^{(\alpha)}(\mathrm{d}x, \mathrm{d}y) := I_{\{r(x,y)^\alpha > 0\}} \frac{J(\mathrm{d}x, \mathrm{d}y)}{r(x,y)^\alpha}, \qquad \alpha > 0.$$

Throughout the paper we adopt the convention $r^0 = 1$ for all $r \geqslant 0$. Thus, when $\alpha = 0$, we return to the original form $J^{(0)} = J$. For jump process with q-pair $(q(x), q(x, \mathrm{d}y))$, one simply takes $r(x,y) = q(x) \vee q(y)$. We can now define a new type of Cheeger's constant as follows.

$$k^{(\alpha)} = \inf_{\pi(A)\in(0,1)} \frac{J^{(\alpha)}(A \times A^c)}{\pi(A) \wedge \pi(A^c)} = \inf_{\pi(A)\in(0,1/2]} \frac{J^{(\alpha)}(A \times A^c)}{\pi(A)}.$$

Now, here is one of our main results.

Theorem 3.1[Chen & Wang (1998)]**.** Under (3.2), we have

$$\lambda_1 \geqslant {k^{(1/2)}}^2 \Big/ \Big[1 + \sqrt{1 - {k^{(1)}}^2}\,\Big].$$

The following simple consequence illustrates the application of the theorem. It has already contained a large number of non-trivial examples.

Corollary 3.2[Chen & Wang (1998)]. Let $j(x,y)$ be a non-negative, symmetric function with $j(x,x)=0$ and $j(x):=\int j(x,y)\,\pi(\mathrm{d}y)<\infty$ for all $x\in E$. Then, for the symmetric form $J(\mathrm{d}x,\mathrm{d}y):=j(x,y)\pi(\mathrm{d}x)\pi(\mathrm{d}y)$, we have

$$\lambda_1 \geqslant \frac{1}{8}\inf_{x\neq y}\frac{j(x,y)^2}{j(x)\vee j(y)}.$$

Proof. Note that

$$\begin{aligned}
k^{(\alpha)} &= \inf_{\pi(A)\in(0,1/2]}\frac{1}{\pi(A)}\int_{A\times A^c}\frac{j(x,y)}{[j(x)\vee j(y)]^\alpha}\pi(\mathrm{d}x)\pi(\mathrm{d}y)\\
&\geqslant \inf_{x\neq y}\frac{j(x,y)}{[j(x)\vee j(y)]^\alpha}\inf_{\pi(A)\in(0,1/2]}\pi(A^c)\\
&\geqslant \frac{1}{2}\inf_{x\neq y}\frac{j(x,y)}{[j(x)\vee j(y)]^\alpha}.
\end{aligned}$$

The conclusion follows from Theorem 3.1 immediately. $\square$

The proof of Theorem 3.1 is based on Cheeger's idea. That is, estimate λ_1 in terms of λ_0 for a more general symmetric form

$$D(f)=\frac{1}{2}\int_{E\times E}J(\mathrm{d}x,\mathrm{d}y)[f(y)-f(x)]^2+\int_E K(\mathrm{d}x)f(x)^2, \tag{3.3}$$

where K is a non-negative measure on $(E,\mathscr{E})$. The study on λ_0 is meaningful since $D(1)\neq 0$ whenever $K\neq 0$. It is called *Dirichlet eigenvalue of* $(D,\mathscr{D}(D))$. Thus, in what follows, when dealing with λ_0 (resp., λ_1), we consider only the symmetric form given by (3.3) (resp., (3.1)). Instead of (3.2), we now require that

$$[J^{(1)}(\mathrm{d}x,E)+K^{(1)}(\mathrm{d}x)]/\pi(\mathrm{d}x)\leqslant 1, \qquad \pi\text{-a.s}, \tag{3.4}$$

where $J^{(\alpha)}$ is the same as before and

$$K^{(\alpha)}(\mathrm{d}x)=I_{\{s(x)^\alpha>0\}}K(\mathrm{d}x)/s(x)^\alpha$$

for some non-negative function $s(x)$. Next, define

$$h^{(\alpha)}=\inf_{\pi(A)>0}[J^{(\alpha)}(A\times A^c)+K^{(\alpha)(A)}]/\pi(A).$$

Theorem 3.3[Chen & Wang (1998)]. For the symmetric form given by (3.3), under (3.4), we have

$$\lambda_0\geqslant {h^{(1/2)}}^2\Big/\Big[1+\sqrt{1-{h^{(1)}}^2}\,\Big].$$

Proof. a) First, we express $h^{(\alpha)}$ by the following functional form

$$h^{(\alpha)}=\inf\bigg\{\frac{1}{2}\int J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y)|f(x)-f(y)|+K^{(\alpha)}(f):\, f\geqslant 0,\,\pi(f)=1\bigg\}.$$

By setting $f = I_A/\pi(A)$, one returns to the original set form of $h^{(\alpha)}$. For the reverse assertion, simply consider the set $A_\gamma = \{f > \gamma\}$ for $\gamma \geqslant 0$. The proof is also not difficult:

$$\begin{aligned}
&\int_{\{f(x)>f(y)\}} J^{(\alpha)}(\mathrm{d}x, \mathrm{d}y)[f(x) - f(y)] + K^{(\alpha)}(f)\\
&= \int_0^\infty \mathrm{d}\gamma \Big\{ J^{(\alpha)}\big(\{f(x) > \gamma \geqslant f(y)\}\big) + K^{(\alpha)}\big(\{f > \gamma\}\big)\Big\}\\
&= \int_0^\infty \big[J^{(\alpha)}\big(A_\gamma \times A_\gamma^c\big) + K^{(\alpha)}(A_\gamma)\big]\mathrm{d}\gamma\\
&\geqslant h^{(\alpha)} \int_0^\infty \pi\big(A_\gamma\big)\mathrm{d}\gamma\\
&= h^{(\alpha)}\pi(f).
\end{aligned}$$

The appearance of K makes the notations heavier. To avoid this, one may enlarge the state space to $E^* = E \cup \{\infty\}$. Regarding K as a killing measure on E^*, the form $D(f, g)$ can be extended to the product space $E^* \times E^*$ but expressed by using a symmetric measure J^* only. At the same time, one can extend f to a function f^* on E^*: $f^* = f I_E$. Then, we have

$$h^{(\alpha)} = \inf\bigg\{\frac{1}{2}\int_{E^*\times E^*} J^{*(\alpha)}(\mathrm{d}x, \mathrm{d}y)|f^*(x) - f^*(y)| : \, f \geqslant 0, \, \pi(f) = 1\bigg\}.$$

However, for simplicity, we will omit the superscript "$*$" everywhere.

b) Take f with $\pi(f^2) = 1$, by a), Cauchy-Schwarz inequality and condition (3.4), we have

$$\begin{aligned}
{h^{(1)}}^2 &\leqslant \bigg\{\frac{1}{2}\int J^{(1)}(\mathrm{d}x, \mathrm{d}y)\big|f(y)^2 - f(x)^2\big|\bigg\}^2\\
&\leqslant \frac{1}{2}D^{(1)}(f)\int J^{(1)}(\mathrm{d}x, \mathrm{d}y)\big[f(y) + f(x)\big]^2\\
&= \frac{1}{2}D^{(1)}(f)\bigg\{2\int J^{(1)}(\mathrm{d}x, \mathrm{d}y)\big[f(y)^2 + f(x)^2\big]\\
&\qquad\qquad - \int J^{(1)}(\mathrm{d}x, \mathrm{d}y)\big[f(y) - f(x)\big]^2\bigg\}\\
&\leqslant D^{(1)}(f)\big[2 - D^{(1)}(f)\big]. \qquad (3.5)
\end{aligned}$$

Solving this quadratic inequality in $D^{(1)}(f)$, one obtains $D^{(1)}(f) \geqslant 1 - \sqrt{1 - {h^{(1)}}^2}$.

c) Repeating the above proof but by a more careful use of Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
{h^{(1/2)}}^2 &\leqslant \bigg\{\frac{1}{2}\int J^{(1/2)}(\mathrm{d}x, \mathrm{d}y)\big|f(y)^2 - f(x)^2\big|\bigg\}^2\\
&= \bigg\{\frac{1}{2}\int J(\mathrm{d}x, \mathrm{d}y)|f(y) - f(x)| \cdot I_{\{r(x,y)>0\}}\frac{|f(y) + f(x)|}{\sqrt{r(x, y)}}\bigg\}^2 \leqslant
\end{aligned}$$

$$\leqslant \frac{1}{2}D(f)\int J^{(1)}(\mathrm{d}x,\mathrm{d}y)\big[f(y)+f(x)\big]^2$$
$$\leqslant D(f)\big[2-D^{(1)}(f)\big].$$

From this and b), the required assertion follows. $\square$

Proof of Theorem 3.1. a) For any $B\subset E$ with $\pi(B)>0$, define a local form as follows.

$$\widetilde{D}_B^{(\alpha)}(f)=\frac{1}{2}\int_{B\times B}J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y)[f(y)-f(x)]^2+\int_B J^{(\alpha)}(\mathrm{d}x,B^c)f(x)^2.$$

Obviously, $\widetilde{D}_B^{(\alpha)}(f)=\widetilde{D}_B^{(\alpha)}(fI_B)$. Moreover,

$$\lambda_0(B):=\inf\{D(f):f|_{B^c}=0,\ \|f\|=1\}=\inf\big\{\widetilde{D}_B(f):\ \pi(f^2I_B)=1\big\}.$$

Let $\pi^B=\pi(\cdot\cap B)/\pi(B)$ and set

$$h_B^{(\alpha)}=\inf_{A\subset B,\,\pi(A)>0}\frac{J^{(\alpha)}(A\times(B\setminus A))+J^{(\alpha)}(A\times B^c)}{\pi(A)}=\inf_{A\subset B,\,\pi(A)>0}\frac{J^{(\alpha)}(A\times A^c)}{\pi(A)}.$$

Applying Theorem 3.3 to the local form on $L^2(B,\mathscr{E}\cap B,\pi^B)$ generated by $J^B=\pi(B)^{-1}J|_{B\times B}$ and $K^B=J(\cdot,B^c)|_B$, we obtain

$$\lambda_0(B)\geqslant {h_B^{(1/2)}}^2\Big/\Big[1+\sqrt{1-{h_B^{(1)}}^2}\,\Big].$$

b) We now come to another key point of the proof. The original Cheeger's proof is based on the estimate

$$\lambda_1\geqslant\inf_B\{\lambda_0(B)\vee\lambda_0(B^c)\}.$$

It is called the *Cheeger's splitting technique*. However, we are unable to prove this in the present setup. Instead, we use the following weaker result which is enough for our purpose.

$$\lambda_1\geqslant\inf_{\pi(B)\leqslant 1/2}\lambda_0(B).\tag{3.6}$$

Assuming (3.6), the proof of Theorem 3.1 is easy to complete. Note that

$$\inf_{\pi(B)\leqslant 1/2}h_B^{(\alpha)}=k^{(\alpha)}.$$

Hence

$$\begin{aligned}\lambda_1&\geqslant\inf_{\pi(B)\leqslant 1/2}\frac{{h_B^{(1/2)}}^2}{1+\sqrt{1-{h_B^{(1)}}^2}}\\&\geqslant\inf_{\pi(B)\leqslant 1/2}\frac{\inf_{\pi(B)\leqslant 1/2}{h_B^{(1/2)}}^2}{1+\sqrt{1-{h_B^{(1)}}^2}}\\&\geqslant\frac{\inf_{\pi(B)\leqslant 1/2}{h_B^{(1/2)}}^2}{1+\sqrt{1-\inf_{\pi(B)\leqslant 1/2}{h_B^{(1)}}^2}}\\&=\frac{{k^{(1/2)}}^2}{1+\sqrt{1-{k^{(1)}}^2}}.\end{aligned}$$

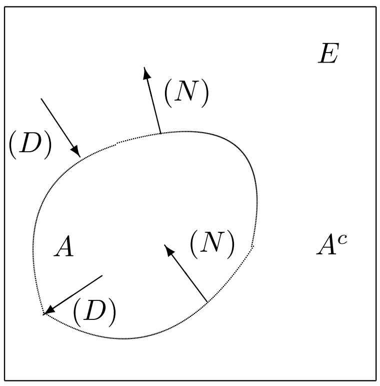
c) It remains to prove (3.6). For each $\varepsilon > 0$, choose f_ε with $\pi(f_\varepsilon) = 0$ and $\pi(f_\varepsilon^2) = 1$ such that $\lambda_1 + \varepsilon \geqslant D(f_\varepsilon)$. Next, choose c_ε such that $\pi(f_\varepsilon < c_\varepsilon)$, $\pi(f_\varepsilon > c_\varepsilon) \leqslant 1/2$. Set $f_\varepsilon^\pm = (f_\varepsilon - c_\varepsilon)^\pm$ and $B_\varepsilon^\pm = \{f_\varepsilon^\pm > 0\}$. Then

$$\begin{aligned}
\lambda_1 + \varepsilon &\geqslant D(f_\varepsilon)\\
&= D(f_\varepsilon - c_\varepsilon)\\
&= \frac{1}{2}\int J(\mathrm{d}x, \mathrm{d}y)\big[\big|f_\varepsilon^+(y) - f_\varepsilon^+(x)\big| + \big|f_\varepsilon^-(y) - f_\varepsilon^-(x)\big|\big]^2\\
&\geqslant \frac{1}{2}\int J(\mathrm{d}x, \mathrm{d}y)\big(f_\varepsilon^+(y) - f_\varepsilon^+(x)\big)^2 + \frac{1}{2}\int J(\mathrm{d}x, \mathrm{d}y)\big(f_\varepsilon^-(y) - f_\varepsilon^-(x)\big)^2\\
&\geqslant \lambda_0\big(B_\varepsilon^+\big)\pi\big(\big(f_\varepsilon^+\big)^2\big) + \lambda_0\big(B_\varepsilon^-\big)\pi\big(\big(f_\varepsilon^-\big)^2\big)\\
&\geqslant \inf_{\pi(B)\leqslant 1/2} \lambda_0(B)\pi\big(\big(f_\varepsilon^+\big)^2 + \big(f_\varepsilon^-\big)^2\big)\\
&= (1 + c_\varepsilon^2) \inf_{\pi(B)\leqslant 1/2} \lambda_0(B)\\
&\geqslant \inf_{\pi(B)\leqslant 1/2} \lambda_0(B).
\end{aligned}$$

Because ε is arbitrary, we obtain the required conclusion. $\square$

Part IV. Existence Criterion of Spectral Gap

For compact state space $(E, \mathscr{E})$, it is often true that $\lambda_1 > 0$. Thus, we need only to consider the non-compact case. The idea is to use the Cheeger's splitting technique. Split the space E into two parts A and A^c. Mainly, there are two boundary conditions: Dirichlet or Neumann boundary condition, that is absorbing or reflecting at the boundary respectively. The corresponding eigenvalue problems are denoted by (D) and (N) respectively. Let A be compact for a moment. Then on A^c, one should consider the problem (D). Otherwise, since A^c is non-compact, the solution to problem (N) is unknown and it is indeed what we are also interested in. On A, we can use either of the boundary conditions. However, it is better to use the Neumann one since the corresponding λ_1 is more closer to the original λ_1 when A becomes larger. In other words, we want to describe the original λ_1 in terms of the local $\lambda_1(A)$ and $\lambda_0(A^c)$.



We now state our criterion informally, which is easier to remember.

Criterion (Heuristic Description). $\lambda_1 > 0$ iff $\lambda_0(A^c) > 0$ for some compact A.

Why the result can be regarded as a criterion is due to the following reason. Consider a second-order differential operator L in $\mathbb{R}^d$. Then there is a variational formula for λ_0:
$\lambda_0(A^c) = \sup_{\varphi>0} \inf_{A^c}(-L\varphi)/\varphi$. Moreover, $\lambda_0 > 0$ iff the Maximum Principle holds. Thus, ones have a lot of knowledge about λ_0. The conclusion also works for other types of Markov processes.

To present a precise description of the criterion, we restrict ourselves to the symmetric form studied above.

Define another local form as follows.

$$D_B^{(\alpha)}(f) = \frac{1}{2}\int_{B\times B} J^{(\alpha)}(\mathrm{d}x, \mathrm{d}y)[f(y) - f(x)]^2.$$

The corresponding spectral gap is denoted by $\lambda_1(B)$. Note that in the result below, we use $\lambda_1(B)$ rather than $\lambda_1(A)$ mentioned above the heuristic description.

Theorem 4.1[Chen & Wang (1998)]**.** Consider the form given by (3.1). For any $A \subset B$ with $0 < \pi(A)$, $\pi(B) < 1$, we have

$$\frac{\lambda_0(A^c)}{\pi(A)} \geqslant \lambda_1 \geqslant \frac{\lambda_1(B)[\lambda_0(A^c)\pi(B) - 2M_A\pi(B^c)]}{2\lambda_1(B) + \pi(B)^2[\lambda_0(A^c) + 2M_A]}, \tag{4.1}$$

where $M_A = \operatorname*{ess\,sup}_{\pi}\limits_{x\in A} J(\mathrm{d}x, A^c)/\pi(\mathrm{d}x)$.

As we mentioned before, usually, $\lambda_1(B) > 0$ for all compact B. Hence the result means, as stated in the heuristic description, that $\lambda_1 > 0$ iff $\lambda_0(A^c) > 0$ for some compact A, because we can first fix such an A and then make B large enough so that the right-hand side of (4.1) becomes positive.

Proof of Theorem 4.1. Let f satisfy $\pi(f) = 0$ and $\pi(f^2) = 1$. Our aim is to bound $D(f)$ in terms of $\lambda_0(A^c)$ and $\lambda_1(B)$.

a) First, we use $\lambda_1(B)$.

$$\begin{aligned} D(f) &\geqslant D_B(fI_B) \geqslant \lambda_1(B)\pi(B)^{-1}\big[\pi\big(f^2I_B\big) - \pi(B)^{-1}\pi\big(fI_B\big)^2\big] \\ &= \lambda_1(B)\pi(B)^{-1}\big[\pi\big(f^2I_B\big) - \pi(B)^{-1}\pi\big(fI_{B^c}\big)^2\big]. \end{aligned} \tag{4.2}$$

Here in the last step, we have used $\pi(f) = 0$.

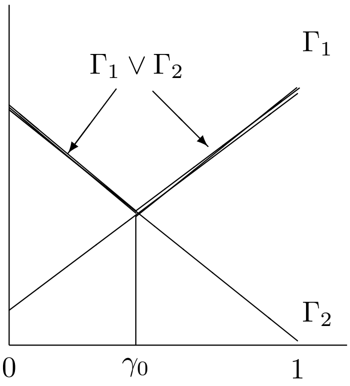
b) Next, we use $\lambda_0(A^c)$. We need the following elementary inequality

$$|(fI_{A^c})(x) - (fI_{A^c})(y)| \leqslant |f(x) - f(y)| + I_{A\times A^c\cup A^c\times A}(x,y)|(fI_A)(x) - (fI_A)(y)|.$$

Then

$$\begin{aligned} \lambda_0(A^c)\pi\big(f^2I_{A^c}\big) &\leqslant D(fI_{A^c}) \\ &= \frac{1}{2}\int J(\mathrm{d}x, \mathrm{d}y)\big[(fI_{A^c})(y) - (fI_{A^c})(x)\big]^2 \\ &\leqslant 2D(f) + 2\int_{A\times A^c} J(\mathrm{d}x, \mathrm{d}y)\big[(fI_A)(y) - (fI_A)(x)\big]^2 \\ &\leqslant 2D(f) + 2M_A\pi\big(f^2I_A\big). \end{aligned} \tag{4.3}$$

c) Estimating the right-hand sides of (4.2) and (4.3) in terms of $\gamma := \pi\big(f^2 I_B\big)$, we obtain two inequalities $D(f) \geqslant c_1\gamma + c_2$ and $D(f) \geqslant -c_3\gamma + c_4$ for some constants c_1, $c_3 > 0$.



Hence

$$D(f) \geqslant \inf_{\gamma\in[0,1]} \max\{c_1\gamma + c_2,\ -c_3\gamma + c_4\}.$$

Clearly, the infimum is achieved at γ_0, which is the intersection of the two lines Γ_1 and Γ_2 in $\{\cdots\}$. Then, the required lower bound of λ_1 is given by $c_1\gamma_0 + c_2$. $\square$

Part V. Logarithmic Sobolev inequality and Nash inequality

In this part, we adopt again the Cheeger's method to study the following two inequalities.

Logarithmic Sobolev inequality: $\pi\big(f^2 \log f^2/\|f\|^2\big) \leqslant 2\sigma^{-1} D(f), \quad f \in L^2(\pi).$

Nash inequality: $\mathrm{Var}(f)(= \|f - \pi(f)\|^2) \leqslant \eta^{-1} D(f)^{1/p} \|f\|_1^{2/q}, \quad f \in L^2(\pi).$

In order to get the ideas quickly, here we discuss only some weaker results and leave the general results to the original papers [Chen (1999 b, c)] and [Wang (1998b)]. There are plenty of publications for the logarithmic Sobolev inequality in the context of diffusions which are not touched here. Refer to Chen & Wang (1997), Wang (1998a) and Wang (1999) for the statue of the present study, related topics and further references.

1. Logarithmic Sobolev inequality.

Theorem 5.1[Chen (1999c)]**.** We have

$$2\kappa \geqslant \sigma \geqslant \frac{2\lambda_1 \kappa^{(1/2)}}{\sqrt{\lambda_1(2-\lambda_1^{(1)})} + 3\kappa^{(1/2)}} \geqslant \frac{1}{8}{\kappa^{(1/2)}}^2,$$

where

$$\kappa^{(\alpha)} = \inf_{\pi(A)\in(0,1)} \frac{J^{(\alpha)}(A\times A^c)}{-\pi(A)\log\pi(A)}$$

and

$$\lambda_1^{(\alpha)} = \inf\{D^{(\alpha)}(f):\ \pi(f)=0,\ \|f\|=1\}.$$

Proof. The proof is partially due to F. Y. Wang. To get the upper bound, simply apply the inequality to the test function $f = I_A/\sqrt{\pi(A)}$, $\pi(A) \in (0,1)$. To prove the lower bound, let $\pi(f) = 0$ and $\|f\| = 1$.

a) Set $\varepsilon = \sqrt{2 - \lambda_1^{(1)}}/[2\kappa^{(1/2)}]$ and $E(f) = \pi\big(f^2 \log f^2\big)$. Then, it can be proved that

$$E(f) \leqslant 2\varepsilon\sqrt{D(f)} + 1. \tag{5.1}$$

Actually, one shows first that

$$I := \frac{1}{2}\int J^{(1/2)}(\mathrm{d}x, \mathrm{d}y)\big|f(y)^2 - f(x)^2\big| \leqslant \sqrt{\big(2 - \lambda_1^{(1)}\big)D(f)}. \tag{5.2}$$

The proof is standard as used before (cf. (3.5)). Next, set $A_t = \big\{f^2 > t\big\}$ and prove that

$$I \geqslant \kappa^{(1/2)}[E(f) - 1]. \tag{5.3}$$

The proof goes as follows. Note that $h_t := \pi(A_t) \leqslant 1 \wedge t^{-1}$. We have

$$\begin{aligned}
I &= \int_0^\infty J^{(1/2)}(A_t \times A_t^c)\mathrm{d}t \\
&\geqslant \kappa^{(1/2)}\int_0^\infty (-h_t \log h_t)\mathrm{d}t \\
&\geqslant \kappa^{(1/2)}\int_0^\infty h_t \log t\mathrm{d}t \\
&= \kappa^{(1/2)}\int_0^\infty h_t(\log t + 1)\mathrm{d}t - \kappa^{(1/2)} \\
&= \kappa^{(1/2)}\int \mathrm{d}\pi \int_0^{f^2}(\log t + 1)\mathrm{d}t - \kappa^{(1/2)} \\
&= \kappa^{(1/2)}[E(f) - 1].
\end{aligned}$$

Combining (5.2) with (5.3), we get (5.1).

b) By (5.1), we have

$$E(f) \leqslant 2\varepsilon\sqrt{D(f)} + 1 \leqslant \gamma\varepsilon D(f) + \varepsilon/\gamma + 1,$$

where $\gamma > 0$ is a constant to be specified below. On the other hand, by [Bakry (1992); Proposition 3.10], the inequality

$$\pi\big(f^2 \log f^2\big) \leqslant C_1 D(f) + C_2, \qquad \pi(f) = 0, \quad \|f\| = 1 \tag{5.4}$$

implies that

$$\sigma \geqslant 2/[C_1 + (C_2 + 2)\lambda_1^{-1}]. \tag{5.5}$$

Combining these facts together, it follows that

$$\sigma \geqslant \frac{2}{\varepsilon\gamma + [\varepsilon/\gamma + 3]/\lambda_1}.$$

Maximizing the right-hand side with respect to γ, we get

$$\sigma \geqslant \frac{2\lambda_1 \kappa^{(1/2)}}{\sqrt{(2-\lambda_1^{(1)})\lambda_1} + 3\kappa^{(1/2)}}. \tag{5.6}$$

On the other hand, applying Theorem 3.1 to $J^{(1)}$, we have $k^{(1/2)} = k^{(1)}$ and hence $\lambda_1^{(1)} \geqslant 1 - \sqrt{1 - {k^{(1)}}^2}$. Combining this with Theorem 3.1 and noting that $k^{(1/2)} \geqslant (\log 2)\kappa^{(1/2)}$, it follows that the right-hand side of (5.6) is bounded below by

$$\frac{2(\log 2)^2 {\kappa^{(1/2)}}^2}{(\log 2 + 3)\big[1 + \sqrt{1 - {k^{(1)}}^2}\,\big]} \geqslant \frac{1}{8}{\kappa^{(1/2)}}^2. \quad \square$$

We now introduce a more powerful result, its proof is much technical and omitted here.

Theorem 5.2[Chen (1999c)]**.** Define

$$\begin{aligned} \xi^{(\delta)} &= \inf_{\pi(A)>0} \frac{J^{(1/2)}(A \times A^c) + \delta\pi(A)}{\pi(A)\sqrt{1 - \log \pi(A)}}, \qquad \delta > 0, \\ \xi^{(\infty)} &= \lim_{\delta\to\infty} \xi^{(\delta)} = \sup_{\delta>0} \xi^{(\delta)} \end{aligned}$$

and

$$A(\delta) = \frac{(2+\delta)(\lambda_1 + \delta)}{(\xi^{(\delta)})^2}.$$

Then, we have

$$2\kappa \geqslant \sigma \geqslant \frac{2\lambda_1}{1 + 16 \inf_{\delta>0} A(\delta)}.$$

2. Upper bounds.

The upper bound given by Theorem 5.1 is usually very rough. Here we introduce two results which are often rather effective. The results show that order one (resp., two) of exponential integrability is required for $\lambda_1 > 0$ (resp., $\sigma > 0$).

Theorem 5.3[Chen & Wang (1998)]**.** Suppose that the function r used in (3.2) is J-a.e. positive. If there exists $\varphi \geqslant 0$ such that

$$\operatorname{ess\,sup}_J |\varphi(x) - \varphi(y)|^2 r(x,y) \leqslant 1, \tag{5.7}$$

then

$$\lambda_1 \leqslant \inf\Big\{\varepsilon^2/4 : \varepsilon \geqslant 0,\ \pi\big(e^{\varepsilon\varphi}\big) = \infty\Big\}. \tag{5.8}$$

Consequently, $\lambda_1 = 0$ if there exists $\varphi \geqslant 0$ satisfying (5.8) such that $\pi\big(e^{\varepsilon\varphi}\big) = \infty$ for all $\varepsilon > 0$.

Proof. We need to show that if $\pi\big(e^{\varepsilon\varphi}\big) = \infty$, then $\lambda_1 \leqslant \varepsilon^2/4$. For $n \geqslant 1$, define $f_n = \exp[\varepsilon(\varphi \wedge n)/2]$. Then, we have

$$\lambda_1 \leqslant D(f_n)\big/\big[\pi\big(f_n^2\big) - \pi(f_n)^2\big]. \tag{5.9}$$

For every $m \geqslant 1$, choose $r_m > 0$ such that $\pi(\varphi \geqslant r_m) \leqslant 1/m$. Then

$$\pi\big(I_{[\varphi \geqslant r_m]} f_n^2\big)^{1/2} \geqslant \sqrt{m}\, \pi\big(I_{[\varphi \geqslant r_m]} f_n\big) \geqslant \sqrt{m}\, \pi(f_n) - \sqrt{m}\, e^{\varepsilon r_m/2}.$$

Hence

$$\pi\big(f_n\big)^2 \leqslant \Big[\sqrt{\pi\big(f_n^2\big)}/\sqrt{m} + e^{\varepsilon r_m/2}\Big]^2. \tag{5.10}$$

On the other hand, by Mean Value Theorem, $|e^A - e^B| \leqslant |A - B| e^{A \vee B} = |A - B|(e^A \vee e^B)$ for all A, $B \geqslant 0$. Hence

$$\begin{aligned} D(f_n) &= \frac{1}{2}\int J(\mathrm{d}x, \mathrm{d}y)[f_n(x) - f_n(y)]^2 \\ &\leqslant \frac{\varepsilon^2}{8}\int J^{(1)}(\mathrm{d}x, \mathrm{d}y)[\varphi(x) - \varphi(y)]^2 r(x,y)\big[f_n(x) \vee f_n(y)\big]^2 \\ &\leqslant \frac{\varepsilon^2}{4}\pi\big(f_n^2\big). \end{aligned} \tag{5.11}$$

Noticing that $\pi\big(f_n^2\big) \uparrow \infty$, combining (5.11) with (5.9) and (5.10) and then letting $n \uparrow \infty$, we obtain $\lambda_1 \leqslant \varepsilon^2/\big[4(1 - m^{-1})\big]$. The proof is completed by setting $m \uparrow \infty$. $\square$

Theorem 5.4[Wang (1999)]**.** Suppose that (5.7) holds. If $\sigma > 0$, then

$$\pi\big(e^{\varepsilon\varphi^2}\big) \leqslant \exp\left[\frac{\varepsilon\pi(\varphi^2)}{1 - 2\varepsilon/\sigma}\right] < \infty, \qquad \varepsilon \in [0, \sigma/2).$$

Proof. a) Given $n \geqslant 1$, let $\varphi_n = \varphi \wedge n$, $f_n = \exp[r\varphi_n^2/2]$ and $h_n(r) = \pi\big(e^{r\varphi_n^2}\big)$. Then, by (3.2), (5.7) and applying the Cauchy's mean value theorem to the function $\exp[rx^2/2]$, we get

$$\begin{aligned} D(f_n) &= \frac{1}{2}\int J(\mathrm{d}x, \mathrm{d}y)[f_n(x) - f_n(y)]^2 \\ &\leqslant \frac{r^2}{2}\int J^{(1)}(\mathrm{d}x, \mathrm{d}y)[\varphi(x) - \varphi(y)]^2 r(x,y) \max\big\{\varphi_n(x) f_n(x), \varphi_n(y) f_n(y)\big\}^2 \\ &\leqslant r^2 \int J^{(1)}(\mathrm{d}x, \mathrm{d}y)\varphi_n(x)^2 f_n(x)^2 \leqslant r^2 h_n'. \end{aligned}$$

b) Next, applying logarithmic Sobolev inequality to the function f_n and using a), it follows that

$$r h_n'(r) \leqslant h_n(r) \log h_n(r) + 2r^2 h_n'(r)/\sigma, \qquad r \geqslant 0.$$

That is,

$$h_n'(r) \leqslant \frac{1}{r(1 - 2r/\sigma)} h_n(r) \log h_n(r), \qquad r \in [0, \sigma/2).$$

Now the required assertion follows from Corollary A5. $\square$

3. Nash inequality.

Theorem 5.5[Chen (1999b)]**.** Define the isoperimetric constant I_ν as follows:

$$I_\nu = \inf_{0<\pi(A)\leqslant 1/2} \frac{J^{(1/2)}(A\times A^c)}{\pi(A)^{(\nu-1)/\nu}} = \inf_{0<\pi(A)<1} \frac{J^{(1/2)}(A\times A^c)}{\big[\pi(A)\wedge\pi(A^c)\big]^{(\nu-1)/\nu}}, \qquad \nu>1.$$

Then

$$\mathrm{Var}(f)^{1+2/\nu} \leqslant 2I_\nu^{-2} D(f)\,\|f\|_1^{4/\nu}, \qquad f\in L^2(\pi). \tag{5.12}$$

Proof. The proof below is quite close to Saloff-Coste (1997). Fix a bounded $g\in\mathscr{D}(D)$. Let c be the median of g. Set $f=\mathrm{sgn}(g-c)|g-c|^2$. Then f has median 0. By using the functional form of I_ν:

$$I_\nu = \inf\left\{\frac{\frac12\int J^{(1/2)}(\mathrm{d}x,\mathrm{d}y)|f(y)-f(x)|}{\inf_{c:\ c\text{ is a median of }f}\|f-c\|_{\nu/(\nu-1)}} :\ f\in L^1(\pi)\text{ is non-constant}\right\} \tag{5.13}$$

which will be proved later, we obtain

$$\|g-c\|_{2q}^2 = \|f\|_q \leqslant \frac12 I_\nu^{-1}\int J^{(1/2)}(\mathrm{d}x,\mathrm{d}y)|f(y)-f(x)|. \tag{5.14}$$

On the other hand, since

$$|a-b|\,(|a|+|b|) = \begin{cases} |a^2-b^2|, & \text{if } ab>0\\ (|a|+|b|)^2, & \text{if } ab<0,\end{cases}$$

we have $|f(y)-f(x)|\leqslant |g(y)-g(x)|\big(|g(y)-c|+|g(x)-c|\big)$. By using this inequality and following the proof of (3.5), we get

$$\begin{aligned}&\int J^{(1/2)}(\mathrm{d}x,\mathrm{d}y)|f(y)-f(x)|\\ &\quad\leqslant \sqrt{2D(g)}\left[\int J^{(1)}(\mathrm{d}x,\mathrm{d}y)[|g(y)-c|+|g(x)-c|]^2\right]^{1/2}\\ &\quad\leqslant 2\sqrt{2D(g)}\ \|g-c\|_2.\end{aligned} \tag{5.15}$$

Combining (5.14) with (5.15) together, we get

$$\|g-c\|_{2q}^2 \leqslant 2I_\nu^{-1}\sqrt{2D(g)}\ \|g-c\|_2.$$

On the other hand, writing $g^2=g^{2/(\nu+1)}\cdot g^{2\nu/(\nu+1)}$ and applying Hölder inequality with $p'=(\nu+1)/2$ and $q'=(\nu+1)/(\nu-1)$, we obtain

$$\|g\|_2 \leqslant \|g\|_1^{1/(\nu+1)}\|g\|_{2q}^{\nu/(\nu+1)}.$$

From these facts, it follows that

$$\|g-c\|_2 \leqslant \Big[I_\nu^{-1}\sqrt{2D(g)}\ \|g-c\|_2\Big]^{\nu/2(\nu+1)}\|g-c\|_1^{1/(\nu+1)}.$$

Thus,

$$\|g-c\|_2^{2(1+2/\nu)} \leqslant 2I_\nu^{-2}D(g)\,\|g-c\|_1^{4/\nu}$$

and hence

$$\mathrm{Var}(g)^{1+2/\nu} \leqslant 2I_\nu^{-2}D(g)\,\|g\|_1^{4/\nu}.$$

We now return to prove (5.13). Denote by J_ν the right-hand side of (5.13). Set $q=\nu/(\nu-1)$ and ignore the superscript "(1/2)" everywhere for simplicity. Take $f=I_A$ with $0<\pi(A)\leqslant 1/2$. Then, f has a median 0. Moreover,

$$\int J(\mathrm{d}x,\mathrm{d}y)|f(y)-f(x)| = 2J(A\times A^c), \qquad \|f\|_q=\pi(A)^{1/q}.$$

This proves that $I_\nu \geqslant J_\nu$.

Conversely, fix f with median c. Set $f_\pm=(f-c)^\pm$. Then $f_+ + f_- = |f-c|$ and $|f(y)-f(x)| = |f_+(y)-f_+(x)|+|f_-(y)-f_-(x)|$. Put $F_t^\pm=\{f_\pm\geqslant t\}$. Then

$$\begin{aligned}\frac{1}{2}\int J(\mathrm{d}x,\mathrm{d}y)|f(y)-f(x)| &= \frac{1}{2}\int J(\mathrm{d}x,\mathrm{d}y)\big[|f_+(y)-f_+(x)|+|f_-(y)-f_-(x)|\big]\\ &= \int_0^{\|f\|_u}\big[J\big(F_t^+\times\big(F_t^+\big)^c\big)+J\big(F_t^-\times\big(F_t^-\big)^c\big)\big]\mathrm{d}t\\ &\qquad\text{(by co-area formula)}\\ &\geqslant I_\nu\int_0^{\|f\|_u}\big[\pi(F_t^+)^{1/q}+\pi(F_t^-)^{1/q}\big]\mathrm{d}t.\end{aligned}$$

Next, we need the following simple result:

Claim. Let $p\geqslant 1$. Then $\|f\|_p\leqslant F$ iff $\|fg\|_1\leqslant FG$ holds for all g satisfying $\|g\|_q\leqslant G$.

It follows that

$$\pi(F_t^\pm)^{1/q}=\big\|I_{F_t^\pm}\big\|_q=\sup_{\|g\|_r\leqslant 1}\big\langle I_{F_t^\pm},\,g\big\rangle, \qquad \frac{1}{r}+\frac{1}{q}=1.$$

Thus, for every g with $\|g\|_r\leqslant 1$, we have

$$\begin{aligned}\frac{1}{2}\int J(\mathrm{d}x,\mathrm{d}y)|f(y)-f(x)| &\geqslant I_\nu\int_0^\infty\big[\langle I_{F_t^+},\,g\rangle+\langle I_{F_t^-},\,g\rangle\big]\mathrm{d}t\\ &= I_\nu\big[\langle f_+,\,g\rangle+\langle f_-,\,g\rangle\big]=I_\nu\langle|f-c|,\,g\rangle.\end{aligned}$$

Making supremum with respect to g, we get

$$\frac{1}{2}\int J(\mathrm{d}x,\mathrm{d}y)|f(y)-f(x)|\geqslant I_\nu\|f-c\|_q. \quad \square$$

Part VI. New Picture of Ergodic Theory

Consider a continuous-time, irreducible Markov chain with transition probability matrix $P(t) = (p_{ij}(t))$ on a countable state space E with stationary distribution $(\pi_i > 0 : i \in E)$. There are traditionally three types of ergodicity:

Ordinary ergodicity $\lim\limits_{t\to\infty} p_{ij}(t) = \pi_j, \qquad i, j \in E$

Exponentially ergodicity $|p_{ij}(t) - \pi_j| \leqslant C_{ij} e^{-\varepsilon t},\ t > 0,\ i, j \in E,\ C_{ij} \geqslant 0,\ \varepsilon > 0$

Strongly (or Uniformly) ergodicity $\lim\limits_{t\to\infty} \sup\limits_i |p_{ij}(t) - \pi_j| = 0.$

The relationship between the three types of ergodicity and the inequalities discussed above is given by the following diagram.

Theorem 6.1[Chen (1999d)]**.** For reversible Markov chains, the following implications hold.

$$
\begin{array}{ccccc}
 & & \text{Nash inequality} & & \\
 & \swarrow\!\!\swarrow & & \searrow\!\!\searrow & \\
\text{Logarithmic Sobolev inequality} & & & & \text{Strong ergodicity} \\
\Downarrow & & & & \Downarrow \\
\text{Poincaré inequality} & & \Longleftrightarrow & & \text{Exponential ergodicity} \\
 & & \Downarrow & & \\
 & & L^2\text{-algebraic ergodicity} & & \\
 & & \Downarrow & & \\
 & & \text{Ordinary ergodicity} & &
\end{array}
$$

Here L^2-algebraic ergodicity means that

$$\mathrm{Var}(P_t f)\big(= \|P_t f - \pi(f)\|^2\big) \leqslant CV(f)/t^{q-1}, \qquad t > 0. \tag{6.1}$$

holds for some $V : L^2(\pi) \to [0, \infty]$ satisfying

$$V(c_1 f + c_2) = c_1^2 V(f) \qquad \text{for all constants } c_1, c_2 \in \mathbb{R}. \tag{6.2}$$

and $V(f) < \infty$ for all functions f with finite support.

The proofs given below show that the diagram is indeed available for very general Markov processes except the equivalence for which some restriction is necessary.

To prove Theorem 6.1, we need one more result which is fundamental in various applications of the inequalities discussed above.

Theorem 6.2[Chen (1991, 1992), Liggett (1989, 1991)]**.** Let V satisfy (6.2) and $(P_t)_{t\geqslant 0}$ be the semigroup determined by the symmetric form $(D, \mathscr{D}(D))$.

(1) Let $V(P_t f) \leqslant V(f)$ for all $t \geqslant 0$ and $f \in L^2(\pi)$ (which is automatic when $V(f) = \|f\|_r^2$). Then (6.1) is implied by

$$\textit{Liggett-Stroock inequality}: \mathrm{Var}(f) \leqslant CD(f)^{1/p} V(f)^{1/q}, \qquad f \in L^2(\pi) \tag{6.3}$$

(2) Conversely, (6.1) $\Longrightarrow$ Liggett-Stroock inequality.

(3) Poincaré inequality $\Longleftrightarrow$ $\mathrm{Var}(P_t f) \leqslant \mathrm{Var}(f) \exp[-2\lambda_1 t]$.

Proof. Here we prove parts (1) and (2) only. The proof of part (3) is similar (cf. [Chen (1992); Theorem 9.1]).

a) Assume that (6.3) holds. Let $f \in \mathscr{D}(D)$ and $\pi(f) = 0$. Then $f_t := P(t)f \in \mathscr{D}(D)$. Set $F_t = \pi\big(f_t^2\big)$. Since

$$F_t' = -2D(f_t) \leqslant -2C^{-p}V(f)^{-p/q}\|f_t\|^{2p} = -2C^{-p}F_t^pV(f)^{-p/q}.$$

Now, part (1) follows from Corollary A4.

b) Conversely, since the process is reversible, the spectral representation theorem gives us

$$\frac{1}{t}(f - P(t)f,\, f) \uparrow D(f), \qquad \text{as } \ t \downarrow 0.$$

Hence

$$\|f\|^2 - tD(f) \leqslant (P(t)f,\, f) \leqslant \|P(t)f\|\,\|f\| \leqslant \|f\|\sqrt{C\,V(f)t^{1-q}}, \quad \pi(f) = 0.$$

Put $A = D(f)$, $B = \|f\|\sqrt{C\,V(f)}$ and $C_1 = \|f\|^2$. It follows that $C_1 - At \leqslant Bt^{(1-q)/2}$. The function $h(t) := At + Bt^{(1-q)/2} - C_1$ ($\geqslant 0$ for all $t > 0$) achieves its minimum

$$h(t_0) = \left[\left(\frac{q-1}{2}\right)^{2/(q+1)} + \left(\frac{2}{q-1}\right)^{(q-1)/(q+1)}\right]A^{(q-1)/(q+1)}B^{2/(q+1)} - C_1$$

at the point

$$t_0 = \left[\frac{2A}{B(q-1)}\right]^{-2/(q+1)} > 0.$$

Now, since $h(t_0) \geqslant 0$, it follows that $\|f\|^2 \leqslant C_2\, D(f)^{1/p}V(f)^{1/q}$ for some constant $C_2 > 0$ and so we have proved part (2). $\square$

Proof of Theorem 6.1. Here, we collect the proofs given in several papers.

Nash inequality $\Longrightarrow$ Strong ergodicity [Chen (1999b)]. Assume that Nash inequality holds. Then (6.1) holds with $V(f) = \|f\|_1^2$. We have

$$\|P_tf - \pi(f)\| \leqslant C\|f\|_1/t^{(q-1)/2}$$

and so

$$\|(P_t - \pi)f\| \leqslant C\|f - \pi(f)\|_1/t^{(q-1)/2}.$$

This means that the operator norm $\|P_t - \pi\|_{1\to 2}$ as a mapping from $L^1(\pi)$ to $L^2(\pi)$ is bounded above by $C/t^{(q-1)/2}$. Because of the symmetry of $P_t - \pi$, we get

$$\|P_t - \pi\|_{1\to\infty} \leqslant \|P_t - \pi\|_{1\to 2}\|P_t - \pi\|_{2\to\infty} = \|P_t - \pi\|_{1\to 2}^2.$$

Hence

$$\begin{aligned}\sup_x \|P_t(x,\cdot) - \pi\|_{\text{Var}} &= \sup_x \sup_{|f|\leqslant 1} |(P_t(x,\cdot) - \pi)f| \\ &\leqslant \sup_x \sup_{\|f\|_1\leqslant 1} |(P_t(x,\cdot) - \pi)f| \\ &= \|P_t - \pi\|_{1\to\infty} \\ &\leqslant C/t^{q-1} \to 0\end{aligned}$$

as $t \to \infty$. This proves strong ergodicity of $(P_t(x, \mathrm{d}y))$.

Nash inequality $\Longrightarrow$Logarithmic Sobolev inequality[Chen (1999b)]. Because $\|f\|_1 \leqslant \|f\|_p$ for all $p \geqslant 1$, Nash inequality $\Longrightarrow$Poincaré inequality and so we have $\lambda_1 > 0$. Next, as we have just proved above, $\|P_t - \pi\|_{1\to 2} < \infty$. Hence

$$\|P_t\|_{p\to 2} \leqslant \|P_t\|_{1\to 2} \leqslant \|P_t - \pi\|_{1\to 2} + \|\pi\|_{1\to 2} < \infty$$

for all $t > 0$ and $p \in [1, 2]$. These two facts are enough to claim Logarithmic Sobolev inequality (cf. [Bakry (1992); Theorem 3.6 and Proposition 3.9]).

Logarithmic Sobolev inequality $\Longrightarrow$ Poincaré inequality[Rothaus (1981)]. Actually, we have more precise result: $\lambda_1 \geqslant \sigma$. Consider $f = 1 + \varepsilon g$ for sufficient small ε. Then $D(f) = \varepsilon^2 D(g)$. Next, expand $f^2 \log f^2$ and $f^2 \log \|f\|^2$ in ε up to order 2. Then, we get

$$\int f^2 \log f^2/\|f\|^2 \mathrm{d}\pi = 2\varepsilon^2 \mathrm{Var}(g) + O(\varepsilon^3).$$

The proof can be done by using the definitions of λ_1 and σ and letting $\varepsilon \to 0$.

Poincaré inequality $\Longleftrightarrow$ L^2-exponential convergence. That is part (3) of Theorem 6.2.

L^2-exponential convergence $\Longrightarrow$ L^2-algebraic ergodicity. Simply take $V(f) = \|f\|^2$ in (6.1) and apply Theorem 6.2.

L^2-algebraic convergence $\Longrightarrow$ Ordinary ergodicity. Simply note that

$$\pi_i |p_{ik}(t) - \pi_k|^2 \leqslant \sum_j \pi_j |p_{jk}(t) - \pi_k|^2 \leqslant CV(I_{\{k\}})/t^{q-1} \to 0$$

as $t \to \infty$.

Strong ergodicity$\Longrightarrow$Exponential ergodicity $\Longrightarrow$Ordinary ergodicity. In the discrete case, the three types of ergodicity defined above are indeed equivalent to the following ones:

$$\begin{aligned} &\textit{Ordinary ergodicity}: && \lim_{t\to\infty} \|p_t(x, \cdot) - \pi\|_{\mathrm{Var}} = 0 \\ &\textit{Exponential ergodicity}: && \|p_t(x, \cdot) - \pi\|_{\mathrm{Var}} \leqslant C(x) e^{-\varepsilon t} \\ &\textit{Strong ergodicity}: && \lim_{t\to\infty} \sup_x \|p_t(x, \cdot) - \pi\|_{\mathrm{Var}} = 0. \end{aligned}$$

Then, the implications hold for general Markov processes when $\mathscr{E}$ is countably generated. In other words, if the Markov process corresponding to the semigroup (P_t) is irreducible and aperiodic in the Harris sense, then the implications hold. To see this, noting that by [Chen (1992); §4.4] and [Down et al (1995)], the continuous-time case can be reduced to the discrete-time one and then the conclusion follows from [Meyn & Tweedie (1993); Chapter 16].

L^2-exponential convergence$\Longrightarrow$Exponential ergodicity[Chen (1991)]. By assumption,

$$e^{-2\lambda_1 t} \|f - \pi(f)\|^2 \geqslant \pi_{i_0} |p_{i_0 j_0}(t) - \pi_{j_0}|^2$$

for the function $f_j = \delta_{jj_0}$ and arbitrary i_0 and j_0. Hence

$$|p_{ij}(t) - \pi_j| \leqslant \sqrt{\pi_j(1-\pi_j)/\pi_i}\, e^{-\lambda_1 t}$$

for all i, j, which proves exponential ergodicity.

This proof is not suitable for more general Markov processes. However, there is a stronger proof given in [Chen (1998b)].

Exponential ergodicity$\Longrightarrow$ L^2-exponential convergence [Chen (1998b)].

Step 1. Let $\Omega\varphi(i) = \sum_j q_{ij}(\varphi_j - \varphi_i)$ and set $A = \{0\}$ for simplicity. Then

$$\begin{aligned}
(p_{ij}(t)) \text{ is exponential ergodic } &\Longleftrightarrow \Omega\varphi \leqslant -\beta\varphi + CI_A \text{ for some constants} \\
&\qquad C \geqslant 0,\ \beta > 0 \text{ and function } \varphi \geqslant 1 \\
&\Longrightarrow \sup_{i\neq 0}(\Omega\varphi/\varphi)(i) \leqslant -\beta \\
&\Longrightarrow \lambda_0(A^c) \geqslant \beta > 0.
\end{aligned}$$

The last step needs some more work.

Step 2. We need only to show that $\lambda_1 \geqslant \lambda_0(A^c)$. For f with $\pi(f) = 0$, let $c = f_0$. Then $(f-c)(0) = 0$. Moreover,

$$\begin{aligned}
D(f) &= D(f-c) \\
&\geqslant \lambda_0(A^c)\|f-c\|^2 \\
&= \lambda_0(A^c)(\|f\|^2 + c^2) \\
&\geqslant \lambda_0(A^c)\|f\|^2 \\
\Longrightarrow \lambda_1 &\geqslant \lambda_0(A^c).
\end{aligned}$$

One may use compacts A instead of $\{0\}$. Then the proof of Step 1 still works. As for Step 2, one may use Theorem 4.1. Therefore, the result is meaningful for more general processes. $\square$

To illustrate the power of the above results, consider a regular birth-death process on $\mathbb{Z}_+$ with birth rates (b_i) and death rates (a_i). Then $J_{ij} = \pi_i b_i$ if $j = i+1$, $J_{ij} = \pi_i a_i$ if $j = i-1$ and $J_{ij} = 0$ otherwise.

Theorem 6.3. For birth-death process, take $r_{ij} = (a_i + b_i) \vee (a_j + b_j)\, (i \neq j)$. Then

(1) For Nash inequality, $I_\nu > 0$ for some $\nu \geqslant 1$ iff there exists a constant $c > 0$ such that

$$\frac{\pi_i a_i}{\sqrt{r_{i,i-1}}} \geqslant c\Big[\sum_{j\geqslant i}\pi_j\Big]^{(\nu-1)/\nu}, \qquad i \geqslant 1.$$

If so, we indeed have $I_\nu \geqslant c$.

(2) For logarithmic Sobolev inequality,

$$\xi^{(\infty)} > 0 \text{ iff } \inf_{i\geqslant 1} \frac{\pi_i a_i}{\sqrt{r_{i,i-1}}} \Big/ \Big(\sum_{j\geqslant i}\pi_j\Big)\sqrt{1-\log\sum_{j\geqslant i}\pi_j} > 0.$$

$$\kappa^{(\alpha)} > 0 \text{ iff } \inf_{i\geqslant 1} \frac{\pi_i a_i}{r_{i,i-1}^{\alpha}} \Big/ \Big(-\sum_{j\geqslant i}\pi_j\Big)\log\sum_{j\geqslant i}\pi_j > 0.$$

(3) For Poincaré inequality, $k^{(\alpha)} > 0$ iff there exists a constant $c > 0$ such that

$$\frac{\pi_i a_i}{[(a_i+b_i)\vee(a_{i-1}+b_{i-1})]^{\alpha}} \geqslant c\sum_{j\geqslant i}\pi_j, \qquad i\geqslant 1.$$

Then, we indeed have $k^{(\alpha)} \geqslant c$.

(4) The process is strong ergodic iff

$$S := \sum_{n=1}^{\infty} a_{n+1}^{-1}\Big\{1+\sum_{k=1}^{n} b_k\cdots b_n/a_k\cdots a_n\Big\} < \infty$$ [Zhang (1999), Zhang et al (1999)].

(5) The rate of the exponentially ergodic convergence coincides with λ_1 [Chen (1991)], which is described by Theorem 2.2.

To conclude this part, we compute the rates for the above inequalities and two types of ergodicity in the simplest situation.

Example 6.4. Let $E = \{0, 1\}$ and consider the Q-matrix

$$Q = \begin{pmatrix} -b & b \\ a & -a \end{pmatrix}.$$

Then the Nash constant

$$\eta = (a+b)\Big(\frac{a\wedge b}{a\vee b}\Big)^{1/q},$$

the logarithmic Sobolev constant

$$\sigma = \frac{2(a+b)(a\vee b - a\wedge b)}{\log[(a\vee b)/(a\wedge b)]},$$

the rates of L^2-exponential convergence, exponential ergodicity, strong ergodicity (must be exponential) are all equal to $\lambda_1 = a+b$.

Proof. a) Note that

$$P(t) = (p_{ij}(t)) = e^{tQ} = \frac{1}{a+b}\begin{pmatrix} a+be^{-\lambda_1 t} & b\big[1-e^{-\lambda_1 t}\big] \\ a\big[1-e^{-\lambda_1 t}\big] & b+ae^{-\lambda_1 t} \end{pmatrix}$$

and

$$\pi_0 = \frac{a}{a+b}, \qquad \pi_1 = \frac{b}{a+b}.$$

Hence

$$|p_{ij}(t) - \pi_j| \leqslant \frac{a \vee b \vee 1}{a+b} e^{-\lambda_1 t}.$$

This proves the last assertion.

b) Write

$$Q = (a+b)\begin{pmatrix} -\theta & \theta \\ 1-\theta & \theta - 1 \end{pmatrix},$$

where

$$\theta = \frac{b}{a+b}.$$

Therefore, it suffices to consider the Q-matrix

$$Q_1 = \begin{pmatrix} -\theta & \theta \\ 1-\theta & \theta - 1 \end{pmatrix}.$$

Without loss of generality, one may assume that $\theta \leqslant 1/2$, i.e., $b \leqslant a$.

c) By [Diaconis $ Saloff-Coste (1996)] or Chen (1997), for Q_1 , we have

$$\sigma = \frac{2(1-2\theta)}{\log(1/\theta - 1)}.$$

From this and b), it is easy to obtain the second assertion.

d) We now show that Nash inequality is equivalent to

$$\|f - \pi(f)\|^2 \leqslant \eta^{-1} D(f)^{1/p} \|f - c\|_1^{2/q}, \qquad f \in L^2(\pi), \tag{6.4}$$

where c is the median of f. To see this, replace f with $f - c$ in the original Nash inequality, we get (6.4). The inverse implication follows from $\|f - c\|_1 = \inf_\alpha \|f - \alpha\|_1 \leqslant \|f\|_1$.

Consider Q_1. Given a function f on $\{0, 1\}$. Without loss of generality, assume that $f_0 > f_1$. Since $\theta \leqslant 1/2$, the median of f is f_0. Set $g = f - f_0$. Then,

$$\begin{aligned} \|g\|_1 &= \theta |g_1| = \theta(f_0 - f_1). \\ \mathrm{Var}(g) &= \pi_1 g_1^2 + (\pi_1 g_1)^2 = \theta(1-\theta)(f_0 - f_1)^2, \\ D(g) &= \pi_0 q_{01}(g_1 - g_0)^2 = (1-\theta)\theta(f_0 - f_1)^2. \end{aligned}$$

Hence,

$$\eta = \inf_g \frac{D(g)^{1/p} \|g\|_1^{2/q}}{\mathrm{Var}(g)} = \left(\frac{\theta}{1-\theta}\right)^{1/q}$$

for Q_1. Applying b) again, we obtain the first assertion. $\square$

Acknowledgement The author acknowledges Prof. F. Y. Wang for fruitful cooperation and Prof. M. P. Qian for encouragement in writing this paper. Without their support, this paper would never be exist.

Appendix. Some Elementary Lemmas

An earlier version of the next result was appeared in [Li (1995)].

Lemma A1. Let u and v be two functions defined on $[a,b]\,(b\leqslant\infty)$. Suppose that

(1) u is non-negative and absolutely continuous with $u(0)>0$.
(2) v is local integrable.
Next, let $[c,d]\supset\{u(t):t\in[a,b]\}$ and suppose that
(3) $g:(c,d)\to(0,\infty)$ is non-decreasing.
(4) $G(u(a))+\int_a^t v(s)\mathrm{d}s\in[G(c),G(d)]$, where

$$G(u)=\int_{u_0}^{u}\frac{\mathrm{d}x}{g(x)},\qquad u,u_0\in(c,d),$$
$$G(c)=\lim_{u\to c}G(u),\qquad G(d)=\lim_{u\to d}G(u).$$

If
(5) $u'(t)\leqslant v(t)g(u(t)),\qquad$ a.e. t,

Then

$$u(t)\leqslant G^{-1}\bigg(G(u(a))+\int_a^t v(s)\mathrm{d}s\bigg),\qquad t\in[a,b]$$

where G^{-1} is the inverse function of G.

Remark A2.

(1) If $u(a)=0$, one may replace u by $u+M$ for some $M>0$, so condition $u(0)>0$ is not really a restriction.
(2) Condition (5) is equivalent to the integral form

$$u(t_2)-u(t_1)\leqslant\int_{t_1}^{t_2}v(s)g(u(s))\mathrm{d}s,\qquad t_1,t_2\in[a,b],\ t_2\geqslant t_1.$$

Actually, since g is local bounded, vg is local integrable. Then condition (5) is deduced by using the absolute continuity of integration.

*Proof of Lemma A*1. By condition (3), G is continuous and increasing. By conditions (2) and (4), it suffices to prove that

$$G(u(t))\leqslant G(u(a))+\int_a^t v(s)\mathrm{d}s.$$

Set

$$F(t)=G(u(t))-G(u(a))-\int_a^t v(s)\mathrm{d}s.$$

Then

$$g(u(t))F'(t)=u'(t)-v(t)g(u(t)),\qquad \text{a.e.}\,t.$$

By conditions (5) and (3), it follows that $F'(t)\leqslant0$, a.e. t. Therefore $F(t)\leqslant F(a)=0$, $t\in[a,b]$. $\square$

Corollary $A3$ **(Exponential form).** If a non-negative function u satisfies $u(0) > 0$ and $u'(t) \leqslant -\alpha u(t)$ on $[0,\infty)$ for some constant $\alpha > 0$, then $u(t) \leqslant u(0)e^{-\alpha t}$ for all $t \geqslant 0$.

Proof. Take $[a,b) = [0,\infty) = [c,d)$, $v(t) \equiv -\alpha$ and $g(x) = x$. Then

$$G(u) = \int_1^u \frac{1}{x}\mathrm{d}x = \log u$$

and $G^{-1}(u) = e^u$. Hence by Lemma A1, we have

$$u(t) \leqslant \exp[\log u(0) - \alpha t] = u(0)e^{-\alpha t}. \quad \square$$

Corollary $A4$ **(Algebraic form).** If a non-negative function u satisfies $u(0) > 0$ and $u'(t) \leqslant -\alpha u(t)^p$ on $(0,\infty)$ for some constants $\alpha > 0$ and $p > 1$, then

$$u(t) \leqslant \big(u(0)^{1-p} + (p-1)\alpha t\big)^{1-q},$$

where $1/p + 1/q = 1$.

Proof. Take $(a,b) = (0,\infty) = (c,d)$, $v(t) \equiv -\alpha$ and $g(x) = x^p$. Then

$$G(u) = \int_1^u x^{-p}\mathrm{d}x = \frac{1}{1-p}\big(u^{1-p} - 1\big)$$

and

$$G^{-1}(u) = [1 + (1-p)u]^{1/(1-p)}.$$

Hence by Lemma A1, we have

$$\begin{aligned} u(t) &\leqslant \left(1 + (1-p)\left[\frac{1}{1-p}\big(u(0)^{1-p} - 1\big) - \alpha t\right]\right)^{1/(1-p)} \\ &= \big(u(0)^{1-p} + (p-1)\alpha t\big)^{1-q}. \quad \square \end{aligned}$$

Corollary $A5$**.** If a non-negative function u satisfies $u(0) > 0$ and

$$u'(t) \leqslant \frac{1}{t(1-2t/\sigma)} u(t)\log u(t)$$

on $[\varepsilon, \sigma/2)$, then

$$u(t) \leqslant u(\varepsilon)^{t(1-2\varepsilon/\sigma)/\varepsilon(1-2t/\sigma)}.$$

Proof. Take $[a,b) = [\varepsilon,\sigma/2)$, $[c,d) = [1,\infty)$, $g(x) = x\log x$, $v(t) = \{t(1-2t/\sigma)\}^{-1}$ and $u_0 = u(\varepsilon)$. Then

$$\begin{aligned} G(u) &= \int_{u_0}^u \frac{\mathrm{d}x}{g(x)} \\ &= \int_{u_0}^u \frac{\mathrm{d}x}{x\log x} \\ &= \int_{u_0}^u \frac{\mathrm{d}(\log x)}{\log x} \\ &= \int_{\log u_0}^{\log u} \frac{\mathrm{d}y}{y} \\ &= \log\log u - \log\log u_0. \end{aligned}$$

Solving the equation $\log\log u - \log\log u_0 = x$, we get $G^{-1}(x) = u_0^{\exp[x]}$. Because

$$v(s) = \frac{1}{s(1-2s/\sigma)} = \frac{1}{s} + \frac{2/\sigma}{1-2s/\sigma}$$

$$\int_\varepsilon^t v(s)\mathrm{d}s = \log t - \log\varepsilon - \log(1-2t/\sigma) + \log(1-2\varepsilon/\sigma) = \log\frac{t(1-2\varepsilon/\sigma)}{\varepsilon(1-2t/\sigma)}.$$

By Lemma A1, we obtain

$$\begin{aligned} u(t) &\leqslant G^{-1}\bigg(G(u(\varepsilon)) + \int_\varepsilon^t v(s)\mathrm{d}s\bigg)\\ &= G^{-1}\bigg(\log\frac{t(1-2\varepsilon/\sigma)}{\varepsilon(1-2t/\sigma)}\bigg)\\ &= u_0^{\exp\big[\log[t(1-2\varepsilon/\sigma)/\varepsilon(1-2t/\sigma)]\big]}\\ &= u(\varepsilon)^{t(1-2\varepsilon/\sigma)/\varepsilon(1-2t/\sigma)}. \quad\square \end{aligned}$$

For the remainder of this part, we consider a Markov semigroup $\{P(t)\}_{t\geqslant 0}$ with weak operator Ω having domain

$$\mathscr{D}_w(\Omega) = \bigg\{f : \frac{\mathrm{d}}{\mathrm{d}t}P(t)f(x) = P(t)\Omega f(x) \text{ for all } x\in E \text{ and } t\geqslant 0\bigg\}$$

The next two results describe the exponential or algebraic decay of the semigroup in terms of its operator.

Lemma *A6* **(Exponential form).** Let $f\in\mathscr{D}_w(\Omega)$ and $\alpha>0$ be a constant. Then $P(t)f\leqslant e^{-\alpha t}f$ iff $\Omega f\leqslant -\alpha f$.

Proof. Let $f_t = P(t)f$. Then

$$f_t' = P(t)\Omega f \leqslant -\alpha P(t)f = -\alpha f_t.$$

The sufficiency now follows from Corollary A3. The necessity follows from

$$\Omega f = \lim_{t\to 0}\frac{P(t)f-f}{t} \leqslant \lim_{t\to 0}\frac{e^{-\alpha t}-1}{t}f = -\alpha f. \quad\square$$

Lemma *A7* **(Algebraic form).** Fix $p>1$. Let $f\in\mathscr{D}_w(\Omega)$, $f\geqslant 0$ and $C>0$ be a constant. Then $P(t)f\leqslant[f^{1-p}+(p-1)Ct]^{1-q}$ iff $\Omega f\leqslant -Cf^p$.

Proof. Again, let $f_t = P(t)f$. Then $f_t' = P(t)\Omega f\leqslant -CP(t)(f^p)$. However, by Hölder inequality, $P(t)(f^p)\geqslant (P(t)f)^p$. Hence $f_t'\leqslant -Cf_t^p$. The sufficiency now follows from Corollary A4. Next, note that $p-1=p/q$ and $q-1=q/p$. The necessity follows from

$$\begin{aligned} \Omega f &= \lim_{t\to 0}\frac{P(t)f-f}{t}\\ &\leqslant \lim_{t\to 0}\frac{\big[f^{1-p}+(p-1)Ct\big]^{1-q}-f}{t}\\ &= \lim_{t\to 0}(1-q)(p-1)C\big[f^{1-p}+(p-1)Ct\big]^{-q}\\ &= -Cf^{q(p-1)}\\ &= -Cf^p. \quad\square \end{aligned}$$

References

1. Bakry, D. (1992), *L'hypercontractivité et son utilisation en théorie des semigroupes*, LNM, Springer-Verlag, **1581**.
2. Chen, M. F. (1991), *Exponential L^2-convergence and L^2-spectral gap for Markov processes*, Acta Math. Sin. New Ser. 7:1, 19–37.
3. Chen, M. F. (1992), *From Markov Chains to Non-Equilibrium Particle Systems*, World Scientific.
4. Chen, M. F. (1994), *Optimal Markovian couplings and application to Riemannian geometry, in Prob. Theory and Math. Stat., Eds. B. Grigelionis et al,* 121–142,, VPS/TEV.
5. Chen, M. F.(1994), *Optimal Markovian couplings and applications*, Acta Math. Sin. New Ser.10:3, 260–275.
6. Chen M. F. (1996), *Estimation of spectral gap for Markov chains*, Acta Math. Sin. New Ser. 12:4, 337–360.
7. Chen, M. F. (1997), *Coupling, spectral gap and related topics*, (I): Chin. Sci. Bulletin, 42:14, 1472–1477 (Chinese Edition); 42:16, 1321–1327 (English Edition). (Ⅱ): 42:15, 1585–1591 (Chinese Edition); 42:17, 1409–1416 (English Edition). (Ⅲ): 42:16, 1696–1703 (Chinese Edition); 1997, 42:18, 1497–1505 (English Edition).
8. Chen M. F. (1998a), *Trilogy of couplings and general formulas for lower bound of spectral gap*, in "Probability Towards 2000", Edited by L. Accardi and C. Heyde, Lecture Notes in Statistics, Vol. **128**, 123–136, Springer-Verlag.
9. Chen, M. F. (1998b), *Equivalence of exponential ergodicity and L^2-exponential convergence for Markov chains*, to appear in Stoch. Proc. & Appl..
10. Chen (1999a), *Analytic proof of dual variational formula for the first eigenvalue in dimension one*, Sci. Sin. (A) 42:8, 805–815.
11. Chen, M. F. (1999b), *Nash inequalities for general symmetric forms*, Acta Math. Sin. Eng. Ser. 15:3, 353–370.
12. Chen, M. F. (1999c), *Logarithmic Sobolev inequality for symmetric forms*, To appear in Sci. Sin. (A).
13. Chen, M. F. (1999d), *A new story of ergodic theory*, to appear in Proceedings of IMS Workshop on Applied Probability, Intern. Press, Hong Kong.
14. Chen, M. F. (1999e), *Eigenvalues, inequalities and ergodic theory*, to appear in Chin. Sci. Bull.
15. Chen, M. F. and Li, S. F. (1989), *Coupling methods for multi-dimensional diffusion processes*, Ann. Prob. 17:1, 151–177.
16. Chen, M. F. and Wang, F. Y. (1993), *Application of coupling method to the first eigenvalue on manifold*, Sci. Sin.(A), 23:11(1993) (Chinese Edition), 1130–1140, 37:1(1994) (English Edition), 1–14.
17. Chen, M. F. and Wang, F. Y. (1997a), *Estimates of logarithmic Sobolev constant — An improvement of Bakry–Emery criterion*, J. Funct. Anal. 144:2, 287–300.
18. Chen, M. F. and Wang, F. Y. (1997b), *Estimation of spectral gap for elliptic operators*, Trans. Amer. Math. Soc. 349, 1239-1267.
19. Chen, M. F. and Wang, F. Y. (1998), *Cheeger's inequalities for general symmetric forms and existence criteria for spectral gap*, To appear in Ann. Prob. Abstract. Chin. Sci. Bulletin 43:14, 1475–1477 (Chinese Edition); 43:18, 1516–1519 (English Edition).
20. Diaconis, P. and Saloff-Coste, L. (1996), *Logarithmic Sobolev inequalities for finite Markov chains*, Ann. Appl. Prob. 6:3, 695–750.
21. Down, D., Meyn, S. P. and Tweedie, R. L. (1995), *Exponential and uniform ergodicity of Markov processes*, Ann. Prob. 23, 1671–1691.
22. Gross, L. (1976), *Logarithmic Sobolev inequalities*, Amer. J. Math. 97, 1061–1083.
23. Li, Y. (1995), *Ergodicity for a class of translation-invariant infinite-dimensional reaction-diffusion processes*, Ann. Math. Sin. 16A:2, 223–229.
24. Liggett, T. M. (1989), *Exponential L_2 convergence of attractive reversible nearest particle systems*, Ann. Prob. 17, 403-432.
25. Liggett, T. M. (1991), *L_2 rates of convergence for attractive reversible nearest particle systems: the critical case*, Ann. Prob. 19:3, 935–959.

26. Lindvall, T. and Rogers, L. C. G. (1986), *Coupling of multidimensional diffusion processes*, Ann. Prob. 14:3, 860–872.
27. Meyn, S. P. & Tweedie, R. L. (1993), *Markov chains and Stochastic Stability*, Springer-Verlag.
28. Rothaus, O. (1981), *Diffusion on compact Riemannian manifolds and logarithmic Sobolev inequalities*, J. Finct. Anal. 42, 102–109.
29. Saloff-Coste, L. (1997), *Lectures on finite Markov chains*, LNM **1665**, 301–413, Springer-Verlag.
30. Sinclair, A. (1993), *Algorithms for Random Generation and Counting: A Markov Chain Approach*, Birkhäuser.
31. Wang, F. Y. (1998a), *Logarithmic Sobolev inequalities: conditions and counterexamples*, to appear in J. Operator Theory.
32. Wang, F. Y. (1998b), *Sobolev type inequalities for general symmetric forms*, to appear in Proc. Amer. Math. Soc.
33. Wang, F. Y. (1999), *Functional inequalities for empty essential spectrum*, to appear in J. Funct. Anal.
34. Zhang, Y. H. (1999), *Strong ergodicity for continuous-time Markov chains*, preprint.
35. Zhang, H. J., Lin, X. and Hou, Z. T. (1998), *Uniformly polynomial convergence for standard transition functions*, preprint.

Dept. of Math., Beijing Normal University, Beijing 100875, PRC.

Acta Math. Sin. Eng. Ser. 2000, 16:3, 361-368

The principal eigenvalue for jump processes

Mu-Fa Chen

(Department of Mathematics, Beijing Normal University, Beijing 100875, P.R. China)
(E-mail: mfchen@bnu.edu.cn Received November 15, 1999; accepted March 9, 2000)

Abstract A variational formula for the lower bound of the principal eigenvalue of general Markov jump processes is presented. The result is complete in the sense that the condition is fulfilled and the resulting bound is sharp for Markov chains under some mild assumptions.

Keywords Principal eigenvalue, jump processes, variational formula for Dirichlet form

1991 MR Subject Classification 60J75, 47A75

1. Introduction.

Let $(E, \mathscr{E})$ be a general measurable space satisfying $\{x\} \in \mathscr{E}$ for all $x \in E$. Given a q-pair $(q(x), q(x, \mathrm{d}y))$ (i.e., $q(x, \mathrm{d}y)$ is a non-negative measurable kernel, $q \geqslant 0$ is a measurable function and moreover $q(x) \geqslant q(x, E) = q(x, E \setminus \{x\})$ for all $x \in E$), denote by $r(x) := q(x) - q(x, E)$ the non-conservative quantity of the q-pair at $x \in E$. Refer to [1] for general terminology, notations and results about jump processes. Suppose that the q-pair is reversible with respect to a probability π, i.e., $\pi_q(\mathrm{d}x, \mathrm{d}y)(\mathrm{d}x, \mathrm{d}y) := \pi(\mathrm{d}x)q(x, \mathrm{d}y)$ is a symmetric measure on $\mathscr{E} \times \mathscr{E}$. Denote by $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ respectively the norm and inner product in $L^2(\pi)$. Let

$$D(f) = \frac{1}{2} \int \pi_q(\mathrm{d}x, \mathrm{d}y)[f(y) - f(x)]^2 + \int \pi_r(\mathrm{d}x) f(x)^2$$

where $\pi_r(\mathrm{d}x) = r(x)\pi(\mathrm{d}x)$. Next, set

$$\|f\|_D^2 = \|f\|^2 + D(f), \qquad E_n = \{x \in E : \ q(x) \leqslant n\}, \quad n \geqslant 1,$$
$$\mathscr{D}_0 = \{f \in L^2(\pi) : f \text{ vanishes out of some } E_n\}.$$

It is easy to check that $\|f\|_D < \infty$ for all $f \in \mathscr{D}_0$ (Lemma 3.1). Let $\mathscr{D}(D)$ be the completion of $\mathscr{D}_0$ with respect to $\|\cdot\|_D$. Note that for the bounded q-pair (i.e., $M := \sup_x q(x) < \infty$), $\mathscr{D}(D) = L^2(\pi) = \mathscr{D}_0$ since $E_n = E$ for all $n \geqslant M$.

The principal eigenvalue studied in the paper is defined by

$$\lambda_0 = \inf\{D(f) : f \in \mathscr{D}(D), \ \|f\| = 1\}.$$

Research supported in part by NSFC (No. 19631060), Math. Tian Yuan Found., Qiu Shi Sci. & Tech. Found., RFDP and MCME.

Actually, our original interest is in the case of $r = 0$. Then $\lambda_0 = 0$ since $1 \in \mathscr{D}(D)$ and $D(1) = 0$. In this case, instead of λ_0, we are interested in

$$\lambda_1 = \inf\{D(f) : f \in \mathscr{D}(D),\ \pi(f) = 0,\ \|f\| = 1\},$$

where $\pi(f) = \int f \mathrm{d}\pi$. However, as proved in [2], a criterion for $\lambda_1 > 0$ or some estimates of λ_1 can all be deduced from λ_0 for some $(D, \mathscr{D}(D))$ with $r \neq 0$. This explains the original purpose of the present study. Of course, the study on λ_0 also has an independent significance.

Recall that there is a one-to-one correspondence of the q-pair and the following operator:

$$\Omega f(x) = \int q(x, \mathrm{d}y) f(y) - q(x) f(x) = \int q(x, \mathrm{d}y)[f(y) - f(x)] - r(x) f(x)$$
$$\mathscr{D}(\Omega) = \left\{ f \in \mathscr{E} : \int q(x, \mathrm{d}y)|f(y)| + q(x)|f(x)| < \infty \text{ for all } x \in E \right\}.$$

The main result of the paper is a variational formula for the lower bound of λ_0, which can be stated as follows:

Theorem 1.1. We have $\lambda_0 \geqslant \sup_{0<g\in\mathscr{E}}\ \pi\text{-ess}\inf\,(-\Omega g/g)$. In words: if there is a positive function $g \in \mathscr{E}$ and a constant $\lambda > 0$ such that $\Omega g \leqslant -\lambda g$, π-a.s., then $\lambda_0 \geqslant \lambda$.

Theorem 1.1 improves on [2; Theorem 3.2] in the present context by removing some extra assumptions. The proof of the theorem is based on a variational formula for Dirichlet form (Theorem 2.1), it is a generalization of [3; Appendix 1, Theorem 10.2]. The proof of Theorem 1.1 is completed in two parts, presented in Sections 2 and 3 separately. We will show in Section 4 that Theorem 1.1 can often be sharp (i.e., $\lambda_0 = \lambda$) for Markov chains.

We remark that the condition "$g > 0$" in Theorem 1.1 can be replaced by "$g > 0$ π-a.s.". To see this, let $A_1 = [g = 0]$. By symmetry,

$$0 = \int_{A_1} \pi(\mathrm{d}x) q(x, E) = \int \pi(\mathrm{d}x) q(x, A_1).$$

Hence $A_2 := [q(\cdot, A_1) = 0]$ is a null set. Denote by A_3 the exceptional set of $\Omega g \leqslant -\lambda g$ and set $\tilde{g} = g I_{A_1^c} + I_{A_1}$. Then $\tilde{g} > 0$ and it is easy to check that $\Omega\tilde{g} \leqslant -\lambda\tilde{g}$ on $(A_1 \cup A_2 \cup A_3)^c$.

The reason why one can use the positive eigenfunction g is explained in the Appendix.

2. Variational formula for Dirichlet form. Proof of Theorem 1.1 (bounded case).

Throughout this section, assume that the q-pair is bounded: $M := \sup_x q(x) < \infty$. The original version of the next result is due to [3; Appendix 1, Theorem 10.2], in which the bounded q-pair with $r = 0$ is treated for countable E.

Theorem 2.1 (Variational formula for the Dirichlet form). Let $(q(x), q(x, \mathrm{d}y))$ be a bounded q-pair. Then for every non-negative $f \in L^2(\pi)$, we have

$$D(f) = \sup_g \langle f^2/g,\, -\Omega g\rangle \tag{2.1}$$

where g varies over all strictly positive (i.e., $g \geqslant c_g > 0$ for some constant c_g), bounded $\mathscr{E}$-measurable functions.

Proof. a) First, we prove that the right-hand side of (2.1) is controlled by the left-hand side. Because

$$\begin{aligned}
\int \pi(\mathrm{d}x)\frac{f(x)^2}{g(x)}(-\Omega g)(x) &= \int_{[f\neq 0]} \pi(\mathrm{d}x)\frac{f(x)^2}{g(x)}\left[q(x)g(x) - \int q(x,\mathrm{d}y)g(y)\right]\\
&\leqslant \int_{[f\neq 0]} \pi(\mathrm{d}x)\frac{f(x)^2}{\big(gI_{[f\neq 0]}\big)(x)}\\
&\qquad\times\left[q(x)\big(gI_{[f\neq 0]}\big)(x) - \int q(x,\mathrm{d}y)\big(gI_{[f\neq 0]}\big)(y)\right]\\
&= \langle f^2/\tilde g, -\Omega\tilde g\rangle,
\end{aligned}$$

where $\tilde g = gI_{[f\neq 0]}$, thus, we may replace g by $\tilde g$ in the present proof.

Define $h = (\tilde g/f)I_{[f\neq 0]}$ and denote by $p(t,x,\mathrm{d}y)$ the jump process determined uniquely by the bounded q-pair (cf. [1; Corollary 3.12]). The corresponding semigroup is denoted by $\{P_t\}_{t\geqslant 0}$. Then, by the symmetry of $\pi(\mathrm{d}x)p(t,x,\mathrm{d}y)$ (cf. [1; Theorem 6.7]), we have

$$\begin{aligned}
\langle f^2/\tilde g,\, P_t\tilde g\rangle &= \int_{[f\neq 0]} \pi(\mathrm{d}x)\frac{f(x)}{h(x)}\int_{[f\neq 0]} p(t,x,\mathrm{d}y)f(y)h(y)\\
&= \frac{1}{2}\int_{[f\neq 0]} \pi(\mathrm{d}x)\int_{[f\neq 0]} p(t,x,\mathrm{d}y)f(x)f(y)\left[\frac{h(x)}{h(y)} + \frac{h(y)}{h(x)}\right]\\
&\geqslant \int_{[f\neq 0]} \pi(\mathrm{d}x)f(x)P_tf(x) = \langle f,\, P_tf\rangle.
\end{aligned}$$

Here, we have used the fact that $\alpha + 1/\alpha \geqslant 2$ for all $\alpha > 0$. Hence

$$\frac{1}{t}\langle f^2/\tilde g,\, \tilde g - P_t\tilde g\rangle \leqslant \frac{1}{t}\langle f,\, f - P_tf\rangle.$$

It is standard from the spectral theory that the right-hand side increases to $D(f)$ as $t \downarrow 0$ (cf. [1; Section 6.7]). Thus, it remains to show that the left-hand side converges to $\langle f^2/\tilde g, -\Omega\tilde g\rangle$ as $t \downarrow 0$. This can be done by using the dominated convergence theorem and the following facts:

$$\begin{aligned}
&\frac{1}{t}(\tilde g - P_t\tilde g)(x) = \frac{1 - p(t,x,\{x\})}{t}\tilde g(x) - \frac{1}{t}\int_{[y\neq x]} p(t,x,\mathrm{d}y)\tilde g(y)\\
&\frac{1 - p(t,x,\{x\})}{t} \leqslant \frac{1 - e^{-q(x)t}}{t} \leqslant q(x) \leqslant M < \infty\\
&\frac{1}{t}\int_{[y\neq x]} p(t,x,\mathrm{d}y)\tilde g(y) \leqslant \frac{\|\tilde g\|_\infty}{t}p(t,x,\{x\}^c) \leqslant \|\tilde g\|_\infty\frac{1 - p(t,x,\{x\})}{t} \leqslant M\|\tilde g\|_\infty < \infty.
\end{aligned}$$

b) When $0 < c \leqslant f \leqslant C < \infty$ for some constants c and C, the inverse inequality holds since one can simply set $g = f$. The general situation can be proved by approximation. Let $f_n = n^{-1} + f \wedge n$. Then, by the reversibility and boundedness of the q-pair, we have

$$-\langle f^2/f_n,\, \Omega f_n\rangle = \frac{1}{2}\int \pi_q(\mathrm{d}x, \mathrm{d}y)\left[\frac{f(y)^2}{f_n(y)} - \frac{f(x)^2}{f_n(x)}\right][f_n(y) - f_n(x)] + \int \pi_r(\mathrm{d}x) f(x)^2.$$

Since

$$\left[\frac{f(y)^2}{f_n(y)} - \frac{f(x)^2}{f_n(x)}\right][f_n(y) - f_n(x)] \geqslant 0 \quad \text{and} \quad f_n(x) \to f(x),$$

by Fatou's lemma, it follows that

$$\varliminf_{n\to\infty} -\langle f^2/f_n,\, \Omega f_n\rangle \geqslant D(f).$$

This completes the proof. □

Because, in general, we have $D(f) \geqslant D(|f|)$ and the strict inequality can happen for some f, it follows that the condition "$f \geqslant 0$" in Theorem 2.1 cannot be removed.

The next result proves Theorem 1.1 in a special case.

Proposition 2.2. Let $(q(x), q(x, \mathrm{d}y))$ be a bounded q-pair. If there is a strictly positive function $g \in \mathscr{E}$ and a constant $\lambda > 0$ such that $\Omega g \leqslant -\lambda g$, π-a.s., then $\lambda_0 \geqslant \lambda$.

Proof. From the assumption, $-\Omega g/g \geqslant \lambda$, π-a.e., it follows that $\langle -\Omega g/g,\, f^2\rangle \geqslant \lambda\|f\|^2$. Since $D(f) \geqslant D(|f|)$, once g is bounded, the conclusion follows from Theorem 2.1 immediately. We now consider the general g. Let $g_n = g \wedge n$. Then, it is easy to check that

$$-\frac{\Omega g_n}{g_n}(x) \geqslant \begin{cases} (-\Omega g/g)(x), & \text{if } \ g(x) \leqslant n \\ 0, & \text{if } \ g(x) > n. \end{cases}$$

Therefore

$$\langle -\Omega g_n/g_n,\, f^2\rangle \geqslant \int_{[g\leqslant n]} \left(-\frac{\Omega g}{g}\right)(x) f(x)^2 \pi(\mathrm{d}x).$$

From the assumption, $-\Omega g/g \geqslant \lambda > 0$, the required assertion now follows by using the monotone convergence theorem. □

For discrete E and the bounded q-pair with $r = 0$, the above proof is the same as the one in [3; Appendix 3, Proposition 0.2]. The proof shows that $D(f) \geqslant \lambda$ for every f with $\|f\| = 1$. This leads to the conclusion "$\lambda_1 \geqslant \lambda$" made in the quoted proposition. Unfortunately, the conclusion is wrong in the case of $r = 0$. Otherwise, one deduces a contradiction: $0 = D(1) \geqslant \lambda > 0$. In other words, Proposition 2.2 has no meaning in the case of $r = 0$, because its assumption can never be satisfied and hence there is nothing that can be done about λ_1.

3. Proof of Theorem 1.1 (general case).

First, we prove a result used in the definition of $\mathscr{D}(D)$.

Lemma 3.1. For each $f \in \mathscr{D}_0$, we have $\|f\|_D < \infty$.

Proof. Take n such that $f|_{E_n^c} = 0$. Then, by the definition of $\mathscr{D}_0$, we have $\|f\| < \infty$. Next, from the symmetry of $\pi_q(\mathrm{d}x, \mathrm{d}y)$, we obtain

$$\begin{aligned}
D(f) &= \frac{1}{2}\int_{E_n\times E_n} \pi_q(\mathrm{d}x,\mathrm{d}y)[f(y)-f(x)]^2 + \int_{E_n\times E_n^c} \pi_q(\mathrm{d}x,\mathrm{d}y) f(x)^2 \\
&\quad + \int_{E_n} \pi_r(\mathrm{d}x) f(x)^2 \\
&\leqslant 2\int_{E_n\times E_n} \pi_q(\mathrm{d}x,\mathrm{d}y) f(x)^2 + \int_{E_n\times E_n^c} \pi_q(\mathrm{d}x,\mathrm{d}y) f(x)^2 + \int_{E_n} \pi_r(\mathrm{d}x) f(x)^2 \\
&\leqslant 2\int_{E_n} \pi(\mathrm{d}x) f(x)^2 q(x,E) + \int_{E_n} \pi(\mathrm{d}x) f(x)^2 r(x) \\
&\leqslant 2\int_{E_n} \pi(\mathrm{d}x) f(x)^2 q(x) \leqslant 2\|f\|^2 \sup_{x\in E_n} q(x) < \infty.
\end{aligned}$$

This gives us $\|f\|_D < \infty$. $\square$

The way in proving Theorem 1.1 is a localizing procedure reducing the general case to the one of Proposition 2.2. To do so, we need some preparations.

From now on, we fix a function g and a constant $\lambda > 0$, as given in Theorem 1.1.

Lemma 3.2. Let $F_m = \{x \in E : g(x) \geqslant 1/m\}$, $m \geqslant 1$, and $\mathscr{D}_1 = \{f I_{F_m} : f \in \mathscr{D}_0,\ m \geqslant 1\}$. Then, $\mathscr{D}_1$ is dense in $\mathscr{D}(D)$ in the norm $\|\cdot\|_D$.

Proof. Given $f \in \mathscr{D}(D)$, choose $\{f_n\} \subset \mathscr{D}_0$ so that $\|f_n - f\|_D \to 0$. Let $f_{nm} = f_n I_{F_m}$. Then,

$$\begin{aligned}
D(f_{nm}-f_n) &= \frac{1}{2}\int \pi_q(\mathrm{d}x,\mathrm{d}y)[f_{nm}(y)-f_n(y)-f_{nm}(x)+f_n(x)]^2 \\
&\quad + \int \pi_r(\mathrm{d}x)[f_{nm}(x)-f_n(x)]^2 \\
&= \frac{1}{2}\int_{F_m^c\times F_m^c} \pi_q(\mathrm{d}x,\mathrm{d}y)[f_n(y)-f_n(x)]^2 + \int_{F_m\times F_m^c} \pi_q(\mathrm{d}x,\mathrm{d}y) f_n(y)^2 \\
&\quad + \int_{F_m^c} \pi_r(\mathrm{d}x) f_n(x)^2 \\
&= \frac{1}{2}\int_{F_m^c\times F_m^c} \pi_q(\mathrm{d}x,\mathrm{d}y)[f_n(y)-f_n(x)]^2 + \int_{F_m^c} \pi(\mathrm{d}x) f_n(x)^2 q(x,F_m) \\
&\quad + \int_{F_m^c} \pi(\mathrm{d}x) f_n(x)^2 r(x) \\
&\leqslant \frac{1}{2}\int_{F_m^c\times F_m^c} \pi_q(\mathrm{d}x,\mathrm{d}y)[f_n(y)-f_n(x)]^2 + \int_{F_m^c} \pi(\mathrm{d}x) f_n(x)^2 q(x).
\end{aligned}$$

For each fixed n, since $\|f_n\|_D < \infty$, $F_m \uparrow E$ and $q(x)$ is bounded on the support of f_n, the right-hand side goes to zero as $m \to \infty$. From the triangle inequality, $\|f_{nm} - f\|_D \leqslant \|f_{nm} - f_n\|_D + \|f_n - f\|_D$, we can first choose a large enough n and then a large enough m so that $\|f_{nm} - f\|_D$ becomes arbitrarily small. $\square$

For each $B \in \mathscr{E}$, define a local q-pair $(q^B(x), q^B(x, \mathrm{d}y))$ and the corresponding operator Ω^B on $(B, B \cap \mathscr{E})$ as follows:

$$\begin{aligned} &q^B(x) = q(x), \qquad q^B(x, \mathrm{d}y) = q(x, \mathrm{d}y) I_B(y) \\ &\Omega^B f(x) = \int q^B(x, \mathrm{d}y) f(y) - q^B(x) f(x), \qquad x \in B. \end{aligned} \tag{3.1}$$

Lemma 3.3. Let g and λ be given by Theorem 1.1. Then for every $B \in \mathscr{E}$, $\Omega^B g \leqslant -\lambda g$, π-a.s. on B.

Proof. By assumption,

$$-\lambda g(x) \geqslant \Omega g(x) = \int q(x, \mathrm{d}y) g(y) - q(x) g(x) \geqslant \int_B q(x, \mathrm{d}y) g(y) - q^B(x) g(x) = \Omega^B g(x)$$

for all $x \in B$. $\square$

For each $n, m \geqslant 1$, let $G_{n,m} = E_n \cap F_m$ and define the q-pair $(q_{n,m}(x), q_{n,m}(x, \mathrm{d}y))$ and operator $\Omega_{n,m}$ as above (by setting $B = G_{n,m}$). Next, define

$$D_{n,m}(f) = \frac{1}{2} \int_{G_{n,m} \times G_{n,m}} \pi_{q_{n,m}}(\mathrm{d}x, \mathrm{d}y)[f(y) - f(x)]^2 + \int_{G_{n,m}} \pi_{r_{n,m}}(\mathrm{d}x) f(x)^2,$$

where

$$r_{n,m}(x) = q(x) - q(x, G_{n,m}) = r(x) + q(x, G_{n,m}^c), \qquad x \in G_{n,m}.$$

Corresponding to the bounded form $D_{n,m}$, we have

$$\begin{aligned} \lambda_0^{(n,m)} &:= \inf\{D_{n,m}(f) : f \in \mathscr{D}(D_{n,m}),\ \|f I_{G_{n,m}}\| = 1\} \\ &= \inf\{D_{n,m}(f) : \|f I_{G_{n,m}}\| = 1\} \end{aligned}$$

since $\mathscr{D}(D_{n,m}) = L^2(G_{n,m}, \pi)$ (the set of square-integrable functions on $G_{n,m}$ with respect to the measure $\pi|_{G_{n,m}}$). A simple computation shows that we also have

$$\lambda_0^{(n,m)} = \inf\{D(f) : f|_{G_{n,m}^c} = 0, \|f\| = 1\}, \tag{3.2}$$

since $D_{n,m}(f) = D(f)$ for every $f \in L^2(\pi)$ with $f|_{G_{n,m}^c} = 0$. In other words, $\lambda_0^{(n,m)}$ is the Dirichlet eigenvalue of the the q-pair on the domain $G_{n,m}$.

Lemma 3.4. $\lambda_0^{(n,m)}$ is decreasing in both n and m. Moreover, $\lim_{\substack{n\to\infty\\ m\to\infty}} \lambda_0^{(n,m)} = \lambda_0$.

Proof. The first assertion follows from (3.2) and the fact that $E_n \uparrow E$, $F_m \uparrow E$ as $n, m \to \infty$. Moreover, it is obvious that $\lambda_0^{(n,m)} \geqslant \lambda_0$.

Next, from the definition of λ_0, for every $\varepsilon > 0$, there is $f_\varepsilon \in \mathscr{D}(D)$ such that $\|f_\varepsilon\| = 1$ and $\lambda_0 \geqslant D(f_\varepsilon) - \varepsilon$. From Lemma 3.2, there exists a sequence $\{f_{nm}\} \subset \mathscr{D}_1$ so that $\|f_{nm} - f_\varepsilon\|_D \to 0$. Without loss of generality, we may also assume that $f_{nm}|_{G_{n,m}^c} = 0$ and $\|f_{nm}\| = 1$. Thus, for large enough n, m, we have $D(f_\varepsilon) \geqslant D(f_{nm}) - \varepsilon$. Hence

$$\lambda_0 \geqslant D(f_{nm}) - 2\varepsilon \geqslant \lambda_0^{(n,m)} - 2\varepsilon$$

by (3.2). Since ε is arbitrary, we have thus proved the required assertion. $\square$

It is now easy to complete the proof of Theorem 1.1:

Proof of Theorem 1.1. Applying Proposition 2.2 to q-pair $(q_{n,m}(x), q_{n,m}(x, \mathrm{d}y))$ on $(G_{n,m}, G_{n,m} \cap \mathscr{E})$ and using Lemma 3.3, it follows that $\lambda_0^{(n,m)} \geqslant \lambda$. Then, the required assertion follows from Lemma 3.4. $\square$

4. Markov chains.

In this section, we discuss Theorem 1.1 in the context of Markov chains which means that E is countable. Then, we use Q-matrix $Q = (q_{ij})$ instead of the q-pair: $q_{ij} \geqslant 0$ for all $i \neq j$,

$$\sum_{j\neq i} q_{ij} \leqslant q_i := -q_{ii} < \infty$$

for all $i \in E$. We show that the theorem is often sharp.

Proposition 4.1. Let E be countable. Suppose that

(1) λ_0 is attainable, i.e., there is $g \in \mathscr{D}(D)$, $g \neq 0$ such that $D(g) = \lambda_0\|g\|^2$. Then $g \geqslant 0$ and $\Omega g = -\lambda_0 g$. If moreover,

(2) $Q = (q_{ij})$ is irreducible, i.e., for each pair $\{i, j\}$, there exist $i_1 = i$, i_2, $\cdots$, $i_n = j$ such that $q_{i_1 i_2} > 0$, $\cdots$, $q_{i_{n-1}, i_n} > 0$,

then $g > 0$ and so the lower bound given by Theorem 1.1 is exact.

Proof. Because $D(f) \geqslant D(|f|)$, we must have $g \geqslant 0$. Next, fix $k \in E$ and let $\tilde g_k = g_k + \varepsilon$ for some $\varepsilon \in \mathbb{R}$, $\tilde g_i = g_i$ for $i \neq k$. Then $\tilde g \in \mathscr{D}(D)$ and $D(\tilde g) - D(g) \geqslant \lambda_0(\|\tilde g\|^2 - \|g\|^2)$ from (1). That is,

$$\varepsilon[-2(\Omega + \lambda_0)g(k) + (q_k - \lambda_0)\varepsilon] \geqslant 0.$$

This implies that $\Omega g(k) = -\lambda_0 g_k$ since ε is arbitrary, and then $\Omega g = -\lambda_0 g$ since k is arbitrary.

Because $g \neq 0$, we may assume that $g_k > 0$ for some k. If $q_{ik} > 0$, then

$$0 < q_{ik} g_k \leqslant \sum_{j\neq i} q_{ij} g_j = (q_i - \lambda_0) g_i$$

and so $q_i > \lambda_0$ and $g_i > 0$. By using the condition (2) and an inductive procedure, one may prove that $g_i > 0$ for all $i \in E$. $\square$

To conclude this section, we introduce an example for which the conditions of Proposition 4.1 do not hold but Theorem 1.1 still works with the exact estimate.

Example 4.2. Consider a Markov chain with state space $\mathbb{Z}_+$. For each $i \geqslant 1$, the chain jumps from 0 to i at rate β_i and from i to 0 at rate q_i. Assume that $\sum_{i\geqslant 1}\beta_i < \infty$ and $\sum_{i\geqslant 1}\beta_i/q_i < \infty$. Take $E = \{1, 2, \cdots\}$. Then the Q-matrix $Q = (q_{ij} : i, j \in \mathbb{Z}_+)$ restricted on E becomes $q_{ij} = 0$ for all $i \neq j$ and $r_i = q_i$ for all $i, j \in E$. Then, the conditions of Proposition 4.1 may be fail to hold. However, the assumption of Theorem 1.1 is always satisfied with the sharp estimate.

Proof. It is trivial to check that the chain has a stationary distribution $\pi_i = \pi_0\beta_i/q_i$, $i \geqslant 1$. Clearly, the restricted Q-matrix is reducible on E and hence condition (2) fails. Note that

$$D(f)/\|f\|^2 = \sum_{i\geqslant 1}\pi_i q_i f_i^2 \bigg/ \sum_{i\geqslant 1}\pi_i f_i^2.$$

Hence, $\lambda_0 = \inf_{i\geqslant 1} q_i$. From this, one sees that condition (1) of Proposition 4.1 does not hold except that there is some $k \geqslant 1$ such that $q_k = \inf_{i\geqslant 1} q_i$. In that case, the only solution is $g_k > 0$ and $g_i = 0$ for all i with $q_i > q_k$. Therefore, the solution to the equation $\Omega g = -\lambda_0 g$ is not positive everywhere. By the way, we mention that the solution is also different from $(f_i^* := \mathbb{E}_i e^{\lambda_0\tau_0} : i \geqslant 0)$ which is the minimal solution to the system

$$\sum_{j\geqslant 0} q_{ij}(f_j - f_i) = -\lambda_0 f_i$$

for $i \geqslant 1$ with boundary condition $f_0 = 1$ (actually, $f_i^* = (\lambda_0 - q_i)^{-1}$ for $i \geqslant 1$ and $f_0^* = 1$), where τ_0 is the hitting time at 0. Here in the last sentence, the Q-matrix $Q = (q_{ij} : i, j \in \mathbb{Z}_+)$ is assumed to be regular.

Next, since $\Omega f(i) \leqslant -\lambda_0 f_i$ iff $q_i f_i \geqslant \lambda_0 f_i$ $(i \geqslant 1)$, any positive sequence (f_i) is a solution to the inequality $\Omega g(i) \leqslant -\lambda_0 g_i$. This proves the last assertion. $\square$

Appendix: The positiveness of the eigenfunction.

In this appendix, we extend part (2) of Proposition 4.1 to the general case under the "irreducible" assumption, which is reasonable since the reducible case can be often reduced to some irreducible ones. The result also shows that the condition "$g > 0$" used in Theorem 1.1 is reasonable.

Proposition A1. Suppose that $\pi_q(A \times A^c) > 0$ for every $A \in \mathscr{E}$ with $\pi(A) > 0$ and $\pi(A^c) > 0$. Then

$$\lambda_0 = \inf\{D(f)\colon f \in \mathscr{D}(D),\ f > 0 \text{ and } \|f\| = 1\}. \tag{A1}$$

Proof. (a) Since $D(|f|) \leqslant D(f)$, it is trivial to add the condition "$f \geqslant 0$" to the original definition of λ_0. Next, if $A := \{f = 0\}$ is a null set, then as remarked at the end of Section 1, one may replace f by

$$\tilde{f} = f I_{A^c} + I_A > 0.$$

Thus, it is enough to prove $(A1)$ with "$f > 0$" replaced by "$f > 0$, π-a.s."

(b) Given $f \in \mathscr{D}(D)$, $\|f\| \neq 0$ and $\pi(A) > 0$, where $A = \{f = 0\}$. Certainly, $\pi(A^c) > 0$. We are going to construct a new function f_1 such that $f_1 > 0$ on the set $A^c \cup \big(\{x : q(x) \leqslant N_1\} \cap A\big)$ for some $N_1 \geqslant 1$ and

$$\frac{D(f_1)}{\|f_1\|^2} < \frac{D(f)}{\|f\|^2}.$$

Let $B_N = \{x : q(x) \leqslant N\} \cap A$ and define

$$f_1 = f_1(N, \varepsilon) = \varepsilon I_{B_N} + f I_{A^c},$$

where N and $\varepsilon > 0$ are constants to be determined later. Because $I_{B_N} \in \mathscr{D}_0 \subset \mathscr{D}(D)$ and $f = f I_{A^c} \in \mathscr{D}(D)$, we have $f_1 \in \mathscr{D}(D)$. Next, since

$$\pi_q(B_N \times B_N^c) \leqslant \pi_r(B_N) + \pi_q(B_N \times E) \leqslant N\pi(B_N) < \infty, \tag{A2}$$

we have

$$\begin{aligned}
D(f_1) &= \frac{1}{2}\int_{A^c\times A^c} \pi_q(\mathrm{d}x, \mathrm{d}y)[f(y) - f(x)]^2 + \int_{A^c\times B_N} \pi_q(\mathrm{d}x, \mathrm{d}y)[\varepsilon - f(x)]^2 \\
&\quad + \int_{A^c\times AB_N^c} \pi_q(\mathrm{d}x, \mathrm{d}y) f(x)^2 + \varepsilon^2 \pi_q(B_N \times AB_N^c) \\
&\quad + \int_{A^c} \pi_r(\mathrm{d}x) f(x)^2 + \varepsilon^2 \pi_r(B_N) \\
&= \frac{1}{2}\int_{A^c\times A^c} \pi_q(\mathrm{d}x, \mathrm{d}y)[f(y) - f(x)]^2 + \int_{A^c} f(x)^2 [\pi_q(\mathrm{d}x, A) + \pi_r(\mathrm{d}x)] \\
&\quad - 2\varepsilon \int_{A^c} f(x)\pi_q(\mathrm{d}x, B_N) + \varepsilon^2[\pi_r(B_N) + \pi_q(B_N \times B_N^c)] \\
&= D(f) - 2\varepsilon \int_{A^c} f(x)\pi_q(\mathrm{d}x, B_N) + \varepsilon^2[\pi_r(B_N) + \pi_q(B_N \times B_N^c)].
\end{aligned}$$

Since

$$\|f_1\|^2 = \|f\|^2 + \varepsilon^2 \pi(B_N), \qquad \|f_1\|^{-2} \leqslant \|f\|^{-2},$$

it follows that

$$\frac{D(f_1)}{\|f_1\|^2} - \frac{D(f)}{\|f\|^2} \leqslant \frac{\varepsilon}{\|f_1\|^2}\bigg\{\varepsilon[\pi_r(B_N) + \pi_q(B_N \times B_N^c)] - 2\int_{A^c} f(x)\pi_q(\mathrm{d}x, B_N)\bigg\}. \tag{A3}$$

Note that $\pi(A) > 0$ and $\pi(A^c) > 0$, $\pi_q(A^c \times A) > 0$ by assumption and $B_N \uparrow A$ as $N \to \infty$. There exists $N_1 \geqslant 1$ such that $\pi(B_{N_1}) > 0$ and $\pi_q(A^c \times B_{N_1}) > 0$. Therefore, we have

$$\int_{A^c} f(x)\pi_q(\mathrm{d}x, B_{N_1}) > 0$$

since $f > 0$ on A^c. Thus, by using (A2), it follows that one can choose small enough

$$\varepsilon_1 \in \big(0,\, 2^{-1} \wedge [N_1 \pi(B_{N_1})]^{-1}\big]$$

so that the right-hand side of (A3) is negative. This completes the construction of f_1.

(c) If $\pi(AB_{N_1}^c) = 0$, then we have already obtained that $f_1 > 0$, π-a.s. and so the proof is done. Otherwise, rewriting $B_1 = B_{N_1}$, replacing A with $A_1 = AB_1^c$ and repeating the above construction, we obtain the second function

$$f_2 = \varepsilon_2 I_{B_2} + f_1 I_{A_1^c} = \varepsilon_2 I_{B_2} + \varepsilon_1 I_{B_1} + f I_{A^c},$$

where

$$B_2 = \{x : q(x) \leqslant N_2\} \cap A_1 = \{x : N_1 < q(x) \leqslant N_2\} \cap A$$

for some $N_2 \geqslant 2$ having property $\pi(B_2) > 0$ and

$$\varepsilon_2 \in \big(0,\, 2^{-2} \wedge \varepsilon_1 \wedge [N_2 \pi(B_2)]^{-1}\big].$$

Moreover,

$$\frac{D(f_2)}{\|f_2\|^2} < \frac{D(f_1)}{\|f_1\|^2}.$$

Now, if $\pi(A_1 B_2^c) = 0$, then we have $f_2 > 0$, π-a.s. and so the proof is done again. Otherwise, we go on by the same procedure. At the n-th step, we have $\pi(B_n) > 0$,

$$\varepsilon_n \in \big(0,\, 2^{-n} \wedge \varepsilon_{n-1} \wedge [N_n \pi(B_n)]^{-1}\big], \qquad N_n \geqslant n$$

and

$$\frac{D(f_n)}{\|f_n\|^2} < \frac{D(f_{n-1})}{\|f_{n-1}\|^2}.$$

The construction will be stopped either in a finite number of steps, or we get at last

$$f_\infty = f I_{A^c} + \sum_{n=1}^{\infty} \varepsilon_n I_{B_n}$$

for some sequences $\{B_n\}$, $\{\varepsilon_n\}$ and moreover

$$\frac{D(f_\infty)}{\|f_\infty\|^2} < \cdots < \frac{D(f_2)}{\|f_2\|^2} < \frac{D(f_1)}{\|f_1\|^2} < \frac{D(f)}{\|f\|^2}.$$

In the latter case, since $B_n \subset \{x : N_{n-1} < q(x) \leqslant N_n\}$ and $N_n \geqslant n \to \infty$, we have $\sum_{n=1}^{\infty} B_n = A$ and hence $f_\infty > 0$ everywhere. Finally, because

$$\|f_\infty\|^2 = \|f\|^2 + \sum_{n \geqslant 1} \varepsilon_n^2 \pi(B_n) < \infty, \qquad f_n \in \mathscr{D}(D)$$

and

$$
\begin{aligned}
D(f_\infty - f_n) &= \frac{1}{2}\int \pi_q(\mathrm{d}x,\mathrm{d}y)\bigg[\sum_{k\geqslant n+1}\varepsilon_k\big(I_{B_k}(y)-I_{B_k}(x)\big)\bigg]^2 \\
&\quad + \int \pi_r(\mathrm{d}x)\bigg[\sum_{k\geqslant n+1}\varepsilon_k I_{B_k}(x)\bigg]^2 \\
&= \int_{\left(A^c+\sum_{\ell=1}^n B_\ell\right)\times\sum_{k\geqslant n+1}B_k}\pi_q(\mathrm{d}x,\mathrm{d}y)\bigg[\sum_{k\geqslant n+1}\varepsilon_k I_{B_k}(y)\bigg]^2 \\
&\quad + \frac{1}{2}\int_{\left(\sum_{k\geqslant n+1}B_k\right)\times\sum_{k\geqslant n+1}B_k}\pi_q(\mathrm{d}x,\mathrm{d}y)\bigg[\sum_{k\geqslant n+1}\varepsilon_k\big(I_{B_k}(y)-I_{B_k}(x)\big)\bigg]^2 \\
&\quad + \int \pi_r(\mathrm{d}x)\bigg[\sum_{k\geqslant n+1}\varepsilon_k I_{B_k}(x)\bigg]^2 \\
&= \sum_{k\geqslant n+1}\varepsilon_k^2\pi_q\bigg(B_k\times\bigg(A^c+\sum_{\ell=1}^n B_\ell\bigg)\bigg) \\
&\quad + \sum_{n+1\leqslant k<\ell}(\varepsilon_k-\varepsilon_\ell)^2\pi_q(B_k\times B_\ell) + \sum_{k\geqslant n+1}\varepsilon_k^2\pi_r(B_k) \\
&\leqslant \sum_{k\geqslant n+1}\varepsilon_k^2[\pi_q(B_k\times E)+\pi_r(B_k)] \\
&\leqslant \sum_{k\geqslant n+1}\varepsilon_k^2 N_k\pi(B_k) \\
&\leqslant \sum_{k\geqslant n+1}\varepsilon_k \to 0
\end{aligned}
$$

as $n\to\infty$, we obtain $f_\infty \in \mathscr{D}(D)$. $\square$

References

[1]. Chen, M F, *From Markov Chains to Non-Equilibrium Particle Systems*, World Scientific, Singapore, 1992.

[2]. Chen, M F, Wang, F Y, *Cheeger's inequalities for general symmetric forms and existence criteria for spectral gap*, Ann Probab 28(1), 2000, **Abstract**: Chin Sci Bull 1998, 43(18), 1516–1519.

[3]. Kipnis, C, Landim, C, *Scaling Limits of Interacting Particle Systems*, Springer-Verlag, Berlin, 1999.

Ann. Appl. Prob. 2003, 13:2, 604–627

ALGEBRAIC CONVERGENCE OF MARKOV CHAINS

Mu-Fa Chen and Ying-Zhe Wang

(Beijing Normal University)
Received January 2000; revised April 2002

Abstract. Algebraic convergence in L^2-sense is studied for general reversible Markov chains and especially for birth-death chains. Some criteria for the convergence are presented. The results are effective since the convergence region can be completely covered, as illustrated by two examples.

1. Introduction.

The paper is devoted to study algebraic (or polynomial) L^2-convergence for reversible Markov chains. Roughly speaking, we are looking for a slower convergence rather than the exponential one, for which there is a great deal of publications (see for instance [1], [6], [12] and the references within). Contrarily, the work on algebraic convergence is still limited, the readers are urged to refer to [6] (II), [8] and [13] for the background and the present status of the study on the topic. Additionally, a referee provides the recent preprints [10] and [11] in which the same topic is studied with different approach for time-discrete Markov processes.

Consider a reversible Markov process on a complete separable metric space $(E, \mathscr{E})$ with probability measure π. The process corresponds in a natural way a strongly continuous semigroup (P_t) on $L^2(\pi)$ with generator L and domain $\mathscr{D}(L)$. It is said that the process has algebraic convergence in L^2-sense if there exists a functional $V : L^2(\pi) \longrightarrow [0, \infty]$ and constants $C > 0$, $q > 1$ so that

$$\|P_t f - \pi(f)\|^2 \leqslant CV(f)/t^{q-1}, \quad t > 0, \ f \in L^2(\pi), \tag{1.1}$$

where $\|\cdot\|$ denotes the L^2-norm and $\pi(f) = \int f d\pi$.

The starting point of our study is the following result, taken from Liggett (1991), which provides some necessary and sufficient conditions for algebraic L^2-convergence.

2000 *Mathematics Subject Classification.* 60J27, 60F25.

Key words and phrases. Algebraic convergence, Markov chains, birth-death chains, coupling.

Research supported in part by NSFC (No. 19631060), Math. Tian Yuan Found., Qiu Shi Sci. & Tech. Found., RFDP and 973 Project.

Typeset by $\mathcal{A}_{\mathcal{M}}\mathcal{S}$-TEX

Theorem A (Liggett-Stroock). Let $1 < p, q < \infty$ such that $1/p + 1/q = 1$ and let $V : L^2(\pi) \longrightarrow [0, \infty]$ satisfy $V(cf + d) = c^2 V(f)$ for all constants c and d. Consider the following two statements:

(a) There exists a constant $C > 0$, may be different from (1.1), so that

$$\|f - \pi(f)\|^2 \leqslant C D(f)^{1/p} V(f)^{1/q} \qquad \text{for all } f \in \mathscr{D}(D), \tag{1.2}$$

where $D(f) := D(f, f)$ is the Dirichlet form of L with domain $\mathscr{D}(D)$.

(b) There exists a constant $C > 0$ so that (1.1) holds.

We have the following conclusions:

(1) If (a) holds and V satisfies the following contraction:

$$V(P_t f) \leqslant V(f), \qquad f \in L^2(\pi), \quad t > 0 \tag{1.3}$$

then (b) holds.

(2) If (b) holds then so does (a) if the process is reversible with respect to π.

Remark

(1) In condition (a), we use $D(f)$ instead of $-\int fLf$ appeared in Liggett (1991)([13]). The advantage of this was explained in Chen([2] , §6.7, §9.1).

(2) If $p = 1$, then the process is in fact exponentially convergent. Hence we restrict ourselves to the case of $p > 1 (\Longleftrightarrow q < \infty)$.

(3) If (a) is satisfied with

$$V(f) = \|f - \pi(f)\|^2, \qquad V(f) = -\int fLf d\pi \quad \text{or} \quad V(f) = D(f)^\alpha, \quad \alpha > -1/p,$$

then the algebraic L^2-convergence is indeed exponential. Thus, none of these choices for V is useful in the present context. We will adopt several different types of V, given in (1.4), (1.7), (1.9) and Theorem C below.

The main purpose of the paper is to work out some more explicit conditions for the Liggett-Stroock theorem in the context of Markov chains.

Let $Q = (q_{ij})$ be a regular and irreducible Q-matrix on a countable set E: $q_{ij} \geqslant 0 (i \neq j)$, $0 < q_i := -q_{ii} = \sum_{j \neq i} q_{ij} < \infty$. Assume that the corresponding Q-process $P(t) = (p_{ij}(t) : i, j \in E)$ is stationary having distribution (π_i), and $\pi_i q_{ij} = \pi_j q_{ji}$ for all $i, j \in E$. Then the corresponding operator

$$\Omega f(i) := \sum_j q_{ij}(f_j - f_i), \qquad i \in E$$

becomes symmetric on $L^2(\pi)$. Denote by $(D, \mathscr{D}(D))$ the Dirichlet form:

$$D(f) = D(f, f) = \frac{1}{2} \sum_{i,j} \pi_i q_{ij} (f_j - f_i)^2.$$

Its domain is assumed to be $\mathscr{D}(D) = \{f \in L^2(\pi) : D(f) < \infty\}$.

Following [9], we define a graph structure associated with the matrix $Q = (q_{ij})$. We call $\langle i, j\rangle$ an edge if $q_{ij} > 0\,(i \neq j)$. The adjacent edges

$$\langle i, i_1\rangle, \langle i_1, i_2\rangle, \dots, \langle i_n, j\rangle \quad (i, j \text{ and } i_k\text{'s are different})$$

consist a path from i to j. Assume that for each pair $i \neq j$, there exists a directional path from i to j. Choose and fix such a path γ_{ij}. We have linear order for the vertices on each path. Then, for $e \in \gamma_{ij}$, we may write $e = \langle e_\ell, e_r\rangle$, where e_ℓ and e_r are the left and right vertices of e respectively. Fix all the selected paths $\{\gamma_{ij}\}$. Define

$$\beta = \sup_i \#\{e: \ i \text{ is the left vertex of } e\}.$$

Next, choose a symmetric function $\phi: \phi_{ij} \equiv \phi(i,j) \geqslant 0$ and $\phi_{ij} = 0$ if and only if $i = j$. For instance, one may take ϕ_{ij} to be the geodesic distance between i and j on the graph. Then define

$$V_\delta(f) = \sup_{i\neq j}(f_j - f_i)^2/\phi_{ij}^{2\delta}, \tag{1.4}$$

where $\delta = 0$ or 1. Note that $V_0(f)$ is independent of ϕ but is meaningful only for bounded f. This is rather restrictive but still enough to deduce the ergodicity of the process under (1.1). As we will prove in Section 2, the contraction (1.3) is automatic for V_0. Next, a sufficient condition for (1.3) with $V = V_1$ is the following: There exists a coupling operator $\widetilde{\Omega}$ so that

$$\widetilde{\Omega}\phi(i,j) \leqslant 0, \quad \text{for all } i \neq j \quad \text{and} \quad \widetilde{\Omega}\phi(i,i) = 0 \text{ for all } i. \tag{1.5}$$

For the readers' convenience, we recall the definition of coupling operators. Because of the one-to-one correspondence of a Q-matrix and its operator just mentioned above, we need only to define the coupling Q-matrices. For a given Q-matrix $Q = (q_{ij})$, a coupling Q-matrix $(q_{(ij),(k\ell)} : (ij), (k\ell) \in E \times E)$ is described by the following marginality: $\sum_\ell q_{(ij),(k\ell)} = q_{ik}$ and $\sum_k q_{(ij),(k\ell)} = q_{j\ell}$. We refer to [2], Chapters 0, 5 and [3] for various coupling operators. Set

$$\sigma_1(e) = \sum_i \frac{\pi_i \phi_{i,e_\ell}^2}{\pi_{e_\ell}^2 q_{e_\ell e_r}} \bigg(\sum_{j:\gamma_{ij}\ni e} \pi_j/\phi_{ij}^2\bigg)^2, \qquad \sigma_2(e) = \frac{1}{\pi_{e_\ell} q_{e_\ell e_r}} \sum_{\{i,j\}:\gamma_{ij}\ni e} \pi_i\pi_j/\phi_{ij}^2,$$

where $\{i, j\}$ denotes the disordered pair of i and j. We remark that the summation appeared in the first formula varies only over the pairs $\{i, j\} : \gamma_{ij} \ni e$.

To state our result, we still need some notations. We say that the process has a finite range R if $q_{ij} = 0$ whenever $|j - i| > R$. We will use some function ρ on $E = \{0, 1, 2, \cdots\}$ having the property:

ρ is increasing, $\rho_0 = 0$ and there exists a constant c such that either $\rho_N \leqslant c\rho_{N/2}$ for all $N \geqslant 1$ or $\rho_{i+R} \leqslant c\rho_i$ for all $i \geqslant 1$ but still $\displaystyle\sum_{N\geqslant 1} \rho_{NR}^{-\epsilon}$ for all $\epsilon > 0$. (1.6)

A typical choice of ρ is $\rho_i = i^\alpha$ $(i \geqslant 1)$ for some constant $\alpha > 0$. Then condition "$\rho_N \leqslant c\rho_{N/2}$ for all $N \geqslant 1$" holds. Otherwise, let $\rho_i = \alpha^i$ $(i \geqslant 1)$ for some constant $\alpha > 1$. Then we do have "$\rho_{i+R} \leqslant c\rho_i$ for all $i \geqslant 1$" and "$\sum_{N\geqslant 1} \rho_{NR}^{-\epsilon}$ for all $\epsilon > 0$".

Now, we can state our first criterion as follows:

Theorem 1.1. (0) If (1.5) is satisfied, then (1.3) holds with $V = V_1$.

(1) Let (1.3) hold. If $\beta < \infty$,

$$\sup_e\{\sigma_1(e)+\sigma_2(e)\}<\infty \quad \text{and} \quad \sum_{i,j}\pi_i\pi_j\phi_{ij}^{2(q+\delta-1)}<\infty$$

for some constant $q>1$, then the Markov chain has algebraic decay with $V = V_\delta$ ($\delta = 0$ or 1) and the same q.

(2) Conversely, let $E=\{0,1,2,\dots\}$ and suppose that the process has algebraic decay with respect to $V_\delta(\delta=0,1)$ and $\phi_{ij}=|\rho_j-\rho_i|$ for some function ρ satisfying (1.6). If moreover $\sup_{i\geqslant 1}\sum_{j=i+1}^{\infty} q_{ij}\phi_{ij}^2<\infty$, then we have $\sum_j \rho_j^k\pi_j<\infty$ for all $k<2(q+\delta-1)$.

The next result is a straightforward consequence of, but more practical than, Theorem 1.1.

Corollary 1.2. Part (1) of Theorem 1.1 holds if $\sigma_1(e)$ and $\sigma_2(e)$ are replaced by

$$\sigma_1'(e)=\sup_i\frac{\phi_{i,e_\ell}}{\pi_{e_\ell}\sqrt{q_{e_\ell e_r}}}\sum_{j:\gamma_{ij}\ni e}\frac{\pi_j}{\phi_{ij}^2} \qquad \text{and} \qquad \sigma_2'(e)=\frac{1}{\pi_{e_\ell}q_{e_\ell e_r}}\sup_i\sum_{j:\gamma_{ij}\ni e}\frac{\pi_j}{\phi_{ij}^2}$$

respectively.

Clearly, algebraic convergence depends heavily on the functional V. We now introduce a different choice of V. Fix a reference point in E, say 0 for simplicity. For each $j\in E\setminus\{0\}$, choose a directional path from 0 to j, denoted by γ_{0j}. Fix the family $\{\gamma_{0j}:\ j\neq 0\}$ and define β as above. Next, choose $\phi:\ \phi_i>0$ for $i\neq 0$ and $\phi_0=0$. Define

$$\widetilde{V}_\delta(f)=\sup_{i\neq 0}\,(f(i)-f(0))^2/\phi_i^{2\delta}. \tag{1.7}$$

When $E=\{0,1,2,\dots\}$ and ϕ_i is increasing, then for $\phi_{ij}:=|\phi_j-\phi_i|$, it is easy to check that $\widetilde{V}_\delta(f)\leqslant V_\delta(f)$ for each $\delta=1$ or 0. Finally, set

$$\tilde\sigma_1(e)=\frac{\sqrt{\phi_{e_\ell}}}{\pi_{e_\ell}\sqrt{q_{e_\ell e_r}}}\sum_{j:\gamma_{0j}\ni e}\frac{\pi_j}{\phi_j^2}, \qquad \tilde\sigma_2(e)=\frac{1}{\pi_{e_\ell}q_{e_\ell e_r}}\sum_{j:\gamma_{0j}\ni e}\frac{\pi_j}{\phi_j^2}.$$

Theorem 1.3. (1) Let

$$\beta<\infty,\qquad \sup_e\{\tilde\sigma_1(e)+\tilde\sigma_2(e)\}<\infty \quad\text{and}\quad \sum_j\pi_j\phi_j^{2(q+\delta-1)}<\infty \tag{1.8}$$

for some $q>1$ and $\delta=0$ or 1. When $\delta=1$, suppose additionally that (1.3) holds with $V=\widetilde{V}_1$. Then the Markov chain has algebraic decay with $V=\widetilde{V}_1$ when $\delta=1$ and $V=V_0$ when $\delta=0$.

In particular, if $E=\{0,1,2,\dots\}$ and ϕ is increasing, whenever (1.5) holds for $\phi_{ij}:=|\phi_i-\phi_j|$, then condition (1.8) with $\delta=1$ implies the algebraic decay of the Markov chain with respect to V_1 defined by (1.4).

(2) Conversely, if $E = \{0, 1, 2, \dots\}$, ϕ is increasing and the Markov chain has algebraic decay with respect to $\widetilde{V}_\delta$ ($\delta = 0,\ 1$), then part (2) of Theorem 1.1 holds with $\rho = \phi$.

Finally, we consider positive recurrent birth-death processes. Then, we have $E = \{0, 1, 2, \dots\}$, birth rate $b_i > 0 (i \geqslant 0)$, death rate $a_i > 0 (i \geqslant 1)$ and reversible measure (π_i). Each edge has the form: $e = \langle k, k+1\rangle,\ k \geqslant 0$. Obviously, $\beta = 1$ and $R = 1$. Let u_n be a positive sequence and set $\phi_{ij} = |\sum_{k<j} u_k - \sum_{k<i} u_k|$. Then, we have

$$V_\delta(f) = \sup_{i \neq j} |f(i) - f(j)|^2/\phi_{ij}^{2\delta} = \sup_{k \geqslant 0} |f(k+1) - f(k)|^2/u_k^{2\delta} \tag{1.9}$$

and for $e = \langle k, k+1\rangle$,

$$\sigma_1(e) = \sum_{i \leqslant k-1} \frac{\pi_i \phi_{ik}^2}{\pi_k^2 b_k} \bigg(\sum_{j \geqslant k+1} \frac{\pi_j}{\phi_{ij}^2} \bigg)^2, \qquad \sigma_2(e) = \frac{1}{\pi_k b_k} \sum_{i \leqslant k} \pi_i \sum_{j \geqslant k+1} \frac{\pi_j}{\phi_{ij}^2},$$

$$\sigma_1'(e) = \sup_{i \leqslant k-1} \frac{\phi_{ik}}{\pi_k \sqrt{b_k}} \sum_{j \geqslant k+1} \frac{\pi_j}{\phi_{ij}^2}, \qquad \sigma_2'(e) = \sup_{i \leqslant k} \sum_{j \geqslant k+1} \frac{\pi_j}{\pi_k b_k \phi_{ij}^2}.$$

As a consequence of Theorem 1.1, we have the following result.

Corollary 1.4. (1) Suppose that $\sup_e\{\sigma_1(e) + \sigma_2(e)\} < \infty$ $\big($or sufficiently, $\sup_e\{\sigma_1'(e) + \sigma_2'(e)\} < \infty\big)$ and $\sum_{i,j} \pi_i \pi_j \phi_{ij}^{2(q+\delta-1)} < \infty$ for some constant $q > 1$. For V_1, suppose additionally that $b_n u_n - a_n u_{n-1}$ is non-increasing $(u_{-1} = 0)$. Then the birth-death process has algebraic decay with respect to V_δ.

(2) Conversely, suppose that the process has algebraic decay with respect to V_δ and $\phi_{ij} := |\rho_i - \rho_j|$ for some ρ satisfying (1.6) with $R = 1$. If moreover $\sup_i b_i u_i^2 < \infty$, then we have $\sum_j \rho_j^k \pi_j < \infty$ for all $k < 2(q + \delta - 1)$.

Let $\phi_n = \sum_{i<n} u_i$ and define $\theta_1(n) = \phi_n \pi_n^{-1} b_n^{-1/2} \sum_{k=n+1}^\infty \pi_k/\phi_k^2$. As a direct consequence of Theorem 1.3, we have the following result.

Corollary 1.5. (1) Suppose that the following conditions hold:

$$(a)\ \ \sup_n \theta_1(n) < \infty, \qquad (b)\ \liminf_{k\to\infty} \phi_k \sqrt{b_k} > 0, \qquad (c)\ \sum_n \pi_n \phi_n^{2(q+\delta-1)} < \infty.$$

For V_1, suppose additionally that $b_n u_n - a_n u_{n-1}$ is non-increasing $(u_{-1} = 0)$. Then the process has algebraic decay with respect to V_δ.

(2) Part (2) of Theorem 1.3 holds.

Remark

(1) Conditions (a) and (b) imply that $\sup_n \theta_2(n) < \infty$, where $\theta_2(n) = \big(\sum_{k=n+1}^\infty \pi_k/\phi_k^2\big)/(\pi_n b_n)$. In fact,

$$\liminf_{k\to\infty} \phi_k \sqrt{b_k} > 0 \Longleftrightarrow \sup_k \frac{1}{\phi_k \sqrt{b_k}} < \infty \Longleftrightarrow \sup_n \frac{\theta_2(n)}{\theta_1(n)} < \infty.$$

This plus condition (a) implies that $\sup_n \theta_2(n) < \infty$.

(2) Obviously, when $e = \langle n, n+1\rangle$, we have $\theta_1(n) \leqslant \sigma_1'(e)$ and $\theta_2(n) \leqslant \sigma_2'(e)$. Hence, $\sup_n\{\theta_1(n) + \theta_2(n)\} \leqslant \sup_e\{\sigma_1'(e) + \sigma_2'(e)\}$.

The next result is a special case of Corollary 1.5.

Corollary 1.6. The birth-death process has algebraic decay with respect to V_0 provided

$$\liminf_{n\to\infty} n\left(\frac{a_{n+1}}{b_n}-1\right)>1,\quad \limsup_{n\to\infty}\frac{1}{\pi_n}\sum_{k\geqslant n+1}\pi_k>0\ \Big(\text{ or } \liminf_{n\to\infty}\phi_n\sqrt{b_n}>0\Big)$$

$$\text{and}\quad \sup_n \frac{1}{\sqrt{b_n}\,n^{\alpha}\pi_n}\sum_{k\geqslant n+1}\pi_k<\infty$$

for some $\alpha>0$.

For birth-death chains, the algebraic convergence was studied by Liggett(1991) [13], as a tool to deal with the critical case of attractive reversible nearest particle systems. In order to compare our results with known ones, we introduce two theorems taken from Liggett(1991) as follows. The first result below was mentioned in the quoted paper without a proof. For completeness, we present a proof at the end of Section 3.

Theorem B. Let (u_k) and ϕ_{ij} be as above. Define V_1 as (1.4). Let $\sigma_n=\sum_{k=n}^{\infty}\pi_k/\pi_n$. If the following conditions hold:
1) $b_nu_n-a_nu_{n-1}$ is non-increasing $(u_{-1}=0)$; 2) $\inf_{i\geqslant 0}b_i>0$;
3) $\sup_n\sigma_n/n<\infty$; 4) $\sum_{n=0}^{\infty}u_n^2n^{2q}\pi_n<\infty$;
then the process has algebraic decay with respect to V_1.

The next result is due to Liggett [13] (Theorem 2.10 and Proposition 2.15):

Theorem C. Define

$$\bar V(f)=\sup_{i\neq j}\left(|f_i-f_j|/|i-j|\right)^2=\sup_k|f(k+1)-f(k)|^2$$

which is nothing new but V_1 with $\phi_{ij}=|i-j|$. If the following conditions hold,
1) $\inf_i b_i>0,\ \ a_i\geqslant b_i,\ \ \sup_i i(a_i-b_i)<\infty$. 2) $\sup_n\sigma(n)/n<\infty$.
3) $\sum_n n\pi_n<\infty$. 4) $\sum_n(\log(n+2))^{3/2}n^{2q}\pi_n<\infty$.
then the process has algebraic decay with respect to $\bar V$.

Conversely, suppose that the process has algebraic decay with respect to $\bar V$ and $\sup_i b_i<\infty$, then we have $\sum_k k^{\alpha}\pi_k<\infty$ for all $\alpha<2q$.

In general, the conditions of Theorem C is stronger than those of Corollary 1.5, as will be shown by Example 4.1, for which Theorem C is not available but Corollary 1.5 is exact. Roughly speaking, the conditions of Corollary 1.5 (resp.,Theorem 1.3) is stronger than those of Theorem B (resp., Theorem 1.1) since $\widetilde{V_\delta}\leqslant V_\delta$. However, the same example shows that in some situation, Corollary 1.5 gives us the power $q\in(1,\infty)$ which can be much larger than $q\in(1,3/2)$ provided by Theorem B. Among the corollaries, the conditions of Corollary 1.6 are the weakest but the corresponding conclusion (1.1) holds for a smaller class of functions. Besides, the two examples discussed in Section 4 are always (resp., partially) algebraically convergent with respect to V_0 (resp., V_1 or $\bar V$). We refer to Section 4 for details.

Finally, we examine a special birth-death process: $a_i = b_i = i^2$ for even number i and $a_i = b_i = i^{3/2}$ for odd i. It is easy to check that Corollary 1.6 fails for such an oscillation model. To handle it, we adopt the following comparison theorem: comparing the original process with the new one having $\tilde{a}_i = \tilde{b}_i = i^{3/2}$.

Theorem 1.7. Let $Q = (q_{ij})$ and $\widetilde{Q} = (\tilde{q}_{ij})$ be two Q-matrices, reversible with respect to the distributions (π_i) and $(\tilde{\pi}_i)$ respectively. Suppose that

$$\sup_{i\neq j} \tilde{\pi}_i\tilde{q}_{ij}/(\pi_i q_{ij}) < \infty \qquad \text{and} \qquad \sup_i \pi_i/\tilde{\pi}_i < \infty.$$

If moreover, the $\widetilde{Q}$-process has algebraic decay with respect to V_δ (resp., $\widetilde{V}_\delta$), then so does the Q-process provided it is V_δ (resp., $\widetilde{V}_\delta$)-contractive.

An immediate consequence of Theorem 1.7 is as follows. With respect to V_0 or $\widetilde{V}_0$, any local perturbation does not interfere the algebraic convergence.

Theorem 1.1 is proved in the next section. The other results are proved in Section 3. In the last section (§4), two examples are discussed to illustrate the power of the results obtained in the paper.

2. Proof of Theorem 1.1.

A) First, we prove (1.3) under (1.5). Obviously, $V_\delta(cf+d) = c^2 V_\delta(f)$ holds for all constants c and d.

Let $(x_t,\ y_t)$ be the Markov chain determined by the coupling operator $\widetilde{\Omega}$, starting from (i,j). Because $\widetilde{\Omega}\phi(i,j) \leqslant 0$ for all $i \neq j$ and $\widetilde{\Omega}\phi(i,i) = 0$ for all i, we have $E^{(i,j)}\phi_{x_t,y_t} \leqslant \phi_{ij}$ (For more details of couplings, refer to [2] , [3] , [6], [7]). Then,

$$\begin{aligned}
\left|\frac{P_t f(i) - P_t f(j)}{\phi_{ij}}\right|^2 &= \left|\frac{E^i f(x_t) - E^j f(y_t)}{\phi_{ij}}\right|^2 \\
&= \left|\frac{E^{(i,j)}(f(x_t) - f(y_t))}{\phi_{ij}}\right|^2 \\
&= \left|E^{(i,j)}\left[\frac{f(x_t) - f(y_t)}{\phi_{x_t,y_t}} \cdot \frac{\phi_{x_t,y_t}}{\phi_{ij}}\right]\right|^2 \\
&\leqslant \sup_{k,\ell\in E}\left|\frac{f(k) - f(\ell)}{\phi_{k\ell}}\right|^2 \left(\frac{E^{(i,j)}\phi_{x_t,y_t}}{\phi_{ij}}\right)^2 \\
&\leqslant V(f), \qquad i \neq j.
\end{aligned}$$

Making the supremum over all $i \neq j$ on the left-hand side yields $V_1(P_t f) \leqslant V_1(f)$.

Next, we prove that (1.3) always holds for V_0. Actually, for any coupled process (x_t, y_t), we have

$$\begin{aligned}
V_0(P_t f) &= \sup_{i\neq j} |P_t f(i) - P_t f(j)| \\
&= \sup_{i\neq j} |E^i f(x_t) - E^j f(y_t)| \\
&\leqslant \sup_{i\neq j} E^{(i,j)}|f(x_t) - f(y_t)| \\
&\leqslant \sup_{i\neq j} E^{(i,j)} V_0(f) = V_0(f), \quad t \geqslant 0.
\end{aligned}$$

However, the proof does not work when V_0 is replaced by $\widetilde{V}_0$ and so we do not consider the contraction for $\widetilde{V}_0$.

B) Next, we prove part (1) of the theorem. Some ideas of the proof are taken from [4] and [13]. Let f satisfy $\pi(f) = 0$ and $\|f\|^2 = 1$. Then, we have

$$\begin{aligned}
\mathrm{Var}_\pi(f) &= \frac{1}{2}\sum_{i,j} \pi_i\pi_j(f_j - f_i)^2 \\
&= \sum_{\{i,j\}} \pi_i\pi_j(f_j - f_i)^2 \\
&\leqslant \left\{\sum_{\{i,j\}} \pi_i\pi_j\left(\frac{f_j - f_i}{\phi_{ij}}\right)^2\right\}^{1/p} \left\{\sum_{\{i,j\}} \pi_i\pi_j\left(\frac{f_j - f_i}{\phi_{ij}^\delta}\right)^2 \phi_{ij}^{2(q+\delta-1)}\right\}^{1/q} \\
&=: \mathrm{I}^{1/p}\cdot \mathrm{II}^{1/q} \qquad (2.1)
\end{aligned}$$

Put $f(e) = f_{e_r} - f_{e_\ell}$. Then,

$$\begin{aligned}
\mathrm{I} &= \sum_{\{i,j\}} \frac{\pi_i\pi_j}{\phi_{ij}^2}\left(\sum_{e\in\gamma_{ij}} f(e)\right)^2 \\
&= \sum_{\{i,j\}} \frac{\pi_i\pi_j}{\phi_{ij}^2}\sum_{e\in\gamma_{ij}} f(e)\left(\sum_{b\in\gamma_{i,e_\ell}} f(b) + \sum_{d\in\gamma_{e_\ell,j}} f(d)\right) \\
&= \sum_{\{i,j\}} \frac{\pi_i\pi_j}{\phi_{ij}^2}\left(\sum_{e\in\gamma_{ij}} f(e)\sum_{b\in\gamma_{i,e_\ell}} f(b) + \sum_{e\in\gamma_{ij}} f(e)\sum_{d\in\gamma_{e_\ell,j}} f(d)\right) \\
&= \sum_{\{i,j\}} \frac{\pi_i\pi_j}{\phi_{ij}^2}\left(\sum_{e\in\gamma_{ij}} f(e)\sum_{b\in\gamma_{i,e_\ell}} f(b) + \sum_{d\in\gamma_{ij}} f(d)\sum_{e\in\gamma_{i,d_r}} f(e)\right) \\
&= \sum_{\{i,j\}} \frac{\pi_i\pi_j}{\phi_{ij}^2}\left(\sum_{e\in\gamma_{ij}} f(e)\sum_{b\in\gamma_{i,e_\ell}} f(b) + \sum_{d\in\gamma_{ij}} f(d)\sum_{e\in\gamma_{i,d_\ell}} f(e) + \sum_{d\in\gamma_{ij}} f(d)\sum_{e=\langle d_\ell,d_r\rangle} f(e)\right) \\
&= \sum_{\{i,j\}} \frac{\pi_i\pi_j}{\phi_{ij}^2}\left(2\sum_{e\in\gamma_{ij}} f(e)\sum_{b\in\gamma_{i,e_\ell}} f(b) + \sum_{e\in\gamma_{ij}} f(e)^2\right) \\
&= 2\sum_e f(e)\sqrt{\pi_{e_\ell}q_{e_\ell e_r}}\cdot\sum_{\{i,j\}:\gamma_{ij}\ni e}\frac{\pi_i\pi_j}{\phi_{ij}^2\sqrt{\pi_{e_\ell}q_{e_\ell e_r}}}\sum_{b\in\gamma_{i,e_\ell}} f(b) + \sum_{\{i,j\}}\frac{\pi_i\pi_j}{\phi_{ij}^2}\sum_{e\in\gamma_{ij}} f(e)^2 \\
&\leqslant 2\left(\sum_e \pi_{e_\ell}q_{e_\ell e_r}f(e)^2\right)^{1/2}\cdot\left(\sum_e\left[\sum_{\{i,j\}:\gamma_{ij}\ni e}\frac{\pi_i\pi_j}{\phi_{ij}^2\sqrt{\pi_{e_\ell}q_{e_\ell e_r}}}\sum_{b\in\gamma_{i,e_\ell}} f(b)\right]^2\right)^{1/2} \\
&\quad + \sum_e \pi_{e_\ell}q_{e_\ell e_r}f(e)^2\frac{1}{\pi_{e_\ell}q_{e_\ell e_r}}\sum_{\{i,j\}:\gamma_{ij}\ni e}\frac{\pi_i\pi_j}{\phi_{ij}^2}. \qquad (2.2)
\end{aligned}$$

Here, we have used Schwarz's inequality in the last step. Note that

$$\sum_{\{i,j\}:\gamma_{ij}\ni e} = \sum_{i\in E}\sum_{j:\gamma_{ij}\ni e}.$$

By using Schwarz's inequality again, we obtain

$$\begin{aligned}
&\sum_e \left[\sum_{\{i,j\}:\gamma_{ij}\ni e} \frac{\pi_i\pi_j}{\phi_{ij}^2\sqrt{\pi_{e_\ell} q_{e_\ell e_r}}} \sum_{b\in\gamma_{i,e_\ell}} f(b)\right]^2\\
&= \sum_e \left[\sum_i \left(\frac{\sqrt{\pi_i\pi_{e_\ell}}}{\phi_{i,e_\ell}}\sum_{b\in\gamma_{i,e_\ell}} f(b)\right)\left(\frac{\sqrt{\pi_i}\phi_{i,e_\ell}}{\pi_{e_\ell}\sqrt{q_{e_\ell,e_r}}}\sum_{j:\gamma_{ij}\ni e}\frac{\pi_j}{\phi_{ij}^2}\right)\right]^2\\
&\leqslant \sum_e \left[\sum_i \frac{\pi_i\pi_{e_\ell}}{\phi_{i,e_\ell}^2}\left(\sum_{b\in\gamma_{i,e_\ell}} f(b)\right)^2 \sum_i \frac{\pi_i\phi_{i,e_\ell}^2}{\pi_{e_\ell}^2 q_{e_\ell,e_r}}\left(\sum_{j:\gamma_{ij}\ni e}\frac{\pi_j}{\phi_{ij}^2}\right)^2\right]\\
&\leqslant \big\{\sup_e \sigma_1(e)\big\}\sum_{e,i}\frac{\pi_i\pi_\ell}{\phi_{i,e_\ell}^2}[f_{e_\ell}-f_i]^2\\
&\leqslant \left\{\sup_e \sigma_1(e)\right\}\beta\cdot \mathrm{I}.
\end{aligned}\tag{2.3}$$

Here in the last step we have used the fact that a point e_ℓ occurs in $\sum_{e,i}$ at most β times. Combining (2.2) and (2.3), we see that

$$\mathrm{I} \leqslant 2\sqrt{\sup_e \sigma_1(e)}\,\sqrt{\beta D(f)\mathrm{I}} + D(f)\sup_e \sigma_2(e) =: 2C_1\sqrt{\mathrm{I}\cdot D(f)} + D(f)C_2.$$

Solving the inequality, we get $\mathrm{I} \leqslant D(f)\left[C_1+\sqrt{C_1^2+C_2}\right]^2$. Next,

$$\mathrm{II} = \sum_{\{i,j\}} \pi_i\pi_j\left(\frac{f_j-f_i}{\phi_{ij}^\delta}\right)^2\phi_{ij}^{2(q+\delta-1)} \leqslant V_\delta(f)\sum_{i,j}\pi_i\pi_j\phi_{ij}^{2(q+\delta-1)}.$$

Hence

$$\mathrm{Var}_\pi(f) \leqslant CD(f)^{1/p}V_\delta(f)^{1/q},$$

where $C = \left(C_1+\sqrt{C_1^2+C_2}\right)^{2/p}\sum_{i,j}\pi_i\pi_j\phi_{ij}^{2(q+\delta-1)}$. By the Liggett-Stroock Theorem, the process has algebraic decay.

C) We now prove part (2) of the theorem. We remark that condition "$\rho_{i+R} \leqslant c'\rho_i$ for all $i \geqslant 1$" holds whenever $\rho_N \leqslant c'\rho_{N/2}$ for all $N \geqslant 1$. To see this, let $i \geqslant R$ and $N = i + R$. Then $\rho_{i+R} = \rho_N \leqslant c'\rho_{(i+R)/2} \leqslant c'\rho_i$ since ρ_i is increasing in i. On the other hand, since the set $\{i : i < R\}$ is finite, the inequality "$\rho_{i+R} \leqslant c''\rho_i$

for all $i < R$" is automatic for some constant $c'' \leqslant c'$. However, to simplify the notation, we will use the same c in these inequalities.

Assume that the process has algebraic decay. Let $m, N \in \mathbb{N}$ so that $\sum_i \rho_i^m \pi_i = \infty$ and let $f(k) = \rho_{k\wedge N}^m$. Then, we have

$$\begin{aligned}
V(f) &= \max_{0\leqslant i,j\leqslant N, i\neq j} (\rho_i^m - \rho_j^m)^2/(\rho_i - \rho_j)^{2\delta} \\
&= \max_{0\leqslant i<j\leqslant N} \Big\{\rho_j^{m-1}\Big(1+\rho_i/\rho_j+(\rho_i/\rho_j)^2+\cdots+(\rho_i/\rho_j)^{m-1}\Big)\Big\}^2 (\rho_j-\rho_i)^{2(1-\delta)} \\
&\leqslant m^2 \rho_N^{2(m-\delta)}.
\end{aligned}$$

We now consider $D(f)$:

$$D(f) = \frac{1}{2}\sum_{0\leqslant i,j\leqslant N} \pi_i q_{ij}(\rho_j^m - \rho_i^m)^2 + \sum_{i<N,\, j>N} \pi_i q_{ij}(\rho_N^m - \rho_i^m)^2.$$

For the first term on the right-hand side, we have

$$\begin{aligned}
&\frac{1}{2}\sum_{0\leqslant i,j\leqslant N} \pi_i q_{ij}(\rho_j^m - \rho_i^m)^2 = \sum_{0\leqslant i<j\leqslant N} \pi_i q_{ij}(\rho_j^m - \rho_i^m)^2 \\
&= \sum_{i=0}^{N} \pi_i \sum_{j=i+1}^{N} q_{ij}(\rho_j^m - \rho_i^m)^2 \\
&= \sum_{i=0}^{N} \pi_i \sum_{j=i+1}^{N} q_{ij}(\rho_j - \rho_i)^2 \left[\rho_j^{m-1} + \rho_j^{m-2}\rho_i + \cdots + \rho_i^{m-1}\right]^2 \\
&= \sum_{i=1}^{N} \pi_i \sum_{j=i+1}^{N\wedge(i+R)} q_{ij}(\rho_j - \rho_i)^2 \left[\rho_j^{m-1} + \rho_j^{m-2}\rho_i + \cdots + \rho_i^{m-1}\right]^2 + \pi_0 \sum_{j=1}^{N\wedge R} q_{0j}\rho_j^{2m} \\
&\leqslant m^2 c_1 \sum_{i=1}^{N} \pi_i \rho_i^{2m-2} \sum_{j=i+1}^{N\wedge(i+R)} q_{ij}(\rho_j - \rho_i)^2 + \pi_0 \sum_{j=1}^{N\wedge R} q_{0j}\rho_j^{2m} \\
&\leqslant m^2 c_1 \left(\sup_{k\geqslant 1} \sum_{j=k+1}^{k+R} q_{kj}(\rho_j - \rho_k)^2\right) \sum_{i=1}^{N} \pi_i \rho_i^{2m-2} + c_2 \sum_{j=1}^{N\wedge R} \pi_j \rho_j^{2m-2} \\
&= m^2 \left(c_1 \sup_{k\geqslant 1} \sum_{j=k+1}^{\infty} q_{kj}(\rho_j - \rho_k)^2 + c_2\right) \sum_{i=1}^{N} \pi_i \rho_i^{2m-2},
\end{aligned}$$

where

$$c_1 = c_1(m) = \sum_{k=0}^{m-1} c^{2k} = (c^{2m} - 1)/(c - 1), \quad c_2 = c_2(m) = m^{-2}\pi_0 \max_{1\leqslant j\leqslant R} q_{0j}\rho_j^2/\pi_j.$$

As for the second term, because of finite range, by condition (1.6) and the remark at the beginning of this part C), we have $\sum_{j>N} q_{0j} = 0$ for $N > R$ and

$\rho_N \leqslant c\rho_i$ for all $1 \vee (N-R) \leqslant i \leqslant N-1$. Then,

$$\begin{aligned}
&\sum_{i<N,\, j>N} \pi_i q_{ij} (\rho_N^m - \rho_i^m)^2 \\
&= \sum_{i=1}^{N-1} \pi_i \left(\rho_N^{m-1} + \rho_N^{m-2}\rho_i + \cdots + \rho_i^{m-1}\right)^2 \cdot (\rho_N - \rho_i)^2 \sum_{j>N} q_{ij} + \pi_0 \sum_{j>N} q_{0j} \rho_N^{2m} \\
&\leqslant \sum_{i=1\vee(N-R)}^{N-1} \pi_i \cdot \tilde{c} \cdot m^2 \rho_i^{2(m-1)} \cdot \sup_{N\geqslant 1} \max_{1\vee(N-R)\leqslant k\leqslant N-1} (\rho_N - \rho_k)^2 \sum_{j>N} q_{kj} \\
&\leqslant c_1 m^2 \left\{ \sup_{N\geqslant 1} \max_{1\vee(N-R)\leqslant k\leqslant N-1} (\rho_N - \rho_k)^2 \sum_{j>N} q_{kj} \right\} \sum_{i=1}^{N} \pi_i \rho_i^{2(m-1)} \\
&\leqslant c_1 m^2 \left\{ \sup_{k\geqslant 1} \sum_{j=k+1}^{\infty} q_{kj} (\rho_j - \rho_k)^2 \right\} \sum_{i=1}^{N} \pi_i \rho_i^{2(m-1)}, \quad N > R.
\end{aligned}$$

The last inequality holds because

$$\sum_{j>N} q_{kj}(\rho_N - \rho_k)^2 \leqslant \sum_{j>N} q_{kj}(\rho_j - \rho_k)^2 \leqslant \sum_{j=k+1}^{\infty} q_{kj}(\rho_j - \rho_k)^2.$$

Finally, we get

$$D(f) \leqslant C_1(m) \sum_{i=1}^{N} \pi_i \rho_i^{2(m-1)}, \tag{2.4}$$

where

$$C_1(m) = m^2 [2c_1(m) \sup_{i\geqslant 1} \sum_{j=i+1}^{\infty} q_{ij} (\rho_j - \rho_i)^2 + c_2(m)] < \infty$$

for all m by assumption.

Before moving further, we need an elementary result about the estimation of variation.

Lemma 2.1. Let f be an increasing function and define $h = f \circ g$ for some function g. Next, let $W > 0$ be a constant and set $h_W = h \wedge W$. Choose γ_M large enough so that $\pi(g > \gamma_M) \leqslant 1/M$. Then we have

$$\mathrm{Var}(h_W) \geqslant \left(\int_{[h\leqslant W]} h^2 d\pi \right) \left\{ 1 - \left(\frac{1}{\sqrt{M}} + \frac{f(\gamma_M)}{\left(\int_{[h\leqslant M]} h^2 d\pi \right)^{1/2}} \right) \right\}^2. \tag{2.5}$$

Proof. Note that

$$\pi(h_W) - \pi\left(I_{[g\leqslant\gamma_M]} h_W\right) = \pi\left(I_{[g>\gamma_M]} h_W\right) = \int_{[g>\gamma_M]} h_W d\pi \leqslant \|h_W\| \sqrt{1/M}.$$

We have

$$\pi(h_W) \leqslant \|h_W\|/\sqrt{M} + \pi(I_{[g \leqslant \gamma_M]} h_W) \leqslant \|h_W\|/\sqrt{M} + f(\gamma_M).$$

Hence

$$\begin{aligned} \operatorname{Var}(h_W) &= \|h_W\|^2 - \pi(h_W)^2 \geqslant \|h_W\|^2 - \Big(\|h_W\|/\sqrt{M} + f(\gamma_M)\Big)^2 \\ &= \|h_W\|^2 \left\{1 - \left(\frac{1}{\sqrt{M}} + \frac{f(\gamma_M)}{\|h_W\|}\right)\right\}^2. \end{aligned}$$

On the other hand,

$$\|h_W\|^2 = \int_{[h \leqslant W]} h^2 d\pi + W^2 \pi[h > W] \geqslant \int_{[h \leqslant W]} h^2 d\pi.$$

From these two facts, we obtain (2.5). $\square$

Now, let $g_k = \rho_k,\ f(x) = x^m$ and $W = \rho_N^m$. Then we come back to $h_W(k) = \rho_{k \wedge N}^m$. The estimate (2.5) yields that

$$\left(\sum_{i=1}^N \pi_i \rho_i^{2m}\right) \left\{1 - \left[\frac{1}{\sqrt{M}} + \frac{\gamma_M^m}{\sqrt{\sum_{i=1}^\infty \pi_i \rho_{i \wedge N}^{2m}}}\right]\right\}^2 \leqslant \operatorname{Var}(f).$$

Take $M = 16$. Since $\pi(\rho^{2m}) = \infty$, there exists $N_0 = N_0(m)$ such that

$$\frac{1}{2} \sum_{i=1}^N \pi_i \rho_i^{2m} \leqslant \operatorname{Var}(f), \qquad \text{for all } N \geqslant N_0. \tag{2.6}$$

By part (2) of Theorem A, (1.2) holds. Combining (2.5), (2.6) with (1.2), we get

$$\begin{aligned} \sum_{j=1}^N \pi_j \rho_j^{2m} &\leqslant C_2(m) \left(\sum_{j=1}^N \pi_j \rho_j^{2m-2}\right)^{1/p} \rho_N^{2(m-\delta)/q} \\ &\leqslant C_2(m) \left(\sum_{j=1}^N \pi_j \rho_j^{2m}\right)^{(m-1)/(mp)} \rho_N^{2(m-\delta)/q} \end{aligned}$$

where in the last step, we have used the Schwarz's inequality. Therefore,

$$\sum_{j=1}^N \pi_j \rho_j^{2m} \leqslant C_3(m) \rho_N^{\frac{2(m-\delta)mp}{q(mp-m+1)}}. \tag{2.7}$$

Now, we consider separately the two cases listed in (1.6). First, assume that there is $\epsilon < 1$ such that $\inf_N \rho_{N/2}/\rho_N \geqslant \epsilon$. Then we have

$$\begin{aligned}\sum_{j=N/2}^{N} \pi_j \rho_j^k &= \sum_{j=N/2}^{N} \pi_j \rho_j^{2m+k-2m} \\ &\leqslant \rho_{N/2}^{k-2m} \sum_{j=N/2}^{N} \pi_j \rho_j^{2m} \\ &\leqslant \epsilon^{k-2m} \rho_N^{k-2m} \sum_{j=N/2}^{N} \pi_j \rho_j^{2m} \\ &\leqslant C_4(m) \rho_N^{\frac{2(m-\delta)mp}{q(mp-m+1)}}.\end{aligned}$$

When $m \longrightarrow \infty$, the power of ρ_N on the right-hand side converges to $k - 2(q + \delta - 1)$. When $k < 2(q+\delta-1)$, since

$$(\rho_N/\rho_{N/2})^{k-2(q+\delta-1)} \leqslant \epsilon^{2(q+\delta-1)-k} < 1,$$

by (2.7) and ratio test, we get

$$\sum_j \rho_j^k \pi_j = \sum_{\ell=0}^{\infty} \sum_{j \in \{2^\ell \leqslant \rho_j \leqslant 2^{\ell+1}-1\}} \rho_j^k \pi_j < \infty.$$

Secondly, by assumption, we have $\rho_{(N+1)R} \leqslant c\rho_{NR}$ and $\sum_{N=1}^{\infty} \rho_{NR}^{-\epsilon} < \infty$ for all $\epsilon > 0$. Hence

$$\begin{aligned}\sum_{j=NR}^{(N+1)R} \pi_j \rho_j^k &= \sum_{j=NR}^{(N+1)R} \pi_j \rho_j^{2m+k-2m} \\ &\leqslant \rho_{NR}^{k-2m} \sum_{j=NR}^{(N+1)R} \pi_j \rho_j^{2m} \\ &\leqslant c^{k-2m} \rho_{(N+1)R}^{k-2m} \sum_{j=NR}^{(N+1)R} \pi_j \rho_j^{2m} \\ &\leqslant C_4(m) \rho_{(N+1)R}^{k-2m+\frac{2(m-\delta)mp}{q(mp-m+1)}}\end{aligned}$$

So by (2.7) we get

$$\sum_j \rho_j^k \pi_j = \sum_{N=0}^{\infty} \sum_{j=NR}^{(N+1)R-1} \pi_j \rho_j^k \leqslant \sum_{N=0}^{\infty} \sum_{j=NR}^{(N+1)R} \pi_j \rho_j^k < \infty.$$

Now, the proof of Theorem 1.1 is completed. $\square$

3. Proofs of Theorem 1.3 and other results.

Proof of Theorem 1.3

Let f satisfy $\pi(f)=0$ and $\|f\|^2=1$. Then, we have

$$\begin{aligned}
\operatorname{Var}_\pi(f) &= \inf_c \sum_j \pi_j\,(f_j-c)^2\\
&\leqslant \sum_j \pi_j\,(f_j-f_0)^2\\
&= \sum_j \pi_j \left(\frac{f_j-f_0}{\phi_j}\right)^2 \phi_j^2\\
&\leqslant \left\{\sum_j \pi_j \left(\frac{f_j-f_0}{\phi_j}\right)^2\right\}^{1/p} \left\{\sum_j \pi_j \left(\frac{f_j-f_0}{\phi_j^\delta}\right)^2 \phi_j^{2(q+\delta-1)}\right\}^{1/q}\\
&=: \mathrm{I}^{1/p}\cdot \mathrm{II}^{1/q}.
\end{aligned}$$

The remainder of the proof is similar to the one of Theorem 1.1. The key point is replacing $\sum_i$ used there by the single point $i=0$. For instance, put $f(e)=f_{e_r}-f_{e_\ell}$. Then we have

$$\begin{aligned}
\mathrm{I} &= \sum_j \frac{\pi_j}{\phi_j^2}\left(\sum_{e\in\gamma_{0j}} f(e)\right)^2\\
&= \sum_j \frac{\pi_j}{\phi_j^2}\sum_{e\in\gamma_{0j}} f(e)\left(\sum_{b\in\gamma_{0,e_\ell}} f(b)+\sum_{d\in\gamma_{e_\ell,j}} f(d)\right)\\
&= \sum_j \frac{\pi_j}{\phi_j^2}\left\{2\sum_{e\in\gamma_{0j}} f(e)\sum_{b\in\gamma_{0,e_\ell}} f(b)+\sum_{e\in\gamma_{0j}} f(e)\right\}\\
&= 2\sum_e f(e)\sqrt{\pi_{e_\ell}q_{e_\ell e_r}}\cdot \sum_{j:\gamma_{0j}\ni e}\frac{\pi_j}{\phi_j^2\sqrt{\pi_{e_\ell}q_{e_\ell e_r}}}\sum_{b\in\gamma_{0,e_\ell}} f(b)+\sum_j\frac{\pi_j}{\phi_j^2}\sum_{e\in\gamma_{0j}} f(e)^2\\
&\leqslant 2\left(\sum_e \pi_{e_\ell}q_{e_\ell e_r} f(e)^2\right)^{1/2}\cdot\left(\sum_e\left[\sum_{j:\gamma_{0j}\ni e}\frac{\pi_j}{\phi_j^2\sqrt{\pi_{e_\ell}q_{e_\ell e_r}}}\sum_{b\in\gamma_{0,e_\ell}} f(b)\right]^2\right)^{1/2}\\
&\quad +\sum_e \pi_{e_\ell}q_{e_\ell e_r} f(e)^2\frac{1}{\pi_{e_\ell}q_{e_\ell e_r}}\sum_{j:\gamma_{0j}\ni e}\frac{\pi_j}{\phi_j^2}
\end{aligned}\tag{3.1}$$

Moreover,

$$\begin{aligned}
&\sum_e\left[\sum_{j,\gamma_{0j}\ni e}\frac{\pi_j}{\phi_j^2\sqrt{\pi_{e_\ell}q_{e_\ell e_r}}}\sum_{b\in\gamma_{0,e_\ell}} f(b)\right]^2\\
&\qquad=\sum_e\left[\left(\frac{\sqrt{\pi_{e_\ell}}}{\phi_{e_\ell}}\sum_{b\in\gamma_{0,e_\ell}} f(b)\right)\left(\frac{\sqrt{\phi}_{e_\ell}}{\pi_{e_\ell}\sqrt{q_{e_\ell,e_r}}}\sum_{j:\gamma_{0j}\ni e}\pi_j\frac{1}{\phi_j^2}\right)\right]^2\leqslant
\end{aligned}$$

$$\leqslant \left\{\sup_e \tilde{\sigma}_1(e)\right\}^2 \sum_e \frac{\pi_{e_\ell}}{\phi_{e_\ell}^2} \left(\sum_{b\in\gamma_{0,e_\ell}} f(b)\right)^2 \beta \cdot \mathrm{I} \tag{3.2}$$

Combining (3.1)and (3.2), we see that

$$\mathrm{I} \leqslant 2\big\{\sup_e \tilde{\sigma}_1(e)\big\}\sqrt{\beta D(f)\mathrm{I}} + D(f)\sup_e \tilde{\sigma}_2(e) =: 2C_1\sqrt{\mathrm{I}\cdot D(f)} + D(f)C_2.$$

Next,

$$\mathrm{II} = \sum_j \pi_j \left(\frac{f_j - f_0}{\phi_j^\delta}\right)^2 \phi_j^{2(q+\delta-1)} \leqslant \widetilde{V}_\delta(f)\sum_j \pi_j \phi_j^{2(q+\delta-1)}$$

In the particular case mentioned in part (1) of Theorem 1.3, one may replace $\widetilde{V}_1$ by V_1 on the right-hand side since $\widetilde{V}_1(f) \leqslant V_1(f)$. The remainder of the proof is almost the same as the one of Theorem 1.1, the only place which needs a slight change is estimating $\widetilde{V}_\delta$ instead of V_δ at the beginning of the proof for part (2) of Theorem 1.1. $\square$

To prove Corollaries 1.4 and 1.5, recall that for a positive recurrent birth-death process with birth rate $b_i > 0 (i \geqslant 0)$ and death rate $a_i > 0 (i \geqslant 1)$, the reversible measure (π_i) is:

$$\pi_i = \frac{\mu_i}{\mu}, \quad \mu_0 = 1, \quad \mu_i = \frac{b_0 b_1 \cdots b_{i-1}}{a_1 a_2 \cdots a_i},\ i \geqslant 1,$$

where $\mu = \sum_i \mu_i$.

Lemma 3.1([3], **Theorem** 3.3)**.** Let (u_k) be a positive sequence on Z_+ and set $F(k) = \sum_{j<k} u_j$. Define $\rho(m,n) = |F(m) - F(n)|$. Then, there exists a coupling operator $\widetilde{\Omega}$ such that

$$\widetilde{\Omega}\rho(i,j) = b_j u_j - a_j u_{j-1} - b_i u_i + a_i u_{i-1}, \qquad u_{-1} := 1. \tag{3.3}$$

Proof of Corollaries 1.4 and 1.5.

First, we prove $V(P_t f) \leqslant V(f)$. By Lemma 3.1, we know that there exists coupling operator $\widetilde{\Omega}$ satisfying (3.1). By the first assumption of the corollaries, we have $\widetilde{\Omega}\rho(i,j) \leqslant 0$, for all $i,j \in E$. Then applying the proof A) of Theorem 1.1 gives the required assertion.

Secondly, note that $\langle i,j\rangle$ is an edge if and only if $|i-j| = 1$. For $k < \ell$, choose and fix a path from k to ℓ: $\langle k, k+1\rangle, \langle k+1, k+2\rangle, \dots, \langle \ell-1, \ell\rangle$. Then, the remainder of the proof of Corollary 1.4 is the same as those of Theorem 1.1, and the proof of Corollary 1.5 is the same as the one of Theorem 1.3. We omit the details here. $\square$

Proof of Corollary 1.6

Take $\phi_n = n^\alpha$, Then

$$\frac{\pi_n \phi_n^{2(q-1)}}{\pi_{n+1}\phi_{n+1}^{2(q-1)}} = \frac{a_{n+1}}{b_n}\left(1-\frac{1}{n+1}\right)^{2\alpha(q-1)}$$
$$= \left(1+\frac{1}{n}\cdot n\left(\frac{a_{n+1}}{b_n}-1\right)\right)\left(1-\frac{1}{n+1}\right)^{2\alpha(q-1)}.$$

By the Gauss' test, we have $\sum_n \pi_n \phi_n^{2(q-1)} < \infty$ once

$$\liminf_{n\to\infty} n\left(\frac{a_{n+1}}{b_n}-1\right) - 2\alpha(q-1) > 1,$$

which is fulfilled for sufficient small $q-1>0$ by assumption. Next,

$$\sum_{k\geqslant n+1} \pi_k k^{-2\alpha} \leqslant (n+1)^{-2\alpha} \sum_{k\geqslant n+1}\pi_k \leqslant n^{-2\alpha}\sum_{k\geqslant n+1}\pi_k.$$

Hence

$$\theta_1(n) = \frac{n^\alpha}{\sqrt{b_n}\pi_n}\sum_{k\geqslant n+1}\pi_k k^{-2\alpha} \leqslant \frac{1}{\sqrt{b_n}n^\alpha \pi_n}\sum_{k\geqslant n+1}\pi_k.$$

By assumption, we have $\sup_n \theta_1(n) < \infty$ and

$$\limsup_{n\to\infty}\frac{1}{\phi_n\sqrt{b_n}} \leqslant \sup_n\left\{\frac{\sum_{k\geqslant n+1}\pi_k}{\sqrt{b_n}n^\alpha\pi_n}\right\}\cdot \limsup_{m\to\infty}\frac{\pi_m}{\sum_{k\geqslant m+1}\pi_k} < \infty.$$

The required conclusion now follows from part (1) of Corollary 1.5. $\square$

Proof of Theorem B

We have already proved that $V(P_t f) \leqslant V(f)$ in the proof of Theorem 1.1.

a) Obviously,

$$D(f) = \sum_{k\geqslant 0}(f_{k+1}-f_k)^2 q_{k,k+1}\pi_k \geqslant \left(\inf_{i\geqslant 0} q_{i,i+1}\right)\sum_{k\geqslant 0}(f_{k+1}-f_k)^2\pi_k.$$

b) Let $f\in L^2(\pi)$. Then,

$$\begin{aligned}
&\sum_{n=0}^{\infty}\pi_n\left\{\sum_{k=0}^{n}|f_{k+1}-f_k|\right\}^2\\
&\leqslant 2\sum_{n=0}^{\infty}\pi_n \sum_{0\leqslant j\leqslant k\leqslant n}|f_{j+1}-f_j||f_{k+1}-f_k|\\
&= 2\sum_{n=0}^{\infty}\sum_{k=0}^{n}\sum_{j=0}^{k} = 2\sum_{k=0}^{\infty}\sum_{j=0}^{k}\sum_{n=k}^{\infty}\\
&= 2\sum_{k=0}^{\infty}|f_{k+1}-f_k|\sigma_k\pi_k\sum_{j=0}^{k}|f_{j+1}-f_j|\\
&\leqslant 2\left\{\sum_{k=0}^{\infty}|f_{k+1}-f_k|^2\sigma_k^2\pi_k\sum_{k=0}^{\infty}\left(\sum_{j=0}^{k}|f_{j+1}-f_j|\right)^2\pi_k\right\}^{1/2}.
\end{aligned}$$

That is
$$\sum_{n=0}^{\infty}\pi_n\left\{\sum_{k=0}^{n}|f_{k+1}-f_k|\right\}^2\leqslant 4\sum_{k=0}^{\infty}|f_{k+1}-f_k|^2\sigma_k^2\pi_k.$$
On the other hand,
$$\begin{aligned}\|f-\pi(f)\|^2&=\frac{1}{2}\sum_{j,k=0}^{\infty}\pi_j\pi_k(f_k-f_j)^2\\&\leqslant\sum_{0\leqslant j<k}\pi_j\pi_k\left\{\sum_{i=j}^{k-1}(f_{i+1}-f_i)\right\}^2\\&\leqslant\sum_{k=1}^{\infty}\pi_k\sum_{j=0}^{k-1}\pi_j\left\{\sum_{i=j}^{k-1}(f_{i+1}-f_i)\right\}^2\\&\leqslant\sum_{k=0}^{\infty}\pi_k\left\{\sum_{i=0}^{k}|f_{i+1}-f_i|\right\}^2\end{aligned}$$
Combining the above two inequalities, it follows that
$$\|f-\pi(f)\|^2\leqslant 4\sum_{k=0}^{\infty}|f_{k+1}-f_k|^2\sigma_k^2\pi_k.$$
c) By Schwarz's inequality, we get
$$\begin{aligned}\|f-\pi(f)\|^2&\leqslant 4\left\{\sum_{k=0}^{\infty}|f_{k+1}-f_k|^2\pi_k\right\}^{1/p}\left\{\sum_{k=0}^{\infty}|f_{k+1}-f_k|^2\sigma_k^{2q}\pi_k\right\}^{1/q}\\&\leqslant CD(f)^{1/p}V(f)^{1/q}\end{aligned}$$
where
$$C=4\left(\inf_i q_{i,i+1}\right)^{-1/p}\left(\sup_k\sigma_k/k\right)^2\left\{\sum_{k=0}^{\infty}u_k^2k^{2q}\pi_k\right\}^{1/q}.\quad\square$$

Proof of Theorem 1.7

Let V denote either V_δ or $\widetilde{V}_\delta$ appeared in the theorem. Because the $\widetilde{Q}$-process has algebraic decay with respect to V, we have for some constants p, q and C that
$$\mathrm{Var}_{\tilde\pi}(f)\leqslant C\widetilde{D}(f)^{1/p}V(f)^{1/q},\qquad f\in\mathscr{D}(\widetilde{D}).$$
Next, by the assumptions of the theorem, we have $L^2(\tilde\pi)\subset L^2(\pi)$ and moreover,
$$\begin{aligned}\frac{D(f)^{1/p}V(f)^{1/q}}{\mathrm{Var}_\pi(f)}&=\frac{\left[\frac{1}{2}\sum_{i,j}\pi_iq_{ij}(f_j-f_i)^2\right]^{1/p}V(f)^{1/q}}{\inf_{c\in\mathbb{R}}\sum_i\pi_i(f_i-c)^2}\\&\geqslant\frac{\inf_{k\neq\ell}\{\pi_kq_{k\ell}/(\tilde\pi_k\tilde q_{k\ell})\}^{1/p}}{\sup_k\{\pi_k/\tilde\pi_k\}}\cdot\frac{\widetilde{D}(f)^{1/p}V(f)^{1/q}}{\mathrm{Var}_{\tilde\pi}(f)}\\&\geqslant\frac{\inf_{k\neq\ell}\{\pi_kq_{k\ell}/(\tilde\pi_k\tilde q_{k\ell})\}^{1/p}}{\sup_k\{\pi_k/\tilde\pi_k\}}\cdot C,\\&\qquad f\in L^2(\tilde\pi)\cap\mathscr{D}(D).\end{aligned}\tag{3.4}$$

The proof will be complete once we remove "$L^2(\tilde{\pi})$" appeared at the end of (3.4). To do so, let $f \in \mathscr{D}(D)$ and set $f_M = (-M) \vee f \wedge M$ for constant $M > 0$. Then, by [2; Lemma 6.47], we have $f_M \in \mathscr{D}(D)$, $\|f_M - f\| \to 0$ and $D(f_M) \to D(f)$ as $M \to \infty$. Hence $\mathrm{Var}_\pi(f_M) \to \mathrm{Var}_\pi(f)$ as $M \to \infty$. The assertion now follows by replacing f with $f_M \in L^2(\tilde{\pi}) \cap \mathscr{D}(D)$ in (3.4) and then letting $M \to \infty$, since $V(f_M) \leqslant V(f)$. $\square$

4. Two examples.

In this section, we examine two examples of birth-death processes.

Example 4.1. $a_i = b_i = i^r$, $1 < r \leqslant 2$.

Example 4.2. $a_i = 1$, $b_i = 1 - c/i$, $\quad i \gg 1$.

It is easy to check that the process of Example 4.1 (resp., Example 4.2) is positive recurrent iff $r > 1$ (resp., $c > 1$). As was proved in [5], the first example has L^2-exponential convergence iff $r \geqslant 2$. However, the second example is never L^2-exponentially convergent for all c.

Proposition 4.3. With respect to V_0, Example 4.1 (resp., 4.2) has algebraic decay for all $r \in (1, 2)$ (resp., $c \in (1, \infty)$).

Proof. Simply take $\epsilon = 1/2$ and $\epsilon = 0$, respectively, for Examples 4.1 and 4.2 and then apply Corollary 1.6. $\square$

For the remainder of this section, we study the region of algebraic convergence with different V.

Proposition 4.4. With respect to V_1 defined by (1.9), Example 4.1 has algebraic decay iff $r > 5/3$.

Proof. Clearly, we need only to prove the assertion for $r \in (1, 2)$ since the process has L^2-exponential convergence for all $r \geqslant 2$.

(I) Set $u_n = (n + 1)^{-s}$ for some constant $s > 0$ to be determined later. We should justify the power of the different results for this typical example.

(A) Use Corollary 1.5.

1) We prove that the additional condition of Corollary 1.5 is satisfied once $s \leqslant r - 1$.

$$b_k u_k - a_k u_{k-1} = k^r \left(\frac{1}{(k+1)^s} - \frac{1}{k^s} \right) = k^{r-s} \left[(1 - \frac{1}{k+1})^s - 1 \right].$$

Let

$$f(x) = x^{r-s} \left[1 - (1 - \frac{1}{x+1})^s \right], \quad x \geqslant 1.$$

Then,

$$f'(x) = (r-s)x^{r-s-1} \left[1 - (1 - \frac{1}{x+1})^s \right] - s x^{r-s} \left(1 - \frac{1}{x+1} \right)^{s-1} \frac{1}{(1+x)^2}$$

It is easy to prove that $f'(x) \geqslant 0$ if and only if $r - s \geqslant 1$. So , when $s \leqslant r - 1$, condition (a) is satisfied with $u_n = (n+1)^{-s}$.

2) We prove that condition (a) of Corollary 1.5 is satisfied for all $s \leqslant r/2$. Since $s < 1$, we have

$$\phi_n = \sum_{k=0}^{n-1} u_k = \sum_{k=0}^{n-1} \frac{1}{(k+1)^s} \sim n^{1-s}.$$

Then,

$$\theta_1(k) = \frac{k^{1-s}}{k^{-r}k^{r/2}} \sum_{n=k+1}^{\infty} \frac{n^{-r}}{n^{2-2r}} \sim \frac{1}{k^{r/2-s}}.$$

So, when $s \leqslant r/2$, we get (a).

3) Because

$$\phi_k \sqrt{q_{k,k+1}} \sim k^{1-s}k^{r/2} = k^{1-s+r/2},$$

condition (b) follows for all $s \leqslant 1 + r/2$.

4) Because

$$\sum_n \pi_n \phi_n^{2q} \sim \sum_n n^{-r} n^{2q(1-s)},$$

if $2q < (r-1)/(1-s)$ $(s < 1)$, then we have

$$\sum_n \pi_n \phi_n^{2q} < \infty.$$

Combining this with condition $q > 1$, we get $(r-1)/(1-s) > 2$, that is $s > (3-r)/2$.

Because of 1)—4), the process has algebraic decay whenever $(3-r)/2 < s \leqslant r-1$, namely $r > 5/3$. Choosing $s = r-1$, we obtain $q < (r-1)/[2(2-r)]$. It is clear that when $r \to 2$, q is allowed to tend to ∞.

(B) Use Corollary 1.4.

1) It is proved in (A) 1) above that $b_n u_n - a_n u_{n-1}$ is non-increasing whenever $s \leqslant r-1$.

2) Note that

$$\phi_{ij} = u_i + u_{i+1} + \cdots + u_{j-1} = (i+1)^{-s} + \cdots + j^{-s},$$

and hence

$$\frac{1}{1-s}\left[(j+1)^{1-s} - (i+1)^{1-s}\right] = \int_{i+1}^{j+1} \frac{dx}{x^s} \geqslant \phi_{ij} \geqslant \int_i^j \frac{dx}{x^s} = \frac{1}{1-s}(j^{1-s} - i^{1-s}).$$

Set $\alpha = 1-s > 0$. Then $1 > \alpha \geqslant 1-r$ by 1).

3) Consider condition

$$\sum_{i,j} \pi_i \pi_j \phi_{ij}^{2q} = 2\sum_{i<j} \pi_i \pi_j \phi_{ij}^{2q} < \infty.$$

Choose $i=0$. We have

$$\begin{aligned}\sum_{j>0}\pi_j\phi_{0j}^{2q}<\infty &\iff \sum_{j>0} j^{-r} j^{\alpha\cdot 2q}<\infty\\ &\iff r-2q\alpha>1\\ &\iff r>1+2\alpha \text{ (since } q>1)\\ &\iff \alpha<\frac{r-1}{2}.\end{aligned}$$

Combining 2) with 3), we get $r>5/3$. Then,

$$\sum_{i<j}\pi_i\pi_j\phi_{ij}^{2q}\leqslant\sum_{i<j}\pi_i\pi_j[(j+1)^\alpha-(i+1)^\alpha]^{2q}\leqslant\sum_i\pi_i\sum_{j\geqslant 1}\pi_j(j+1)^{2q\alpha}=\sum_{j\geqslant 1}\pi_j(j+1)^{2q\alpha}.$$

The last sum is finite if and only if $q<(r-1)/(2\alpha)$.

4) Now, we consider condition $\sup_e\sigma_2'(e)<\infty$. Let $e=\langle k,k+1\rangle$. Then

$$\begin{aligned}\sigma_2'(e)&=\sup_{i\leqslant k}\sum_{j\geqslant k+1}\frac{\pi_j}{\pi_k q_{k,k+1}\phi_{ij}^2}\\ &\sim\sup_{i\leqslant k}\sum_{j\geqslant k+1}\pi_j/\phi_{ij}^2\\ &=\sum_{j\geqslant k+1}\pi_j/\phi_{k,j}^2\\ &\sim\int_{k+1}^\infty\frac{dx}{x^r[x^\alpha-k^\alpha]^2}\\ &\sim-\int_{k+1}^\infty\frac{1}{x^{r+\alpha-1}}d\left(\frac{1}{x^\alpha-k^\alpha}\right)\\ &=\frac{1}{(k+1)^{r+\alpha-1}[(k+1)^\alpha-k^\alpha]}-(r+\alpha-1)\int_{k+1}^\infty\frac{dx}{x^{r+\alpha}[x^\alpha-k^\alpha]}\\ &\sim\frac{1}{k^{r+\alpha-1+\alpha-1}}+\int_{k+1}^\infty\frac{dx}{x^{r+\alpha+\alpha-1}}\\ &\sim k^{-r-2\alpha+2}\end{aligned}$$

Because $r+2\alpha\geqslant r+4-2\alpha=4-r\geqslant 2$, the last term is bounded.

5) Finally, consider condition $\sup_e\sigma_1'(e)<\infty$.

$$\sigma_1'(e)=\sup_{i\leqslant k-1}\frac{\phi_{ik}}{\pi_k\sqrt{q_{k,k+1}}}\sum_{j\geqslant k+1}\pi_j/\phi_{ij}^2\sim\sup_{i\leqslant k-1}\frac{\phi_{ik}}{\sqrt{\pi_k}}\sum_{j\geqslant k+1}\frac{\pi_j}{\phi_{ij}^2}.$$

On the other hand,

$$\frac{\phi_{ik}}{\sqrt{\pi_k}}\sum_{j\geqslant k+1}\frac{\pi_j}{\phi_{ij}^2}\leqslant\text{const.}\frac{(k+1)^\alpha-(i+1)^\alpha}{k^{-r/2}}\sum_{j\geqslant k+1}\frac{1}{j^r[j^\alpha-i^\alpha]^2}.$$

We now adopt the continuous approximation. Note that

$$\sup_k f(k)/g(k) \leqslant \sup_k f'(k)/g'(k).$$

For $x \leqslant k-1$, we get

$$\begin{aligned}&\frac{(k+1)^\alpha-(x+1)^\alpha}{k^{-r/2}}\int_{k+1}^\infty \frac{dy}{y^r(y^\alpha-x^\alpha)^2}\\ &\leqslant \frac{1}{(-r/2)k^{-r/2+1}}\left[\alpha(k+1)^{\alpha-1}\int_{k+1}^\infty \frac{dy}{y^r(y^\alpha-x^\alpha)^2}-\frac{(k+1)^\alpha-(x+1)^\alpha}{(k+1)^r((k+1)^\alpha-x^\alpha)^2}\right].\end{aligned}$$

Note that

$$\begin{aligned}\frac{(k+1)^\alpha-(x+1)^\alpha}{(k+1)^r((k+1)^\alpha-x^\alpha)^2}&\leqslant \frac{1}{(k+1)^r[(k+1)^\alpha-x^\alpha]}\\ &\leqslant \frac{1}{(k+1)^r[(k+1)^\alpha-(k-1)^\alpha]}\\ &\sim k^{-r-\alpha+1}\\ &=k^{-r+s}<\infty\end{aligned}$$

and

$$\alpha(k+1)^{\alpha-1}\int_{k+1}^\infty \frac{dy}{y^r(y^\alpha-x^\alpha)^2}\leqslant k^{-r-\alpha+2}.$$

When $r<2$, we have

$$\frac{k^{-r-\alpha+2}}{k^{-r/2+1}}=k^{-r/2-\alpha+1}<\infty$$

and so $\sup_e \sigma_1'(e)<\infty$. Because

$$\sup_e(\sigma_1'(e)+\sigma_2'(e))<\infty \Longrightarrow \sup_e(\sigma_1(e)+\sigma_2(e))<\infty,$$

by Corollary 1.4, the process has algebraic decay for $r\in(1,\ 5/3)$.

(C) Use Theorem B.

As shown in (A) 1), we may take $u_n=(n+1)^{r-1}$. In order for

$$\sum_{n=0}^\infty u_n^2 n^{2q}\pi_n<\infty,$$

we need $r>(2q+3)/3>5/3$. Namely, $q<(3r-3)/2$. When $r\in(0,1)$, we get $q\in(1,\ 3/2)$, which is obviously not good.

(II) Finally, we prove that when $r\leqslant 5/3$, the process is not algebraic convergent with respect to the functional $\widetilde{V}_1$ and hence V_1 defined by (1.9).

Assume that the process has algebraic decay when $1<r\leqslant 5/3$, and the convergence power is $q-1>0$. Let

$$\phi_n=\sum_{k=0}^{n-1}(k+1)^{-s},\qquad s\leqslant r-1$$

which comes from 1) of part (I). In order to have $\sup_n b_n(\phi_{n+1} - \phi_n)^2 < \infty$, we must take $s < r/2$. Then, by Corollary 1.5, we have $\sum_n \pi_n \phi_n^k < \infty$ for all $k < 2q$. Because $q > 1$, we have $\sum_n \pi_n \phi_n^k < \infty$ for all $k \leqslant 2$. In particular,

$$\sum_n \pi_n \phi_n^2 \sim \sum_n n^{-r+2(1-s)} < \infty$$

implies that $-r + 2(1-s) < -1$. That is $s > (3-r)/2$. When $r \leqslant 5/3$, this gives us $s > 2/3$ which is in contradiction with $s \leqslant r - 1 \leqslant 2/3$. $\square$

Proposition 4.5. With respect to $\bar{V}$ defined in Theorem C, Example 4.2 has algebraic decay iff $c > 3$.

Proof. Choose $\phi_n = n$ and apply the Kummer's test to $\sum_n n^{2q}\pi_n$. Set $u_n = n^{2q}\pi_n$ and $v_n = n$. Then,

$$v_n \frac{u_n}{u_{n+1}} - v_{n+1} = \frac{n \cdot n^{2q}}{(n+1)^{2q}(1-c/n)} - (n+1) \sim \frac{(c-2q-1)n^{2q+1}}{n^{2q+1}} = c - 2q + 1$$

Where "$\sim$" comes from

$$(n+1)^{2q+1} \sim n^{2q+1} + (2q+1)n^{2q} + \cdots .$$

So $\sum_n n^{2q}\pi_n$ is finite whenever $c > 2q + 1$. That is, the process is algebraic convergent when $c > 3$.

Now we prove the process is not algebraic convergent when $c \leqslant 3$. Suppose that the process has algebraic decay. Since $\sup_k q_{k,k+1} \leqslant 1 < \infty$, by Theorem C, for all $\alpha < 2q$, we must have $\sum_k k^\alpha \pi_k < \infty$. But $q > 1$, the conclusion should hold for $\alpha = 2$, i.e., $\sum_k k^2\pi_k < \infty$. We prove that this is impossible when $c \leqslant 3$. Let $x_n = n^2\pi_n$ and apply the Gauss' test. We have

$$\begin{aligned}\frac{x_n}{x_{n+1}} &= \frac{n^2}{(n+1)^2(1-c/n)}\\ &= 1 + \frac{c-2}{n} + \frac{1}{n^2}\frac{3(c-1) + (3c-1)/n + c/n^2}{(1+1/n)^2(1-c/n)}\\ &\sim 1 + (c-2)/n + M/n^2.\end{aligned}$$

So, $\sum_n x_n$ is finite if and only if $c - 2 > 1 \iff c > 3$. $\square$

References

[1]. Chen, M. F., *Exponential L^2-convergence and L^2-spectral gap for Markov processes*, Acta Math. Sin., New Ser. 71(1991), 19–37.

[2]. Chen, M. F., *From Markov to Non-Equilibrium Particle Systems*, Singapore, World Scientific, 1992.

[3]. Chen, M. F., *Optimal Markovian coupling and applications*, Acta. Math. Sin., New Ser. 10:3(1994), 260–275.

[4]. Chen, M. F., *Analytic proof of dual variational formula for the first eigenvalue in dimension one*, Sci. Sin. 42:8(1999), 805-815.

[5]. Chen, M. F., *Estimation of spectral gap for Markov chains*, Acta Math. Sin. New Series 12:4 (1996), 337-360.
[6]. Chen, M. F., *Coupling, spectral gap and related topics*, (I): Chin.Sci.Bulletin(1997), 42:14, 1472–1477 (Chinese Edition); 42:16, 1321–1327 (English Edition). (II): 42:15, 1585–1591 (Chin. Ed.); 42:17,1409–1416 (Eng. Ed.). (III):42:16, 1696–1703 (Chin. Ed.); 42:18, 1497–1505 (Eng. Ed.).
[7]. Chen, M. F. and Li, S. F., *Coupling methods for multidimensional diffusion processes*, Ann. Prob. 17:1 (1989), 151–177.
[8]. Deuschel, J. D., *Algebraic L^2 decay of attractive critical processes on the lattice*, Ann. Prob. 22:1(1994), 264–283.
[9]. Diaconis, P. and Stroock, D. W., *Geometric bounds for eigenvalues of Markov chains*, Ann. Appl. Prob. 1:1 (1991) , 36–61.
[10]. Fort, G. and Moulines, E., *V-subgeometric ergodicity for a Hastings-Metropolis algorithm*, 1999 preprint, http://www.statslab.cam.ac.uk/~mcmc/.
[11]. Jarner, S. F. and Roberts, G., *Polynomial convergence rates of Markov chains*, 2000 preprint, http://www.statslab.cam.ac.uk/~mcmc/.
[12]. Liggett, T. M., *Exponential L^2 convergence of attractive reversible nearest particle systems*, Ann. Prob.17:2(1989), 403–432.
[13]. Liggett, T. M., *L^2 Rates of convergence for attractive reversible nearest particle systems: the critical case*, Ann. Prob. 19:3(1991), 935–959.

Department of mathematics, Beijing Normal University, Beijing 100875, China.
E-mail: mfchen@ bnu.edu.cn

Science in China (A) 2000, 39:9 (Chinese Edition), 769–776
43:10 (English Edition), 1051–1059

EXPLICIT BOUNDS OF THE FIRST EIGENVALUE

Mu-Fa Chen

(Dept. of Math., Beijing Normal University, Beijing 100875)
Received January 28, 2000

Abstract *It is proved that the general formulas, obtained recently for the lower bound of the first eigenvalue, can be further bounded by one or two constants depending on the coefficients of the corresponding operators only. Moreover, the ratio of the upper and the lower bounds is no more than four.*

Keywords: First eigenvalue, elliptic operator, Riemannian manifolds, birth-death process

Some general formulas of the first eigenvalue are presented in refs. [1–4] for elliptic operators, Laplacian on Riemannian manifolds and Markov chains. The formulas are expressed in terms of some class of functions. That is making variation with respect to test functions. Several explicit bounds are further presented here, avoiding the use of test functions. It is surprising that the bounds not only control all the essential estimates produced by the formulas but also deduce a simple criterion for the positiveness of the eigenvalue in one-dimensional situation. Further improvement of the bounds will be presented in a subsequent paper.

1 Special case: Illustration of the results and the proofs

The main results and their proofs are illustrated in this section in a particular situation.

Consider differential operator $L = a(x)\mathrm{d}^2/\mathrm{d}x^2 + b(x)\mathrm{d}/\mathrm{d}x$ on $(0, D)$, where $a(x)$ is positive everywhere, with Dirichlet and Neumann boundary at 0 and D (if $D < \infty$) respectively. Assume that

$$\int_0^D \mathrm{d}x e^{C(x)}/a(x) < \infty, \tag{1.1}$$

where $C(x) = \int_0^x b/a$. Consider the (generalized) eigenvalue of L:

$$\lambda_0 = \inf\big\{D(f) : f \in C^1(0,D) \cap C[0,D],\ f(0) = 0,\ \|f\| = 1\big\},$$

where

$$D(f) = \int_0^D a(x)f'(x)^2\pi(\mathrm{d}x), \qquad \pi(\mathrm{d}x) = (a(x)Z)^{-1}e^{C(x)}\mathrm{d}x,$$

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

here and in what follows, Z denotes the normalizing constant, and $\|\cdot\|$ denotes the L^2-norm with respect to π. The following variational formula was presented by Theorem 2.2 in ref. [4]:

$$\lambda_0 \geqslant \xi_0 := \sup_{f\in\mathscr{F}} \inf_{x\in(0,D)} I(f)(x)^{-1}, \tag{1.2}$$

where $\mathscr{F} = \big\{f \in C^1(0,D)\cap C[0,D] : f(0)=0,\ f'|_{(0,D)}>0\big\}$ and

$$I(f)(x) = \frac{e^{-C(x)}}{f'(x)} \int_x^D \frac{f(u)e^{C(u)}}{a(u)}\mathrm{d}u, \qquad x\in(0,D). \tag{1.3}$$

Moreover, it was proved in ref. [4] that the equality in (1.2) holds under mild assumption.

The test function f used in (1.2) is a mimic of the eigenfunction of λ_0. Note that there is no explicit solution of the eigenfunction. More seriously, the eigenvalues and eigenfunctions are very sensitive. For instance, let $D=\infty$, $a(x)\equiv 1$ and $b(x) = -(x+c)$. Then, for the specific value of constant c: 0, 1, $\sqrt{3}$, $\sqrt{3+\sqrt{6}}$, both the eigenvalue λ_0 and the order of its eigenfunction(polynomial) change from 1 to 4 successively. And for the other values of c between the above ones, the eigenfunctions are even not polynomial. Thus, it is hardly imaginable to get a good estimate without using test functions. However, we do have the following result.

Theorem 1.1 Let (1.1) hold and define

$$\begin{aligned} Q(x) &= \int_0^x e^{-C(y)}\mathrm{d}y \int_x^D a(y)^{-1}e^{C(y)}\mathrm{d}y,\\ \delta &= \sup_{x\in(0,D)} Q(x),\\ \delta' &= 2\sup_{x\in(0,D)}\int_0^x Q\,\mathrm{d}\nu^{(x)}, \end{aligned}$$

where $\nu^{(x)}$ is a probability measure on $(0,x)$ with density $e^{-C(y)}/Z^{(x)}$ (and $Z^{(x)}$ is the normalizing constant). Then

$$\delta'^{-1} \geqslant \lambda_0 \geqslant \xi_0 \geqslant (4\delta)^{-1}, \tag{1.4}$$

and moreover $\delta\leqslant\delta'\leqslant 2\delta$. In particular, when $D=\infty$, we have $\lambda_0>0$ iff $\delta<\infty$.

When $D=\infty$, in order to justify $\lambda_0>0$, it suffices to consider the limiting behavior of $Q(x)$ as $x\to\infty$. For this, there are some simpler sufficient conditions. Let the corresponding process be non-explosive on $[0,\infty)$ (with reflecting boundary at 0):

$$\int_0^\infty e^{-C(s)}\mathrm{d}s\int_0^s a(u)^{-1}e^{C(u)}\mathrm{d}u = \infty.$$

By using the l'Hospital's rule, from (1.4), it follows that whenever the limit

$$\kappa := \lim_{x\to\infty}\big[e^C/\sqrt{a}\,\big](x)\int_0^x e^{-C}\ (\leqslant\infty)$$

exists, then $\lambda_0 > 0$ iff $\kappa < \infty$. Especially, if $a(x) \in C^1$, $\lim_{x\to\infty} \big[\sqrt{a}\, e^{-C}\big](x) = \infty$ and the limit $\kappa' := \lim_{x\to\infty} \big[\sqrt{a}/(a'/2 - b)\big](x)$ exists, then $\lambda_0 > 0$ iff $\kappa' < \infty$. Furthermore, recall the Mean Value Theorem: if $f(0) = g(0) = 0$ or $f(D) = g(D) = 0$ but $g'|_{(0,D)} \neq 0$, then

$$\sup_{x\in(0,D)} f(x)/g(x) \leqslant \sup_{x\in(0,D)} f'(x)/g'(x).$$

Thus, if $a \in C^1$, then $\lambda_0 > 0$ once $\sup_{x>0} \big[\sqrt{a}/(a'/2 - b)\big](x) < \infty$.

We point out that the result is meaningful for the three situations mentioned in the abstract. This is due to the coupling method, which reduces the higher-dimensional case to dimension one. To avoid the use of too much notations at the same time, the results are not listed here but discussed case by case in the subsequent sections.

When $b(x) \equiv 0$, the estimate

$$\delta^{-1} \geqslant \lambda_0 \geqslant (4\delta)^{-1}$$

was obtained in ref. [5] and is also true for λ_1 (see ref. [1]). The conclusion also holds for birth-death processes (cf. ref. [3]). But in general δ^{-1} is not an upper bound of λ_1 (see Example 3.10).

Proof of Theorem 1.1 The original motivation comes from ref. [6], in which the weighted Hardy's inequality

$$\int_0^\infty f(x)^2 \nu(\mathrm{d}x) \leqslant A \int_0^\infty f'(x)^2 \lambda(\mathrm{d}x), \qquad f \in C^\infty, \quad f(0) = 0$$

was studied, where the optimal constant A obeys the following estimates:

$$B \leqslant A \leqslant 4B, \tag{1.5}$$

here ν and λ are non-negative Borel measures on $[0,\infty)$,

$$B = \sup_{x>0} \nu[x,\infty) \int_0^x p_\lambda(u)^{-1} \mathrm{d}u,$$

and p_λ is the derivative of the absolutely continuous part of λ with respect to the Lebesgue measure. However, (1.4) is more precise than (1.5) and so a different proof is needed. The methods of the proofs adopted here mainly come from refs. [1]–[4].

(a) The second inequality in (1.4) is just (1.2), proved in ref. [4].

(b) To prove the last inequality in (1.4), we need the following result which is an analog of Lemma 6.1 (2) in ref. [1].

Lemma 1.2 Let m, n be non-negative functions satisfying $\int_0^D m(x)\mathrm{d}x < \infty$ and let

$$c := \sup_{x\in(0,D)} \varphi(x) \int_x^D m(y)\mathrm{d}y < \infty,$$

where $\varphi(x)=\int_0^x n(y)\mathrm{d}y$. Then for every $\gamma\in(0,1)$, we have

$$\int_x^D \varphi(y)^\gamma m(y)\mathrm{d}y \leqslant c(1-\gamma)^{-1}\varphi(x)^{\gamma-1}$$

for all $x\in(0,D)$.

Proof Let $M(x)=\int_x^D m(y)\mathrm{d}y$ and $\gamma\in(0,1)$. Then, by assumption, $M(x)\leqslant c\varphi(x)^{-1}$. By using the integration by parts formula, we get

$$\begin{aligned}\int_x^D \varphi(y)^\gamma m(y)\mathrm{d}y &= -\int_x^D \varphi(y)^\gamma \mathrm{d}M(y)\\ &\leqslant [\varphi^\gamma M](x)+\gamma\int_x^D [\varphi^{\gamma-1}\varphi' M](y)\mathrm{d}y\\ &\leqslant c\varphi(x)^{\gamma-1}+c\gamma\int_x^D \varphi^{\gamma-2}\varphi'\\ &= c\varphi(x)^{\gamma-1}+\frac{c\gamma}{\gamma-1}\int_x^D \mathrm{d}\varphi(y)^{\gamma-1}\\ &\leqslant \frac{c}{1-\gamma}\varphi(x)^{\gamma-1}, \qquad x\in(0,D).\end{aligned}$$

The first and the last inequalities cannot be replaced by equalities because one may ignore a negative term in the case of $D=\infty$. $\square$

Now, take $m(x)=e^{C(x)}/a(x)$ and $n(x)=e^{-C(x)}$. Because of (1.1) and $\delta<\infty$, the assumptions of Lemma 1.2 are satisfied. Then

$$\int_x^D \left[a^{-1}\varphi^\gamma e^C\right](u)\mathrm{d}u \leqslant c(1-\gamma)^{-1}\varphi(x)^{\gamma-1}.$$

Next, take $f(x)=\varphi(x)^\gamma$. Then

$$\begin{aligned}I(f)(x) &= \frac{e^{-C(x)}}{f'(x)}\int_x^D \frac{fe^C}{a}(u)\mathrm{d}u\\ &\leqslant \frac{e^{-C}}{\gamma\varphi^{\gamma-1}e^{-C}}(x)\cdot\frac{\delta}{1-\gamma}\varphi(x)^{\gamma-1}\\ &= \frac{\delta}{\gamma(1-\gamma)}.\end{aligned}$$

Optimizing the right-hand side with respect to γ, we obtain $\gamma=1/2$ and then the required assertion follows.

(c) We now prove the first inequality in (1.4). Fix $x\in(0,D)$. Take

$$f(y)=f_x(y)=\int_0^{x\wedge y} e^{-C(s)}\mathrm{d}s, \qquad y\in(0,D).$$

Then $f'(y) = e^{-C(y)}$ if $y < x$ and $f'(y) = 0$ if $y \in (x, D)$. Furthermore,

$$\|f\|^2 = \int_0^x f(y)^2 \pi(\mathrm{d}y) + f(x)^2 \pi[x, D),$$
$$D(f) = \int_0^x e^{-2C(y)} e^{C(y)} \mathrm{d}y / Z = f(x)/Z,$$

where $\pi[p, q] = \int_p^q \mathrm{d}\pi$. Hence

$$\begin{aligned}
\lambda_0^{-1} &\geqslant \|f\|^2 / D(f) \\
&= Z f(x)^{-1} \int_0^x f(y)^2 \pi(\mathrm{d}y) + Z f(x) \pi[x, D) \\
&= -Z f(x)^{-1} \int_0^x f(y)^2 \mathrm{d}(\pi(y, D)) + Q(x) \\
&= -Z f(x)^{-1} \big[f(y)^2 \pi(y, D)\big]\big|_0^x + Q(x) + 2 f(x)^{-1} \int_0^x e^{-C(y)} Q(y) \mathrm{d}y \\
&= 2 f(x)^{-1} \int_0^x e^{-C(y)} Q(y) \mathrm{d}y \\
&= 2 \int_0^x Q \mathrm{d}\nu^{(x)}.
\end{aligned} \tag{1.6}$$

Making supremum with respect to x, it follows that $\lambda_0 \leqslant {\delta_0'}^{-1}$.

(d) The equalities in (1.6) from the second to the last show that

$$2 \int_0^x Q \mathrm{d}\nu^{(x)} = Z f(x)^{-1} \int_0^x f(y)^2 \pi(\mathrm{d}y) + Q(x) > Q(x).$$

Hence $\delta' \geqslant \delta$. On the other hand, from the definitions, it follows immediately that $\delta' \leqslant 2\delta$. Usually, we have $\delta < \delta'$ unless $\delta = \infty$. $\square$

2 Higher dimensional case: Euclidean space and compact manifolds

This section applies Theorem 1.1 to the higher-dimensional Euclidean space and compact Riemannian manifolds. First, consider elliptic operator

$$L = \sum_{i,j=1}^d a_{ij}(x) \partial_i \partial_j \ + \sum_{i=1}^d b_i(x) \partial_i, \qquad \partial_i = \partial / \partial x_i$$

in $\mathbb{R}^d$, where $a(x) := (a_{ij}(x))$ is positive definite, $a_{ij} \in C^2(\mathbb{R}^d)$,

$$b_i = \sum_{j=1}^d (a_{ij} \partial_j V + \partial_j a_{ij}), \qquad V \in C^2(\mathbb{R}^d).$$

Assume additionally that the corresponding diffusion process is non-explosive, having stationary distribution $\pi(\mathrm{d}x) = Z^{-1} \exp[V(x)] \mathrm{d}x$, where

$$Z := \int \exp[V(x)] \mathrm{d}x < \infty,$$

and its Dirichlet form $(D, \mathscr{D}(D))$ is regular:

$$D(f) = \int \langle a\nabla f, \nabla f\rangle \mathrm{d}\pi, \qquad \mathscr{D}(D) \supset C_0^\infty(\mathbb{R}^d).$$

Since L has trivial maximum eigenvalue 0 in the present situation, we are interested only in the first non-trivial one (i.e., the spectral gap):

$$\lambda_1 = \inf\big\{D(f) : f \in \mathscr{D}(D), \pi(f) = 0, \pi(f^2) = 1\big\},$$

where $\pi(f) = \int f \mathrm{d}\pi$.

The main steps of the study on λ_1 by couplings are as follows. Take and fix a distance $d(x, y)$ in $\mathbb{R}^d$, it belongs to C^2, out of the diagonal. Set $D = \sup_{x,y} d(x, y)$. For each coupling operator $\widetilde{L}$ and $f \in C^2[0, D)$, there always exist two functions A and B in $\mathbb{R}^d \times \mathbb{R}^d$ such that

$$\widetilde{L} f \circ d(x, y) = A(x, y)f''(d(x, y)) + B(x, y)f'(d(x, y)), \qquad x \neq y.$$

The key step of the method is finding a coupling operator $\widetilde{L}$ and a function $f \in C^2[0, D)$ satisfying $f(0) = 0$, $f'|_{(0,D)} > 0$ and $f'' \leqslant 0$ so that for some constant $\delta > 0$,

$$\widetilde{L} f \circ d(x, y) \leqslant -\delta f \circ d(x, y), \qquad x \neq y. \tag{2.1}$$

We now choose α, $\beta \in C(0, D)$ such that $\alpha(r) \leqslant \inf_{d(x,y)=r} A(x, y)$ and $\beta(r) \geqslant \sup_{d(x,y)=r} B(x, y)$. Then, (2.1) holds provided

$$\alpha(r)f''(r) + \beta(r)f'(r) \leqslant -\delta f(r)$$

for $r \in (0, D)$. Thus, the higher-dimensional case is reduced to dimension one.

Replacing $a(x)$ and $b(x)$ used in the last section by $\alpha(r)$ and $\beta(r)$ respectively, define the correspondent function $C(r)$, operator $I(f)$ and the class $\mathscr{F}$ of test functions. Then, the variational formula given by Theorem 4.1 in ref. [1] is as follows:

$$\lambda_1 \geqslant \xi_1 := \sup_{f\in\mathscr{F}} \inf_{r\in(0,D)} I(f)(r)^{-1}. \tag{2.2}$$

Now, define δ and δ' as in Theorem 1.1. From which, one deduces immediately the following result.

Theorem 2.1 $\delta'^{-1} \geqslant \xi_1 \geqslant (4\delta)^{-1}$.

Comparing this theorem with Theorem 1.1, the difference is that here we have upper bound only for ξ_1 rather than λ_1.

We now turn to manifolds. Let M be a compact, connected Riemannian manifold, without or with convex boundary ∂M. Let $L = \Delta + \nabla V$, $V \in C^2(M)$. When $\partial M \neq \emptyset$, we adopt Neumann boundary condition. Next, let $\mathrm{Ric}_M \geqslant -K$ for some $K \in \mathbb{R}$. Denote by d, D and ρ respectively the dimension, diameter and the Riemannian distance. Let $K(V) = \inf\{r : \mathrm{Hess}_V - \mathrm{Ric}_M \leqslant r\}$ and denote by $\mathrm{cut}(x)$ the cut locus of x. Define

$$a_1(r) = \sup\{\langle\nabla\rho(x,\cdot)(y), \nabla V(y)\rangle + \langle\nabla\rho(\cdot,y)(x), \nabla V(x)\rangle : \rho(x, y) = r,\ y \notin \mathrm{cut}(x)\}, \qquad r \in (0, D].$$

By convention, $a_1(0) = 0$. Choose $\gamma \in C[0, D]$ so that

$$\gamma(r) \geqslant \min\left\{K(V)r,\ a_1(r) + 2\sqrt{|K|(d-1)}\, a_2(r)\right\},$$

where $a_2(r) = \tanh\left[\frac{r}{2}\sqrt{K/(d-1)}\right]$ if $K \geqslant 0$ and $a_2(r) = -\tan\left[\frac{r}{2}\sqrt{-K/(d-1)}\right]$ if $K \leqslant 0$. Redefine

$$C(r) = \frac{1}{4}\int_0^r \gamma(s)\mathrm{d}s, \qquad r \in [0, D].$$

Then, the variational formula obtained by ref. [2] can be stated as follows.

$$\lambda_1 \geqslant 4\xi_1 := 4 \sup_{f \in \mathscr{F}} \inf_{r \in (0,D)} f(r) \left\{ \int_0^r e^{-C(s)}\mathrm{d}s \int_s^D \left[e^C f\right](u)\mathrm{d}u \right\}^{-1}, \tag{2.3}$$

where $\mathscr{F} = \{f \in C[0, D] : f|_{(0,D)} > 0\}$. Note that $C(r)$ was used in ref. [2] instead of $e^{C(r)}$ used here. We now have the following result.

Theorem 2.2 Define δ and δ' as in Theorem 1.1 but set $a(x) \equiv 1$ and $b(x) = \gamma(x)/4$. Then $\delta^{-1} \geqslant \delta'^{-1} \geqslant \xi_1 \geqslant (4\delta)^{-1}$.

Proof The proof is similar to the one of Theorem 1.1, but there are two places needed to be modified. The first one is the proof (b). Let $\varphi(r) = \int_0^r e^{-C(s)}\mathrm{d}s$. By Lemma 1.2 $\big($with $n(s) = e^{-C(s)}$, $m(s) = e^{C(s)}$ and $c = \delta\big)$, we have

$$\int_r^D \varphi^\gamma e^C \leqslant \delta(1-\gamma)^{-1}\varphi(r)^{\gamma-1}, \qquad \gamma \in (0, 1).$$

Hence

$$\begin{aligned}
\int_0^r e^{-C(s)}\mathrm{d}s \int_s^D \varphi^\gamma e^C &\leqslant \frac{\delta}{1-\gamma}\int_0^r e^{-C}\varphi^{\gamma-1} \\
&= \frac{\delta}{\gamma(1-\gamma)}\int_0^r \mathrm{d}\varphi^\gamma \\
&= \frac{\delta}{\gamma(1-\gamma)}\varphi(r)^\gamma, \qquad r \in (0, D).
\end{aligned}$$

In particular, setting $\gamma = 1/2$ and $f(r) = \varphi(r)^\gamma$, we obtain $\xi_1 \geqslant (4\delta)^{-1}$.

To complete the proof, one needs to show that ξ_1 is a lower bound of the eigenvalue of operator $L = \mathrm{d}^2/\mathrm{d}r^2 + [\gamma(r)/4]\mathrm{d}/\mathrm{d}r$. Then the upper bound $\xi_1 \leqslant \delta'^{-1}$ follows from Theorem 1.1. The proof for the required assertion is similar to the one of (1.2), but is left to a subsequent paper[1]. $\square$

Example 2.3 Consider the case of zero curvature. Let $V = 0$. Then $\delta = D^2/4$, $\delta' = 3D^2/8$. The precise solution is $4D^2/\pi^2$, which can be deduced by using the test function $\sin(r\pi/2D)$.

[1]Chen M F, Variational formulas and approximation theorems for the first eigenvalue in dimension one, Science in China, Ser. A, 2000, in press

Of course, the above idea is also meaningful for Dirichlet eigenvalue in higher-dimensional situation.

3 The general relation between λ_0 and λ_1 and one-dimensional case.

The main purpose of this section is to deal with λ_1, by comparing it with λ_0. We now study a general relation between λ_0 and λ_1.

Let $(D, \mathscr{D}(D))$ be a Dirichlet form on a general probabilistic space $(E, \mathscr{E}, \pi)$, it determines a Markov transition probability $p(t, x, \mathrm{d}y)$. Assume that $p(t, x, E) = 1$ for all $t \geqslant 0$ and $x \in E$. Define

$$\lambda_1 = \inf\{D(f) : f \in \mathscr{D}(D), \pi(f) = 0, \pi(f^2) = \|f\|^2 = 1\}.$$

For each $A \in \mathscr{E}$ with $\pi(A) \in (0, 1)$, let

$$\lambda_0(A) = \inf\{D(f) : f \in \mathscr{D}(D), f|_{A^c} = 0, \|f\| = 1\}.$$

Then, we have the following result.

Theorem 3.1

$$\begin{aligned}\inf_{\pi(A)\in(0,1/2]} \lambda_0(A) \leqslant \lambda_1 \leqslant \inf_{\pi(A)\in(0,1)} \min\left\{\frac{\lambda_0(A)}{\pi(A^c)}, \frac{\lambda_0(A^c)}{\pi(A)}\right\} \\ \leqslant 2 \inf_{\pi(A)\in(0,1/2]} \lambda_0(A).\end{aligned}$$

In particular, $\lambda_1 > 0$ iff $\inf_{\pi(A)\in(0,1/2]} \lambda_0(A) > 0$.

The theorem also holds for general symmetric forms studied in ref. [7], and improves Theorem 1.4 there.

Proof of Theorem 3.1 First, by spectral representation theorem,

$$D(f) = \lim_{t\downarrow 0} \frac{1}{2t} \int \pi(\mathrm{d}x) \int p(t, x, \mathrm{d}y)[f(y) - f(x)]^2, \qquad f \in L^2(\pi),$$

(cf. [8], §6.7). Replacing $J(\mathrm{d}x, \mathrm{d}y)$ by $\frac{1}{2t}\pi(\mathrm{d}x)p(t, x, \mathrm{d}y)$, in proof (b) of Theorem 1.2 in ref. [7], or in the last paragraph of part 3 in ref. [9], then setting $t \downarrow 0$, it follows that $\lambda_1 \geqslant \inf_{\pi(A)\in(0,1/2]} \lambda_0(A)$.

Next, by Theorem 3.1 in ref. [7], we know that $\lambda_1 \leqslant \lambda_0(A)/\pi(A^c)$ for all A: $\pi(A) \in (0, 1)$. Hence

$$\begin{aligned}\lambda_1 &\leqslant \inf_{\pi(A)\in(0,1)} \min\left\{\lambda_0(A)/\pi(A^c), \lambda_0(A^c)/\pi(A)\right\} \\ &= \inf_{\pi(A)\in(0,1/2]} \min\left\{\lambda_0(A)/\pi(A^c), \lambda_0(A^c)/\pi(A)\right\} \\ &\leqslant \inf_{\pi(A)\in(0,1/2]} \lambda_0(A)/\pi(A^c) \\ &\leqslant 2 \inf_{\pi(A)\in(0,1/2]} \lambda_0(A). \quad \square\end{aligned}$$

The simplest case is that A consists of a single point, say $A = \{0\} \subset E$ for instance. Then the proof becomes rather easy. For simplicity, let $\lambda_0 = \lambda_0(\{0\}^c)$ (but not $\{0\}$). Then, we have the following result.

Proposition 3.2 $\lambda_1 \geqslant \lambda_0$.

Proof Simply noting that $\mathrm{Var}(f) = \|f - \pi(f)\|^2 = \inf_{c\in\mathbb{R}} \|f - c\|^2$, we have

$$\lambda_1 = \inf_{f\neq\text{const.}\, \|f\|<\infty} \frac{D(f)}{\mathrm{Var}(f)} \geqslant \inf_{f\neq\text{const.}\, \|f\|<\infty} \frac{D(f)}{\|f - f(0)\|^2} = \inf_{f(0)=0,\, \|f\|=1} D(f) = \lambda_0. \quad \square$$

In one-dimensional situation, because of the linear order, Theorem 3.1 takes a much simpler form. For instance, the proof of Theorem 3.1 and the property of linear order give us immediately that

$$\lambda_1 \leqslant \inf_{c\in(p,q)} \big\{\big[\lambda_0(p,c)\pi(c,q)^{-1}\big] \wedge \big[\lambda_0(c,q)\pi(p,c)^{-1}\big]\big\}.$$

However, we have a much stronger result as follows.

Theorem 3.3 Let $L = a(x)\mathrm{d}^2/\mathrm{d}x^2 + b(x)\mathrm{d}/\mathrm{d}x$ be an elliptic operator on the interval (p, q), where $a(x)$ is positive everywhere. When p (resp., q) is finite, we adopt Neumann boundary condition. Assume that the process is non-explosive and (1.1) holds. Then,

$$\sup_{c\in(p,q)} \big\{\lambda_0(p,c) \wedge \lambda_0(c,q)\big\} \leqslant \lambda_1 \leqslant \inf_{c\in(p,q)} \big\{\lambda_0(p,c) \vee \lambda_0(c,q)\big\}.$$

Note that when $c \uparrow$, we have $\lambda_0(p, c) \downarrow$ and $\lambda_0(c, q) \uparrow$. Thus, once the two curves $\lambda_0(p, \cdot)$ and $\lambda_0(\cdot, q)$ intersect, the two inequalities become equalities. The conclusion holds once both $a(x)$ and $b(x)$ are continuous. Actually, denoting by x_0 the unique point at which the eigenfunction of λ_1 vanishes, we have $\lambda_1 = \lambda_0(p, x_0) = \lambda_0(x_0, q)$ (the proof needs Theorem 1.1 in the subsequent paper[2]).

Theorem 3.4 Consider birth-death processes. Let $b_i > 0 (i \geqslant 0)$, and $a_i > 0 (i \geqslant 1)$ be the birth and death rates respectively. Define

$$\pi_i = \frac{\mu_i}{\mu}, \qquad \mu_0 = 1,\ \mu_i = \frac{b_0 \cdots b_{i-1}}{a_1 \cdots a_i}, \qquad \mu = \sum_i \mu_i$$

and

$$D(f) = \sum_i \pi_i b_i [f_{i+1} - f_i]^2, \qquad \mathscr{D}(D) = \{f \in L^2(\pi) : D(f) < \infty\}.$$

Assume that $\mu < \infty$ and the process is non-explosive, i.e.

$$\sum_{k=0}^{\infty} (b_k\mu_k)^{-1} \sum_{i=0}^{k} \mu_i = \infty.$$

---

[2]See the footnote in the previous page.
Let c be the medium of π. Then the theorem gives us immediately that

$$\lambda_0(p,c) \wedge \lambda_0(c,q) \leqslant \lambda_1 \leqslant 2\big\{\lambda_0(p,c) \wedge \lambda_0(c,q)\big\}.$$

Reset $\lambda_0([0,k]) = \lambda_0'(k)$, $\lambda_0([k,\infty)) = \lambda_0''(k)$ and adopt the convention $\lambda_0'(-1) = \infty$, here $\lambda_0(A)$ and λ_1 are defined at the beginning of this section. Then we have

$$\sup_{k\geqslant 0}\big\{\lambda_0'(k-1)\wedge\lambda_0''(k+1)\big\}\leqslant \lambda_1\leqslant \inf_{k\geqslant 1}\big\{\lambda_0'(k-1)\vee\lambda_0''(k+1)\big\}.\quad [3]$$

Proof Here, we prove Theorem 3.4 only, the proof of Theorem 3.3 is similar and even simpler. Given $f\in\mathscr{D}(D)$ and $k\geqslant 0$, let $\tilde f = f - f_k$. Then

$$\begin{aligned}
D(f) = D(\tilde f) &= \sum_{i\leqslant k-1}\pi_i b_i[\tilde f_{i+1}-\tilde f_i]^2 + \sum_{i\geqslant k}\pi_i b_i[\tilde f_{i+1}-\tilde f_i]^2\\
&\geqslant \lambda_0'(k-1)\sum_{i\leqslant k-1}\pi_i\tilde f_i^2 + \lambda_0''(k+1)\sum_{i\geqslant k+1}\pi_i\tilde f_i^2\\
&\geqslant [\lambda_0'(k-1)\wedge\lambda_0''(k+1)]\sum_i \pi_i\tilde f_i^2\\
&\geqslant [\lambda_0'(k-1)\wedge\lambda_0''(k+1)]\mathrm{Var}(\tilde f)\\
&= [\lambda_0'(k-1)\wedge\lambda_0''(k+1)]\mathrm{Var}(f).
\end{aligned}$$

Making supremum with respect to k and infimum with respect to f, the required lower bound follows.

We now prove the upper estimate. Given $\varepsilon>0$, take $f_1,\ f_2\geqslant 0$ such that $f_1|_{[k,\infty)}=0,\ f_2|_{[0,k]}=0,\ \|f_1\|=\|f_2\|=1$ and $D(f_1)\leqslant\lambda_0'(k-1)+\varepsilon$, $D(f_2)\leqslant \lambda_0''(k+1)+\varepsilon$. Set $f=-f_1+\alpha f_2$, where α is the constant so that $\pi(f)=0$. Then

$$\begin{aligned}
D(f) &= \sum_{i\geqslant 0}\pi_i b_i[f_{i+1}-f_i]^2\\
&= D(f_1)+\alpha^2 D(f_2)\\
&\leqslant \lambda_0'(k-1)+\varepsilon+(\lambda_0''(k+1)+\varepsilon)\alpha^2\\
&\leqslant (\lambda_0'(k-1)\vee\lambda_0''(k+1)+\varepsilon)\|f\|^2.
\end{aligned}$$

Letting $\varepsilon\to 0$ and then making infimum with respect to $k\geqslant 1$, we obtain the required assertion. $\square$

For birth-death processes, the following variational formulas were presented in refs. [3] and [4].

$$\begin{aligned}
\lambda_0 &= \sup_{w\in\mathscr{W}_0}\ \inf_{i\geqslant 0} I_i(w)^{-1},\\
\lambda_1 &= \sup_{w\in\mathscr{W}_1}\ \inf_{i\geqslant 0} I_i(w)^{-1},
\end{aligned}$$

[3] Addition to the original proof: As mentioned above Theorem 3.3, there is a rough upper bound. If $\pi_0\leqslant 1/2$, then $\pi[0,m-1]\leqslant 1/2$ and $\pi[m+1,\infty)\leqslant 1/2$ for some $m\geqslant 1$. Hence by Theorem 3.1 we have $\lambda_1\leqslant 2[\lambda_0'(m-1)\wedge\lambda_0''(m+1)]$. On the other hand, if $\pi_0>1/2$, then by Theorem 3.1 again, $\lambda_1\leqslant 2\lambda_0''(1)=2[\lambda_0'(-1)\wedge\lambda_0''(1)]$. Therefore, we always have $\lambda_1\leqslant 2\sup_{k\geqslant 0}[\lambda_0'(k-1)\wedge\lambda_0''(k+1)]$. This upper bound matches with the lower bound and so by weighted Hardy inequality, one obtains an estimate up to a factor 8, which is the assertion proved in [10] for the process on the whole line with different proof. Clearly, the assertion made in the theorem is stronger since one gets sharp estimate whenever $\lambda_0'(k-1)=\lambda_0''(k+1)$ for some k. In the extremal case that $\lambda_0''(k+1)=0$, the original bound is still sharp since $\lambda_0'(k-1)\to 0$ as $k\to\infty$ by the ergodicity of the process.

where

$$\begin{aligned} I_i(w) &= [\mu_i b_i (w_{i+1} - w_i)]^{-1} \sum_{j \geqslant i+1} \mu_j w_j, \\ \mathscr{W}_0 &= \{w : w_0 = 0, w_i \text{ is increasing in } i\}, \\ \mathscr{W}_1 &= \{w : w_i \text{ is strictly increasing in } i \text{ and } \pi(w) \geqslant 0\}. \end{aligned}$$

Our new result is as follows.

Theorem 3.5 Let $\mu < \infty$, $Q_i = \sum_{j \leqslant i-1} (\mu_j b_j)^{-1} \sum_{j \geqslant i} \mu_j$ and

$$Q_i' = \left[\sum_{j \leqslant i-1} (\mu_j b_j)^{-1} + (2\mu_i b_i)^{-1} \right] \sum_{j \geqslant i+1} \mu_j.$$

Next, let $\delta = \sup_{n>0} Q_n$ and $\delta' = 2 \sup_{n>0} \sum_{j=0}^{n-1} Q_j' \nu_j^{(n)}$, where $\nu^{(k)}$ is a probability measure on $\{0, 1, \cdots, k-1\}$ with density $\nu_j^{(k)} = (\mu_j b_j)^{-1} / Z^{(k)}$ (and $Z^{(k)}$ is the normalizing constant). Then $\delta'^{-1} \geqslant \lambda_0 \geqslant (4\delta)^{-1}$ and moreover $\delta \leqslant \delta' \leqslant 2\delta$. Assume additionally that the process is non-explosive, then $\lambda_0 / \pi_0 \geqslant \lambda_1 \geqslant \lambda_0$. In particular, λ_0 (resp., λ_1) > 0 iff $\delta < \infty$.

Proof The lower bound of λ_1 comes from Proposition 3.2 (or Theorem 3.4 with $k = 0$). The proof of Theorem 3.1 shows that

$$\lambda_1 \leqslant \inf_{k \geqslant 0} \left\{ \left[\lambda_0'(k-1) \pi[k, \infty)^{-1} \right] \wedge \left[\lambda_0''(k+1) \pi[0, k]^{-1} \right] \right\}.$$

Then the upper bound follows by setting $k = 0$. The proof for the estimates of λ_0 is similar to the one of Theorem 1.1. First, prove the following result, which is the discrete version of Lemma 1.2 and improves Lemma 2.2 (2) in ref. [3].

Lemma 3.6 Let (m_i) and $(n_i \neq 0)$ are non-negative sequences satisfying $\sup_{n>0} \varphi_n \sum_{j=n}^{\infty} m_j =: c < \infty$, where $\varphi_n = \sum_{i=0}^{n-1} n_i$. Then for every $\gamma \in (0, 1)$, we have $\sum_{j \geqslant i} \varphi_j^\gamma m_j \leqslant c(1-\gamma)^{-1} \varphi_i^{\gamma-1}$.

Proof Let $M_n = \sum_{j \geqslant n} m_j$. Fix $N > i$. Then by summation by parts formula and $M_n \leqslant c \varphi_n^{-1}$, we get

$$\sum_{j=i}^{N} \varphi_j^\gamma m_j = \varphi_i^\gamma M_i + \sum_{j=i}^{N} [\varphi_{j+1}^\gamma - \varphi_j^\gamma] M_{j+1} \leqslant c \left\{ \varphi_i^{\gamma-1} + \sum_{j=i}^{N} [\varphi_{j+1}^\gamma - \varphi_j^\gamma] / \varphi_{j+1} \right\}.$$

By using the elementary inequality $\gamma(1-\gamma)^{-1}(x^{\gamma-1} - 1) + x^\gamma \geqslant 1 \, (x > 0)$, it is easy to check that $\varphi_{j+1}^{\gamma-1} - \varphi_j^\gamma / \varphi_{j+1} \leqslant \gamma(1-\gamma)^{-1} [\varphi_j^{\gamma-1} - \varphi_{j+1}^{\gamma-1}]$. Combining this with the last estimate gives us the required assertion. □

We now take $\gamma = 1/2$, $m_i = \mu_i$, $n_i = (\mu_i b_i)^{-1}$ and $c = \delta$. Then

$$I_i(\sqrt{\varphi}) = \frac{1}{b_i \mu_i \left(\sqrt{\varphi_{i+1}} - \sqrt{\varphi_i} \right)} \sum_{j \geqslant i+1} \mu_j \sqrt{\varphi_j} \leqslant \frac{2\delta}{b_i \mu_i \left(\sqrt{\varphi_{i+1}} - \sqrt{\varphi_i} \right)} \cdot \frac{1}{\sqrt{\varphi_{i+1}}} \leqslant 4\delta,$$

since $\left(\sqrt{\varphi_{i+1}} - \sqrt{\varphi_i}\right)\sqrt{\varphi_{i+1}} \geqslant (\varphi_{i+1} - \varphi_i)/2$. Therefore $\lambda_0 \geqslant (4\delta)^{-1}$.

It remains is to show that $\lambda_0 \leqslant \delta'^{-1}$. Fix $k \geqslant 1$ and take

$$f_i = f_i^{(k)} = \sum_{j=0}^{(i-1)\wedge(k-1)} (\pi_j b_j)^{-1}.$$

Then

$$\|f\|^2 = \sum_{i\leqslant k-1} \pi_i f_i^2 + f_k^2 \sum_{i\geqslant k} \pi_i,$$
$$D(f) = \sum_{i\leqslant k-1} \pi_i b_i [f_{i+1} - f_i]^2 = \sum_{i\leqslant k-1} (\pi_i b_i)^{-1} = f_k.$$

By using the summation by parts formula again, we get

$$\begin{aligned}\frac{1}{\lambda_0} &\geqslant \frac{\|f\|^2}{D(f)}\\ &= \frac{1}{f_k}\sum_{i=0}^{k-1} \pi_i f_i^2 + f_k \sum_{i=k}^{\infty} \pi_i\\ &= \frac{1}{f_k}\sum_{i=0}^{k-1} \left(f_{i+1}^2 - f_i^2\right) \sum_{j=i+1}^{\infty} \pi_j\\ &= \frac{2}{f_k}\sum_{i=0}^{k-1} \frac{\mu Q_i'}{\mu_i b_i}\\ &= 2\sum_{i=0}^{k-1} Q_i' \nu_i^{(k)}.\end{aligned}$$

Making supremum with respect to $k \geqslant 1$ gives us $\lambda_0 \leqslant \delta'^{-1}$. Similar to the continuous case, from the above formula, it follows that

$$2\sum_{i=0}^{k-1} Q_i' \nu_i^{(k)} = \frac{1}{f_k}\sum_{i=0}^{k-1} \pi_i f_i^2 + f_k \sum_{i=k}^{\infty} \pi_i > f_k \sum_{i=k}^{\infty} \pi_i = Q_k.$$

Hence $\delta' \geqslant \delta$. The conclusion $\delta' \leqslant 2\delta$ is easy because $Q_i' < Q_{i+1}$ for all $i \geqslant 0$ and so

$$\delta' = \sup_{k\geqslant 1} 2\sum_{i=0}^{k-1} Q_i' \nu_i^{(k)} \leqslant 2\sup_{i\geqslant 0} Q_i' \leqslant 2\sup_{i\geqslant 1} Q_i = 2\delta. \quad \square$$

Because λ_1 coincides with exponential convergence rate (cf. Theorem 9.21 in ref. [8]), Theorem 3.5 gives us at the same time (and is indeed for the first time) an explicit criterion for exponential ergodicity. By using comparison method (cf. Theorem 4.58 in ref. [8]), this result can be further applied to a class of multidimensional Markov chains. Finally, we return to the case of half-line discussed at the beginning of the paper.

Theorem 3.7 Consider the operator $L = a(x)\mathrm{d}^2/\mathrm{d}x^2 + b(x)\mathrm{d}/\mathrm{d}x$ on $[0,\infty)$, where $a(x)$ is positive everywhere. Let the process be non-explosive (equivalently, $\int_0^\infty e^{-C(s)}\mathrm{d}s \int_0^s a(u)^{-1}e^{C(u)}\mathrm{d}u = \infty$) and let (1.1) hold. Then

$$(4\delta'(c_0))^{-1} \leqslant \lambda_1 \leqslant \delta'(c_0)^{-1},$$

where

$$\delta'(c) = \sup_{x\in(0,c)} \int_c^x e^{-C}\int_x^0 e^C/a, \qquad \delta''(c) = \sup_{x\in(c,\infty)} \int_c^x e^{-C}\int_x^\infty e^C/a,$$

and c_0 is the unique solution to the equation $\delta'(c) = \delta''(c)$, $c \in [0,\infty]$. In particular, $\lambda_1 > 0$ iff $\delta < \infty$.

Proof First, when $c\uparrow$, we have $\delta'(c)\uparrow$ and $\delta''(c)\downarrow$. Obviously,

$$\lim_{c\to 0}\delta'(c) = 0, \qquad \lim_{c\to 0}\delta''(c) = \delta$$

and moreover $\lim_{c\to\infty}\delta''(c) \leqslant \delta$. On the other hand, since the process is non-explosive, when $x\uparrow\infty$, we have $\varphi(x) = \int_0^x e^{-C}\uparrow\infty$. It follows that

$$\delta'(c) \geqslant \int_1^c e^{-C}\int_0^1 e^C/a \to \infty \quad \text{as} \quad c\to\infty.$$

Next, when $c_1 < c_2$, we have

$$0 < \int_x^{c_2} e^{-C}\int_0^x e^C/a - \int_x^{c_1} e^{-C}\int_0^x e^C/a \leqslant \left[\int_0^{c_2} e^C/a\right]\int_{c_1}^{c_2} e^{-C} \to 0, \quad \text{if } c_2 - c_1 \to 0;$$

$$0 < \int_{c_1}^x e^{-C}\int_x^\infty e^C/a - \int_{c_2}^x e^{-C}\int_x^\infty e^C/a \leqslant \left[\int_{c_1}^\infty e^C/a\right]\int_{c_1}^{c_2} e^{-C} \to 0, \quad \text{if } c_2 - c_1 \to 0.$$

Hence both $\delta'(c)$ and $\delta''(c)$ are continuous in c. Therefore, the equation $\delta'(c) = \delta''(c)$ has a unique solution. Then the first assertion follows from Theorem 3.3. Clearly $\delta < \infty$ iff $\delta''(c) < \infty$. Hence we obtain the last assertion. □

In a similar way, one can deduce a criterion for the existence of spectral gap of diffusion on the full-line (cf. sec. 3 in ref. [1]). One may also study the bounds for the processes on finite intervals.

Example 3.8 Take $b(x) \equiv 0$. Define δ as in Theorem 1.1. Then, by Theorem 1.1 and Corollary 2.5 (5) in ref. [1], we know that $\delta^{-1} \geqslant \lambda_1 \geqslant \lambda_0 \geqslant (4\delta)^{-1}$. In particular, when $a(x) = (1+x)^2$, we have $\delta = 1$ (but $\delta' = 2$) and $\lambda_1 = \lambda_0 = 1/4$.[4] Hence, our lower bound is exact.

Example 3.9 Take $a(x) \equiv 1$ and $b(x) = -x$. Then Example 2.10 in ref. [1] gives us $\lambda_1 = 2$. It is easy to check that $\lambda_0 = 1$ (having eigenfunction $g(x) = x$) and $\delta \approx 0.4788$ (but $\delta' \approx 0.9285$). Hence $\delta^{-1} > \lambda_1 > \lambda_0 > (4\delta)^{-1}$.

[4]The eigenfunctions of λ_0 and λ_1 are $\sqrt{x+1}\,\log(x+1)$ and $\sqrt{x+1}\,(\log(x+1)-2)$, respectively.

Example 3.10 An extreme example is the space with two points $\{0,1\}$ only. Then $\lambda_1 = \lambda_0/\pi_0$. Therefore the upper bound of λ_1 in Theorem 3.5 is exact but $\delta^{-1} = \lambda_0 < \lambda_1$. Thus, δ^{-1} is not an upper bound of λ_1 in general.

Added in proof In the recent paper [10], the estimate $\delta^{-1} \geqslant \lambda_0 \geqslant (4\delta)^{-1}$ for birth-death processes is also obtained by using the discrete Hardy's inequality. Refer also to refs. [11–13] for related study and further references.

Acknowledgements The author acknowledges Dr. Mao Yong-Hua for his help, he also obtains independently some results in the case of half-line. This work was supported in part by National Natural Science Foundation of China (No. 19631060), Mathematical Tian Yuan Foundation, Qiu Shi Science & Technology Foundation, RFDP and MCEC.

References

[1]. Chen M F Wang F Y, *Estimation of spectral gap for elliptic operators*, Trans Amer Math Soc 1997, 349(3), 1239-1267.

[2]. Chen M F Wang F Y, *General formula for lower bound of the first eigenvalue on Riemannian manifolds*, Sci. Sin. 1997, 40:4, 384–394.

[3]. Chen M F, *Estimation of spectral gap for Markov chains*, Acta Math Sin New Series 1996, 12(4), 337-360.

[4]. Chen M F, *Analytic proof of dual variational formula for the first eigenvalue in dimension one*, Sci. in China (A) 1999, 42:8, 805–815..

[5]. Kac I S and Krein M G, *Criteria for the discreteness of the spectrum of a singular string*, Izv Vyss Ucebn Zaved Mat 1958, 2, 136–153 (In Russian).

[6]. Muckenhoupt, B, *Hardy's inequality with weights*, Studia Math 1972, XLIV, 31–38.

[7]. Chen M F Wang F Y, *Cheeger's inequalities for general symmetric forms and existence criteria for spectral gap*, Ann. Prob., 28:1(2000), 235–257. **Abstract**: Chin. Sci. Bulletin 1998, 43:18, 1516–1519.

[8]. Chen M F, *From Markov Chains to Non-Equilibrium Particle Systems*, World Scientific, Singapore.

[9]. Chen M F, *Eigenvalues, inequalities and ergodic theory* (II), Advances in Math. (China) 28:6 (1999), 481–505.

[10]. Miclo, L, *An example of application of discrete Hardy's inequalities*, Markov Processes Relat. Fields 5 (1999), 319–330.

[11]. Miclo L, *Relations entre isopérimétrie et trou spectral pour les chaînes de Markov finies*, Probab Th Rel Fields 114 (1999), 431–485.

[12]. Bobkov S G and Götze F, *Discrete isoperimetric and Poincaré inequalities*, Probab Th Rel Fields 114 (1999), 245–277.

[13]. Bobkov S G and Götze F, *Exponential integrability and transportation cost related to logarithmic Sobolev inequalities*, J Funct Anal 163 (1999), 1–28.

Science in China (A) 2001, 31:1 (Chinese Edition), 28–36
44:4 (English Edition), 409–418

VARIATIONAL FORMULAS AND APPROXIMATION THEOREMS FOR THE FIRST EIGENVALUE IN DIMENSION ONE

Mu-Fa Chen

(Dept. of Math., Beijing Normal University, Beijing 100875)
Received March 15, 2000

Abstract *Some complete variational formulas and approximation theorems for the first eigenvalue of elleptic operators in dimension one or a class of Markov chains are presented.*

Keywords: First eigenvalue, variational formula, elliptic operator, birth-death process

As a continuation of ref. [1], some complete variational formulas and approximation theorems for the first eigenvalue in dimension one are presented. The upper bound part of the formulas are dual of the variational formulas for the lower bound introduced by refs. [2]–[5], but they are completely different to the classical ones. As shown in the mentioned papers, the results obtained in the paper can be immediately applied to higher-dimensional situation and also to Riemannian manifolds. This will be also discussed in a subsequent paper.

1 Continuous case

Consider differential operator $L = a(x)\mathrm{d}^2/\mathrm{d}x^2 + b(x)\mathrm{d}/\mathrm{d}x$ on $(0, D)$, where $a(x)$ is positive everywhere, with Dirichlet and Neumann boundary at 0 and D (if $D < \infty$) respectively. Assume that

$$\int_0^D \mathrm{d}x e^{C(x)}/a(x) < \infty, \tag{1.1}$$

where $C(x) = \int_0^x b/a$. Two eigenvalues we are interested in are as follows:

$$\lambda_0 = \inf\{D(f) : f \in C^1(0, D) \cap C[0, D],\ f(0) = 0,\ \|f\| = 1\},$$
$$\lambda_1 = \inf\{D(f) : f \in C^1(0, D) \cap C[0, D],\ \pi(f) = 0,\ \|f\| = 1\},$$

where

$$D(f) = \int_0^D a(x)f'(x)^2\pi(\mathrm{d}x), \qquad \pi(\mathrm{d}x) = (a(x)Z)^{-1}e^{C(x)}\mathrm{d}x,$$

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

Z is the normalizing constant, $\pi(f) = \int f \mathrm{d}\pi$ and $\|\cdot\|$ is the L^2-norm with respect to π. When $D < \infty$, λ_0 and λ_1 are nothing but the ordinary eigenvalues of the operator L, and the above formulas are called the classical variational formula of λ_0 and λ_1 respectively. In the study on λ_1, we always assume that the process is non-explosive:

$$\int_0^\infty e^{-C(s)}\mathrm{d}s \int_0^s a(u)^{-1}e^{C(u)}\mathrm{d}u = \infty. \tag{1.2}$$

To state the main results, we need some notations. First, there are two operators

$$\begin{aligned} I(f)(x) &= \frac{e^{-C(x)}}{f'(x)}\int_x^D \big[fe^C/a\big](u)\mathrm{d}u, \\ I\!I(f)(x) &= \frac{1}{f(x)}\int_0^x \mathrm{d}y e^{-C(y)}\int_y^D \big[fe^C/a\big](u)\mathrm{d}u \end{aligned}$$

(here I and $I\!I$ represent respectively single and double integrals.[1]). Next, define[2]

$$\begin{aligned} &\mathscr{F}' = \big\{f \in C[0,D] : f(0) = 0,\ \text{there exists } x_0 \in (0,D] \text{ so that} \\ &\qquad\qquad f = f(\cdot \wedge x_0) \text{ and } f|_{(0,x_0)} > 0\big\}, \\ &\mathscr{F}'' = \big\{f \in C[0,D] : f(0) = 0, f|_{(0,D)} > 0\big\}, \\ &\xi_0' = \inf_{f\in\mathscr{F}'}\sup_{x\in(0,D)} I\!I(f)(x)^{-1}, \qquad \xi_0'' = \sup_{f\in\mathscr{F}''}\inf_{x\in(0,D)} I\!I(f)(x)^{-1}, \\ &\widetilde{\mathscr{F}}' = \big\{f \in C[0,D] : f(0) = 0,\ \text{there exists } x_0 \in (0,D] \text{ so that} \\ &\qquad\qquad f = f(\cdot \wedge x_0),\ f \in C^1(0,x_0) \text{ and } f'|_{(0,x_0)} > 0\big\}, \\ &\widetilde{\mathscr{F}}'' = \big\{f \in C[0,D] \cap C^1(0,D) : f(0) = 0,\ f'|_{(0,D)} > 0\big\}, \\ &\tilde{\xi}_0' = \inf_{f\in\widetilde{\mathscr{F}}'}\sup_{x\in(0,D)} I(f)(x)^{-1}, \qquad \tilde{\xi}_0'' = \sup_{f\in\widetilde{\mathscr{F}}''}\inf_{x\in(0,D)} I(f)(x)^{-1}. \end{aligned}$$

When $D = \infty$, one should replace $[0,D]$ by $[0,\infty)$ but we will not mention again in what follows. Throughout the paper, the superscript single-prime or double-prime denote respectively the lower and upper bounds. Note that ξ_0' is the dual of

---

[1][August 16, 2007: Remarks are added in this paper by footnotes. Some of them as well as some improvements will be published in a subsequent paper.]

This becomes more clear in the expression $fI\!I(f)$ which is our mimic of the eigenfunction and is used often subsequently. If we write $I\!I(f)$ as g/f, then the operator $I(f)$ simply means that g'/f'

[2]Note that $\mathscr{F}'$ is simply a modification of $\mathscr{F}''$ by stopping the functions in $\mathscr{F}''$ somewhere. Hence the latter one is more essential than the former one. The quantities ξ_0' and ξ_0'', defined in terms of the operator $I\!I$, are used to estimate λ_0 from above and below, respectively. The single prime and double prime are designed for the upper estimate and the lower one, respectively. We have thus explained the meaning of the notations in the first four lines. Similarly, by using the operator I instead of $I\!I$, we define the notations with additional tilde in the last four lines.

Here and also above Theorem 1.3, when $x_0 = \infty$, in the definition of $\mathscr{F}'$ and $\widetilde{\mathscr{F}}'$, and additional condition $f \in L^2(\pi)$ is required. Similar change ($f \in L^2(\pi)$ if $k = \infty$) is needed for the discrete case (above Theorems 2.1 and 2.3), but we will not mention again.

ξ_0'', they use operator $II(f)$; and $\tilde{\xi}_0'$ is the dual of $\tilde{\xi}_0''$, they use operator $I(f)$. Since different operators have different domains, they use different classes of functions. The set $\mathscr{F}''$ is natural and the weaker condition used in the set $\mathscr{F}'$ is more convenient for applications. The variational formula for λ_0 is as follows.

Theorem 1.1 Let (1.1) hold. Then $\tilde{\xi}_0' = \xi_0' \geqslant \lambda_0 \geqslant \xi_0'' = \tilde{\xi}_0''$. If additionally, a and b are continuous, then the two inequalities all become equalities.

The explicit estimates $\delta^{-1} \geqslant \lambda_0 \geqslant (4\delta)^{-1}$, where

$$\delta = \sup_{x\in(0,D)} \int_0^x e^{-C} \int_x^D e^C/a,$$

are presented in ref. [1]. The next result is further an approximation procedure.

Theorem 1.2 Let (1.1) hold. Set $\varphi(x) = \int_0^x e^{-C(y)}\mathrm{d}y$.

(1) Define $f_1 = \sqrt{\varphi}$, $f_n = f_{n-1} II(f_{n-1})$ and $\delta_n'' = \sup_{x\in(0,D)} II(f_n)\ (x)$. Then δ_n'' is decreasing in n and

$$\lambda_0 \geqslant \lim_{n\to\infty} \delta_n''^{-1} \geqslant \delta_1''^{-1} \geqslant (4\delta)^{-1}.$$

(2) Fix $x_0 \in (0, D)$. Define[3, 4]

$$f_1^{(x_0)} = \varphi(\cdot \wedge x_0), \qquad f_n^{(x_0)} = f_{n-1}^{(x_0)}(\cdot \wedge x_0) II\big(f_{n-1}^{(x_0)}(\cdot \wedge x_0)\big)$$

and

$$\delta_n' = \sup_{x_0\in(0,D)} \inf_{x\in(0,D)} II\big(f_n^{(x_0)}(\cdot \wedge x_0)\big)(x).$$

[3]More explicitly, we have

$$\begin{aligned} f_n^{(x_0)}(x) &= \int_0^x e^{-C(y)}\mathrm{d}y \int_y^D f_{n-1}^{(x_0)}(\cdot \wedge x_0)\frac{e^C}{a} \\ &= \int_0^D f_{n-1}^{(x_0)}(\cdot \wedge x_0)\, \varphi(\cdot \wedge x)\frac{e^C}{a} \\ &= \int_0^{x_0} f_{n-1}^{(x_0)}\, \varphi(\cdot \wedge x)\frac{e^C}{a} + f_{n-1}^{(x_0)}(x_0) \int_{x_0}^D \varphi(\cdot \wedge x)\frac{e^C}{a}. \end{aligned}$$

In particular,

$$f_n^{(x_0)}(x \wedge x_0) = \int_0^{x\wedge x_0} f_{n-1}^{(x_0)}\, \varphi \frac{e^C}{a} + \varphi(x\wedge x_0) \int_{x\wedge x_0}^{x_0} f_{n-1}^{(x_0)} \frac{e^C}{a} + f_{n-1}^{(x_0)}(x_0)\, \varphi(x\wedge x_0) \int_{x_0}^D \frac{e^C}{a}.$$

By integration by parts formula, we also have another expression:

$$f_n^{(x_0)}(x) = \int_0^{x\vee x_0} \Big(f_{n-1}^{(x_0)}(\cdot \wedge x_0)\, \varphi(\cdot \wedge x)\Big)' \psi,$$

where $\psi(x) = \int_x^D e^C/a$.

[4]see Appendix.

Then δ'_n is increasing in n and[5, 6]

$$\delta^{-1} \geqslant {\delta'_1}^{-1} \geqslant \lim_{n\to\infty} {\delta'_n}^{-1} \geqslant \lambda_0.$$

$(2)'$ Replace the initial function in (2) by $f_1 = \sqrt{\varphi}$, and define

$$f_n = f_{n-1} I\!I(f_{n-1}) \quad \text{and} \quad \delta'_n = \inf_{x\in(0,D)} I\!I(f_n)(x),$$

then δ'_n is increasing in n and

$${\delta'_n}^{-1} \geqslant \lambda_0, \qquad n \geqslant 2.$$

Here we adopt the convention $1/0 = \infty$.

Assertions (1) and $(2)'$ can be restated as follows: the inverse of the supremum (infimum) of function $I\!I(f_n)$ is a lower (upper) bound of λ_0.

Finally, in the definitions of δ'_n and $\delta''_n (n \geqslant 1)$, replacing $I\!I(f)$ by $I(f)$ everywhere, one obtains $\tilde{\delta}'_n$ and $\tilde{\delta}''_n$, then

$$\delta''_n \leqslant \tilde{\delta}''_n, \quad \delta'_n \geqslant \tilde{\delta}'_n \,(n \geqslant 1) \quad \text{and} \quad \tilde{\delta}''_n \leqslant \delta''_{n-1}, \quad \tilde{\delta}'_n \geqslant \delta'_{n-1} \,(n \geqslant 2).$$

In particular, the modified assertions (1)–$(2)'$ all hold, the only change is replacing ${\delta'_1}^{-1}$ at the end of (2) by $\tilde{\delta}_2'^{-1}$.

For Theorem 1.1, the only known result is $\lambda_0 \geqslant \tilde{\xi}''_0$ (cf. Theorem 2.2 in ref. [5]). Thus, Theorem 1.1 completes the whole variational formula for λ_0. In this theorem, $\tilde{\xi}'_0$ and ξ'_0 are defined by using $I(f)$ and $I\!I(f)$ respectively. The difference is that for fixed f, $I(f)$ is easier to compute but is not sharper than $I\!I(f)$. The idea of iteration given in Theorem 1.2 comes from refs. [2] and sec. 2.3 in ref. [6], the unified initial function comes from ref. [1]. Assertions (1) and $(2)'$ are completely symmetric and so we will ignore $(2)'$ in what follows. Because $I\!I(f_n)$ is bounded above, the lower bound is always non-trivial. However, the minimum of some $I\!I(f_n)$ can be zero and the function φ may not be integrable in general, this leads to the modified form (2), providing a non-trivial upper bound. Theorem 1.2 is also an approximation result for the eigenfunction of λ_0. Actually, each f_n serves as an approximation of the eigenfunction.

In the later part of this section, we deal with λ_1 in the continuous situation and the next section is devoted to handle with λ_0 and λ_1 in the discrete case. There are two theorems in each case for each λ_0 or λ_1.

Proof of Theorem 1.1 Given h with $h|_{(0,D)} > 0$, then for every g: $g(0) = 0$,

[5]We remark that δ'_1 here coincides with δ' introduced in [1; Theorem 1.1]. See the footnote of Theorem 2.2 below.

[6]see Appendix.

$\|g\| = 1$, by Cauchy-Schwarz inequality, we have

$$\begin{aligned}
1 &= \int_0^D g(x)^2 \pi(\mathrm{d}x) \\
&= \int_0^D \pi(\mathrm{d}x) \bigg[\int_0^x g'(u)\mathrm{d}u\bigg]^2 \\
&\leqslant \int_0^D \pi(\mathrm{d}x) \int_0^x \big[g'^2 e^C h^{-1}\big](u)\mathrm{d}u \int_0^x \big[he^{-C}\big](\xi)\mathrm{d}\xi \\
&= \int_0^D a(u) g'(u)^2 \pi(\mathrm{d}u) \frac{Z}{h(u)} \int_u^D \pi(\mathrm{d}x) \int_0^x he^{-C} \\
&\leqslant D(g) \sup_{x\in(0,D)} \frac{1}{h(x)} \int_x^D \frac{e^{C(y)}}{a(y)}\mathrm{d}y \int_0^y he^{-C} \\
&=: D(g) \sup_{x\in(0,D)} H(x).
\end{aligned} \tag{1.3}$$

Now, let $f \in \mathscr{F}''$ satisfy $\sup_{x\in(0,D)} I\!I(f)(x) =: c < \infty$. Take $h(x) = \int_x^D f a^{-1} e^C$. Then we have not only

$$\int_0^x e^{-C(y)}\mathrm{d}y \int_y^D \frac{fe^C}{a} \leqslant cf(x) < \infty, \qquad \int_x^D \frac{fe^C}{a} < \infty$$

for all $x \in (0, D)$, but also

$$H(x) = \int_x^D \frac{e^C}{a}(y)\mathrm{d}y \int_0^y e^{-C(u)}\mathrm{d}u \int_u^D \frac{fe^C}{a} \leqslant c\int_x^D \frac{fe^C}{a} < \infty.^{7}$$

Since $\lim_{x\to D} h(x) = 0$, by (Cauchy's differential) mean value theorem, we get

$$\sup_{x\in(0,D)} H(x) \leqslant \sup_{x\in(0,D)} \bigg[-\frac{e^C}{ah'}(x)\bigg] \int_0^x he^{-C} = \sup_{x\in(0,D)} I\!I(f)(x). \tag{1.4}$$

Because g is arbitrary, by (1.3) and (1.4), it follows that $\lambda_0 \geqslant \xi_0''$.

For each $f \in \widetilde{\mathscr{F}}''$, without loss of generality, assume that

$$\sup_{x\in(0,D)} I(f)(x) < \infty.$$

By the mean value theorem, $\sup_{x\in(0,D)} I\!I(f)(x) \leqslant \sup_{x\in(0,D)} I(f)(x)$. But $\mathscr{F}'' \supset \widetilde{\mathscr{F}}''$, so

$$\xi_0'' = \sup_{f\in\mathscr{F}''} \inf_{x\in(0,D)} I\!I(f)(x)^{-1} \geqslant \sup_{f\in\widetilde{\mathscr{F}}''} \inf_{x\in(0,D)} I(f)(x)^{-1} = \tilde{\xi}_0''.$$

[7]This integrability was missed in the original paper where instead of this, a truncating argument was proposed.

Conversely, for a given $f \in \mathscr{F}''$ with $\sup_{x\in(0,D)} I\!I(f)(x) =: c < \infty$, let $g = fI\!I(f)$. Then we have $g \in \widetilde{\mathscr{F}}''$ and as in the last paragraph that

$$\int_x^D \frac{ge^C}{a} = \int_x^D \frac{e^C}{a}(y)\mathrm{d}y \int_0^y e^{-C(u)}\mathrm{d}u \int_u^D \frac{fe^C}{a} \leqslant c\int_x^D \frac{fe^C}{a} < \infty.$$

By using the mean value theorem again, we obtain

$$\begin{aligned} I(g)(x) &= \int_x^D ga^{-1}e^C \Big/ \int_x^D fa^{-1}e^C \\ &\leqslant \sup_{x\in(0,D)} (g/f)(x) \\ &= \sup_{x\in(0,D)} I\!I(f)(x). \end{aligned}$$

Hence $\inf_{x\in(0,D)} I\!I(f)(x)^{-1} \leqslant \inf_{x\in(0,D)} I(g)(x)^{-1} \leqslant \tilde{\xi}_0''$. Making supremum with respect to $f \in \mathscr{F}''$, it follows that $\xi_0'' \leqslant \tilde{\xi}_0''$. An alternative proof of the this assertion is using the identity

$$\big(e^C g'\big)' = -fe^C/a, \tag{1.5}$$

[8] We have thus proved that $\xi_0'' = \tilde{\xi}_0''$.

As for $\tilde{\xi}_0' = \xi_0'$, the proof is a dual of the above one, exchanging supremum and infimum, making inverse order of the inequalities and redefining $g = [fI\!I(f)](\cdot \wedge x_0)$.

Let $f \in \mathscr{F}'$ satisfy $f = f(\cdot \wedge x_0)$ and $c := \sup_{x\in(0,D)} I\!I(f)(x)^{-1} < \infty$ and let $g_0 = [fI\!I(f)](\cdot \wedge x_0)$. Then g_0 is bounded and (1.5) holds on $(0, x_0)$. By integration by parts formula, we get

$$\begin{aligned} \int_0^D {g_0'}^2 e^C &= [g_0 g_0' e^C](x_0-) - \int_0^{x_0} g_0\big(e^C g_0'\big)' \\ &= \int_0^{x_0} g_0 fe^C/a + g_0(x_0)\int_{x_0}^D fe^C/a \\ &= \int_0^D g_0 fe^C/a \\ &\leqslant \int_0^D \big(g_0^2 e^C/a\big) \sup_{x\in(0,D)} f/g_0 \\ &= c\int_0^D g_0^2 e^C/a. \end{aligned}$$

[8]since

$$I(g)(x) = \int_x^D \frac{ge^C}{a} \Big/ (e^C g')(x) \leqslant \sup_x \Big[-\frac{ge^C}{a}\Big/(e^C g')'\Big](x) = \sup_x \frac{g}{f}(x).$$

Here we have used the fact that

$$\sup_{x\in(0,D)} f/g_0 = \sup_{x\in(0,x_0)} f/g_0 = \sup_{x\in(0,D)} I\!I(f)(x)^{-1}.$$

Hence $\lambda_0 \leqslant$ and furthermore $\lambda_0 \leqslant \xi_0'$.[9]

Finally, we prove the last assertion. Let a and b are continuous and $a > 0$. By existence theorem of solution to the Sturm-Liouville eigenvalue problem (if $D < \infty$) or to a system of linear differential equations (if $D = \infty$), it follows that there is a non-trivial solution to the equation $Lf = -\lambda_0 f$, $f(0) = 0$ and $f'(D) = 0$ if $D < \infty$.[10] If $\lambda_0 > 0$, then by Theorem 2.2 in ref. [5] and sec. 6 of ref. [2], one may assume that $f'|_{(0,D)} > 0$. From this, it is easy to check that $I(f)(x) \equiv I\!I(f)(x) \equiv \lambda_0^{-1}$. Examining the proofs for the lower and upper bounds, one sees that the equalities should hold everywhere[11]. If $\lambda_0 = 0$, take $D_n \uparrow \infty$ and denote by $\lambda_0(D_n)$ the corresponding λ_0 determined by $L|_{(0,D_n)}$. By using the proof of Lemma 5.1 in ref. [2], we obtain $\lambda_0(D_n) \downarrow \lambda_0$. Thus, when n is large enough, $\lambda_0(D_n) \leqslant \varepsilon$. Now, let f_n be a solution to the eigen-equation $Lf_n = -\lambda_0(D_n) f_n$ $\big(f_n(0) = 0,\ f_n'(D_n) = 0\big)$ on $(0, D_n)$. Let $f_n = f_n(D_n)$ on (D_n, ∞). Then the above proof shows that

$$\begin{aligned}
\lambda_0(D_n)^{-1} &= \inf_{x\in(0,D_n)} f_n(x)^{-1} \int_0^x e^{-C(y)}\mathrm{d}y \int_y^{D_n} f_n e^C/a \\
&\leqslant \inf_{x\in(0,D_n)} I\!I(f_n)(x) \\
&= \inf_{x\in(0,D)} I\!I(f_n)(x).
\end{aligned}$$

It follows that

$$\lambda_0 = 0 \leqslant \inf_{f\in\mathscr{F}''} \sup_{x\in(0,D)} I\!I(f)(x)^{-1} \leqslant \sup_{x\in(0,D)} I\!I(f_n)(x)^{-1} \leqslant \lambda_0(D_n) \leqslant \varepsilon.$$

Since ε is arbitrary, the first inequality can be also replaced by equality. □

Proof of Theorem 1.2 Condition $\delta < \infty$ implies that

$$\int_0^D \sqrt{\varphi} e^C/a \leqslant \sqrt{\delta} \int_0^D \bigg(\int_x^D e^C/a\bigg)^{-1/2} e^C/a = 2\sqrt{\delta Z} < \infty.$$

[9]When $x_0 = \infty$, an additional condition that $g_0 \in L^2(\pi)$ is required here. This is the reason why we made a change to $\mathscr{F}'$ and $\widetilde{\mathscr{F}}'$.

[10]The problem is easier here since we fix the constant λ_0 given by the classical variational formula (i.e., the formula below (1.1)). Actually, only the continuity of a and C, but not b, is required in the analytic proof, hence the measurability of b is enough for the assertion. This is proved in the paper "Dual variational formulas for the first Dirichlet eigenvalue on half-line, Sci. China (A) 2003, 46:6, 847–861", Proposition 1.2, by Chen, M.F., Zhang, Y.H. and Zhao, X.L.

[11]Because of the truncating argument, for the equality in the upper estimate, one needs an approximating procedure. This is not difficult and will be given in a subsequent paper.

Hence $\sqrt{\varphi} \in L^1(\pi)$ [these two conditions are needed for the initial function $\sqrt{\varphi}$. In practice, one can certainly choose some more convenient functions]. Furthermore, as did in the second paragraph in the last proof and by using induction, it follows that $f_n \in L^1(\pi)$ for all n. By Theorem 1.1, $\lambda_0 \geqslant \xi_0'' \geqslant \delta_n''^{-1}$. Then by the mean value theorem and the proof (b) of Theorem 1.1 in ref. [1], we get $\delta_1'' \leqslant 4\delta$. On the other hand, by definition of f_n and (1.5), we have

$$\big(-e^C f_n'\big)' = a^{-1} e^C f_{n-1} \geqslant a^{-1} f_n e^C \delta_{n-1}''^{-1}. \tag{1.6}$$

That is, $f_n e^C/a \leqslant \delta_{n-1}''\big(-e^C f_n'\big)'$. Hence

$$f_{n+1}(x) \leqslant \delta_{n-1}'' \int_0^x e^{-C(y)} \mathrm{d}y \int_y^D \big(-e^C f_n'\big)'(u)\mathrm{d}u \leqslant \delta_{n-1}'' f_n(x). \tag{1.7}$$

From this, one deduces that $\delta_n'' \leqslant \delta_{n-1}''$. Similarly,

$$\int_x^D f_n e^C/a \leqslant \delta_{n-1}'' \int_x^D \big(-e^C f_n'\big)' \leqslant \delta_{n-1}''\big[e^C f_n'\big](x)$$

and so $\tilde{\delta}_n'' \leqslant \delta_{n-1}''$. By using the mean value theorem again, $\delta_n'' \leqslant \tilde{\delta}_n''$.

We now consider the second part of the theorem. First, consider (2). By identity

$$[f I\!I(f)](x) = \int_0^D f\varphi(\cdot \wedge x) e^C/a = \int_0^x f\varphi e^C/a + \varphi(x)\int_x^D f e^C/a,$$

we get

$$f_2^{(x_0)}(x \wedge x_0) \geqslant \varphi(x \wedge x_0)\varphi(x_0)\int_{x_0}^D e^C/a$$

and so

$$\sup_{x\in(0,D)} \frac{f_1^{(x_0)}}{f_2^{(x_0)}}(x \wedge x_0) = \sup_{x \in (0,x_0)} \frac{f_1^{(x_0)}}{f_2^{(x_0)}}(x) \leqslant \left[\varphi(x_0)\int_{x_0}^D e^C/a\right]^{-1}.$$

This implies that $\delta_1' \geqslant \delta$ $\big[$and $\tilde{\delta}_2' \geqslant \delta$ at the same time$\big]$. Here, the reason one needs the local procedure "stopping at x_0" is the possibility of $\varphi \notin L^1(\pi)$ which then implies that $D(\varphi) = \infty$. Next, consider (2)$'$. In this case, the speed of approximating to the eigenvalue is slower than the previous one. In particular, the upper bound obtained by the first iteration is always trivial. Because, when $x \to 0$, we have

$$\big[f_1' e^C\big](x)^{-1} \int_x^D f_1 e^C/a = 2\sqrt{\varphi(x)}\int_x^D \sqrt{\varphi} e^C/a \to 0$$

and so $\delta_1' = 0$. On the other hand, when $x \to 0$,

$$\big[f_2' e^C\big](x)^{-1} \int_x^D f_2 e^C/a = \int_x^D f_2 e^C a^{-1} \Big/ \int_x^D f_1 e^C a^{-1} \to \frac{\pi(f_2)}{\pi(f_1)} > 0.$$

Thus, when $D < \infty$, we have $\delta_2' > 0$. However, when $D = \infty$, it is still possible that $\delta_2' = 0$ [Noting that the case of $D = \infty$ can be approximated arbitrarily by finite D, and it seems that δ_n' can not be all vanished].

We now prove the monotonicity of δ_n's. Applying the mean value theorem twice, we obtain

$$\begin{aligned}
\sup_{x\in(0,D)} \big[f_n^{(x_0)}/f_{n+1}^{(x_0)}\big](x\wedge x_0) &= \sup_{x\in(0,x_0)} \big[f_n^{(x_0)}/f_{n+1}^{(x_0)}\big](x)\\
&\leqslant \sup_{x\in(0,x_0)} \big[f_n^{(x_0)\prime}/f_{n+1}^{(x_0)\prime}\big](x)\\
&= \sup_{x\in(0,x_0)} \int_x^D f_{n-1}^{(x_0)}(\cdot\wedge x_0)\frac{e^C}{a} \bigg/ \int_x^D f_n^{(x_0)}(\cdot\wedge x_0)\frac{e^C}{a}\\
&\leqslant \sup_{x\in(0,D)} \big[f_{n-1}^{(x_0)}/f_n^{(x_0)}\big](x\wedge x_0).
\end{aligned}$$

This implies that ${\delta_n'}^{-1} \leqslant {\tilde\delta_n'}^{-1} \leqslant {\delta_{n-1}'}^{-1}$. The inequality ${\delta_n'}^{-1} \geqslant \xi_0' \geqslant \lambda_0$ comes from Theorem 1.1. Ignoring x_0, we obtain the monotonicity mentioned in $(2)'$. $\square$

Note that $II(f_{n-1}) = f_n/f_{n-1} \equiv \delta_{n-1}''$ implies $\lambda_0 = {\delta_{n-1}''}^{-1}$. Otherwise, it contradicts the minimal property of λ_0. On the other hand, if $II(f_{n-1}) \not\equiv \delta_{n-1}''$, then inequality (1.6) holds in some interval and so the inequality in (1.7) holds on whole $(0, D)$. Hence, one often has $\delta_n'' < \delta_{n-1}''$. But in the case that the supremum is achieved at $D = \infty$, one may still have equality here. See Examples 1.5 and 1.6. Several numerical examples (Example 1.5, for instance) show that one should have $\delta_n' - \delta_n'' < \delta_{n-1}' - \delta_{n-1}''$ whenever $II(f_{n-1})$ is not a constant. Therefore, one would have

$$\lim_{n\to\infty} \delta_n' = \lim_{n\to\infty} \delta_n'' = \lambda_0^{-1}$$

(has not proved yet). Let us make one more remark on this point. Because infinite D can be approximated arbitrarily by finite ones, we may assume that $D < \infty$. Then, the spectrum of L is discrete and so λ_0 is the ordinary eigenvalue of L. Let $\underline{\delta} := \lim_{n\to\infty} \delta_n''$ and $\lambda_0 > \underline{\delta}^{-1}$. Without changing δ_n'', at each step of iteration, one may replace f_n by $f_n/f_n(D)$. Thus, whenever $\{f_n\}_{n\geqslant 1}$ has a limit point, it should be non-degenerated. Then, by taking the limit to show that the equation $f II(f) = -\underline{\delta} f$ does have a non-degenerated solution. This shows that $\underline{\delta}$ is also an eigenvalue and it contradicts the minimal property of λ_0. Therefore $\lambda_0 = \underline{\delta}^{-1} = \lim_n {\delta_n''}^{-1}$. Applying Theorem 1.1 again, we get $\lim_{n\to\infty} \delta_n' = \lambda_0^{-1}$.

We now study λ_1. Let $\bar f = f - \pi(f)$ and set

$$\begin{aligned}
\mathscr{F}' &= \big\{f \in C[0,D] : f(0) = 0,\ \text{there exists } x_0 \in (0,D) \text{ so that}\\
&\qquad\qquad f = f(\cdot\wedge x_0),\ f \in C^1(0,x_0) \text{ and } f'|_{(0,x_0)} > 0\big\},\\
\mathscr{F}'' &= \{f \in C[0,D]\cap C^1(0,D) : f(0) = 0,\ f'|_{(0,D)} > 0\},\\
\xi_1' &= \inf_{f\in\mathscr{F}'} \sup_{x\in(0,D)} I(\bar f)(x)^{-1},\\
\xi_1'' &= \sup_{f\in\mathscr{F}''} \inf_{x\in(0,D)} I(\bar f)(x)^{-1}.
\end{aligned}$$

Theorem 1.3 Assume that (1.1) and (1.2) hold. Then $\xi_1' \geqslant \lambda_1 \geqslant \xi_1''$. If additionally a and b are continuous, then the inequalities all become equalities.

Proof First, if $f'|_{(0,x_0)} > 0$, then $\int_x^D \bar{f} e^C/a > 0$ for all $x \in (0, x_0)$. Otherwise, $\bar{f}(x) \leqslant 0$ and

$$0 = \int_0^D \bar{f} e^C/a \leqslant \int_0^x \bar{f} e^C/a < \bar{f}(x) \int_0^x e^C/a \leqslant 0,$$

which is a contradiction. Next, let $f \in \mathscr{F}'$ and $c := \sup_{x\in(0,D)} I(\bar{f})(x)^{-1} < \infty$. Set $c_0 = \sup_{x\in(0,x_0)} I(\bar{f})(x)^{-1}$. Then

$$\begin{aligned}
\int_0^D f'^2 e^C &= \int_0^{x_0} \left[I(\bar{f})(x)^{-1} \int_x^D \bar{f} e^C/a \right] \mathrm{d}\bar{f}(x) \\
&\leqslant c_0 \int_0^{x_0} \left[\int_x^D \bar{f} e^C/a \right] \mathrm{d}\bar{f}(x) \\
&= c_0 \bar{f}(x_0) \int_{x_0}^D \bar{f} e^C/a + c_0 \int_0^{x_0} \bar{f}^2 e^C/a \\
&= c_0 \int_0^D \bar{f}^2 e^C a^{-1} \\
&\leqslant c \int_0^D \bar{f}^2 e^C a^{-1}.
\end{aligned}$$

Here the positivity of $\int_x^D \bar{f} e^C/a$ is used in the inequalities. This implies that

$$\lambda_1 \leqslant \inf_{f\in\mathscr{F}'} \sup_{x\in(0,D)} I(\bar{f})(x)^{-1}.$$

The lower bound is just Theorem 2.1 (2) given in ref. [2]. $\square$

The main difference between Theorems 1.3 and 1.1 is that one can only use operator $I(\bar{f})$ but not $I\!I(f)$. This is not a technical but an essential difference between λ_1 and λ_0. The eigenfunction of λ_0 always keeps its sign but the one of λ_1 must change its sign and hence one can not use division by the eigenfunction. Based on this, the above proofs for the monotonicity of δ_n'' and δ_n' are no longer suitable for η_n'' and η_n', even though the assertion seems still to be true (cf. Example 1.7). The starting point of the next result is as follows: Because f_n's are good approximation of the eigenfunction of λ_0 and λ_0 and λ_1 are zero or not simultaneously, one expects that $f_n - \pi(f_n)$ (the eigenfunction of λ_1 should have mean zero, refer to Lemmas 2.2 and 2.3 in ref. [7]) should be also a reasonable approximation to the eigenfunction of λ_1.

Theorem 1.4 Let (1.1) and (1.2) hold. Set $\varphi(x) = \int_0^x e^{-C(y)}\mathrm{d}y$ and $\bar{f} = f - \pi(f)$.

(1) Define $f_1 = \sqrt{\varphi}$, $f_n = \bar{f}_{n-1} I\!I(\bar{f}_{n-1})$ and $\eta_n'' = \sup_{x\in(0,D)} I(\bar{f}_n)(x)$. Then $\lambda_1 \geqslant \eta_n''^{-1} \geqslant (4\delta)^{-1}$.[12]

[12] see Appendix.

(2) Fix $x_0 \in (0, D)$ and define[13]

$$\begin{aligned} f_1^{(x_0)} &= \varphi(\cdot \wedge x_0), \\ f_n^{(x_0)} &= \bar{f}_{n-1}^{(x_0)}(\cdot \wedge x_0) I\!I\big(\bar{f}_{n-1}^{(x_0)}(\cdot \wedge x_0)\big), \\ \eta_n' &= \sup_{x_0 \in (0,D)} \inf_{x \in (0,D)} I\big(\bar{f}_n^{(x_0)}(\cdot \wedge x_0)\big)(x). \end{aligned}$$

Then ${\eta_n'}^{-1} \geqslant \lambda_1 (n \geqslant 2)$. By convention, $1/0 = \infty$.

Proof By Theorem 1.3, Proposition 3.2 in ref. [1] and Theorem 1.1, we have

$${\eta_n'}^{-1} \geqslant \lambda_1 \geqslant \lambda_0 \vee {\eta_n''}^{-1} \geqslant \lambda_0 \wedge {\eta_n''}^{-1} \geqslant (4\delta)^{-1}.$$

Note that when $x \to 0$,

$$I\big(\bar{f}_1^{(x_0)}\big)(x) = \int_x^D \bar{f}_1^{(x_0)} e^C a^{-1} \big/ \big[\bar{f}_1^{(x_0)\prime} e^C\big](x) \to 0$$

and hence $\eta_1' = 0$.[14] The remainder assertions can be deduced from Theorem 1.3. □

Example 1.5 Consider interval $[0, 1]$, $b(x) \equiv 0$ and $a(x) \equiv 1$. Then $(\delta, \delta_1', \delta_2') \approx (0.25, 0.375, 0.4005)$. But 4δ and $\delta_1'', \cdots, \delta_5''$ are 1, 0.4275, 0.4074, 0.4056, 0.405322, 0.405289 successively. The precise solution is $4/\pi^2 \approx 0.405285$ which can be obtained by using the test function $\sin(x\pi/2)$. Even though this function is rather different to the initial one $f_1(x) = \sqrt{x}$, it is very close to the resulting function $f_3(x)/f_3(1)$ in two iterations. This suggests us to use a simpler function instead of f_n to simplify the computation and fast the convergence speed. In particular, one may use some approximating function of $f_2/f_2(D)$ instead of f_1 to avoid the computation on multi-integrals. Moreover, the identity

$$\begin{aligned} f_n(x) &= [f_{n-1} I\!I(f_{n-1})](x) \\ &= \int_0^D f_{n-1} \varphi(\cdot \wedge x) \frac{e^C}{a} \\ &= \int_0^x f_{n-1} \varphi \frac{e^C}{a} + \varphi(x) \int_x^D f_{n-1} \frac{e^C}{a} \end{aligned} \tag{1.8}$$

is also helpful to decrease the computations. This example shows that ${\delta_n'}^{-1}$ and ${\delta_n''}^{-1}$ all approximate to λ_0.

---

[13]There is a similar question as in part (2) of Theorem 1.2 since the sequence $\{f_n^{(x_0)}\}_{n \geqslant 1}$ is usually not contained in $\mathscr{F}'$ and may have the integrability problem. It is more natural to use

$$f_n^{(x_0)} = \big[\bar{f}_{n-1}^{(x_0)} I\!I\big(\bar{f}_{n-1}^{(x_0)}\big)\big](\cdot \wedge x_0), \quad n \geqslant 2$$

instead of the original one.

This remark is also meaningful in the discrete case (part (2) of Theorem 2.4).

[14]see Appendix.

Example 1.6 Take $b(x) \equiv 0$. From ref. [1], we know that $\delta^{-1} \geqslant \lambda_1 \geqslant \lambda_0 \geqslant (4\delta)^{-1}$. In particular, when $a(x) = (1+x)^2$, we have

$$\delta = 1, \qquad \lambda_1 = \lambda_0 = 1/4.$$

Hence the lower bound is sharp for this example.[15]

Example 1.7 This is a continuation of Example 1.5 but considering λ_1. Applying the procedure given in Theorem 1.4 (1), in the first 4 steps of the iterations, the (maximum, minimum) of $I(\bar{f}_n)$ are the following:

$$(0.1406, 0), \quad (0.1078, 0.0525), \quad (0.103166, 0.09573), \quad (0.101961, 0.100295).$$

The precise solution is

$$\lambda_1^{-1} = \pi^{-2} \approx 0.101321$$

which corresponds to the eigenfunction g: $g(x) = \cos(\pi x)$ with $\pi(g) = 0$. Noticing that $\min_x I(\bar{f}_n)$ with $f_1 = \sqrt{\varphi}$ (corresponding to part $(2)'$ of Theorem 1.1) is usually less powerful than ${\eta_n'}^{-1}$,[16] it follows that ${\eta_n'}^{-1}$ and ${\eta_n''}^{-1}$ all approximate to λ_1.[17]

2 Discrete case

Consider birth-death processes. Let $b_i > 0 (i \geqslant 0)$ and $a_i > 0 (i \geqslant 1)$ be the birth and death rates respectively. Define

$$\pi_i = \frac{\mu_i}{\mu}, \qquad \mu_0 = 1, \ \mu_i = \frac{b_0 \cdots b_{i-1}}{a_1 \cdots a_i}.$$

Throughout this section, assume that $\mu = \sum_i \mu_i < \infty$. The process is non-explosive iff

$$\sum_{k=0}^{\infty} (b_k \mu_k)^{-1} \sum_{i=0}^{k} \mu_i = \infty. \tag{2.1}$$

The corresponding Dirichlet form is

$$D(f) = \sum_i \pi_i b_i [f_{i+1} - f_i]^2, \qquad \mathscr{D}(D) = \{f \in L^2(\pi) : D(f) < \infty\}.$$

[15] We have $\bar{\eta}_1 = \eta_1^* = 2$ (in both cases, $x_0 = \infty$, ε_1 and ε_2 play no role), $\eta_1'' = 4$. Hence the first lower estimate is exact and furthermore $\eta_1''/\bar{\eta}_1 = \eta_1''/\eta_1^* = 2$.

[16] If we use the procedure given in Theorem 1.4 (2) with fixed $x_0 = 1$, then the first four upper bonds are 0, 0.08333, 0.1, 0.10119 which are clearly better than the ones just mentioned. Using the procedure of Theorem 1.4 (2), the computation becomes more complicated since it contains a maximization with respect to x_0. However the result is the same since the maximum achieves at $x_0 = 1$ for this example.

[17] see Appendix.

Let

$$I_i(f) = \frac{1}{\mu_i b_i (f_{i+1} - f_i)} \sum_{j \geqslant i+1} \mu_j f_j, \qquad II_i(f) = \frac{1}{f_i} \sum_{j \leqslant i-1} \frac{1}{\mu_j b_j} \sum_{k \geqslant j+1} \mu_k f_k,$$

$$\mathscr{F}' = \{f : f_0 = 0, \text{ there exists } k : 1 \leqslant k < \infty \text{ such that } f_i = f_{i\wedge k} \text{ and } f > 0 \text{ on } (0,k]\},$$

$$\mathscr{F}'' = \{f : f_0 = 0,\ f_i > 0,\ \forall i \geqslant 1\},$$

$$\xi_0' = \inf_{f \in \mathscr{F}'} \sup_{i \geqslant 1} II_i(f)^{-1}, \qquad \xi_0'' = \sup_{f \in \mathscr{F}''} \inf_{i \geqslant 1} II_i(f)^{-1},$$

$$\widetilde{\mathscr{F}}' = \{f : f_0 = 0, \text{ there exists } k : 1 \leqslant k < \infty \text{ such that } f_i = f_{i\wedge k} \text{ and } f \text{ is strictly increasing in } [0,k]\},$$

$$\widetilde{\mathscr{F}}'' = \{f : f_0 = 0,\ f \text{ is strictly increasing}\},$$

$$\tilde{\xi}_0' = \inf_{f \in \widetilde{\mathscr{F}}'} \sup_{i \geqslant 0} I_i(f)^{-1}, \qquad \tilde{\xi}_0'' = \sup_{f \in \widetilde{\mathscr{F}}''} \inf_{i \geqslant 0} I_i(f)^{-1}.$$

Theorem 2.1 $\tilde{\xi}_0' = \xi_0' = \lambda_0 = \xi_0'' = \tilde{\xi}_0''$.

Proof Let g satisfy $g_0 = 0$ and $\|g\| = 1$. Next, let (h_i) be a positive sequence. Then

$$\begin{aligned}
1 = \sum_i g_i^2 \pi_i &= \sum_i \pi_i (g_i - g_0)^2 = \sum_i \pi_i \bigg(\sum_{j \leqslant i-1} (g_{j+1} - g_j) \bigg)^2 \\
&\leqslant \sum_i \pi_i \sum_{j \leqslant i-1} \frac{(g_{j+1} - g_j)^2 \pi_j b_j}{h_j} \sum_{k \leqslant i-1} \frac{h_k}{\pi_k b_k} \\
&= \sum_j \pi_j b_j (g_{j+1} - g_j)^2 \frac{1}{h_j} \sum_{i \geqslant j+1} \pi_i \sum_{k \leqslant i-1} \frac{h_k}{\pi_k b_k} \\
&\leqslant D(g) \sup_{j \geqslant 0} \frac{1}{h_j} \sum_{i \geqslant j+1} \pi_i \sum_{k \leqslant i-1} \frac{h_k}{\pi_k b_k} \\
&=: D(g) \sup_{j \geqslant 0} H_j.
\end{aligned}$$

Let $f \in \mathscr{F}''$ satisfy $\sup_{j \geqslant 1} II_j(f) < \infty$. Instead of the mean value theorem, we adopt some elementary proportion property. Take $h_i = \sum_{j \geqslant i+1} \mu_j f_j$. Then

$$\sup_{j \geqslant 0} H_j \leqslant \sup_{j \geqslant 0} \bigg[-\frac{\pi_{j+1}}{h_{j+1} - h_j} \sum_{k \leqslant j} \frac{h_k}{\pi_k b_k} \bigg] = \sup_{j \geqslant 0} \frac{1}{f_{j+1}} \sum_{k \leqslant j} \frac{h_k}{\mu_k b_k} = \sup_{j \geqslant 1} II_j(f).$$

Combining these two facts, we obtain $\lambda_0 \geqslant \xi_0''$.

Let $f \in \mathscr{F}''$ satisfy $\sup_{j \geqslant 1} II_j(f) < \infty$. Set

$$g_i = \sum_{j \leqslant i-1} \frac{1}{\mu_j b_j} \sum_{k \geqslant j+1} \mu_k f_k.$$

Then

$$g_{i+1} - g_i = \frac{1}{\mu_i b_i} \sum_{k \geqslant i+1} \mu_k f_k, \qquad i \geqslant 0 \tag{2.2}$$

and furthermore $g \in \widetilde{\mathscr{F}}''$. Hence $\Omega g(i) = b_i(g_{i+1} - g_i) + a_i(g_{i-1} - g_i) = -f_i$ for $i \geqslant 1$. Or,

$$\pi_i b_i (g_{i+1} - g_i) - \pi_{i-1} b_{i-1}(g_i - g_{i-1}) = -\pi_i f_i, \qquad i \geqslant 1. \tag{2.3}$$

The right-hand side is controlled from above by $-\pi_i g_i \big(\sup_{i \geqslant 1} II_i(f)\big)^{-1}$. Summing up in i from $k+1$ to ∞ gives

$$\sum_{j \geqslant k+1} \pi_j g_j \leqslant \pi_k b_k (g_{k+1} - g_k) \sup_{i \geqslant 1} II_i(f), \qquad k \geqslant 0. \tag{2.4}$$

Hence $\sup_{k \geqslant 0} I_k(g) \leqslant \sup_{k \geqslant 1} II_k(f)$. This implies that

$$\inf_{g \in \widetilde{\mathscr{F}}''} \sup_{k \geqslant 0} I_k(g) \leqslant \sup_{k \geqslant 1} II_k(f)$$

and furthermore

$$\inf_{g \in \widetilde{\mathscr{F}}''} \sup_{k \geqslant 0} I_k(g) \leqslant \inf_{f \in \mathscr{F}''} \sup_{k \geqslant 1} II_k(f)$$

since f is arbitrary. The inverse inequality follows immediately from the proportional property. This proves that $\xi_0'' = \tilde{\xi}_0''$. Dually, $\xi_0' = \tilde{\xi}_0'$.

Let $f \in \mathscr{F}'$ satisfy $f_i = f_{i \wedge k}$ and set $g = [f II(f)](\cdot \wedge k)$. Then by (2.2) we have

$$\begin{aligned}
D(g) &= \sum_{0 \leqslant i \leqslant k-1} \pi_i b_i (g_{i+1} - g_i)^2 \\
&= \sum_{0 \leqslant i \leqslant k-1} (g_{i+1} - g_i) \sum_{j \geqslant i+1} \pi_j f_j \\
&= \sum_{j \geqslant 1} \pi_j f_j \sum_{i \leqslant (k-1) \wedge (j-1)} (g_{i+1} - g_i) \\
&= \sum_{i \geqslant 1} \pi_i f_i g_{k \wedge i} \\
&\leqslant \|g\|^2 \sup_{i \geqslant 1} II_i(f)^{-1}.
\end{aligned}$$

This gives $\lambda_0 \leqslant \xi_0'$. It was proved in ref. [5] that $\lambda_0 \geqslant \tilde{\xi}_0''$ and moreover the equality must hold. Dually, we have $\lambda_0 = \tilde{\xi}_0'$.

Theorem 2.2 Let $\varphi_i = \sum_{j \leqslant i-1} (\mu_j b_j)^{-1}$.

(1) Define $f_1 = \sqrt{\varphi}$, $f_n = f_{n-1} II(f_{n-1})$ and $\delta_n'' = \sup_{i \geqslant 1} II_i(f_n)$. Then δ_n'' is decreasing in n and

$$\lambda_0 \geqslant \lim_{n \to \infty} \delta_n''^{-1} \geqslant \delta_1''^{-1} \geqslant (4\delta)^{-1},$$

where $\delta = \sup_{i \geqslant 1} \sum_{j \leqslant i-1} (\mu_j b_j)^{-1} \sum_{j \geqslant i} \mu_j$.

(2) For fixed $k \geqslant 1$ define

$$\begin{aligned} f_1^{(k)} &= \varphi(\cdot \wedge k), \\ f_n^{(k)} &= f_{n-1}^{(k)}(\cdot \wedge k)\, I\!I\big(f_{n-1}^{(k)}(\cdot \wedge k)\big), \end{aligned}$$

and then define $\delta_n' = \sup_{k \geqslant 1} \inf_{i \geqslant 1} I\!I_i\big(f_n^{(k)}(\cdot \wedge k)\big)$. Then δ_n' is increasing in n and[18]

$$\delta^{-1} \geqslant {\delta_1'}^{-1} \geqslant \lim_{n \to \infty} {\delta_n'}^{-1} \geqslant \lambda_0.$$

In definitions of δ_n' and $\delta_n''(n \geqslant 1)$, replacing $I\!I(f)$ by $I(f)$ everywhere, one obtains $\tilde{\delta}_n'$ and $\tilde{\delta}_n''$. Then $\delta_n'' \leqslant \tilde{\delta}_n'', \delta_n' \geqslant \tilde{\delta}_n'(n \geqslant 1)$ and $\tilde{\delta}_n'' \leqslant \delta_{n-1}'', \tilde{\delta}_n' \geqslant \delta_{n-1}'(n \geqslant 2)$. In particular, the modified assertions (1) and (2) still hold, the only change is replacing ${\delta_1'}^{-1}$ at the end of (2) by $\tilde{\delta}_2'^{-1}$.

Proof Recall that $f_{n+1}(i) = \sum_{j \leqslant i-1} (\mu_j b_j)^{-1} \sum_{k \geqslant j+1} \mu_k f_n(k)$. Applying (2.4) to $g = f_n$ and $f = f_{n-1}$, we get

$$f_{n+1}(i) \leqslant \delta_{n-1}'' \sum_{j \leqslant i-1} \big[f_n(j+1) - f_n(j)\big] = \delta_{n-1}'' f_n(i), \qquad i \geqslant 1.$$

This proves that $\delta_n'' \leqslant \delta_{n-1}''$. In the same way, we obtain $\tilde{\delta}_n'' \leqslant \delta_{n-1}''$.

Next, by (2.2) and the proportional property, it follows that

$$\begin{aligned} \sup_{i \geqslant 1} \big[f_n^{(k)}/f_{n+1}^{(k)}\big](i \wedge k) &= \sup_{1 \leqslant i \leqslant k} \big[f_n^{(k)}/f_{n+1}^{(k)}\big](i) \\ &\leqslant \sup_{1 \leqslant i \leqslant k} \sum_{j \geqslant i+1} \mu_j f_{n-1}^{(k)}(j \wedge k) \Big/ \sum_{j \geqslant i+1} \mu_j f_n^{(k)}(j \wedge k) \\ &\leqslant \sup_{i \geqslant 1} \sum_{j \geqslant i+1} \mu_j f_{n-1}^{(k)}(j \wedge k) \Big/ \sum_{j \geqslant i+1} \mu_j f_n^{(k)}(j \wedge k) \\ &\quad \Big[= \sup_{i \geqslant 1} I_i\big(f_n^{(k)}\big)^{-1}\Big] \\ &\leqslant \sup_{i \geqslant 1} \big[f_{n-1}^{(k)}/f_n^{(k)}\big](i \wedge k). \end{aligned}$$

[18] It is proved in the paper "Computable bounds for the decay parameter of a birth-death process" by D. Sirl, H. Zhang, and P. Pollett [J. Appl. Prob. 44(2): 476-491, 2007] that δ_1' here coincides with δ' introduced in [1; Theorem 3.5]. The proof goes as follows. First show that

$$\frac{1}{\varphi_{i \wedge k}} \sum_{j=1}^{i} \frac{1}{\mu_j a_j} \sum_{m=j}^{\infty} \mu_m \varphi_{m \wedge k}$$

achieves its minimum at $i = k$. Then, by exchanging the order of the summation, it follows that the minimum is equal to

$$\frac{1}{\varphi_k} \sum_{m=1}^{\infty} \mu_m \varphi_{m \wedge k}^2.$$

Finally, the proof of [1; Theorem 3.5] gives us the required assertion.

This implies that ${\delta'_n}^{-1} \leqslant \tilde{\delta}_n'^{-1} \leqslant {\delta'_{n-1}}^{-1}$. The proofs for the remainder assertions are similar to the ones of Theorem 1.2. $\square$

We now study λ_1. Write $\bar{f} = f - \pi(f)$ and define

$$\begin{aligned}
\mathscr{F}' &= \{f : f_0 = 0,\ \text{there exists } k : 1 \leqslant k < \infty \text{ so that } f_i = f_{i\wedge k} \\
&\qquad\qquad \text{and } f \text{ is strictly increasing in } [0,k]\},\\
\mathscr{F}'' &= \{f : f_0 = 0,\ f \text{ is strictly increasing}\},\\
\xi_1' &= \inf_{f\in\mathscr{F}'} \sup_{i\geqslant 0} I_i(\bar{f})^{-1}, \qquad \xi_1'' = \sup_{f\in\mathscr{F}''} \inf_{i\geqslant 0} I_i(\bar{f})^{-1}.
\end{aligned}$$

Theorem 2.3 Assume that (2.1) holds. Then $\xi_1' = \lambda_1 = \xi_1''$.

Proof Let $f \in \mathscr{F}'$, then $f_0 = 0$, $f_i = f_{i\wedge k}$, and

$$\sum_{j=1}^{\infty} \pi_j \bar{f}_j = -\pi_0 \bar{f}_0. \tag{2.5}$$

Write $c_0 = \sup_{0\leqslant i\leqslant k-1} I_i(\bar{f})^{-1}$ and $c = \sup_{i\geqslant 0} I_i(\bar{f})^{-1}$. Then

$$\begin{aligned}
D(f) &= \sum_{i=0}^{k-1} \pi_i b_i [f_{i+1} - f_i]^2 \\
&\leqslant c_0 \sum_{i=0}^{k-1} \bigg[\sum_{j\geqslant i+1} \pi_j \bar{f}_j\bigg][\bar{f}_{i+1} - \bar{f}_i] \\
&= c_0 \sum_{j=1}^{\infty} \pi_j \bar{f}_j \sum_{i=0}^{(j-1)\wedge(k-1)} [\bar{f}_{i+1} - \bar{f}_i] \\
&= c_0 \sum_{j=1}^{\infty} \pi_j \bar{f}_j [\bar{f}_{j\wedge k} - \bar{f}_0] \\
&= c_0 \sum_{j=0}^{\infty} \pi_j \bar{f}_j^2 \quad (\text{by } (2.5)) \\
&\leqslant c \mathrm{Var}(f).
\end{aligned}$$

Hence $\lambda_1 \leqslant c$ and furthermore $\lambda_1 \leqslant \xi_1'$. It was proved in ref. [4], Theorem 1.1 that $\lambda_1 = \xi_1''$ and moreover the equality $\lambda_1 = \xi_1'$ holds. $\square$

Now, the next result follows directly from Theorem 2.3.

Theorem 2.4 Assume that (2.1) holds. Write $\varphi_i = \sum_{j\leqslant i-1} (\mu_j b_j)^{-1}$.

(1) Define $f_1 = \sqrt{\varphi}$, $f_n = \bar{f}_{n-1} I\!I(\bar{f}_{n-1})$ and $\eta_n'' = \sup_{i\geqslant 0} I_i(\bar{f}_n)$. Then $\lambda_1 \geqslant {\eta_n''}^{-1} \geqslant (4\delta)^{-1}$.[19]

[19] Here is the explicit expression of η_1'':

$$\eta_1'' = \sup_{i\geqslant 0} \frac{1}{\mu_i b_i \big(\sqrt{\varphi_{i+1}} - \sqrt{\varphi_i}\big)} \bigg[\sum_{j\geqslant i+1} \mu_j \sqrt{\varphi_j} - \frac{1}{\mu}\bigg(\sum_{j\geqslant i+1} \mu_j\bigg)\bigg(\sum_{i\geqslant 0} \mu_i \sqrt{\varphi_i}\bigg)\bigg].$$

(2) Fix $k \geqslant 1$ and define

$$\begin{aligned}
f_1^{(k)} &= \varphi(\cdot \wedge k),\\
f_n^{(k)} &= \bar{f}_{n-1}^{(k)}(\cdot \wedge k)\, II_{\cdot\wedge k}\big(\bar{f}_{n-1}^{(k)}(\cdot \wedge k)\big),\\
\eta_n' &= \sup_{k\geqslant 1}\inf_{i\geqslant 0} I_i\big(\bar{f}_n^{(k)}(\cdot \wedge k)\big).
\end{aligned}$$

Then ${\eta_n'}^{-1} \geqslant \lambda_1 (n \geqslant 2)$. By convention, $1/0 = \infty$.[20]

Acknowledgements This work was supported in part by National Natural Science Foundation of China (No. 19631060), Mathematical Tian Yuan Foundation, Qiu Shi Science & Technology Foundation, RFDP and MCEC. The author acknowledges Profs. Yuan, Rong and Zhang, Yu-Hui for their help.

References

[1]. Chen M F, *Explicit bounds of the first eigenvalue*, Sci. in China, Ser. A, 2000, 43(10), 1051–1059.

[2]. Chen M F Wang F Y, *Estimation of spectral gap for elliptic operators*, Trans Amer Math Soc 1997, 349(3), 1239-1267.

[3]. Chen M F Wang F Y, *General formula for lower bound of the first eigenvalue on Riemannian manifolds*, Sci. in China, Ser. A, 1997, 40(4), 384–394.

[4]. Chen M F, *Estimation of spectral gap for Markov chains*, Acta Math Sin New Series 1996, 12(4), 337-360.

[5]. Chen M F, *Analytic proof of dual variational formula for the first eigenvalue in dimension one*, Sci. in China, Ser. (A) 1999, 42(8), 805–815..

[6]. Chen M F, *Coupling, spectral gap and related topics* (I), Chin. Sci. Bulletin, 1997, 42(16), 1321–1327.

[7]. Chen M F, *Estimate of exponential convergence rate in total variation by spectral gap*, Acta Math. Sin. Ser. (A) 1998, 41(1), 1–6; Acta Math. Sin. New Ser. 1998, 14(1), 9–16..

[20] see Appendix.

Appendix: Some footnotes.

Footnote 4. One may argue about the sequence $\{f_n^{(x_0)}\}_{n\geqslant 1}$ since it is usually not contained in $\mathscr{F}'$. However, the modified sequence:

$$\tilde f_1^{(x_0)} = \varphi(\cdot \wedge x_0), \qquad \tilde f_n^{(x_0)} = \big[\tilde f_{n-1}^{(x_0)} I\!I\big(\tilde f_{n-1}^{(x_0)}\big)\big](\cdot \wedge x_0), \quad n \geqslant 2$$

is clearly contained in $\mathscr{F}'$. Moreover,

$$\begin{aligned}\delta_n' &= \sup_{x_0\in(0,D)} \inf_{x\in(0,D)} I\!I\big(f_n^{(x_0)}(\cdot\wedge x_0)\big)(x)\\ &= \sup_{x_0\in(0,D)} \inf_{x\in(0,x_0)} I\!I\big(f_n^{(x_0)}(\cdot\wedge x_0)\big)(x)\\ &= \sup_{x_0\in(0,D)} \inf_{x\in(0,x_0)} I\!I\big(\tilde f_n^{(x_0)}\big)(x)\\ &= \sup_{x_0\in(0,D)} \inf_{x\in(0,D)} I\!I\big(\tilde f_n^{(x_0)}\big)(x).\end{aligned}$$

Here in the last step we have used the convention $1/0=\infty$. Hence these two sequences produce the same result.

This comment is also meaningful in the discrete case (part (2) of Theorem 2.2).

Footnote 6. The last footnote suggests us to use directly $\bar\delta_n^{-1}$:

$$\bar\delta_n = \sup_{x\in(0,D)} \frac{\big\|f_n^{(x)}\big\|^2}{D\big(f_n^{(x)}\big)} \quad (n\geqslant 1) \tag{A.1}$$

as an upper bound of λ_0, where $f_n^{(x)} = \big[f_{n-1}^{(x)} I\!I\big(f_{n-1}^{(x)}\big)\big](\cdot\wedge x)$. The computation of $\bar\delta_n$ is easier than δ_n' since the minimizing procedure with respect to $x\in(0,x_0)$ is avoided. For Example 1.5, we have $\bar\delta_1=\delta_1'=0.375$, $\delta_2'\approx 0.400509$, $\bar\delta_2\approx 0.404762$, $\delta_3'\approx 0.404762$, $\bar\delta_3\approx 0.405279$. It seems that in general one has $\bar\delta_n \geqslant \delta_n'$ for all $n\geqslant 1$ but we are unable to prove it. What we have is $\bar\delta_n\geqslant\tilde\delta_n'$ for all $n\geqslant 1$, where $\tilde\delta_n'$ is defined in the last paragraph of Theorem 1.2. This is a consequence of the second assertion of Lemma A.2 below. Furthermore, the first one $\delta_1'=\delta'$ can be easily improved by δ_1^*:

$$\delta_1^* = \sup_{x\in(0,D)} \sup_{0\leqslant\varepsilon\leqslant x} \frac{\big\|f_1^{(x,\varepsilon)}\big\|^2}{D\big(f_1^{(x,\varepsilon)}\big)}, \tag{A.2}$$

where

$$f_1^{(x,\varepsilon)}(y) = \int_0^{y\wedge x} \gamma e^{-C}, \qquad \gamma(z) = \left(\frac{1}{\varepsilon}(x-z)^+\right)\wedge 1.$$

For instance, for Example 1.5, we have $\delta_1'=0.375$, $\delta_1^*\approx 0.4045>\delta_2'$; for Example 1.6 with Dirichlet condition at 0, we have $\lambda_0^{-1}=4$, $\delta_1'=2$, and $\delta_1^*=2.5$; and for the standard O.U.-process on $(0,\infty)$ with Dirichlet condition at 0, we have $\lambda_0=1$, $\delta_1'\approx 0.7973$, and $\delta_1^*\approx 0.97$.

This remark is also meaningful in the discrete case (part (2) of Theorem 2.2), but we will not write down again.

Footnote 12. Very often, we have $\eta_n''^{-1} > (4\delta)^{-1}$. Here is the explicit formula of η_1'':

$$\eta_1'' = 2 \sup_{x\in(0,D)} \sqrt{\varphi(x)}\bigg[\int_x^D \frac{\sqrt{\varphi}\, e^C}{a} - \frac{1}{Z}\int_x^D \frac{e^C}{a}\int_0^D \frac{\sqrt{\varphi}\, e^C}{a}\bigg].$$

One can improve η_1'' by modifying f_1 as follows. When $D < \infty$, define

$$f_1^{\varepsilon_1,\varepsilon_2}(x) = \sqrt{\int_0^x \gamma_1\gamma_2 e^{-C}},$$

where

$$\gamma_1(z) = \frac{z^2}{\varepsilon_1^2}\wedge 1, \qquad \gamma_2(z) = \bigg(\frac{1}{\varepsilon_2}(x_0 - z)^+\bigg)\wedge 1, \quad \varepsilon_1, \varepsilon_2 \geqslant 0,\ \varepsilon_1 + \varepsilon_2 \leqslant D.$$

When $D = \infty$, one simply set $\varepsilon_2 = 0$, ignoring the factor γ_2. Then, it is obvious that $\eta_1'' \geqslant \eta_{1*}$:

$$\eta_{1*} = \inf_{\varepsilon_1,\,\varepsilon_2\geqslant 0,\,\varepsilon_1+\varepsilon_2\leqslant D}\ \sup_{x\in(0,D)} I\big(\bar f_1^{\varepsilon_1,\varepsilon_2}\big). \tag{A.3}$$

For instance, for Example 1.7, we have $\eta_1'' \approx 0.1406$ but $\eta_{1*} \approx 0.1072$ (with $\varepsilon_1 = 0.5461$, $\varepsilon_2 = 0.207$), and the exact value is 0.101321. For the standard O.U.-process on $(0,\infty)$ with reflecting boundary at 0, $\eta_1'' \approx 0.7$ but $\eta_{1*} \approx 0.5785$ (with $\varepsilon_1 = 1.445435$), and the exact value is 1/2.

Footnote 14. (a) This is the reason to start at η_2' in the theorem. However, the test function $f_1^{(x_0)} = \varphi(\cdot \wedge x_0)$ already provides us a meaningful upper estimate: $\lambda_1 \leqslant \bar\eta_1^{-1}$, where

$$\begin{aligned}\bar\eta_1 &= \sup_{x_0\in(0,D)} \frac{\big\|\bar f_1^{(x_0)}\big\|^2}{D\big(f_1^{(x_0)}\big)}\\ &= \sup_{x\in(0,D)} \frac{1}{\varphi(x)}\bigg\{2\int_0^x e^{-C}\varphi\psi - \frac{1}{\psi(0)}\bigg[\int_0^x e^{-C}\psi\bigg]^2\bigg\}\end{aligned} \tag{A.4}$$

and

$$\varphi(x) = \int_0^x e^{-C}, \qquad \psi(x) = \int_x^D \frac{e^C}{a}.$$

To see the expression of $\bar\eta_1$ given above, write $\big\|\bar f_1^{(x_0)}\big\|^2 = \big\|f_1^{(x_0)}\big\|^2 - \pi\big(f_1^{(x_0)}\big)^2$. Then the first term on the right-hand side comes from $\big\|f_1^{(x_0)}\big\|^2/D\big(f_1^{(x_0)}\big)$ and the computation given in the proof (c) of [1; Theorem 1.1]. The second term is simply an interchange of the order of the integrals

$$\int_0^D\bigg(\int_0^{x_0\wedge y} e^{-C}\bigg)\frac{e^C}{a}(y)\mathrm{d}y = \int_0^{x_0} e^{-C(z)}\mathrm{d}z\int_z^D \frac{e^C}{a}$$

(see also (A.10) below).

The above discussion suggests that in practice, one may use the sequence

$$\bar\eta_n = \sup_{x_0\in(0,D)} \frac{\big\|\bar f_n^{(x_0)}\big\|^2}{D\big(f_n^{(x_0)}\big)}, \qquad n\geqslant 1 \tag{A.5}$$

instead of η_n' for the upper estimate of λ_1, where $f_n^{(x_0)} = \big[f_{n-1}^{(x_0)} I\!I\big(f_{n-1}^{(x_0)}\big)\big](\cdot\wedge x_0)$. Actually, since "inf" is rather sensitive, it is not surprising that $\bar\eta_n \geqslant \eta_n'$ for all $n\geqslant 1$, as a consequence of the next result.

Lemma A.1. For every $f\in\widetilde{\mathscr F}'$, we have

$$\frac{\|\bar f\|^2}{D(f)} \geqslant \inf_{x\in(0,D)} I(\bar f)(x).$$

Similarly,

$$\frac{\|f\|^2}{D(f)} \geqslant \inf_{x\in(0,D)} I(f)(x).$$

Proof. Let $\gamma = \inf_{x\in(0,D)} I(\bar f)(x)$. Then we have

$$-\int_0^x \bar f e^C/a \geqslant \gamma f'(x) e^{C(x)}, \qquad x\geqslant 0.$$

Since f' is positive on an interval $(0,x_0)$ (here we assume that f is stopped at x_0), we can multiply f' in the both sides to obtain

$$-f'(x)\int_0^x \bar f e^C/a \geqslant \gamma f'(x)^2 e^{C(x)}, \qquad x\in(0,x_0).$$

By making an integration, we get

$$-\int_0^{x_0} f'(x)\mathrm{d}x\int_0^x \bar f e^C/a \geqslant \gamma\int_0^{x_0} f'^2 e^C.$$

By using the integration by parts formula and noting that $\int_0^D \bar f e^C/a = 0$, and f is a constant on (x_0,D), it follows that the left-hand side is equal to

$$\begin{aligned}
-\int_0^{x_0}\left[\int_0^x \bar f e^C/a\right]\mathrm{d}\bar f &= -\bar f(x_0)\int_0^{x_0}\bar f e^C/a + \int_0^{x_0}\bar f^2 e^C/a\\
&= \bar f(x_0)\int_{x_0}^D \bar f e^C/a + \int_0^{x_0}\bar f^2 e^C/a\\
&= \int_0^D \bar f^2 e^C/a.
\end{aligned}$$

Collecting the last two results together, we obtain the required assertion.

For the second assertion, we start at

$$\int_x^D fe^C/a \geqslant \gamma f'(x)e^{C(x)}, \qquad x \geqslant 0.$$

Then

$$\int_0^{x_0} f'(x)\mathrm{d}x \int_x^D fe^C/a \geqslant \gamma \int_0^{x_0} f'^2 e^C.$$

Applying the integration by parts formula to the left-hand side, we obtain the required assertion. □

The effectiveness of $\{\bar{\eta}_n\}$ is shown in the footnotes of Examples 1.6 and 1.7 below.

(b) An alternative modification to use η_1' is simply replace the original

$$f_1^{(x_0)}(x) = \int_0^{x\wedge x_0} e^{-C(y)}\mathrm{d}y, \qquad x \geqslant 0$$

with

$$f_1^{(x_0)}(x) = \int_0^{x\wedge x_0} y(x_0 - y)e^{-C(y)}\mathrm{d}y, \qquad x \geqslant 0.$$

Then the resulting new sequence $\{\eta_n'\}_{n\geqslant 1}$ provide us rather effective upper estimates. Certainly, by using this test function, one gets a new $\bar{\eta}_1$, defined by (A.5). The new one is often better than the original one (as checked by Example 1.7 and the standard Gaussian case) but may be worse, for instance, the new bound for Example 1.6 is 3/2 but the original $\bar{\eta}_1$ is 2. The idea of this modification is to make the derivative of the test f' to be zero at 0 and x_0, which is the main property of the eigenfunction on the interval $(0, x)$ with reflecting boundary. This is the reason why the factor $y(x_0 - y)$ appeared in the last formula. Based on this, a further improvement goes as follows. Define

$$\gamma_1(z) = \frac{z}{\varepsilon_1} \wedge 1, \qquad \gamma_2(z) = \left(\frac{1}{\varepsilon_2}(x_0 - z)^+\right) \wedge 1, \quad \varepsilon_1, \varepsilon_2 \geqslant 0, \varepsilon_1 + \varepsilon_2 \leqslant x_0$$

and redefine our first test function as

$$f_1^{(x_0,\varepsilon_1,\varepsilon_2)}(x) = \int_0^{x\wedge x_0} \gamma_1\gamma_2 e^{-C}.$$

When $\varepsilon_j = 0$, the factor γ_j $(j = 1, 2)$ is ignored. Then the resulting new η_1' works very well, even though the computation becomes little more complicated, but is often easier than computing the original η_2'.

We now define

$$\eta_1^* = \sup_{x_0\in(0,D)} \sup_{\varepsilon_1,\varepsilon_2\geqslant 0,\, \varepsilon_1+\varepsilon_2\leqslant x_0} \frac{\big\|\bar{f}_1^{(x_0,\varepsilon_1,\varepsilon_2)}\big\|^2}{D\big(f_1^{(x_0,\varepsilon_1,\varepsilon_2)}\big)}. \tag{A.6}$$

Clearly, we have $\bar\eta_1 \leqslant \eta_1^*$. In practice, we suggest to use ${\eta_1^*}^{-1}$ or more simply $\bar\eta_1^{-1}$ as an upper bound of λ_1.

(c) Finally, we make some remarks about the computation of η_n' and $\bar\eta_n$.

First, we compute $\inf_{y\in(0,x)} I\big(\bar f\big)$ more explicitly for general f with $f(0)=0$ and stopping at x: $f(y)=f(y\wedge x)$. To do so, set $\psi(y)=\int_y^D e^C/a$ and use the integration by parts formula

$$\int_p^q he^C/a = -[h\psi]|_p^q + \int_p^q h'\psi.$$

Then

$$\begin{aligned}\int_0^y \bar f e^C/a &= -\bar f(y)\psi(y) + \bar f(0)\psi(0) + \int_0^y f'\psi\\ &= -f(y)\psi(y) + \pi(f)(\psi(y)-\psi(0)) + \int_0^y f'\psi, \qquad y\leqslant x.\end{aligned}$$

Next,

$$\pi(f) = \psi(0)^{-1}\bigg[\int_0^x fe^C/a + f(x)\psi(x)\bigg] = \psi(0)^{-1}\int_0^x f'\psi. \tag{A.7}$$

Thus

$$\begin{aligned}I\big(\bar f\big)(y) &= \frac{e^{-C(y)}\int_y^D \bar f e^C/a}{f'(y)}\\ &= \frac{-\int_0^y \bar f e^C/a}{e^{C(y)}f'(y)}\\ &= \frac{1}{e^{C(y)}f'(y)}\bigg[[f\psi](y) + \int_y^x f'\psi - \frac{\psi(y)}{\psi(0)}\int_0^x f'\psi\bigg], \qquad y\leqslant x.\end{aligned} \tag{A.8}$$

In particular, if we take

$$f(y) = \int_0^{y\wedge x} e^{-C} = \varphi(y\wedge x),$$

then

$$\inf_{y\in(0,x)} I\big(\bar f\big)(y) = \inf_{y\in(0,x)}\bigg[[\varphi\psi](y) + \int_y^x e^{-C}\psi - \frac{\psi(y)}{\psi(0)}\int_0^x e^{-C}\psi\bigg];$$

and if we take

$$f(y) = \int_0^{y\wedge x} z(x-z)e^{-C(z)}\text{d}z,$$

then

$$\inf_{y\in(0,x)} I\big(\bar f\big)(y) = \inf_{y\in(0,x)} \frac{1}{y(x-y)}\bigg[[f\psi](y) + \int_y^x f'\psi - \frac{\psi(y)}{\psi(0)}\int_0^x f'\psi\bigg].$$

Next, we discuss $\sup_{x\in(0,D)} \|\bar f\|^2/D(f)$, again for f with $f(0)=0$ and stopping at x. First, by using (A.7). we have

$$\bar f = f - \psi(0)^{-1}\int_0^x f'\psi,$$

and

$$\begin{aligned}\psi(0)\|\bar f\|^2 &= \int_0^x \bar f^2 e^C/a + \bar f(x)^2\int_x^D e^C/a\\ &= \bar f(0)^2\psi(0) + 2\int_0^x \bar f f'\psi\\ &= 2\int_0^x \bar f f'\psi + \pi(f)^2\psi(0)\\ &= 2\int_0^x \bar f f'\psi + \frac{1}{\psi(0)}\bigg(\int_0^x f'\psi\bigg)^2\\ &= 2\int_0^x f f'\psi - \frac{1}{\psi(0)}\bigg(\int_0^x f'\psi\bigg)^2.\end{aligned}$$

Next, we have

$$\psi(0)D(f) = \int_0^x f'^2 e^C.$$

Therefore

$$\sup_{x\in(0,D)}\frac{\|\bar f\|^2}{D(f)} = \sup_{x\in(0,D)}\bigg[2\int_0^x f f'\psi - \frac{1}{\psi(0)}\bigg(\int_0^x f'\psi\bigg)^2\bigg]\bigg/\int_0^x f'^2 e^C. \tag{A.9}$$

In particular, for $f=\varphi(\cdot\wedge x)$, we obtain

$$\bar\eta_1 = \sup_{x\in(0,D)}\frac{1}{\varphi(x)}\bigg\{2\int_0^x \varphi\psi e^{-C} - \frac{1}{\psi(0)}\bigg(\int_0^x \psi e^{-C}\bigg)^2\bigg\}. \tag{A.10}$$

Footnote 17. As we have seen in Example 1.5 that $\delta = 1/4$, hence $(4\delta)^{-1} < \pi^2 = \lambda_1 \not\leq \delta^{-1}$. But $\bar\eta_1 = 64/729 \approx 0.08779$, $\eta_1^* \approx 0.101127$ (with $x_0 = 1$, $\varepsilon_1 = \varepsilon_2 \approx 0.345492$, due to the symmetry of the boundaries), $\eta_1'' = 9/64 \approx 0.1406$, $\eta_{1*} \approx 0.1072$, and so

$$\frac{\eta_1''}{\bar\eta_1} = \frac{9}{64}\bigg/\frac{64}{729} = \frac{6561}{4096} \approx 1.6, \qquad \frac{\eta_1''}{\eta_1^*} \approx 1.4, \qquad \frac{\eta_{1*}}{\eta_1^*} \approx 1.06.$$

Furthermore, we have $\bar\eta_2 \approx 0.10119$, $\bar\eta_3 \approx 0.10132$. The last one is almost sharp.

Here are additional examples on the half line we have computed so far. For all of them, we have $\eta_1''/\bar\eta_1 < 2$. For instance, for the second example below, we have $\bar\eta_1 \approx 0.37$, $\eta_1^* \approx 0.48$ (with $x_0 = 2.895$, $\varepsilon_1 = 0.675$, $\varepsilon_2 = 1.85$), $\eta_1'' \approx 0.7$, and $\eta_{1*} \approx 0.5789$. Hence $\eta_{1*}/\eta_1^* \approx 1.206$. In the table below, $P_n(x)$ means a polynomial in x with degree n.

$a(x)$	$b(x)$	λ_1	Egenfunction
1	-1	$1/4$	$e^{x/2}(x/2-1)$
1	$-x$	2	$P_2(x) = x^2 - 1$
1	$-(x+1)$	3	$P_3(x)$
1	$-(x+\sqrt{3}\,)$	4	$P_4(x)$
1	$-(x+\sqrt{3+\sqrt{6}}\,)$	5	$P_5(x)$
1	$-(x+\sqrt{5+\sqrt{10}}\,)$	6	$P_6(x)$
1	$-4x^3$	unknown	unknown

Clearly, it should be rather hard to get a good and general upper estimate for the ratios $\eta_1''/\bar{\eta}_1$, η_1''/η_1^*, or η_{1*}/η_1^*, except the obvious fact that

$$\frac{\eta_{1*}}{\eta_1^*} \leqslant \frac{\eta_1''}{\eta_1^*} \leqslant \frac{\eta_1''}{\bar{\eta}_1}.$$

Footnote 20. Similar to the footnote given in the proof of Theorem 1.4, we may redefine $f_n^{(k)} = \big[f_{n-1}^{(k)} I\!I\big(f_{n-1}^{(k)}\big)\big](\cdot \wedge k)$ for $n \geqslant 2$. Then we also have the upper estimate $\bar{\eta}_n^{-1}$ for λ_1, where

$$\bar{\eta}_n = \sup_{k \geqslant 1} \frac{\big\|\bar{f}_n^{(k)}\big\|^2}{D\big(f_n^{(k)}\big)}, \qquad n \geqslant 1, \tag{A.12.}$$

In particular,

$$\bar{\eta}_1 = \sup_{k \geqslant 1} \frac{1}{\varphi_k}\left\{2\sum_{i=0}^{k-1}\frac{Q_i'}{\mu_i b_i} - \frac{1}{\mu}\left[\sum_{j=0}^{k-1}\frac{1}{\mu_j b_j}\sum_{i \geqslant j+1}\mu_i\right]^2\right\}, \tag{A.13}$$

where the first term on the right-hand comes from the proof of [1; Theorem 3.5] and

$$Q_i' = \left[\sum_{j \leqslant i-1}\frac{1}{\mu_j b_j} + \frac{1}{2\mu_i b_i}\right]\sum_{j \geqslant i+1}\mu_j.$$

Finally, we mention that the discrete analog of Lemma A.1 is meaningful in the present context.

Up to now, we have not worked on the convergence of all the approximating sequences. The reason is that the computation of these sequences is generally not practical. Fortunately, we need only a few of the steps of the iteration in practice, or even easier just modify the initial test function based on the first few of the iterated functions or on some rough knowledge of the eigenfunction, as illustrated by the use of $f_1^{(x,\varepsilon)}$, $f_1^{(\varepsilon_1,\varepsilon_2)}$ and $f_1^{(x_0,\varepsilon_1,\varepsilon_2)}$. These modifications are meaningful for general situation, in the next paper one can see much more examples of the design of the test functions for a particular model.

Sci. China (A) 2001, 31:9 (Chinese Edition), 807–816;
2002, 45:4(English Edition), 450-461

Linear approximation of the first eigenvalue on compact manifolds

Mu-Fa Chen
(Dept. of Math., Beijing Normal University, Beijing 100875, PRC)

E. Scacciatelli
(Dip. di Mat., Univ. "La Sapienza", 00815 Rome, Italy)

Liang Yao
(Dept. of Math., Beijing Normal University, Beijing 100875, PRC)

(Received March 19, 2001; accepted May 20, 2001)

Abstract For compact, connected Riemannian manifolds with Ricci curvature bounded below by a constant, what is the linear approximation of the first eigenvalue of Laplacian? The answer is presented with computer assisted proof and the result is optimal in certain sense.

Keywords: First eigenvalue, Riemannian manifolds, linear approximation

1 Main result

Let M be a compact, connected Riemannian manifold, without or with convex boundary ∂M. When $\partial M \neq \emptyset$, we adopt Neumann boundary condition. Next, let $\mathrm{Ric}_M \geqslant K$ for some $K \in \mathbb{R}$. Denote by d and D respectively the dimension and diameter of M. We are interested in the estimates of the first non-trivial eigenvalue of Laplacian. On this topic, there is a great deal of publications (see refs. [1–3] and references within). One of the problems is to find out a linear estimate, formally independent of the dimension d (cf. Problem 1 in ref. [4]):

$$\lambda_1 \geqslant \pi^2/D^2 + \delta K \quad (\delta \in \mathbb{R}). \tag{1.1}$$

Typeset by $\mathcal{A}_{\mathcal{M}}\mathcal{S}$-TEX

The linear estimate has been improved step by step as follows:

Zhung & Yang (1984)[5].	$\dfrac{\pi^2}{D^2}$, if $K \geqslant 0$.
Yang (1999)[6].	$\dfrac{\pi^2}{D^2} + \dfrac{K}{4}$, if $K \geqslant 0$.
Chen &Wang(1997)[3].	$\dfrac{\pi^2}{D^2} + \max\left\{\dfrac{\pi}{4d}, 1 - \dfrac{2}{\pi}\right\}K$, if $K \geqslant 0$ $(1 - 2/\pi \approx 0.36338)$
Cai (1991)[7].	$\dfrac{\pi^2}{D^2} + K$, if $K \leqslant 0$.
Chen & Wang (1997)[3].	$\dfrac{\pi^2}{D^2} + \left(\dfrac{\pi}{2} - 1\right)K$, if $K \leqslant 0$ $(\pi/2 - 1 \approx 0.5708)$
Zhao (1999)[8].	$\dfrac{\pi^2}{D^2} + 0.52K$, if $K \leqslant 0$

The first estimate is optimal at $K = 0$. It was also proved in ref. [8] that $\lambda_1 \geqslant \pi^2/D^2 + K/2$ whenever $-5\pi^2/(3D^2) \leqslant K \leqslant 0$. The main result of the paper is as follows.

Theorem 1.1 (computer assisted).[1] In general, we have $\lambda_1 \geqslant \pi^2/D^2 + K/2$ for all K. Especially, we have

$$\lambda_1 \geqslant \pi^2/D^2 + (3 - \pi^2/4)K, \quad \text{if } K \geqslant 4/D^2. \tag{1.2}$$

$$\lambda_1 \geqslant \pi^2/D^2 + (\pi^2/4 - 2)K, \quad \text{if } K \leqslant -4/D^2. \tag{1.3}$$

$$\lambda_1 \geqslant \pi^2/D^2 + (\pi^2/8\alpha_0 - 1)K, \quad \text{if } K \leqslant -8\alpha_0/D^2, \tag{1.4}$$

where $\alpha_0 \approx 0.85403$ is the unique solution to the equation

$$e^{\alpha} = 2\alpha \int_0^1 e^{\alpha y^2}, \qquad \alpha \in [0, 3).$$

More explicitly, $3 - \pi^2/4 \approx 0.532599$, $\pi^2/4 - 2 \approx 0.467401$ and $\pi^2/8\alpha_0 - 1 \approx 0.444563$. Actually, the result can be further refined by using more exact solutions,

---

[1] [Sep. 23, 2007] Actually, this paper (especially, Lemma 2.5) proves the following result. Let $\alpha = D^2K/8$. Define

$$H(x) = \sum_{m=0}^{\infty} \frac{1}{(2m)!} \prod_{k=1}^{m} [2(2k-1)\alpha - x], \qquad x \geqslant 0$$

and let λ_0 be the first root of H on $(0, \infty)$. Then the optimal linear approximation of λ_1 is as follows.

$$\lambda_1 \geqslant 4\lambda_0/D^2. \qquad \text{(Lemma 2.1)}$$

Certainly, this result is not explicit and the aim of Theorem 1.1 is to present some explicit estimates.

given in the next section, to an ordinary differential equation. For instance, with a partially numerical proof, we will show in the last section that

$$\lambda_1 \geqslant \pi^2/D^2 + K/2 + (5 - \pi^2/2)D^2K^2/8, \qquad \text{if } |K| \leqslant 4/D^2 \tag{1.5}$$

with equality holds at $K = 0$ or $|K| = 4/D^2$. It is explained in the next section that the above result is optimal in certain sense. This may yield some confusion since to our knowledge, there is still no concrete geometric example with negative curvature for which λ_1 is precisely known, and moreover, one is looking for the dimension-free estimate here. On the other hand, in order to determine the precise constant, one has to handle several double or triple integrals, they are rather technical and may have no special value and so are left to computer. Hence, the result is computer assisted. Of course, the method used here can be also applied to improving the other corollaries given in [3].

2 The ideas of the proof

The proof consists of four steps. First, apply the variational formula for the lower bound of λ_1, given in ref. [3], to reduce the higher-dimensional case to dimension one. Next, simplify further the one-dimensional problem in terms of the dimension-free consideration. Thirdly, find out some particular solutions to the Sturm–Liouville eigenvalue problem, this provides us a possibility to determine the constants given in (1.2)–(1.5). Finally, apply an approximation procedure, introduced in ref. [9], to proving that the linear approximation holds. Actually, one purpose of the present study is to justify the power of the approximation procedure. In the last two steps, a duality between $\pm\alpha$ ($|\alpha| \leqslant 1/2$) (see Lemma 2.6 below) plays a critical role. The first three steps are completed in this section and the last step will be completed in the next section.

The variational formula given in ref. [3] is as follows:

$$\lambda_1 \geqslant 4 \sup_{f\in\mathscr{F}} \inf_{r\in(0,D)} f(r)\left\{\int_0^r e^{-C(s)}\mathrm{d}s \int_s^D \left[e^C f\right](u)\mathrm{d}u\right\}^{-1}, \tag{2.1}$$

where

$$\begin{aligned}
C(r) &= \frac{1}{4}\int_0^r \gamma(s)\mathrm{d}s,\\
\gamma(r) &= \begin{cases} -2\sqrt{K(d-1)}\,\tan\left[\frac{r}{2}\sqrt{K/(d{-}1)}\,\right] & \text{if } K \geqslant 0,\\ 2\sqrt{-K(d-1)}\,\tanh\left[\frac{r}{2}\sqrt{-K/(d-1)}\,\right] & \text{if } K \leqslant 0,\end{cases}\\
\mathscr{F} &= \{f \in C[0,D] : f|_{(0,D)} > 0\}.
\end{aligned}$$

Here $\mathscr{F}$ is the set of test functions. The estimate given in (2.1) is essentially a comparison theorem for eigenvalues (cf. refs. [3] and [10]). Actually, if we denote by $\lambda_0^{(1)}$ the first mixed eigenvalue of the operator $L_1 = 4\mathrm{d}^2/\mathrm{d}r^2 + \gamma(r)\mathrm{d}/\mathrm{d}r$ (with boundary conditions $f(0) = 0$, $f'(D) = 0$), then

$$\lambda_1 \geqslant \lambda_0^{(1)} \geqslant \text{ the right-hand side of (2.1).}$$

In details, if there exist $f : f(0) = 0, f'(D) = 0, f'|_{(0,D)} > 0$ and constant $\varepsilon > 0$ such that

$$4f'' + \gamma f' + \varepsilon f \leqslant 0, \tag{2.2}$$

then $\lambda_1 \geqslant \varepsilon$ (actually, whenever $\lambda_0^{(1)} > 0$, the eigenfunction of $\lambda_0^{(1)}$ must satisfy the conditions as those of f just listed above). In fact, the last assertion is equivalent to $\lambda_0^{(1)} \geqslant$ the right-hand side of (2.1).

Before moving on, let us mention an equivalent result. Since γ is odd: $\gamma(-r) = -\gamma(r)$, the first mixed eigenvalue of L_1 on $(0, D)$ coincides with the first Neumann eigenvalue of L_1 on $(-D, D)$. Hence, the last eigenvalue lower bounds λ_1. This is the main result presented in, Theorem 2 and its Remark of ref. [11] and Theorem 14 of ref. [12].

As pointed in ref. [4], when $d \uparrow \infty$, $\gamma(r) \uparrow -Kr$. Thus, it suffices to consider the mixed eigenvalue of the operator

$$L_2 = 4\frac{\mathrm{d}^2}{\mathrm{d}r^2} - Kr\frac{\mathrm{d}}{\mathrm{d}r},$$

because every solution to the differential inequality

$$4f'' - Krf' + \varepsilon f \leqslant 0 \tag{2.3}$$

with the same boundary conditions must satisfy (2.2). Conversely, if (2.2) holds for all d, then so does (2.3). On the other hand, by making a change of the variable, one reduces D to be 1. In details, if we denote by $\lambda_0 = \lambda_0(\alpha)$ the first mixed eigenvalue of operator

$$L = \frac{\mathrm{d}^2}{\mathrm{d}x^2} - 2\alpha x\frac{\mathrm{d}}{\mathrm{d}x}$$

on the interval $(0, 1)$ (with boundary conditions $f(0) = 0$ and $f'(1) = 0$), where $\alpha = D^2K/8$, then we have the following result.

Lemma 2.1. $\lambda_1 \geqslant 4\lambda_0/D^2$.

From now on, fix the notations α, L and $\lambda_0 = \lambda_0(\alpha)$ just used above. It is well known that there is no explicit solution of λ_0 for general α. Fortunately, we do have some particular solutions. The first one below is well known.

Lemma 2.2. When $\alpha = 0$, we have eigenvalue $\lambda_0 = \pi^2/4$ with eigenfunction $g(x) = \sin(\pi x/2)$.

Indeed, we have infinitely many particular solutions.

Lemma 2.3. For each integer $n \geqslant 2$, let α_n be the minimal positive root of the polynomial

$$\sum_{k=1}^{n} \frac{(-4\alpha)^{k-1}}{(n-k)!(2k-2)!},$$

then we have eigenvalue $\lambda_0(\alpha_n) = 2(2n-1)\alpha_n$ with eigenfunction

$$g_n(x) = \sum_{k=1}^{n} \frac{1}{(n-k)!(2k-1)!}(-4\alpha_n)^{k-1}x^{2k-1}.$$

Here are the first four particular solutions.

n	α_n	$\lambda_0(\alpha_n)$
2	$1/2$	3
3	$\big(3-\sqrt{6}\,\big)/2 \approx 0.275255$	$5\big(3-\sqrt{6}\,\big) \approx 2.75255$
4	$\frac{1}{2}\big(5-\sqrt{10}\,\cos\theta - \sqrt{30}\,\sin\theta\big) \approx 0.190164$ $\theta := \frac{1}{3}\arctan\sqrt{3/2}$	$14\alpha_4 \approx 2.66229$
5	$\frac{7}{2} - \frac{B}{\sqrt{2}} - \frac{1}{2^{1/4}}\sqrt{-\frac{14}{B} + \frac{21-B^2}{\sqrt{2}}} \approx 0.145304$ $B := \sqrt{7+\sqrt{70}\,\cos\left(\frac{1}{3}\arctan\sqrt{3/7}\right)}$	$18\alpha_5 \approx 2.61546$

When $n \geqslant 6$, one has to use numerical computation. Note that α_n is strictly decreasing as n increases. This result corresponds to the positive curvature. It is interesting that a "dual" of the result gives us particular solutions to the case of negative curvature.

Lemma 2.4. Let $n \geqslant 2$ and α_n be the same as above. Then, corresponding to $\alpha = -\alpha_n$, we have eigenvalue $\lambda_0(-\alpha_n) = 4(n-1)\alpha_n$ with eigenfunction

$$g_n(x) = e^{-\alpha_n x^2} \sum_{k=1}^{n-1} \frac{1}{(n-1-k)!(2k-1)!}(-4\alpha_n)^{k-1}x^{2k-1}.$$

Additionally, corresponding to $\alpha = -\alpha_0$ (given in Theorem 1.1), we have $\lambda_0(-\alpha_0) = 2\alpha_0$ with eigenfunction

$$g(x) = e^{-\alpha_0 x^2} \int_0^x e^{\alpha_0 y^2}.$$

Having the lemmas in mind, it is easy to understand the meaning of Theorem 1.1. The general estimate means that the curve $\lambda_0 = \lambda_0(\alpha)$ has the tangent line $\lambda(\alpha) = \pi^2/4 + \alpha$ from below. The estimate (1.2) means that the curve $\lambda_0(\alpha)$ on $[1/2, \infty)$ is above the straight line connecting the two points $(0, \pi^2/4)$ and $(1/2, 3)$. Similarly, the curve on $(-\infty, -\alpha_0]$ is above the straight line connecting $(0, \pi^2/4)$ and $(-\alpha_0, 2\alpha_0)$. These facts explain the meaning of the term "optimal". It should be clear now that one may further refine Theorem 1.1 by using the other particular solutions given by Lemmas 2.3 and 2.4. For instance, the last straight line may be replaced by the tangent line to the curve $\lambda_0(\alpha)$ at $-\alpha_0$.

To prove the lemmas, we need some preparations. The first one below is a characterization of the eigenfunctions.

Lemma 2.5. The eigenfunction g of $\lambda_0(\alpha)$ can be expressed as

$$g(x) = \sum_{m=1}^{\infty} \frac{x^{2m-1}}{(2m-1)!} \prod_{k=1}^{m-1} [2(2k-1)\alpha - \lambda_0(\alpha)]. \tag{2.4}$$

Moreover,

$$\sum_{m=0}^{\infty} \frac{x^{2m}}{(2m)!} \prod_{k=1}^{m} [2(2k-1)\alpha - \lambda_0(\alpha)] \geqslant 0, \tag{2.5}$$

the sign of equality holding iff $x = 1$:

$$\sum_{m=0}^{\infty} \frac{1}{(2m)!} \prod_{k=1}^{m} [2(2k-1)\alpha - \lambda_0(\alpha)] = 0. \tag{2.6}$$

Proof. Let $\alpha \in \mathbb{R}$. Consider the differential equation $f'' - 2\alpha x f' + \lambda f = 0$ on $(0,1)$ with boundary conditions $f(0) = 0$ and $f'(1) = 0$. Without loss of generality, one may assume that the eigenfunction g of λ_0 also satisfies $g'|_{(0,1)} > 0$. Let f be a solution to the equation with power series expansion $f(x) = \sum_{n=0}^{\infty} a_n x^n$. Then, the boundary condition $f(0) = 0$ gives us

$$a_0 = 0, \qquad a_{2n} = 0, \qquad a_{n+2} = \frac{2\alpha n - \lambda}{(n+2)(n+1)} a_n, \qquad n \geqslant 1.$$

Without loss of generality, let $a_1 = 1$. Set $\beta_m = a_{2m-1}$, $m \geqslant 1$. Then, we have

$$\beta_{m+1} = \frac{2\alpha(2m-1) - \lambda}{2m(2m+1)} \beta_m, \qquad m \geqslant 1.$$

By induction, we obtain

$$\beta_m = \frac{1}{(2m-1)!} \prod_{k=1}^{m-1} [2(2k-1)\alpha - \lambda], \qquad m \geqslant 1 \tag{2.7}$$

and then

$$f(x) = \sum_{m=1}^{\infty} \beta_m x^{2m-1}.$$

The series always converges absolutely since $x \in [0,1]$ and $|\beta_{m+1}/\beta_m| \sim m^{-1}$. The same conclusion holds for

$$f'(x) = \sum_{m=1}^{\infty} (2m-1)\beta_m x^{2m-2},$$

as well as for $f''(x)$. Now, the boundary condition $f'(1) = 0$ gives us

$$\sum_{m=1}^{\infty}(2m-1)\beta_m = 0.$$

Next, when $\lambda = \lambda_0(\alpha)$, we have $f'(x) \geqslant 0$, the sign of equality holds iff at the boundary $x = 1$. From this, one deduces (2.5) and (2.6) by simple computations. $\square$

The next result describes the "duality" mentioned before.

Lemma 2.6. Let $|\alpha| \leqslant 1/2$. Then $\lambda_0(\alpha) = \lambda_0(-\alpha) + 2\alpha$ and its eigenfunction has the expression $g(x) = e^{\alpha x^2}\sum_{m=1}^{\infty}\frac{x^{2m-1}}{(2m-1)!}\prod_{k=1}^{m-1}[-2(2k+1)\alpha - \lambda_0(-\alpha)]$. Moreover,

$$\sum_{m=0}^{\infty}\frac{x^{2m}}{(2m)!}\prod_{k=1}^{m}[-2(2k-1)\alpha - \lambda_0(-\alpha)] \geqslant 0, \tag{2.8}$$

the sign of equality holds iff $x = 1$:

$$\sum_{m=0}^{\infty}\frac{1}{(2m)!}\prod_{k=1}^{m}[-2(2k-1)\alpha - \lambda_0(-\alpha)] = 0. \tag{2.9}$$

Proof. Let $f(x) = e^{-\alpha x^2}g(x)$. Then the equation

$$g'' - 2\alpha x g' + \lambda g = 0$$

becomes

$$f'' + 2\alpha x f' + (\lambda + 2\alpha)f = 0.$$

This provides us the duality between α and $-\alpha$. However, the condition $g'|_{(0,1)} > 0$ becomes $f' > -2\alpha x f$ on $(0, 1)$, and

$$g'(1) = 0 \Longleftrightarrow f'(1) = -2\alpha f(1).$$

From the first paragraph of the proof of Lemma 2.5, it follows that

$$f(x) = \sum_{m=1}^{\infty}\beta_m x^{2m-1} \quad \text{and} \quad \beta_m = \frac{1}{(2m-1)!}\prod_{k=1}^{m-1}[-2(2k-1)\alpha - \lambda - 2\alpha].$$

Replacing λ by $\lambda' + 2\alpha$, we get

$$\beta_m = \frac{1}{(2m-1)!}\prod_{k=1}^{m-1}[-2(2k+1)\alpha - \lambda'] = \frac{1}{(2m-1)!}\prod_{k=2}^{m}[-2(2k-1)\alpha - \lambda'].$$

Now,

$$f'(x) = \sum_{m=1}^{\infty}c_m x^{2m-2}, \qquad c_m := (2m-1)\beta_m.$$

$$f' > -2\alpha x f \Longleftrightarrow c_1 > \sum_{m=1}^{\infty}(-c_{m+1} - 2\alpha\beta_m)x^{2m}. \tag{2.10}$$

Note that $c_{m+1} + 2\alpha\beta_m = \frac{1}{(2m)!}\prod_{k=1}^m[-2(2k-1)\alpha-\lambda']$ and $c_1 = 1$. By (2.10), the inequality $f' > -2\alpha x f$ holds iff $1 > -\sum_{m=1}^\infty \frac{x^{2m}}{(2m)!}\prod_{k=1}^m[-2(2k-1)\alpha-\lambda']$, that is,

$$\sum_{m=0}^\infty \frac{x^{2m}}{(2m)!}\prod_{k=1}^m[-2(2k-1)\alpha-\lambda'] > 0. \tag{2.11}$$

Thus, $f'(1) = -2\alpha f(1)$ iff

$$\sum_{m=0}^\infty \frac{1}{(2m)!}\prod_{k=1}^m[-2(2k-1)\alpha-\lambda'] = 0. \tag{2.12}$$

By Lemma 2.5, the last two conditions (2.11) and (2.12) mean that λ' is the first eigenvalue of the operator $L = \mathrm{d}^2/\mathrm{d}x^2 + 2\alpha x\mathrm{d}/\mathrm{d}x$ with eigenfunction

$$\sum_{m=1}^\infty \frac{x^{2m-1}}{(2m-1)!}\prod_{k=1}^{m-1}[-2(2k-1)\alpha-\lambda']$$

and so $\lambda' = \lambda_0(-\alpha)$. Returning to the original $\lambda (= \lambda' + 2\alpha)$ and g, we claim that $\lambda_0(\alpha) = \lambda_0(-\alpha) + 2\alpha$ and its eigenfunction $g = e^{\alpha x^2} f$ has the expression given in the lemma. $\square$

Proof of Lemma 2.3. Let $n \geqslant 2$ be an integer and set $\lambda = 2(2n-1)\alpha$ with $\alpha > 0$ to be determined later. Then we have $\beta_k = 0$ for all $k \geqslant n+1$. Furthermore, it is easy to check that

$$\beta_k = \frac{(n-1)!}{(n-k)!(2k-1)!}(-4\alpha)^{k-1}, \qquad 1 \leqslant k \leqslant n. \tag{2.13}$$

Thus,

$$f(x) = (n-1)!\sum_{k=1}^n \frac{1}{(n-k)!(2k-1)!}(-4\alpha)^{k-1}x^{2k-1}$$

and $f'(1) = 0$ iff

$$\sum_{k=1}^n \frac{(-4\alpha)^{k-1}}{(n-k)!(2k-2)!} = 0. \tag{2.14}$$

From this, one finds the minimal root α_n, which then gives us the eigenfunction $g_n = f/(n-1)!$. $\square$

Proof of Lemma 2.4. Let α_n $(n \geqslant 2)$ be given as in Lemma 2.3. Then

$$\prod_{k=1}^{m-1}[2(2k+1)\alpha_n - \lambda_0(\alpha_n)] = \begin{cases} (-4\alpha_n)^{m-1}(n-2)!/(n-1-m)!, & \text{if } m \leqslant n-1 \\ 0, & \text{otherwise.} \end{cases}$$

Hence

$$
\begin{aligned}
& e^{-\alpha_n x^2} \sum_{m=1}^{\infty} \frac{x^{2m-1}}{(2m-1)!} \prod_{k=2}^{m} [2(2k-1)\alpha_n - \lambda_0(\alpha_n)] \\
&= e^{-\alpha_n x^2} (n-2)! \sum_{m=1}^{n-1} \frac{(-4\alpha_n)^{m-1} x^{2m-1}}{(n-1-m)!(2m-1)!}.
\end{aligned}
$$

By Lemma 2.6, we have $\lambda_0(-\alpha_n) = \lambda_0(\alpha_n) - 2\alpha_n = 4(n-1)\alpha_n$ with the required eigenfunction. We have thus proved the main part of the lemma.

The proof of the last assertion is much easier. Again, one needs to show that $g(0) = 0$, $g'|_{(0,1)} > 0$ and $g'(1) = 0$. $\square$

3 Proof of Theorem 1.1 (computer assisted)

By Lemma 2.1, it suffices to show that

$$
\begin{aligned}
&\lambda_0 \geqslant \pi^2/4 + \alpha, \qquad \text{for all } \alpha && (3.1)\\
&\lambda_0 \geqslant \pi^2/4 + (6 - \pi^2/2)\alpha, \qquad \text{if } \alpha \geqslant 1/2 && (3.2)\\
&\lambda_0 \geqslant \pi^2/4 + (\pi^2/2 - 4)\alpha, \qquad \text{if } \alpha \leqslant -1/2 && (3.3)\\
&\lambda_0 \geqslant \pi^2/4 + (\pi^2/4\alpha_0 - 2)\alpha, \qquad \text{if } \alpha \leqslant -\alpha_0 \ . && (3.4)
\end{aligned}
$$

By Lemma 2.2, the four inequalities all become equalities at $\alpha = 0$. By Lemmas 2.3 and 2.4, the equality in (3.2)–(3.4) also holds at $\alpha = 1/2$, $-1/2$ and $-\alpha_0$ respectively. We need to show that the inequalities hold for all other α. The idea is using an approximation procedure proposed in ref. [9].

Let

$$
\begin{aligned}
& C(x) = -\alpha x^2, \\
& \mathscr{F} = \{ f \in C[0, D] \colon f(0) = 0, f|_{(0,1)} > 0 \}, \\
& II(f)(x) = f(x)^{-1} \int_0^x e^{-C(y)} \mathrm{d}y \int_y^1 f e^C.
\end{aligned}
$$

Clearly, $II(f) \in \mathscr{F}$ for every $f \geqslant 0$. Define

$$
\begin{aligned}
\varphi(x) &= \int_0^x e^{-C(y)} \mathrm{d}y, \\
f_1 &= \sqrt{\varphi}, \qquad f_{n+1} = f_n II(f_n), \ n \geqslant 1, \\
\delta_n &= \inf_{x \in (0,1)} II(f_n)(x)^{-1}, \\
\delta_n' &= \sup_{x \in (0,1)} II(f_n)(x)^{-1}.
\end{aligned}
$$

It is proved in ref. [9] that $\delta_n \downarrow$ and $\delta_n' \uparrow$ as $n \uparrow$, ${\delta_n'}^{-1} \geqslant \lambda_0 \geqslant \delta_n^{-1}$ for all n. Thus, by (3.1)–(3.4), it suffices to show that there exists some $n \geqslant 1$ such that

$$\sup_{x\in(0,1)} II(f_n)(x) \leqslant [\pi^2/4+\alpha]^{-1}, \qquad \text{if } |\alpha| \leqslant 1/2 \tag{3.5}$$

$$\sup_{x\in(0,1)} II(f_n)(x) \leqslant [\pi^2/4+(6-\pi^2/2)\alpha]^{-1}, \qquad \text{if } \alpha \geqslant 1/2 \tag{3.6}$$

$$\sup_{x\in(0,1)} II(f_n)(x) \leqslant [\pi^2/4+(\pi^2/2-4)\alpha]^{-1}, \qquad \text{if } -\alpha_0 \leqslant \alpha \leqslant -1/2 \tag{3.7}$$

$$\sup_{x\in(0,1)} II(f_n)(x) \leqslant [\pi^2/4+(\pi^2/4\alpha_0-2)\alpha]^{-1}, \qquad \text{if } -\alpha_0\pi^2/(\pi^2-8\alpha_0) \leqslant \alpha \leqslant -\alpha_0. \tag{3.8}$$

The difficulty comes from the fact that one has to compute one- or two-more multiple of integrals in each iteration. Thus, it becomes impractical for more than three iterations, unless the integrands are simple (polynomials, for instance). Because the iteration is not only an approximation for the eigenvalue but also for eigenfunction, it is natural to use certain modification of the explicitly known eigenfunctions as initial function instead of the original one: $f_1 = \sqrt{\varphi}$. Correspondingly, divide the interval real line into five parts: $[1/2,\infty)$, $(0,1/2)$, $(-1/2,0)$, $(-\alpha_0,-1/2)$ and $(-\alpha_0\pi^2/(\pi^2-8\alpha_0),-\alpha_0)$.

(a) First, we consider the case that $\alpha \geqslant 1/2$. This is a little far from the linear approximation and so it is easier to handle. Take $f_1(x) = x - x^3/3$, which is the eigenfunction of λ_0 when $\alpha = 1/2$. Define

$$J_m(x) = \int_0^x e^{\alpha y^2}\mathrm{d}y \int_y^1 u^{2m-1}e^{-\alpha u^2}\mathrm{d}u, \qquad m \geqslant 1.$$

Then

$$J_{m+1}(x) = -\frac{1}{2\alpha}e^{-\alpha}\varphi(x) + \frac{1}{2(2m+1)\alpha}x^{2m+1} + \frac{m}{\alpha}J_m(x), \tag{3.9}$$

where $\varphi(x) = \int_0^x e^{\alpha y^2}$, and so

$$J_m(x) = -\frac{e^{-\alpha}\varphi(x)}{2}\sum_{k=0}^{m-1}\frac{(m-1)!}{(m-1-k)!\alpha^{k+1}} + \frac{1}{2}\sum_{k=0}^{m-1}\frac{(m-1)!x^{2(m-k)-1}}{(m-1-k)!\alpha^{k+1}[2(m-k)-1]}. \tag{3.10}$$

Hence

$$\begin{aligned} f_2(x) &= \int_0^x e^{\alpha y^2}\mathrm{d}y \int_y^1 f_1(u)e^{-\alpha u^2}\mathrm{d}u \\ &= J_1(x) - J_2(x)/3 \\ &= \frac{1-2\alpha}{6\alpha^2}e^{-\alpha}\varphi(x) + \frac{3\alpha-1}{6\alpha^2}x - \frac{1}{18\alpha}x^3. \end{aligned} \tag{3.11}$$

Then we claim that $II(f_1) = f_2/f_1$ satisfies (3.6) on the interval $[0.725, \infty)$. That is, the curve $[f_2/f_1](x)$ should be located below the straight line $y(x) \equiv [\pi^2/4 + \alpha]^{-1}$. It is just at this place, we use mathematical software to plot the functions (All the checks in the paper are done by using Mathematica 3.0 on PC 266). We remark that it is often true that once the required conclusion holds at some α_c, which is in the interior of the considered interval, then it should also hold for all $\alpha > \alpha_c$ on the interval. We will no longer repeat this fact in what follows. The reason we use $f_2/f_1 \leqslant C$ rather than the equivalent inequality $f_2 \leqslant Cf_1$, which is easier in computation, is that the inverse of the supremum (infimum) of f_2/f_1 represents a lower (upper) bound of λ_0. Hence, the oscillation $\operatorname{osc}(f_2/f_1)$ describes the difference of f_2 and the eigenfunction (the smaller oscillation is the closer one). In other words, if $\operatorname{osc}(f_2/f_1)$ is not very smaller, then there is still a room for an improvement in the next iteration.

To cover the interval $(1/2, 0.725)$, we use one more iteration. Because

$$\begin{aligned} f_{n+1}(x) &= \int_0^x e^{-C(y)}\mathrm{d}y \int_y^1 f_n(u)e^{C(u)}\mathrm{d}u \\ &= \int_0^1 f_n(u)e^{C(u)}\varphi(x \wedge u)\mathrm{d}u \\ &= \int_0^x f_n(u)e^{C(u)}\varphi(u)\mathrm{d}u + \varphi(x)\int_x^1 f_n(u)e^{C(u)}\mathrm{d}u, \end{aligned} \tag{3.12}$$

combining this with (3.9)–(3.11), we get

$$\begin{aligned} f_3(x) &= \int_0^x e^{\alpha y^2}\mathrm{d}y \int_y^1 f_2(u)e^{-\alpha u^2}\mathrm{d}u \\ &= \frac{3\alpha - 1}{6\alpha^2}J_1(x) - \frac{1}{18\alpha}J_2(x) \\ &\quad + \frac{1-2\alpha}{6\alpha^2}e^{-\alpha}\left[\int_0^x \varphi(u)^2 e^{-\alpha u^2}\mathrm{d}u + \varphi(x)\int_x^1 \varphi(u)e^{-\alpha u^2}\mathrm{d}u\right] \\ &= \frac{1-2\alpha}{9\alpha^3}e^{-\alpha}\varphi(x) + \frac{9\alpha - 4}{36\alpha^3}x - \frac{1}{108\alpha^2}x^3 \\ &\quad + \frac{1-2\alpha}{6\alpha^2}e^{-\alpha}\left[\int_0^x \varphi(u)^2 e^{-\alpha u^2}\mathrm{d}u + \varphi(x)\int_x^1 \varphi(u)e^{-\alpha u^2}\mathrm{d}u\right]. \end{aligned}$$

Only double integrals are met here. The computations made above decrease the multiplicity of integrals, which are critical in using mathematical software. Then, we need only to check by computer that $II(f_2) = f_3/f_2$ satisfies (3.6) on the interval $(0.5, 0.725)$. Certainly, it is helpful to simplify the expression first before going to plot it.

(b) Next, consider $\alpha \leqslant -\alpha_0$. This is also an easier case. Take $f_1(x) = e^{\alpha x^2}\int_0^x e^{-\alpha y^2}$. Then, by (3.12),

$$f_2(x) = \int_0^x f_1(u)\varphi(u)e^{-\alpha u^2}\mathrm{d}u + \varphi(x)\int_x^1 f_1(u)e^{-\alpha u^2}\mathrm{d}u =$$

$$
\begin{aligned}
&= \int_0^x \mathrm{d}u\varphi(u)\int_0^u e^{-\alpha y^2}\mathrm{d}y + \varphi(x)\int_x^1 \mathrm{d}u\int_0^u e^{-\alpha y^2}\mathrm{d}y \\
&= \int_0^x \varphi\psi + \varphi(x)\big[\psi(1) - x\psi(x) + \big(e^{-\alpha} - e^{-\alpha x^2}\big)/2\alpha\big],
\end{aligned}
$$

where $\varphi(x) = \int_0^x e^{\alpha y^2}$ and $\psi(x) = \int_0^x e^{-\alpha y^2}$. Then $I\!I(f_1) = f_2/f_1$ covers the interval $(-\alpha_0\pi^2/(\pi^2 - 8\alpha_0), -1.55)$. To cover the other part of the interval, one has to use the next iteration (cf. (3.12)):

$$
f_3(x) = \int_0^x \varphi(u)f_2(u)e^{-\alpha u^2}\mathrm{d}u + \varphi(x)\int_x^1 f_2(u)e^{-\alpha u^2}\mathrm{d}u. \tag{3.13}
$$

Here, a triple integral is used. In plotting the functions, one has to be careful about the computation errors. For instance, at the point $-\alpha_0$, the ratio f_3/f_2 should be the constant $2\alpha_0$ in the interval $[0,1]$, but the result picture can be different.

(c) Now, consider $\alpha \in [-\alpha_0, -1/2)$. Recall that $C(x) = -\alpha x^2$. It is natural to take $f_1(x) = xe^{\alpha x^2}$. Then

$$
\begin{aligned}
f_2(x) &= \int_0^x f_1\varphi e^C + \varphi(x)\int_x^1 f_1 e^C \\
&= \int_0^x u\varphi(u)\mathrm{d}u + \varphi(x)\int_x^1 u\mathrm{d}u \\
&= \frac{1}{2}\varphi(x) - \frac{1}{2}\int_0^x u^2 e^{\alpha u^2}\mathrm{d}u \\
&= \frac{1}{2}\left(1 + \frac{1}{2\alpha}\right)\varphi(x) - \frac{x}{4\alpha}e^{\alpha x^2}.
\end{aligned}
$$

However, $I\!I(f_1)$ satisfies (3.7) only at one point $\alpha = -1/2$. So, we have to go to the next iteration.

$$
\begin{aligned}
f_3(x) &= \int_0^x f_2\varphi e^C + \varphi(x)\int_x^1 f_2 e^C \\
&= \frac{1}{2}\left[1 + \frac{1}{2\alpha}\right]\left[\int_0^x \varphi^2 e^C + \varphi(x)\int_x^1 \varphi e^C\right] - \frac{1}{4\alpha}\left[\int_0^x u\varphi(u)\mathrm{d}u + \varphi(x)\int_x^1 u\mathrm{d}u\right] \\
&= \frac{1}{2}\left[1 + \frac{1}{2\alpha}\right]\left[\int_0^x \varphi^2 e^C + \varphi(x)\int_x^1 \varphi e^C - \frac{\varphi(x)}{4\alpha}\right] + \frac{x}{16\alpha^2}e^{\alpha x^2}.
\end{aligned}
$$

Again, only double integrals are met here. Then, check that $I\!I(f_2) = f_3/f_2$ satisfies (3.7) by using computer.

(d) We now go to the harder part of the proof: $|\alpha| \leqslant 1/2$. First, we mention that the conclusion holds in virtue of ref. [8] and Lemma 2.6. Here, we propose several different ways to check it.

Let $\alpha \in [0, 1/2]$ for a moment. As we did before, it is natural to take $f_1 = g_n$ defined by Lemma 2.3 as the test function. When $n = 3$, the first iteration f_2/f_1 covers $[0.23, 0.5]$. The computations are rather easy in terms of (3.9) and (3.10). When $n = 4$, f_2/f_1 covers $[0.166, 0.23]$ and so on. Alternatively, one may take

$$f_1(x) = \sum_{m=1}^{M} \frac{x^{2m-1}}{(2m-1)!} \prod_{k=1}^{m-1} [2(2k-1)\alpha - \lambda']$$

for large enough M, where $\lambda' = \pi^2/4 + \alpha + (10 - \pi^2)\alpha^2$, and check that f_2/f_1 (recall (3.10) again) covers an interval containing α. This method is based on (2.4) and (1.5) (see also (3.17) below). However, we have a more simpler test function described below.

Let $\alpha \in [-1/2, 0]$, which is easier to handle than $[0, 1/2]$ learnt from practice. To begin with, we explain how to choose a new test function. When $\alpha = 0$, the eigenfunction is $g(x) = \sin(\pi x/2)$. Inserting this into the differential inequality

$$g'' - 2\alpha x g' + \varepsilon g \leqslant 0 \tag{3.14}$$

gives $\lambda_0 \geqslant \varepsilon_{\max} = \pi^2/4 + 2\alpha$. Unfortunately, unless $\alpha = 0$, $\varepsilon_{\max}$ is less than the estimate required by (3.1). One may use g as the initial function instead of $\sqrt{\varphi}$, then the region of α for which (3.5) holds can be enlarged step by step by the iterations. But there is a more effective method. That is, optimizing g first, so that (3.14) produces a better lower bound and then go to the iterations. In the present situation, we take $f_1(x) = e^{\alpha x^2} \sin(\beta x)$, $\beta \in (0, \pi/2]$. It is a modification of the eigenfunction regarding $-\alpha x\, (-\alpha \ll 1)$ as a constant. Then, $f_1'(x)$ has minimum $f_1'(1)$. Moreover, $f_1'(1) \geqslant 0$ iff $\beta \cot \beta \geqslant -2\alpha$. Note that when β increases from 0 to $\pi/2$, $\beta \cot \beta$ decreases from 1 to 0. Use β as a parameter and let $-\alpha = \beta \cot \beta/2$. Inserting this f_1 into (3.14), one deduces the estimate $\varepsilon_{\max} = \beta^2 + \beta \cot \beta + (\beta \cot \beta)^2$ which is better than the one deduced by using g. In conclusion, we take

$$\begin{aligned}
f_1(x) &= e^{\alpha x^2} \sin(\beta x), \qquad \alpha = -\beta \cot \beta/2, \qquad \beta \in (0, \pi/2]. \\
f_2(x) &= \int_0^x e^{\alpha y^2} \mathrm{d}y \int_y^1 \sin(\beta u) \mathrm{d}u \\
&= \frac{1}{\beta} \int_0^x e^{\alpha y^2} [\cos(\beta y) - \cos \beta] \\
&= \frac{1}{\beta} \int_0^x e^{\alpha y^2} \cos(\beta y) - \frac{\cos \beta}{\beta} \varphi(x).
\end{aligned}$$

Only single integral is used now. Then, use computer to check that $I\!I(f_1) = f_2/f_1$ satisfies (3.5) on the interval $\beta \in (0, 1.195]$. By using (3.13), the next iteration covers the interval $[1.195, 1.51]$. Thus, in two steps of iterations, we cover 0.96 part of the interval $[0, \pi/2]$. The only remainder part is $(1.51, \pi/2]$. For this, one has to go to the third iteration. Here, we mention that it is possible to reduce the multiplicity of the integrals. Recall that $C(x) = -\alpha x^2$, $\varphi(x) = \int_0^x e^{-C}$ and $\psi(x) = \int_0^x e^C$. Define $\xi(x) = \int_0^x e^C \varphi$. Then

$$\int_0^x f_n e^C \varphi = \int_0^x f_n \mathrm{d}\xi = (f_n \xi)(x) - \int_0^x \xi f_n',$$
$$\int_x^1 f_n e^C = \int_x^1 f_n \mathrm{d}\psi = (f_n\psi)(1) - (f_n\psi)(x) - \int_x^1 \psi f_n'.$$

Thus, by (3.12), we have

$$f_{n+1}(x) = [\xi f_n](x) - \int_0^x \xi f_n' + \varphi(x)\bigg[(f_n\psi)(1) - (f_n\psi)(x) - \int_x^1 \psi f_n'\bigg],$$
$$f_{n+1}'(x) = \varphi'(x)\bigg[(f_n\psi)(1) - (f_n\psi)(x) - \int_x^1 \psi f_n'\bigg]. \tag{3.15}$$

The required assertion $I\!I(f_n) \leqslant \delta$ now becomes

$$[\xi(x) - \delta] f_n(x) + \varphi(x)\bigg[(f_n\psi)(1) - (f_n\psi)(x) - \int_x^1 \psi f_n\bigg] \leqslant \int_0^x \xi f_n'. \tag{3.16}$$

When $n \geqslant 2$, the multiplicity of the integrals in (3.16) is the same as that of f_n, but is smaller than the one of f_{n+1}. At the present, we have $n = 3$, $\delta = (\pi^2/4 + \alpha)^{-1}$,

$$f_2'(x) = e^{\alpha x^2}[\cos(\beta x) - \cos\beta]/\beta,$$
$$f_3'(x) = e^{\alpha x^2}\bigg[(f_2\psi)(1) - (f_2\psi)(x) - \int_x^1 \psi f_2'\bigg],$$
$$f_3(x) = [\xi f_2](x) - \int_0^x \xi f_2' + \varphi(x)\bigg[(f_2\psi)(1) - (f_2\psi)(x) - \int_x^1 \psi f_2'\bigg]$$

by (3.15). Next, denote by $F(x)$ the difference the right- and left-hand sides of (3.16) for fixed α. To show that $F(x) \geqslant 0$ on $[0, 1]$, it is not necessary to plot F on the whole interval. Because, by using numerical integration at a few points of x, it is easy to see that the function F first increases and then decreases. Thus, since $F(0) = 0$, it suffices to show that $F(1) \geqslant 0$ which becomes much easier in view of (3.16). This iteration extends the available interval to $[0, 1.564]$ (the corresponding α-interval is $[0.0053, 0.5]$) which covers 0.996 part of the whole interval $[0, \pi/2]$. We are satisfactory to stop at this step in view of the limitation of the accuracy of the computations. Another different way to check the conclusion will be discussed in the last part of the proof.

Before moving on, let us make some remarks about the test functions used above. Note that the restriction $\beta > 0$ is used in (d) and so $-\alpha < 1/2$. When $\beta = 0$, f_1 is degenerated. However, if one replaces $f_1(x)$ by $e^{\alpha x^2}\sin(\beta x)/\beta$, then $f_1(x) \to x e^{-x^2/2}$ as $\beta \to 0$. The limit is just the eigenfunction at $\alpha = -1/2$. Since the change by a constant does not change $I\!I(f_n)$, the above proof is still valid without any change, but now the available region of β can be extended to the left-end point $\alpha = -1/2$ on the interval.

In case (b), one may use a more general initial function

$$f_1(x) = e^{\alpha x^2} \int_0^x e^{\beta y^2} \ (\beta \leqslant -\alpha).$$

Then $f_1' > 0$ iff $e^{\beta} \geqslant -2\alpha \int_0^1 e^{\beta y^2}$. When $-\alpha \downarrow 1/2$, one has $\beta \downarrow 0$. Hence we also have $f_1(x) \to x e^{-x^2/2}$. This explains the relation between the three explicit eigenfunctions at $\alpha = 0$, $\alpha = -1/2$ and $\alpha = -\alpha_0$.

In the proof (c), one may also use

$$f_1(x) = e^{\alpha x^2} \int_0^x e^{\beta y^2} (\beta \leqslant -\alpha).$$

Similar to (d), regard $\beta\,(\beta \geqslant 0)$ as a parameter and let $\alpha = -e^{\beta}/2 \int_0^1 e^{\beta y^2}$. Then, one step of iteration covers the region $(-\alpha_0, -0.605)$, but unfortunately not $(-0.605, -0.5)$. Now, one may go to the next iteration. However, this concerns triple integrals and so is less convenient than the one used in proof (c).

(e) Finally, we prove (1.5). First, by using Mathematica, it is easy to write a program in a few of lines to compute α_n and then $\lambda_0(\alpha_n)$ defined in Lemma 2.3, and check that

$$\lambda_0(\alpha_n) = 2(2n-1)\alpha_n \geqslant \pi^2/4 + \alpha_n + (10 - \pi^2)\alpha_n^2 \tag{3.17}$$

for all $n (\geqslant 2)$ up to a large number, depending on the limitation of a computer. In other words, (1.5) holds at each point $\alpha = \alpha_n > 0$, and the sign of equality holds at $\alpha = 0$ and $\alpha = \alpha_2 = 1/2$. The same conclusion holds if α_n is replaced by $-\alpha_n$, because of Lemma 2.6.

Next, since the differences between the eigenvalues $\lambda_0(\alpha)$ and the quadratic function $y(\alpha) = \pi^2/4 + \alpha + (10 - \pi^2)\alpha^2$ along the sequence $\{\pm\alpha_n\}$ are all rather small, at most $\sim 10^{-5}$, and the curve $\lambda_0(\alpha)$ is regular, it should be believable that the curve $\lambda_0(\alpha)$ is located above the curve $y(\alpha)$. One may check this by using the standard power series solution to the eigenvalue problem, since we are now in the smaller region: $|\alpha| \leqslant 1/2$. To do so, define $\{\beta_m\}$ as in (2.5). For each $\alpha \in [0, 1/2]$, find the minimal root of

$$\sum_{m=1}^{M} (2m-1)\beta_m$$

for large enough M (say, 33) (Again, one needs a program here). Then the root can be regarded as an approximation of the eigenvalue $\lambda_0(\alpha)$.

Finally, since the straight line $z(\alpha) = \pi^2/4 + \alpha$ tangents to the curve $y(\alpha)$, the straight line $z(\alpha)$ should also be located below the curve $\lambda_0(\alpha)$ on the interval $[-1/2, 1/2]$.

We have thus completed the proof of the theorem.

Acknowledgements The first author acknowledges Prof. Wang Zi-Kun for his attention on the topic and also for his encouragement, and acknowledges the financial support for the visit during June 3–July 3, 2000 by University of Rome I. This work was supported in part by National Natural Science Foundation of China (No. 19631060), 973 project, RFDP, C.N.R. and University of Rome I, Italy.

References

[1]. Chavel, I., *Eigenvalues in Riemannian Geometry*, Academic Press, 1984.

[2]. Yau, S. T., Schoen, R., *Differential Geometry* (in Chinese), Beijing: Science Press, 1988.

[3]. Chen, M. F., Wang, F. Y., *General formula for lower bound of the first eigenvalue on Riemannian manifolds*, Science in China, Ser. A, 1997, 40(4): 384–394.

[4]. Chen, M. F., *Coupling, spectral gap and related topics* (Ⅱ), Chinese Science Bulletin, 1997, 42(17): 1409–1416.

[5]. Zhong, J. Q., Yang, H. C., *On estimate of the first eigenvalue of a compact Riemannian manifold*, Science in China, Ser. A, 1984, 12: 1251–1265.

[6]. Yang, D. G., *Lower bound estimates of the first eigenvalue for compact manifolds with positive Ricci curvature*, Pacific. J. Math., 1999, 190(2): 383–398.

[7]. Cai, K. R., *Estimate on lower bound of the first eigenvalue of a compact Riemannian manifold*, Chin. Ann. Math. (B), 1991, 12(3): 267–271.

[8]. Zhao, D., *Lower estimate of the first eigenvalue on compact Riemannian manifolds*, Science in China, Ser. A, 1999, 42(9): 897–904.

[9]. Chen, M. F., *Variational formulas and approximation theorems for the first eigenvalue in dimension one*, Science in China, Ser. A, 2001, 44(4): 409–418.

[10]. Chen, M. F., *Analytic proof of dual variational formula for the first eigenvalue in dimension one*, Science in China, Ser. A, 1999, 42(8): 805–815.

[11]. Kröger, P., *On the spectral gap for compact manifolds*, J. Differ. Geom., 1992, 36: 315–330.

[12]. Bakry, D., Qian, Z. M., *Some new results on eigenvectors via dimension, diameter and Ricci curvature*, Adv. Math., 2000, 155(1): 98–153.

Sci. China (A) 2003, 33:4 (Chinese Edition), 371–383; (English Edition),46:6, 847–861

Dual Variational Formulas for the First Dirichlet Eigenvalue on Half-Line

Chen Mufa, Zhang Yuhui and Zhao Xiaoliang

(Beijing Normal University, Beijing, 100875)
(Received March 18, 2003)

Abstract. The aim of the paper is to establish two dual variational formulas for the first Dirichlet eigenvalue of second order elliptic operators on half-line. Some explicit bounds of the eigenvalue depending only on the coefficients of the operators are presented. Moreover, the corresponding problems in the discrete case and the higher-order eigenvalues in the continuous case are also studied.

The Sturm-Liouville eigenvalue problem consists a classical topic in analytics, in which the first Dirichlet eigenvalue problem on finite intervals is the most essential and typical one. There is a great deal of publications on this problem (cf. ref. [1; Chapter 5] and references therein). As a complement of the classical theory and a continuation of refs. [2] and [3], the variational formulas and explicit bounds for the eigenvalue are presented in the paper, and the higher-order eigenvalues are studied at the end of the paper. Our investigation is split into two parts: the continuous case and the discrete one.

1. Continuous case.

Consider the differential operator $L = a(x)\mathrm{d}^2/\mathrm{d}x^2 + b(x)\mathrm{d}/\mathrm{d}x$ on a finite interval or half-line $(p, q)(-\infty < p < q \leqslant +\infty)$, where $a(x)$ is positive everywhere ($\lim_{x\to+\infty} a(x) > 0$ if $q = +\infty$), with Dirichlet boundary at p and q. We adopt the convention $f(+\infty) = \lim_{x\to+\infty} f(x)$ when $q = +\infty$. Throughout the paper, assume that

$$\int_p^q \frac{\mathrm{e}^{C(x)}}{a(x)}\mathrm{d}x < \infty \qquad \text{and} \qquad \int_p^q \mathrm{e}^{-C(x)}\mathrm{d}x < \infty, \tag{1.0}$$

Key words and phrases. First Dirichlet eigenvalue, variational formula, elliptic operator, birth-death process.

Research supported in part by 973 Project, NSFC(No. 10121101, 10101003) and RFDP

Typeset by $\mathcal{A}_\mathcal{M}\mathcal{S}$-TEX where $C(x) = \int_p^x b/a$. From probabilistic point of view, under the first condition in (1.0), the second condition is equivalent to that the process is transient. Otherwise, the process is positive recurrent. Then, the situation can be reduced to the one dealt in ref. [2] since the infinity boundary plays no role. See Remark 1.7 for more detailed analytic remark on (1.0). If $q < \infty$ and a is continuous, then these two conditions are trivial. The classical variational formula for the first Dirichlet eigenvalue of L is as follows:

$$\lambda_0 = \inf\{D(f) : f \in C^1(p,q) \cap C[p,q], f(p) = f(q) = 0, \pi(f^2) = 1\},$$

where

$$D(f) = \int_p^q a(x) f'(x)^2 \pi(\mathrm{d}x), \quad \pi(\mathrm{d}x) = (a(x)Z)^{-1} \mathrm{e}^{C(x)} \mathrm{d}x,$$

here and in what follows, Z denotes the normalizing constant (i.e. the first integration in (1.0)) and $\pi(g) = \int_p^q g(x)\pi(\mathrm{d}x)$.

The main idea of the study goes as follows. Due to the uniqueness of the extreme points for the eigenfunction g of λ_0 (see Proposition 1.3), the interval (p,q) can be divided into two sub-intervals according to the extreme point of g, which means that the study on λ_0 can be reduced to the mixed eigenvalue problem dealt in refs. [2] and [3] on sub-intervals. In detail, given $x_0 \in (p,q)$, consider the differential operator L on (p, x_0) with Dirichlet and Neumann boundary at p and x_0 respectively, and the operator L on (x_0, q) with Neumann and Dirichlet boundary on x_0 and q respectively. Corresponding π and D, define π_1, π_2 and D_1, D_2 on $[p, x_0]$ and $[x_0, q]$ respectively:

$$\pi_1(\mathrm{d}x) = \frac{\pi(\mathrm{d}x \cap [p, x_0])}{\pi[p, x_0]}, \qquad D_1(f) = \frac{\pi\big(af'^2 I_{[p,x_0]}\big)}{\pi[p, x_0]}, \quad f \in C^1(p, x_0),$$

$$\pi_2(\mathrm{d}x) = \frac{\pi(\mathrm{d}x \cap [x_0, q])}{\pi[x_0, q]}, \qquad D_2(f) = \frac{\pi\big(af'^2 I_{[x_0,q]}\big)}{\pi[x_0, q]}, \quad f \in C^1(x_0, q),$$

where $\pi[s,t] = \int_s^t \mathrm{d}\pi$. Recall the definition of the mixed eigenvalues:

$$\lambda_0[p, x_0] = \inf\{D_1(f) : f \in C^1(p, x_0) \cap C[p, x_0], f(p) = 0, \pi_1(f^2) = 1\},$$
$$\lambda_0[x_0, q] = \inf\{D_2(f) : f \in C^1(x_0, q) \cap C[x_0, q], f(q) = 0, \pi_2(f^2) = 1\}.$$

To state the main results, we need some notations. First, we need two operators (double integrals) which are defined on $C[p, x_0]$ and on $C[x_0, q]$ respectively:

$$\mathrm{II}_1(f)(x) = \frac{1}{f(x)} \int_p^x \mathrm{d}y\, \mathrm{e}^{-C(y)} \int_y^{x_0} [f\mathrm{e}^C/a](u)\mathrm{d}u, \quad x \in (p, x_0],$$
$$\mathrm{II}_2(f)(x) = \frac{1}{f(x)} \int_x^q \mathrm{d}y\, \mathrm{e}^{-C(y)} \int_{x_0}^y [f\mathrm{e}^C/a](u)\mathrm{d}u, \quad x \in [x_0, q).$$

Next, define

$$\mathscr{F}[p, x_0] = \{f \in C^1(p, x_0) \cap C[p, x_0] : f(p) = 0, f'|_{(p,x_0)} > 0\},$$
$$\xi_0'[p, x_0] = \inf_{f \in \mathscr{F}[p,x_0]} \sup_{x \in (p,x_0)} \mathrm{II}_1(f)(x)^{-1},$$
$$\xi_0''[p, x_0] = \sup_{f \in \mathscr{F}[p,x_0]} \inf_{x \in (p,x_0)} \mathrm{II}_1(f)(x)^{-1},$$

$$\mathscr{F}[x_0, q] = \{f \in C^1(x_0, q) \cap C[x_0, q] : f(q) = 0, f'|_{(x_0,q)} < 0\},$$
$$\xi_0'[x_0, q] = \inf_{f \in \mathscr{F}[x_0,q]} \sup_{x \in (x_0,q)} \Pi_2(f)(x)^{-1},$$
$$\xi_0''[x_0, q] = \sup_{f \in \mathscr{F}[x_0,q]} \inf_{x \in (x_0,q)} \Pi_2(f)(x)^{-1}.$$

As explained in ref. [3], ξ_0' is used for upper bounds and ξ_0'' is for lower bounds. The difference is that some sets larger than $\mathscr{F}$ are adopted in ref. [3] to guarantee the integrability but at the present situation, under (1.0), the set $\mathscr{F}$ is large enough. The upper and lower bounds are dual mutually in the following variational formulas, i.e. the one can be deduced from the another by exchanging "sup" with "inf" in the definition of ξ_0' and ξ_0'' and (1.1) below.

Theorem 1.1. The variational formulas for λ_0 are as follows.

$$\begin{aligned}&\inf_{x_0 \in (p,q)} \big(\xi_0'[p, x_0] \vee \xi_0'[x_0, q]\big) \geqslant \inf_{x_0 \in (p,q)} \big(\lambda_0[p, x_0] \vee \lambda_0[x_0, q]\big) \geqslant \lambda_0,\\ &\lambda_0 \geqslant \sup_{x_0 \in (p,q)} \big(\lambda_0[p, x_0] \wedge \lambda_0[x_0, q]\big) \geqslant \sup_{x_0 \in (p,q)} \big(\xi_0''[p, x_0] \wedge \xi_0''[x_0, q]\big).\end{aligned} \tag{1.1}$$

Moreover, the inequalities in (1.1) all become equalities once a and b are continuous on $[p, q]$. Furthermore, if the interval $[p, q]$ is finite and a is continuous, then the last assertion holds for all Lebesgue measurable function b.

Proof. The proof consists of four steps.

(i) Lower bounds. Let $f \in C^1(p, q) \cap C[p, q]$ with $f(p) = f(q) = 0$ and let $x_0 \in (p, q)$. Then, we have

$$\begin{aligned}D(f) &= \int_p^{x_0} [af'^2](x)\pi(\mathrm{d}x) + \int_{x_0}^q [af'^2](x)\pi(\mathrm{d}x)\\ &\geqslant \lambda_0[p, x_0] \int_p^{x_0} f(x)^2\pi(\mathrm{d}x) + \lambda_0[x_0, q] \int_{x_0}^q f(x)^2\pi(\mathrm{d}x)\\ &\geqslant \big(\lambda_0[p, x_0] \wedge \lambda_0[x_0, q]\big)\pi(f^2).\end{aligned}$$

So we get $\lambda_0 \geqslant \sup_{x_0 \in (p,q)} \big(\lambda_0[p, x_0] \wedge \lambda_0[x_0, q]\big)$ that is just the third inequality in (1.1). The fourth inequality follows from [3; Theorem 1.1] immediately.

(ii) Upper bounds. Let $\varepsilon > 0$. By definition, there exists an $\tilde{f} \in C^1(p, x_0) \cap C[p, x_0]$ with $\tilde{f}(p) = 0$ and $\pi_1(\tilde{f}^2) = 1$ such that $D_1(\tilde{f}) < \lambda_0[p, x_0] + \varepsilon$. If necessary, by modifying $\tilde{f}$ properly on a sufficiently small neighborhood of x_0, one may construct an $f \in C^1(p, x_0) \cap C[p, x_0]$ with $f(p) = 0$ and $\pi_1(f^2) = 1$ satisfying $f'(x_0) = 0$ and $D_1(f) < \lambda_0[p, x_0] + \varepsilon$. Similarly, there exists a $g \in C^1(x_0, q) \cap C[x_0, q]$ satisfying $g(q) = 0$, $g'(x_0) = 0$ and $\pi_2(g^2) = 1$ such that $D_2(g) < \lambda_0[x_0, q] + \varepsilon$. Furthermore, we can assume that $f(x_0) \neq 0$. Otherwise, modify f as follows. Since $f \in C[p, x_0]$, f takes its maximum and minimum at some point x_1 and x_2 respectively on $[p, x_0]$. Without loss of generality, assume that $|f(x_1)| \geqslant |f(x_2)|$. Then $f(x_1) \neq 0$ (otherwise, $f = 0$ which contradicts with

$\pi_1(f^2)=1)$. Let $\bar f = fI_{[p,x_1)} + f(x_1)I_{[x_1,x_0]}$. Then $\bar f \in C^1(p,x_0)\cap C[p,x_0]$ and

$$\pi_1(\bar f^2) = \pi_1(f^2 I_{[p,x_1)}) + f(x_1)^2\pi_1[x_1,x_0] \geqslant \pi_1(f^2)=1,$$
$$D_1(\bar f) = \int_p^{x_1} af'^2 \mathrm{d}\pi_1 \leqslant D_1(f).$$

Set $\hat f = \pi_1(\bar f^2)^{-1/2}\bar f$. Now it follows that

$$\hat f(p)=0, \quad \hat f'(x_0)=0, \quad \pi_1(\hat f^2)=1, \quad \hat f(x_0)\neq 0$$

and

$$D_1(\hat f) = \pi_1(\bar f^2)^{-1} D_1(\bar f) \leqslant D_1(f) < \lambda_0[p,x_0]+\varepsilon.$$

Hence, we can replace f by $\hat f$ when $f(x_0)=0$.

Let $h = cfI_{[p,x_0]} + gI_{(x_0,q]}$, where $c = g(x_0)/f(x_0)$. Then $h\in C^1(p,q)\cap C[p,q]$, $h(p)=h(q)=0$ and

$$\begin{aligned}
\pi(ah'^2) &= c^2\int_p^{x_0} af'^2\mathrm{d}\pi + \int_{x_0}^q ag'^2\mathrm{d}\pi\\
&= c^2 D_1(f)\pi[p,x_0] + D_2(g)\pi[x_0,q]\\
&< c^2\big(\lambda_0[p,x_0]+\varepsilon\big)\pi[p,x_0] + \big(\lambda_0[x_0,q]+\varepsilon\big)\pi[x_0,q]\\
&\leqslant \big(\lambda_0[p,x_0]\vee\lambda_0[x_0,q]+\varepsilon\big)\big(c^2\pi[p,x_0]+\pi[x_0,q]\big),\\
\pi(h^2) &= c^2\pi\big(f^2 I_{[p,x_0]}\big) + \pi\big(g^2 I_{(x_0,q]}\big)\\
&= c^2\pi_1(f^2)\pi[p,x_0] + \pi_2(g^2)\pi[x_0,q]\\
&= c^2\pi[p,x_0]+\pi[x_0,q].
\end{aligned}$$

Hence, it follows that $\lambda_0 < \lambda_0[p,x_0]\vee\lambda_0[x_0,q]+\varepsilon$. Since ε is arbitrary, we get $\lambda_0 \leqslant \lambda_0[p,x_0]\vee\lambda_0[x_0,q]$, which implies the second inequality in (1.1). The first inequality in (1.1) follows from [3; Theorem 1.1] directly.

(iii) To prove the equalities in (1.1) for continuous a and b on $[p,q]$, we need the following Proposition 1.3 that is the main credit of Theorem 1.1. Note that (6.1) and Lemma 6.3 in ref. [4] still hold. For convenience, we copy them as (1.2) and Lemma 1.2 below.

$$(f'\mathrm{e}^C)' = (af''+bf')\mathrm{e}^C/a = (Lf)\mathrm{e}^C/a. \tag{1.2}$$

Lemma 1.2. Let $Lf=-\lambda f$ for some $f\in C^2[p,q]$ and $\lambda\geqslant 0$. If there exist α and β with $p\leqslant\alpha<\beta\leqslant q$ such that $f=0$ on $[\alpha,\beta]$, then $f=0$ on $[p,q]$.

Proposition 1.3. Let $f\geqslant 0$ be a non-constant solution to the equation $Lf=-\lambda_0 f$ with $f\in C^2[p,q]$ and $f(p)=f(q)=0$, then $\lambda_0>0$ and there exists an unique extreme point of f on (p,q).

Proof. (i) First, assume that $\lambda_0=0$. Then we have

$$\pi(af'^2) = -\int_p^q fLf\mathrm{d}\pi = 0.$$

Furthermore, from $a > 0$, it follows that $f' = 0$. By this and the fact that $f(p) = f(q) = 0$, we have $f = 0$ which means f is a constant solution. This is a contradiction with the assumptions. Hence, we must have $\lambda_0 > 0$.

(ii) Next, it is obvious that there exist extreme points of f. We will prove that there exists an unique extreme point of f on (p, q). Assume that x_1 and x_2 both are extreme points of f on (p, q) with $x_1 < x_2$. From this, a conclusion contradicted with the minimum of λ_0 will be deduced.

(a) At first, we claim that $f \neq$ constant on $[p, x_1]$. Otherwise, we have $f = -\lambda_0^{-1} Lf = 0$ on $[p, x_1]$ which implies that $f = 0$ by Lemma 1.2.

(b) Next, we prove that $f(x_1) \neq 0$. To this end, let

$$g = fI_{[p,x_1]} + f(x_1)I_{(x_1,q]}.$$

If $f(x_1) = 0$, combining it with $f'(x_1) = 0$ and $a > 0$, from the equation, we derive that $f''(x_1) = 0$; furthermore, $g \in C^2[p, q]$ satisfies $Lg = -\lambda_0 g$ and $g = 0$ on $[x_1, q]$. By Lemma 1.2, we have $g = 0$ and in particular $f = 0$ on $[p, x_1]$. This implies that $f = 0$ by Lemma 1.2 again.

(c) Similarly, we can prove that $f \neq$ constant on $[x_2, q]$ and $f(x_2) \neq 0$. Finally, let

$$h = cfI_{[p,x_1]} + f(x_2)I_{(x_1,x_2]} + fI_{(x_2,q]},$$

where $c = f(x_2)/f(x_1)$. Then $h \neq$ constant and $h \in C^1(p, q) \cap C[p, q]$ with $h(p) = h(q) = 0$. From (1.2), it follows that

$$\begin{aligned}
\pi(ah'^2) &= c^2 \int_p^{x_1} af'^2 \mathrm{d}\pi + \int_{x_2}^q af'^2 \mathrm{d}\pi \\
&= -c^2 \int_p^{x_1} fLf \mathrm{d}\pi - \int_{x_2}^q fLf \mathrm{d}\pi \\
&= \lambda_0 \Big(c^2 \pi\big(f^2 I_{[p,x_1]}\big) + \pi\big(f^2 I_{(x_2,q]}\big) \Big) \\
\pi(h^2) &= c^2 \pi\big(f^2 I_{[p,x_1]}\big) + f(x_2)^2 \pi[x_1, x_2] + \pi\big(f^2 I_{(x_2,q]}\big).
\end{aligned}$$

Therefore,

$$\lambda_0 \leqslant \frac{\pi(ah'^2)}{\pi(h^2)} = \frac{\lambda_0 \Big(c^2 \pi\big(f^2 I_{[p,x_1]}\big) + \pi\big(f^2 I_{(x_2,q]}\big) \Big)}{c^2 \pi\big(f^2 I_{[p,x_1]}\big) + f(x_2)^2 \pi[x_1, x_2] + \pi\big(f^2 I_{(x_2,q]}\big)} < \lambda_0.$$

This is a contradiction. □

We now return the proof of the theorem. Let a, b be continuous and $a > 0$. By the existence theorem of solution to the Sturm-Liouville eigenvalue problem, it follows that there is a non-trivial solution f to the equation $Lf = -\lambda_0 f$ with $f(p) = f(q) = 0$. From Proposition 1.3, it follows that $\lambda_0 > 0$ and there exists an unique extreme point of f on (p, q). Let x_0 denote the unique extreme point. Then, $f_1 = fI_{[p,x_0]} \in \mathscr{F}[p, x_0]$ and $f_2 = fI_{[x_0,q]} \in \mathscr{F}[x_0, q]$. Noting that $f'(x_0) = 0$ and (1.2), we have

$$-\int_x^{x_0} \lambda_0 f \mathrm{e}^C / a = (f' \mathrm{e}^C)\big|_x^{x_0} = -(f' \mathrm{e}^C)(x).$$

Hence, $\Pi_1(f_1)(x) = \lambda_0^{-1}$ $(x \in (p, x_0))$ and $\Pi_2(f_2)(x) = \lambda_0^{-1}$ $(x \in (x_0, q))$. From ref. [3; Theorem 1.1], it follows that

$$\xi_0'[p, x_0] = \lambda_0[p, x_0] = \xi_0''[p, x_0].$$

In a similar way, we get

$$\xi_0'[x_0, q] = \lambda_0[x_0, q] = \xi_0''[x_0, q].$$

Collecting up these facts together, we see that all four inequalities in (1.1) become equalities.

(iv) To prove the last assertion, we need the following proposition.

Proposition 1.4. Let a be positive, continuous and b be Lebesgue measurable on the finite interval $[p, q]$. Denote by λ_0 the first eigenvalue determined by a and b. Then there exist two sequences $\{a_n\}$ and $\{b_n\}$ of continuous functions such that b_n converges to b pointwise, a_n and C_n converge respectively to a and C as $n \to \infty$, uniformly on $[p, q]$, and the corresponding first eigenvalue $\lambda_0^{(n)}$ converges to λ_0 as $n \to \infty$.

Proof. By assumptions, there exist two positive constants δ and N such that $\delta \leqslant a \leqslant N$. We now adopt the smooth approximation. For convenience, we extend a and C to the full line $\mathbf{R}$: $\widetilde{C}(x) = C(x \vee p \wedge q)$ and $\tilde{a}(x) = a(x \vee p \wedge q)$. For simplicity, omit the superscript "$\sim$". As usual, let

$$\eta(x) = \begin{cases} A \exp\big[-\big((q-p)^2/4 - (x-(p+q)/2)^2\big)^{-1}\big], & \text{if } |x-(p+q)/2| < 1; \\ 0, & \text{if } |x-(p+q)/2| \geqslant 1, \end{cases}$$

where A is the normalizing constant: $\int_{\mathbf{R}} \eta(x)\mathrm{d}x = 1$. First, set $\eta_\varepsilon(x) = \varepsilon^{-1}\eta(x/\varepsilon)$, $C_\varepsilon = C * \eta_\varepsilon$ and $a_\varepsilon = a * \eta_\varepsilon$ (i.e. $C_\varepsilon(x) = \int_{\mathbf{R}} C(y)\eta_\varepsilon(x-y)\mathrm{d}y$ and so on). Then, we have C_ε, $a_\varepsilon \in C^\infty(\mathbf{R})$ and

$$C_\varepsilon \to C, \quad a_\varepsilon \to a \quad \text{and} \quad \delta \leqslant a_\varepsilon \leqslant N \tag{1.3}$$

as $\varepsilon \downarrow 0$, uniformly on compact sets. By (1.3), there exists a sequence $\varepsilon(n) \to 0$ such that

$$C_{\varepsilon(n)} \to C \qquad \text{and} \qquad a_{\varepsilon(n)} \to a \tag{1.4}$$

as $n \to \infty$, uniformly on $[p, q]$. Next, let $b_\varepsilon = a_\varepsilon C_\varepsilon'$. Then we have a_ε, $b_\varepsilon \in C^\infty(\mathbf{R})$ and

$$Z_\varepsilon = \int_p^q \mathrm{e}^{C_\varepsilon}/a_\varepsilon \leqslant \int_p^q \mathrm{e}^{C_\varepsilon}/\delta < \infty.$$

From (1.3), it follows that

$$\int_p^q \frac{f^2\mathrm{e}^{C_{\varepsilon(n)}}}{a_{\varepsilon(n)}} \leqslant \frac{N}{\delta} \sup_{p \leqslant x \leqslant q} \mathrm{e}^{C_{\varepsilon(n)}(x) - C(x)} \int_p^q \frac{f^2\mathrm{e}^C}{a}.$$

Hence, we have $L^2(\pi) \subset L^2(\pi_{\varepsilon(n)})$. Similarly, we have $L^2(\pi) \supset L^2(\pi_{\varepsilon(n)})$. Hence, $L^2(\pi) = L^2(\pi_{\varepsilon(n)})$. Note that

$$\inf_{p\leqslant x\leqslant q} \mathrm{e}^{C_{\varepsilon(n)}(x)-C(x)} \frac{a(x)}{a_{\varepsilon(n)}} \leqslant \frac{\int_p^q f^2 \mathrm{e}^{C_{\varepsilon(n)}}/a_{\varepsilon(n)}}{\int_p^q f^2 \mathrm{e}^C/a} \leqslant \sup_{p\leqslant x\leqslant q} \mathrm{e}^{C_{\varepsilon(n)}(x)-C(x)} \frac{a(x)}{a_{\varepsilon(n)}}$$

and

$$\inf_{p\leqslant x\leqslant q} \mathrm{e}^{C_{\varepsilon(n)}(x)-C(x)} \leqslant \frac{\int_p^q f'^2 \mathrm{e}^{C_{\varepsilon(n)}}}{\int_p^q f'^2 \mathrm{e}^C} \leqslant \sup_{p\leqslant x\leqslant q} \mathrm{e}^{C_{\varepsilon(n)}(x)-C(x)}.$$

By these facts, (1.3) and (1.4), it follows that

$$\frac{\int_p^q f^2 \mathrm{e}^{C_{\varepsilon(n)}}/a_{\varepsilon(n)}}{\int_p^q f^2 \mathrm{e}^C/a} \to 1 \quad \text{and} \quad \frac{\int_p^q f'^2 \mathrm{e}^{C_{\varepsilon(n)}}}{\int_p^q f'^2 \mathrm{e}^C} \to 1$$

as $n \to \infty$, uniformly with respect to f. Hence,

$$\frac{D_{\varepsilon(n)}(f)}{\pi_{\varepsilon(n)}(f^2)}\Big/\frac{D(f)}{\pi(f^2)} = \frac{Z_{\varepsilon(n)}^{-1}\int_p^q f'^2 \mathrm{e}^{C_{\varepsilon(n)}}}{Z_{\varepsilon(n)}^{-1}\int_p^q f^2 \mathrm{e}^{C_{\varepsilon(n)}}/a_{\varepsilon(n)}}\Big/\frac{Z^{-1}\int_p^q f'^2 \mathrm{e}^C}{Z^{-1}\int_p^q f^2 \mathrm{e}^C/a} \to 1$$

as $n \to \infty$, uniformly with respect to f. Having this in mind and noting that $\{a_{\varepsilon(n)}\}$, $\{b_{\varepsilon(n)}\}$ are continuous functions, it is easy to prove that $\lim_{n\to\infty} \lambda_0^{(n)} = \lambda_0$. $\square$

It is the position to prove the last assertion of the theorem. Let a be positive, continuous and b be Lebesgue measurable on the finite interval $[p, q]$. By Proposition 1.4, there exist two sequences of continuous functions $\{a_n\}$ and $\{b_n\}$ such that $a_n \to a$, $b_n \to b$ and the corresponding first eigenvalue $\lambda_0^{(n)} \to \lambda_0$ as $n \to \infty$. Define $\Pi_i^{(n)}(f)$, $\xi_0'^{(n)}$ and $\xi_0''^{(n)}$ correspondingly. By the continuity of a_n and b_n, the equalities in the corresponding (1.1) hold, i.e.

$$\inf_{x_0\in(p,q)} \big(\xi_0'^{(n)}[p, x_0] \vee \xi_0'^{(n)}[x_0, q]\big) = \lambda_0^{(n)} = \sup_{x_0\in(p,q)} \big(\xi_0''^{(n)}[p, x_0] \wedge \xi_0''^{(n)}[x_0, q]\big). \tag{1.5}$$

Using (1.3), (1.4) and

$$\begin{aligned}
&\inf_{x\in(p,q)} \mathrm{e}^{C(x)-C_n(x)} \frac{a_n(x)}{a(x)} \inf_{x\in(p,q)} \mathrm{e}^{C(x)-C_n(x)} \\
&\quad\leqslant \frac{\Pi_i^{(n)}(f)(x)^{-1}}{\Pi_i(f)(x)^{-1}} \\
&\quad\leqslant \sup_{x\in(p,q)} \mathrm{e}^{C(x)-C_n(x)} \frac{a_n(x)}{a(x)} \sup_{x\in(p,q)} \mathrm{e}^{C(x)-C_n(x)},
\end{aligned}$$

we obtain

$$\frac{\Pi_i^{(n)}(f)(x)^{-1}}{\Pi_i(f)(x)^{-1}} \to 1, \qquad i = 1, 2$$

as $n \to \infty$, uniformly with respect to x, f and x_0. From these facts and the following inequalities

$$\inf_{f\in\mathscr{F}[p,x_0]}\ \inf_{x\in(p,x_0)} \frac{\Pi_1^{(n)}(f)(x)^{-1}}{\Pi_1(f)(x)^{-1}} \leqslant \frac{\xi_0'^{(n)}[p,x_0]}{\xi_0'[p,x_0]} \leqslant \sup_{f\in\mathscr{F}[p,x_0]}\ \sup_{x\in(p,x_0)} \frac{\Pi_1^{(n)}(f)(x)^{-1}}{\Pi_1(f)(x)^{-1}},$$
$$\inf_{f\in\mathscr{F}[x_0,q]}\ \inf_{x\in(x_0,q)} \frac{\Pi_2^{(n)}(f)(x)^{-1}}{\Pi_2(f)(x)^{-1}} \leqslant \frac{\xi_0'^{(n)}[x_0,q]}{\xi_0'[x_0,q]} \leqslant \sup_{f\in\mathscr{F}[x_0,q]}\ \sup_{x\in(x_0,q)} \frac{\Pi_2^{(n)}(f)(x)^{-1}}{\Pi_2(f)(x)^{-1}},$$

it follows that

$$\frac{\xi_0'^{(n)}[p,x_0]}{\xi_0'[p,x_0]} \to 1 \qquad \text{and} \qquad \frac{\xi_0'^{(n)}[x_0,q]}{\xi_0'[x_0,q]} \to 1$$

as $n \to \infty$, uniformly with respect to x_0. Note that

$$\begin{aligned}\inf_{x_0\in(p,q)} \left(\frac{\xi_0'^{(n)}[p,x_0]}{\xi_0'[p,x_0]} \wedge \frac{\xi_0'^{(n)}[x_0,q]}{\xi_0'[x_0,q]}\right) &\leqslant \frac{\inf_{x_0\in(p,q)}\big(\xi_0'^{(n)}[p,x_0] \vee \xi_0'^{(n)}[x_0,q]\big)}{\inf_{x_0\in(p,q)}\big(\xi_0'[p,x_0] \vee \xi_0'[x_0,q]\big)}\\ &\leqslant \sup_{x_0\in(p,q)} \left(\frac{\xi_0'^{(n)}[p,x_0]}{\xi_0'[p,x_0]} \vee \frac{\xi_0'^{(n)}[x_0,q]}{\xi_0'[x_0,q]}\right).\end{aligned}$$

Collecting these facts together, we obtain

$$\inf_{x_0\in(p,q)} \big(\xi_0'^{(n)}[p,x_0] \vee \xi_0'^{(n)}[x_0,q]\big) \to \inf_{x_0\in(p,q)} \big(\xi_0'[p,x_0] \vee \xi_0'[x_0,q]\big)$$

as $n \to \infty$. In a similar way, one shows that

$$\sup_{x_0\in(p,q)} \big(\xi_0''^{(n)}[p,x_0] \wedge \xi_0''^{(n)}[x_0,q]\big) \to \sup_{x_0\in(p,q)} \big(\xi_0''[p,x_0] \wedge \xi_0''[x_0,q]\big).$$

Now, the last assertion follows immediately from (1.5) and the fact that $\lambda_0^{(n)} \to \lambda_0$. $\square$

By examining the proof above, it is not difficult to see that the variational formulas with single integrals and explicit bounds in ref. [2], and the iteration method in ref. [3] can all be applied to the present situation. Here we discuss only the explicit bounds of λ_0. Define

$$\begin{aligned}
&Q_1(x) = \int_p^x \mathrm{e}^{-C(y)}\mathrm{d}y \int_x^c [\mathrm{e}^C/a](u)\mathrm{d}u,\\
&\delta_1(c) = \sup_{x\in(p,c)} Q_1(x), \qquad \delta_1'(c) = 2\sup_{x\in(p,c)} \int_p^x Q_1 \mathrm{d}\nu_1^{(x)},\\
&Q_2(x) = \int_x^q \mathrm{e}^{-C(y)}\mathrm{d}y \int_c^x [\mathrm{e}^C/a](u)\mathrm{d}u,\\
&\delta_2(c) = \sup_{x\in(c,q)} Q_2(x), \qquad \delta_2'(c) = 2\sup_{x\in(c,q)} \int_x^q Q_2 \mathrm{d}\nu_2^{(x)},
\end{aligned}$$

where $\nu_1^{(x)}$ and $\nu_2^{(x)}$ are probability measures on (p,x) and (x,q) with density $\mathrm{e}^{-C(y)}/Z^{(x)}$ respectively ($Z^{(x)}$ is the normalizing constant). Let

$$f_1(x)=\sqrt{\int_p^x \mathrm{e}^{-C(y)}\mathrm{d}y} \quad \text{and} \quad f_2(x)=\sqrt{\int_x^q \mathrm{e}^{-C(y)}\mathrm{d}y}.$$

Replacing x_0 with c in $\Pi_1(f)(x)$ and $\Pi_2(f)(x)$ defined at the beginning of this section and define

$$\delta_1''(c)=\sup_{x\in(p,c)}\Pi_1(f_1)(x) \qquad \text{and} \qquad \delta_2''(c)=\sup_{x\in(c,q)}\Pi_2(f_2)(x).$$

Corollary 1.5. The explicit bounds of λ_0 are given as follows.

$$\begin{aligned}&\Big(\sup_{c\in(p,q)}\delta_1(c)\wedge\delta_2(c)\Big)^{-1}\geqslant\Big(\sup_{c\in(p,q)}\delta_1'(c)\wedge\delta_2'(c)\Big)^{-1}\geqslant\lambda_0,\\ &\lambda_0\geqslant\Big(\inf_{c\in(p,q)}\delta_1''(c)\vee\delta_2''(c)\Big)^{-1}\geqslant\Big(4\inf_{c\in(p,q)}\delta_1(c)\vee\delta_2(c)\Big)^{-1}. \qquad (1.6)\end{aligned}$$

Furthermore,

$$\delta_1(c_0)^{-1}\geqslant\lambda_0\geqslant\big(4\delta_1(c_0)\big)^{-1}, \qquad (1.7)$$

where c_0 is the unique solution of the equation $\delta_1(c)=\delta_2(c)$ on (p,q).

Proof. From refs. [2; Theorem 1.1] and [3; Theorem 1.2], it follows that

$$\delta_1'(c)^{-1}\geqslant\lambda_0[p,c]\geqslant\delta_1''(c)^{-1}\geqslant(4\delta_1(c))^{-1}$$

and

$$\delta_1(c)\leqslant\delta_1'(c)\leqslant 2\delta_1(c).$$

Similarly, we have

$$\delta_2'(c)^{-1}\geqslant\lambda_0[c,q]\geqslant\delta_2''(c)^{-1}\geqslant(4\delta_2(c))^{-1}$$

and

$$\delta_2(c)\leqslant\delta_2'(c)\leqslant 2\delta_2(c).$$

Then (1.6) follows from Theorem 1.1 immediately.

To prove the last assertion, note that $\delta_1(c)$ and $\delta_2(c)$ are strictly increasing and strictly decreasing in c respectively. Obviously, $\lim_{c\to p}\delta_1(c)=0$, $\lim_{c\to q}\delta_2(c)=0$, $\delta_1(c)>0$ and $\delta_2(c)>0$ for all $c\in(p,q)$. Moreover, when $c_1<c_2$, we have

$$\begin{aligned}&0<\int_p^x\mathrm{e}^{-C}\int_x^{c_2}\mathrm{e}^{C}/a-\int_p^x\mathrm{e}^{-C}\int_x^{c_1}\mathrm{e}^{C}/a\leqslant\int_p^{c_2}\mathrm{e}^{-C}\int_{c_1}^{c_2}\mathrm{e}^{C}/a\to 0,\\ &\quad\text{as } c_2-c_1\to 0;\\ &0<\int_x^q\mathrm{e}^{-C}\int_{c_1}^{x}\mathrm{e}^{C}/a-\int_x^q\mathrm{e}^{-C}\int_{c_2}^{x}\mathrm{e}^{C}/a\leqslant\int_{c_1}^{q}\mathrm{e}^{-C}\int_{c_1}^{c_2}\mathrm{e}^{C}/a\to 0,\\ &\quad\text{as } c_2-c_1\to 0.\end{aligned}$$

So $\delta_1(c)$ and $\delta_2(c)$ are both continuous in c. Hence, there exists uniquely a solution to the equation $\delta_1(c) = \delta_2(c)$. Combining these with (1.6) and the monotonicity of $\delta_1(c)$ and $\delta_2(c)$, we obtain (1.7). $\square$

Remark 1.6. P. Gurka's explicit bounds of λ_0 are presented in ref [5; Page 93][1]:

$$2B^{-1} \geqslant \lambda_0 \geqslant (4B)^{-1}, \tag{1.8}$$

where

$$B = \sup_{p<c<d<q} \gamma(c,d) \quad \text{and}$$

$$\gamma(c,d) = \Big(\int_p^c \mathrm{e}^{-C(y)}\mathrm{d}y \bigwedge \int_d^q \mathrm{e}^{-C(y)}\mathrm{d}y\Big) \int_c^d [\mathrm{e}^C/a](u)\mathrm{d}u. \tag{1.9}$$

We now shows that the bounds in Corollary 1.5 are indeed sharper than those in (1.8). For this, it suffices to prove that $B/2 \leqslant \delta_1(c_0) \leqslant B$.

First, since $\delta_1(x)$ and $\delta_2(x)$ are strictly increasing and strictly decreasing in x respectively, we see that

$$\gamma(c,d) \leqslant \delta_1(d) \wedge \delta_2(c) \leqslant \delta_1(d) \wedge \delta_2(c_0) = \delta_1(c_0) \ \text{ if } \ c_0 \leqslant c < d < q,$$

$$\gamma(c,d) \leqslant \delta_1(d) \wedge \delta_2(c) \leqslant \delta_1(c_0) \wedge \delta_2(c) = \delta_1(c_0) \ \text{ if } \ p < c < d \leqslant c_0,$$

and

$$\begin{aligned}\frac{1}{2}\gamma(c,d) &\leqslant \Big(\int_p^c \mathrm{e}^{-C} \bigwedge \int_d^q \mathrm{e}^{-C}\Big)\Big(\int_c^{c_0} \mathrm{e}^C/a \bigvee \int_{c_0}^d \mathrm{e}^C/a\Big)\\ &\leqslant \Big(\int_p^c \mathrm{e}^{-C} \int_c^{c_0} \mathrm{e}^C/a\Big) \bigvee \Big(\int_d^q \mathrm{e}^{-C} \int_{c_0}^d \mathrm{e}^C/a\Big)\\ &\leqslant \delta_1(c_0) \vee \delta_2(c_0) = \delta_1(c_0) \ \text{ if } \ p < c < c_0 < d < q.\end{aligned}$$

Collecting these facts together, we obtain $B/2 \leqslant \delta_1(c_0)$.

Secondly, to prove the upper bound, we adopt the reductio ad absurdum proof. Suppose that $\delta_1(c_0) > B$. If $\int_p^{c_0} \mathrm{e}^{-C} \leqslant \int_{c_0}^q \mathrm{e}^{-C}$, then there exists $x_0 \in (p, c_0)$ such that $\int_p^{x_0} \mathrm{e}^{-C} \int_{x_0}^{c_0} \mathrm{e}^C/a > B$. However, this contradicts with

$$B \geqslant \gamma(x_0, c_0) \geqslant \int_p^{x_0} \mathrm{e}^{-C} \int_{x_0}^{c_0} \mathrm{e}^C/a.$$

If $\int_p^{c_0} \mathrm{e}^{-C} > \int_{c_0}^q \mathrm{e}^{-C}$, due to the assumption that $\delta_2(c_0) = \delta_1(c_0) > B$, then there exists $x_1 \in (c_0, q)$ such that $\int_{x_1}^q \mathrm{e}^{-C} \int_{c_0}^{x_1} \mathrm{e}^C/a > B$ which contradicts with

$$B \geqslant \gamma(c_0, x_1) \geqslant \int_{x_1}^q \mathrm{e}^{-C} \int_{c_0}^{x_1} \mathrm{e}^C/a.$$

[1]There is a citation error in (1.8) and (1.10) below: λ_0 should be replaced by $\sqrt{\lambda_0}$. Hence (1.8) reads as $4B^{-1} \geqslant \lambda_0 \geqslant (16B)^{-1}$. The corresponding change is needed in Example 1.9.

Hence, we always have $\delta_1(c_0) \leqslant B$. $\square$

Remark 1.7. The estimates (1.8) presented in ref. [5] do not require condition (1.0) and work for general intervals $(p, q) \subset \mathbf{R}$. From analytic point of view, if the first condition in (1.0) is satisfied but the second one fails, noting that $p > -\infty$, then we have

$$B = \sup_{x \in (p,\infty)} \int_p^x e^{-C(y)} \mathrm{d}y \int_x^\infty [e^C/a](u) \mathrm{d}u,$$

which coincides with the constant δ introduced in ref. [2; Theorem 1.1]. So this belongs to the cases dealt in ref. [2]. We can deal with the case of $p = -\infty$ similarly: given $x_0 \in (p, q)$ and redefine the function C as $C(x) = \int_{x_0}^x b/a$. Let

$$M_\pm = \pm \int_{x_0}^{\pm\infty} e^{-C(y)} \mathrm{d}y.$$

If one of M_+ and M_- is finite, $M_- < \infty$ for example, then this can be similarly included into ref. [2] as the case of $p > -\infty$; if both are infinite, then by (1.8), it follows that $B = \infty$ which is just the trivial case of $\lambda_0 = 0$.

If the first condition in (1.0) is not satisfied, we discuss only the situation that $p = -\infty$ and $q = +\infty$ since other situations can be dealt with analogically. If M_+ and M_- both are infinite, then it is the trivial case by (1.8). If one of them is infinite, then it can be included into refs. [2] and [5] (Let $M_- < \infty$ and $M_+ = \infty$ for instance. Using (1.8) and imitating the discussion above, we can reduce it to the single boundary case. At the same time, if $\int_{x_0}^\infty e^C/a < \infty$, then it was discussed above. Otherwise, by ref. [5; Theorem 1.14], it follows that $\lambda_0 = 0$). If both are infinite, then either it is in the trivial case or we get $B < \infty$ which is the only case not handled in the paper.

Remark 1.8. In fact, ref. [5; Theorem 8.2] presents the more exact bounds than those in (1.8)[2]:

$$\sqrt{2}B^{-1} \geqslant \lambda_0 \geqslant 4\sqrt{2}(\sqrt{5}+1)^{-5/2}B^{-1} (\approx (3.33B)^{-1}), \tag{1.10}$$

where B is the same as in (1.9). It is regretted that we do not know how to compare (1.10) with (1.7) directly[3]. Instead, we present the following example as an illustration.

---

[2] As mentioned in the last footnote, the formula (1.10) becomes

$$2B^{-1} \geqslant \lambda_0 \geqslant \omega^{-5}B^{-1} \approx 0.09B^{-1},$$

where ω is the golden ratio $(\sqrt{5}+1)/2$ and so $\omega^5 = 3 + 5\omega$. Based on a splitting technique as used here, a better result was given in [5; Theorem 8.8]:

$$2\widetilde{B} \geqslant \lambda_0 \geqslant \widetilde{B}/4$$

for some constant $\widetilde{B} \leqslant \infty$. The comparison with this result should be clear.

[3] Having the correction, it is clear that (1.7) improves (1.10).

Example 1.9. Let $a(x) \equiv 1$ and $b(x) \equiv 0$ on $[0,1]$. This is the simplest case. By (1.8), (1.10), (1.7) and (1.6), we obtain respectively the bounds: $2 \leqslant \lambda_0 \leqslant 16$,[4] $2.4023 \leqslant \lambda_0 \leqslant 11.3137$,[5] $4 \leqslant \lambda_0 \leqslant 16$ and $16/\sqrt[3]{5}\,(\approx 9.3569) \leqslant \lambda_0 \leqslant 32/3\,(\approx 10.6667)$. Given $x_0 = 1/2$, set the test function $f_1(x) = x$ if $x \in [0,1/2]$, $f_1(x) = 1-x$ if $x \in [1/2,1]$, then we have $8 \leqslant \lambda_0 \leqslant 12$ by Theorem 1.1. Next, let

$$f_n(x) = \begin{cases} f_{n-1}(x)\Pi_1(f_{n-1})(x) & \text{if } x \in [0,1/2], \\ f_n(x) = f_{n-1}(x)\Pi_2(f_{n-1})(x) & \text{if } x \in [1/2,1]. \end{cases}$$

Then as $n = 2,3,4$, we obtain respectively the estimates: $9.6 \leqslant \lambda_0 \leqslant 10$, $9.8361 \leqslant \lambda_0 \leqslant 9.9188$ and $9.8657 \leqslant \lambda_0 \leqslant 9.8710$. Here the choosing of test functions is just the iteration method stated in ref. [4]. The exact value is $\lambda_0 = \pi^2 \approx 9.8696$ with eigenfunction $f(x) = \sin(\pi x)$.

Example 1.10. Let $a(x) = x/2$ and $b(x) = -x$ on $[0,1]$. This is non-trivial because of the variable coefficients. Given $x_0 = 1/2$, let the test function

$$f(x) = \begin{cases} x(1-x)(n-1+x) & \text{if } x \in [0,1/2], \\ x(1-x)(n-x) & \text{if } x \in [1/2,1]. \end{cases}$$

By Theorem 1.1, it follows that

$$\frac{4n-4}{2n-4+\mathrm{e}} \leqslant \lambda_0 \leqslant \frac{6n\mathrm{e}-3\mathrm{e}}{3n\mathrm{e}-\mathrm{e}-3}.$$

In particular, when $n = 2,10,50,100$, we get respectively the bounds: $1.4715 \leqslant \lambda_0 \leqslant 2.3099$, $1.9233 \leqslant \lambda_0 \leqslant 2.0433$, $1.9855 \leqslant \lambda_0 \leqslant 2.0082$ and $1.9928 \leqslant \lambda_0 \leqslant 2.0041$. Here the test functions consist an approximation of the eigenfunction. The exact value of λ_0 equals 2 and the corresponding eigenfunction $f(x)$ is $x - x^2$.

2. Discrete case.

Consider the birth-death processes on $E = \{0,1,2,\cdots,N\}$ $(3 \leqslant N \leqslant \infty)$ (The case of $N = 2$ is trivial since it is easy to get $\lambda_0 = a_1 + b_1$). Let $b_i > 0\,(0 \leqslant i < N)$ and $a_i > 0\,(0 < i \leqslant N)$ be the birth and death rates respectively. For convenience, let $a_0 = 0$ and $b_N = 0$. Define $\mu_0 = 1$,

$$\mu_i = \frac{b_0 \cdots b_{i-1}}{a_1 \cdots a_i}, \qquad 1 \leqslant i \leqslant N,$$

$\mu = \sum_{i=0}^N \mu_i$ and $\pi_i = \mu_i/\mu$. We adopt the convention $f_\infty = \lim_{n\to+\infty} f_n$. Assume that[6]

$$\mu < \infty \qquad \text{and} \qquad \sum_{i=0}^N \frac{1}{\mu_i b_i} < \infty. \tag{2.0}$$

[4] Correction: $1/2 \leqslant \lambda_0 \leqslant 32$.

[5] Correction: $0.72 \leqslant \lambda_0 \leqslant 16$.

[6] Under hypothesis (2.0), it is essentially in the case with finite state space. Otherwise, the Q-processes are not unique. For infinite state space, since we are interested in the double Dirichlet boundaries, it follows naturally that $\mu = \infty$. Then, much work is required as shown in a subsequent paper.

The operator Ω is defined by

$$\Omega f(i) = b_i(f_{i+1} - f_i) + a_i(f_{i-1} - f_i).$$

The corresponding Dirichlet form is

$$D(f) = \sum_{i=0}^{N-1} \pi_i b_i [f_{i+1} - f_i]^2 = -\sum_{i=0}^{N} \pi_i (f\Omega f)(i).$$

Consider the principal eigenvalue of Ω:

$$\lambda_0 = \inf\{D(f) : f_0 = f_N = 0, \pi(f^2) = 1\},$$

where $\pi(f^2) = \sum_{i=0}^{N} \pi_i f_i^2$. Enlightened by the continuous case, it is hopeful to reduce the study of λ_0 to the mixed eigenvalue problem on sub-intervals, but in the discrete case, the corresponding eigenfunction g of λ_0 has some different modality (cf. Proposition 2.4), and then the final variational formulas are a little different from those in the continuous case. To state the results, we need some notations. First, define

$$\begin{aligned}
\overline{\Pi}_i^{(1)}(f) &= \frac{1}{f_i} \sum_{j=0}^{i-1} \frac{1}{\mu_j b_j} \bigg(\sum_{s=j+1}^{k-1} \mu_s f_s + \frac{1}{c} \mu_k f_k \bigg), \quad 0 < i \leqslant k, \\
\overline{\Pi}_i^{(2)}(f) &= \frac{1}{f_i} \sum_{j=i+1}^{N} \frac{1}{\mu_j a_j} \bigg(\sum_{s=k+1}^{j-1} \mu_s f_s + \frac{c-1}{c} \mu_k f_k \bigg), \quad k \leqslant i < N,
\end{aligned}$$

where $c > 1$ is a new parameter added just for the discrete case. Next, define

$$\begin{aligned}
&\mathscr{F}[0,k] = \{f : 0 = f_0 < f_1 < \cdots < f_k\}, \\
&\mathscr{F}[k,N] = \{f : f_k > f_{k+1} > \cdots > f_N = 0\}, \\
&\zeta[0,k] = \inf_{f \in \mathscr{F}[0,k]} \max_{0 < i \leqslant k} \overline{\Pi}_i^{(1)}(f)^{-1}, \qquad \eta[0,k] = \sup_{f \in \mathscr{F}[0,k]} \min_{0 < i \leqslant k} \overline{\Pi}_i^{(1)}(f)^{-1}, \\
&\zeta[k,N] = \inf_{f \in \mathscr{F}[k,N]} \max_{k \leqslant i < N} \overline{\Pi}_i^{(2)}(f)^{-1}, \qquad \eta[k,N] = \sup_{f \in \mathscr{F}[k,N]} \min_{k \leqslant i < N} \overline{\Pi}_i^{(2)}(f)^{-1}.
\end{aligned}$$

where ζ and η are used to estimate the upper and lower bounds respectively.

Theorem 2.1. The variational formulas of λ_0 are as follows.

$$\inf_{1 \leqslant k < N} \inf_{c > 1} \big(\zeta[0,k] \vee \zeta[k,N]\big) = \lambda_0 = \sup_{1 \leqslant k < N} \sup_{c > 1} \big(\eta[0,k] \wedge \eta[k,N]\big). \tag{2.1}$$

Before proving the theorem in detail, we present two lemmas and make some explanations for the ideas of the proof.

Lemma 2.2. Let $\Omega g(i) = -\lambda g_i$ $(0 < i < N)$ for some g and $\lambda \geqslant 0$. If there exist m and n with $0 \leqslant m < n \leqslant N$ such that $g = 0$ on $[m, n]$, then $g = 0$ on $[0, N]$.

Proof. Following the proof of ref. [2; Theorem 3.4], we obtain

$$\begin{aligned}
-\lambda\sum_{s=i}^{j}\pi_s g_s &= \sum_{s=i}^{j}\pi_s\Omega g(s)\\
&= \sum_{s=i}^{j}[\pi_s a_s(g_{s-1}-g_s)+\pi_s b_s(g_{s+1}-g_s)]\\
&= \sum_{s=i}^{j}[-\pi_s a_s(g_s-g_{s-1})+\pi_{s+1}a_{s+1}(g_{s+1}-g_s)]\\
&= \pi_{j+1}a_{j+1}(g_{j+1}-g_j)-\pi_i a_i(g_i-g_{i-1}),\\
&\qquad 0<i\leqslant j<N. \qquad (2.2)
\end{aligned}$$

Let $i=n$ in (2.2), since $g_{n-1}=g_n=0$, by induction, we have $g_j=0\,(n<j\leqslant N)$. Next, set $j=m$ in (2.2), because of $g_{m+1}=g_m=0$, by induction, it follows that $g_i=0\,(0<i<m)$. Hence, we have $g=0$. □

Lemma 2.3. Let g be a non-constant solution to the equation $\Omega g(i)=-\lambda_0 g_i$ $(0<i<N)$ with $g_0=g_N=0$, then $\lambda_0>0$ and there is no zero-point of g on $(0,N)$. Moreover, g is strictly positive or strictly negative on $(0,N)$.

Proof. (i) Assume that $\lambda_0=0$. Then we have

$$D(g)=\sum_{i=0}^{N-1}\pi_i b_i[g_{i+1}-g_i]^2=-\sum_{i=0}^{N}\pi_i(g\Omega g)(i)=0.$$

Hence, it follows that $g=0$. This is a contradiction by the assumptions.

(ii) Assume that $g_i=0$ for some i: $0<i<N$. By Lemma 2.2, we know that $g_{i-1}\neq 0$ and $g_{i+1}\neq 0$. If g_{i-1} and g_{i+1} are positive or negative simultaneously, without loss of generality, assume that $g_{i-1}\geqslant g_{i+1}>0$ and let

$$\overline{g}=gI_{[j\neq i]}+g_{i+1}I_{[j=i]}.$$

Otherwise, without loss of generality, assume that $g_{i-1}\geqslant -g_{i+1}>0$ and set

$$\overline{g}=gI_{[0,i-1]}+(-g_{i+1})I_{[j=i]}+(-g)I_{[i+1,N]}.$$

Then it always holds that $\overline{g}_0=\overline{g}_N=0$, $\pi(\overline{g}^2)>\pi(g^2)$ and $D(\overline{g})<D(g)$. Hence, we get

$$\lambda_0\leqslant\frac{D(\overline{g})}{\pi(\overline{g}^2)}<\frac{D(g)}{\pi(g^2)}=\lambda_0.$$

This is a contradiction.

(iii) Assume that g is not strictly positive on $(0,N)$. By Lemma 2.2, we know that $g_1\neq 0$. Without loss of generality, assume that $g_1>0$ and let $m=\max\{i: g_i>0\}$. By (ii), we have $g_{m+1}<0$. Set $\overline{g}=gI_{[0,m]}+(-g)I_{[m+1,N]}$. Then we get $\overline{g}_0=\overline{g}_N=0$, $\pi(\overline{g}^2)=\pi(g^2)$ and $D(\overline{g})<D(g)$. This induces a contradiction, similarly in (ii). □

The following result represents a character of the eigenfunction g of the principal eigenvalue λ_0: there maybe exist just two maximal points of g, so it is different from the continuous case.

Proposition 2.4. Let $g \geqslant 0$ be a non-constant solution to the equation $\Omega g(i) = -\lambda_0 g_i\,(0 < i < N)$ with $g_0 = g_N = 0$, then one of the following two statements for g must hold:

(1) there exists a $k \in (0, N-1)$ such that $0 = g_0 < g_1 < \cdots < g_k = g_{k+1} > \cdots > g_{N-1} > g_N = 0$;
(2) there exists a $k \in (0, N)$ such that $0 = g_0 < g_1 < \cdots < g_k > \cdots > g_{N-1} > g_N = 0$.

Proof. Suppose that k and $n\,(> k+1)$ are two maximal points of g on $(0, N)$, i.e. $g_{k-1} < g_k \geqslant g_{k+1}$ and $g_{n-1} \leqslant g_n > g_{n+1}$. By Lemma 2.3, we have $g_n > 0$. Let

$$\overline{g} = gI_{[0,k-1]} + g_k I_{[k,n]} + cgI_{[n+1,N]},$$

where $c = g_k/g_n$. Then we get $\overline{g}_0 = \overline{g}_N = 0$ and

$$\begin{aligned} \pi(\overline{g}^2) &= \sum_{i=0}^{k-1} \pi_i g_i^2 + g_k^2 \sum_{i=k}^{n} \pi_i + c^2 \sum_{i=n+1}^{N} \pi_i g_i^2, \\ -\sum_{i=0}^{N} \pi_i (\overline{g}\Omega\overline{g})(i) &= \lambda_0 \sum_{i=0}^{k-1} \pi_i g_i^2 + \pi_k g_k a_k (g_k - g_{k-1}) \\ &\quad + c^2 \pi_n g_n b_n (g_n - g_{n+1}) + \lambda_0 c^2 \sum_{i=n+1}^{N} \pi_i g_i^2. \end{aligned}$$

Note that

$$\lambda_0 g_k = -\Omega g(k) \geqslant a_k (g_k - g_{k-1})$$

and

$$\lambda_0 g_n = -\Omega g(n) \geqslant b_n (g_n - g_{n+1}).$$

Hence, it follows that

$$\pi_k g_k a_k (g_k - g_{k-1}) + c^2 \pi_n g_n b_n (g_n - g_{n+1}) \leqslant \lambda_0 \pi_k g_k^2 + \lambda_0 c^2 \pi_n g_n^2 < \lambda_0 g_k^2 \sum_{i=k}^{n} \pi_i.$$

Therefore, we obtain

$$\lambda_0 \leqslant -\sum_{i=0}^{N} \frac{\pi_i (\overline{g}\Omega\overline{g})(i)}{\pi(\overline{g}^2)} < \lambda_0.$$

This is a contradiction. So either $k+1 = n$ which is just the first statement, or $k = n$ which implies that there exists uniquely an extreme (maximal) point k and so the second statement holds. Hence, the required assertion follows. $\square$

We now study the principal eigenvalue λ_0 of Ω. If imitating the continuous case completely, we can obtain the variational formulas similar to (1.1) in which the inequalities become equalities only for those eigenfunctions for which part (1) of Proposition 2.4 holds. To deal with the case of part (2) of Proposition 2.4, our idea goes as follows. First, insert a point behind k and enlarge the state space to $\overline{E} = \{0, 1, \cdots, N, N+1\}$ (if $N = \infty$, then $\overline{E} = E$ and so there is no change). Next, construct a new birth-death chain $(\overline{a}_i, \overline{b}_i)$ on $\overline{E}$ such that its eigenfunction $\overline{g}$ satisfies $\overline{g}_i = g_i\,(0 \leqslant i \leqslant k)$ and $\overline{g}_i = g_{i-1}\,(k+1 \leqslant i \leqslant N+1)$ (note that $\overline{g}_k = \overline{g}_{k+1}$) and the eigenvalue $\overline{\lambda}_0 = \lambda_0$. In other words, by enlarging the state space, the second case in Proposition 2.4 can be reduced to the first one. Here, it is not difficult to show that

$$\frac{\overline{a}_k}{a_k} = \frac{\lambda_0 g_k}{a_k(g_k - g_{k-1})} = 1 + \frac{b_k(g_k - g_{k+1})}{a_k(g_k - g_{k-1})} > 1$$

by the fact that

$$-\overline{\Omega}\overline{g}(k) = \overline{\lambda}_0\overline{g}_k = \lambda_0 g_k = -\Omega g(k).$$

This is just the derivation of the constant c. Based on the consideration above, construct the new birth-death chain as follows. Given $k \in [1, N)$, let

$$\overline{a}_i = \begin{cases} a_i, & 1 \leqslant i \leqslant k-1; \\ ca_k, & i = k; \\ 1, & i = k+1; \\ a_{i-1}, & k+2 \leqslant i \leqslant N+1, \end{cases} \qquad \overline{b}_i = \begin{cases} b_i, & 0 \leqslant i \leqslant k-1; \\ c-1, & i = k; \\ cb_k/(c-1), & i = k+1; \\ b_{i-1}, & k+2 \leqslant i \leqslant N, \end{cases}$$

where $c > 1$. Then, we get

$$\overline{\pi}_i = \pi_i\,(0 \leqslant i \leqslant k-1), \quad \overline{\pi}_k = \frac{1}{c}\pi_k, \quad \overline{\pi}_{k+1} = \frac{c-1}{c}\pi_k, \quad \overline{\pi}_i = \pi_{i-1}\,(k+2 \leqslant i \leqslant N+1).$$

Of course, There are same relations between $\overline{\mu}_i$ and μ_i. By (2.0), we have

$$\overline{\mu} = \mu < \infty, \qquad \sum_{i=0}^{N+1} \frac{1}{\overline{\mu}_i\overline{b}_i} = \sum_{i=0}^{N} \frac{1}{\mu_i b_i} + \frac{c}{(c-1)\mu_k} < \infty. \tag{2.3}$$

The final result (Theorem 2.1) unifies the two cases of eigenfunctions. We now start the detail proof.

Proof of Theorem 2.1. (i) Given $k \in [1, N)$, $c > 1$ and two positive sequences $\{\ell_i\}_{i=0}^{k-1}$ and $\{r_i\}_{i=k+1}^{N}$. For every h satisfying $h_0 = h_N = 0$ and $\pi(h^2) = 1$, define $\overline{h}$ on $\overline{E}$: $\overline{h}_i = h_i\,(0 \leqslant i \leqslant k)$ and $\overline{h}_i = h_{i-1}\,(k+1 \leqslant i \leqslant N+1)$. Then we have

$\overline{\pi}\big(\overline{h}^2\big)=1$ and $\overline{D}(\overline{h})=D(h)$. By Cauchy-Schwarz inequality, it follows that

$$\begin{aligned}
1&=\sum_{i=0}^{N+1}\overline{\pi}_i\overline{h}_i^2=\sum_{i=1}^{k}\overline{\pi}_i\big(\overline{h}_i-\overline{h}_0\big)^2+\sum_{i=k+1}^{N}\overline{\pi}_i\big(\overline{h}_{N+1}-\overline{h}_i\big)^2\\
&=\sum_{i=1}^{k}\overline{\pi}_i\bigg(\sum_{j=0}^{i-1}\big(\overline{h}_{j+1}-\overline{h}_j\big)\bigg)^2+\sum_{i=k+1}^{N}\overline{\pi}_i\bigg(\sum_{j=i}^{N}\big(\overline{h}_{j+1}-\overline{h}_j\big)\bigg)^2\\
&\leqslant\sum_{i=1}^{k}\overline{\pi}_i\sum_{j=0}^{i-1}\frac{\big(\overline{h}_{j+1}-\overline{h}_j\big)^2\overline{\pi}_j\overline{b}_j}{\ell_j}\sum_{s=0}^{i-1}\frac{\ell_s}{\overline{\pi}_s\overline{b}_s}+\sum_{i=k+1}^{N}\overline{\pi}_i\sum_{j=i}^{N}\frac{\big(\overline{h}_{j+1}-\overline{h}_j\big)^2\overline{\pi}_j\overline{b}_j}{r_j}\sum_{s=i}^{N}\frac{r_s}{\overline{\pi}_s\overline{b}_s}\\
&=\sum_{j=0}^{k-1}\overline{\pi}_j\overline{b}_j\big(\overline{h}_{j+1}-\overline{h}_j\big)^2\frac{1}{\ell_j}\sum_{i=j+1}^{k}\overline{\pi}_i\sum_{s=0}^{i-1}\frac{\ell_s}{\overline{\pi}_s\overline{b}_s}\\
&\quad+\sum_{j=k+1}^{N}\overline{\pi}_j\overline{b}_j\big(\overline{h}_{j+1}-\overline{h}_j\big)^2\frac{1}{r_j}\sum_{i=k+1}^{j}\overline{\pi}_i\sum_{s=i}^{N}\frac{r_s}{\overline{\pi}_s\overline{b}_s}\\
&\leqslant\overline{D}(\overline{h})\bigg[\max_{0\leqslant j\leqslant k-1}\bigg(\frac{1}{\ell_j}\sum_{i=j+1}^{k}\overline{\pi}_i\sum_{s=0}^{i-1}\frac{\ell_s}{\overline{\pi}_s\overline{b}_s}\bigg)\bigvee\sup_{k+1\leqslant j\leqslant N}\bigg(\frac{1}{r_j}\sum_{i=k+1}^{j}\overline{\pi}_i\sum_{s=i+1}^{N+1}\frac{r_{s-1}}{\overline{\pi}_s\overline{a}_s}\bigg)\bigg]\\
&=:D(h)\Big[\max_{0\leqslant j\leqslant k-1}\overline{L}_j\bigvee\sup_{k+1\leqslant j\leqslant N}\overline{R}_j\Big].
\end{aligned}$$

Let $f\in\mathscr{F}[0,k]$, $g\in\mathscr{F}[k,N]$ and set $\ell_j=\sum_{i=j+1}^{k}\overline{\mu}_if_i$ and $r_j=\sum_{i=k+1}^{j}\overline{\mu}_ig_{i-1}$. Instead of Mean Value Theorem, we adopt the proportion property and get

$$\begin{aligned}
\max_{0\leqslant j\leqslant k-1}\overline{L}_j&\leqslant\max_{1\leqslant i\leqslant k}\frac{1}{f_i}\sum_{s=0}^{i}\frac{\ell_s}{\overline{\mu}_s\overline{b}_s}=\max_{0<i\leqslant k}\overline{\Pi}_i^{(1)}(f),\\
\sup_{k+1\leqslant j\leqslant N}\overline{R}_j&\leqslant\sup_{k+1\leqslant i\leqslant N}\frac{1}{g_{i-1}}\sum_{s=i+1}^{N+1}\frac{r_{s-1}}{\overline{\mu}_s\overline{a}_s}=\sup_{k\leqslant i<N}\overline{\Pi}_i^{(2)}(g).
\end{aligned}$$

Collecting these facts together, it follows that $\lambda_0\geqslant\eta[0,k]\wedge\eta[k,N]$. Therefore, we have

$$\lambda_0\geqslant\sup_{1\leqslant k<N}\sup_{c>1}\big(\eta[0,k]\wedge\eta[k,N]\big).\tag{2.4}$$

(ii) Given $k\in[1,N)$, $c>1$, let $f\in\mathscr{F}[0,k]$, $g\in\mathscr{F}[k,N]$ and set $\alpha=f_k\overline{\Pi}_k^{(1)}(f)/\big(g_k\overline{\Pi}_k^{(2)}(g)\big)$. Define $h_0=h_N=0$, $h_i=f_i\overline{\Pi}_i^{(1)}(f)\,(1\leqslant i\leqslant k)$ and $h_i=\alpha g_i\overline{\Pi}_i^{(2)}(g)\,(k\leqslant i\leqslant N-1)$. Then we obtain

$$\begin{aligned}
D(h)&=\sum_{i=0}^{k-1}\pi_ib_i(h_{i+1}-h_i)^2+\sum_{i=k+1}^{N}\pi_ia_i(h_{i-1}-h_i)^2\\
&=\sum_{i=0}^{k-1}(h_{i+1}-h_i)\bigg(\sum_{s=i+1}^{k-1}\pi_sf_s+\frac{1}{c}\pi_kf_k\bigg)\\
&\quad+\alpha\sum_{i=k+1}^{N}(h_{i-1}-h_i)\bigg(\sum_{s=k+1}^{i-1}\pi_sg_s+\frac{c-1}{c}\pi_kg_k\bigg)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{s=1}^{k-1} \pi_s f_s \sum_{i=0}^{s-1} (h_{i+1} - h_i) + \frac{1}{c} \pi_k f_k h_k + \alpha \sum_{s=k+1}^{N-1} \pi_s g_s \sum_{i=s+1}^{N} (h_{i-1} - h_i) \\
&\quad + \frac{c-1}{c} \alpha \pi_k g_k h_k \\
&= \sum_{s=1}^{k-1} \pi_s f_s h_s + \frac{1}{c} \pi_k f_k h_k + \alpha \sum_{s=k+1}^{N-1} \pi_s g_s h_s + \frac{c-1}{c} \alpha \, \pi_k g_k h_k \\
&\leqslant \Big(\sum_{s=1}^{k-1} \pi_s h_s^2 + \frac{\pi_k h_k^2}{c} \Big) \max_{1 \leqslant i \leqslant k} \overline{\Pi}_i^{(1)}(f)^{-1} \\
&\quad + \Big(\sum_{s=k+1}^{N-1} \pi_s h_s^2 + \frac{c-1}{c} \pi_k h_k^2 \Big) \sup_{k \leqslant i < N} \overline{\Pi}_i^{(2)}(g)^{-1} \\
&\leqslant \pi(h^2) \Big(\max_{1 \leqslant i \leqslant k} \overline{\Pi}_i^{(1)}(f)^{-1} \bigvee \sup_{k \leqslant i < N} \overline{\Pi}_i^{(2)}(g)^{-1} \Big).
\end{aligned}
$$

Hence, we have $\lambda_0 \leqslant \zeta[0,k] \vee \zeta[k,N]$. Therefore, we have

$$\lambda_0 \leqslant \inf_{1 \leqslant k < N} \inf_{c > 1} \big(\zeta[0,k] \vee \zeta[k,N] \big). \tag{2.5}$$

(iii) No matter the eigenfunction g belongs to the case (1) or (2) of Proposition 2.4, we always set $f_1 = g I_{[0,k]} \in \mathscr{F}[0,k]$ and $f_2 = g I_{[k,N]} \in \mathscr{F}[k,N]$. By (2.2), it follows that

$$\lambda_0 \sum_{s=j+1}^{k-1} \mu_s g_s = \mu_j b_j (g_{j+1} - g_j) - \mu_k a_k (g_k - g_{k-1}) \, (0 \leqslant j \leqslant i-1)$$

and

$$\lambda_0 \sum_{s=k+1}^{j-1} \mu_s g_s = \mu_j a_j (g_{j-1} - g_j) - \mu_k b_k (g_{k+1} - g_k) \, (i+1 \leqslant j \leqslant N)$$

.

If g belongs to the case (2) of Proposition 2.4, set $c = \lambda_0 g_k / (a_k (g_k - g_{k-1}))$, then $\overline{\Pi}_i^{(1)}(f_1) = \lambda_0^{-1} \, (0 < i \leqslant k)$, $\overline{\Pi}_i^{(2)}(f_2) = \lambda_0^{-1} \, (k \leqslant i < N)$. Hence, the equalities in (2.4) and (2.5) hold.

If g belongs to the case (1) of Proposition 2.4, then we have

$$
\begin{aligned}
\overline{\Pi}_i^{(1)}(f_1) &= \frac{1}{\lambda_0} - \Big(1 - \frac{1}{c}\Big) \mu_k g_k \frac{1}{g_i} \sum_{j=0}^{i-1} \frac{1}{\mu_j b_j}, \qquad 0 < i \leqslant k, \\
\overline{\Pi}_i^{(2)}(f_2) &= \frac{1}{\lambda_0} + \Big(1 - \frac{1}{c}\Big) \mu_k g_k \frac{1}{g_i} \sum_{j=i+1}^{N} \frac{1}{\mu_j a_j}, \qquad k \leqslant i < N.
\end{aligned}
$$

By these equalities, it is easy to get that

$$\inf_{c>1}\Big(\max_{0<i\leqslant k}\overline{\Pi}_i^{(1)}(f_1)^{-1}\bigvee\sup_{k\leqslant i<N}\overline{\Pi}_i^{(2)}(f_2)^{-1}\Big)=\lambda_0,$$
$$\sup_{c>1}\Big(\min_{0<i\leqslant k}\overline{\Pi}_i^{(1)}(f_1)^{-1}\bigwedge\inf_{k\leqslant i<N}\overline{\Pi}_i^{(2)}(f_2)^{-1}\Big)=\lambda_0.$$

So the right-hand side of (2.4) $\geqslant \lambda_0$ and the right-hand side of (2.5) $\leqslant \lambda_0$. Hence, the formulas in (2.1) hold.

We have thus proved the required assertion. $\square$

We now study the explicit bounds of λ_0. In the proof of Theorem 2.1, we construct a new birth-death chain $(\overline{a}_i,\overline{b}_i)$ on $\overline{E}=\{0,1,\cdots,N,N+1\}$. The mixed eigenvalues on $[0,k]$ and $[k+1,N+1]$ are denoted by $\lambda_0^{(c)}[0,k]$ and $\lambda_0^{(c)}[k+1,N+1]$ respectively. It is easy to see that

$$\overline{\Pi}_i^{(1)}(f)=\frac{1}{f_i}\sum_{j=0}^{i-1}\frac{1}{\overline{\mu}_j\overline{b}_j}\sum_{s=j+1}^{k}\overline{\mu}_s f_s,\qquad 0<i\leqslant k.$$

Let $\overline{f}_i=f_{i-1}\,(k+1\leqslant i\leqslant N+1)$. Then we easily know that

$$\overline{\Pi}_i^{(2)}(f)=\frac{1}{\overline{f}_{i+1}}\sum_{j=i+2}^{N+1}\frac{1}{\overline{\mu}_j\overline{a}_j}\sum_{s=k+1}^{j-1}\overline{\mu}_s\overline{f}_s,\qquad k\leqslant i<N.$$

Hence, from ref. [3; Theorem 2.1], it follows that $\zeta[0,k]=\lambda_0^{(c)}[0,k]=\eta[0,k]$ and $\zeta[k,N]=\lambda_0^{(c)}[k+1,N+1]=\eta[k,N]$. So (2.1) can be rewritten as

$$\inf_{1\leqslant k<N}\inf_{c>1}\big(\lambda_0^{(c)}[0,k]\vee\lambda_0^{(c)}[k+1,N+1]\big)=\lambda_0=\sup_{1\leqslant k<N}\sup_{c>1}\big(\lambda_0^{(c)}[0,k]\wedge\lambda_0^{(c)}[k+1,N+1]\big).\tag{2.6}$$

Let

$$\delta'(k,c)=\max_{1\leqslant i\leqslant k}\sum_{j=0}^{i-1}\frac{1}{\mu_j b_j}\Big(\sum_{j=i}^{k-1}\mu_j+\frac{1}{c}\mu_k\Big)\ \Big(=\max_{1\leqslant i\leqslant k}\sum_{j=0}^{i-1}\frac{1}{\overline{\mu}_j\overline{b}_j}\sum_{j=i}^{k}\overline{\mu}_j\Big),$$
$$\delta''(k,c)=\sup_{k+1\leqslant i\leqslant N}\sum_{j=i}^{N}\frac{1}{\mu_j a_j}\Big(\sum_{j=k+1}^{i-1}\mu_j+\frac{c-1}{c}\mu_k\Big)$$
$$\Big(=\sup_{k+1\leqslant i\leqslant N}\sum_{j=i+1}^{N+1}\frac{1}{\overline{\mu}_j\overline{a}_j}\sum_{j=k+1}^{i}\overline{\mu}_j\Big).$$

Then, by ref. [2; Theorem 3.4], it follows that

$$\delta'(k,c)^{-1}\geqslant\lambda_0^{(c)}[0,k]\geqslant\big(4\delta'(k,c)\big)^{-1}$$

and

$$\delta''(k,c)^{-1}\geqslant\lambda_0^{(c)}[k+1,N+1]\geqslant\big(4\delta''(k,c)\big)^{-1}.$$

From (2.6), the following corollary follows immediately.

Corollary 2.5.

$$\inf_{1\leqslant k<N}\inf_{c>1}\big(\delta'(k,c)\wedge\delta''(k,c)\big)^{-1}\geqslant\lambda_0\geqslant\sup_{1\leqslant k<N}\sup_{c>1}\big(4(\delta'(k,c)\vee\delta''(k,c))\big)^{-1}.$$

Three examples are illustrated as follows.

Example 2.6. Let $N=3$, $a_i=1\,(i=1,2,3)$ and $b_i=1\,(i=0,1,2)$. Then $\mu_i=1\,(i=0,1,2,3)$. Given $k=1$, we have

$$\begin{aligned}\overline{\Pi}_1^{(1)}(f)&=\frac{1}{c},\\ \overline{\Pi}_1^{(2)}(f)&=2-\frac{2}{c}+\frac{f_2}{f_1},\\ \overline{\Pi}_2^{(2)}(f)&=1+\left(1-\frac{1}{c}\right)\frac{f_1}{f_2}.\end{aligned}$$

Set $f_1=f_2$. We get $\overline{\Pi}_1^{(2)}(f)=3-2/c$ and $\overline{\Pi}_2^{(2)}(f)=2-1/c$. By (2.1) plus some computation, one gets the exact estimate $\lambda_0=1$. In fact, the eigenfunction g here satisfies $g_1=g_2$. By Corollary 2.5, we have $1\geqslant\lambda_0\geqslant 1/4$.

Example 2.7. Let $N=3$, $a_1=a_2=1$, $a_3=3$ and $b_i=i+1\,(i=0,1,2)$. Then $\mu_0=\mu_1=1$ and $\mu_2=\mu_3=2$. Given $k=1$, we have

$$\begin{aligned}\overline{\Pi}_1^{(1)}(f)&=\frac{1}{c},\\ \overline{\Pi}_1^{(2)}(f)&=\frac{2}{3}(1-1/c)+\frac{f_2}{3f_1},\\ \overline{\Pi}_2^{(2)}(f)&=\frac{1}{3}+(1-1/c)\frac{f_1}{6f_2}.\end{aligned}$$

Set $f_1=2f_2$. We have

$$\overline{\Pi}_1^{(2)}(f)=\frac{5}{6}-\frac{2}{3c},\qquad\overline{\Pi}_2^{(2)}(f)=\frac{2}{3}-\frac{1}{3c}.$$

Let $c=2$. The exact estimate $\lambda_0=2$ is provided by (2.1). The eigenfunction g here satisfies $g_1=2g_2$ (so $g_1\neq g_2$). Corollary 2.5 gives us $7/3\geqslant\lambda_0\geqslant 7/12$.

Example 2.8. Let $N=3$, $a_1=(2-\varepsilon^2)/(1+\varepsilon)$, $a_2=a_3=1$, $b_0=b_1=1$ and $b_2=2$, where $\varepsilon\in[0,\sqrt{2})$. Then $\mu_0=1$ and

$$\mu_1=\mu_2=\mu_3/2=(1+\varepsilon)/(2-\varepsilon^2).$$

Here the eigenfunction g satisfies $g_1=(1+\varepsilon)g_2$ and $\lambda_0=2-\varepsilon$. Given $k=1$, for each f, we have

$$\begin{aligned}\overline{\Pi}_1^{(1)}(f)&=\frac{1+\varepsilon}{c(2-\varepsilon^2)},\\ \overline{\Pi}_1^{(2)}(f)&=\frac{3}{2}(1-1/c)+\frac{f_2}{2f_1},\\ \overline{\Pi}_2^{(2)}(f)&=\frac{1}{2}+(1-1/c)\frac{f_1}{2f_2}.\end{aligned}$$

In particular, let $f_1 = (1+\varepsilon)f_2$. We see that

$$\overline{\Pi}_1^{(2)}(f) = \frac{3}{2}(1-1/c) + \frac{1}{2(1+\varepsilon)} \quad \text{and} \quad \overline{\Pi}_2^{(2)}(f) = \frac{1}{2} + (1-1/c)\frac{1+\varepsilon}{2}.$$

If $\varepsilon = 0$, then by (2.1), it follows that

$$\sup_{c>1} \frac{2}{4-3/c} \leqslant \lambda_0 \leqslant \inf_{c>1} 2c,$$

i.e. we get the exact estimate $\lambda_0 = 2$ (by letting $c \to 1$). At the same time, by Corollary 2.5, we have $2 \geqslant \lambda_0 \geqslant 1/2$. If $\varepsilon > 0$, set $c = (2-\varepsilon)(1+\varepsilon)/(2-\varepsilon^2)$, then we get the exact estimate $\lambda_0 = 2-\varepsilon$ by (2.1) plus some computation. If $\varepsilon = 1$, from Corollary 2.5, it follows that $7/6 \geqslant \lambda_0 \geqslant 7/24$.

3. Higher-order eigenvalues in continuous case.

Consider the higher-order Dirichlet eigenvalues of the differential operator L on the finite interval (p, q):

$$\begin{aligned}\lambda_n = \inf\{D(f) : f \in C^1(p,q) \cap C[p,q],\ f(p) = f(q) = 0,\\ \pi(f^2) = 1,\ \pi(fg_i) = 0,\ i = 0, 1, \cdots, n-1\},\end{aligned}$$

where g_i is the corresponding eigenfunction of λ_i on (p, q), i.e. $Lg_i = -\lambda_i g_i$. The idea is that since there exist exactly number n of zero-points of the corresponding eigenfunction g_n of λ_n on (p, q) (cf. ref. [5; Theorem 19 of Chapter 5]), we divided the interval (p, q) into $n+1$ sub-intervals according to the zero-points of g_n. Then reduce the study of λ_n to the first Dirichlet eigenvalue problem in sub-intervals. In detail, given n points on (p, q), denoted by $x_1, \cdots, x_n$, with $x_1 < \cdots < x_n$. Let $x_0 = p$ and $x_{n+1} = q$. Consider the differential operator L on (x_i, x_{i+1}) with Dirichlet boundary. Denote by $\lambda_0(x_i, x_{i+1})$ the first Dirichlet eigenvalue on (x_i, x_{i+1}).

Theorem 3.1. Let a and b are continuous on the finite interval $[p, q]$ and $a > 0$. Then

$$\inf_{p<x_1<\cdots<x_n<q} \max_{0\leqslant i\leqslant n} \lambda_0(x_i, x_{i+1}) = \lambda_n = \sup_{p<x_1<\cdots<x_n<q} \min_{0\leqslant i\leqslant n} \lambda_0(x_i, x_{i+1})$$

Proof. Applying ref. [5; Theorem 19 of Chapter 5] to the functions $p(x) = \mathrm{e}^{C(x)}$, $q(x) \equiv 0$ and $\rho(x) = p(x)/a(x)$, we know that there exist number n of zero-points on (p, q) of the eigenfunction g_n corresponding to λ_n. Denoted them by $x'_1, \cdots, x'_n$. Let $x'_0 = p$ and $x'_{n+1} = q$. Then there must exist i and j such that $x_i \leqslant x'_i < x'_{i+1} \leqslant x_{i+1}$ and $x'_j \leqslant x_j < x_{j+1} \leqslant x'_{j+1}$, i.e. $(x'_i, x'_{i+1}) \subseteq (x_i, x_{i+1})$ and $(x_j, x_{j+1}) \subseteq (x'_j, x'_{j+1})$. On the one hand, by the monotonicity of the first Dirichlet eigenvalue with respect to intervals, we have $\lambda_0(x'_i, x'_{i+1}) \geqslant \lambda_0(x_i, x_{i+1})$ and $\lambda_0(x_j, x_{j+1}) \geqslant \lambda_0(x'_j, x'_{j+1})$. On the other hand, note that we have $Lg_n = -\lambda_n g_n$, but there is no zero-point of g_n on every sub-interval $(x'_k, x'_{k+1})\,(k = 0, \cdots, n)$. This implies that g_n is just the eigenfunction corresponding to the first

Dirichlet eigenvalue on the sub-interval and $\lambda_n = \lambda_0(x'_k, x'_{k+1})$. Hence, we obtain $\lambda_0(x_j, x_{j+1}) \geqslant \lambda_n \geqslant \lambda_0(x_i, x_{i+1})$. Furthermore, since $x_1 < \cdots < x_n$ are arbitrary, we get

$$\inf_{p<x_1<\cdots<x_n<q} \max_{0\leqslant i\leqslant n} \lambda_0(x_i, x_{i+1}) \geqslant \lambda_n \geqslant \sup_{p<x_1<\cdots<x_n<q} \min_{0\leqslant i\leqslant n} \lambda_0(x_i, x_{i+1}).$$

In particular, if $x_i = x'_i \,(i = 1, 2, \cdots, n)$, then the inequalities above all become equalities and so the required assertion holds. $\square$

The next explicit bounds of λ_n follow from Theorem 3.1 and Corollary 1.5 immediately.

Corollary 3.2. Let

$$\delta'_i(c) = \sup_{x\in(x_i,c)} \int_{x_i}^{x} \mathrm{e}^{-C(y)}\mathrm{d}y \int_x^c \frac{\mathrm{e}^{C(u)}}{a(u)}\mathrm{d}u \quad \text{and}$$
$$\delta''_i(c) = \sup_{x\in(c,x_{i+1})} \int_{x}^{x_{i+1}} \mathrm{e}^{-C(y)}\mathrm{d}y \int_c^x \frac{\mathrm{e}^{C(u)}}{a(u)}\mathrm{d}u.$$

Then under the assumptions of Theorem 3.1, the following conclusion holds

$$\inf_{p<x_1<\cdots<x_n<q} \max_{0\leqslant i\leqslant n} \delta'_i(c_i)^{-1} \geqslant \lambda_n \geqslant \sup_{p<x_1<\cdots<x_n<q} \min_{0\leqslant i\leqslant n} \big(4\delta'_i(c_i)\big)^{-1},$$

where c_i is the unique solution to the equation $\delta'_i(c) = \delta''_i(c)$ on (x_i, x_{i+1}) $(i = 0, 1, \cdots, n)$.

Example 3.3. Let $a(x) \equiv 1$ and $b(x) \equiv 0$. Set $x_i = i/(n+1)\,(i = 1, 2, \cdots, n)$. Then

$$c_i = \frac{x_i + x_{i+1}}{2} = \frac{2i+1}{2(n+1)} \quad \text{and} \quad \delta'_i(c_i) = \frac{1}{16(n+1)^2}, \qquad i = 0, 1, \cdots, n.$$

By Corollary 3.2, it follows that $4(n+1)^2 \leqslant \lambda_n \leqslant 16(n+1)^2$. The exact value is $\lambda_n = (n+1)^2\pi^2 \approx 9.8696(n+1)^2$ and the corresponding eigenfunction is $f(x) = \sin\big((n+1)\pi x\big)$.

Acknowledgements The authors would like to thank the referee for their helpful comments on the first version of the paper.

References

[1] Egorov Y and Kondratiev V, *On Spectral Theory of Elliptic Operators*, Birkhäuser Verlag, Berlin, 1996.
[2] Chen M F, *Explicit bounds of the first eigenvalue*, Sci. in China (A) 2000, 43:10, 1051-1059.
[3] Chen M F, *Variational formulas and approximation theorems for the first eigenvalue in dimension one*, Sci. in China (A) 2001, 44:4, 409-418.
[4] Chen M F and Wang F Y, *Estimation of spectral gap for elliptic operators*, Trans. Amer. Math. Soc. 1997, 349:3, 1239-1267.
[5] Opic B and Kufner A, *Hardy-type Inequalities*, Longman, New York, 1990.
[6] Chen M F, *Analytic proof of dual variational formula for the first eigenvalue in dimension one*, Sci. in China (A) 1999, 42:8, 805-815.

Appendix [unpublished]**.** We show a result about an approximating procedure.

Proposition. Fix $\varepsilon > 0$. Let C be bounded on the interval $[0, \varepsilon]$. Given $\delta \neq 0$, then for every absolutely continuous function g with $g(0) = 0$ and $g(\varepsilon) = \delta$, we have

$$\int_0^\varepsilon {g'}^2 e^C \geqslant \frac{1}{\varepsilon}\delta^2 \exp\Big[\inf_{x\in[0,\varepsilon]} C(x)\Big].$$

If moreover g is increasing, then

$$\int_0^\varepsilon g^2 e^C \leqslant \varepsilon\delta^2 \exp\Big[\sup_{x\in[0,\varepsilon]} C(x)\Big].$$

In particular, if $\delta = \delta(\varepsilon) \not\to 0$ as $\varepsilon \to 0$, then

$$\varliminf_{\varepsilon\to 0} \int_0^\varepsilon {g'}^2 e^C = \infty.$$

Proof. Represent g by

$$g(x) = \delta \int_0^x h \Big/ \int_0^\varepsilon h, \qquad x \leqslant \varepsilon.$$

Then $g' = \delta h \big/ \int_0^\varepsilon h$. By the Cauchy-Schwarz inequality, we have

$$\bigg(\int_0^\varepsilon h\bigg)^2 = \bigg(\int_0^\varepsilon e^{-C/2} h e^{C/2}\bigg)^2 \leqslant \int_0^\varepsilon e^{-C} \int_0^\varepsilon h^2 e^C.$$

Hence

$$\int_0^\varepsilon {g'}^2 e^C = \frac{\delta^2 \int_0^\varepsilon h^2 e^C}{\big(\int_0^\varepsilon h\big)^2} \geqslant \frac{\delta^2}{\int_0^\varepsilon e^{-C}} \geqslant \frac{1}{\varepsilon}\delta^2 \exp\Big[\inf_{x\in[0,\varepsilon]} C(x)\Big].$$

In the case that g is increasing, we can assume that $h \geqslant 0$. Then

$$\int_0^\varepsilon g^2 e^C = \delta^2 \int_0^\varepsilon \bigg(\int_0^\bullet h \Big/ \int_0^\varepsilon h\bigg)^2 e^C \leqslant \delta^2 \int_0^\varepsilon e^C \leqslant \varepsilon\delta^2 \exp\Big[\sup_{x\in[0,\varepsilon]} C(x)\Big]. \quad \square$$

Corollary. If $\varepsilon \to 0$ but $\delta = \delta(\varepsilon) \not\to 0$, then $\varliminf_{\varepsilon\to 0} \int_0^\varepsilon {g'}^2 e^C = \infty$.

Ergodic Convergence Rates of Markov Processes — Eigenvalues, Inequalities and Ergodic Theory

Mu-Fa Chen*

Abstract

This paper consists of four parts. In the first part, we explain what eigenvalues we are interested in and show the difficulties of the study on the first (non-trivial) eigenvalue through examples. In the second part, we present some (dual) variational formulas and explicit bounds for the first eigenvalue of Laplacian on Riemannian manifolds or Jacobi matrices (Markov chains). Here, a probabilistic approach—the coupling methods is adopted. In the third part, we introduce recent lower bounds of several basic inequalities; these are based on a generalization of Cheeger's approach which comes from Riemannian geometry. In the last part, a diagram of nine different types of ergodicity and a table of explicit criteria for them are presented. These criteria are motivated by the weighted Hardy inequality which comes from Harmonic analysis.

2000 Mathematics Subject Classification: 35P15, 47A75, 49R50, 60J99.
Keywords and Phrases: Eigenvalue, variational formula, inequality, convergence rate, ergodic theory, Markov process.

Part I. Introduction

We will start by explaining what eigenvalues we are interested in.

1.1 Definition. Consider a birth-death process with a state space

$$E = \{0, 1, 2, \cdots, n\} \ (n \leqslant \infty)$$

and an intensity matrix $Q = (q_{ij})$: $q_{k,k-1} = a_k > 0\,(1 \leqslant k \leqslant n)$, $q_{k,k+1} = b_k > 0\,(0 \leqslant k \leqslant n-1)$, $q_{k,k} = -(a_k + b_k)$, and $q_{ij} = 0$ for other $i \neq j$. Since the sum of each row equals 0, we have $Q1 = 0 = 0 \cdot 1$. This means that the Q-matrix has an eigenvalue 0 with an eigenvector 1. Next, consider the finite case of $n < \infty$. Then, the eigenvalues of $-Q$ are discrete: $0 = \lambda_0 < \lambda_1 \leqslant \cdots \leqslant \lambda_n$. We are interested in the first (non-trivial) eigenvalue $\lambda_1 = \lambda_1 - \lambda_0$ (also called spectral gap of Q). In

*Department of Mathematics, Beijing Normal University, Beijing 100875, The People's Republic of China. E-mail: mfchen@bnu.edu.cn
Home page: http://www.bnu.edu.cn/˜chenmf/main_eng.htm

the infinite case ($n = \infty$), λ_1 can be 0. Certainly, one can consider a self-adjoint elliptic operator in $\mathbb{R}^d$, the Laplacian Δ on manifolds, or an infinite-dimensional operator as in the study of interacting particle systems.

1.2 Difficulties. To get a concrete feeling about the difficulties of this topic, let us first look at the following examples with a finite state space. When $E = \{0, 1\}$, it is trivial that $\lambda_1 = a_1 + b_0$. The result is nice because when either a_1 or b_0 increases, so does λ_1. When $E = \{0, 1, 2\}$, we have four parameters b_0, b_1, a_1, a_2 and

$$\lambda_1 = 2^{-1}\big[a_1 + a_2 + b_0 + b_1 - \sqrt{(a_1 - a_2 + b_0 - b_1)^2 + 4a_1 b_1}\,\big].$$

When $E = \{0, 1, 2, 3\}$, we have six parameters: $b_0, b_1, b_2, a_1, a_2, a_3$. In this case, the expression for λ_1 is too lengthy to write. The roles of the parameters are inter-related in a complicated manner. Clearly, it is impossible to compute λ_1 explicitly when the size of the matrix is greater than five.

Next, consider the infinite state space $E = \{0, 1, 2, \cdots\}$. Denote the eigenfunction of λ_1 by g and the degree of g by $D(g)$ when g is polynomial. Three examples of the perturbation of λ_1 and $D(g)$ are listed in Table 1.1.

$b_i (i \geqslant 0)$	$a_i (i \geqslant 1)$	λ_1	$D(g)$
$i + c\,(c > 0)$	$2i$	1	1
$i + 1$	$2i + 3$	2	2
$i + 1$	$2i + \big(4 + \sqrt{2}\,\big)$	3	3

Table 1.1 Three examples of the perturbation of λ_1 and $D(g)$

The first line is the well known linear model for which $\lambda_1 = 1$, independent of the constant $c > 0$, and g is linear. Keeping the same birth rate, $b_i = i + 1$, changes the death rate a_i from $2i$ to $2i + 3$ (resp. $2i + 4 + \sqrt{2}$), which leads to the change of λ_1 from one to two (resp. three). More surprisingly, the eigenfunction g is changed from linear to quadratic (resp. triple). For the other values of a_i between $2i$, $2i + 3$ and $2i + 4 + \sqrt{2}$, λ_1 is unknown since g is non-polynomial. As seen from these examples, the first eigenvalue is very sensitive. Hence, in general, it is very hard to estimate λ_1.

In the next section, we find that this topic is studied extensively in Riemannian geometry.

II. New variational formula for the first eigenvalue

2.1 Story of estimating λ_1 in geometry. At first, we recall the study of λ_1 in geometry.

Consider Laplacian Δ on a compact Riemannian manifold (M, g), where g is the Riemannian metric. The spectrum of Δ is discrete:

$$\cdots \leqslant -\lambda_2 \leqslant -\lambda_1 < -\lambda_0 = 0$$

(may be repeated). Estimating these eigenvalues λ_k (especially λ_1) is very important in modern geometry. As far as we know, five books, excluding those books on general spectral theory, have been devoted to this topic: Chavel (1984), Bérard (1986), Schoen and Yau (1988), Li (1993) and Ma (1993). For a manifold M, denote its dimension, diameter and the lower bound of Ricci curvature by d, D, and K ($\text{Ricci}_M \geqslant Kg$), respectively. We are interested in estimating λ_1 in terms of these three geometric quantities. It is relatively easy to obtain an upper bound by applying a test function $f \in C^1(M)$ to the classical variational formula:

$$\lambda_1 = \inf\left\{\int_M \|\nabla f\|^2 \mathrm{d}x :\ f \in C^1(M),\ \int f \mathrm{d}x = 0,\ \int f^2 \mathrm{d}x = 1\right\}, \tag{2.0}$$

where "dx" is the Riemannian volume element. To obtain the lower bound, however, is much harder. In Table 2.1, we list eight of the strongest lower bounds that have been derived in the past, using various sophisticated methods.

Author(s)	**Lower bound**
A. Lichnerowicz (1958)	$\boldsymbol{\dfrac{d}{d-1}K, \quad K \geqslant 0.}$ (2.1)
P. H. Bérard, G. Besson & S. Gallot (1985)	$\boldsymbol{d\left\{\dfrac{\int_0^{\pi/2}\cos^{d-1}t\mathrm{d}t}{\int_0^{D/2}\cos^{d-1}t\mathrm{d}t}\right\}^{2/d}, \quad K = d-1 > 0.}$ (2.2)
P. Li & S. T. Yau (1980)	$\dfrac{\pi^2}{2\,D^2}, \quad K \geqslant 0.$ (2.3)
J. Q. Zhong & H. C. Yang (1984)	$\boldsymbol{\dfrac{\pi^2}{D^2}, \quad K \geqslant 0.}$ (2.4)
P. Li & S. T. Yau (1980)	$\dfrac{1}{D^2(d-1)\exp\left[1+\sqrt{1+16\alpha^2}\right]}, \quad K \leqslant 0.$ (2.5)
K. R. Cai (1991)	$\boldsymbol{\dfrac{\pi^2}{D^2} + K, \quad K \leqslant 0.}$ (2.6)
H. C. Yang (1989) & F. Jia (1991)	$\boldsymbol{\dfrac{\pi^2}{D^2}e^{-\alpha}, \quad \text{if } d \geqslant 5, \quad K \leqslant 0.}$ (2.7)
H. C. Yang (1989) & F. Jia (1991)	$\dfrac{\pi^2}{2\,D^2}e^{-\alpha'}, \quad \text{if } 2 \leqslant d \leqslant 4, \quad K \leqslant 0,$ (2.8)

Table 1.2 Ten lower bounds of λ_1

In Table 2.1, the two parameters α and α' are defined as $\alpha = D\sqrt{|K|(d-1)}/2$ and $\alpha' = D\sqrt{|K|((d-1)\vee 2)}/2$. Among these estimates, five ((2.1), (2.2), (2.4), (2.6) and (2.7)) are sharp. The first two are sharp for the unit sphere in two or higher dimensions but fail for the unit circle; the fourth, the sixth, and the seventh are all sharp for the unit circle. As seen from this table, the picture is now very complete, due to the efforts of many geometers in the past 40 years. Our original starting point is to learn from the geometers and to study their methods, especially the recent new developments. In the next section, we will show that one can go in the opposite direction, i.e., studying the first eigenvalue by using probabilistic methods. Exceeding our expectations, we find a general formula for the lower bound.

2.2 New variational formula. Before stating our new variational formula, we introduce two notations:

$$C(r) = \cosh^{d-1}\left[\frac{r}{2}\sqrt{\frac{-K}{d-1}}\right],\ r \in (0, D). \qquad \mathscr{F} = \{f \in C[0, D] : f > 0 \text{ on } (0, D)\}.$$

Here, we have used all the three quantities: the dimension d, the diameter D, and the lower bound K of Ricci curvature.

Theorem 2.1[General formula] (Chen & Wang (1997a)).

$$\lambda_1 \geqslant \sup_{f\in\mathscr{F}} \inf_{r\in(0,D)} \frac{4f(r)}{\int_0^r C(s)^{-1}\mathrm{d}s \int_s^D C(u)f(u)\mathrm{d}u} =: \xi_1. \tag{2.9}$$

The new variational formula has its essential value in estimating the lower bound. It is a dual of the classical variational formula in the sense that “inf” in (2.0) is replaced by “sup” in (2.9). The classical formula can be traced to Lord S. J. W. Rayleigh (1877) and E. Fischer (1905). Noticing that these two formulas (2.0) and (2.9) look very different, which explains that why such a formula (2.9) has never appeared before. This formula can produce many new lower bounds. For instance, the one corresponding to the trivial function $f \equiv 1$ is non-trivial in geometry. Applying the general formula to the test functions $\sin(\alpha r)$ and $\cosh^{1-d}(\alpha r)\sin(\beta r)$ with $\alpha = 2^{-1}\sqrt{|K|/(d-1)}$ and $\beta = \pi/(2D)$, we obtain the following:

Corollary 2.2 (Chen & Wang (1997a)).

$$\lambda_1 \geqslant \frac{dK}{d-1}\left\{1 - \cos^d\left[\frac{D}{2}\sqrt{\frac{K}{d-1}}\right]\right\}^{-1}, \qquad d > 1, \quad K \geqslant 0 \tag{2.10}$$

$$\lambda_1 \geqslant \frac{\pi^2}{D^2}\sqrt{1 - \frac{2D^2K}{\pi^4}}\cosh^{1-d}\left[\frac{D}{2}\sqrt{\frac{-K}{d-1}}\right], \quad d > 1,\ K \leqslant 0. \tag{2.11}$$

Applying this formula to some very complicated test functions, we can prove the following result:

Corollary 2.3 (Chen, Scacciatelli and Yao (2002)).

$$\lambda_1 \geqslant \pi^2/D^2 + K/2, \qquad K \in \mathbb{R}. \tag{2.12}$$

The corollaries improve all the estimates (2.1)—(2.8). Especially, (2.10) improves (2.1) and (2.2), (2.11) improves (2.7) and (2.8), and (2.12) improves

(2.3) and (2.6). Moreover, the linear approximation in (2.12) is optimal in the sense that the coefficient $1/2$ of K is exact.

A test function is indeed a mimic of the eigenfunction, so it should be chosen appropriately in order to obtain good estimates. A question arises naturally: does there exist a single representative test function such that we can avoid the task of choosing a different test function each time? The answer is seemingly negative since we have already seen that the eigenvalue and the eigenfunction are both very sensitive. Surprisingly, the answer is affirmative. The representative test function, though very tricky to find, has a rather simple form: $f(r) = \sqrt{\int_0^r C(s)^{-1}\mathrm{d}s}$. This is motivated from the study of the weighted Hardy inequality, a powerful tool in harmonic analysis (cf. Muckenhoupt (1972), Opic and Kufner (1990)).

Corollary 2.4 (Chen (2000)). For the lower bound ξ_1 of λ_1 given in Theorem 2.1, we have $4\delta^{-1} \geqslant \xi_1 \geqslant \delta^{-1}$, where

$$\delta = \sup_{r\in(0,D)} \left(\int_0^r C(s)^{-1}\mathrm{d}s\right)\left(\int_r^D C(s)\mathrm{d}s\right),$$
$$C(s) = \cosh^{d-1}\left[\frac{s}{2}\sqrt{\frac{-K}{d-1}}\,\right]. \tag{2.13}$$

Theorem 2.1 and its corollaries are also valid for manifolds with a convex boundary endowed with the Neumann boundary condition. In this case, the estimates (2.1)—(2.8) are conjectured by the geometers to be correct. However, only the Lichnerowicz's estimate (2.1) was proven by J. F. Escobar in 1990. The others in (2.2)—(2.8) and furthermore in (2.10)—(2.13) are all new in geometry.

On the one hand, the proof of this theorem is quite straightforward, based on the coupling introduced by Kendall (1986) and Cranston (1991). On the other hand, the derivation of this general formula requires much effort. The key point is to find a way to mimic the eigenfunctions. For more details, refer to Chen (1997).

Applying similar proof techniques to general Markov processes, we also obtain variational formulas for non-compact manifolds, elliptic operators in $\mathbb{R}^d$ (Chen and Wang (1997b)), and Markov chains (Chen (1996)). It is more difficult to derive the variational formulas for the elliptic operators and Markov chains due to the presence of infinite parameters in these cases. In contrast, there are only three parameters (d , D, and K) in the geometric case. In fact, formula (2.9) is a particular example of our general formula (which is complete in dimensional one) for elliptic operators.

To conclude this part, we return to the matrix case introduced at the beginning of the paper.

2.3 Birth-death processes. Let $b_i > 0(i \geqslant 0)$ and $a_i > 0(i \geqslant 1)$ be the birth and death rates, respectively. Define

$$\mu_0 = 1, \quad \mu_i = \frac{b_0 \cdots b_{i-1}}{a_1 \cdots a_i} \quad (i \geqslant 1).$$

Assume that the process is non-explosive:

$$\sum_{k=0}^{\infty}(b_k\mu_k)^{-1}\sum_{i=0}^{k}\mu_i=\infty \qquad \text{and moreover} \qquad \mu=\sum_i \mu_i<\infty. \tag{2.14}$$

The corresponding Dirichlet form is

$$D(f)=\sum_i \pi_i b_i (f_{i+1}-f_i)^2, \qquad \mathscr{D}(D)=\{f\in L^2(\pi): D(f)<\infty\}.$$

Here and in what follows, only the diagonal elements $D(f)$ are written, but the non-diagonal elements can be computed from the diagonal ones by using the quadrilateral role. We then have the classical formula

$$\lambda_1=\big\{D(f):\pi(f)=0,\pi\big(f^2\big)=1\big\}.$$

Define

$$\begin{aligned}\mathscr{F}'=\{f: f_0=0, &\text{there exists } k: 1\leqslant k\leqslant\infty \text{ so that } f_i=f_{i\wedge k}\\ &\text{and } f \text{ is strictly increasing in } [0,k]\},\\ \mathscr{F}''=\{f: f_0=0,\ &f \text{ is strictly increasing}\},\end{aligned}$$

and

$$I_i(f)=\frac{1}{\mu_i b_i(f_{i+1}-f_i)}\sum_{j\geqslant i+1}\mu_j f_j.$$

Let $\bar f=f-\pi(f)$. Then we have the following results:

Theorem 2.5 (Chen (1996, 2000, 2001))[1]. Under (2.14), we have

(1) Dual variational formula.

$$\inf_{f\in\mathscr{F}'}\sup_{i\geqslant 1} I_i(\bar f)^{-1}=\lambda_1=\sup_{f\in\mathscr{F}''}\inf_{i\geqslant 0} I_i(\bar f)^{-1}.$$

(2) Explicit estimate. $\mu\delta^{-1}\geqslant\lambda_1\geqslant(4\delta)^{-1}$, where

$$\delta=\sup_{i\geqslant 1}\sum_{j\leqslant i-1}(\mu_j b_j)^{-1}\sum_{j\geqslant i}\mu_j.$$

(3) Approximation procedure. There exist explicit sequences η_n' and η_n'' such that

$${\eta_n'}^{-1}\geqslant\lambda_1\geqslant{\eta_n''}^{-1}\geqslant(4\delta)^{-1}.$$

Here the word "dual" means that the upper and lower bounds are interchangeable if one exchanges "sup" and "inf". With slight modifications, this result is also valid for finite matrices, refer to Chen (1999).

[1]Due to the limitation of the space, the most of the author's papers during 1993–2001 are not listed in References, the readers are urged to refer to [11].

Ⅲ. Basic inequalities and new forms of Cheeger's constants

3.1 Basic inequalities. We now go to a more general setup. Let $(E, \mathcal{E}, \pi)$ be a probability space satisfying $\{(x,x) : x \in E\} \in \mathcal{E} \times \mathcal{E}$. Denote by $L^p(\pi)$ the usual real L^p-space with norm $\|\cdot\|_p$. Write $\|\cdot\| = \|\cdot\|_2$.

For a given Dirichlet form $(D, \mathcal{D}(D))$, the classical variational formula for the first eigenvalue λ_1 can be rewritten in the form of (3.1) below with an optimal constant $C = \lambda_1^{-1}$. From this point of view, it is natural to study other inequalities. Two additional basic inequalities appear in (3.2) and (3.3) below.

$$\textit{Poincaré inequality}: \qquad \mathrm{Var}(f) \leqslant CD(f), \qquad f \in L^2(\pi), \tag{3.1}$$

$$\textit{Logarithmic Sobolev inequality}: \int f^2 \log \frac{f^2}{\|f\|^2} \mathrm{d}\pi \leqslant CD(f),\ f \in L^2(\pi), \tag{3.2}$$

$$\textit{Nash inequality}: \qquad \mathrm{Var}(f) \leqslant CD(f)^{1/p} \|f\|_1^{2/q},\ f \in L^2(\pi), \tag{3.3}$$

where $\mathrm{Var}(f) = \pi(f^2) - \pi(f)^2$, $\pi(f) = \int f \mathrm{d}\pi$, $p \in (1, \infty)$ and $1/p + 1/q = 1$. The last two inequalities are due to Gross (1976) and Nath (1958), respectively.

Our main object is a symmetric (not necessarily Dirichlet) form $(D, \mathcal{D}(D))$ on $L^2(\pi)$, corresponding to an integral operator (or symmetric kernel) on $(E, \mathcal{E})$:

$$D(f) = \frac{1}{2} \int_{E \times E} J(\mathrm{d}x, \mathrm{d}y)[f(y) - f(x)]^2, \quad \mathcal{D}(D) = \{f \in L^2(\pi) : D(f) < \infty\}, \tag{3.4}$$

where J is a non-negative, symmetric measure having no charge on the diagonal set $\{(x,x) : x \in E\}$. A typical example is the reversible jump process with a q-pair $(q(x),\ q(x, \mathrm{d}y))$ and a reversible measure π. Then $J(\mathrm{d}x, \mathrm{d}y) = \pi(\mathrm{d}x) q(x, \mathrm{d}y)$.

For the remainder of this part, we restrict our discussions to the symmetric form of (3.4).

3.2 Status of the research. An important topic in this research area is to study under what conditions on the symmetric measure J do the above inequalities hold. In contrast with the probabilistic method used in Part (Ⅱ), here we adopt a generalization of Cheeger's method (1970), which comes from Riemannian geometry. Naturally, we define

$$\lambda_1 := \inf\{D(f) : \pi(f) = 0,\ \|f\| = 1\}.$$

For bounded jump processes, the fundamental known result is the following:

Theorem 3.1 (Lawler & Sokal (1988)).

$$\lambda_1 \geqslant \frac{k^2}{2M},$$

where

$$k = \inf_{\pi(A)\in(0,1)} \frac{\int_A \pi(\mathrm{d}x)q(x,A^c)}{\pi(A)\wedge\pi(A^c)} \qquad \text{and} \qquad M = \sup_{x\in E} q(x).$$

In the past years, the theorem has been collected into six books: Chen (1992), Sinclair (1993), Chung (1997), Saloff-Coste (1997), Colin de Verdière (1998), Aldous, D. G. & Fill, J. A. (1994–). From the titles of the books, one can see a wide range of the applications. However, this result fails for the unbounded operator. Thus, it has been a challenging open problem in the past ten years to handle the unbounded case.

As for the logarithmic Sobolev inequality, there have been a large number of publications in the past twenty years for differential operators. (For a survey, see Bakry (1992) or Gross (1993)). Still, there are very limited results for integral operators.

3.3 New results. Since the symmetric measure can be unbounded, we choose a symmetric, non-negative function $r(x,y)$ such that

$$J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y) := I_{\{r(x,y)^\alpha>0\}}\frac{J(\mathrm{d}x,\mathrm{d}y)}{r(x,y)^\alpha}\ (\alpha>0)\quad\text{satisfies}\quad \frac{J^{(1)}(\mathrm{d}x,E)}{\pi(\mathrm{d}x)}\leqslant 1,\ \pi\text{-a.s.}$$

For convenience, we use the convention $J^{(0)} = J$. Corresponding to the three inequalities above, we introduce the following new forms of Cheeger's constants.

Inequality	**Constant** $k^{(\alpha)}$	
Poincaré	$\inf_{\pi(A)\in(0,1)} \frac{J^{(\alpha)}(A\times A^c)}{\pi(A)\wedge\pi(A^c)}$	(Chen and Wang(1998))
Log. Sobolev	$\lim_{r\to 0}\inf_{\pi(A)\in(0,r]} \frac{J^{(\alpha)}(A\times A^c)}{\pi(A)\sqrt{\log[e+\pi(A)^{-1}]}}$	(Wang (2001a))
Log. Sobolev	$\lim_{\delta\to\infty}\inf_{\pi(A)>0} \frac{J^{(\alpha)}(A\times A^c)+\delta\pi(A)}{\pi(A)\sqrt{1-\log\pi(A)}}$	(Chen (2000b))
Nash	$\inf_{\pi(A)\in(0,1)} \frac{J^{(\alpha)}(A\times A^c)}{[\pi(A)\wedge\pi(A^c)]^{(2q-3)/(2q-2)}}$	(Chen (1999b))

Table 3.1 New forms of Cheeger's constants

Our main result can be easily stated as follows.

Theorem 3.2. $k^{(1/2)} > 0 \Longrightarrow$ the corresponding inequality holds.

In other words, we use $J^{(1/2)}$ and $J^{(1)}$ to handle the unbounded J. The first two kernels come from the use of Schwarz inequality. This result is proven in four papers quoted in Table (3.1). In these papers, some estimates which are sharp or qualitatively sharp for the upper or lower bounds are also presented.

IV. New picture of ergodic theory and explicit criteria

4.1 Importance of the inequalities. Let $(P_t)_{t\geqslant 0}$ be the semigroup determined by a Dirichlet form $(D, \mathcal{D}(D))$. Then, various applications of the inequalities are based on the following results:

Theorem 4.1 (Liggett (1989), Gross (1976) and Chen (1999)).

(1) Poincaré inequality

$$\Longleftrightarrow \|P_t f - \pi(f)\|^2 = \mathrm{Var}(P_t f) \leqslant \mathrm{Var}(f)\exp[-2\lambda_1 t].$$

(2) Logarithmic Sobolev inequality $\Longrightarrow$ exponential convergence in entropy:

$$\mathrm{Ent}(P_t f) \leqslant \mathrm{Ent}(f)\exp[-2\sigma t],$$

where

$$\mathrm{Ent}(f) = \pi(f\log f) - \pi(f)\log\|f\|_1.$$

(3) Nash inequality $\Longleftrightarrow \mathrm{Var}(P_t f) \leqslant C\|f\|_1/t^{1-q}$.

In the context of diffusions, one can replace "$\Longrightarrow$" by "$\Longleftrightarrow$" in part (2). Therefore, the above inequalities describe some type of L^2-ergodicity for the semigroup $(P_t)_{t\geqslant 0}$. These inequalities have become powerful tools in the study on infinite-dimensional mathematics (phase transitions, for instance) and the effectiveness of random algorithms.

4.2 Three traditional types of ergodicity. The following three types of ergodicity are well known for Markov processes.

Ordinary ergodicity: $\lim_{t\to\infty}\|p_t(x,\cdot)-\pi\|_{\mathrm{Var}} = 0$

Exponential ergodicity: $\|p_t(x,\cdot)-\pi\|_{\mathrm{Var}} \leqslant C(x)e^{-\alpha t}$ for some $\alpha > 0$

Strong ergodicity: $\lim_{t\to\infty}\sup_x\|p_t(x,\cdot)-\pi\|_{\mathrm{Var}} = 0$

$\Longleftrightarrow \lim_{t\to\infty} e^{\beta t}\sup_x\|p_t(x,\cdot)-\pi\|_{\mathrm{Var}} = 0$ for some $\beta > 0$

where $p_t(x,\mathrm{d}y)$ is the transition function of the Markov process and $\|\cdot\|_{\mathrm{Var}}$ is the total variation norm. They obey the following implications:

$$\text{Strong ergodicity} \Longrightarrow \text{Exponential ergodicity} \Longrightarrow \text{Ordinary ergodicity}.$$

It is natural to ask the following question. does there exist any relation between the above inequalities and the three traditional types of ergodicity?

4.3 New picture of ergodic theory.

Theorem 4.2 (Chen (1999), ...). For reversible Markov processes with densities, we have the diagram shown in Figure 4.1. In Figure 4.1, L^2-*algebraic ergodicity* means that

$$\mathrm{Var}(P_t f) \leqslant CV(f)t^{1-q}\,(t>0)$$

holds for some V having the properties (cf. Liggett (1991)): V is homogeneous of degree two (in the sense that

$$V(cf+d) = c^2V(f)$$

for any constants c and d) and $V(f) < \infty$ for all functions f with finite support.

$$\begin{array}{ccc}
 & \text{Nash inequality} & \\
\swarrow\!\!\swarrow & & \searrow\!\!\searrow \\
\text{Logarithmic Sobolev inequality} & & L^1\text{-exponential convergence} \\
\Downarrow & & \Updownarrow \\
\text{Exponential convergence in entropy} & & \text{Strong ergodicity} \\
\Downarrow & & \Downarrow \\
\text{Poincaré inequality} & \Longleftrightarrow & \text{Exponential ergodicity} \\
 & \Downarrow & \\
 & L^2\text{-algebraic ergodicity} & \\
 & \Downarrow & \\
 & \text{Ordinary ergodicity} &
\end{array}$$

Figure 4.1 Diagram of nine types of ergodicity

In Figure 4.1, L^2-*algebraic ergodicity* means that

$$\mathrm{Var}(P_t f) \leqslant CV(f)t^{1-q} \quad (t > 0)$$

holds for some V having the properties (cf. Liggett (1991)): V is homogeneous of degree two (in the sense that $V(cf + d) = c^2 V(f)$ for any constants c and d), $V(f) < \infty$ for all functions f with finite support. The L^1-exponential convergence means that

$$\|P_t f - \pi(f)\|_1 \leqslant C\|f - \pi(f)\|_1 e^{-\varepsilon t}$$

for some constants $\varepsilon > 0$ and $C\ (\geqslant 1)$ and for all $t \geqslant 0$.

The diagram is complete in the following sense: each single-side implication can not be replaced by double-sides one. Moreover, strong ergodicity and logarithmic Sobolev inequality (resp. exponential convergence in entropy) are not comparable. With exception of the equivalences, all the implications in the diagram are suitable for more general Markov processes. Clearly, the diagram extends the ergodic theory of Markov processes.

The diagram was presented in Chen (1999), originally for Markov chains only. Recently, the equivalence of L^1-exponential convergence and strong ergodicity was mainly proven by Y. H. Mao. A counter-example of diffusion was constructed by Wang (2001b) to show that strong ergodicity does not imply exponential convergence in entropy. For other references and a detailed proof of the diagram, refer to Chen (1999).

4.4 Explicit criteria for several types of ergodicity. As an application of the diagram in Figure 4.1, we obtain a criterion for the exponential ergodicity of birth-death processes, as listed in Table 4.2. To achieve this, we use the equivalence of exponential ergodicity and Poincaré inequality, as well as the explicit criterion for Poincaré inequality given in part (3) of Theorem 2.5. This solves a long standing open problem in the study of Markov chains (cf. Anderson (1991), §6.6 and Chen (1992), §4.4).

Next, it is natural to look for some criteria for other types of ergodicity. To do so, we consider only the one-dimensional case. Here we focus on the birth-death processes since the one-dimensional diffusion processes are in parallel. The criterion for strong ergodicity was obtained recently by Zhang, Lin and Hou (2000), and extended by Zhang (2001), using a different approach, to a larger class of Markov chains. The criteria for logarithmic Sobolev, Nash inequalities, and the discrete spectrum (no continuous spectrum and all eigenvalues have finite multiplicity) were obtained by Bobkov and Götze (1999) and Mao (2000, 2002a,b), respectively, based on the weighted Hardy inequality (see also Miclo (1999), Wang (2000), Gong and Wang (2002)). It is understood now the results can also be deduced from generalizations of the variational formulas discussed in this paper (cf. Chen (2001b)). Finally, we summarize these results in Theorem 4.3 and Table 4.2. The table is arranged in such an order that the property in the latter line is stronger than the property in the former line. The only exception is that even though the strong ergodicity is often stronger than the logarithmic Sobolev inequality, they are not comparable in general, as mentioned in Part Ⅲ.

Theorem 4.3 (Chen (2001a)). For birth-death processes with birth rates $b_i (i \geqslant 0)$ and death rates $a_i (i \geqslant 1)$, ten criteria are listed in Table 4.2. Recall the sequence (μ_i) defined in Part II and set $\mu[i,k] = \sum_{i \leqslant j \leqslant k} \mu_j$. The notion "$(*)$ & $\cdots$" appeared in Table 4.2 means that one requires the uniqueness condition in the first line plus the condition "$\cdots$".

Property	**Criterion**
Uniqueness	$\displaystyle\sum_{n\geqslant 0} \frac{1}{\mu_n b_n} \mu[0,n] = \infty \quad (*)$
Recurrence	$\displaystyle\sum_{n\geqslant 0} \frac{1}{\mu_n b_n} = \infty$
Ergodicity	$(*)$ & $\mu[0,\infty) < \infty$
Exponential ergodicity L^2-exp. convergence	$(*)$ & $\displaystyle\sup_{n\geqslant 1} \mu[n,\infty) \sum_{j\leqslant n-1} \frac{1}{\mu_j b_j} < \infty$
Discrete spectrum	$(*)$ & $\displaystyle\lim_{n\to\infty} \mu[n,\infty) \sum_{0\leqslant j\leqslant n-1} \frac{1}{\mu_j b_j} = 0$
Log. Sobolev inequality	$(*)$ & $\displaystyle\sup_{n\geqslant 1} \mu[n,\infty) \log[\mu[n,\infty)^{-1}] \sum_{j\leqslant n-1} \frac{1}{\mu_j b_j} < \infty$
Strong ergodicity L^1-exp. convergence	$(*)$ & $\displaystyle\sum_{n\geqslant 0} \frac{1}{\mu_n b_n} \mu[n+1,\infty) = \sum_{n\geqslant 1} \mu_n \sum_{j\leqslant n-1} \frac{1}{\mu_j b_j} < \infty$
Nash inequality	$(*)$ & $\displaystyle\sup_{n\geqslant 1} \mu[n,\infty)^{(q-2)/(q-1)} \sum_{j\leqslant n-1} \frac{1}{\mu_j b_j} < \infty$

Table 4.2 Ten criteria for birth-death processes

Added in Proof. We remark that in the original paper, for the Nash inequality there is an extra condition which is removed by the following paper:
Wang, J. (2009), Criteria for functional inequalities for ergodic birth-death processes, Acta Math. Sin. 2012, 28:2, 357–370.

References

[1] Anderson, W. J. (1991), *Continuous-Time Markov Chains*, Springer Series in Statistics.

[2] Aldous, D. G. & Fill, J. A. (1994–), *Reversible Markov Chains and Random Walks on Graphs* **www. stat.Berkeley.edu/users/aldous/book.html**.

[3] Bakry, D. (1992), *L'hypercontractivité et son utilisation en théorie des semigroupes*, LNM, Springer-Verlag, **1581**.

[4] Bérard, P. H. (1986), *Spectral Geometry: Direct and Inverse Problem*, LNM. vol. 1207, Springer-Verlag.

[5] Bérard, P. H., Besson, G., Gallot, S. (1985), *Sur une inéqualité isopérimétrique qui généralise celle de Paul Lévy-Gromov*, *Invent. Math.* 80: 295–308.

[6] Bobkov, S. G. and Götze, F. (1999), *Exponential integrability and transportation cost related to logarithmic Sobolev inequalities*, J. Funct. Anal. 163, 1–28.

[7] Cai, K. R. (1991), *Estimate on lower bound of the first eigenvalue of a compact Riemannian manifold*, Chin. Ann. of Math. 12(B):3, 267–271.

[8] Chavel, I. (1984), *Eigenvalues in Riemannian Geometry*, Academic Press.

[9] Cheeger, J. (1970), *A lower bound for the smallest eigenvalue of the Laplacian, Problems in analysis, a symposium in honor of S. Bochner*, 195–199, Princeton U. Press, Princeton.

[10] Chen, M. F. (1992), *From Markov Chains to Non-Equilibrium Particle Systems*, World Scientific.

[11] Chen, M. F. (1993–2001), *Ergodic Convergence Rates of Markov Processes* (Collection of papers)*:* **www.bnu.edu.cn/~chenmf/main_eng.htm**.

[12] Chen, M. F. (2001a), *Explicit criteria for several types of ergodicity*, Chin. J. Appl. Prob. Stat. 17:2, 1–8.

[13] Chen, M. F. (2001b), *Variational formulas of Poincaré-type inequalities in Banach spaces of functions on the line*, to appear in Acta Math. Sin. Eng. Ver.

[14] Chen, M. F., Scacciatelli, E. and Yao, Liang (2002), *Linear approximation of the first eigenvalue on compact manifolds*, Sci. Sin. (A) 45:4, 450–461.

[15] Chen, M. F. and Wang, F. Y. (1997a), *General formula for lower bound of the first eigenvalue on Riemannian manifolds*, Sci. Sin. 40:4, 384–394.

[16] Chen, M. F. and Wang, F. Y. (1997b), *Estimation of spectral gap for elliptic operators*, Trans. Amer. Math. Soc. 349, 1239-1267.

[17] Chen, M. F. and Wang, F. Y. (1998), *Cheeger's inequalities for general symmetric forms and existence criteria for spectral gap*, Abstract. Chin. Sci. Bulletin 43:18, 1516–1519. Ann. Prob. 2000, 28:1, 235–257.

[18] Chung, F. R. K. (1997), *Spectral Graph Theory*, CBMS, **92**, AMS, Providence, Rhode Island.

[19] Colin de Verdière, Y. (1998), *Spectres de Graphes*, Publ. Soc. Math. France.

[20] Cranston, M. (1991), *Gradient estimates on manifolds using coupling*, J. Funct. Anal. 99, 110–124.

[21] Escobar, J. F. (1990), *Uniqueness theorems on conformal deformation of metrics, Sobolev inequalities, and an eigenvalue estimate*, Comm. Pure and Appl. Math. XLIII: 857–883.

[22] Gong, F. Z. Wang, F. Y.(2002), *Functional inequalities for uniformly integrable semigroups and application to essential spectrums*, Forum Math. 14, 293–313.

[23] Gross, L. (1976), *Logarithmic Sobolev inequalities*, Amer. J. Math. 97, 1061–1083.

[24] Gross, L. (1993), *Logarithmic Sobolev inequalities and contractivity properties of semigroups*, LNM **1563**, Springer-Verlag.

[25] Jerrum, M. R. and Sinclair, A. J. (1989), *Approximating the permanent*, SIAM J. Comput.18, 1149–1178.

[26] Jia, F. (1991), *Estimate of the first eigenvalue of a compact Riemannian manifold with Ricci curvature bounded below by a negative constant* (In Chinese), Chin. Ann. Math. 12(A):4, 496–502.

[27] Kendall, W. (1986), *Nonnegative Ricci curvature and the Brownian coupling property*, Stochastics 19, 111–129.

[28] Lawler, G. F. and Sokal, A. D.(1988), *Bounds on the L^2 spectrum for Markov chain and Markov processes: a generalization of Cheeger's inequality*, Trans. Amer. Math. Soc.309, 557–580.

[29] Li, P. (1993), *Lecture Notes on Geometric Analysis*, Seoul National U., Korea.

[30] Li, P. and Yau, S. T. (1980), *Estimates of eigenvalue of a compact Riemannian manifold*, Ann. Math. Soc. Proc. Symp. Pure Math. 36, 205–240.

[31] Lichnerowicz, A. (1958), *Geometrie des Groupes des Transformationes*, Dunod.

[32] Liggett, T. M. (1989), *Exponential L_2 convergence of attractive reversible nearest particle systems*, Ann. Prob. 17, 403-432.

[33] Liggett, T. M. (1991), *L_2 rates of convergence for attractive reversible nearest particle systems: the critical case*, Ann. Prob. 19:3, 935–959.

[34] Ma, C. Y. (1993), *The Spectrum of Riemannian Manifolds* (In Chinese), Press of Nanjing U., Nanjing.

[35] Mao, Y. H. (2000), *On empty essential spectrum for Markov processes in dimension one*, preprint.

[36] Mao, Y. H. (2002a), *The logarithmic Sobolev inequalities for birth-death process and diffusion process on the line*, Chin. J. Appl. Prob. Statis., 18(1), 94–100.

[37] Mao, Y. H. (2002b), *Nash inequalities for Markov processes in dimension one*, Acta. Math. Sin. Eng. Ser., 18(1), 147–156.

[38] Miclo, L. (1999), *An example of application of discrete Hardy's inequalities*, Markov Processes Relat. Fields 5, 319–330.

[39] Muckenhoupt, B. (1972), *Hardy's inequality with weights*, Studia Math. XLIV, 31–38.

[40] Nash, J. (1958), *Continuity of solutions of parabolic and elliptic equations*, Amer. J. Math., 80, 931–954.

[41] Opic, B. and Kufner, A. (1990), *Hardy-type Inequalities*, Longman, New York.

[42] Saloff-Coste, L. (1997), *Lectures on finite Markov chains*, LNM **1665**, 301–413, Springer-Verlag.

[43] Schoen, R. and Yau, S. T. (1988), *Differential Geometry* (In Chinese), Science Press, Beijing, China.

[44] Sinclair, A. (1993), *Algorithms for Random Generation and Counting: A Markov Chain Approach*, Birkhäuser.

[45] Wang, F. Y. (2000), *Functional inequalities for empty essential spectrum*, J. Funct. Anal. 170, 219–245.

[46] Wang, F. Y. (2001a), *Sobolev type inequalities for general symmetric forms*, Proc. Amer. Math. Soc. 128:12, 3675–3682.

[47] Wang, F. Y. (2001b), *Convergence rates of Markov semigroups in probability distances*, preprint.

[48] Yang, H. C. (1989), *Estimate of the first eigenvalue of a compact Riemannian manifold with Ricci curvature bounded below by a negative constant* (In Chi- nese), Sci. Sin.(A) 32:7, 698–700.

[49] Zhang, H. J., Lin, X. and Hou, Z. T. (2000), *Uniformly polynomial convergence for standard transition functions*, In "*Birth-death Processes*" by Hou, Z. T. et al (2000), Hunan Sci. Press, Hunan.

[50] Zhang, Y. H. (2001), *Strong ergodicity for continuous-time Markov chains*, J. Appl. Prob. 38, 270–277.

[51] Zhong, J. Q. and Yang, H. C. (1984), *Estimates of the first eigenvalue of a compact Riemannian manifolds*, Sci. Sin. 27:12, 1251–1265.

Acta Mathematica Sinica, English Series
July, 2002, Vol. 18, No. 3, pp. 417–436

VARIATIONAL FORMULAS OF POINCARÉ-TYPE INEQUALITIES IN BANACH SPACES OF FUNCTIONS ON THE LINE

Mu-Fa Chen

Department of Mathematics, Beijing Normal University, Beijing 100875, P. R. China
E-mail: mfchen@bnu.edu.cn
Home page: http://math.bnu.edu.cn/~chenmf/main_eng.htm)
Received December 13, 2001; accepted March 26, 2002.

Abstract. Motivated from the study on the logarithmic Sobolev, Nash and other functional inequalities, the variational formulas for Poincaré inequalities are extended to a large class of Banach (Orlicz) spaces of functions on the line. Explicit criteria for the inequalities to hold and explicit estimates for the optimal constants in the inequalities are presented. As a typical application, the logarithmic Sobolev constant is carefully examined.

1. Introduction. In this section, we explain the background of the study and prove one of the main results in the paper to illustrate the ideas.

Let

$$L = a(x)\mathrm{d}^2/\mathrm{d}x^2 + b(x)\mathrm{d}/\mathrm{d}x$$

be an elliptic operator on an interval $(0, D)$ $(D \leqslant \infty)$ with Dirichlet boundary at 0 and Neumann boundary at D when $D < \infty$, where a and b are Borel measurable functions and a is positive everywhere. Set $C(x) = \int_0^x b/a$, where the Lebesgue measure $\mathrm{d}x$ is often omitted. Throughout the paper, assume that

$$Z := \int_0^D e^C/a < \infty.$$

2000 *Mathematics Subject Classification.* 49R50, 34L15, 26D10.
Key words and phrases. Variational formula, Poincaré inequality, logarithmic Sobolev inequality, Orlicz space.
Research supported in part NSFC (No. 10121101), 973 Project and RFDP.

Hence, $d\mu := a^{-1}e^C dx$ is a finite measure, which is crucial in the paper. It is well known that the first Dirichlet eigenvalue λ_0 of L is equal to the reciprocal of the optimal constant A in Poincaré inequality

$$\int_0^D f^2 d\mu \leqslant A \int_0^D f'^2 e^C, \qquad f \in \mathbb{C}_d[0, D], \quad f(0) = 0, \tag{1.1}$$

where $\mathbb{C}_d$ is the set of all continuous functions, differentiable almost everywhere and having compact supports. When $D = \infty$, one should replace $[0, D]$ by $[0, D)$ but we will not mention again in what follows. The starting point of the paper is the following variational (or min-max) formulas proved in [1; Theorem 1.1] (see also [2] and [3]).

$$A \leqslant \inf_{f \in \mathscr{F}} \sup_{x \in (0,D)} I(f)(x) \tag{1.2}$$

$$A \geqslant \sup_{f \in \mathscr{F}'} \inf_{x \in (0,D)} I(f)(x) \tag{1.3}$$

where

$$I(f)(x) = (f'(x)e^{C(x)})^{-1} \int_x^D f d\mu,$$

$$\mathscr{F} = \{f \in C[0, D] \cap C^1(0, D) : f(0) = 0, f'|_{(0,D)} > 0\}$$

and $\mathscr{F}'$ is a suitable modification of $\mathscr{F}$ (see [1] for more details). Moreover, when a and b are both continuous on $[0, D]$, the inequalities in (1.2) and (1.3) all become equalities. Note that the same notation f is used in (1.1) and $\mathscr{F}$, but this should yield no confusion.

Next, consider a Banach space $(\mathbb{B}, \|\cdot\|_{\mathbb{B}}, d\mu)$ of Borel measurable functions $f : [0, D] \to \mathbb{R}$ with norm

$$\|f\|_{\mathbb{B}} = \sup_{g \in \mathscr{G}} \int |f| g d\mu, \tag{1.4}$$

for a fixed set $\mathscr{G}$, to be specified case by case later, of non-negative Borel measurable functions on $[0, D]$. Throughout the paper, assume that $1 \in \mathbb{B}$ and $\mathbb{B}$ is ideal: If $h \in \mathbb{B}$ and $|f| \leqslant |h|$, μ-a.e., then $f \in \mathbb{B}$ and $\|f\|_{\mathbb{B}} \leqslant \|h\|_{\mathbb{B}}$. Then $\mathbb{C}_d[0, D] \subset L^\infty(\mu) \subset \mathbb{B}$. Note that if a Banach space $(\mathbb{B}, \|\cdot\|_{\mathbb{B}}, \mu)$ is ideal and normal ($f_1, f_2 \in \mathbb{B}$ and $|f_1| \leqslant |f_2| \Longrightarrow \|f_1\|_{\mathbb{B}} \leqslant \|f_2\|_{\mathbb{B}}$), then by Nakano-Amemiya-Mori theorem, the dual representation (1.4) is equivalent to the order semicontinuous of the norm:

$$0 \leqslant f_n \uparrow f \in \mathbb{B} \Longrightarrow \|f_n\|_{\mathbb{B}} \to \|f\|_{\mathbb{B}}$$

(cf. [4; [page 190]). The main goal of the study is to replace the L^2-norm on the left-hand side of (1.1) with the norm $\|\cdot\|_{\mathbb{B}}$. That is, extending (1.1) into the Banach form:

$$\|f^2\|_{\mathbb{B}} \leqslant A' \int_0^D f'^2 e^C =: A' D(f), \qquad f \in \mathbb{C}_d[0, D], \quad f(0) = 0, \tag{1.5}$$

where A' is a constant. The meaning of the extension is, as did in [5] and [6], that one can establish a criterion for the logarithmic Sobolev inequality

$$\int_0^D f^2 \log\big(f^2/\pi\big(f^2\big)\big)\mathrm{d}\mu \leqslant A'' D(f), \qquad f \in \mathbb{C}_d[0,D],$$

where

$$\pi(f) = \int f \mathrm{d}\pi = \mu(f)/Z$$

and A'' is a constant, by choosing a suitable Banach (Orlicz) space $\mathbb{B}$. Along this line, other criteria are also obtained in [7–9] for the Nash inequality, empty essential spectrum and super-contractivity of Markov semigroups. Readers are urged to refer to [10] for a survey of the study on these topics. Certainly, the idea should be meaningful for other inequalities, for instance for F-Sobolev inequalities (i.e., replacing the logarithmic factor in the above inequality by $F\big(f^2/\pi\big(f^2\big)\big)$ for a suitable functional F), or equivalently, the functional inequalities introduced in [11] (see also [12]). These facts lead us to study the general form (1.5), and furthermore to estimate the optimal constant $A_{\mathbb{B}}$ in (1.5):

$$A_{\mathbb{B}} = \sup\big\{\big\|f^2\big\|_{\mathbb{B}}/D(f) : f \in \mathbb{C}_d[0,D],\ f(0)=0,\ D(f)\neq 0\big\}. \tag{1.6}$$

The expression (1.6) is an analog of the classical variational formula for the first Dirichlet eigenvalue, and is especially powerful for the lower bounds of $A_{\mathbb{B}}$. The next result (1.7) is a variational formula for the upper bounds of $A_{\mathbb{B}}$. In practice, upper bounds are more useful but much harder to handle. The formulas (1.2) and (1.3) are originally based on the eigenequation

$$Lf = -\lambda_0 f,$$

which has no meaning in the present setup. Thus, it is not obvious at all that (1.2) can be generalized to (1.7) but a generalization of (1.3) fails. The explicit bounds are presented in (1.8) and (1.9) below, they lead to the main criterion.

Theorem 1.1. Let $\mathscr{G} \ni g_0$ with $\inf g_0 > 0$. Then, we have

$$A_{\mathbb{B}} \leqslant \inf_{f\in\mathscr{F}} \sup_{x\in(0,D)} \big[f'(x)e^{C(x)}\big]^{-1}\|fI_{(x,D)}\|_{\mathbb{B}}. \tag{1.7}$$

In particular,

$$A_{\mathbb{B}} \leqslant 4 \sup_{x\in(0,D)} \varphi(x)\|I_{(x,D)}\|_{\mathbb{B}} =: 4B_{\mathbb{B}}, \tag{1.8}$$

$$A_{\mathbb{B}} \geqslant \sup_{x\in(0,D)} \varphi(x)^{-1}\big\|\varphi(x\wedge\cdot)^2\big\|_{\mathbb{B}} =: C_{\mathbb{B}}, \tag{1.9}$$

where $\varphi(x) = \int_0^x e^{-C}$. Moreover,

$$B_{\mathbb{B}} \leqslant C_{\mathbb{B}} \leqslant 2B_{\mathbb{B}}. \tag{1.10}$$

Hence $A_{\mathbb{B}} < \infty$ iff $B_{\mathbb{B}} < \infty$.

The idea of extending (1.1) into (1.5) is due to [5], where the estimates $B_{\mathbb{B}} \leqslant A_{\mathbb{B}} \leqslant 4B_{\mathbb{B}}$ were presented. The last result is based on Muckenhoupt's estimates for weighted Hardy inequality given in [13] (see also [14] and [15] for various generalizations including the $L^p(\mu)$-cases). However, this paper adopts a completely different tool: the variational formulas (1.2) and (2.1) below which have been known only recently. Thus, the most important credit of Theorem 1.1 is (1.7) which provides a new variational formula for a large class of Banach spaces but can not be deduced from the known results on the weighted Hardy inequality. The variational formulas (1.7) and (2.2) below give us not only new explicit bounds ((1.9) and (2.3)) but also sharper estimates, as illustrated by Example 3.4 below. In brief, this paper is an extension of [1] and [16] to the setup of Banach spaces.

The remainder of the paper is organized as follows. In the next section, some refinements and extensions of Theorem 1.1 are presented (Theorems 2.1 and 2.2). In section 3, the results developed in previous sections are specified to the Orlicz spaces (Theorems 3.1 and 3.2). The application of these results to the logarithmic Sobolev inequality with rather good bounds is presented in the last section (Theorem 4.3). We now conclude this section with the following proof.

Proof of Theorem 1.1. First, we remark that the condition "with compact supports" of functions used in (1.1) (resp. (1.5)) can be replaced by "contained in $L^2(\mu)$ (resp. $f^2 \in \mathbb{B}$)". To see this, let f be a continuous, a.e. differentiable function such that $f^2 \in \mathbb{B}$ and let $D = \infty$. Set $f_N = f I_{[0,N]}$ for $N < \infty$. Then, it is immediately to see that

$$\infty > D(f_N) \uparrow D(f) \leqslant \infty \quad \text{as} \quad N \uparrow \infty.$$

Moreover, since $0 \leqslant f_N^2 \uparrow f^2 \in \mathbb{B}$, we have

$$\left\|f_N^2\right\|_{\mathbb{B}} \to \left\|f^2\right\|_{\mathbb{B}} < \infty \quad \text{as} \quad N \to \infty.$$

This proves the required assertion. Furthermore, when $\inf a > 0$, one may replace $\mathbb{C}_d$ in (1.1) by smooth functions with compact supports since f_N can be approximated in the $\mathbb{H}^{1,2}$-sense by those functions on $[0, N+1]$ and

$$0 < e^C/a \leqslant (\inf a)^{-1} e^C$$

is locally bounded.

(a) Note that $1 \in \mathbb{B}$ and so

$$\sup_{g \in \mathscr{G}} \int_0^D g \mathrm{d}\mu = \|1\|_{\mathbb{B}} < \infty.$$

Hence $\mathscr{G} \subset L^1(\mu)$. Let $g_0 \in \mathscr{G}$ satisfy $\inf g_0 =: \varepsilon > 0$. Since

$$\varepsilon \mu(|f|) \leqslant \mu(|f| g_0) \leqslant \|f\|_{\mathbb{B}},$$

we have $\mathbb{B} \subset L^1(\mu)$. To prove (1.7), one may assume that the right-hand side of (1.7) is finite. Otherwise, there is nothing to do. Then, for g_0 given above, we have

$$\begin{aligned}\varepsilon \inf_{f\in\mathscr{F}} \sup_{x\in(0,D)} \frac{e^{-C(x)}}{f'(x)} \int_x^D f \mathrm{d}\mu &\leqslant \inf_{f\in\mathscr{F}} \sup_{x\in(0,D)} \frac{e^{-C(x)}}{f'(x)} \int_x^D f g_0 \mathrm{d}\mu \\ &\leqslant \inf_{f\in\mathscr{F}} \sup_{x\in(0,D)} \frac{e^{-C(x)}}{f'(x)} \|f I_{(x,D)}\|_{\mathbb{B}} < \infty.\end{aligned}$$

Hence

$$A \leqslant \inf_{f\in\mathscr{F}} \sup_{x\in(0,D)} I(f)(x) < \infty.$$

Let $g \in \mathscr{G}$ and set $g_n = g + 1/n$. At the moment, we do not require that $g_n \in \mathscr{G}$. Note that if we replace a and b by a/g_n and b/g_n respectively, then the function C remains the same and a/g_n is also positive everywhere. Because $g \in L^1(\mu)$, we have $\int_0^D g_n \mathrm{d}\mu < \infty$. The corresponding functional $I(f)$ becomes

$$\frac{e^{-C(x)}}{f'(x)} \int_x^D \frac{f e^C}{a/g_n} = \frac{e^{-C(x)}}{f'(x)} \int_x^D f g_n \mathrm{d}\mu.$$

Next, since

$$\begin{aligned}\sup_{x\in(0,D)} \frac{e^{-C(x)}}{f'(x)} \int_x^D f g \mathrm{d}\mu &\leqslant \sup_{x\in(0,D)} \frac{e^{-C(x)}}{f'(x)} \int_x^D f g_n \mathrm{d}\mu \\ &\leqslant \sup_{x\in(0,D)} \frac{e^{-C(x)}}{f'(x)} \int_x^D f g \mathrm{d}\mu + \frac{1}{n} \sup_{x\in(0,D)} I(f)(x),\end{aligned}$$

for each $f \in \mathscr{F}$ with $\sup_{x\in(0,D)} I(f)(x) < \infty$, we have

$$\lim_{n\to\infty} \sup_{x\in(0,D)} \frac{e^{-C(x)}}{f'(x)} \int_x^D f g_n \mathrm{d}\mu = \sup_{x\in(0,D)} \frac{e^{-C(x)}}{f'(x)} \int_x^D f g \mathrm{d}\mu.$$

Denote by $A(g_n)$ the correspondent optimal constant in (1.1):

$$A(g_n) = \sup_{f\in\mathscr{D}} \int_0^D f^2 g_n \mathrm{d}\mu / D(f),$$

where

$$\mathscr{D} = \{f \in \mathrm{C}_d[0, D] : f(0) = 0 \text{ and } D(f) \neq 0\}.$$

Note that $\mathscr{D}$ is independent of a and $\mathscr{G}$. Similarly, we have $A(g)$. Clearly, $A(g) \leqslant A(g_n) \downarrow$. Next,

$$A(g_n) \leqslant A(g) + n^{-1} \sup_{f\in\mathscr{D}} \int_0^D f^2 \mathrm{d}\mu / D(f) = A(g) + A/n.$$

Therefore, $A(g_n) \downarrow A(g)$ as $n \to \infty$. By (1.2), we have

$$A(g) = \lim_{n\to\infty} A(g_n) \leqslant \lim_{n\to\infty} \sup_{x\in(0,D)} \frac{e^{-C(x)}}{f'(x)} \int_x^D f g_n \mathrm{d}\mu = \sup_{x\in(0,D)} \frac{e^{-C(x)}}{f'(x)} \int_x^D f g \mathrm{d}\mu.$$

Making infimum with respect to $f \in \mathscr{F}$, it follows that

$$A(g) \leqslant \inf_{f\in\mathscr{F}} \sup_{x\in(0,D)} [f'(x)e^{C(x)}]^{-1} \int_x^D f g \mathrm{d}\mu$$

for all $g \in \mathscr{G}$.

It is now easy to complete the proof of (1.7). By definition, we have

$$\begin{aligned} A_{\mathbb{B}} &= \sup_{f\in\mathscr{D}} \left\|f^2\right\|_{\mathbb{B}}/D(f) \\ &= \sup_{f\in\mathscr{D}} \sup_{g\in\mathscr{G}} \int f^2 g \mathrm{d}\mu / D(f) \\ &= \sup_{g\in\mathscr{G}} \sup_{f\in\mathscr{D}} \int f^2 g \mathrm{d}\mu / D(f) \\ &= \sup_{g\in\mathscr{G}} A(g). \end{aligned}$$

Hence

$$\begin{aligned} A_{\mathbb{B}} &\leqslant \sup_{g\in\mathscr{G}} \inf_{f\in\mathscr{F}} \sup_{x\in(0,D)} \frac{e^{-C(x)}}{f'(x)} \int_x^D f g \mathrm{d}\mu \\ &\leqslant \inf_{f\in\mathscr{F}} \sup_{g\in\mathscr{G}} \sup_{x\in(0,D)} \frac{e^{-C(x)}}{f'(x)} \int_x^D f g \mathrm{d}\mu \\ &= \inf_{f\in\mathscr{F}} \sup_{x\in(0,D)} \frac{e^{-C(x)}}{f'(x)} \sup_{g\in\mathscr{G}} \int_x^D f g \mathrm{d}\mu \\ &= \inf_{f\in\mathscr{F}} \sup_{x\in(0,D)} \frac{e^{-C(x)}}{f'(x)} \left\|f I_{(x,D)}\right\|_{\mathbb{B}}. \end{aligned}$$

This proves (1.7). One key point here is that "sup inf $\leqslant$ inf sup", which is the main argument in the proof, is where one may lose a lot in the estimation. Fortunately, it is not our case as guaranteed by (1.8)–(1.10) and (2.3) below.

(b) For the explicit bounds, by [16], we have

$$B_1 \leqslant A \leqslant 4B_2, \qquad B_2 \leqslant B_1 \leqslant 2B_2 \tag{1.11}$$

where

$$B_2 = \sup_{x\in(0,D)} \varphi(x) \int_x^D \mathrm{d}\mu, \qquad B_1 = 2 \sup_{x\in(0,D)} \int_0^x \varphi(y) \int_y^D \mathrm{d}\mu \, \nu^{(x)}(\mathrm{d}y),$$

and

$$\nu^{(x)}(\mathrm{d}y) = \frac{e^{-C(y)}\mathrm{d}y}{\varphi(x)}$$

is a probability measure on $[0, x]$. Note that $\nu^{(x)}$ is invariant under the transform $(a, b) \to (a/g, b/g)$ used above.

If $B_2 = \infty$, equivalently, $B_1 = \infty$ by (1.11), take $g_0 \in \mathscr{G}$ so that $\inf g_0 = \varepsilon > 0$, then

$$\begin{aligned}
B_{\mathbb{B}} &= \sup_{x\in(0,D)} \varphi(x)\|I_{(x,D)}\|_{\mathbb{B}} \\
&\geqslant \sup_{x\in(0,D)} \varphi(x)\int_x^D g_0\mathrm{d}\mu \\
&\geqslant \varepsilon \sup_{x\in(0,D)} \varphi(x)\int_x^D \mathrm{d}\mu \\
&= \varepsilon B_2 = \infty.
\end{aligned}$$

Similarly,

$$C_{\mathbb{B}} = \sup_{x\in(0,D)} \varphi(x)^{-1}\big\|\varphi(x\wedge\cdot)^2\big\|_{\mathbb{B}} = \infty.$$

Hence (1.8) and (1.10) are all trivial. We will prove later that (1.9) is also trivial in this case.

Unless otherwise stated, we assume that $B_2 < \infty$ and $B_1 < \infty$. Again, set $g_n = g + 1/n$ for $g \in \mathscr{G}$. Then

$$\lim_{n\to\infty} B_2(g_n) = \lim_{n\to\infty}\sup_{x\in(0,D)} \varphi(x)\int_x^D g_n\mathrm{d}\mu = \sup_{x\in(0,D)} \varphi(x)\int_x^D g\mathrm{d}\mu = B_2(g).$$

At the same time, $\lim_{n\to\infty} B_1(g_n) = B_1(g)$. Thus, for each $g \in \mathscr{G}$, replacing B_i with $B_i(g)\,(i = 1, 2)$ and A with $A(g)$ in (1.11), the conclusions still hold. Now, the proof of (1.8) is easy:

$$\sup_{g\in\mathscr{G}} B_2(g) = \sup_{x\in(0,D)} \varphi(x)\sup_{g\in\mathscr{G}}\int_x^D g\mathrm{d}\mu = \sup_{x\in(0,D)} \varphi(x)\|I_{(x,D)}\|_{\mathbb{B}} = B_{\mathbb{B}}$$

and

$$A_{\mathbb{B}} = \sup_{g\in\mathscr{G}} A(g) \leqslant 4\sup_{g\in\mathscr{G}} B_2(g) = 4B_{\mathbb{B}}.$$

An alternative proof for the upper estimate goes as follows. First, applying (1.7) to $f = \sqrt{\varphi}$, we obtain

$$A_{\mathbb{B}} \leqslant 2\sup_{x\in(0,D)} \sqrt{\varphi(x)}\,\|\sqrt{\varphi}\, I_{(x,D)}\|_{\mathbb{B}}.$$

Then the right is bounded above by $4B_{\mathbb{B}}$ as an application of [16; Lemma 1.2].

For the lower bound, the first step is also easy:

$$\begin{aligned}\sup_{g\in\mathscr{G}} B_1(g) &= 2\sup_{g\in\mathscr{G}}\sup_{x\in(0,D)}\int_0^x \varphi(y)\int_y^D g\mathrm{d}\mu\,\nu^{(x)}(\mathrm{d}y)\\ &= 2\sup_{x\in(0,D)}\sup_{g\in\mathscr{G}}\int_0^x \varphi(y)\int_y^D g\mathrm{d}\mu\,\nu^{(x)}(\mathrm{d}y).\end{aligned}$$

However, one can not exchange the order of "$\sup_g$" and "$\int_0^x$" here. Alternatively, we use the Fubini theorem,

$$\int_0^x \varphi(y)\int_y^D g\mathrm{d}\mu\,\nu^{(x)}(\mathrm{d}y) = \int_0^D \mu(\mathrm{d}z)g(z)\int_0^{x\wedge z}\varphi(y)\nu^{(x)}(\mathrm{d}y).$$

Hence

$$\begin{aligned}\sup_{g\in\mathscr{G}} B_1(g) &= 2\sup_{x\in(0,D)}\sup_{g\in\mathscr{G}}\int_0^D\bigg[\int_0^{x\wedge z}\varphi(y)\mathrm{d}\nu^{(x)}(\mathrm{d}y)\bigg]g(z)\mu(\mathrm{d}z)\\ &= \sup_{x\in(0,D)}\varphi(x)^{-1}\big\|\varphi(x\wedge\cdot)^2\big\|_{\mathbb{B}}\\ &= C_{\mathbb{B}}.\end{aligned}$$

Here in the last step, we have used the identity:

$$\begin{aligned}\int_0^{x\wedge z}\varphi(y)\nu^{(x)}(\mathrm{d}y) &= \frac{1}{\varphi(x)}\int_0^{x\wedge z}\bigg(e^{-C(y)}\int_0^y e^{-C}\bigg)\mathrm{d}y\\ &= \frac{1}{2\varphi(x)}\int_0^{x\wedge z}\mathrm{d}\bigg(\int_0^y e^{-C}\bigg)^2\\ &= \frac{1}{2\varphi(x)}\varphi(x\wedge z)^2.\end{aligned}$$

Since the sign of equality holds at each step in the above proof when making supremum with respect to $g\in\mathscr{G}$, by (1.11), we obtain (1.10).

We now prove (1.9). Fix $x\in(0,D)$ and let $f=\varphi(x\wedge\cdot)$. Then $f\in L^\infty(\mu)\subset\mathbb{B}$. The function f is clearly absolutely continuous and so (1.5) is meaningful by the remark at the beginning of the proof. Then some simple computation shows that $D(f)=\varphi(x)$ and $\big\|f^2\big\|_{\mathbb{B}}=\|\varphi(x\wedge\cdot)^2\|_{\mathbb{B}}$. Therefore,

$$A_{\mathbb{B}}\geqslant\big\|f^2\big\|_{\mathbb{B}}/D(f)=\varphi(x)^{-1}\|\varphi(x\wedge\cdot)^2\|_{\mathbb{B}}.$$

Making supremum with respect to $x\in(0,D)$, we get (1.9) (and then also $C_{\mathbb{B}}\geqslant B_{\mathbb{B}}$). Here, $C_{\mathbb{B}}=\infty$ is allowed and in particular

$$A_{\mathbb{B}}\geqslant C_{\mathbb{B}}\geqslant B_{\mathbb{B}}=\infty$$

when $B_2=\infty$. This proves the promised assertion.

(c) The last assertion follows from (1.8)–(1.10). $\square$

Recall that the formula (1.7) is an extension of (1.2). However, the extension of (1.3) to Banach spaces $\mathbb{B}$ is still unclear since the similar proof does not work: the inequality $\sup_g \inf_x \leqslant \inf_x \sup_g$ goes to the opposite direction than what we need, even though we do have the nice expression $A_{\mathbb{B}} = \sup_{g \in \mathscr{G}} A(g)$ and both (1.2) and (1.3) are meaningful for $A(g)$. This fact and the proof of Theorem 1.1 indicate the limitation of the approach: In order to obtain some bounds in terms of $\|\cdot\|_{\mathbb{B}}$, one needs a good enough representation of the constant A in (1.1), as shown in Theorem 1.1. In particular, the multidimensional analogue of Theorem 1.1 is still unknown at the moment. Nevertheless, the one-dimensional situation often plays a critical role in the study on higher- or even infinite-dimensional cases.

2. Extension. Neumann Case. This section consists of three parts. The first one is an alternative formula of (1.7) and an iterative procedure for estimating upper bounds of $A_{\mathbb{B}}$. The second one is an extension of Theorem 1.1 to the case where the Dirichlet boundary at 0 is replaced by the Neumann one. The last one is about a variant of (1.5) and some comparisons of the corresponding norms.

Instead of $I(f)$, define

$$I\!I(f)(x) = f(x)^{-1} \int_0^x \mathrm{d}y e^{-C(y)} \int_y^D f \mathrm{d}\mu = f(x)^{-1} \int_0^D f\varphi(x \wedge \cdot) \mathrm{d}\mu,$$

where

$$f \in \mathscr{F}'' := \{f \in C[0, D], f(0) = 0, f|_{(0,D)} > 0\}.$$

It is proved in [1; Theorem 1.1] that

$$A \leqslant \inf_{f \in \mathscr{F}''} \sup_{x \in (0,D)} I\!I(f)(x). \tag{2.1}$$

Using (2.1) instead of (1.2) and following the proof of Theorem 1.1, we obtain (2.2) below.

Theorem 2.1. Let $\mathscr{G} \ni g_0$ with $\inf g_0 > 0$. Then, we have

$$A_{\mathbb{B}} \leqslant \inf_{f \in \mathscr{F}''} \sup_{x \in (0,D)} f(x)^{-1} \big\| f\varphi(x \wedge \cdot)\big\|_{\mathbb{B}}. \tag{2.2}$$

Next, let $B_{\mathbb{B}} < \infty$. Define $f_0 = \sqrt{\varphi}$, $f_n(x) = \|f_{n-1}\varphi(x \wedge \cdot)\|_{\mathbb{B}}$ and set $D_{\mathbb{B}}(n) = \sup_{x \in (0,D)} f_n/f_{n-1}$ for $n \geqslant 1$. Then, we have

$$4B_{\mathbb{B}} \geqslant D_{\mathbb{B}}(n) \downarrow \lim_{n \to \infty} D_{\mathbb{B}}(n) \geqslant A_{\mathbb{B}}. \tag{2.3}$$

Proof. For a fixed test function $f \in \mathscr{F} \subset \mathscr{F}''$ (independent of $g \in \mathscr{G}$), by the Mean Value Theorem, we always have $\sup_{x \in (0,D)} I\!I(f)(x) \leqslant \sup_{x \in (0,D)} I(f)(x)$. The same inequality holds if $\mathrm{d}\mu$ is replaced by $\mathrm{d}\mu_g := g\mathrm{d}\mu$ and so

$$\sup_{x \in (0,D)} f(x)^{-1} \big\| f\varphi(x \wedge \cdot)\big\|_{\mathbb{B}} \leqslant \sup_{x \in (0,D)} (f'(x)e^{C(x)})^{-1} \|fI_{(x,D)}\|_{\mathbb{B}}.$$

Applying this inequality to the test function $f = \sqrt{\varphi}$, we obtain $D_{\mathbb{B}}(1) \leqslant 4B_{\mathbb{B}}$ since the original upper bound $4B_2$ comes also from the same specific test function (cf. Proof of [16; Theorem 1.1]). The monotonicity of $D_{\mathbb{B}}(n)$ is simple: By definition, $f_n \leqslant D_{\mathbb{B}}(n) f_{n-1}$. Hence

$$\begin{aligned} D_{\mathbb{B}}(n+1) &= \sup_{x\in(0,D)} \|f_n\varphi(x\wedge\cdot)\|_{\mathbb{B}}/f_n(x)\\ &\leqslant D_{\mathbb{B}}(n) \sup_{x\in(0,D)} \|f_{n-1}\varphi(x\wedge\cdot)\|_{\mathbb{B}}/f_n(x)\\ &= D_{\mathbb{B}}(n) \end{aligned}$$

for all $n \geqslant 1$. On the other hand, by assumption,

$$D_{\mathbb{B}}(1) \leqslant 4B_{\mathbb{B}} < \infty,$$

we have $f_0\varphi(x\wedge\cdot) \in \mathbb{B}$. Clearly, $f_1|_{(0,D)} > 0$ by the assumption on $\mathscr{G}$. Moreover, it is easy to see that $f_1 \in C[0,D]$ by triangle inequality of the norm and the locally uniform continuity of φ. Therefore, $f_1 \in \mathscr{F}''$. Furthermore, by induction, we have $f_n \in \mathscr{F}''$ for all $n \geqslant 1$. This gives us, by (2.2), that $D_{\mathbb{B}}(n) \geqslant A_{\mathbb{B}}$ and then $\lim_{n\to\infty} D_{\mathbb{B}}(n) \geqslant A_{\mathbb{B}}$. $\square$

The estimate (2.2) can be stronger than (1.7), as shown in Part (c) of Example 3.4 below, it says that (2.2), but not (1.7), is applicable for a specific test function. The iterative procedure in Theorem 2.1 is a modification of those introduced in [1]. As a dual of the iterative procedure, one may define $f_0 = \varphi$, $f_n(x) = \|f_{n-1}\varphi(x\wedge\cdot)\|_{\mathbb{B}}$ (with some localizing procedure if necessary) and set $C_{\mathbb{B}}(n) = \inf_{x\in(0,D)} f_n/f_{n-1}$ for $n \geqslant 1$. Then we have $C_{\mathbb{B}}(n)\uparrow$ as $n\uparrow$. However, we do not have $A_{\mathbb{B}} \geqslant C_{\mathbb{B}}(n)$, due to the reason explained at the end of the last section, and so this procedure does not work in general. Fortunately, lower bounds can be obtained from (1.6) easily. As illustrated in Examples 3.4 and 4.4 below, the estimates $D_{\mathbb{B}} := D_{\mathbb{B}}(1)$ and $C_{\mathbb{B}}$ are already quite satisfactory. Thus, in what follows, we will usually write down the related definition of $D_{\mathbb{B}}$, and then the successive procedure should be automatic. Besides, as we mentioned before, for continuous a and b (on $[0,D]$), (1.2) and (2.1) are all equalities. Thus, there are some chances for which the variational formulas (1.7) and (2.2) may be exact and then we do have complete variational formulas for $A_{\mathbb{B}}$.

Next, consider the Neumann case. Replace the interval $(0,D)$ by a general one $(p,q)\,(-\infty \leqslant p < q \leqslant \infty)$. When p (resp., q) is finite, at which we adopt Neumann boundary condition. Then the Poincaré inequality becomes

$$\int_p^q \big(f-\pi(f)\big)^2 \mathrm{d}\mu \leqslant \overline{A}\int_p^q f'^2 e^C, \qquad f\in \mathbb{C}_d[p,q], \tag{2.4}$$

where $\pi(f) = \int f\mathrm{d}\pi$ and $\mathrm{d}\pi = \mathrm{d}\mu/Z$. Since it is in the ergodic situation, we assume the non-explosive condition

$$\begin{aligned} &\int_p^c e^{-C(s)}\mathrm{d}s\int_s^c e^C/a = \infty \quad \text{if } p=-\infty \quad \text{and}\\ &\int_c^q e^{-C(s)}\mathrm{d}s\int_c^s e^C/a = \infty \quad \text{if } q=\infty \end{aligned} \tag{2.5}$$

for some (equivalently, all) $c \in (p,q)$. Similar to the last section, we are looking for the extension of (2.4):

$$\big\|\big(f-\pi(f)\big)^2\big\|_{\mathbb{B}} \leqslant \overline{A}_{\mathbb{B}} \int_p^q f'^2 e^C =: \overline{A}_{\mathbb{B}} D(f), \qquad f \in \mathbb{C}_d[p,q]. \tag{2.6}$$

For a fixed point $c \in (p,q)$, according to the last section, one may consider the similar inequalities as (1.5) on (p,c) and (c,q), respectively, with Dirichlet boundary at c. Let $A_{\mathbb{B}^{1c}}$ and $A_{\mathbb{B}^{2c}}$ denote the corresponding optimal constants, respectively. Our goal is to estimate $\overline{A}_{\mathbb{B}}$ in terms of $A_{\mathbb{B}^{1c}}$ and $A_{\mathbb{B}^{2c}}$. More precisely, let

$$\mathbb{B}^{1c} = \{fI_{(p,c)} : f \in \mathbb{B}\}.$$

Then $1 \in \mathbb{B}^{1c} \subset \mathbb{B}$ since $1 \in \mathbb{B}$ and $\mathbb{B}$ is ideal. Next, let

$$\mathscr{G}^{1c} = \{gI_{(p,c)} : g \in \mathscr{G}\}, \qquad \mu^{1c} = \mu|_{(p,c)}$$

and define

$$\|f\|_{\mathbb{B}^{1c}} = \sup_{g \in \mathscr{G}^{1c}} \int_p^c |f| g \mathrm{d}\mu^{1c}.$$

Suppose that $\mathscr{G}^{1c} \subset \mathscr{G}$. Then $(\mathbb{B}^{1c}, \|\cdot\|_{\mathbb{B}^{1c}})$ is complete since so is $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$. Therefore, $(\mathbb{B}^{1c}, \|\cdot\|_{\mathbb{B}^{1c}}, \mu^{1c})$ is an ideal Banach space containing 1, and furthermore the optimal constant $A_{\mathbb{B}^{1c}}$ in (1.5) (or (1.6)) (using the same notation $D(f)$) is well defined. Similarly, one has $(\mathbb{B}^{2c}, \|\cdot\|_{\mathbb{B}^{2c}}, \mu^{2c})$ and $A_{\mathbb{B}^{2c}}$ corresponding to the interval (c,q). In terms of (1.7) and (2.7) below, it is a simple matter to write down a variational formula for the upper estimate of $\overline{A}_{\mathbb{B}}$, but we will often omit this in what follows to save notations.

Theorem 2.2. Let $\mathscr{G} \ni g_0$ with $\inf g_0 > 0$.

(1) In general, we have

$$\overline{A}_{\mathbb{B}} \leqslant \inf_{c \in (p,q)} A_{\mathbb{B}^{1c}} \vee A_{\mathbb{B}^{2c}} \leqslant \inf_{c \in (p,q)} D_{\mathbb{B}^{1c}} \vee D_{\mathbb{B}^{2c}} \leqslant 4 \inf_{c \in (p,q)} B_{\mathbb{B}^{1c}} \vee B_{\mathbb{B}^{2c}}, \tag{2.7}$$

where the constants $B_{\mathbb{B}}$ and $D_{\mathbb{B}}$'s are defined in terms of $\varphi^{1c}(x) = \int_x^c e^{-C}$ and $\varphi^{2c}(x) = \int_c^x e^{-C}$:

$$B_{\mathbb{B}^{1c}} = \sup_{x \in (p,c)} \varphi^{1c}(x) \|I_{(p,x)}\|_{\mathbb{B}^{1c}},$$

$$B_{\mathbb{B}^{2c}} = \sup_{x \in (c,q)} \varphi^{2c}(x) \|I_{(x,q)}\|_{\mathbb{B}^{2c}},$$

$$D_{\mathbb{B}^{1c}} = \sup_{x \in (p,c)} \varphi^{1c}(x)^{-1/2} \big\|\sqrt{\varphi^{1c}}\, \varphi^{1c}(x \vee \cdot)\big\|_{\mathbb{B}^{1c}},$$

$$D_{\mathbb{B}^{2c}} = \sup_{x \in (c,q)} \varphi^{2c}(x)^{-1/2} \big\|\sqrt{\varphi^{2c}}\, \varphi^{2c}(x \wedge \cdot)\big\|_{\mathbb{B}^{2c}}.$$

(2) Assume additionally that $(g_1+g_2)/2 \in \mathscr{G}$ for every $g_1 \in \mathscr{G}^{1c}$ and $g_2 \in \mathscr{G}^{2c}$. Then, we have

$$\overline{A}_{\mathbb{B}} \geqslant \frac{1}{2} \sup_{c \in (p,q)} A_{\mathbb{B}^{1c}} \wedge A_{\mathbb{B}^{2c}} \geqslant \frac{1}{2} \sup_{c \in (p,q)} C_{\mathbb{B}^{1c}} \wedge C_{\mathbb{B}^{2c}} \geqslant \frac{1}{2} \sup_{c \in (p,q)} B_{\mathbb{B}^{1c}} \wedge B_{\mathbb{B}^{2c}}, \tag{2.8}$$

where

$$C_{\mathbb{B}^{1c}}=\sup_{x\in(p,c)}\varphi^{1c}(x)^{-1}\big\|\varphi^{1c}(x\vee\cdot)^2\big\|_{\mathbb{B}^{1c}},\quad C_{\mathbb{B}^{2c}}=\sup_{x\in(c,q)}\varphi^{2c}(x)^{-1}\big\|\varphi^{2c}(x\wedge\cdot)^2\big\|_{\mathbb{B}^{2c}}.$$

(3) Let $c(A)=\sup_{f\in\mathbb{B}}\pi(|f|I_A)/\|fI_A\|_{\mathbb{B}}$. Under the above assumption on $\mathscr{G}^{1c}$, $\mathscr{G}^{2c}$ and $\mathscr{G}$, if $c(A)<\infty$ for all A with $\pi(A)\ll 1$ (i.e., $\pi(A)$ is sufficient small), then $\overline{A}_{\mathbb{B}}<\infty$ iff $B_{\mathbb{B}^{1c}}\vee B_{\mathbb{B}^{2c}}<\infty$ for some (equivalent, for all) $c\in(p,q)$. If so,

$$B_{\mathbb{B}^{1c_0}}/2\leqslant C_{\mathbb{B}^{1c_0}}/2\leqslant\overline{A}_{\mathbb{B}}\leqslant D_{\mathbb{B}^{1c_0}}\leqslant 4B_{\mathbb{B}^{1c_0}},$$

where c_0 is the unique solution to the equation $B_{\mathbb{B}^{1c}}=B_{\mathbb{B}^{2c}}$, $c\in(p,q)$.

Proof. (a) First, we prove (2.7). Let $c\in(p,q)$ and $g\in\mathscr{G}$ with $g>0$. Define $g_1=gI_{(p,c)}$ and $g_2=gI_{(c,q)}$. Then $g_1\in\mathscr{G}^{1c}$ and $g_2\in\mathscr{G}^{2c}$, by assumption. It is proved in [16; Theorem 3.3] that the optimal constant $\overline{A}(g)$ in (2.4), replacing a with a/g, satisfies $\overline{A}(g)\leqslant\inf_{c\in(p,q)}A^{1c}(g_1)\vee A^{2c}(g_2)$, where $A^{1c}(g_1)$ and $A^{2c}(g_2)$ are the corresponding optimal constants in (1.1) with respect to the intervals (p,c) and (c,q), respectively. As we did in the proof of Theorem 1.1, by a suitable approximation, one may ignore the condition "$g>0$". Then,

$$\begin{aligned}\overline{A}_{\mathbb{B}}&\leqslant\sup_{g\in\mathscr{G}}\inf_{c\in(p,q)}A^{1c}(g_1)\vee A^{2c}(g_2)\\&\leqslant\inf_{c\in(p,q)}\sup_{g\in\mathscr{G}}A^{1c}(g_1)\vee A^{2c}(g_2)\\&\leqslant\inf_{c\in(p,q)}\Big[\sup_{g\in\mathscr{G}^{1c}}A^{1c}(g)\Big]\vee\Big[\sup_{g\in\mathscr{G}^{2c}}A^{2c}(g)\Big]\\&=\inf_{c\in(p,q)}A_{\mathbb{B}^{1c}}\vee A_{\mathbb{B}^{2c}}.\end{aligned}$$

Combining this with Theorem 2.1, we obtain (2.7).

(b) Next, we prove (2.8). Note that for (2.4), there is a dual result

$$\overline{A}(g)\geqslant\sup_{c\in(p,q)}A^{1c}(g_1)\wedge A^{2c}(g_2)$$

([16; Theorem 3.2]). However, one cannot use this to prove (2.8) since the orders of "sup" and "min" are not exchangeable. Here we follow the original proof with some modification. The price for the general $\mathbb{B}$ is a new factor 1/2 and so is less sharp than the original one. Fix $c\in(p,q)$ and $\varepsilon>0$. Choose $f_1,\ f_2\geqslant 0$ such that $f_1|_{(c,q)}=0,\ f_2|_{(p,c)}=0$, $\big\|f_1^2\big\|_{\mathbb{B}^{1c}}=\big\|f_2^2\big\|_{\mathbb{B}^{2c}}=1$ and $D(f_1)\leqslant A_{\mathbb{B}^{1c}}^{-1}+\varepsilon$, $D(f_2)\leqslant A_{\mathbb{B}^{2c}}^{-1}+\varepsilon$. Next, choose $g_1\in\mathscr{G}^{1c}$ and $g_2\in\mathscr{G}^{2c}$ such that $\int_p^c f_1^2g_1\mathrm{d}\mu\geqslant 1-\varepsilon$ and $\int_c^q f_2^2g_2\mathrm{d}\mu\geqslant 1-\varepsilon$. Set $f=-\sqrt{\lambda}\,f_1+\sqrt{1-\lambda}\,f_2$, where $\lambda=\pi(f_2)^2/\big[\pi(f_1)^2+$

$\pi(f_2)^2)\big]$ is the constant so that $\pi(f) = 0$. Then

$$\begin{aligned}
D(f) &= \lambda D(f_1) + (1-\lambda) D(f_2) \\
&\leqslant \lambda\big(A_{\mathbb{B}^{1c}}^{-1} + \varepsilon\big) + (1-\lambda)\big(A_{\mathbb{B}^{2c}}^{-1} + \varepsilon\big) \\
&\leqslant \big(A_{\mathbb{B}^{1c}}^{-1} \vee A_{\mathbb{B}^{2c}}^{-1} + \varepsilon\big)\big(\lambda + (1-\lambda)\big) \\
&\leqslant \big(A_{\mathbb{B}^{1c}}^{-1} \vee A_{\mathbb{B}^{2c}}^{-1} + \varepsilon\big)\bigg(\lambda \int_p^c f_1^2 g_1 \mathrm{d}\mu + (1-\lambda)\int_c^q f_2^2 g_2 \mathrm{d}\mu + \varepsilon\bigg) \\
&= \big(A_{\mathbb{B}^{1c}}^{-1} \vee A_{\mathbb{B}^{2c}}^{-1} + \varepsilon\big)\bigg(\int_p^q \Big[\lambda f_1^2 g_1 + (1-\lambda) f_2^2 g_2\Big]\mathrm{d}\mu + \varepsilon\bigg) \\
&= 2\big(A_{\mathbb{B}^{1c}}^{-1} \vee A_{\mathbb{B}^{2c}}^{-1} + \varepsilon\big)\bigg(\int_p^q f^2\big[g_1/2 + g_2/2\big]\mathrm{d}\mu + \varepsilon/2\bigg) \\
&\leqslant 2\big(A_{\mathbb{B}^{1c}}^{-1} \vee A_{\mathbb{B}^{2c}}^{-1} + \varepsilon\big)\Big(\big\|f^2\big\|_{\mathbb{B}} + \varepsilon/2\Big).
\end{aligned}$$

Here in the last step, we have used the fact that $(g_1 + g_2)/2 \in \mathscr{G}$. Letting $\varepsilon \to 0$ and then making infimum with respect to c, we obtain the first inequality in (2.8). Then, the second and the third ones follow from Theorem 1.1.

(c) The proof of part (3) of the theorem is similar to [16; proof of Theorem 3.7]. Here, we sketch the ideas only.

Note that $I_{(p,x)}, I_{(x,q)} \in \mathbb{B}$ and so $B_{\mathbb{B}^{1c}}, B_{\mathbb{B}^{2c}} < \infty$ when p and q are both finite. We now assume that $B_{\mathbb{B}^{1c}} \vee B_{\mathbb{B}^{2c}} < \infty$ but allow p and q to be infinite. Then, it is not difficult to show that $B_{\mathbb{B}^{1c}}$ and $B_{\mathbb{B}^{2c}}$ are both continuous in c with different values at $c = p$ and $c = q$. Noting that when $c' \geqslant c$,

$$\sup_{g\in\mathscr{G}^{1c}} \int_p^x g \mathrm{d}\mu^{1c} = \sup_{g\in\mathscr{G}^{1c}} \int_p^x g \mathrm{d}\mu^{1c'} \leqslant \sup_{g\in\mathscr{G}} \int_p^x g \mathrm{d}\mu^{1c'} = \sup_{g\in\mathscr{G}^{1c'}} \int_p^x g \mathrm{d}\mu^{1c'}$$

for all $x \in (p,c)$, we have $\big\|I_{(p,x)}\big\|_{\mathbb{B}^{1c}} \leqslant \big\|I_{(p,x)}\big\|_{\mathbb{B}^{1c'}}$ for all $x \in (p,c)$. Hence, $B_{\mathbb{B}^{1c}} \uparrow$ and $B_{\mathbb{B}^{2c}} \downarrow$ strictly as $c \uparrow$. Thus, as c varies, the two curves $B_{\mathbb{B}^{1c}}$ and $B_{\mathbb{B}^{2c}}$ must intersect at a point $c_0 \in (p,q)$ uniquely, and then the required estimates follows from (2.7) and (2.8).

Next, assume that $p > -\infty$ and $q = \infty$. We need to consider only the case where $B_{\mathbb{B}^{2c}} = \infty$ since $B_{\mathbb{B}^{1c}} < \infty$. Following the proof of [16; Theorem 3.7] (or the ideas in the last paragraph), by condition (2.5), the two curves $B_{\mathbb{B}^{1c}}$ and $B_{\mathbb{B}^{2c}}$ must intersect at $c_0 = q = \infty$. Moreover, the lower bound given in (2.8) equals ∞. Therefore, the required assertions all hold. The case of $p = -\infty$ and $q < \infty$ is symmetric and so can be proved in a similar way.

It remains to consider only the case that $(p,q) = \mathbb{R}$ and one of $B_{\mathbb{B}^{1c}}$ and $B_{\mathbb{B}^{2c}}$ is infinite. We will come back to the proof soon.

Roughly speaking, Theorem 2.2 is a comparison of $\big\|f^2\big\|_{\mathbb{B}}$ and $\big\|\big(f - \pi(f)\big)^2\big\|_{\mathbb{B}}$. Alternatively, one may compare $\|f\|_{\mathbb{B}}^2$ with $\|f - \pi(f)\|_{\mathbb{B}}^2$ (If we replace $\big\|f^2\big\|_{\mathbb{B}}$ with $\|f\|_{\mathbb{B}}^2$ in (1.5), then by the Cauchy-Schwarz inequality, $\|f\|_{\mathbb{B}}^2 \leqslant \big\|f^2\big\|_{\mathbb{B}}\|1\|_{\mathbb{B}}$, hence the resulting inequality is weaker than (1.5). However, they are equivalent each other in the context of the Orlicz spaces studied in the subsequent sections). This is also used in the study of the logarithmic Sobolev and the Nash inequalities (cf. [5], [6], [7] and §4 below). The key of the comparison is as follows.

Proposition 2.3. Let $(E, \mathscr{E}, \pi)$ be a probability space and $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ be an arbitrary Banach space of Borel measurable functions on $(E, \mathscr{E}, \pi)$, containing the constant function 1.

(1) Assume that there is a constant c_1 such that $\sqrt{\pi(f^2)} \leqslant c_1 \|f\|_{\mathbb{B}}$ for all $f \in \mathbb{B}$. Then

$$\|f - \pi(f)\|_{\mathbb{B}} \leqslant \big(1 + c_1 \|1\|_{\mathbb{B}}\big) \|f\|_{\mathbb{B}}. \tag{2.9}$$

(2) Next, for a given $A \in \mathscr{E}$, let $c_2(A)$ be the constant such that $\sqrt{\pi(f^2 I_A)} \leqslant c_2(A) \|f I_A\|_{\mathbb{B}}$ for all $f \in \mathbb{B}$. If $c_2(A)\sqrt{\pi(A)}\,\|1\|_{\mathbb{B}} < 1$, then for every f with $f|_{A^c} = 0$ we have

$$\|f\|_{\mathbb{B}} \leqslant \big\|f - \pi(f)\big\|_{\mathbb{B}} \big/ \big[1 - c_2(A)\sqrt{\pi(A)}\,\|1\|_{\mathbb{B}}\big]. \tag{2.10}$$

Proof. The proof is quite easy. We follow the above quoted papers with a slight modification. First, we have

$$\|f - \pi(f)\|_{\mathbb{B}} \leqslant \|f\|_{\mathbb{B}} + |\pi(f)|\,\|1\|_{\mathbb{B}} \leqslant \|f\|_{\mathbb{B}} + \sqrt{\pi(f^2)}\,\|1\|_{\mathbb{B}}$$

by the Cauchy-Schwarz inequality. This proves (2.9).

Next, by using the Cauchy-Schwarz inequality again, for every f with $f|_{A^c} = 0$, we have $\pi(f)^2 \leqslant \pi(A)\,\pi(f^2)$. Therefore, by triangle inequality, we have

$$\|f\|_{\mathbb{B}} \leqslant \big\|f - \pi(f)\big\|_{\mathbb{B}} + |\pi(f)|\|1\|_{\mathbb{B}} \leqslant \big\|f - \pi(f)\big\|_{\mathbb{B}} + c_2(A)\sqrt{\pi(A)}\,\|1\|_{\mathbb{B}}\,\|f\|_{\mathbb{B}}.$$

Collecting the terms of $\|f\|_{\mathbb{B}}$, we obtain (2.10). $\square$

Similarly, we have the following result.

Proposition 2.4. Everything in the premise is the same as in Proposition 2.3.

(1) Assume that there is a constant c_1 such that $\pi(|f|) \leqslant c_1 \|f\|_{\mathbb{B}}$ for all $f \in \mathbb{B}$. Then

$$\big\|(f - \pi(f))^2\big\|_{\mathbb{B}} \leqslant \big(1 + \sqrt{c_1 \|1\|_{\mathbb{B}}}\,\big)^2 \big\|f^2\big\|_{\mathbb{B}}. \tag{2.11}$$

(2) Next, for a given $A \in \mathscr{E}$, let $c_2(A)$ be the constant such that $\pi(|f| I_A) \leqslant c_2(A) \|f I_A\|_{\mathbb{B}}$ for all $f \in \mathbb{B}$. If $c_2(A)\pi(A)\,\|1\|_{\mathbb{B}} < 1$, then for every f with $f|_{A^c} = 0$ we have

$$\big\|f^2\big\|_{\mathbb{B}} \leqslant \big\|(f - \pi(f))^2\big\|_{\mathbb{B}} \big/ \big[1 - \sqrt{c_2(A)\pi(A)\,\|1\|_{\mathbb{B}}}\,\big]^2. \tag{2.12}$$

Proof of Theorem 2.2 (*continued*). Let $(p, q) = \mathbb{R}$. We need to prove that $A_{\mathbb{B}} < \infty$ iff $B_{\mathbb{B}^{1c}} \vee B_{\mathbb{B}^{2c}} < \infty$. By assumption, for sufficiently small c, condition (2) of Proposition 2.4 is satisfied with $A = (-\infty, c)$. Let (2.6) hold. Then we have for every f with $f|_{A^c} = 0$, that $\big\|f^2\big\|_{\mathbb{B}^{1c}} \leqslant c' \big\|\big(f - \pi(f)\big)^2\big\|_{\mathbb{B}}$ for some $c' < \infty$. Thus, once (2.6) holds, we must have $B_{\mathbb{B}^{1c}} \leqslant A_{\mathbb{B}^{1c}} < \infty$ first for sufficiently small c, and then for all $c \in \mathbb{R}$, since $B_{\mathbb{B}^{1c}}$ is continuous in $c \in \mathbb{R}$. By symmetry, the same conclusion holds for $B_{\mathbb{B}^{2c}}$ and so $B_{\mathbb{B}^{1c}} \vee B_{\mathbb{B}^{2c}} < \infty$ for all $c \in \mathbb{R}$. This proves the necessity of the condition $B_{\mathbb{B}^{1c}} \vee B_{\mathbb{B}^{2c}} < \infty$ for $\overline{A}_{\mathbb{B}} < \infty$. The sufficiency comes from (2.7). If the equation $B_{\mathbb{B}^{1c}} = B_{\mathbb{B}^{2c}} = \infty$ holds for some $c \in \mathbb{R}$, then it also holds for all $c \in \mathbb{R}$. Otherwise, the solution $c_0 \in [p, q]$ must be unique as shown before. Of course, the case that $c_0 = \pm\infty$ is useless for us. $\square$

3. Orlicz form. In this section, the above results are specialized to the Orlicz spaces. To do so, we recall some basic notions and facts. A function Φ: $\mathbb{R} \to \mathbb{R}$ is called an N-function if it is non-negative, continuous, convex, even (i.e., $\Phi(-x) = \Phi(x)$) and satisfies the following conditions:

$$\Phi(x) = 0 \ \text{ iff } \ x = 0, \qquad \lim_{x\to 0} \Phi(x)/x = 0, \qquad \lim_{x\to\infty} \Phi(x)/x = \infty.$$

We will often assume the following growth condition (or Δ_2-condition) for Φ:

$$\sup_{x\gg 1} \Phi(2x)/\Phi(x) < \infty\Big(\iff \sup_{x\gg 1} x\Phi'_-(x)/\Phi(x) < \infty\Big),$$

where Φ'_- is the left derivative of Φ. It is interesting that this condition is also essential in the study of F-Sobolev inequalities by Cheeger's method (cf. [17; condition (2.3)], see also [18]).

Corresponding to each N-function, we have a complementary N-function:

$$\Phi_c(y) := \sup\{x|y| - \Phi(x) : x \geqslant 0\}, \qquad y \in \mathbb{R}.$$

Alternatively, let φ_c be the inverse function of Φ'_-, then $\Phi_c(y) = \int_0^{|y|} \varphi_c$. The two typical examples of pairs of N-functions are as follows. First, $\Phi(x) = |x|^p/p$ and $\Phi_c(y) = |y|^q/q$, $1/p+1/q = 1$. This corresponds to the standard L^p-spaces. Next, $\Phi(x) = (1+|x|)\log(1+|x|) - |x|$ and $\Phi_c(y) = e^{|y|} - |y| - 1$. This is related to the logarithmic Sobolev inequality mentioned before. Very often, Φ_c is not explicitly known for a given Φ, for instance when $\Phi(x) = |x|\log(1+|x|)$.

Given an N-function and a finite measure μ on $E := (p,q) \subset \mathbb{R}$, we define an Orlicz space as follows:

$$L^\Phi(\mu) = \Big\{f\,\big(E \to \mathbb{R}\big) : \int_E \Phi(f)\mathrm{d}\mu < \infty\Big\}, \qquad \|f\|_\Phi = \sup_{g\in\mathscr{G}} \int_E |f|g\mathrm{d}\mu, \qquad (3.1)$$

where

$$\mathscr{G} = \Big\{g \geqslant 0 : \int_E \Phi_c(g)\mathrm{d}\mu \leqslant 1\Big\}.$$

Under Δ_2-condition, $(L^\Phi(\mu), \|\cdot\|_\Phi, \mu)$ is a Banach space. For this, the Δ_2-condition is indeed necessary. Clearly, $L^\Phi(\mu) \ni 1$ and is ideal.

Having these preparations in mind, it is rather simple to state our first result in this section.

Theorem 3.1. For every N-function Φ satisfying Δ_2-condition, the conclusions of Theorems 1.1 and 2.1 hold with $\mathbb{B} = L^\Phi(\mu)$. Moreover,

$$\|I_{(x,D)}\|_\Phi = \mu(x,D)\Phi_c^{-1}(\mu(x,D)^{-1}) = \inf_{\alpha>0} \big(1 + \mu(x,D)\Phi(\alpha)\big)/\alpha,$$

where Φ_c^{-1} is the inverse function of Φ_c.

Proof. Clearly, $L^\Phi(\mu)$ is an ideal Banach space. Since $\lim_{x\to 0} \Phi_c(x)/x = 0$ and μ is a finite measure, $\mathscr{G}$ contains all sufficiently small constants and furthermore

$L^\Phi(\mu) \subset L^1(\mu)$. Next, for an indicator function I_B of B, by [19; §3.4, Corollary 7], we have

$$\|I_B\|_\Phi = \mu(B)\Phi_c^{-1}(\mu(B)^{-1}) < \infty. \tag{3.2}$$

This gives us again $1 \in L^\Phi(\mu)$ and then $\mathscr{G} \subset L^1(\mu)$. Combining (3.2) with (3.4) below, we obtain the last assertion. $\square$

The next result is a consequence of Theorem 2.2 and Proposition 2.4. We will use the notations introduced there. For simplicity, we write $c_2(A) = c_2(p, c)$ and $\pi(A) = \pi(p, c)$ when $A = (p, c)$.

Theorem 3.2. For every N-function Φ satisfying Δ_2-condition, the conclusions of Theorem 2.2 hold with $\mathbb{B} = L^\Phi(\mu)$. Additionally, if

$$\big[[c_2(p,c)\pi(p,c)] \vee [c_2(c,q)\pi(c,q)]\big]\, \|1\|_{\mathbb{B}} < 1,$$

then we have

$$\overline{A}_{\mathbb{B}} \geqslant \sup_{c\in(p,q)} \Big(1 - \sqrt{\|1\|_{\mathbb{B}}}\Big[[c_2(p,c)\pi(p,c)] \vee [c_2(c,q)\pi(c,q)]\Big]^{1/2}\Big)^2 \big(A_{\mathbb{B}^{1c}} \vee A_{\mathbb{B}^{2c}}\big).$$

To prove this theorem and also for the later use, we recall some known facts about the norm $\|\cdot\|_\Phi$. The next result is taken from [19; §3.3, Theorem 13 and Proposition 14].

Proposition 3.3. We have

$$\|f\|_\Phi = \inf_{\alpha>0} \frac{1}{\alpha}\bigg(1 + \int_E \Phi(\alpha f)\mathrm{d}\mu\bigg). \tag{3.3}$$

In particular,

$$\|I_B\|_\Phi = \inf_{\alpha>0} \frac{1}{\alpha}\big(1 + \mu(B)\Phi(\alpha)\big). \tag{3.4}$$

Furthermore, if there is $\alpha^*(= \alpha^*(f)) > 0$ such that

$$\int_E \big[\alpha^*|f|\Phi'_-(\alpha^*|f|) - \Phi(\alpha^* f)\big]\mathrm{d}\mu = 1, \tag{3.5}$$

where Φ'_- is the left derivative of Φ as usual, then

$$\|f\|_\Phi = \frac{1}{\alpha^*}\bigg(1 + \int_E \Phi(\alpha^* f)\mathrm{d}\mu\bigg) = \int_E |f|\Phi'_-(\alpha^*|f|)\mathrm{d}\mu, \tag{3.6}$$

Next, a more practical but equivalent norm is as follows:

$$\|f\|_{(\Phi)} = \inf\bigg\{\alpha > 0 : \int_E \Phi(f/\alpha)\mathrm{d}\mu \leqslant 1\bigg\}. \tag{3.7}$$

In particular,

$$\|I_B\|_{(\Phi)} = 1/\Phi^{-1}(\mu(B)^{-1}) \tag{3.8}$$

(cf. [19; §3.4, Corollary 7]). The comparison of these two norms is as follows:

$$\|f\|_{(\Phi)} \leqslant \|f\|_\Phi \leqslant 2\|f\|_{(\Phi)} \tag{3.9}$$

(cf. [19; §3.3, Proposition 4]). However, $\|f\|_{(\Phi)} = \|f\|_\Phi$ iff $f=0$, μ-a.e. (cf. [19; §3.3, Proposition 17]).

Proof of Theorem 3.2. Because Φ_c is convex, $\mathscr{G}$ is a convex set, the assumption in Part (2) of Theorem 2.2 about $\mathscr{G}^{1c}$, $\mathscr{G}^{1c}$ and $\mathscr{G}$ is satisfied. It remains to check that $c(A) < \infty$ for all A with $\pi(A) \ll 1$. Let $f \in L^\Phi(\mu)$ with $f \geqslant 0$. By convexity of Φ, we have

$$\Phi\bigg(\frac{\alpha}{\mu(A)}\int_A f\,\mathrm{d}\mu\bigg) \leqslant \frac{1}{\mu(A)}\int_A \Phi(\alpha f)\,\mathrm{d}\mu.$$

Recall $Z = \mu(p,q) := \int_p^q \mathrm{d}\mu$. By (3.3), we obtain

$$\big\|fI_A\big\|_\Phi = \inf_{\alpha>0}\frac{1}{\alpha}\bigg(1+\int_A \Phi(\alpha f)\,\mathrm{d}\mu\bigg) \geqslant \inf_{\alpha>0}\bigg[\frac{1}{\alpha}+\frac{\mu(A)}{\alpha}\Phi\bigg(\frac{\alpha Z\pi(fI_A)}{\mu(A)}\bigg)\bigg]. \tag{3.10}$$

Next, because of $\lim_{x\to\infty}\Phi(x)/x=\infty$ and the continuity of Φ, there exists $c'>0$ such that $\sup_{x\geqslant 1}\Phi(x)/x \geqslant c'$. Thus,

$$\begin{aligned}
&\frac{1}{\alpha}+\frac{\mu(A)}{\alpha}\Phi\bigg(\frac{\alpha Z\pi(fI_A)}{\mu(A)}\bigg)\\
&\quad\geqslant \begin{cases} \dfrac{1}{\alpha} \geqslant \dfrac{Z\pi(fI_A)}{\mu(A)}, & \text{if } \dfrac{\alpha Z\pi(fI_A)}{\mu(A)} \leqslant 1\\[2ex] \dfrac{\mu(A)}{\alpha}\cdot c'\dfrac{\alpha Z\pi(fI_A)}{\mu(A)} = c'Z\pi(fI_A), & \text{if } \dfrac{\alpha Z\pi(fI_A)}{\mu(A)} \geqslant 1. \end{cases}
\end{aligned}$$

Therefore $\big\|fI_A\big\|_\Phi \geqslant \min\{Z/\mu(A),\, c'Z\}\pi(fI_A)$. Hence the required assertion holds indeed for all A.

To prove the last assertion of Theorem 3.2, fix $c\in(p,q)$. Applying (2.6) to the function f_1 with $f_1|_{(c,q)}=0$, we have $\overline{A}_{\mathbb{B}}D(f_1) \geqslant \big\|(f_1-\pi(f_1))^2\big\|_{\mathbb{B}}$. On the other hand, by Part (2) of Proposition 2.4, we get $\big\|f_1^2\big\|_{\mathbb{B}^{1c}} \leqslant K_1\big\|(f_1-\pi(f_1))^2\big\|_{\mathbb{B}}$, where $K_1 = \big[1-\sqrt{c_2(p,c)\pi(p,c)\|1\|_{\mathbb{B}}}\,\big]^{-2}$. Therefore, $K_1\overline{A}_{\mathbb{B}}D(f) \geqslant \big\|f_1^2\big\|_{\mathbb{B}^{1c}}$. From definition of $A_{\mathbb{B}^{1c}}$, it follows that $\overline{A}_{\mathbb{B}} \geqslant K_1^{-1}A_{\mathbb{B}^{1c}}$. Symmetrically, $\overline{A}_{\mathbb{B}} \geqslant K_2^{-1}A_{\mathbb{B}^{2c}}$, where $K_2 = \big[1-\sqrt{c_2(c,q)\pi(c,q)\|1\|_{\mathbb{B}}}\,\big]^{-2}$. Thus,

$$\begin{aligned}
\overline{A}_{\mathbb{B}} &\geqslant \big(K_1^{-1}A_{\mathbb{B}^{1c}}\big)\vee\big(K_2^{-1}A_{\mathbb{B}^{2c}}\big)\\
&\geqslant \big(K_1^{-1}\wedge K_2^{-1}\big)\big(A_{\mathbb{B}^{1c}}\vee A_{\mathbb{B}^{2c}}\big)\\
&= \Big[\big(1-\sqrt{c_2(p,c)\pi(p,c)\|1\|_{\mathbb{B}}}\,\big)\wedge\big(1-\sqrt{c_2(c,q)\pi(c,q)\|1\|_{\mathbb{B}}}\,\big)\Big]^2\big(A_{\mathbb{B}^{1c}}\vee A_{\mathbb{B}^{2c}}\big).
\end{aligned}$$

Making supremum with respect to c, we obtain the required assertion. $\square$

To have a feeling about the above results, we now consider a very simple example.

Example 3.4. Let $D = 1$, $a = 1$ and $b = 0$. Then the operator L has the first Dirichlet eigenvalue $\lambda_0 = \pi^2/4$ with eigenfunction $g(x) = \sin(\pi x/2)$, and so the optimal constant in (1.1) is $A = 4/\pi^2 \approx 0.4053$.

Consider the extension from $\mathbb{B} = L^1(\mu)$ to $\mathbb{B} = L^p(\mu)$ for all $p > 1$ in the form of (1.5). That is, taking $\Phi(x) = |x|^p/p$ and then $\Phi_c(y) = |y|^q/q$, $1/p + 1/q = 1$. The case of $p = 1$ is nothing but the original (1.1) and the case of $p > 1$ corresponds to the Nash inequalities (cf. [7]). We are going to compute the bounds provided by Theorems 3.1 and 3.2.

(a) Upper bound. Since $\Phi_c^{-1}(y) = (qy)^{1/q}$ for $y \geqslant 0$, we have

$$\mu(x,1)\Phi_c^{-1}(\mu(x,1)^{-1}) = (1-x)\Phi_c^{-1}((1-x)^{-1}) = q^{1/q}(1-x)^{1/p}.$$

Next, $\varphi(x) = x$. Thus, by (1.8), we have

$$A_\Phi \leqslant 4q^{1/q} \sup_{x\in(0,1)} x(1-x)^{1/p} = \frac{4pq^{1/q}}{(p+1)^{1+1/p}} < \infty. \tag{3.11}$$

Thus, the Orlicz form of the inequality holds for all $p > 1$.

(b) Lower bounds. In view of (1.9) and (1.10), we already have a lower bound:

$$A_\Phi \geqslant \frac{pq^{1/q}}{(p+1)^{1+1/p}}. \tag{3.12}$$

We are going to compute another one by using (1.9). For this, we need to compute the norm $\|fI_{(x,1)}\|_\Phi$.

The equation (3.5):

$$1 = \int_x^1 \big[\alpha|f|\Phi'_-(\alpha|f|) - \Phi(\alpha f)\big]\mathrm{d}\mu = \frac{1}{q}\int_x^1 (\alpha|f|)^p$$

has a solution

$$\alpha^* = \alpha^*(x) = q^{1/p}/\|fI_{(x,1)}\|_p,$$

where $\|\cdot\|_p$ is the usual $L^p(\mu)$-norm. Thus, by (3.6), we have

$$\|fI_{(x,1)}\|_\Phi = \frac{1}{\alpha^*} + \frac{1}{\alpha^*}\int_x^1 (\alpha^*|f|)^p/p = \frac{1}{\alpha^*}(1+q/p) = \frac{q}{\alpha^*} = q^{1/q}\|fI_{(x,1)}\|_p. \tag{3.13}$$

Fix $x \in (0,1)$ and take $f(y) = (x\wedge y)^2$. Then

$$\|fI_{(x,1)}\|_p^p = \int_0^x y^{2p} + x^{2p}(1-x) = \frac{1}{2p+1}x^{2p}(2p+1-2px).$$

Combining this with (1.9) and (3.13), we get

$$A_\Phi \geqslant q^{1/q}\frac{1}{(2p+1)^{1/p}} \sup_{x\in(0,1)} x(2p+1-2px)^{1/p} = \frac{(2p+1)q^{1/q}}{2(p+1)^{1+1/p}}. \tag{3.14}$$

Clearly, this lower bound is bigger than (3.12).

(c) Improvement of the bounds. For upper bound, one may use (1.7). The first candidate of test function should be $f(x) = \sin(\pi x/2)$ since it is the eigenfunction of the original inequality. Surprisingly, since

$$\lim_{x\to 1} \|fI_{(x,1)}\|_p / f'(x) = \infty,$$

by (3.13), it leads to the trivial upper bound

$$A_\Phi \leqslant \sup_{x\in(0,1)} \|fI_{(x,1)}\|_\Phi / f'(x) = \infty.$$

The reason is that even though

$$4/\pi^2 \equiv I(f) \leqslant \sup_{x\in(0,1)} I\big(\sqrt{\varphi}\big)(x) \quad (\leqslant 4B_2),$$

this inequality is no longer true when $\mathrm{d}\mu$ in $I(f)$ is replaced with $\mathrm{d}\mu_g$. This shows that on the one hand the constant A_Φ is quite sensitive to test functions, and on the other hand we are lucky to have the same representative test functions ($f = \varphi^\gamma$ for $\gamma = 1$ or $1/2$, independent of a) which deduce the explicit bounds for all $\mathbb{B}$ (cf. the last paragraph of proof (b) of Theorem 1.1 and the proof of Theorem 2.1).

However, Theorem 2.1 is applicable for this test function $f(x) = \sin(\pi x/2)$. First, by (3.13), we have

$$\|f\varphi(x\wedge\cdot)\|_\Phi = q^{1/q}\|f\varphi(x\wedge\cdot)\|_p = q^{1/q}\bigg[\int_0^x \Big(y\sin\frac{\pi y}{2}\Big)^p + x^p\int_x^1 \Big(\sin\frac{\pi y}{2}\Big)^p\bigg]^{1/p}.$$

By Theorem 2.1,

$$A_\Phi \leqslant \sup_{x\in(0,1)} f(x)^{-1}\|f\varphi(x\wedge\cdot)\|_\Phi.$$

Numerical computation shows that the supremum is attained at $x = 1$. Therefore,

$$A_\Phi \leqslant \|f\varphi\|_\Phi = q^{1/q}\bigg[\int_0^1 \Big(x\sin\frac{\pi x}{2}\Big)^p\bigg]^{1/p}.$$

On the other hand, since $D(f) = \pi^2/8$, by (1.6) and (3.3), we have

$$A_\Phi \geqslant \frac{q^{1/q}\|f\|_p}{D(f)} = \frac{8q^{1/q}}{\pi^2}\bigg[\int_0^1 \Big(\sin\frac{\pi x}{2}\Big)^p\bigg]^{1/p}.$$

Combining these facts together, we obtain

$$\frac{8q^{1/q}}{\pi^2}\bigg[\int_0^1 \Big(\sin\frac{\pi x}{2}\Big)^p\bigg]^{1/p} \leqslant A_\Phi \leqslant q^{1/q}\bigg[\int_0^1 \Big(x\sin\frac{\pi x}{2}\Big)^p\bigg]^{1/p}. \tag{3.15}$$

The estimates in (3.15) are quite good for smaller $p > 1$ and are indeed exact when $p \to 1$.

Finally, we compute the upper bound D_Φ. Since

$$\big\|\sqrt{\varphi}\,\varphi(x \wedge \cdot)\big\|_p^p = \frac{2x^p}{p+2}\left[1 - \frac{2px^{p/2+1}}{3p+2}\right],$$

by (2.3) and (3.13), we have

$$\begin{aligned} A_\Phi &\leqslant q^{1/q}\left(\frac{2}{p+2}\right)^{1/p} \sup_{x\in(0,1)} \sqrt{x}\left[1 - \frac{2px^{p/2+1}}{3p+2}\right]^{1/p} \\ &= \frac{q^{1/q}}{(p+1)^{1/p}}\left(\frac{3p+2}{4(p+1)}\right)^{1/(p+2)}. \end{aligned} \tag{3.16}$$

This bound is much better than (3.11), which can still be improved by using $D_\Phi(2) = D_{\mathbb{B}}(2)$ given in Theorem 2.1:

$$D_\Phi(2) = q^{1/q}[(7p+4)/(10p^2+13p+4)]^{1/p}.$$

Note that the ratio of the upper bound in (3.16) (resp. $D_\Phi(2)$) and the lower bound in (3.14) is bounded above by $2\cdot 5^{1/3}/3 \approx 1.13998$ (resp. $88/80 \approx 1.08642$) and decreases to 1 as $p \to \infty$. Moreover, when p varies from 1 to ∞, $D_\Phi/D_\Phi(2)$ starts at $27\cdot 5^{1/3}/44 \approx 1.0493$ and decreases to 1 rapidly. In the worst case of $p = 1$, applying (1.6) to the same test function

$$x\left[1 - 2px^{p/2+1}/(3p+2)\right]^{1/p}$$

with $p = 1$, we obtain the lower bound $162/405 > 3/8$. Therefore, we have

$$162/405 \approx 0.4049 < A_\Phi = 4/\pi^2 \approx 0.4053 < D_\Phi(2) = 11/27 \approx 0.4075$$

and so there is not much room for further improvement. Actually, in this case, the iterative procedure works well as shown in [1; Example 1.5]. We have thus illustrated the power of the variational formulas.

4. Logarithmic Sobolev inequality. As a typical application of the above general setup, this section studies the Orlicz spaces with N-functions $\Phi(x) = |x|\log(1+|x|)$ and $\Psi(x) = x^2\log(1+x^2)$, and apply to the logarithmic Sobolev inequality on $\mathbb{R}$ (or subintervals of $\mathbb{R}$ with obvious modification). The starting point is the following observation which is a slight improvement of [5; Proposition 4.1].

Lemma 4.1. For every $f \in L^\Phi(\mathbb{R}, \mu)$, we have

$$\frac{4}{5}\big\|f - \pi(f)\big\|_{(\Psi)}^2 \leqslant \mathscr{L}(f) \leqslant \frac{51}{20}\big\|f - \pi(f)\big\|_{(\Psi)}^2,$$

where $\mathscr{L}(f) = \sup_{c\in\mathbb{R}} \mathrm{Ent}\big((f+c)^2\big)$ and

$$\mathrm{Ent}(f) = \int_{\mathbb{R}} f \log\big(f/\|f\|_{L^1(\pi)}\big)\mathrm{d}\mu$$

for $f \geqslant 0$.

Proof. Note that if we replace μ with π in definitions of $\mathscr{L}(f)$, $\mathrm{Ent}(f)$ and $\|\cdot\|_{(\Psi)}$ used here, then the conclusion is the same and we return to the context of [5] (the μ used in [5] is the π used here).

Let $\|f\|_{(\Psi)} = 1$ and $\pi(f) = 0$. By a result essentially due to [20; Lemma 9], we have $\mathscr{L}(f) \leqslant \mathrm{Ent}\big(f^2\big) + 2\pi\big(f^2\big)$. Express the right as

$$\int f^2\big(\delta + \log f^2\big)\mathrm{d}\pi + \pi\big(f^2\big)\big[2 - \delta - \log \pi\big(f^2\big)\big]$$

for some $\delta \in [0,2]$. Note that $x(2-\delta-\log x) \leqslant e^{1-\delta}$ for all $x > 0$. Let $c(\delta)$ be the bound so that $\delta + \log x \leqslant c(\delta)\log(1+x)$ for all $x > 0$. Then, we have

$$\mathscr{L}(f) \leqslant c(\delta)\int f^2\log\big(1+f^2\big)\mathrm{d}\pi + e^{1-\delta} \leqslant c(\delta) + e^{1-\delta}.$$

Minimizing the right in δ and noting that $c(\delta)$ satisfies the equation

$$c\log c - (c-1)\log(c-1) = \delta \; (c > 1)$$

(which comes from the equation $c'(\delta) = 0$), we obtain $\delta \approx 1.02118$, $c(\delta) \approx 1.56271$ and then obtain the required upper bound.

For the lower bound, the idea is to find the smallest constant $\delta \approx 0.4408$ so that $x\log\big(1 + x/(2+\delta)\big) \leqslant \delta + x\log x$ for all $x > 0$. Then

$$\int (f^2/(2+\delta))\log\big(1 + f^2/(2+\delta)\big)\mathrm{d}\pi \leqslant \big(\delta + \int f^2 \log f^2 \mathrm{d}\pi\big)/(2+\delta) \leqslant 1,$$

since

$$\int f^2\log f^2\mathrm{d}\pi \leqslant 2 = \mathscr{L}(f)$$

by assumption, and the remainder of the proof is the same as the original one given in [5]. $\square$

Lemma 4.1 leads to the use of Ψ. Next, since $\|f\|_{(\Psi)}^2 = \big\|f^2\big\|_{(\Phi)}$ (the similar relation $\|f\|_{\Psi}^2 = \big\|f^2\big\|_{\Phi}$ seems not to be true), it is also natural to use Φ. Note that the use of the norm $\|\cdot\|_{\Phi}$ is necessary because of the representation (3.1). This point was missed in the previous papers and so it is worthy to re-examine the estimates of the optimal constants. Actually, we will produce a new and much precise result (Theorem 4.3).

According to Lemma 4.1, it suffices to estimate the constant $\overline{A}_{\Phi}$ in (2.6). To do so, we first study A_{Φ} on the interval $(0, D)$. Again, we concentrate on explicit bounds without examining the variational formulas (1.7) and (2.2).

Theorem 4.2. Consider the interval $(0, D)$. For $\Phi(x) = |x|\log(1+|x|)$, we have

$$B_\Phi \leqslant C_\Phi \leqslant A_\Phi \leqslant D_\Phi \leqslant 4B_\Phi, \tag{4.1}$$

where

$$\begin{aligned}
B_\Phi &= \sup_{x\in(0,D)} \varphi(x) M(\mu(x,D)),\\
M(x) &:= x\bigg[\frac{2}{1+\sqrt{1+4x}} + \log\bigg(1+\frac{1+\sqrt{1+4x}}{2x}\bigg)\bigg],\\
C_\Phi &= \sup_{x\in(0,D)} \varphi(x)^{-1}\big\|\varphi(x\wedge\cdot)^2\big\|_\Phi,\\
D_\Phi &= \sup_{x\in(0,D)} \varphi(x)^{-1/2}\big\|\sqrt{\varphi}\,\varphi(x\wedge\cdot)\big\|_\Phi.
\end{aligned} \tag{4.2}$$

In particular, the Poincaré-type inequality (1.5) in the Orlicz space $L^\Phi(\mu)$ holds iff

$$\sup_{x\in(0,D)} \varphi(x)\mu(x,D)\log\big(1/\mu(x,D)\big) < \infty. \tag{4.3}$$

Proof. First, we compute $\|I_B\|_\Phi$. The equation (3.5) with $f = I_B$ becomes

$$\int_B \alpha^2/(1+\alpha)\mathrm{d}\mu = 1,$$

from which we obtain the solution $\alpha^* = \big(1+\sqrt{1+4\mu(B)}\,\big)/(2\mu(B))$. Inserting this into (3.6), we obtain

$$\|I_B\|_\Phi = M(\mu(B)) = \frac{1}{2}\big(\sqrt{1+4\mu(B)}-1\big) + \mu(B)\log\bigg(1+\frac{1+\sqrt{1+4\mu(B)}}{2\mu(B)}\bigg).$$

Combining this with Theorem 3.1, we obtain (4.1). Obviously, we have the following simpler estimates:

$$\mu(B)\log\big(1+1/\mu(B)\big) \leqslant \|I_B\|_\Phi \leqslant \mu(B)\big[1+\log\big(1+\omega/\mu(B)\big)\big],$$

where $\omega = \big(\sqrt{1+4\mu(0,D)}+1\big)/2$. Then, the last assertion follows from (4.1). $\square$

The computations of C_Φ and D_Φ are usually non-trivial. For this, we introduce some approximation procedures of $\|f\|_\Phi$ for general $f \in L^\Phi(\mu)$. Replacing f by $|f|$ if necessary, assume that $f \geqslant 0$. Again, let E be an open subinterval of $\mathbb{R}$. Then, by (3.3), we have

$$\|f\|_\Phi = \inf_{\alpha>0}\bigg[\frac{1}{\alpha} + \int_E f\log(1+\alpha f)\mathrm{d}\mu\bigg] =: \inf_{\alpha>0} H(\alpha).$$

Then $H'(\alpha) \geqslant 0$ iff

$$\alpha^{-2} \leqslant \int_E f^2/(1+\alpha f)\mathrm{d}\mu.$$

That is,

$$\alpha \geqslant \big[1+\mu(E)-\mu\big((1+\alpha f)^{-1}\big)\big]/\mu(f) =: J(\alpha)/\mu(f).$$

Clearly, $H(\alpha)$ attains its infimum (i.e., $\alpha\mu(f)=J(\alpha)$) at some α^*:

$$0<\underline{\alpha}_1 := \mu(f)^{-1} \leqslant \alpha^* \leqslant \big(1+\mu(E)\big)/\mu(f) =: \bar{\alpha}_1.$$

Define $\underline{\alpha}_n = J(\underline{\alpha}_{n-1})/\mu(f)$ and $\bar{\alpha}_n = J(\bar{\alpha}_{n-1})/\mu(f)$. Then, we have $\underline{\alpha}_n \uparrow \alpha^* \downarrow \bar{\alpha}_n$. Therefore,

$$\underline{\alpha}_n \leqslant J\big(\underline{\alpha}_n\big)/\mu(f) \leqslant J\big(\bar{\alpha}_n\big)/\mu(f) \leqslant \bar{\alpha}_n$$

since $J(\alpha)$ is increasing in α. Then, $H'\big(\underline{\alpha}_n\big) \leqslant 0 \leqslant H'\big(\bar{\alpha}_n\big)$ and so $\|f\|_\Phi \leqslant H\big(\underline{\alpha}_n\big)\wedge H\big(\bar{\alpha}_n\big)$ for all $n\geqslant 1$ and $\|f\|_\Phi = \lim_{n\to\infty} H\big(\underline{\bar{\alpha}}_n\big)$. This leads to our first approximation procedure:

$$\|f\|_\Phi \lessapprox H\big((\underline{\alpha}_n+\bar{\alpha}_n)/2\big), \qquad \|f\|_\Phi = \lim_{n\to\infty} H\big((\underline{\alpha}_n+\bar{\alpha}_n)/2\big). \tag{4.4}$$

Here, "$\lessapprox$" means that "$\leqslant$" and "$\approx$". In practice, the second approximation procedure below, called Bolzano's method, is even more effective: Set $\alpha_1=\mu(f)^{-1}$ and $\alpha_2=(1+\mu(E))/\mu(f)$. Noting that $\alpha\mu(f)-J(\alpha)$, as well as $H'(\alpha)$, has different sign at α_1 and α_2, we make a test at the middle: $\alpha_3 := (\alpha_1+\alpha_2)/2$. Next, if $\alpha\mu(f)-J(\alpha)$ have different sign at α_1 and α_3 for instance, eliminating the subinterval (α_3,α_2) and choose the middle of (α_1,α_3): $\alpha_4 := (\alpha_1+\alpha_3)/2$ as the new test point, and so on. At the n-th step, we have a smaller subinterval left with endpoints α_n and α_{n+1} at which $\alpha\mu(f)-J(\alpha)$ has different sign, then we may stop here by choosing $(\alpha_n+\alpha_{n+1})/2$ as an approximation of α^*, and furthermore:

$$\|f\|_\Phi \lessapprox H\big((\alpha_n+\alpha_{n+1})/2\big), \qquad \|f\|_\Phi = \lim_{n\to\infty} H\big((\alpha_n+\alpha_{n+1})/2\big). \tag{4.5}$$

When $n=1$, the two approximations in (4.4) and (4.5) coincide with each other. Having (4.4) and (4.5) at hand, it is not difficult to estimate C_Φ and D_Φ. Actually, even for $n\leqslant 2$, both (4.4) and (4.5) often produce good enough estimates. In these cases, it is not difficult to write down the analytic estimates for C_Φ and D_Φ. One may wonder about the accuracy of the upper bound of C_Φ. For this, we mention an analytic but rough lower bound of C_Φ. By (3.10), we have

$$\big\|\varphi(x\wedge\cdot)^2\big\|_\Phi \geqslant \inf_{\alpha>0}\Big[\alpha^{-1}+\mu\big(\varphi(x\wedge\cdot)^2\big)\log\big[1+\alpha\pi\big(\varphi(x\wedge\cdot)^2\big)\big]\Big].$$

Finding the infimum on the right, we obtain the following estimate.

$$\begin{aligned} C_\Phi &\geqslant \sup_{x\in(0,D)}\varphi(x)^{-1}\bigg[\frac{1}{\alpha^*}+\mu\big(\varphi(x\wedge\cdot)^2\big)\log\big[1+\alpha^*\pi\big(\varphi(x\wedge\cdot)^2\big)\big]\bigg],\\ \alpha^* &= \frac{1+\sqrt{1+\mu(0,D)}}{2\mu\big(\varphi(x\wedge\cdot)^2\big)}. \end{aligned} \tag{4.6}$$

The next step is splitting $\mathbb{R}$ into the half lines $\mathbb{R}_1 := (-\infty, 0)$ and $\mathbb{R}_2 := (0, \infty)$. For this, we need some notations. Denote by $\|\cdot\|_{1,\Phi}$ and $\|\cdot\|_{2,\Phi}$ the norms in $L^\Phi(\mathbb{R}_1, \mu)$ and $L^\Phi(\mathbb{R}_2, \mu)$, respectively. Actually, $\|f\|_{1,\Phi} = \|fI_{\mathbb{R}_1}\|_\Phi$ and $\|f\|_{2,\Phi} = \|fI_{\mathbb{R}_2}\|_\Phi$. The corresponding constants in inequality (1.5) are denoted by A_Φ^1 and A_Φ^2, respectively. Similarly, we have $A_{(\Phi)}^1$, $A_{(\Phi)}^2$ when the norm $\|\cdot\|_\Phi$ is replaced by $\|\cdot\|_{(\Phi)}$, and so on. Recall that $\|\cdot\|_\Phi$ and $\overline{A}_\Phi$ are used for the whole line, i.e., for the space $L^\Phi(\mathbb{R}, \mu)$.

Let B_Φ^k and $C_\Phi^k (k = 1, 2)$ be given by Theorem 4.2 in terms of $\varphi_1(x) = \int_x^0 e^{-C}$ and $\varphi_2(x) = \int_0^x e^{-C}$:

$$\begin{aligned}
B_\Phi^1 &= \sup_{x\in(-\infty,0)} \varphi_1(x) M(\mu(-\infty, x)),\\
B_\Phi^2 &= \sup_{x\in(0,\infty)} \varphi_2(x) M(\mu(x, \infty)),\\
C_\Phi^1 &= \sup_{x\in(-\infty,0)} \varphi_1(x)^{-1} \big\|\varphi_1(x\vee\cdot)^2\big\|_{1,\Phi},\\
C_\Phi^2 &= \sup_{x\in(0,\infty)} \varphi_2(x)^{-1} \big\|\varphi_2(x\wedge\cdot)^2\big\|_{2,\Phi},\\
D_\Phi^1 &= \sup_{x\in(-\infty,0)} \varphi_1(x)^{-1/2} \big\|\sqrt{\varphi_1}\,\varphi_1(x\vee\cdot)\big\|_{1,\Phi},\\
D_\Phi^2 &= \sup_{x\in(0,\infty)} \varphi_2(x)^{-1/2} \big\|\sqrt{\varphi_2}\,\varphi_2(x\wedge\cdot)\big\|_{2,\Phi}.
\end{aligned} \tag{4.7}$$

Then, by Lemma 4.1, Theorems 4.2 and 3.2, (3.9), and choosing $c = 0$ in (2.7) and (2.8), we obtain the following result, which solves the main problem of this section.

Theorem 4.3. Let $\Phi(x) = |x| \log(1 + |x|)$.

(1) The inequality

$$\big\|f^2\big\|_{(\Phi)} \leqslant A_{(\Phi)} D(f), \qquad f \in \mathbb{C}_d[0, D], \qquad f(0) = 0 \tag{4.8}$$

holds iff $B_\Phi < \infty$. The optimal constant $A_{(\Phi)}$ satisfies[1]

$$B_\Phi \leqslant C_\Phi \leqslant A_\Phi \leqslant 2A_{(\Phi)} \leqslant 2A_\Phi \leqslant 2D_\Phi \leqslant 8B_\Phi. \tag{4.9}$$

(2) The inequality

$$\big\|(f - \pi(f))^2\big\|_{(\Phi)} \leqslant \overline{A}_{(\Phi)} D(f), \qquad f \in \mathbb{C}_d(\mathbb{R}) \tag{4.10}$$

holds iff $B_\Phi^1 \vee B_\Phi^2 < \infty$. The optimal constant $\overline{A}_{(\Phi)}$ satisfies

$$\frac{1}{2}\big(B_\Phi^1 \wedge B_\Phi^2\big) \leqslant \frac{1}{2}\big(C_\Phi^1 \wedge C_\Phi^2\big) \leqslant \overline{A}_\Phi \leqslant 2\overline{A}_{(\Phi)} \leqslant 2\overline{A}_\Phi \leqslant 2\big(D_\Phi^1 \vee D_\Phi^2\big) \leqslant 8\big(B_\Phi^1 \vee B_\Phi^2\big). \tag{4.11}$$

[1] Some corrections on the coefficients are made in (4.9), (4.11)–(4.13) and Theorem 4.5

Furthermore,

$$\frac{1}{4}B_{\mathbb{B}^1 c_0} \leqslant \frac{1}{4}C_{\mathbb{B}^1 c_0} \leqslant \overline{A}_{(\Phi)} \leqslant D_{\mathbb{B}^1 c_0} \leqslant 4B_{\mathbb{B}^1 c_0}, \tag{4.12}$$

where $B_{\mathbb{B}^1 c_0}$ is given by Theorem 2.2 with $(p,q)=\mathbb{R}$. In particular, the optimal constant A'' in the logarithmic Sobolev inequality satisfies

$$\frac{1}{5}B_{\mathbb{B}^1 c_0} \leqslant \frac{1}{5}C_{\mathbb{B}^1 c_0} \leqslant A'' \leqslant \frac{51}{20}D_{\mathbb{B}^1 c_0} \leqslant \frac{51}{5}B_{\mathbb{B}^1 c_0}. \tag{4.13}$$

Example 4.4. Everything is the same as in Example 3.4 except that the interval $[0,1]$ is replaced by $[-1/2,1/2]$. We now study the inequality (2.6) with $\mathbb{B} = L^\Phi(\mu)$, $\Phi(x)=|x|\log(1+|x|)$. By Theorems 3.2 and 2.2, the constant $\overline{A}_\Phi$ can be estimated by A_Φ^k, B_Φ^k, C_Φ^k and D_Φ^k $(k=1,2)$ given in Theorem 4.2 corresponding to the subintervals $[-1/2,0]$ and $[0,1/2]$, respectively. By symmetry, it is clear that $A_\Phi^1 = A_\Phi^2$, $B_\Phi^1 = B_\Phi^2$ and so on. Hence, we need to consider the constants B_Φ, C_Φ and D_Φ defined by Theorem 4.2 on $[0,1/2]$. By using the first formula in (4.2), some numerical computations give us $B_\Phi \approx 0.1668$. On the other hand, in using (4.4) with $n=1,\ 2$, we obtain $D_\Phi \lessapprox 0.2402,\ 0.2401$, respectively. Next, in using (4.4) and (4.5) with $n=1,\ 2$ again, we have almost the same bound $C_\Phi \lessapprox 0.2216$. Finally, (4.6) gives us 0.1921. Therefore, the estimates in (4.4) and (4.5), and furthermore those in (4.1), (1.9), (2.3), (2.7) and (2.8) are all quite satisfactory in the present situation.

We are now going to prove a different lower bound which is quite rough, not really needed for our purpose, but has the same form as the upper bound in (4.11). For this, we need Proposition 2.3. We follow the last part of the proof of Theorem 3.2. Applying (4.10) to an arbitrary function f_1 with $f_1|_{\mathbb{R}_2}=0$, we get

$$\overline{A}_{(\Phi)}D(f_1) \geqslant \big\|\big(f_1-\pi(f_1)\big)^2\big\|_{(\Phi)} = \big\|f_1-\pi(f_1)\big\|_{(\Psi)}^2.$$

Next, applying Part (2) of Proposition 2.3 to the space $\mathbb{B}=L^\Psi(\mathbb{R},\mu)$ with norm $\|\cdot\|_{(\Psi)}$ and the set $A=\mathbb{R}_1$, we get $\|f_1\|_{(\Psi)} \leqslant K_1\|(f_1-\pi(f_1)\|_{(\Psi)}$, where $K_1 = \big[1-c_2(\mathbb{R}_1)\sqrt{\pi(\mathbb{R}_1)}\,\|1\|_{(\Psi)}\big]^{-1}$. Hence, by (3.9),

$$K_1^2\overline{A}_{(\Phi)}D(f_1) \geqslant \|f_1\|_{(\Psi)}^2 = \big\|f_1^2\big\|_{1,\Phi}.$$

Combining this with definition of $A_{(\Phi)}^1$, it follows that $\overline{A}_{(\Phi)} \geqslant A_{(\Phi)}^1/K_1^2$. Similarly, applying (4.10) to the function f_2 with $f_2|_{\mathbb{R}_1}=0$, we obtain $\overline{A}_{(\Phi)} \geqslant A_{(\Phi)}^2/K_2^2$, where $K_2 = \big[1-c_2(\mathbb{R}_2)\sqrt{\pi(\mathbb{R}_2)}\,\|1\|_{(\Psi)}\big]^{-1}$. Collecting these facts together, it follows that

$$\overline{A}_{(\Phi)} \geqslant \max\big\{A_{(\Phi)}^1/K_1^2,\ A_{(\Phi)}^2/K_2^2\big\} \geqslant \big(K_1^{-2}\wedge K_2^{-2}\big)\big(A_{(\Phi)}^1 \vee A_{(\Phi)}^2\big). \tag{4.14}$$

Finally, we compute K_1 and K_2. For simplicity, let $Z_1 = \mu(\mathbb{R}_1)$ and $Z_2 = \mu(\mathbb{R}_2)$. Because of $\Psi(f) = \Phi(f^2)$, by the convexity of Φ and Jensen's inequality, we have

$$\begin{aligned}\|f\|_{2,(\Psi)} &= \inf\left\{\alpha > 0 : \int_0^\infty \Phi(f^2/\alpha^2)\mathrm{d}\mu \leqslant 1\right\}\\ &\geqslant \inf\left\{\alpha > 0 : \frac{1}{\alpha^2 Z_2}\int_0^\infty f^2\mathrm{d}\mu \leqslant \Phi^{-1}\big(Z_2^{-1}\big)\right\}\\ &\geqslant \left[\frac{Z\pi(f^2)}{Z_2\Phi^{-1}\big(Z_2^{-1}\big)}\right]^{1/2}.\end{aligned}$$

Therefore, $c_2(\mathbb{R}_2) = \big[Z_2\Phi^{-1}\big(Z_2^{-1}\big)/Z\big]^{1/2}$. On the other hand, by (3.8), we have

$$\|1\|_{(\Psi)} = 1/\Psi^{-1}\big(Z^{-1}\big) = 1/\sqrt{\Phi^{-1}\big(Z^{-1}\big)}.$$

Thus,

$$\begin{aligned}c_2(\mathbb{R}_2)\sqrt{\pi(\mathbb{R}_2)}\,\|1\|_{(\Psi)} &= \left[\frac{Z_2\Phi^{-1}\big(Z_2^{-1}\big)}{Z}\right]^{1/2}\cdot\left[\frac{Z_2}{Z}\right]^{1/2}\cdot\left[\frac{1}{\Phi^{-1}\big(Z^{-1}\big)}\right]^{1/2}\\ &= \frac{Z_2}{Z}\left[\frac{\Phi^{-1}\big(Z_2^{-1}\big)}{\Phi^{-1}\big(Z^{-1}\big)}\right]^{1/2}.\end{aligned}$$

By symmetry, we get

$$c_2(\mathbb{R}_1)\sqrt{\pi(\mathbb{R}_1)}\,\|1\|_{\Psi} = \frac{Z_1}{Z}\left[\frac{\Phi^{-1}\big(Z_1^{-1}\big)}{\Phi^{-1}\big(Z^{-1}\big)}\right]^{1/2} = \frac{Z_1\Psi^{-1}\big(Z_1^{-1}\big)}{Z\Psi^{-1}\big(Z^{-1}\big)} < 1$$

since $\Psi^{-1}(x)/x$ is decreasing in x. Inserting these into (4.14), we arrive at

$$K_1^{-2}\wedge K_2^{-2} = \left[\left(1 - \frac{Z_1\Psi^{-1}\big(Z_1^{-1}\big)}{Z\Psi^{-1}\big(Z^{-1}\big)}\right)\bigwedge\left(1 - \frac{Z_2\Psi^{-1}\big(Z_2^{-1}\big)}{Z\Psi^{-1}\big(Z^{-1}\big)}\right)\right]^2.$$

The right-hand side achieves its maximum at $Z_1 = Z_2 = Z/2$, i.e., 0 is the median of μ. Then we have

$$K_1^{-2}\wedge K_2^{-2} = \left[1 - \frac{\Psi^{-1}\big(2Z^{-1}\big)}{2\Psi^{-1}\big(Z^{-1}\big)}\right]^2.$$

To estimate this constant, set $z = \Psi^{-1}(Z^{-1})$. Then

$$\Psi^{-1}\big(2Z^{-1}\big)\big/\big[2\Psi^{-1}\big(Z^{-1}\big)\big] \leqslant \delta$$

iff

$$\Psi^{-1}\big(2\Psi(z)\big)\big/(2z) \leqslant \delta$$

for some $\delta > 0$. That is, $2\Psi(z) \leqslant \Psi(2\delta z)$. Equivalently,

$$2\delta^2\log\big(1 + 4\delta^2 z^2\big) \geqslant \log(1 + z^2).$$

From this, one sees immediately that $\delta \leqslant 1/\sqrt{2}$. Hence the coefficient is bounded below by $\big(\sqrt{2}-1\big)^2/2 \approx 0.085$. Returning to (4.14) and using part (2) of Theorem 4.3, we obtain, at last, the following result.

Theorem 4.5. By a translation if necessary, assume that 0 is the median of μ. Let B_Φ^1 and B_Φ^2 be given by (4.7). Then the optimal constant $\overline{A}_{(\Phi)}$ in (4.10) satisfies

$$\frac{\big(\sqrt{2}-1\big)^2}{4}\big(B_\Phi^1 \vee B_\Phi^2\big) \leqslant \frac{\big(\sqrt{2}-1\big)^2}{4}\big(C_\Phi^1 \vee C_\Phi^2\big) \leqslant \overline{A}_{(\Phi)} \leqslant \big(D_\Phi^1 \vee D_\Phi^2\big) \leqslant 4\big(B_\Phi^1 \vee B_\Phi^2\big),$$

and the logarithmic Sobolev constant A'' satisfies

$$\frac{\big(\sqrt{2}-1\big)^2}{5}\big(B_\Phi^1 \vee B_\Phi^2\big) \leqslant \frac{\big(\sqrt{2}-1\big)^2}{5}\big(C_\Phi^1 \vee C_\Phi^2\big) \leqslant A'' \leqslant \frac{51}{20}\big(D_\Phi^1 \vee D_\Phi^2\big) \leqslant \frac{51}{5}\big(B_\Phi^1 \vee B_\Phi^2\big).$$

Noticing that $\big(\sqrt{2}-1\big)^2 \approx 0.17$, the largest ratio of the coefficients for A'' is approximately 300. Thus, Theorem 4.5 improves considerably the result [5; Theorem 5.3], where the coefficients of the lower and upper bounds for A'' are $1/150$ and 468, respectively, with a quantity different from $B_\Phi^1 \vee B_\Phi^2$.

We remark that it is not necessary to use the norm $\|\cdot\|_{(\Phi)}$. One may estimate $\mathscr{L}(f)$ in terms of $\|\cdot\|_\Phi$, rather than $\|\cdot\|_{(\Phi)}$ appeared in Lemma 4.1. Then, the remains of the proofs are parallel.

Finally, we mention that all the results in the paper are meaningful in the discrete case for birth-death processes. Actually, all the facts we need here for Poincaré inequalities are presented in [16], [1] and [6]. The details will be published in a subsequent paper.

Acknowledgement. Thanks are given to the referees for their corrections of the earlier version of the paper.

References

[1]. Chen M F., *Variational formulas and approximation theorems for the first eigenvalue in dimension one*, Sci Chin (A), 2001, 44(4), 409–418.

[2]. Chen M F., *Analytic proof of dual variational formula for the first eigenvalue in dimension one*, Sci. Chin. (A), 1999, 42(8): 805–815.

[3]. Chen M F. Wang F Y., *Estimation of spectral gap for elliptic operators*, Trans Amer Math Soc, 1997, 349(3), 1239-1267.

[4]. Kantorovich L V. Akilov G P., *Functional Analysis* (2ed), Pergamon Press, 1982, Oxford, New York.

[5]. Bobkov S G. Götze F., *Exponential integrability and transportation cost related to logarithmic Sobolev inequalities*, J Funct Anal, 1999, 163, 1–28.

[6]. Mao Y H., *The logarithmic Sobolev inequalities for birth-death process and diffusion process on the line*, Chin J Appl Prob Statis, 2002, 18(1), 94–100.

[7]. Mao Y H., *Nash inequalities for Markov processes in dimension one*, Acta Math Sin Eng Ser, 2002, 18(1), 147–156.

[8]. Mao Y H., *On empty essential spectrum for Markov processes in dimension one*, preprint, 2000.

[9]. Mao Y H., *On super-contractivity for Markov semigroups*, preprint, 2001.

[10]. Gross L., *Logarithmic Sobolev inequalities and contractivity of semigroups*, LNM. 1993, **1563**.

[11]. Wang F Y., *Functional inequalities for empty essential spectrum*, J Funct Anal, 2000, 170, 219–245.

[12]. Gong F Z. Wang F Y., *Functional inequalities for uniformly integrable semigroups and application to essential spectrums*, Forum Math, 2002, 14, 293–313.

[13]. Muckenhoupt B., *Hardy's inequality with weights*, Studia Math, 1972(XLIV), 31–38.
[14]. Mazya V G., *Sobolev Spaces*, Springer-Valerg, 1985.
[15]. Opic B. Kufner A., *Hardy-type Inequalities*, Longman, New York, 1990.
[16]. Chen M F., *Explicit bounds of the first eigenvalue*, Sci in China, Ser A, 2000, 43(10), 1051–1059.
[17]. Chen M F., *Logarithmic Sobolev inequality for symmetric forms*, Sci Chin, 2000, 43(6), 601–608.
[18]. Wang F Y., *Sobolev type inequalities for general symmetric forms*, Proc Amer Math Soc, 2001, 128(12), 3675–3682.
[19]. Rao M M. Ren Z D., *Theory of Orlicz Spaces*, Marcel Dekker, Inc New York, 1991.
[20]. Rothaus O S., *Analytic inequalities, isoperimetric inequalities and logarithmic Sobolev inequalities*, J Funct Anal, 1985, 64, 296–313.

Acta Mathematica Sinica, English Series 19(4), 2003, 625–644

VARIATIONAL FORMULAS OF POINCARÉ-TYPE INEQUALITIES FOR BIRTH-DEATH PROCESSES

Mu-Fa Chen

Department of Mathematics, Beijing Normal University, Beijing 100875, P. R. China
E-mail: mfchen@bnu.edu.cn
Home page: http://www.bnu.edu.cn/˜chenmf/main_eng.htm)
Received August 8, 2002; accepted February 17, 2002.

Abstract. In author's one previous paper, the same topic was studied for one dimensional diffusions. As a continuation, this paper studies the discrete case, that is the birth-death processes. The explicit criteria for the inequalities, the variational formulas and explicit bounds of the corresponding constants in the inequalities are presented. As typical applications, the Nash inequalities and logarithmic Sobolev inequalities are examined.

This paper, being a continuation of [1], deals with the discrete case.

1. Introduction. Consider a birth-death process with birth rates $b_i > 0\,(0 \leqslant i \leqslant N-1)$ and death rates $a_i > 0\,(1 \leqslant i \leqslant N \leqslant \infty)$. When $N = \infty$, one should obviously use "$i \geqslant 1$" instead of "$1 \leqslant i \leqslant N$" in the last bracket and elsewhere. However, we will not repeat this in what follows. Let

$$\mu_0 = 1, \quad \mu_n = \frac{b_0 b_1 \cdots b_{n-1}}{a_1 a_2 \cdots a_n}, \quad 0 \leqslant n \leqslant N.$$

Throughout this paper, when $N = \infty$, assume that $Z := \sum_{n=0}^{N} \mu_n < \infty$. Let $\pi_n = \mu_n / Z$, $0 \leqslant n \leqslant N$.

Throughout the paper, let $(\mathbb{B}, \|\cdot\|_{\mathbb{B}}, \mu)$ be a Banach space of functions $f : E_1 := \{1, 2, \cdots, N\} \to \mathbb{R}$ satisfying the following conditions:

2000 *Mathematics Subject Classification.* 60J27, 49R50, 34L15.

Key words and phrases. Variational formula, Poincaré inequality, Nash inequality, Logarithmic Sobolev inequality, Orlicz space, Birth-death process.

Research supported in part NSFC (No. 10121101), 973 Project and RFDP.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

$$\begin{aligned}
&(1)\quad 1\in\mathbb{B};\\
&(2)\quad \mathbb{B}\text{ is ideal: If } h\in\mathbb{B} \text{ and } |f|\leqslant |h|,\ \text{ then } f\in\mathbb{B};\\
&(3)\quad \|f\|_{\mathbb{B}}=\sup_{g\in\mathscr{G}}\sum_{i\in E_1}|f_i|g_i\mu_i,\\
&(4)\quad \mathscr{G}\ni g^{(0)} \text{ with } \inf g^{(0)}>0,
\end{aligned}\tag{1.1}$$

where $\mathscr{G}$ is a fixed set, to be specified case by case later, of non-negative functions on E_1. The first two conditions mean that $\mathbb{B}$ is rich enough and the last one means that $\mathscr{G}$ is not trivial, it contains at least one strictly positive function. The third condition is essential in this paper, which means that the norm $\|\cdot\|_{\mathbb{B}}$ has a "dual" representation.

The aim of this paper is to study the following Poincaré-type inequality:

$$\big\|f^2\big\|_{\mathbb{B}}\leqslant A_{\mathbb{B}}D(f),\qquad f_0=0\tag{1.2}$$

where

$$D(f)=\sum_{i=1}^N\mu_i a_i(f_i-f_{i-1})^2.$$

Set

$$\mathscr{D}(D)=\{f\in L^2(E_1;\mu): f_0=0,\ D(f)<\infty\}.$$

Especially, we will study the estimation of the optimal constant $A_{\mathbb{B}}$ in (1.2):

$$A_{\mathbb{B}}=\sup_{f\in\mathscr{D}_0}\frac{\|f^2\|_{\mathbb{B}}}{D(f)},\qquad \mathscr{D}_0:=\{f\in\mathscr{D}(D): f_0=0,\ 0<D(f)<\infty\}.\tag{1.3}$$

To do so, define

$$\begin{aligned}
\mathscr{W}'&=\{w: w_0=0, w_i \text{ is strictly increasing}\},\\
\mathscr{W}''&=\{w: w_0=0, w_i>0 \text{ for all } i\in E_1\}.
\end{aligned}$$

Now, the main results about (1.2) can be stated as follows.

Theorem 1.1. Let (1.1) hold. Then, we have

$$A_{\mathbb{B}}\leqslant \inf_{w\in\mathscr{W}'}\sup_{1\leqslant i\leqslant N}\frac{1}{\mu_i a_i(w_i-w_{i-1})}\big\|wI_{[i,N]}\big\|_{\mathbb{B}},\tag{1.4}$$

where $[i,N]=\{i,i+1,\cdots,N\}$. In particular,

$$A_{\mathbb{B}}\leqslant 4\sup_{1\leqslant i\leqslant N}\varphi_i\|I_{[i,N]}\|_{\mathbb{B}}=:4B_{\mathbb{B}},\tag{1.5}$$

$$A_{\mathbb{B}}\geqslant \sup_{1\leqslant i\leqslant N}\varphi_i^{-1}\big\|\varphi(i\wedge\cdot)^2\big\|_{\mathbb{B}}=:C_{\mathbb{B}},\tag{1.6}$$

where $\varphi_i=\sum_{j=1}^i(\mu_j a_j)^{-1}$ and $i\wedge k=\min\{i,k\}$. Moreover,

$$C_{\mathbb{B}}\geqslant B_{\mathbb{B}}.\tag{1.7}$$

Hence $A_{\mathbb{B}}<\infty$ iff $B_{\mathbb{B}}<\infty$.

Theorem 1.2. Let (1.1) hold. Then, we have

$$A_{\mathbb{B}} \leqslant \inf_{w\in\mathscr{W}''} \sup_{1\leqslant i\leqslant N} w_i^{-1}\big\|w\varphi(i\wedge\cdot)\big\|_{\mathbb{B}}. \tag{1.8}$$

Next, let $B_{\mathbb{B}} < \infty$. Define

$$w^{(0)} = \sqrt{\varphi}, \quad w_i^{(n)} = \|w^{(n-1)}\varphi(i\wedge\cdot)\|_{\mathbb{B}}, \quad n\geqslant 1$$

and set

$$D_{\mathbb{B}}(n) = \sup_{1\leqslant i\leqslant N} w^{(n)}/w^{(n-1)}, \qquad n\geqslant 1.$$

Then, we have

$$4B_{\mathbb{B}} \geqslant D_{\mathbb{B}}(n)\downarrow \lim_{n\to\infty} D_{\mathbb{B}}(n) \geqslant A_{\mathbb{B}}. \tag{1.9}$$

The above theorems present a criterion for the Poincaré-type inequality (1.2) and two variational formulas (1.4) and (1.8) for upper bounds of $A_{\mathbb{B}}$. In general, the latter formula is stronger, but harder to compute than the former one. From these formulas, one deduces the explicit bounds of $A_{\mathbb{B}}$, given by (1.5)–(1.7), and an approximating procedure (1.9). In contrast to the continuous situation, here we do not have the estimate $C_{\mathbb{B}} \leqslant 2B_{\mathbb{B}}$.

The remainder of the paper is organized as follows. The proofs of Theorems 1.1 and 1.2 are presented in the next section. The Neumann case is treated in Sections 3 and 4, respectively, first for the state space $\{0,1,\cdots,N\}$ ($N\leqslant\infty$) and then for general space $\{M, M+1,\cdots,N-1,N\}$, $-\infty\leqslant M<N\leqslant\infty$. The results obtained in the first four sections are then specified to the Orlicz spaces in Section 5. As typical applications of the general setup, the Nash inequalities and the logarithmic Sobolev inequalities are studied in Sections 6 and 7, respectively, for the two state spaces just mentioned above.

Certainly, a large part of the paper is parallel to [1]. Nevertheless, there are still quite a number of differences from the continuous situation and so it is worthy to write down the details. Besides, the application to the Nash inequalities is newly added.

2. Proofs of Theorems 1.1 and 1.2. The proofs of Theorems 1.1 and 1.2 are quite similar to those of [1; Theorems 1.1 and 2.1]. The detailed proofs are presented here not only for completeness but also for an illustration of the necessary modifications from the continuous case. Besides, there are some simplifications.

Proof of Theorem 1.1. (a) The starting point of the our study is the following. Consider the ordinary Poincaré inequality

$$\|f\|_2^2 \leqslant AD(f), \qquad f_0 = 0, \tag{2.1}$$

where $\|\cdot\|_p$ denotes the $L^p(\mu)$-norm. To save the notation, we simply use A to denote the optimal constant in (2.1):

$$A = \sup_{f\in\mathscr{D}_0} \|f\|_2^2/D(f). \tag{2.2}$$

By [2; Theorem 3.4] or [3; Theorem 2.1] with a slight change of notations, we have a variational formula as follows.

$$A = \inf_{w\in\mathscr{W}'} \sup_{1\leqslant i\leqslant N} I_i(w), \tag{2.3}$$

where $I_i(w) = \big(\mu_i a_i(w_i - w_{i-1})\big)^{-1} \sum_{j=i}^N \mu_j w_j$.

(b) Note that $1 \in \mathbb{B}$ and so $\sup_{g\in\mathscr{G}} \sum_{i=1}^N g_i\mu_i = \|1\|_{\mathbb{B}} < \infty$. Hence $\mathscr{G} \subset L^1(\mu)$. Let $g^{(0)} \in \mathscr{G}$ satisfy $\inf g^{(0)} =: \varepsilon > 0$. Since $\varepsilon\mu(|f|) \leqslant \mu(|f|g^{(0)}) \leqslant \|f\|_{\mathbb{B}}$, we have $\mathbb{B} \subset L^1(\mu)$. To prove (1.4), one may assume that the right-hand side of (1.4) is finite. Otherwise, there is nothing to do. Then, for $g^{(0)}$ given above, we have

$$\begin{aligned}
\varepsilon \inf_{w\in\mathscr{W}'} \sup_{1\leqslant i\leqslant N} I_i(w) &\leqslant \inf_{w\in\mathscr{W}'} \sup_{1\leqslant i\leqslant N} \frac{1}{\mu_i a_i(w_i - w_{i-1})} \sum_{j=i}^N \mu_j w_j g_j^{(0)} \\
&\leqslant \inf_{w\in\mathscr{W}'} \sup_{1\leqslant i\leqslant N} \frac{1}{\mu_i a_i(w_i - w_{i-1})} \big\|wI_{[i,N]}\big\|_{\mathbb{B}} \\
&< \infty.
\end{aligned}$$

Hence $A = \inf_{w\in\mathscr{W}'} \sup_{1\leqslant i\leqslant N} I_i(w) < \infty$.

We now introduce a transform of (a_i, b_i, μ_i) which will be used several times in the paper. Let $g \in \mathscr{G}$ and set $g^{(n)} = g + 1/n$, $g^{(\infty)} = g$. At the moment, we do not require that $g_n \in \mathscr{G}$. Define

$$\begin{aligned}
&a_i^{(n)} = a_i/g_i^{(n)}, && b_i^{(n)} = b_i/g_i^{(n)}, \\
&\mu_0^{(n)} = g_0^{(n)} = \mu_0 g_0^{(n)}, && \mu_i^{(n)} = g_0^{(n)} \frac{b_0^{(n)}\cdots b_{i-1}^{(n)}}{a_1^{(n)}\cdots a_i^{(n)}} = \mu_i g_i^{(n)}, \quad 1 \leqslant i \leqslant N.
\end{aligned} \tag{2.4}$$

Then $\mu_i^{(n)} a_i^{(n)} = \mu_i a_i$ for all $i \in E_1$ and so the corresponding $D(f)$ and $\mathscr{D}_0$ are all invariant under this transform $\big(a_i, b_i, \mu_i\big) \to \big(a_i^{(n)}, b_i^{(n)}, \mu_i^{(n)}\big)$. However, the corresponding $I_i(w)$ is changed into

$$\begin{aligned}
I_i^{(n)}(w) &:= \frac{1}{\mu_i^{(n)} a_i^{(n)}(w_i - w_{i-1})} \sum_{j=i}^N \mu_j^{(n)} w_j \\
&= \frac{1}{\mu_i a_i(w_i - w_{i-1})} \sum_{j=i}^N \mu_j w_j g_j^{(n)}, \quad i \in E_1.
\end{aligned}$$

The formula

$$I_i^{(n)}(w) = \frac{1}{\mu_i a_i(w_i - w_{i-1})} \sum_{j=i}^N \mu_j w_j g_j^{(n)}, \quad i \in E_1, \quad 1 \leqslant n \leqslant \infty \tag{2.5}$$

is meaningful even if $n = \infty$. Note that

$$I_i^{(\infty)}(w) \leqslant I_i^{(n)}(w) \leqslant I_i^{(\infty)}(w) + \frac{1}{n} I_i(w).$$

Thus, once $\sup_{1\leqslant i\leqslant N} I_i(w) < \infty$, we must have

$$\lim_{n\to\infty} \sup_{1\leqslant i\leqslant N} \big|I_i^{(n)}(w) - I_i^{(\infty)}(w)\big| = 0.$$

On the other hand, define $A\big(g^{(n)}\big)$ and $A(g)$ as in (2.2) but replacing $\mathrm{d}\mu$ with $\mathrm{d}\mu_{g^{(n)}} := g^{(n)}\mathrm{d}\mu$ and $\mathrm{d}\mu_g := g\mathrm{d}\mu$, respectively. Clearly, $A(g) \leqslant A(g_n)\downarrow$. Next,

$$A(g_n) \leqslant A(g) + n^{-1} \sup_{f\in\mathscr{D}} \|f\|_2^2/D(f) = A(g) + A/n.$$

Therefore, $A(g_n)\downarrow A(g)$ as $n\to\infty$. We have thus proved that

$$A(g) = \lim_{n\to\infty} A\big(g^{(n)}\big) \leqslant \lim_{n\to\infty} \sup_{1\leqslant i\leqslant N} I_i^{(n)}(w) = \sup_{1\leqslant i\leqslant N} I_i^{(\infty)}(w)$$

for every $w\in\mathscr{W}'$. Hence

$$A(g) \leqslant \inf_{w\in\mathscr{W}'} \sup_{1\leqslant i\leqslant N} I_i^{(\infty)}(w). \tag{2.6}$$

On the other hand,

$$A_{\mathbb{B}} = \sup_{f\in\mathscr{D}_0} \frac{\|f^2\|_{\mathbb{B}}}{D(f)} = \sup_{f\in\mathscr{D}_0} \sup_{g\in\mathscr{G}} \frac{\sum_{i=1}^N \mu_i f_i^2 g_i}{D(f)} = \sup_{g\in\mathscr{G}} \sup_{f\in\mathscr{D}} \frac{\sum_{i=1}^N \mu_i f_i^2 g_i}{D(f)} = \sup_{g\in\mathscr{G}} A(g). \tag{2.7}$$

Combining this with (2.6), we obtain

$$\begin{aligned}
A_{\mathbb{B}} &\leqslant \sup_{g\in\mathscr{G}} \inf_{w\in\mathscr{W}'} \sup_{1\leqslant i\leqslant N} I_i^{(\infty)}(w)\\
&\leqslant \inf_{w\in\mathscr{W}'} \sup_{g\in\mathscr{G}} \sup_{1\leqslant i\leqslant N} I_i^{(\infty)}(w)\\
&= \inf_{w\in\mathscr{W}'} \sup_{1\leqslant i\leqslant N} \frac{1}{\mu_i a_i(w_i - w_{i-1})} \sup_{g\in\mathscr{G}} \sum_{j=i}^N \mu_j w_j g_j\\
&= \inf_{w\in\mathscr{W}'} \sup_{1\leqslant i\leqslant N} \frac{1}{\mu_i a_i(w_i - w_{i-1})} \big\|w I_{[i,N]}\big\|_{\mathbb{B}}.
\end{aligned}$$

This proves (1.4).

(c) We now prove the explicit bounds given in (1.5)–(1.7). To prove (1.5), applying (1.4) to the test sequence $w = \sqrt{\varphi}$ ($\varphi_0 = 0$ by convention), we get

$$A_{\mathbb{B}} \leqslant \sup_{1\leqslant i\leqslant N} \frac{1}{\mu_i a_i\big(\sqrt{\varphi_i} - \sqrt{\varphi_{i-1}}\big)} \big\|\sqrt{\varphi}\, I_{[i,N]}\big\|_{\mathbb{B}}.$$

To estimate $\big\|\sqrt{\varphi}\, I_{[i,N]}\big\|_{\mathbb{B}}$, we follow [4; Lemma 3.6]. Set

$$M_i = M_i(g) = \sum_{j=i}^N \mu_j g_j.$$

Then $M_i < \infty$ since $g \in \mathscr{G} \subset L^1(\mu)$. Furthermore, since $\varphi_i M_i \leqslant B_{\mathbb{B}}$ for all $i \in E_1$, we have for finite N, that

$$\begin{aligned}
\big\|\sqrt{\varphi}\, I_{[i,N]}\big\|_{\mathbb{B}} &= \sup_{g\in\mathscr{G}} \sum_{j=i}^{N} \sqrt{\varphi_j}\, g_j\mu_j \\
&= \sup_{g\in\mathscr{G}} \sum_{j=i}^{N} \sqrt{\varphi_j}\,(M_j - M_{j+1}) \\
&= \sup_{g\in\mathscr{G}} \bigg[\sqrt{\varphi_i}\, M_i + \sum_{j=i}^{N-1} \big(\sqrt{\varphi_{j+1}} - \sqrt{\varphi_j}\,\big) M_{j+1}\bigg] \\
&\leqslant B_{\mathbb{B}} \sup_{g\in\mathscr{G}} \bigg[1\big/\sqrt{\varphi_i} + \sum_{j=i}^{N-1} \big(1\big/\sqrt{\varphi_{j+1}} - \sqrt{\varphi_j}\big/\varphi_{j+1}\,\big)\bigg] \\
&\leqslant B_{\mathbb{B}} \sup_{g\in\mathscr{G}} \bigg[1\big/\sqrt{\varphi_i} + \sum_{j=i}^{N-1} \big(1\big/\sqrt{\varphi_j} - 1\big/\sqrt{\varphi_{j+1}}\,\big)\bigg] \\
&\leqslant 2B_{\mathbb{B}}\big/\sqrt{\varphi_i}, \qquad 1 \leqslant i \leqslant N.
\end{aligned}$$

Here in the second to last step, we have used the fact that

$$1\big/\sqrt{\varphi_{j+1}} - \sqrt{\varphi_j}\big/\varphi_{j+1} \leqslant 1\big/\sqrt{\varphi_j} - 1\big/\sqrt{\varphi_{j+1}}$$

and the convention $\sum_{\emptyset} = 0$. By passing limit, one can show that the above proof also works for $N = \infty$. Noticing that

$$\big(\sqrt{\varphi_i} - \sqrt{\varphi_{i-1}}\big)\sqrt{\varphi_i} \geqslant (\varphi_i - \varphi_{i-1})/2 = (2\mu_i a_i)^{-1},$$

we obtain

$$A_{\mathbb{B}} \leqslant \sup_{1\leqslant i\leqslant N} \frac{2B_{\mathbb{B}}}{\mu_i a_i\big(\sqrt{\varphi_i} - \sqrt{\varphi_{i-1}}\big)} \cdot \frac{1}{\sqrt{\varphi_i}} \leqslant 4B_{\mathbb{B}}.$$

This gives us the required assertion.

To prove (1.6), fix $i \in E_1$ and let $f = \varphi(i \wedge \cdot)$. We have

$$D(f) = \sum_{j=1}^{N} \mu_j a_j \big(\varphi(i\wedge j) - \varphi(i \wedge (j-1))\big)^2 = \varphi_i,$$

and hence $f \in \mathscr{D}_0$. By (1.3), we get $A_{\mathbb{B}} \geqslant \|\varphi(i\wedge\cdot)^2\|_{\mathbb{B}}/\varphi_i$. Making supremum with respect to $i \in E_1$, we obtain $A_{\mathbb{B}} \geqslant C_{\mathbb{B}}$. That is (1.6).

At the same time,

$$\begin{aligned}
C_{\mathbb{B}} &= \sup_{1\leqslant i\leqslant N} \frac{1}{\varphi_i} \sup_{g\in\mathscr{G}} \sum_{j=1}^{N} \varphi(i\wedge j)^2 \mu_j g_j \\
&\geqslant \sup_{1\leqslant i\leqslant N} \frac{1}{\varphi_i} \sup_{g\in\mathscr{G}} \sum_{j=i}^{N} \varphi_i^2 \mu_j g_j \\
&= \sup_{1\leqslant i\leqslant N} \varphi_i \big\|I_{[i,N]}\big\|_{\mathbb{B}} \\
&= B_{\mathbb{B}}.
\end{aligned}$$

This proves (1.7). $\square$

We mention that there is an alternative proof of (1.5)–(1.7) as presented in [1]. Using the estimates $C \leqslant A \leqslant 4B$ for some constants C and B given in [4], and the transform (2.4), we get $C(g) \leqslant A(g) \leqslant 4B(g)$ by passing limit. The required assertions then follow by making supremum with respect to $g \in \mathscr{G}$, plus some computations.

Proof of Theorem 1.2. (a) For $w \in \mathscr{W}''$, let

$$II_i(w) = \frac{1}{w_i} \sum_{j=1}^{i} \frac{1}{\mu_j a_j} \sum_{k=j}^{N} \mu_k w_k, \qquad i \in E_1. \tag{2.8}$$

Then, a variational formula was proven in [3; Theorem 2.1] as follows.

$$A = \inf_{w \in \mathscr{W}''} \sup_{1 \leqslant i \leqslant N} II_i(w) \tag{2.9}$$

Under the transform $\big(a_i, b_i, \mu_i\big) \to \big(a_i^{(n)}, b_i^{(n)}, \mu_i^{(n)}\big)$ given in (2.4), $II_i(w)$ becomes

$$II_i^{(n)}(w) := \frac{1}{w_i} \sum_{j=1}^{i} \frac{1}{\mu_j a_j} \sum_{k=j}^{N} \mu_k w_k g_k^{(n)}, \qquad i \in E_1, \qquad 1 \leqslant n \leqslant \infty. \tag{2.10}$$

We adopt the same notations $A\big(g^{(n)}\big)$ and $A(g)$ introduced in the proof of Theorem 1.1. Without lost of generality, assume that the right-hand side of (1.8) is finite. Then we have

$$A(g) = \lim_{n \to \infty} A\big(g^{(n)}\big) < \infty$$

and moreover,

$$\lim_{n \to \infty} \sup_{1 \leqslant i \leqslant N} \big| II_i^{(n)}(w) - II_i^{(\infty)}(w) \big| = 0.$$

Hence, applying (2.9) to $\big(A\big(g^{(n)}\big), II^{(n)}\big)$, we get

$$\begin{aligned}
A_{\mathbb{B}} &= \sup_{g \in \mathscr{G}} A(g) \\
&= \sup_{g \in \mathscr{G}} \lim_{n \to \infty} A\big(g^{(n)}\big) \\
&= \sup_{g \in \mathscr{G}} \lim_{n \to \infty} \inf_{w \in \mathscr{W}''} \sup_{1 \leqslant i \leqslant N} II_i^{(n)}(w) \\
&\leqslant \sup_{g \in \mathscr{G}} \inf_{w \in \mathscr{W}''} \sup_{1 \leqslant i \leqslant N} II_i^{(\infty)}(w) \\
&\leqslant \inf_{w \in \mathscr{W}''} \sup_{g \in \mathscr{G}} \sup_{1 \leqslant i \leqslant N} II_i^{(\infty)}(w) \\
&= \inf_{w \in \mathscr{W}''} \sup_{1 \leqslant i \leqslant N} \frac{1}{w_i} \sup_{g \in \mathscr{G}} \sum_{k=1}^{N} \mu_k w_k g_k \varphi(k \wedge i) \\
&= \inf_{w \in \mathscr{W}''} \sup_{1 \leqslant i \leqslant N} \frac{1}{w_i} \big\| w \varphi(i \wedge \cdot) \big\|_{\mathbb{B}}.
\end{aligned}$$

Here, in the second to last step, we have used the fact that

$$\sum_{j=1}^{i} \frac{1}{\mu_i a_i} \sum_{k=j}^{N} \mu_k w_k g_k = \sum_{k=1}^{N} \mu_k w_k g_k \varphi(i \wedge k).$$

This proves (2.9).

(b) We now prove the second part of the theorem. By an elementary proportion property, for every $w \in \mathscr{W}' \subset \mathscr{W}''$, we have

$$\sup_{1 \leqslant i \leqslant N} II_i(w) \leqslant \sup_{1 \leqslant i \leqslant N} I_i(w).$$

With the help of the transform (2.4), we get

$$\sup_{1 \leqslant i \leqslant N} II_i^{(\infty)}(w) \leqslant \sup_{1 \leqslant i \leqslant N} I_i^{(\infty)}(w).$$

By setting $w = \sqrt{\varphi}$ and then making supremum with respect to $g \in \mathscr{G}$, we get

$$D_{\mathbb{B}}(1) = \sup_{1 \leqslant i \leqslant N} \frac{1}{\sqrt{\varphi_i}} \big\| \sqrt{\varphi}\, \varphi(i \wedge \cdot) \big\|_{\mathbb{B}} \leqslant \sup_{1 \leqslant i \leqslant N} \frac{1}{\mu_i a_i \big(\sqrt{\varphi_i} - \sqrt{\varphi_{i-1}}\,\big)} \big\| \sqrt{\varphi} I_{[i,N]} \big\|_{\mathbb{B}}.$$

From the first part of the proof (c) of Theorem 1.1, it follows that the right-hand side is controlled by $4B_{\mathbb{B}}$. Hence, we have $D_{\mathbb{B}}(1) \leqslant 4B_{\mathbb{B}}$.

The monotonicity of $D_{\mathbb{B}}(n)$ is simple: By definition, $w^{(n)} \leqslant D_{\mathbb{B}}(n) w^{(n-1)}$. Hence

$$D_{\mathbb{B}}(n+1) = \sup_{1 \leqslant i \leqslant N} \frac{\|w^{(n)} \varphi(i \wedge \cdot)\|_{\mathbb{B}}}{w_i^{(n)}} \leqslant D_{\mathbb{B}}(n) \sup_{1 \leqslant i \leqslant N} \frac{\|w^{(n-1)} \varphi(i \wedge \cdot)\|_{\mathbb{B}}}{w_i^{(n)}} = D_{\mathbb{B}}(n)$$

for all $n \geqslant 1$. On the other hand, by assumption, $D_{\mathbb{B}}(1) \leqslant 4B_{\mathbb{B}} < \infty$. From this and induction, it follows that $w^{(n)} \in \mathscr{W}''$ for all $n \geqslant 0$. This gives us by (1.8) that $D_{\mathbb{B}}(n) \geqslant A_{\mathbb{B}}$ and then $\lim_{n \to \infty} D_{\mathbb{B}}(n) \geqslant A_{\mathbb{B}}$. $\square$

3. Neumann Case: Case 1. Instead of the Dirichlet boundary condition (i.e., $f_0 = 0$), we consider the Neumann case in this and the next sections. In this section, only single infinity (i.e., $N = \infty$) is allowed but in the next section we may have double infinities: $\{\cdots, -2, -1, 0, 1, 2, \cdots\}$. Now, instead of (1.2), we study the following inequality

$$\|\bar{f}^2\|_{\mathbb{B}} \leqslant \overline{A}_{\mathbb{B}} D(f), \tag{3.1}$$

where $\bar{f} = f - \pi(f)$. Since $\bar{f}$ usually does not vanish at the boundary 0, this boundary can not be ignored, the state space now becomes $E = \{0, 1, \cdots, N\}$ rather than $E_1 := \{1, 2, \cdots, N\}$ used in the last section. Thus, the Banach space $(\mathbb{B}, \|\cdot\|_{\mathbb{B}}, \mu)$ is assumed to be the functions of $E \to \mathbb{R}$ satisfying the same conditions mentioned in the first section. In the study of (3.1), without loss of generality, we may and will assume that $f_0 = 0$. Define a projection of the Banach space $(\mathbb{B}, \|\cdot\|_{\mathbb{B}}, \mu)$ to E_1 as follows.

$$\mathbb{B}^1 = \{fI_{\mathbb{E}_1} : f \in \mathbb{B}\}, \quad \mu^1 = \mu|_{E^1}, \quad \mathscr{G}^1 = \{gI_{E_1} : g \in \mathscr{G}\}.$$

Clearly, with the norm

$$\|f\|_{\mathbb{B}^1} = \sup_{g\in\mathscr{G}^1} \sum_{i\in E_1} |f_i|\mu_i^1 g_i,$$

$(\mathbb{B}^1, \|\cdot\|_{\mathbb{B}^1}, \mu^1)$ is, when restricted to E_1, a Banach space satisfying the conditions listed in Section 1. Therefore, Theorems 1.1 and 1.2 are available for $(\mathbb{B}^1, \|\cdot\|_{\mathbb{B}^1}, \mu^1)$. Note that $\|f\|_{\mathbb{B}^1} = \|f\|_{\mathbb{B}}$ for all $f \in \mathbb{B}$ with $f_0 = 0$ (i.e., $f \in \mathbb{B}^1$). Because of this, throughout this section, when $f_0 = 0$, we simply write $\|f\|_{\mathbb{B}}$ instead of $\|f\|_{\mathbb{B}^1}$. The main purpose of this section is to compare the optimal constant $\overline{A}_{\mathbb{B}}$ in (3.1) with $A_{\mathbb{B}} := A_{\mathbb{B}^1}$ given in Section 1. Here is our first result.

Theorem 3.1. Let (1.1) hold and let c_1 and c_2 be constants such that $|\pi(f)| \leqslant c_1\|f\|_{\mathbb{B}}$ and $|\pi(fI_{E_1})| \leqslant c_2\|fI_{E_1}\|_{\mathbb{B}}$ for all $f \in \mathbb{B}$. Then, we have

$$\max\big\{\|1\|_{\mathbb{B}}^{-1},\ \big(1 - \sqrt{c_2(1-\pi_0)\|1\|_{\mathbb{B}}}\,\big)^2\big\}A_{\mathbb{B}} \leqslant \overline{A}_{\mathbb{B}} \leqslant \big(1 + \sqrt{c_1\|1\|_{\mathbb{B}}}\,\big)^2 A_{\mathbb{B}}, \quad (3.2)$$

here, for the second lower bound, it is assumed that $c_2(1-\pi_0)\|1\|_{\mathbb{B}} < 1$. In particular, $\overline{A}_{\mathbb{B}} < \infty$ iff $B_{\mathbb{B}} < \infty$.

Proof. As shown in proof (b) of Theorem 1.1, the first assumption implies that $|\mu(f)| \leqslant \|f\|_{\mathbb{B}}/\inf g^{(0)}$ and so we may assume that $c_1, c_2 \leqslant \big(Z \inf g^{(0)}\big)^{-1} < \infty$.

Note that $\pi_0 = \mu_0/Z = 1/Z$, where $Z = \sum_{i=0}^N \mu_i = \mu(1)$. By [4; Theorem 3.5], the optimal constant $\overline{A}$ in the ordinary form of the Poincaré inequality

$$\|\bar{f}\|_2^2 \leqslant \overline{A}D(f), \quad (3.3)$$

satisfies $\overline{A} \geqslant A/Z$, where A is the optimal constant in (2.1). With the help of the transform (2.4), we obtain

$$\overline{A}(g) \geqslant A(g)/\mu(g). \quad (3.4)$$

At the same time, the left-hand side of (3.3) is changed to

$$\sum_{i=0}^N \big(f_i - \pi_g(f)\big)^2 g_i\mu_i = \inf_{c\in\mathbb{R}} \sum_{i=0}^N (f_i - c)^2 g_i\mu_i,$$

where $\pi_g(f) = \sum_i g_i f_i \mu_i/\mu(g)$. Note that

$$\begin{aligned}\overline{A}(g) &= \sup_{f\in\mathscr{D}_0} \frac{\sum_i (f_i - \pi_g(f))^2\mu_i g_i}{D(f)}\\ &= \sup_{f\in\mathscr{D}} \inf_{c\in\mathbb{R}} \frac{\sum_i (f_i - c)^2\mu_i g_i}{D(f)}\\ &\leqslant \inf_{c\in\mathbb{R}} \sup_{f\in\mathscr{D}_0} \frac{\sum_i (f_i - c)^2\mu_i g_i}{D(f)}.\end{aligned}$$

We have

$$\begin{aligned}\sup_{g\in\mathscr{G}}\overline{A}(g)&\leqslant \sup_{g\in\mathscr{G}}\inf_{c\in\mathbb{R}}\sup_{f\in\mathscr{D}_0}\frac{\sum_i (f_i-c)^2\mu_i g_i}{D(f)}\\&\leqslant \inf_{c\in\mathbb{R}}\sup_{f\in\mathscr{D}_0}\frac{\|(f-c)^2\|_{\mathbb{B}}}{D(f)}\\&\leqslant \overline{A}_{\mathbb{B}}\wedge A_{\mathbb{B}},\end{aligned}$$

by setting $c=\pi(f)$ and $c=f_0$, respectively, in the last step. This is clearly very different from (2.7). Combining this with (3.4), we obtain

$$\overline{A}_{\mathbb{B}}\geqslant \sup_{g\in\mathscr{G}}\overline{A}(g)\geqslant \sup_{g\in\mathscr{G}}\frac{A(g)}{\mu(g)}\geqslant \frac{\sup_{g\in\mathscr{G}}A(g)}{\sup_{g\in\mathscr{G}}\mu(g)}=\frac{A_{\mathbb{B}}}{\|1\|_{\mathbb{B}}}.$$

This gives us the first lower bound in (3.2). The other assertions of the theorem can be deduced from the comparison result, Proposition 3.2 (cf. [1] and references within) below. Actually, let $f^2\in\mathbb{B}$ satisfy $f_0=0$. Then the upper bound follows from part (1) of Proposition 3.2. The second lower bound follows from part (2) of the proposition with $A=E_1$. The last assertion follows from Theorem 1.1. $\square$

Proposition 3.2. Let $(E,\mathscr{E},\pi)$ be a probability space and $(\mathbb{B},\|\cdot\|_{\mathbb{B}})$ be a Banach space, satisfying conditions (1) and (2) in (1.1), of Borel measurable functions on $(E,\mathscr{E},\pi)$.

(1) Assume that there is a constant c_1 such that $|\pi(f)|\leqslant c_1\|f\|_{\mathbb{B}}$ for all $f\in\mathbb{B}$. Then

$$\big\|\bar{f}^2\big\|_{\mathbb{B}}\leqslant\big(1+\sqrt{c_1\|1\|_{\mathbb{B}}}\,\big)^2\big\|f^2\big\|_{\mathbb{B}}. \tag{3.5}$$

(2) Next, for a given $A\in\mathscr{E}$, let $c_2(A)$ be the constant such that $|\pi(fI_A)|\leqslant c_2(A)\|fI_A\|_{\mathbb{B}}$ for all $f\in\mathbb{B}$. If $c_2(A)\pi(A)\,\|1\|_{\mathbb{B}}<1$, then for every f with $f|_{A^c}=0$ we have

$$\big\|f^2\big\|_{\mathbb{B}}\leqslant\big\|\bar{f}^2\big\|_{\mathbb{B}}\big/\big[1-\sqrt{c_2(A)\pi(A)\,\|1\|_{\mathbb{B}}}\,\big]^2. \tag{3.6}$$

Proof. (a) By assumption, we have $\pi(f)^2\leqslant\pi\big(f^2\big)\leqslant c_1\big\|f^2\big\|_{\mathbb{B}}$. Thus, for every pair $p,q>1$ with $(p-1)(q-1)=1$, we have

$$\big\|\bar{f}^2\big\|_{\mathbb{B}}=\big\|(f-\pi(f))^2\big\|_{\mathbb{B}}\leqslant p\big\|f^2\big\|_{\mathbb{B}}+q\pi(f)^2\|1\|_{\mathbb{B}}\leqslant\big(p+c_1q\|1\|_{\mathbb{B}}\big)\big\|f^2\big\|_{\mathbb{B}}.$$

Minimizing the coefficients on the right-hand side with respect to p and q, we get $p=1+\sqrt{c_1\|1\|_{\mathbb{B}}}>1$, $q=1+\big(c_1\|1\|_{\mathbb{B}}\big)^{-1/2}>1$. The first assertion follows.

(b) Similarly, the assumption gives us

$$\pi(f)^2=\pi\big(fI_A\big)^2\leqslant\pi(A)\pi\big(f^2\big)\leqslant\pi(A)c_2(A)\big\|f^2\big\|_{\mathbb{B}}.$$

Thus, for every pair $p,q>1$ with $(p-1)(q-1)=1$, we have

$$\big\|f^2\big\|_{\mathbb{B}}\leqslant p\big\|(f-\pi(f))^2\big\|_{\mathbb{B}}+q\pi(f)^2\|1\|_{\mathbb{B}}\leqslant p\big\|\bar{f}^2\big\|_{\mathbb{B}}+q\pi(A)c_2(A)\|1\|_{\mathbb{B}}\big\|f^2\big\|_{\mathbb{B}}.$$

Hence

$$\|f^2\|_{\mathbb{B}} \leqslant \frac{p}{1 - q\pi(A)c_2(A)\|1\|_{\mathbb{B}}}\|\bar{f}^2\|_{\mathbb{B}},$$

provided $q\pi(A)c_2(A)\|1\|_{\mathbb{B}} < 1$. Minimizing the coefficients on the right-hand side, we get $p = \big(1 - \sqrt{\pi(A)c_2(A)\|1\|_{\mathbb{B}}}\,\big)^{-1} > 1$, $q = \big(\pi(A)c_2(A)\|1\|_{\mathbb{B}}\big)^{-1/2} > 1$ and $q\,\pi(A)c_2(A)\|1\|_{\mathbb{B}} = \big(\pi(A)c_2(A)\|1\|_{\mathbb{B}}\big)^{1/2} < 1$ by assumption. Then, we obtain the second assertion. $\square$

In parallel, one may consider the inequalities modified from (1.2) and (3.1), respectively, for general Banach space $(\mathbb{B}, \|\cdot\|_{\mathbb{B}}, \mu)$ of functions $f\colon E \to \mathbb{R}$:

$$\|f\|_{\mathbb{B}}^2 \leqslant A'_{\mathbb{B}} D(f), \qquad f_0 = 0 \tag{3.7}$$

$$\|\bar{f}\|_{\mathbb{B}}^2 \leqslant \overline{A}'_{\mathbb{B}} D(f). \tag{3.8}$$

Then, we have the following result.

Theorem 3.3. Everything in premise is the same as in Theorem 3.1, but assuming only conditions (1) and (2) in (1.1). For the optimal constants $\overline{A}'_{\mathbb{B}}$ and $A'_{\mathbb{B}} = A'_{\mathbb{B}^1}$ in (3.7) and (3.8), respectively, we have

$$\big(1 - c_2\|1\|_{\mathbb{B}}\big)^2 A'_{\mathbb{B}} \leqslant \overline{A}'_{\mathbb{B}} \leqslant \big(1 + c_1\|1\|_{\mathbb{B}}\,\big)^2 A'_{\mathbb{B}}, \tag{3.9}$$

here, for the lower bound, it is assumed that $c_2\|1\|_{\mathbb{B}} < 1$. In particular, $\overline{A}'_{\mathbb{B}} < \infty$ iff $A'_{\mathbb{B}} < \infty$.

Actually, Theorem 3.3 is an immediate consequence of the following proposition (which is a slight modification of [1; Proposition 2.3]).

Proposition 3.4. Everything in premise is the same as in Proposition 3.2.

(1) Assume that there is a constant c_1 such that $|\pi(f)| \leqslant c_1\|f\|_{\mathbb{B}}$ for all $f \in \mathbb{B}$. Then

$$\|\bar{f}\|_{\mathbb{B}} \leqslant \big(1 + c_1\|1\|_{\mathbb{B}}\big)\|f\|_{\mathbb{B}}. \tag{3.10}$$

(2) Next, for a given $A \in \mathscr{E}$, let $c_2(A)$ be the constant such that $|\pi(fI_A)| \leqslant c_2(A)\|fI_A\|_{\mathbb{B}}$ for all $f \in \mathbb{B}$. If $c_2(A)\|1\|_{\mathbb{B}} < 1$, then for every f with $f|_{A^c} = 0$ we have

$$\|f\|_{\mathbb{B}} \leqslant \|\bar{f}\|_{\mathbb{B}} / \big[1 - c_2(A)\|1\|_{\mathbb{B}}\big]. \tag{3.11}$$

Theorem 3.1 is often powerful to provide a criterion for $\overline{A}_{\mathbb{B}} < \infty$, in terms of $B_{\mathbb{B}}$. However, the estimates of $\overline{A}_{\mathbb{B}}$ given by the theorem are usually quite rough. To improve them, one needs a different approach. Note that we do have a formula for the optimal constant $\overline{A}$ in (3.3), similar to (2.2):

$$\overline{A} = \inf_{w \in \mathscr{W}'} \sup_{1 \leqslant i \leqslant N} I_i(\bar{w})$$

(cf. [2; Theorem 3.2] or [3; Theorem 2.3]). However, we do not know how to extend this result to the Banach space. As we have seen from the proof of Theorem 3.1, on the left-hand side of (3.3), the term $\pi(f)$ is not invariant under the transform (2.4). Moreover, since $\pi(\bar{w}) = \sum_{i=0}^N \bar{w}_i\pi_i = 0$, it is easy to check that for each fixed $w \in \mathscr{W}'$, $I_i(\bar{w})$ is positive for all $i \geqslant 1$. But this property is no longer true when $\mathrm{d}\mu$ is replaced by $\mathrm{d}\mu_g = g\mathrm{d}\mu$.

Fortunately, there is another approach which works well in the present context. The intuitive idea goes as follows: Since a function that attains the optimal constant $\overline{A}_{\mathbb{B}}$ must change signs, it may vanish somewhere, say θ for instance. If so, it is natural to divide the interval $(0, N]$ into two parts: $(0, \theta)$ and $(\theta, N]$. Then, one compares $\overline{A}_{\mathbb{B}}$ with the optimal constants of inequality (1.2) on $(0, \theta)$ and $(\theta, N]$, respectively. It can also happen that the function does not vanish anywhere in the discrete case (but not in the continuous case). However, we do not care about the existence of the vanishing point θ. Such θ is unknown, even it exists. In practice, we regard θ as a reference point and then apply an optimization procedure to θ. This is the goal of the study in the next section.

Similar remarks are valid for Theorem 3.3.

4. Neumann Case: Case 2. In this section, we consider the state space $E = \{M, M-1, \cdots, N-1, N\}$, $-\infty \leqslant M,\ N \leqslant \infty$. Again, we often denote the set $\{m+1, m, \cdots, n-1\}$ by (m, n), and similarly, we have $[m, n]$ and so on. The Q-matrix now is $q_{i,i+1} = b_i > 0$, $q_{i,i-1} = a_i > 0$ and $q_{ij} = 0$ if $|i-j| > 1$ with $b_N = 0$ if $N < \infty$ and $a_M = 0$ if $M > -\infty$. Fix a reference point $\theta \in (M, N)$. Define

$$\mu_\theta = \frac{1}{a_\theta b_\theta},\ \ \mu_{\theta+1} = \frac{1}{a_\theta a_{\theta+1}},\ \ \mu_{\theta+n} = \frac{b_{\theta+1}b_{\theta+2}\cdots b_{\theta+n-1}}{a_\theta a_{\theta+1}\cdots a_{\theta+n}}, \quad 2 \leqslant n \leqslant N-\theta$$
$$\mu_{\theta-1} = \frac{1}{b_\theta b_{\theta-1}},\ \ \mu_{\theta+n} = \frac{a_{\theta-1}a_{\theta-2}\cdots a_{\theta+n+1}}{b_\theta b_{\theta-1}\cdots b_{\theta+n}},\ \ M-\theta \leqslant n \leqslant -2.$$

Since we are working in the ergodic situation, it is natural to assume that the process is non-explosive:

$$\sum_{n\geqslant\theta+1}\frac{1}{\mu_n a_n}\sum_{j=\theta}^{n-1}\mu_j = \infty \text{ if } N=\infty \text{ and } \sum_{n\leqslant\theta-1}\frac{1}{\mu_n b_n}\sum_{j=n+1}^{\theta}\mu_j = \infty \text{ if } M=-\infty. \tag{4.0}$$

Given a Banach space $(\mathbb{B}, \|\cdot\|_{\mathbb{B}}, \mu)$ of functions $E \to \mathbb{R}$ with norm $\|f\|_{\mathbb{B}} = \sup_{g\in\mathscr{G}}\sum_{i\in E}|f_i|g_i\mu_i$, define

$$\mathbb{B}^{1\theta} = \{fI_{[\theta+1,N]} : f \in \mathbb{B}\}, \quad \mu^{1\theta} = \mu|_{[\theta+1,N]}, \quad \mathscr{G}^{1\theta} = \{gI_{[\theta+1,N]} : g \in \mathscr{G}\}$$

and

$$\|f\|_{\mathbb{B}^{1\theta}} = \sup_{g\in\mathscr{G}^{1\theta}}\sum_{i=\theta+1}^{N}|f_i|g_i\mu_i^{1\theta} = \sup_{g\in\mathscr{G}^{1\theta}}\sum_{i=\theta+1}^{N}|f_i|g_i\mu_i.$$

It is easy to check that $1_{[\theta+1,N]} \in \mathbb{B}^{1\theta}$ and $(\mathbb{B}^{1\theta}, \|\cdot\|_{\mathbb{B}^{1\theta}}, \mu^{1\theta})$ is an ideal space. Similarly, we can define $(\mathbb{B}^{2\theta}, \|\cdot\|_{\mathbb{B}^{2\theta}}, \mu^{2\theta})$, corresponding to $[M, \theta-1]$.

The Ponicaré-type inequality that we are interested in this section is formally the same as (3.1):

$$\left\|\bar{f}^2\right\|_{\mathbb{B}} \leqslant \overline{A}_{\mathbb{B}}D(f), \tag{4.1}$$

where $\bar f = f - \pi(f)$,

$$\begin{aligned}
&D(f) = \sum_{i=\theta+1}^{N} \mu_i a_i (f_i - f_{i-1})^2 + \sum_{i=M}^{\theta-1} \mu_i b_i (f_{i+1} - f_i)^2,\\
&Z_{1\theta} = \sum_{i=\theta+1}^{N} \mu_i, \quad Z_{2\theta} = \sum_{i=M}^{\theta-1} \mu_i, \quad Z = Z_{1\theta} + Z_{2\theta} + \mu_\theta < \infty,\\
&\pi_i = \frac{\mu_i}{Z}.
\end{aligned} \tag{4.2}$$

The expression of $D(f)$ may look strange but it is indeed standard, since

$$\sum_{M\leqslant i\leqslant N-1} \pi_i b_i (f_{i+1} - f_i)^2 = \sum_{M\leqslant i\leqslant \theta-1} \pi_i b_i (f_{i+1} - f_i)^2 + \sum_{\theta\leqslant i\leqslant N-1} \pi_i b_i (f_{i+1} - f_i)^2$$

and

$$\sum_{\theta\leqslant i\leqslant N-1} \pi_i b_i (f_{i+1} - f_i)^2 = \sum_{\theta\leqslant i\leqslant N-1} \pi_{i+1} a_{i+1} (f_{i+1} - f_i)^2 = \sum_{\theta+1\leqslant i\leqslant N} \pi_i a_i (f_i - f_{i-1})^2.$$

To state our result, define the following quantities which will be used several times subsequently.

$$\begin{aligned}
&\varphi_i^{1\theta} = \sum_{j=\theta+1}^{i} \frac{1}{\mu_j a_j}, \quad \theta+1 \leqslant i \leqslant N, && \varphi_i^{2\theta} = \sum_{j=i}^{\theta-1} \frac{1}{\mu_j b_j}, \quad M \leqslant i \leqslant \theta-1.\\
&B_{\mathbb{B}^{1\theta}} = \sup_{\theta+1\leqslant i\leqslant N} \varphi_i^{1\theta} \big\| I_{[i,N]} \big\|_{\mathbb{B}^{1\theta}}, && B_{\mathbb{B}^{2\theta}} = \sup_{M\leqslant i\leqslant \theta-1} \varphi_i^{2\theta} \big\| I_{[M,i]} \big\|_{\mathbb{B}^{2\theta}}.\\
&C_{\mathbb{B}^{1\theta}} = \sup_{\theta+1\leqslant i\leqslant N} \frac{\big\| \varphi^{1\theta}(i\wedge \cdot)^2 \big\|_{\mathbb{B}^{1\theta}}}{\varphi_i^{1\theta}}, && C_{\mathbb{B}^{2\theta}} = \sup_{M\leqslant i\leqslant \theta-1} \frac{\big\| \varphi^{2\theta}(i\vee \cdot)^2 \big\|_{\mathbb{B}^{2\theta}}}{\varphi_i^{2\theta}}.\\
&D_{\mathbb{B}^{1\theta}} = \sup_{\theta+1\leqslant i\leqslant N} \frac{\big\| \sqrt{\varphi^{1\theta}}\, \varphi^{1\theta}(i\wedge \cdot) \big\|_{\mathbb{B}^{1\theta}}}{\sqrt{\varphi_i^{1\theta}}}, && D_{\mathbb{B}^{2\theta}} = \sup_{M\leqslant i\leqslant \theta-1} \frac{\big\| \sqrt{\varphi^{2\theta}}\, \varphi^{2\theta}(i\vee \cdot) \big\|_{\mathbb{B}^{2\theta}}}{\sqrt{\varphi_i^{2\theta}}}.
\end{aligned} \tag{4.3}$$

From now on, we often state only explicit bounds, the corresponding variational formula follows from (1.4) and (1.8) immediately.

Theorem 4.1. Let (1.1) hold and assume that $\mathscr{G}^{1\theta}$, $\mathscr{G}^{2\theta} \subset \mathscr{G}$ for all $\theta \in (M,N)$.

(1) In general, we have

$$\overline{A}_{\mathbb{B}} \leqslant \inf_{\theta\in(M,N)} A_{\mathbb{B}^{1\theta}} \vee A_{\mathbb{B}^{2\theta}} \leqslant \inf_{\theta\in(M,N)} D_{\mathbb{B}^{1\theta}} \vee D_{\mathbb{B}^{2\theta}} \leqslant 4 \inf_{\theta\in(M,N)} B_{\mathbb{B}^{1\theta}} \vee B_{\mathbb{B}^{2\theta}}. \tag{4.4}$$

(2) Assume additionally that $(g^{(1)} I_{[\theta+1,N]} + g^{(2)} I_{[M,\theta-1]})/2 \in \mathscr{G}$ for every $g^{(1)} \in \mathscr{G}^{1\theta}$ and $g^{(2)} \in \mathscr{G}^{2\theta}$. Then, we have

$$\overline{A}_{\mathbb{B}} \geqslant \frac{1}{2} \sup_{\theta\in(M,N)} A_{\mathbb{B}^{1\theta}} \wedge A_{\mathbb{B}^{2\theta}} \geqslant \frac{1}{2} \sup_{\theta\in(M,N)} C_{\mathbb{B}^{1\theta}} \wedge C_{\mathbb{B}^{2\theta}} \geqslant \frac{1}{2} \sup_{\theta\in(M,N)} B_{\mathbb{B}^{1\theta}} \wedge B_{\mathbb{B}^{2\theta}}. \tag{4.5}$$

(3) For each A, let $c(A)$ be a constant satisfying $|\pi(fI_A)| \leqslant c(A)\|fI_A\|_{\mathbb{B}}$ for all $f \in \mathbb{B}$. Under the above assumption on $\mathscr{G}^{1\theta}, \mathscr{G}^{2\theta}$ and $\mathscr{G}$, if $c(A) < \infty$ for all A with $\pi(A) \ll 1$ (i.e., $\pi(A)$ is sufficient small), then $\overline{A}_{\mathbb{B}} < \infty$ iff $B_{\mathbb{B}^{1\theta}} \vee B_{\mathbb{B}^{2\theta}} < \infty$ for some (equivalent, for all) $\theta \in (M,N)$.

Proof. (a) First, we prove (4.4). Let $\theta \in (M, N)$ and $g \in \mathscr{G}$ with $g > 0$. Define $g^{(1)} = gI_{(\theta,N]}$ and $g^{(2)} = gI_{[M,\theta)}$. Then $g^{(1)} \in \mathscr{G}^{1\theta}$ and $g^{(2)} \in \mathscr{G}^{2\theta}$ by assumption. The proof of [4; Theorem 3.4] shows that the optimal constant $\overline{A}(g)$, obtained by replacing a with a/g and b with b/g in (3.1), satisfies $\overline{A}(g) \leqslant \inf_{\theta\in(M,N)} A^{1\theta}(g^{(1)}) \vee A^{2\theta}(g^{(2)})$, where $A^{1\theta}(g^{(1)})$ and $A^{2\theta}(g^{(2)})$ are the corresponding optimal constants in (1.2) with respect to the intervals (θ, N) and (M, θ), respectively. As we did in the proof of Theorem 1.1, by using (2.4) and passing to the limit, one may ignore the condition "$g > 0$". Then,

$$\begin{aligned}\overline{A}_{\mathbb{B}} &\leqslant \sup_{g\in\mathscr{G}} \inf_{\theta\in(M,N)} A^{1\theta}(g^{(1)}) \vee A^{2\theta}(g^{(2)})\\ &\leqslant \inf_{\theta\in(M,N)} \sup_{g\in\mathscr{G}} A^{1\theta}(g^{(1)}) \vee A^{2\theta}(g^{(2)})\\ &\leqslant \inf_{\theta\in(M,N)} \Big[\sup_{g\in\mathscr{G}^{1\theta}} A^{1\theta}(g)\Big] \vee \Big[\sup_{g\in\mathscr{G}^{2\theta}} A^{2\theta}(g)\Big]\\ &= \inf_{\theta\in(M,N)} A_{\mathbb{B}^{1\theta}} \vee A_{\mathbb{B}^{2\theta}}.\end{aligned}$$

Combining this with Theorem 1.2, we obtain (4.4).

(b) Fix $\theta \in (M, N)$ and $\varepsilon > 0$. Choose $f^{(1)},\ f^{(2)} \geqslant 0$ such that

$$f^{(1)}|_{(M,\theta)} = 0,\ f^{(2)}|_{(\theta,N)} = 0, \quad \big\|\big(f^{(1)}\big)^2\big\|_{\mathbb{B}^{1\theta}} = \big\|\big(f^{(2)}\big)^2\big\|_{\mathbb{B}^{2\theta}} = 1$$

and

$$D(f^{(1)}) \leqslant A_{\mathbb{B}^{1\theta}}^{-1} + \varepsilon, \qquad D(f^{(2)}) \leqslant A_{\mathbb{B}^{2\theta}}^{-1} + \varepsilon.$$

Next, choose $g^{(1)} \in \mathscr{G}^{1\theta}$ and $g^{(2)} \in \mathscr{G}^{2\theta}$ such that

$$\sum_{i=\theta+1}^{N} \big(f_i^{(1)}\big)^2 g_i^{(1)} \mu_i \geqslant 1 - \varepsilon \quad \text{and} \quad \sum_{i=M}^{\theta-1} \big(f_i^{(2)}\big)^2 g_i^{(2)} \mu_i \geqslant 1 - \varepsilon.$$

Set

$$f = -\sqrt{\lambda}\, f^{(1)} I_{[\theta+1,N]} + \sqrt{1-\lambda}\, f^{(2)} I_{[M,\theta-1]},$$

where

$$\lambda = \pi\Big(\big(f^{(2)}\big)^2\Big)\Big/\Big[\pi\Big(\big(f^{(1)}\big)^2\Big) + \pi\Big(\big(f^{(2)}\big)^2\Big)\Big]$$

is the constant so that $\pi(f) = 0$. Then

$$\begin{aligned}
D(f) &= \lambda D(f^{(1)}) + (1-\lambda) D(f^{(2)}) \\
&\leqslant \lambda\big(A_{\mathbb{B}^{1\theta}}^{-1} + \varepsilon\big) + (1-\lambda)\big(A_{\mathbb{B}^{2\theta}}^{-1} + \varepsilon\big) \\
&\leqslant \big(A_{\mathbb{B}^{1\theta}}^{-1} \vee A_{\mathbb{B}^{2\theta}}^{-1} + \varepsilon\big)\big(\lambda + (1-\lambda)\big) \\
&\leqslant \big(A_{\mathbb{B}^{1\theta}}^{-1} \vee A_{\mathbb{B}^{2\theta}}^{-1} + \varepsilon\big)\bigg(\lambda \sum_{i=\theta+1}^{N} \big(f_i^{(1)}\big)^2 g_i^{(1)} \mu_i + (1-\lambda) \sum_{i=M}^{\theta-1} \big(f_i^{(2)}\big)^2 g_i^{(2)} \mu_i + \varepsilon\bigg) \\
&= \big(A_{\mathbb{B}^{1\theta}}^{-1} \vee A_{\mathbb{B}^{2\theta}}^{-1} + \varepsilon\big)\bigg(\sum_{i=M}^{N} \Big[\lambda\big(f_i^{(1)}\big)^2 g_i^{(1)} + (1-\lambda)\big(f_i^{(2)}\big)^2 g_i^{(2)}\Big]\mu_i + \varepsilon\bigg) \\
&= 2\big(A_{\mathbb{B}^{1\theta}}^{-1} \vee A_{\mathbb{B}^{2\theta}}^{-1} + \varepsilon\big)\bigg(\sum_{i=M}^{N} f_i^2 \big[g_i^{(1)}/2 + g_i^{(2)}/2\big]\mu_i + \varepsilon/2\bigg) \\
&\leqslant 2\big(A_{\mathbb{B}^{1\theta}}^{-1} \vee A_{\mathbb{B}^{2\theta}}^{-1} + \varepsilon\big)\Big(\big\|f^2\big\|_{\mathbb{B}} + \varepsilon\Big).
\end{aligned}$$

Here in the last step, we have used the fact that $(g^{(1)} I_{(\theta,N]} + g^{(2)} I_{[M,\theta)})/2 \in \mathscr{G}$. Letting $\varepsilon \to 0$ and then making infimum with respect to θ, we obtain the first inequality in (4.5). Then, the second and the third ones follow from Theorems 1.1.

(c) To prove part (3) of the theorem, note that $I_{[M,i)},\ I_{(i,N)} \in \mathbb{B}$ and so $B_{\mathbb{B}^{1\theta}}, B_{\mathbb{B}^{2\theta}} < \infty$ when M and N are both finite. In general, if $B_{\mathbb{B}^{1\theta}} \vee B_{\mathbb{B}^{2\theta}} < \infty$, then $\overline{A}_{\mathbb{B}} < \infty$ by (4.4).

Next, consider the case where $M > -\infty$ and $N = \infty$. We need only to handle with the case that $B_{\mathbb{B}^{1\theta}} = \infty$ since $B_{\mathbb{B}^{2\theta}} < \infty$. Noting that when $\theta' \geqslant \theta$,

$$\sup_{g \in \mathscr{G}^{2\theta}} \sum_{i=M}^{k} g_i \mu_i^{2\theta} = \sup_{g \in \mathscr{G}^{2\theta}} \sum_{i=M}^{k} g_i \mu^{2\theta'} \leqslant \sup_{g \in \mathscr{G}} \sum_{i=M}^{k} g_i \mu_i^{2\theta'} = \sup_{g \in \mathscr{G}^{2\theta'}} \sum_{i=M}^{k} g_i \mu_i^{2\theta'}$$

for all $i \in (M,\theta)$, we have $\big\|I_{[M,i]}\big\|_{\mathbb{B}^{2\theta}} \leqslant \big\|I_{[M,i]}\big\|_{\mathbb{B}^{2\theta'}}$ for all $i \in [M,\theta]$. Hence, $B_{\mathbb{B}^{2\theta}} \uparrow$ and $B_{\mathbb{B}^{1\theta}} \downarrow$ strictly as $\theta \uparrow$. Note that $B_{\mathbb{B}^{1\theta}} = \infty$ for some (equivalently, for all) $\theta \in (M,N)$. On the other hand, by condition (4.0) and the ergodicity, $\varphi_i \uparrow \infty$ as $i \to \infty$. We have $B_{\mathbb{B}^{2\theta}} \uparrow \infty$ as $\theta \uparrow \infty$. Clearly, $B_{\mathbb{B}^{1\theta}}$ and $B_{\mathbb{B}^{2\theta}}$ have different values at $\theta = M$ and $\theta = N$. Thus, as θ varies, the two curves $B_{\mathbb{B}^{2\theta}}$ and $B_{\mathbb{B}^{1\theta}}$ must intersect uniquely at $N = \infty$. Furthermore, the lower bound given in (4.5) equals ∞. Therefore, $\overline{A}_{\mathbb{B}} = \infty$ by (4.5). The case of $M = -\infty$ and $N < \infty$ is symmetric and so can be proven in a similar way.

It remains to consider only the case where $(M,N) = \mathbb{Z}$. We need to prove that $B_{\mathbb{B}^{1\theta}} \vee B_{\mathbb{B}^{2\theta}} < \infty$ if $A_{\mathbb{B}} < \infty$. By assumption, for sufficiently small θ, condition (2) of Proposition 3.4 is satisfied with $A = (-\infty,\theta)$. Then we have for every f with $f|_{A^c} = 0$ that $\|f^2\|_{\mathbb{B}^{1\theta}} \leqslant c' \|\bar{f}^2\|_{\mathbb{B}}$ for some constant $c' < \infty$. Thus, once (4.1) holds, we must have $B_{\mathbb{B}^{1\theta}} \leqslant A_{\mathbb{B}^{1\theta}} < \infty$ first for sufficient small θ and then for all $\theta \in (M,N)$. By symmetry, the same conclusion holds for $B_{\mathbb{B}^{2\theta}}$ and so $B_{\mathbb{B}^{1\theta}} \vee B_{\mathbb{B}^{2\theta}} < \infty$ for all $\theta \in \mathbb{Z}$. This proves the necessity of the condition $B_{\mathbb{B}^{1\theta}} \vee B_{\mathbb{B}^{2\theta}} < \infty$ for $\overline{A}_{\mathbb{B}} < \infty$. $\square$

5. Orlicz form. In this section, the above results are specialized to Orlicz spaces. The idea goes back to [5]. A function Φ: $\mathbb{R}\to\mathbb{R}$ is called an N-function if it is non-negative, continuous, convex, even (i.e., $\Phi(-x)=\Phi(x)$) and satisfies the following conditions:

$$\Phi(x)=0 \ \text{ iff } \ x=0, \qquad \lim_{x\to 0}\Phi(x)/x=0, \qquad \lim_{x\to\infty}\Phi(x)/x=\infty.$$

In what follows, we assume the following growth condition (or Δ_2-condition) for Φ:

$$\sup_{x\gg 1}\Phi(2x)/\Phi(x)<\infty\Big(\iff \sup_{x\gg 1}x\Phi_-'(x)/\Phi(x)<\infty\Big),$$

where Φ_-' is the left derivative of Φ.

Corresponding to each N-function, we have a complementary N-function:

$$\Phi_c(y):=\sup\{x|y|-\Phi(x): x\geqslant 0\}, \qquad y\in\mathbb{R}.$$

Alternatively, let φ_c be the inverse function of Φ_-', then $\Phi_c(y)=\int_0^{|y|}\varphi_c$ (cf. [6]).

Given an N-function and a finite measure μ on $E:=[M,N]\subset\mathbb{Z}$, we define an Orlicz space as follows:

$$L^{\Phi}(\mu)=\Big\{f\,\big(E\to\mathbb{R}\big):\sum_{i\in E}\Phi(f_i)\mu_i<\infty\Big\}, \qquad \|f\|_{\Phi}=\sup_{g\in\mathscr{G}}\sum_{i\in E}|f_i|g_i\mu_i, \tag{5.1}$$

where $\mathscr{G}=\big\{g\geqslant 0:\sum_{i\in E}\Phi_c(g_i)\mu_i\leqslant 1\big\}$, which is the set of non-negative functions in the unit ball of $L^{\Phi_c}(\mu)$. Under Δ_2-condition, $(L^{\Phi}(\mu),\|\cdot\|_{\Phi},\mu)$ is a Banach space. Clearly, $L^{\Phi}(\mathrm{d}\mu)\ni 1$ and is ideal.

Having these preparations in mind, it is rather simple to state and prove our first result in this section (cf. [1; Proof of Theorem 3.1]).

Corollary 5.1. For every N-function Φ satisfying Δ_2-condition, the conclusions of Theorems 1.1 and 1.2 hold with $\mathbb{B}=L^{\Phi}(\mu)$.

The next result is a consequence of Theorem 4.1 and Proposition 3.2 with $\mathbb{B}=L^{\Phi}(\mu)$. We will use the notations introduced there. For simplicity, we write $c_2(A)=c_2[M,\theta)$ and $\pi(A)=\pi[M,\theta)$ when $A=[M,\theta)$.

Corollary 5.2. For every N-function Φ satisfying Δ_2-condition, the conclusions of Theorem 4.1 hold with $\mathbb{B}=L^{\Phi}(\mu)$. Additionally, if

$$\big[c_2[M,\theta)\pi[M,\theta)\big]\vee\big[c_2(\theta,N]\pi(\theta,N]\big]<\|1\|_{\mathbb{B}}^{-1},$$

then we have

$$\overline{A}_{\mathbb{B}}\geqslant\sup_{\theta\in(M,N)}\Big(1-\sqrt{\|1\|_{\mathbb{B}}}\Big[\big[c_2[M,\theta)\pi[M,\theta)\big]\vee\big[c_2(\theta,N]\pi(\theta,N]\big]\Big]^{1/2}\Big)^2\big(A_{\mathbb{B}^{1\theta}}\vee A_{\mathbb{B}^{2\theta}}\big).$$

The proof of Corollary 5.2 is almost the same as that of [1; Theorem 3.2]. We omit the details here.

6. Nash inequalities and Sobolev-type inequalities. As a typical application of the above general setup, this section studies the Orlicz spaces with N-functions $\Phi(x) = |x|^p/p\,(p > 1)$, and apply to the Nash (or Sobolev-type) inequalities on $[M, N]$, $-\infty \leqslant M \leqslant N \leqslant \infty$. Here are Nash inequalities on $\{0, 1, \cdots, N\}$:

$$\|f\|_2^{1+4/\nu} \leqslant AD(f)\|f\|_1^{4/\nu}, \qquad f_0 = 0$$
$$\|f - \pi(f)\|_2^{1+4/\nu} \leqslant \overline{A}D(f)\|f\|_1^{4/\nu},$$

where $\|\cdot\|_p$ denotes the usual $L^p(\mu)$-norm, $D(f)$ is the same as in (1.2), and $\nu > 0$. It is known, when $\nu > 2$, that these inequalities are, respectively, equivalent to the following Sobolev-type inequalities (cf. [7]—[10]):

$$\|f\|^2_{2\nu/(\nu-2)} \leqslant A_\nu D(f), \qquad f_0 = 0 \tag{6.1}$$
$$\|f - \pi(f)\|^2_{2\nu/(\nu-2)} \leqslant \overline{A}_\nu D(f). \tag{6.2}$$

The main purpose of this section is to estimate the optimal constants A_ν and $\overline{A}_\nu$.

Corollary 6.1. Let $E_1 = \{1, 2, \cdots, N\}$. Then the optimal constant A_ν in (6.1) satisfies

$$B_\nu \leqslant C_\nu \leqslant A_\nu \leqslant D_\nu \leqslant 4B_\nu,$$

where

$$\varphi_i = \sum_{j=1}^{i} \frac{1}{\mu_j a_j}, \quad 1 \leqslant i \leqslant N, \qquad B_\nu = \sup_{1\leqslant i\leqslant N} \bigg(\sum_{j=i}^{N} \mu_j\bigg)^{(\nu-2)/\nu} \varphi_i,$$
$$C_\nu = \sup_{1\leqslant i\leqslant N} \varphi_i^{-1}\|\varphi(i\wedge\cdot)^2\|_{\nu/(\nu-2)}, \qquad D_\nu = \sup_{1\leqslant i\leqslant N} \varphi_i^{-1/2}\|\sqrt{\varphi}\,\varphi(i\wedge\cdot)\|_{\nu/(\nu-2)}$$

and $\|\cdot\|_p$ is the $L^p(E_1, \mu)$-norm.

Proof. It is natural to use the Orlicz spaces $L^\Phi(\mu)$ with N-function $\Phi(x) = |x|^p/p$ and study the following inequalities:

$$\big\|f^2\big\|_\Phi \leqslant A_\Phi D(f), \qquad f_0 = 0 \tag{6.3}$$
$$\big\|(f - \pi(f))^2\big\|_\Phi \leqslant \overline{A}_\Phi D(f). \tag{6.4}$$

Now, $L^\Phi(E_1, \mu) = \big\{f : p^{-1}\sum_{i=1}^N |f_i|^p \mu_i < \infty\big\} = L^p(E_1, \mu)$. Since $\Phi_c(y) = |y|^q/q$, $1/p + 1/q = 1$, we have $\mathscr{G} = \{g \geqslant 0 : \|g\|_q \leqslant q^{1/q}\}$ and so

$$\|\cdot\|_\Phi = q^{1/q}\|\cdot\|_p. \tag{6.5}$$

Applying Corollary 5.1 to the function $\Phi(x) = |x|^p/p$, we get the estimates B_Φ, C_Φ and D_Φ of A_Φ, corresponding to the explicit bounds $B_{\mathbb{B}}, C_{\mathbb{B}}$ and $D_{\mathbb{B}}$ given in Theorems 1.1 and 1.2. Then, the required estimates follows from (6.5) by setting $p = \nu/(\nu - 2)$ (and then $q = \nu/2$). $\square$

We now turn to study $\overline{A}_\nu$. The idea is to use (6.4) and Theorem 3.1. The next result with different coefficients is proven in [11], based on the weighted Hardy inequality.

Theorem 6.2. Let $E = \{0, 1, \cdots, N\}$. Then the optimal constant $\overline{A}_\nu$ in (6.2) satisfies

$$\max\left\{\left(\frac{2}{\nu Z^{\nu/2-1}}\right)^{2/\nu}, \left[1-\left(\frac{Z-1}{Z}\right)^{1/2+1/\nu}\right]^2\right\}A_\nu \leqslant \overline{A}_\nu \leqslant 4A_\nu. \tag{6.6}$$

In particular,

$$\max\left\{\left(\frac{2}{\nu Z^{\nu/2-1}}\right)^{2/\nu}, \left[1-\left(\frac{Z-1}{Z}\right)^{1/2+1/\nu}\right]^2\right\}B_\nu \leqslant \overline{A}_\nu \leqslant 16B_\nu, \tag{6.7}$$

where $Z = \sum_{i=0}^N \mu_i$, B_ν is defined in Corollary 6.1, and so $\overline{A}_\nu < \infty$ iff $B_\nu < \infty$.

Proof. Clearly, it suffices to study the optimal constant $\overline{A}_\Phi$ in (6.4).

First, we compute the constants c_1 and c_2 used in Theorem 3.1. This can be easily done by using the Hölder inequality: $c_1 = Z^{-1/p}q^{-1/q}$, $c_2 = Z^{-1}(Z_1/q)^{1/q}$, where $Z_1 = \sum_{i=1}^N \mu_i = Z - 1$. Moreover, by (6.5), we have $\|1\|_\Phi = q^{1/q}Z^{1/p}$. Thus,

$$c_1\|1\|_\Phi = 1, \qquad c_2(1-\pi_0)\|1\|_\Phi = (Z_1/Z)^{1+1/q} < 1.$$

By Theorem 3.1, we obtain

$$\max\left\{\frac{1}{q^{1/q}Z^{1/p}}, \left(1-\left(\frac{Z_1}{Z}\right)^{1/2+1/(2q)}\right)^2\right\}A_\Phi \leqslant \overline{A}_\Phi \leqslant 4A_\Phi.$$

Combining this with (6.5) and $p = \nu/(\nu-2)$, we get (6.6). The second assertion of the theorem follows from Corollary 6.1. $\square$

To improve the estimates given in (6.6) and also for handling with the general case where $M \geqslant -\infty$, we adopt the idea explained at the end of Section 3. That is splitting $[M, N]$ into two parts $E_{1\theta} := \{\theta+1, \theta+2, \cdots, N\} =: (\theta, N]$ and $E_{2\theta} := \{M, M+1, \cdots, \theta-1\} =: [M, \theta)$, but leaving θ as a boundary of both $E_{1\theta}$ and $E_{2\theta}$. Denote by $\|\cdot\|_{1\theta,\Phi}$ and $\|\cdot\|_{2\theta,\Phi}$ the norms in $L^\Phi(E_{1\theta}, \mu^{1\theta})$ and $L^\Phi(E_{2\theta}, \mu^{2\theta})$, respectively. Actually, $\|f\|_{1\theta,\Phi} = \|fI_{E_{1\theta}}\|_\Phi$ and $\|f\|_{2\theta,\Phi} = \|fI_{E_{2\theta}}\|_\Phi$. The corresponding constants in inequality (1.2) are denoted by $A_\Phi^{1\theta}$ and $A_\Phi^{2\theta}$, respectively. Similarly, we have the $L^p(E_{k\theta}, \mu^{k\theta})$-norm $\|\cdot\|_{k\theta,p}$ $(k = 1, 2)$.

Define $\varphi^{k\theta}$, $C_\nu^{k\theta}$, $D_\nu^{k\theta}$ as in (4.3), but replacing the norm $\|\cdot\|_{\mathbb{B}^{k\theta}}$ by $\|\cdot\|_{k\theta,p}$ $(k = 1, 2)$. Next, define $B_\nu^{k\theta}$ $(k = 1, 2)$ as follows.

$$B_\nu^{1\theta} = \sup_{\theta+1\leqslant i\leqslant N}\left(\sum_{j=i}^N \mu_j\right)^{(\nu-2)/\nu}\varphi_i^{1\theta}, \qquad B_\nu^{2\theta} = \sup_{M\leqslant i\leqslant \theta-1}\left(\sum_{j=M}^i \mu_j\right)^{(\nu-2)/\nu}\varphi_i^{2\theta} \tag{6.8}$$

Theorem 6.3. Consider the general state space $\{M, M+1, \cdots, N-1, N\}$.

(1) Let $D(f)$ be defined by (4.2). Then inequality (6.2) holds iff $B_\nu^{1\theta} \vee B_\nu^{2\theta} < \infty$ for some (equivalently, for all) $\theta : M < \theta < N$. The optimal constant $\overline{A}_\nu$ satisfies

$$\frac{B_\nu^{1\theta} \wedge B_\nu^{2\theta}}{2} \leqslant \frac{C_\nu^{1\theta} \wedge C_\nu^{2\theta}}{2} \leqslant \frac{A_\nu^{1\theta} \wedge A_\nu^{2\theta}}{2} \leqslant \overline{A}_\nu,$$
$$\overline{A}_\nu \leqslant A_\nu^{1\theta} \vee A_\nu^{2\theta} \leqslant D_\nu^{1\theta} \vee D_\nu^{2\theta} \leqslant 4\big(B_\nu^{1\theta} \vee B_\nu^{2\theta}\big). \tag{6.9}$$

(2) Moreover,

$$\overline{A}_\nu \geqslant H_{\nu,\theta}\big(A_\nu^{1\theta} \vee A_\nu^{2\theta}\big) \geqslant H_{\nu,\theta}\big(C_\nu^{1\theta} \vee C_\nu^{2\theta}\big) \geqslant H_{\nu,\theta}\big(B_\nu^{1\theta} \vee B_\nu^{2\theta}\big), \tag{6.10}$$

where

$$H_{\nu,\theta} = \left[1 - \left(\frac{Z_{1\theta} \vee Z_{2\theta}}{Z}\right)^{1/2+1/\nu}\right]^2.$$

In particular, when θ is the median of μ, we have $H_{\nu,\theta} \geqslant \big[1-\big(1/2\big)^{1/2+1/\nu}\big]^2$.

Proof. Applying Corollary 5.2 to the inequality (6.4) and then using Corollary 6.1 and (6.5) with $p = \nu/(\nu-2)$, we obtain part (1).

Next, we compute the constants used in the second assertion of Corollary 5.2. By Hölder inequality and (6.5), we have

$$\pi(|f|I_A) = \frac{1}{Z}\int_A |f|\mathrm{d}\mu \leqslant \frac{1}{Z}\|fI_A\|_p\,\|I_A\|_q = \frac{1}{Z}\left(\frac{\mu(A)}{q}\right)^{1/q}\|fI_A\|_\Phi.$$

This gives us $c_2(A) = Z^{-1}(\mu(A)/q)^{1/q}$. Recall that $Z_{1\theta} = \sum_{\theta+1\leqslant i\leqslant N}\mu_j$ and $Z_{2\theta} = \sum_{M\leqslant i\leqslant\theta-1}\mu_j$. We have

$$\begin{aligned}
&c_2(\theta,N] = \frac{1}{Z}\left(\frac{Z_{1\theta}}{q}\right)^{1/q}, \qquad \pi(\theta,N] = \frac{Z_{1\theta}}{Z},\\
&c_2[M,\theta) = \frac{1}{Z}\left(\frac{Z_{2\theta}}{q}\right)^{1/q}, \qquad \pi[M,\theta) = \frac{Z_{2\theta}}{Z},\\
&\|1\|_\Phi = q^{1/q}\|1\|_p = q^{1/q}Z^{1/p}.
\end{aligned}$$

Hence

$$\|1\|_\Phi\big[\big(c_2[M,\theta)\pi[M,\theta)\big) \vee \big(c_2(\theta,N]\pi(\theta,N]\big)\big] = \left(\frac{Z_{1\theta}\vee Z_{2\theta}}{Z}\right)^{1+1/q} < 1.$$

Thus, by Corollary 5.2, we obtain

$$\overline{A}_\Phi \geqslant \left[1 - \left(\frac{Z_{1\theta}\vee Z_{2\theta}}{Z}\right)^{1/2+1/(2q)}\right]^2\big(A_\Phi^{1\theta} \vee A_\Phi^{2\theta}\big).$$

Combining this with (6.5) and setting $p = \nu/(\nu-2)$, we obtain (6.10). □

The results in this section are also meaningful for diffusions on the intervals, as did in [10] and [1; Example 3.4]. However, the first lower bound in (6.6) works only in the discrete situation.

7. Logarithmic Sobolev inequality. This section studies the Orlicz spaces with N-functions $\Phi(x) = |x|\log(1+|x|)$ and $\Psi(x) = x^2\log(1+x^2)$, and their application to the logarithmic Sobolev inequality on $[M,N]$, $-\infty \leqslant M \leqslant N \leqslant \infty$:

$$\sum_{i=M}^N f_i^2 \log\frac{f_i^2}{\pi(f^2)}\mu_i \leqslant A'' D(f), \tag{7.1}$$

where $D(f)$ is defined either by (4.2) for general $M \geqslant -\infty$ or as in (1.2) when $M = 0$. For the Orlicz space $\mathbb{B} = L^{\Phi}(\mu)$, when $M = 0$, we use A_Φ, B_Φ, C_Φ and D_Φ, respectively, to denote the constants $A_{\mathbb{B}}$, $B_{\mathbb{B}}$, $C_{\mathbb{B}}$ and $D_{\mathbb{B}} := D_{\mathbb{B}}(1)$ used in Theorems 1.1 and 1.2.

The next result follows from Corollary 5.1 plus some computation, on the basis of the formula $\|I_B\|_\Phi = \inf_{\alpha>0} \big(1+\mu(B)\Phi(\alpha)\big)/\alpha$ (cf. [1; Proof of Theorem 4.2]).

Corollary 7.1. Let $E_1 = \{1, 2, \cdots, N\}$ and $\Phi(x) = |x| \log(1 + |x|)$. Then Theorems 1.1 and 1.2 hold with $\mathbb{B} = L^{\Phi}(E_1, \mu)$. Moreover,

$$\begin{aligned} B_\Phi &= \sup_{1\leqslant i\leqslant N} \varphi_i M(\mu[i,N]), \\ M(x) &:= x\left[\frac{2}{1+\sqrt{1+4x}} + \log\left(1 + \frac{1+\sqrt{1+4x}}{2x}\right)\right]. \end{aligned} \tag{7.2}$$

In particular, the Poincaré-type inequality (1.2) in the Orlicz space $L^{\Phi}(E_1, \mu)$ holds iff

$$\sup_{1\leqslant i\leqslant N} \varphi_i \mu[i,N] \log \frac{1}{\mu[i,N]} < \infty. \tag{7.3}$$

We are now ready to study the logarithmic Sobolev inequality on $\{0, 1, \cdots, N\}$. For this, it is helpful to use an equivalent norm of $\|\cdot\|_\Phi$:

$$\|f\|_{(\Phi)} = \inf\left\{\alpha > 0 : \sum_i \Phi(f_i/\alpha)\mu_i \leqslant 1\right\}, \tag{7.4}$$

for which, we have

$$\|f\|_{(\Phi)} \leqslant \|f\|_\Phi \leqslant 2\|f\|_{(\Phi)} \tag{7.5}$$

(cf. [6; §3.3, Proposition 4]).

Corollary 7.2. Let $\Phi(x) = |x| \log(1+|x|)$ and $D(f)$ be defined as in (1.2). Then, on $E_1 = \{1, 2, \cdots, N\}$, the inequality

$$\left\|f^2\right\|_{(\Phi)} \leqslant A_{(\Phi)} D(f), \qquad f_0 = 0 \tag{7.6}$$

holds iff (7.3) is satisfied. The optimal constant $A_{(\Phi)}$ satisfies

$$B_\Phi/2 \leqslant C_\Phi/2 \leqslant A_\Phi/2 \leqslant A_{(\Phi)} \leqslant A_\Phi \leqslant D_\Phi \leqslant 4B_\Phi. \tag{7.7}$$

Proof. Simply use Corollary 7.1 and (7.8). □

In view of (7.7), the coefficients in [1; (4.9), (4.11)–(4.13) and Theorem 4.5] need a small correction.

The next result is an analogue of [11; Theorem 2.1] with different coefficients. The proof given in [11] is based on the weighted Hardy inequality and hence different from here.

Theorem 7.3. Consider the state space $E = \{0, 1, \cdots, N\}$. The logarithmic Sobolev constant A'' in (7.1) satisfies

$$\frac{2}{5}\max\left\{\frac{\sqrt{4Z+1}-1}{2},\ \left(1-\frac{Z_1\Psi^{-1}\big(Z_1^{-1}\big)}{Z\Psi^{-1}\big(Z^{-1}\big)}\right)^2\right\}B_\Phi \leqslant A'' \leqslant \frac{51\times 4}{5}B_\Phi, \quad (7.8)$$

where $Z = \sum_{i=0}^N \mu_i$, $Z_1 = \sum_{i=1}^N \mu_i = Z - 1$, Ψ^{-1} is the inverse function of Ψ, and B_Φ is given in (7.2). In particular, $A'' < \infty$ iff (7.3) holds.

Proof. (a) First, we computer the constants c_1 and c_2 used in Theorem 3.3 with $\mathbb{B} = L^\Psi(E, \mu)$ and the norm $\|\cdot\|_{(\Psi)}$. Because of the convexity of Φ, we have

$$\begin{aligned}
\|f\|_{(\Phi)} &= \inf\left\{\alpha > 0 : \sum_{i=0}^N \Phi\big(|f_i|/\alpha\big)\mu_i \leqslant 1\right\}\\
&= \inf\left\{\alpha > 0 : \sum_{i=0}^N \Phi\big(|f_i|/\alpha\big)\pi_i \leqslant \frac{1}{Z}\right\}\\
&\geqslant \inf\left\{\alpha > 0 : \Phi\left(\sum_{i=0}^N |f_i|\pi_i/\alpha\right) \leqslant \frac{1}{Z}\right\}\\
&= \inf\left\{\alpha > 0 : \sum_{i=0}^N |f_i|\pi_i/\alpha \leqslant \Phi^{-1}\big(Z^{-1}\big)\right\}\\
&= \pi(|f|)/\Phi^{-1}\big(Z^{-1}\big).
\end{aligned}$$

Hence, $\|f\|_{(\Phi)} \geqslant \pi(|f|)/\Phi^{-1}\big(Z^{-1}\big)$. Because $\|f\|_{(\Psi)}^2 = \big\|f^2\big\|_{(\Phi)}$, we obtain

$$\|f\|_{(\Psi)} \geqslant \sqrt{\pi\big(f^2\big)/\Phi^{-1}\big(Z^{-1}\big)} = \sqrt{\pi\big(f^2\big)}\Big/\Psi^{-1}\big(Z^{-1}\big) \geqslant |\pi(f)|/\Psi^{-1}\big(Z^{-1}\big).$$

This means that one can choose $c_1 = \Psi^{-1}\big(Z^{-1}\big)$.

Next, we compute c_2. Recall that $E_1 = \{1, 2, \cdots, N\}$. Again, by the convexity of Φ, we have

$$\begin{aligned}
\|fI_{E_1}\|_{(\Phi)} &= \inf\left\{\alpha > 0 : \sum_{i=1}^N \Phi\big(|f_i|/\alpha\big)\mu_i \leqslant 1\right\}\\
&= \inf\left\{\alpha > 0 : \frac{1}{Z_1}\sum_{i=1}^N \Phi\big(|f_i|/\alpha\big)\mu_i \leqslant \frac{1}{Z_1}\right\}\\
&\geqslant \inf\left\{\alpha > 0 : \Phi\left(\frac{1}{Z_1}\sum_{i=1}^N |f_i|\mu_i/\alpha\right) \leqslant \frac{1}{Z_1}\right\}\\
&= \inf\left\{\alpha > 0 : \frac{1}{Z_1}\sum_{i=1}^N |f_i|\mu_i/\alpha \leqslant \Phi^{-1}\big(Z_1^{-1}\big)\right\}\\
&= \frac{Z\pi(|f|I_{E_1})}{Z_1\Phi^{-1}\big(Z_1^{-1}\big)}.
\end{aligned}$$

Hence

$$\begin{aligned}\|fI_{E_1}\|^2_{(\Psi)} &= \big\|f^2 I_{E_1}\big\|_{(\Phi)}\\ &\geqslant \frac{Z}{Z_1\Phi^{-1}\big(Z_1^{-1}\big)}\pi\big(f^2 I_{E_1}\big)\\ &\geqslant \frac{Z}{(1-\pi_0)Z_1\Phi^{-1}\big(Z_1^{-1}\big)}\big[\pi(fI_{E_1})\big]^2.\end{aligned}$$

Thus, we can choose $c_2=\Psi^{-1}\big(Z_1^{-1}\big)Z_1/Z$ since $1-\pi_0=Z_1/Z$.

(b) On the other hand, we have $\|1\|_{(\Psi)}=\big(\Phi^{-1}\big(Z^{-1}\big)\big)^{-1/2}=1/\Psi^{-1}\big(Z^{-1}\big)$. Therefore,

$$c_2\|1\|_{(\Psi)}=\frac{Z_1}{Z}\cdot\frac{\Psi^{-1}\big(Z_1^{-1}\big)}{\Psi^{-1}\big(Z^{-1}\big)}<1,$$

since $\Psi^{-1}(x)/x$ is decreasing in x. By Theorem 3.3, we obtain

$$\bigg(1-\frac{Z_1\Psi^{-1}\big(Z_1^{-1}\big)}{Z\Psi^{-1}\big(Z^{-1}\big)}\bigg)^2 A'_{(\Psi)}\leqslant \overline{A}'_{(\Psi)}\leqslant 4\,A'_{(\Psi)},$$

Next, by [1; Lemma 4.1], we have

$$\frac{4}{5}\big\|f-\pi(f)\big\|^2_{(\Psi)}\leqslant \mathscr{L}(f)\leqslant \frac{51}{20}\big\|f-\pi(f)\big\|^2_{(\Psi)}, \tag{7.9}$$

where $\mathscr{L}(f)=\sup_{c\in\mathbb{R}}\operatorname{Ent}\big((f+c)^2\big)$ and $\operatorname{Ent}(f)=\sum_{i=M}^N f_i\log\big(f_i/\|f\|_{L^1(\pi)}\big)\mu_i$ for $f\geqslant 0$. Therefore, the logarithmic constant A'' satisfies

$$\frac{4}{5}\bigg(1-\frac{Z_1\Psi^{-1}\big(Z_1^{-1}\big)}{Z\Psi^{-1}\big(Z^{-1}\big)}\bigg)^2 A'_{(\Psi)}\leqslant A''\leqslant \frac{51}{5}\,A'_{(\Psi)}.$$

Because $\|f\|^2_{(\Psi)}=\big\|f^2\big\|_{(\Phi)}$, we have $A'_{(\Psi)}=A_{(\Phi)}$. Now, the assertions of the theorem, except the first lower bound, follow from Corollary 7.2.

(c) To get the first lower bound, we apply (7.9), (7.5), Theorems 3.1 and 1.1:

$$\begin{aligned}A'' &= \sup_{f\in\mathscr{D}_0}\frac{\mathscr{L}(f)}{D(f)}\\ &\geqslant \frac{4}{5}\sup_{f\in\mathscr{D}_0}\frac{\|f-\pi(f)\|^2_{(\Psi)}}{D(f)}\\ &= \frac{4}{5}\sup_{f\in\mathscr{D}_0}\frac{\|(f-\pi(f))^2\|_{(\Phi)}}{D(f)}\\ &\geqslant \frac{2}{5}\sup_{f\in\mathscr{D}_0}\frac{\|(f-\pi(f))^2\|_{\Phi}}{D(f)}\\ &= \frac{2}{5}\overline{A}_\Phi\geqslant \frac{2}{5}\|1\|_\Phi^{-1}A_\Phi\\ &\geqslant \frac{2}{5}\|1\|_\Phi^{-1}B_\Phi.\end{aligned}$$

The required estimate then follows from

$$\|1\|_\Phi = \inf_{\alpha>0}\{1+Z\Phi(\alpha)\}/\alpha = \big(\sqrt{4Z+1}+1\big)/(2Z). \quad \square$$

Finally, we study the general state space $\{M, M+1, \cdots, N-1, N\}$, in terms of the splitting technique. Recall the notations $\|\cdot\|_{k\theta,\Phi}$ and $A_\Phi^{k\theta}$ $(k=1,2)$ are defined in the last section. Define $\varphi^{k\theta}$, $C_\Phi^{k\theta}$, $D_\Phi^{k\theta}$ as in (4.3), but replacing the norm $\|\cdot\|_{\mathbb{B}^{k\theta}}$ by $\|\cdot\|_{k\theta,\Phi}$ $(k=1,2)$. Next, define $B_\Phi^{k\theta}$ $(k=1,2)$ as in (7.2):

$$B_\Phi^{1\theta} = \sup_{\theta+1\leqslant i\leqslant N} \varphi_i^{1\theta} M(\mu[i,N]), \qquad B_\Phi^{2\theta} = \sup_{M\leqslant i\leqslant \theta-1} \varphi_i^{2\theta} M(\mu[M,i]). \tag{7.10}$$

Theorem 7.4. Let $\Phi(x) = |x|\log(1+|x|)$. Consider the general state space $\{M, M+1, \cdots, N-1, N\}$.

(1) The inequality

$$\big\|(f-\pi(f))^2\big\|_{(\Phi)} \leqslant A_{(\Phi)} D(f), \tag{7.11}$$

holds iff $B_\Phi^{1\theta}\vee B_\Phi^{2\theta}<\infty$ for some (equivalently, for all) $\theta: M<\theta<N$. The optimal constant $\overline{A}_{(\Phi)}$ satisfies

$$\begin{aligned} &\frac{1}{4}\big(B_\Phi^{1\theta}\wedge B_\Phi^{2\theta}\big) \leqslant \frac{1}{4}\big(C_\Phi^{1\theta}\wedge C_\Phi^{2\theta}\big) \leqslant \frac{1}{2}\overline{A}_\Phi \leqslant \overline{A}_{(\Phi)} \leqslant \overline{A}_\Phi \\ &\overline{A}_\Phi \leqslant D_\Phi^{1\theta}\vee D_\Phi^{2\theta} \leqslant 4\big(B_\Phi^{1\theta}\vee B_\Phi^{2\theta}\big). \end{aligned} \tag{7.12}$$

(2) In particular, the optimal constant A'' in logarithmic Sobolev inequality (7.1) satisfies

$$\frac{1}{5}\big(B_\Phi^{1\theta}\wedge B_\Phi^{2\theta}\big) \leqslant \frac{1}{5}\big(C_\Phi^{1\theta}\wedge C_\Phi^{2\theta}\big) \leqslant A'' \leqslant \frac{51}{20}\big(D_\Phi^{1\theta}\vee D_\Phi^{2\theta}\big) \leqslant \frac{51}{5}\big(B_\Phi^{1\theta}\vee B_\Phi^{2\theta}\big). \tag{7.13}$$

Proof. Part (1) follows from Corollaries 5.2 and 7.2. Then, part (2) follows from (7.9). $\square$

Finally, we are going to prove a different lower bound, which is quite rough, but has the same form as the upper bound in (7.12). For this, we need Proposition 3.4. Applying (7.11) to an arbitrary function $f^{(1)}$ with $f^{(1)}|_{[M,\theta]}=0$, we get

$$\overline{A}_{(\Phi)} D(f^{(1)}) \geqslant \big\|\big(f^{(1)}-\pi(f^{(1)})\big)^2\big\|_{(\Phi)} = \big\|f^{(1)}-\pi(f^{(1)})\big\|_{(\Psi)}^2.$$

Next, applying part (2) of Proposition 3.4 to the space $\mathbb{B}=L^\Psi(\mathbb{Z},\mu)$ with norm $\|\cdot\|_{(\Psi)}$ and the set $A=E_{1\theta}$, we get $\|f^{(1)}\|_{(\Psi)} \leqslant K_1\|(f^{(1)}-\pi(f^{(1)})\|_{(\Psi)}$, where $K_1=\big[1-c_2(E_{1\theta})\|1\|_{(\Psi)}\big]^{-1}$. Hence, by (7.5),

$$K_1^2\overline{A}_{(\Phi)} D(f^{(1)}) \geqslant \|f^{(1)}\|_{(\Psi)}^2 = \big\|(f^{(1)})^2\big\|_{1\theta,\Phi} \geqslant \big\|(f^{(1)})^2\big\|_{1\theta,(\Phi)}.$$

Combining this with definition of $A_{(\Phi)}^{1\theta}$, it follows that $\overline{A}_{(\Phi)} \geqslant A_{(\Phi)}^{1\theta}/K_1^2$. Similarly, applying (7.11) to the function $f^{(2)}$ with $f^{(2)}|_{[\theta,N]} = 0$, we obtain $\overline{A}_{(\Phi)} \geqslant A_{(\Phi)}^{2\theta}/K_2^2$, where $K_2 = \big[1 - c_2(E_{2\theta})\|1\|_{(\Psi)}\big]^{-1}$. Collecting these facts together, it follows that

$$\overline{A}_{(\Phi)} \geqslant \max\big\{A_{(\Phi)}^{1\theta}/K_1^2,\, A_{(\Phi)}^{2\theta}/K_2^2\big\} \geqslant \big(K_1^{-2} \wedge K_2^{-2}\big)\big(A_{(\Phi)}^{1\theta} \vee A_{(\Phi)}^{2\theta}\big). \tag{7.14}$$

To estimate $K_1^{-2}\wedge K_2^{-2}$, recall that $Z_1 = Z_{1\theta} = \mu(E_{1\theta})$ and $Z_2 = Z_{2\theta} = \mu(E_{2\theta})$. By using the same technique as used in proof (b) of Theorem 7.3 (or referring to the proof of [1; Theorem 4.5]), we arrive at

$$K_1^{-2} \wedge K_2^{-2} \geqslant \bigg[\bigg(1 - \frac{Z_1\Psi^{-1}\big(Z_1^{-1}\big)}{Z\Psi^{-1}\big(Z^{-1}\big)}\bigg) \bigwedge \bigg(1 - \frac{Z_2\Psi^{-1}\big(Z_2^{-1}\big)}{Z\Psi^{-1}\big(Z^{-1}\big)}\bigg)\bigg]^2.$$

Now, let θ be the median of μ. That is, $Z_1, Z_2 \leqslant Z/2$. Again, since $\Psi^{-1}(x)/x$ is decreasing in x, we have

$$K_1^{-2} \wedge K_2^{-2} \geqslant \bigg[1 - \frac{\Psi^{-1}\big(2Z^{-1}\big)}{2\Psi^{-1}\big(Z^{-1}\big)}\bigg]^2 \geqslant \frac{(\sqrt{2}-1)^2}{2}.$$

The last constant was computed in [1; Proof of Theorem 4.5]. We have thus obtained the following result.

Theorem 7.5. Let θ be the median of μ and let B_Φ^k, C_Φ^k, $D_\Phi^k\,(k=1,2)$ be the same as in Theorem 7.4, ignoring the superscript θ. Then the optimal constant $\overline{A}_{(\Phi)}$ in (7.11) satisfies

$$\frac{\big(\sqrt{2}-1\big)^2}{4}\big(B_\Phi^1 \vee B_\Phi^2\big) \leqslant \frac{\big(\sqrt{2}-1\big)^2}{4}\big(C_\Phi^1 \vee C_\Phi^2\big) \leqslant \overline{A}_{(\Phi)} \leqslant D_\Phi^1 \vee D_\Phi^2 \leqslant 4\big(B_\Phi^1 \vee B_\Phi^2\big),$$

and the logarithmic Sobolev constant A'' satisfies

$$\begin{aligned}&\frac{\big(\sqrt{2}-1\big)^2}{5}\big(B_\Phi^1 \vee B_\Phi^2\big) \leqslant \frac{\big(\sqrt{2}-1\big)^2}{5}\big(C_\Phi^1 \vee C_\Phi^2\big) \leqslant A'',\\ &A'' \leqslant \frac{51}{20}\big(D_\Phi^1 \vee D_\Phi^2\big) \leqslant \frac{51}{5}\big(B_\Phi^1 \vee B_\Phi^2\big).\end{aligned}$$

References

[1]. Chen, M. F., *Variational formulas of Poincaré-type inequalities in Banach spaces of functions on the line*, Acta Math. Sin. Eng. Ser. **18**(3), 417–436 (2002).

[2]. Chen, M. F., *Analytic proof of dual variational formula for the first eigenvalue in dimension one*, Sci. Chin. (A), **42**(8), 805–815 (1999).

[3]. Chen, M. F., *Variational formulas and approximation theorems for the first eigenvalue in dimension one*, Sci. Chin. (A), **44**(4), 409–418 (2001).

[4]. Chen, M. F., *Explicit bounds of the first eigenvalue*, Sci. Chin. (A), **43**(10), 1051–1059 (2000).

[5]. Bobkov, S. G., Götze, F., *Exponential integrability and transportation cost related to logarithmic Sobolev inequalities*, J. Funct. Anal. **163**, 1–28 (1999).

[6]. Rao, M. M.. Ren, Z. D., *Theory of Orlicz Spaces*, Marcel Dekker, Inc. New York, 1991.

[7]. Varopoulos, N., *Hardy-Littlewood theory for semigroups*, J. Funct. Anal. **63**, 240–260 (1985).

[8]. Carlen, E. A., Kusuoka, S., Stroock, D. W., *Upper bounds for symmetric Markov transition functions*, Ann. Inst. Henri Poincaré **2**, 245–287 (1987).

[9]. Bakry, D., Coulhon, T., Ledoux, M., Saloff-Coste, L., *Sobolev inequalities in disguise*, Indiana Univ. Math. J. **44**(4), 1033–1074 (1995).

[10]. Mao, Y. H., *Nash inequalities for Markov processes in dimension one*, Acta Math. Sin. Eng. Ser. **18**(1), 147–156 (2002).

[11]. Mao, Y. H., *The logarithmic Sobolev inequalities for birth-death process and diffusion process on the line*, Chin. J. Appl. Prob. Statis. **18**(1), 94–100 (2002).

IMS Lecture Notes – Monograph Series, Volume 41, 81–96
Probability, Statistics and their Applications:
Papers in Honor of Rabi Bhattacharya, 2003

VARIATIONAL FORMULAS AND EXPLICIT BOUNDS OF POINCARÉ-TYPE INEQUALITIES FOR ONE-DIMENSIONAL PROCESSES

Mu-Fa Chen

(Beijing Normal University)
August 8, 2002

Dedicated to Professor
R. Bhattacharya on his 65th birthday

Abstract. This paper serves as a quick and elementary overview of the recent progress on a large class of Poincaré-type inequalities in dimension one. The explicit criteria for the inequalities, the variational formulas and explicit bounds of the corresponding constants in the inequalities are presented. As typical applications, the Nash inequalities and logarithmic Sobolev inequalities are examined.

1. Introduction.

The one-dimensional processes in this paper mean either one-dimensional diffusions or birth-death Markov processes. Let us begin with diffusions.

Let
$$L = a(x)\mathrm{d}^2/\mathrm{d}x^2 + b(x)\mathrm{d}/\mathrm{d}x$$
be an elliptic operator on an interval $(0, D)$ $(D \leqslant \infty)$ with Dirichlet boundary at 0 and Neumann boundary at D when $D < \infty$, where a and b are Borel measurable functions and a is positive everywhere. Set $C(x) = \int_0^x b/a$, here and in what follows, the Lebesgue measure $\mathrm{d}x$ is often omitted. Throughout the paper, assume that
$$Z := \int_0^D e^C/a < \infty. \tag{1.0}$$

2000 *Mathematics Subject Classification.* 49R50, 34L15, 26D10, 60J27.

Key words and phrases. Variational formula, Poincaré inequality, Nash inequality, logarithmic Sobolev inequality, Orlicz space, one-dimensional diffusion, birth-death process.

Research supported in part by NSFC (No. 10121101), RFDP and 973 Project.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

Hence, $\mathrm{d}\mu := a^{-1}e^{C}\mathrm{d}x$ is a finite measure, which is crucial in the paper. We are interested in the first Poincaré inequality

$$\|f\|^2 := \int_0^D f^2 \mathrm{d}\mu \leqslant A \int_0^D {f'}^2 e^C := AD(f), \quad f \in \mathbb{C}_d[0,D],\ f(0)=0, \tag{1.1}$$

where $\mathbb{C}_d$ is the set of all continuous functions, differentiable almost everywhere and having compact supports. When $D=\infty$, one should replace $[0,D]$ by $[0,D)$ but we will not mention again in what follows. Next, we are also interested in the second Poincaré inequality

$$\|f-\pi(f)\|^2 := \int_0^D (f-\pi(f))^2 \mathrm{d}\mu \leqslant \overline{A} D(f) \qquad f \in \mathbb{C}_d[0,D], \tag{1.2}$$

where $\pi(f) = \mu(f)/Z = \int f \mathrm{d}\mu/Z$. To save the notations, we use the same A (resp., $\overline{A}$) to denote the optimal constant in (1.1) (resp., (1.2)).

The aim of the study on these inequalities is looking for a criterion under which (1.1) (resp., (1.2)) holds, i.e., the optimal constant $A<\infty$ (resp., $\overline{A}<\infty$), and for the estimations of A (resp., $\overline{A}$). The reason why we are restricted in dimension one is looking for some explicit criteria and explicit estimates. Actually, we have dual variational formulas for the upper and lower bounds of these constants. Such explicit story does not exist in higher dimensional situation.

Next, replacing the L^2-norm on the right-hand sides of (1.1) and (1.2) with a general norm $\|\cdot\|_{\mathbb{B}}$ in a suitable Banach space (the details are delayed to §3), respectively, we obtain the following Poincaré-type inequalities

$$\big\|f^2\big\|_{\mathbb{B}} \leqslant A_{\mathbb{B}} D(f), \quad f \in \mathbb{C}_d[0,D],\ f(0)=0. \tag{1.3}$$

$$\big\|(f-\pi(f))^2\big\| \leqslant \overline{A}_{\mathbb{B}} D(f), \qquad f \in \mathbb{C}_d[0,D]. \tag{1.4}$$

For which, it is natural to study the same problems as above. The main purpose of this paper is to answer these problems. By using this general setup, we are able to handle with the following Nash inequalities[23]

$$\|f-\pi(f)\|^{2+4/\nu} \leqslant A_N D(f) \|f\|_1^{4/\nu} \tag{1.5}$$

in the case of $\nu>2$, and the logarithmic Sobolev inequality[18]:

$$\mathrm{Ent}\big(f^2\big) := \int_0^D f^2 \log \frac{f^2}{\pi(f^2)} \mathrm{d}\mu \leqslant A_{LS} D(f). \tag{1.6}$$

To see the importance of these inequalities, define the first Dirichlet eigenvalue λ_0 and the first Neumann eigenvalue λ_1, respectively, as follows.

$$\begin{aligned} \lambda_0 &= \inf\{D(f) : f \in C^1(0,D) \cap C[0,D],\ f(0)=0,\ \pi\big(f^2\big)=1\}, \\ \lambda_1 &= \inf\{D(f) : f \in C^1(0,D) \cap C[0,D],\ \pi(f)=0,\ \pi\big(f^2\big)=1\}. \end{aligned} \tag{1.7}$$

Then, it is clear that $\lambda_0 = 1/A$ and $\lambda_1 = 1/\overline{A}$. Furthermore, it is known that

$$\begin{aligned}
&\text{The second Poincaré inequality } \Longleftrightarrow \mathrm{Var}(P_t f) \leqslant \mathrm{Var}(f)\, e^{-2\lambda_1 t}.\\
&\text{Logarithmic Sobolev inequality } \Longleftrightarrow \mathrm{Ent}(P_t f) \leqslant \mathrm{Ent}(f)\, e^{-2t/A_{LS}},\\
&\text{Nash inequality } \Longleftrightarrow \mathrm{Var}(P_t f) \leqslant C\|f\|_1^2\, t^{-\nu},
\end{aligned} \tag{1.8}$$

where $\|f\|_r$ is the $L^r(\mu)$-norm (cf., [8], [13], [18] and references within). It is clear now that the convergence in the first line is also equivalent to the exponential ergodicity for any reversible Markov processes with density (cf. [10]), i.e., $\|P_t(x,\cdot) - \pi\|_{\mathrm{Var}} \leqslant C(x)e^{-\varepsilon t}$ for some constants $\varepsilon > 0$ and $C(x)$, where $P_t(x,\cdot)$ is the transition probability. The study on the existence of the equilibrium π and on the speed of convergence to equilibrium, by Bhattacharya and his cooperators, consists a fundamental contribution in the field. See for instance [2]–[6] and references within. The second line in (1.8) is correct for diffusions but incorrect in the discrete situation. In general, one has to replace "$\Longleftrightarrow$" by "$\Longrightarrow$".

Here are three examples which distinguish the different inequalities.

Examples: Diffusions on $[0, \infty)$

$\lambda_1 = \hat{\alpha} = \hat{\beta}$	**Log Sobolev** σ	**Nash** η
$a + b \geqslant$	$\dfrac{2(a \vee b - a \wedge b)}{\log a \vee b - \log a \wedge b}$	$(a+b)\left(\dfrac{a \wedge b}{a \vee b}\right)^{1+2/\nu}$

Table 1.1

Here in the first line, "LogS" means the logarithmic Sobolev inequality, "L^1-exp." means the L^1-exponential convergence which will not be discussed in this paper. "$\surd$" means always true and "$\times$" means never true, with respect to the parameters. Once known the criteria presented in this paper, it is easy to check Table 1.1 except the L^1-exponential convergence.

The remainder of the paper is organized as follows. In the next section, we review the criteria for (1.1) and (1.2), the dual variational formulas and explicit estimates of A and $\overline{A}$. Then, we extend partially these results to Banach spaces first for the Dirichlet case and then for the Neumann one. For a very general setup of Banach spaces, the resulting conclusions are still rather satisfactory. Next, we specify the results to Orlicz spaces and finally apply to the Nash inequalities and logarithmic Sobolev inequality.

Since each topic discussed subsequently has a long history and contains a large number of publications, it is impossible to collect in the present paper a complete list of references. We emphasize on recent progress and related references only. For the applications to the higher dimensional case and much more results, the readers are urged to refer to the original papers listed in References, and the informal book [13], in particular.

2. Ordinary Poincaré inequalities.

In this section, we introduce the criteria for (1.1) and (1.2), the dual variational formulas and explicit estimates of A and $\overline{A}$.

To state the main results, we need some notations. Write $x \wedge y = \min\{x, y\}$ and similarly, $x \vee y = \max\{x, y\}$. Define

$$\begin{aligned}
\mathscr{F} &= \big\{f \in C[0, D] \cap C^1(0, D) : f(0) = 0,\ f'|_{(0,D)} > 0\big\},\\
\widetilde{\mathscr{F}} &= \big\{f \in C[0, D] : f(0) = 0,\ \text{there exists } x_0 \in (0, D] \text{ so that}\\
&\qquad f = f(\cdot \wedge x_0),\ f \in C^1(0, x_0) \text{ and } f'|_{(0,x_0)} > 0\big\},\\
\mathscr{F}' &= \big\{f \in C[0, D] : f(0) = 0, f|_{(0,D)} > 0\big\},\\
\widetilde{\mathscr{F}}' &= \big\{f \in C[0, D] : f(0) = 0,\ \text{there exists } x_0 \in (0, D] \text{ so that}\\
&\qquad f = f(\cdot \wedge x_0) \text{ and } f|_{(0,x_0)} > 0\big\}.
\end{aligned} \tag{2.1}$$

Here the sets $\mathscr{F}$ and $\mathscr{F}'$ are essential, they are used, respectively, to define below the operators of single and double integrals, and are used for the upper bounds. The sets $\widetilde{\mathscr{F}}$ and $\widetilde{\mathscr{F}}'$ are less essential, simply the modifications of $\mathscr{F}$ and $\mathscr{F}'$, respectively, to avoid the integrability problem, and are used for the lower bounds. Define

$$\begin{aligned}
I(f)(x) &= \frac{e^{-C(x)}}{f'(x)} \int_x^D \big[f e^C/a\big](u)\mathrm{d}u, \qquad f \in \mathscr{F},\\
I\!I(f)(x) &= \frac{1}{f(x)} \int_0^x \mathrm{d}y e^{-C(y)} \int_y^D \big[f e^C/a\big](u)\mathrm{d}u, \qquad f \in \mathscr{F}'.
\end{aligned} \tag{2.2}$$

The next result is taken from [12; Theorems 1.1 and 1.2]. The word "dual" below means that the upper and lower bounds are interchangeable if one exchanges the orders of "sup" and "inf" with a slight modification of the set $\mathscr{F}$ (resp., $\mathscr{F}'$) of test functions.

Theorem 2.1. Let (1.0) hold. Define

$$\varphi(x) = \int_0^x e^{-C}, \qquad B = \sup_{x\in(0,D)} \varphi(x) \int_x^D \frac{e^C}{a}.$$

Then, we have the following assertions.

(1) Explicit criterion: $A < \infty$ iff $B < \infty$.

(2) Dual variational formulas:

$$\begin{aligned}
A &\leqslant \inf_{f\in\mathscr{F}'} \sup_{x\in(0,D)} I\!I(f)(x) = \inf_{f\in\mathscr{F}} \sup_{x\in(0,D)} I(f)(x),\\
A &\geqslant \sup_{f\in\widetilde{\mathscr{F}}'} \inf_{x\in(0,D)} I\!I(f)(x) = \sup_{f\in\widetilde{\mathscr{F}}} \inf_{x\in(0,D)} I(f)(x).
\end{aligned} \tag{2.3}$$

The two inequalities all become equalities whenever both a and b are continuous on $[0, D]$.

(3) Approximating procedure and explicit bounds:

(a) Define $f_1 = \sqrt{\varphi}$,

$$f_n = f_{n-1} I\!I(f_{n-1})$$

and

$$D_n = \sup_{x\in(0,D)} I\!I(f_n)(x).$$

Then D_n is decreasing in n and $A \leqslant D_n \leqslant 4B$ for all $n \geqslant 1$.

(b) Fix $x_0 \in (0, D)$. Define $f_1^{(x_0)} = \varphi(\cdot \wedge x_0)$,

$$f_n^{(x_0)} = f_{n-1}^{(x_0)}(\cdot \wedge x_0) I\!I\big(f_{n-1}^{(x_0)}(\cdot \wedge x_0)\big)$$

and

$$C_n = \sup_{x_0\in(0,D)} \inf_{x\in(0,D)} I\!I\big(f_n^{(x_0)}(\cdot \wedge x_0)\big)(x).$$

Then C_n is increasing in n and $A \geqslant C_n \geqslant B$ for all $n \geqslant 1$.

We mention that the explicit estimates "$B \leqslant A \leqslant 4B$" were obtained previously in the study on the weighted Hardy's inequality by [22].

We now turn to study $\overline{A}$, for which it is natural to assume that

$$\int_0^D e^{-C(s)} \mathrm{d}s \int_0^s a(u)^{-1} e^{C(u)} \mathrm{d}u = \infty. \tag{2.4}$$

Theorem 2.2. Let (1.0) and (2.4) hold and set $\bar{f} = f - \pi(f)$. Then, we have the following assertions.

(1) Explicit criterion: $\overline{A} < \infty$ iff $B < \infty$, where B is given by Theorem 1.1.

(2) Dual variational formulas:

$$\sup_{f\in\widetilde{\mathscr{F}}} \inf_{x\in(0,D)} I(\bar{f})(x) \leqslant \overline{A} \leqslant \inf_{f\in\mathscr{F}} \sup_{x\in(0,D)} I(\bar{f})(x). \tag{2.5}$$

The two inequalities all become equalities whenever both a and b are continuous on $[0, D]$.

(3) Approximating procedure and explicit bounds:

(a) Define $f_1 = \sqrt{\varphi}$, $f_n = \bar{f}_{n-1} I\!I(\bar{f}_{n-1})$ and $\overline{D}_n = \sup_{x\in(0,D)} I\!I(\bar{f}_n)\ (x)$. Then $\overline{A} \leqslant \overline{D}_n \leqslant 4B$ for all $n \geqslant 1$.

(b) Fix $x_0 \in (0, D)$. Define $f_1^{(x_0)} = \varphi(\cdot \wedge x_0)$,

$$f_n^{(x_0)} = \bar{f}_{n-1}^{(x_0)}(\cdot \wedge x_0) I\!I\big(\bar{f}_{n-1}^{(x_0)}(\cdot \wedge x_0)\big)$$

and

$$\overline{C}_n = \sup_{x_0\in(0,D)} \inf_{x\in(0,D)} I\!I\big(\bar{f}_n^{(x_0)}(\cdot \wedge x_0)\big)(x).$$

Then $\overline{A} \geqslant \overline{C}_n$ for all $n \geqslant 2$. By convention, $1/0 = \infty$.

Part (1) of the theorem is taken from [11; Theorem 3.7]. The upper bound in (2.5) is due to [16]. The other parts are taken from [12; Theorems 1.3 and 1.4].

Finally, we consider inequality (1.2) on a general interval $(p, q)\,(-\infty \leqslant p < q \leqslant \infty)$. When p (resp., q) is finite, at which the Neumann boundary condition is endowed. We adopt a splitting technique. The intuitive idea goes as follows: Since the eigenfunction corresponding to $\overline{A}$, if exists, must change signs, it should vanish somewhere in the present continuous situation, say θ for instance. Thus, it is natural to divide the interval (p, q) into two parts: (p, θ) and (θ, q). Then, one compares $\overline{A}$ with the optimal constants in the inequality (1.1), denoted by $A_{1\theta}$ and $A_{2\theta}$, respectively, on (θ, q) and (p, θ) having the common Dirichlet boundary at θ. Actually, we do not care about the existence of the vanishing point θ. Such θ is unknown, even if it exists. In practice, we regard θ as a reference point and then apply an optimization procedure with respect to θ. We now redefine $C(x) = \int_\theta^x b/a$. Again, since it is in the ergodic situation, we assume the following (non-explosive) conditions:

$$\begin{aligned}
&Z_{1\theta} := \int_\theta^q e^C/a < \infty, \qquad Z_{2\theta} := \int_p^\theta e^C/a < \infty.\\
&\int_p^\theta e^{-C(s)}\mathrm{d}s \int_s^\theta e^C/a = \infty \text{ if } p = -\infty \quad \text{ and }\\
&\int_\theta^q e^{-C(s)}\mathrm{d}s \int_\theta^s e^C/a = \infty \text{ if } q = \infty
\end{aligned} \tag{2.6}$$

for some (equivalently, all) $\theta \in (p, q)$. Corresponding to the intervals (θ, q) and (p, θ), respectively, we have constants $B_{1\theta}$ and $B_{2\theta}$, given by Theorem 1.1.

Theorem 2.3. Let (2.6) hold. Then, we have

(1) $\inf_{\theta\in(p,q)} \big(A_{1\theta} \wedge A_{2\theta}\big) \leqslant \overline{A} \leqslant \sup_{\theta\in(p,q)} \big(A_{1\theta} \vee A_{2\theta}\big)$.

(2) Let θ be the median of μ, then $\big(A_{1\theta} \vee A_{2\theta}\big)/2 \leqslant \overline{A} \leqslant A_{1\theta} \vee A_{2\theta}$.

In particular, $\overline{A} < \infty$ iff $B_{1\theta} \vee B_{2\theta} < \infty$.

Comparing the variational formulas (2.3) and (2.5) with the classical variational formulas given in (1.7), one sees that there are no common points. This explains why the new formulas (2.3) and (2.5) have not appeared before. The key here is the discover of the formulas rather than their proofs, which are usually simple due to the advantage of dimension one. As an illustration, here we present parts of the proofs.

Proof of the upper bound in (2.5).

Originally, the assertion was proved in [16] by using the coupling methods. Here we adopt the analytic proof given in [9].

Let $g \in C[0,D] \cap C^1(0,D)$, $\pi(g)=0$ and $\pi(g^2)=1$. Then, for every $f \in \mathscr{F}$ with $\pi(f) \geqslant 0$, we have

$$\begin{aligned}
1 &= \frac{1}{2}\int_0^D \pi(\mathrm{d}x)\pi(\mathrm{d}y)[g(y)-g(x)]^2 \\
&= \int_{\{x\leqslant y\}} \pi(\mathrm{d}x)\pi(\mathrm{d}y)\bigg(\int_x^y \frac{g'(u)\sqrt{f'(u)}}{\sqrt{f'(u)}}\mathrm{d}u\bigg)^2 \\
&\leqslant \int_{\{x\leqslant y\}} \pi(\mathrm{d}x)\pi(\mathrm{d}y)\int_x^y \frac{g'(u)^2}{f'(u)}\mathrm{d}u\int_x^y f'(\xi)\mathrm{d}\xi \\
&\qquad \text{(by Cauchy-Schwarz inequality)} \\
&= \int_{\{x\leqslant y\}} \pi(\mathrm{d}x)\pi(\mathrm{d}y)\int_x^y g'(u)^2 e^{C(u)}\frac{e^{-C(u)}}{f'(u)}\mathrm{d}u\big[f(y)-f(x)\big] \\
&= \int_0^D a(u)g'(u)^2\pi(\mathrm{d}u)\frac{Ze^{-C(u)}}{f'(u)}\int_0^u \pi(\mathrm{d}x)\int_u^D \pi(\mathrm{d}y)\,\big[f(y)-f(x)\big] \\
&\leqslant D(g)\sup_{u\in(0,D)}\frac{Ze^{-C(u)}}{f'(u)}\int_0^u \pi(\mathrm{d}x)\int_u^D \pi(\mathrm{d}y)\,\big[f(y)-f(x)\big] \\
&\leqslant D(g)\sup_{x\in(0,D)} I(f)(x) \quad \big(\text{since } \pi(f)\geqslant 0\big).
\end{aligned}$$

Thus, $D(g)^{-1} \leqslant \sup_{x\in(0,D)} I(\bar f)(x)$, and so

$$\overline{A} = \sup_{g:\ \pi(g)=0,\ \pi(g^2)=1} D(g)^{-1} \leqslant \sup_{x\in(0,D)} I(\bar f)(x).$$

This gives us the required assertion:

$$\overline{A} \leqslant \inf_{f\in\mathscr{F}}\sup_{x\in(0,D)} I(\bar f)(x).$$

The proof of the sign of the equality holds for continuous a and b needs more work, since it requires some more precise properties of the corresponding eigenfunctions. $\square$

Proof of the explicit upper bound "$A \leqslant 4B$".

As mentioned before, this result is due to [22]. Here we adopt the proof given in [11], as an illustration of the power of our variational formulas.

Recall that $B = \sup_{x\in(0,D)}\int_0^x e^{-C}\int_x^D e^C/a$. By using the integration by parts formula, it follows that

$$\begin{aligned}
\int_x^D \frac{\sqrt{\varphi}\,e^C}{a} &= -\int_x^D \sqrt{\varphi}\,\mathrm{d}\bigg(\int_\bullet^D \frac{e^C}{a}\bigg) \\
&\leqslant \frac{B}{\sqrt{\varphi(x)}} + \frac{B}{2}\int_x^D \frac{\varphi'}{\varphi^{3/2}} \leqslant \frac{2B}{\sqrt{\varphi(x)}}.
\end{aligned}$$

Hence

$$I\big(\sqrt{\varphi}\big)(x) = \frac{e^{-C(x)}}{\big(\sqrt{\varphi}\big)'(x)} \int_x^D \frac{\sqrt{\varphi}e^C}{a} \leqslant \frac{e^{-C(x)}\sqrt{\varphi(x)}}{(1/2)e^{-C(x)}} \cdot \frac{2B}{\sqrt{\varphi(x)}} = 4B$$

as required. □

3. Extension. Banach spaces.

Starting from this section, we introduce the recent results obtained in [14] and [15], but we will not point out time by time subsequently.

In this section, we study the Poincaré-type inequality (1.3). Clearly, the Banach spaces used here can not be completely arbitrary since we are dealing with a topic of hard mathematics. From now on, let $(\mathbb{B}, \|\cdot\|_{\mathbb{B}}, \mu)$ be a Banach space of functions $f:\ [0, D] \to \mathbb{R}$ satisfying the following conditions:

$$\begin{aligned} &(1)\quad 1 \in \mathbb{B};\\ &(2)\quad \mathbb{B} \text{ is ideal: If } h \in \mathbb{B} \text{ and } |f| \leqslant |h|, \text{ then } f \in \mathbb{B};\\ &(3)\quad \|f\|_{\mathbb{B}} = \sup_{g \in \mathscr{G}} \int_0^D |f| g \mathrm{d}\mu,\\ &(4)\quad \mathscr{G} \ni g_0 \text{ with } \inf g_0 > 0, \end{aligned} \tag{3.1}$$

where $\mathscr{G}$ is a fixed set, to be specified case by case later, of non-negative functions on $[0, D]$. The first two conditions mean that $\mathbb{B}$ is rich enough and the last one means that $\mathscr{G}$ is not trivial, it contains at least one strictly positive function. The third condition is essential in this paper, which means that the norm $\|\cdot\|_{\mathbb{B}}$ has a "dual" representation. A typical example of the Banach space is $\mathbb{B} = L^r(\mu)$, then $\mathscr{G} =$ the unit ball in $L_+^{r'}(\mu)$, $1/r + 1/r' = 1$.

The optimal constant $\overline{A}$ in (1.3) can be expressed as a variational formula as follows.

$$A_{\mathbb{B}} = \sup\bigg\{\frac{\|f^2\|_{\mathbb{B}}}{D(f)} : f \in \mathbb{C}_{\mathrm{d}}[0, D], f(0) = 0,\ 0 < D(f) < \infty\bigg\}. \tag{3.2}$$

Clearly, this formula is powerful mainly for the lower bounds of $\overline{A}$. However, the upper bounds are more useful in practice but much harder to handle. Fortunately, for which we have quite complete results.

Define $\varphi(x) = \int_0^x e^{-C}$ as before and let

$$\begin{aligned} B_{\mathbb{B}} &= \sup_{x \in (0,D)} \varphi(x) \big\| I_{(x,D)} \big\|_{\mathbb{B}},\\ C_{\mathbb{B}} &= \sup_{x \in (0,D)} \frac{\big\|\varphi(x \wedge \cdot)^2\big\|_{\mathbb{B}}}{\varphi(x)},\\ D_{\mathbb{B}} &= \sup_{x \in (0,D)} \frac{\big\|\sqrt{\varphi}\,\varphi(x \wedge \cdot)^2\big\|_{\mathbb{B}}}{\sqrt{\varphi(x)}}. \end{aligned} \tag{3.3}$$

Theorem 3.1. Let (1.0) and (3.1) hold. Then we have the following assertions.

(1) Explicit criterion: $A_{\mathbb{B}} < \infty$ iff $B_{\mathbb{B}} < \infty$.

(2) Variational formulas for the upper bounds:

$$\begin{aligned} A_{\mathbb{B}} &\leqslant \inf_{f \in \mathscr{F}'} \sup_{x \in (0,D)} f(x)^{-1} \big\| f\varphi(x \wedge \cdot) \big\|_{\mathbb{B}} \\ &\leqslant \inf_{f \in \mathscr{F}} \sup_{x \in (0,D)} \frac{e^{-C(x)}}{f'(x)} \| f I_{(x,D)} \|_{\mathbb{B}}. \end{aligned} \tag{3.4}$$

(3) Approximating procedure and explicit bounds: Let $B_{\mathbb{B}} < \infty$. Define $f_0 = \sqrt{\varphi}$, $f_n(x) = \|f_{n-1}\varphi(x \wedge \cdot)\|_{\mathbb{B}}$ and $D_{\mathbb{B}}(n) = \sup_{x\in(0,D)} f_n/f_{n-1}$ for $n \geqslant 1$. Then, $D_{\mathbb{B}}(n)$ is decreasing in n and

$$B_{\mathbb{B}} \leqslant C_{\mathbb{B}} \leqslant A_{\mathbb{B}} \leqslant D_{\mathbb{B}}(n) \leqslant D_{\mathbb{B}} \leqslant 4B_{\mathbb{B}} \tag{3.5}$$

for all $n \geqslant 1$.

We are now going to sketch the proof of the second variational formula in (3.4), from which the explicit upper bound $A_{\mathbb{B}} \leqslant 4B_{\mathbb{B}}$ follows immediately, as we did at the end of the last section. The explicit estimates "$B_{\mathbb{B}} \leqslant A_{\mathbb{B}} \leqslant 4B_{\mathbb{B}}$" were previously obtained in [7] in terms of the weighted Hardy's inequality [22]. The lower bounds follows easily from (3.2).

Sketch of the proof of the second variational formula in (3.4).

The starting point is the variational formula for A (cf. (2.3)):

$$A \leqslant \inf_{f\in\mathscr{F}} \sup_{x\in(0,D)} \frac{e^{-C(x)}}{f'(x)} \int_x^D \frac{fe^C}{a} = \inf_{f\in\mathscr{F}} \sup_{x\in(0,D)} \frac{e^{-C(x)}}{f'(x)} \int_x^D f \mathrm{d}\mu.$$

Fix $g > 0$ and introduce a transform as follows.

$$b \to b/g, \qquad a \to a/g > 0. \tag{3.6}$$

Under which, $C(x)$ is transformed into

$$C_g(x) = \int_0^x \frac{b/g}{a/g} = C(x).$$

This means that the function C is invariant of the transform, and so is the Dirichlet form $D(f)$. The left-hand side of (1.1) is changed into

$$\int_0^D f^2 g e^C/a = \int_0^D f^2 g \mathrm{d}\mu.$$

At the same time, the constant A is changed into

$$A_g \leqslant \inf_{f\in\mathscr{F}} \sup_{x\in(0,D)} \frac{e^{-C(x)}}{f'(x)} \int_x^D fg\mathrm{d}\mu.$$

Making supremum with respect to $g \in \mathscr{G}$, the left-hand side becomes

$$\sup_{g\in\mathscr{G}} \int_0^D f^2 g\mathrm{d}\mu = \|f^2\|_{\mathbb{B}}$$

and the constant becomes

$$\begin{aligned}
A_{\mathbb{B}} &= \sup_g A_g \\
&\leqslant \sup_g \inf_f \sup_x \frac{e^{-C(x)}}{f'(x)} \int_x^D fg\mathrm{d}\mu \\
&\leqslant \inf_f \sup_g \sup_x \\
&= \inf_f \sup_x \frac{e^{-C(x)}}{f'(x)} \sup_g \int_0^D f I_{(x,D)} g\mathrm{d}\mu. \\
&\quad = \inf_f \sup_x \frac{e^{-C(x)}}{f'(x)} \|f I_{(x,D)}\|_{\mathbb{B}}.
\end{aligned}$$

We are done! Of course, more details are required for completing the proof. For instance, one may use $g + 1/n$ instead of g to avoid the condition "$g > 0$" and then pass limit. $\square$

The lucky point in the proof is that "sup inf $\leqslant$ inf sup", which goes to the correct direction. However, we do not know at the moment how to generalize the dual variational formula for lower bounds, given in the second line of (2.3), to the general Banach spaces, since the same procedure goes to the opposite direction.

4. Neumann Case. Orlicz Spaces.

In the Neumann case, the boundary condition becomes $f'(0) = 0$, rather than $f(0) = 0$. Then $\lambda_0 = 0$ is trivial. Hence, we study λ_1 (called spectral gap of L), that is the inequality (1.2). We now consider its generalization (1.4). Naturally, one may play the same game as in the last section extending (2.5) to the Banach spaces. However, it does not work this time. Note that on the left-hand side of (1.4), the term $\pi(f)$ is not invariant under the transform (3.6). Moreover, since $\pi(\bar{f}) = 0$, it is easy to check that for each fixed $f \in \mathscr{F}$, $I(\bar{f})(x)$ is positive for all $x \in (0, D)$. But this property is no longer true when $\mathrm{d}\mu$ is replaced by $g\mathrm{d}\mu$. Our goal is to adopt the splitting technique explained in Section 2.

Let $\theta \in (p, q)$ be a reference point and let $A_{\mathbb{B}}^{k\theta}$, $B_{\mathbb{B}}^{k\theta}$, $C_{\mathbb{B}}^{k\theta}$, $D_{\mathbb{B}}^{k\theta}$ $(k = 1, 2)$ be the constants defined in (3.2) and (3.3) corresponding to the intervals (θ, q) and (p, θ), respectively. By Theorem 3.1, we have

$$B_{\mathbb{B}}^{k\theta} \leqslant C_{\mathbb{B}}^{k\theta} \leqslant A_{\mathbb{B}}^{k\theta} \leqslant D_{\mathbb{B}}^{k\theta} \leqslant 4B_{\mathbb{B}}^{k\theta}, \qquad k = 1, 2.$$

Theorem 4.1. Let (2.6) and (3.1) hold. Then, we have the following assertions.

(1) Explicit criterion: $\overline{A}_{\mathbb{B}} < \infty$ iff $B_{\mathbb{B}}^{1\theta} \vee B_{\mathbb{B}}^{2\theta} < \infty$.

(2) Estimates:

$$\max\left\{\frac{1}{2}\big(A_{\mathbb{B}}^{1\theta} \wedge A_{\mathbb{B}}^{2\theta}\big),\ K_\theta\big(A_{\mathbb{B}}^{1\theta} \vee A_{\mathbb{B}}^{2\theta}\big)\right\} \leqslant \overline{A}_{\mathbb{B}} \leqslant A_{\mathbb{B}}^{1\theta} \vee A_{\mathbb{B}}^{2\theta},$$

where K_θ is a constant.

It is the position to consider briefly the discrete case, i.e., the birth-death process. Let b_i $(i \geqslant 0)$ be the birth rates and a_i $(i \geqslant 1)$ be the death rates of the process. Define

$$\mu_0 = 1, \quad \mu_n = \frac{b_0 \cdots b_{n-1}}{a_1 \cdots a_n}, \quad Z = \sum_{n=0}^{\infty} \mu_n, \quad \pi_n = \frac{\mu_n}{Z},\ n \geqslant 1.$$

Consider a Banach space $(\mathbb{B}, \|\cdot\|_{\mathbb{B}}, \mu)$ of functions $E := \{0, 1, 2, \cdots\} \to \mathbb{R}$ satisfying (3.1). Define

$$\varphi_i = \sum_{j=1}^{i} \frac{1}{\mu_j a_j},\ i \geqslant 1; \qquad B_{\mathbb{B}} = \sup_{i \geqslant 1} \varphi_i \big\| I_{\{i, i+1, \cdots\}} \big\|_{\mathbb{B}}.$$

Clearly, the inequalities (1.3) and (1.4) are meaningful with a slight modification.

Theorem 4.2. Consider birth-death processes with state space E. Assume that $Z < \infty$.

(1) Explicit criterion for (1.3): $A_{\mathbb{B}} < \infty$ iff $B_{\mathbb{B}} < \infty$.
(2) Explicit bounds for $A_{\mathbb{B}}$: $B_{\mathbb{B}} \leqslant A_{\mathbb{B}} \leqslant 4B_{\mathbb{B}}$.
(3) Explicit criterion for (1.4): Let the birth-death process be non-explosive:

$$\sum_{i=0}^{\infty} \frac{1}{\mu_i b_i} \sum_{j=0}^{i} \mu_j = \infty. \tag{4.1}$$

Then $\overline{A}_{\mathbb{B}} < \infty$ iff $B_{\mathbb{B}} < \infty$.
(4) Estimates for $\overline{A}_{\mathbb{B}}$: Let $E_1 = \{1, 2, \cdots\}$ and let c_1 and c_2 be two constants such that $|\pi(f)| \leqslant c_1 \|f\|_{\mathbb{B}}$ and $|\pi(f I_{E_1})| \leqslant c_2 \|f I_{E_1}\|_{\mathbb{B}}$ for all $f \in \mathbb{B}$. Then,

$$\begin{aligned} &\max\Big\{\|1\|_{\mathbb{B}}^{-1},\ \big(1 - \sqrt{c_2(1-\pi_0)\|1\|_{\mathbb{B}}}\,\big)^2\Big\} A_{\mathbb{B}} \\ &\quad \leqslant \overline{A}_{\mathbb{B}} \leqslant \big(1 + \sqrt{c_1 \|1\|_{\mathbb{B}}}\,\big)^2 A_{\mathbb{B}}. \end{aligned} \tag{4.2}$$

Similarly, one can handle the birth-death processes on $\mathbb{Z}$.

An interesting point here is that the first lower bound in (4.2) is meaningful only in the discrete situation.

Orlicz spaces. The results obtained so far can be specialized to Orlicz spaces. The idea also goes back to [7]. A function Φ: $\mathbb{R} \to \mathbb{R}$ is called an N-function if it is non-negative, continuous, convex, even (i.e., $\Phi(-x) = \Phi(x)$) and satisfies the following conditions:

$$\Phi(x) = 0 \text{ iff } x = 0, \qquad \lim_{x \to 0} \Phi(x)/x = 0, \qquad \lim_{x \to \infty} \Phi(x)/x = \infty.$$

In what follows, we assume the following growth condition (or Δ_2-condition) for Φ:

$$\sup_{x \gg 1} \Phi(2x)/\Phi(x) < \infty \ \Big(\Longleftrightarrow \sup_{x \gg 1} x\Phi'_-(x)/\Phi(x) < \infty\Big),$$

where Φ'_- is the left derivative of Φ. Corresponding to each N-function, we have a complementary N-function:

$$\Phi_c(y) := \sup\{x|y| - \Phi(x) : x \geqslant 0\}, \qquad y \in \mathbb{R}.$$

Alternatively, let φ_c be the inverse function of Φ'_-, then $\Phi_c(y) = \int_0^{|y|} \varphi_c$ (cf. [24]).

Given an N-function and a finite measure μ on $E := (p, q) \subset \mathbb{R}$, define an Orlicz space as follows:

$$L^\Phi(\mu) = \left\{f\,(E \to \mathbb{R}) : \int_E \Phi(f)\mathrm{d}\mu < \infty\right\}, \qquad \|f\|_\Phi = \sup_{g\in\mathscr{G}} \int_E |f|g\mathrm{d}\mu, \qquad (4.3)$$

where

$$\mathscr{G} = \left\{g \geqslant 0 : \int_E \Phi_c(g)\mathrm{d}\mu \leqslant 1\right\},$$

which is the set of non-negative functions in the unit ball of $L^{\Phi_c}(\mu)$. Under Δ_2-condition, $(L^\Phi(\mu), \|\cdot\|_\Phi, \mu)$ is a Banach space. For this, the Δ_2-condition is indeed necessary. Clearly, $L^\Phi(\mu) \ni 1$ and is ideal. Obviously, $(L^\Phi(\mu), \|\cdot\|_\Phi, \mu)$ satisfies condition (3.1) and so we have the following result.

Corollary 4.3. For any N-function Φ satisfying the growth condition, if (1.0) (resp., (2.6)) holds, then Theorem 3.1 (resp., 4.1) is available for the Orlicz space $(L^\Phi(\mu), \|\cdot\|_\Phi, \mu)$.

5. Nash inequality and Sobolev-type inequality.

It is known that when $\nu > 2$, the Nash inequality (1.5):

$$\|f - \pi(f)\|^{2+4/\nu} \leqslant A_N D(f)\|f\|_1^{4/\nu}$$

is equivalent to the Sobolev-type inequality:

$$\|f - \pi(f)\|^2_{\nu/(\nu-2)} \leqslant A_S D(f),$$

where $\|\cdot\|_r$ is the $L^r(\mu)$-norm. Refer to [1], [8] and [26]. This leads to the use of the Orlicz space $L^\Phi(\mu)$ with $\Phi(x) = |x|^r/r$, $r = \nu/(\nu-2)$:

$$\left\|(f - \pi(f))^2\right\|_\Phi \leqslant \overline{A}_\nu D(f). \qquad (5.1)$$

The results in this section were obtained in [19], based on the weighted Hardy's inequalities.

Define $C(x) = \int_\theta^x b/a$, $\mu(m, n) = \int_m^n e^C/a$ and

$$\varphi^{1\theta}(x) = \int_\theta^x e^{-C} \qquad B_\nu^{1\theta} = \sup_{x>\theta} \varphi^{1\theta}(x)\,\mu(x, q)^{(\nu-2)/\nu},$$
$$\varphi^{2\theta}(x) = \int_x^\theta e^{-C} \qquad B_\nu^{2\theta} = \sup_{x<\theta} \varphi^{2\theta}(x)\,\mu(p, x)^{(\nu-2)/\nu}.$$

Here $B_\nu^{k\theta}$ $(k = 1, 2)$ is specified from $B_{\mathbb{B}}$ given in (3.3) with $\mathbb{B} = L^\Phi((\theta, q), \mu)$ or $\mathbb{B} = L^\Phi((p, \theta), \mu)$, since $\|\cdot\|_\Phi = (r')^{1/r'}\|\cdot\|_r$, $1/r + 1/r' = 1$.

Theorem 5.1. Let (2.6) hold and $\nu > 2$.

(1) Explicit criterion: Nash inequality (equivalently, (5.1)) holds on (p, q) iff $B_\nu^{1\theta} \vee B_\nu^{2\theta} < \infty$.

(2) Explicit bounds:

$$\begin{aligned}&\max\left\{\frac{1}{2}\big(B_\nu^{1\theta} \wedge B_\nu^{2\theta}\big),\ \left[1 - \left(\frac{Z_{1\theta} \vee Z_{2\theta}}{Z_{1\theta} + Z_{2\theta}}\right)^{1/2+1/\nu}\right]^2 \big(B_\nu^{1\theta} \vee B_\nu^{2\theta}\big)\right\}\\ &\quad\leqslant \overline{A}_\nu \leqslant 4\big(B_\nu^{1\theta} \vee B_\nu^{2\theta}\big).\end{aligned} \tag{5.2}$$

In particular, if θ is the median of μ, then

$$\big[1 - (1/2)^{1/2+1/\nu}\big]^2 \big(B_\nu^{1\theta} \vee B_\nu^{2\theta}\big) \leqslant \overline{A}_\nu \leqslant 4\big(B_\nu^{1\theta} \vee B_\nu^{2\theta}\big).$$

We now consider birth-death processes with state space $\{0, 1, 2, \cdots\}$. Define

$$\varphi_i = \sum_{j=1}^{i} \frac{1}{\mu_j a_j},\ i \geqslant 1; \qquad B_\nu = \sup_{i \geqslant 1} \varphi_i \left(\sum_{j=i}^{\infty} \mu_j\right)^{(\nu-2)/\nu}.$$

Theorem 5.2. For birth-death processes, let (4.1) hold and assume that $Z < \infty$. Then, we have

$$\max\left\{\left(\frac{2}{\nu Z^{\nu/2-1}}\right)^{2/\nu},\ \left[1 - \left(\frac{Z-1}{Z}\right)^{1/2+1/\nu}\right]^2\right\} B_\nu \leqslant \overline{A}_\nu \leqslant 16 B_\nu. \tag{5.3}$$

Hence, when $\nu > 2$, the Nash inequality holds iff $B_\nu < \infty$.

6. Logarithmic Sobolev inequality.

The starting point of the study is the following observation.

$$\frac{2}{5}\big\|(f - \pi(f))^2\big\|_\Phi \leqslant \mathscr{L}(f) \leqslant \frac{51}{20}\big\|(f - \pi(f))^2\big\|_\Phi, \tag{6.1}$$

where

$$\begin{aligned}\Phi(x) &= |x| \log(1 + |x|),\\ \mathscr{L}(f) &= \sup_{c \in \mathbb{R}} \mathrm{Ent}\big((f + c)^2\big),\\ \mathrm{Ent}(f) &= \int_{\mathbb{R}} f \log \frac{f}{\pi(f)} \mathrm{d}\mu, \quad f \geqslant 0.\end{aligned}$$

Refer to [7] and [17; page 247], which go back to [25]. A modification of the coefficients is made in [12]. The observation leads to the use of the Orlicz space $\mathbb{B} = L^\Phi(\mu)$ with $\Phi(x) = |x| \log(1 + |x|)$. The results in this section were obtained in [20], based again on the weighted Hardy's inequalities. Refer also to [21] for the related study.

Define

$$\begin{aligned}
&C(x)=\int_\theta^x \frac{b}{a}, \qquad \mu(m,n)=\int_m^n e^C/a;\\
&\varphi^{1\theta}(x)=\int_\theta^x e^{-C}, \qquad \varphi^{2\theta}(x)=\int_x^\theta e^{-C};\\
&M(x)=x\left[\frac{2}{1+\sqrt{1+4x}}+\log\left(1+\frac{1+\sqrt{1+4x}}{2x}\right)\right];\\
&B_\Phi^{1\theta}=\sup_{x\in(\theta,q)}\varphi^{1\theta}(x)M(\mu(\theta,x)), \qquad B_\Phi^{2\theta}=\sup_{x\in(p,\theta)}\varphi^{2\theta}(x)M(\mu(x,\theta)).
\end{aligned} \tag{6.2}$$

Again, here $B_\Phi^{k\theta}$ $(k=1,2)$ is specified from $B_{\mathbb{B}}$ given in (3.3).

Theorem 6.1. Let (2.6) hold.

(1) Explicit criterion: The logarithmic Sobolev inequality on $(p,q)\subset\mathbb{R}$ holds iff

$$\begin{aligned}
&\sup_{x\in(\theta,q)}\mu(x,q)\log\frac{1}{\mu(x,q)}\int_\theta^x e^{-C}<\infty \text{ and}\\
&\sup_{x\in(p,\theta)}\mu(p,x)\log\frac{1}{\mu(p,x)}\int_x^\theta e^{-C}<\infty
\end{aligned} \tag{6.3}$$

hold for some (equivalently, all) $\theta\in(p,q)$.

(2) Explicit bounds: Let $\bar\theta$ be the root of $B_\Phi^{1\theta}=B_\Phi^{2\theta}$, $\theta\in[p,q]$. Then, we have

$$\frac{1}{5}B_\Phi^{1\bar\theta}\leqslant A_{LS}\leqslant\frac{51}{5}B_\Phi^{1\bar\theta}. \tag{6.4}$$

By a translation if necessary, assume that $\theta=0$ is the median of μ. Then, we have

$$\frac{\big(\sqrt{2}-1\big)^2}{5}\big(B_\Phi^{1\theta}\vee B_\Phi^{2\theta}\big)\leqslant A_{LS}\leqslant\frac{51}{5}\big(B_\Phi^{1\theta}\vee B_\Phi^{2\theta}\big). \tag{6.5}$$

We now consider birth-death processes with state space $\{0,1,2,\cdots\}$. Define

$$\varphi_i=\sum_{j=1}^i\frac{1}{\mu_j a_j},\ i\geqslant 1; \qquad B_\Phi=\sup_{i\geqslant 1}\varphi_i M(\mu[i,\infty)),$$

where $\mu[i,\infty)=\sum_{j\geqslant i}\mu_j$ and $M(x)$ is defined in (6.2).

Theorem 6.2. For birth-death processes, let (4.1) hold and assume that $Z<\infty$. Then, we have

$$\begin{aligned}
&\frac{2}{5}\max\left\{\frac{\sqrt{4Z+1}-1}{2},\ \left(1-\frac{Z_1\Psi^{-1}\big(Z_1^{-1}\big)}{Z\Psi^{-1}\big(Z^{-1}\big)}\right)^2\right\}B_\Phi\\
&\qquad\leqslant A_{LS}\leqslant\frac{51}{5}\Big(1+\Psi^{-1}\big(Z^{-1}\big)\Big)^2 B_\Phi,
\end{aligned}$$

where $Z_1 = Z - 1$ and Ψ^{-1} is the inverse function of Ψ: $\Psi(x) = x^2 \log(1 + x^2)$. In particular, $A_{LS} < \infty$ iff

$$\sup_{i \geqslant 1} \varphi_i \, \mu[i, \infty) \log \frac{1}{\mu[i, \infty)} < \infty.$$

Acknowledgement. This paper is based on the talks given at "Stochastic analysis on large scale interacting systems", Shonan Village Center, Hayama, Japan (July 17-26, 2002) and "Stochastic analysis and statistical mechanics", Yukawa Institute, Kyoto University, Japan (July 29–30, 2002). The author is grateful for the kind invitation, financial support and the warm hospitality made by the organization committee: Profs. T. Funaki, H. Osada, N. Yosida, T. Kumagai and their colleagues and students.

References

[1] Bakry, D., Coulhon, T., Ledoux, M., Saloff-Coste, L., *Sobolev inequalities in disguise*, Indiana Univ. Math. J. **44**(4), 1033–1074 (1995)

[2] Bhattacharya, R. N., *Criteria for recurrence and existence of invariant measures for multidimensional diffusions*, Ann. Probab. **3**, 541–553 (1978). Correction, ibid. **8**, 1194–1195 (1980)

[3] Bhattacharya, R. N., *Multiscale diffusion processes with periodic coefficients and an application to solute transport in porous media*, Ann. Appl. Probab. **9**(4), 951-1020 (1999)

[4] Bhattacharya, R. N., Denker, M. and Goswami, A., *Speed of convergence to equilibrium and to normality for diffusions with multiple periodic scales*, Stoch. Proc. Appl. **80**, 55-86 (1999)

[5] Bhattacharya, R. N. and Götze, F. *Time-scales for Gaussian approximation and its break down under a hierarchy of periodic spatial heterogeneities*, Bernoulli **1**, 81-123 (1995)

[6] Bhattacharya, R. N. and Waymire, C. *Iterated random maps and some classes of Markov processes*, in "Handbook of Statistics", Vol. 19, pp.145-170, Eds. Shanbhag, D. N. and Rao, C. R., Elsevier Sci. B. V., 2001.

[7] Bobkov, S. G., Götze, F., *Exponential integrability and transportation cost related to logarithmic Sobolev inequalities*, J. Funct. Anal. **163**, 1–28 (1999)

[8] Carlen, E. A., Kusuoka, S., Stroock, D. W., *Upper bounds for symmetric Markov transition functions*, Ann. Inst. Henri Poincaré **2**, 245–287 (1987)

[9] Chen, M. F., *Analytic proof of dual variational formula for the first eigenvalue in dimension one*, Sci. Chin. (A) **42**(8), 805–815 (1999)

[10] Chen, M. F., *Equivalence of exponential ergodicity and L^2-exponential convergence for Markov chains*, Stoch. Proc. Appl. **87**, 281-297 (2000)

[11] Chen, M. F., *Explicit bounds of the first eigenvalue*, Sci. Chin. (A) **43**(10), 1051–1059 (2000)

[12] Chen, M. F., *Variational formulas and approximation theorems for the first eigenvalue in dimension one*, Sci. Chin. (A) **44**(4), 409–418 (2001)

[13] Chen, M. F., *Ergodic Convergence Rates of Markov Processes — Eigenvalues, Inequalities and Ergodic Theory*, Collection of papers, 1993–2001. http://math.bnu.edu.cn/~chenmf

[14] Chen, M. F., *Variational formulas of Poincaré-type inequalities in Banach spaces of functions on the line*, Acta Math. Sin. Eng. Ser. **18**(3), 417–436 (2002)

[15] Chen, M. F., *Variational formulas of Poincaré-type inequalities for birth-death processes*, preprint (2002), submitted to Acta Math. Sin. Eng. Ser.

[16] Chen, M. F. and Wang, F. Y., *Estimation of spectral gap for elliptic operators*, Trans. Amer. Math. Soc. **349**(3), 1239-1267 (1997)
[17] Deuschel, J. D. and Stroock, D. W., *Large Deviations*, Academic Press, New York, 1989
[18] Gross, L., *Logarithmic Sobolev inequalities*, Amer. J. Math. **97**, 1061–1083 (1976)
[19] Mao, Y. H., *Nash inequalities for Markov processes in dimension one*, Acta Math. Sin. Eng. Ser. **18**(1), 147–156 (2002)
[20] Mao, Y. H., *The logarithmic Sobolev inequalities for birth-death process and diffusion process on the line*, Chin. J. Appl. Prob. Statis. **18**(1), 94–100 (2002)
[21] Miclo, L., *An example of application of discrete Hardy's inequalities*, Markov Processes Relat. Fields **5**, 319–330, (1999)
[22] Muckenhoupt B., *Hardy's inequality with weights*, Studia Math. **XLIV**, 31–38 (1972)
[23] Nash, J., *Continuity of solutions of parabolic and elliptic equations*, Amer. J. Math. **80**, 931–954 (1958)
[24] Rao, M. M. and Ren, Z, D., *Theory of Orlicz Spaces*, Marcel Dekker, Inc. New York, 1991
[25] Rothaus, O. S., *Analytic inequalities, isoperimetric inequalities and logarithmic Sobolev inequalities*, J. Funct. Anal. **64**, 296–313 (1985)
[26] Varopoulos, N., *Hardy-Littlewood theory for semigroups*, J. Funct. Anal. **63**, 240–260 (1985)

Department of Mathematics, Beijing Normal University, Beijing 100875, The People's Republic of China. E-mail: mfchen@bnu.edu.cn
Home page: http://www.bnu.edu.cn/~chenmf/main_eng.htm

Proceedings of the Conference on Stochastic Analysis on Large Scale Interacting Systems, Shonan Village Center, Hayama, Japan (July 17–26, 2002)
Advanced Studies in Pure Mathematics, vol. 39, 89–114, 2004, Mathematical Society of Japan.

TEN EXPLICIT CRITERIA OF ONE-DIMENSIONAL PROCESSES

Mu-Fa Chen

(Beijing Normal University)
Received November 21, 2002. Revised December 20, 2002.

Abstract. The traditional ergodicity consists a crucial part in the theory of stochastic processes, plays a key role in practical applications. The ergodicity has much refined recently, due to the study on some inequalities, which are especially powerful in the infinite dimensional situation. The explicit criteria for various types of ergodicity for birth-death processes and one-dimensional diffusions are collected in Tables 8.1 and 8.2, respectively. In particular, an interesting story about how to obtain one of the criteria for birth-death processes is explained in details. Besides, a diagram for various types of ergodicity for general reversible Markov processes is presented.

The paper is organized as follows. First, we recall the study on an exponential convergence from different point of view in different subjects: probability theory, spectral theory and harmonic analysis. Then we show by examples the difficulties of the study and introduce the explicit criterion for the convergence, the variational formulas and explicit estimates for the convergence rates. Some comparison with the known results and an application are included. Next, we present ten (eleven) criteria for the two classes of processes, respectively, with some remarks. In particular, a diagram of various types of ergodicity for general reversible Markov processes is presented. For which, partial proofs are included in Appendix. Finally, we indicate a generalization to Banach spaces, this enables us to cover a large class of inequalities (equivalently, various types of ergodicity).

Let us begin with the paper by recalling the three traditional types of ergodicity.

1. Three traditional types of ergodicity. Let $Q = (q_{ij})$ be a regular Q-matrix on a countable set $E = \{i, j, k, \cdots\}$. That is, $q_{ij} \geqslant 0$ for all $i \neq j$,

---

2000 *Mathematics Subject Classification.* 60J27, 60J80.
Key words and phrases. Birth-death process, diffusion, ergodicity, principle eigenvalue.
Research supported in part by NSFC (No. 10121101), RFDP and 973 Project.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

$q_i := -q_{ii} = \sum_{j\ne i} q_{ij} < \infty$ for all $i \in E$ and Q determines uniquely a transition probability matrix $P_t = (p_{ij}(t))$ (which is also called a Q-process or a Markov chain). Denote by $\pi = (\pi_i)$ a stationary distribution of P_t: $\pi P_t = \pi$ for all $t \geqslant 0$. From now on, assume that the Q-matrix is irreducible and hence the stationary distribution π is unique. Then, the three types of ergodicity are defined respectively as follows.

$$\textit{Ordinary ergodicity:} \qquad \lim_{t\to\infty} |p_{ij}(t) - \pi_j| = 0 \tag{1.1}$$

$$\textit{Exponential ergodicity:} \quad \lim_{t\to\infty} e^{\hat\alpha t}|p_{ij}(t) - \pi_j| = 0 \tag{1.2}$$

$$\textit{Strong ergodicity:} \quad \lim_{t\to\infty} \sup_i |p_{ij}(t) - \pi_j| = 0$$

$$\Longleftrightarrow \lim_{t\to\infty} e^{\hat\beta t} \sup_i |p_{ij}(t) - \pi_j| = 0, \tag{1.3}$$

where $\hat\alpha$ and $\hat\beta$ are (the largest) positive constants and i, j varies over whole E. The equivalence in (1.3) is well known but one may refer to Proof (b) in the Appendix of this paper. These definitions are meaningful for general Markov processes once the pointwise convergence is replaced by the convergence in total variation norm. The three types of ergodicity were studied in a great deal during 1953–1981. Especially, it was proved that

$$\text{strong ergodicity} \Longrightarrow \text{exponential ergodicity} \Longrightarrow \text{ordinary ergodicity.}$$

Refer to Anderson (1991), Chen (1992, Chapter 4) and Meyn and Tweedie (1993) for details and related references. The study is quite complete in the sense that we have the following criteria which are described by the Q-matrix plus a test sequence (y_i) only, except the exponential ergodicity for which one requires an additional parameter λ.

Theorem 1.1 (Criteria). Let $H \ne \emptyset$ be an arbitrary but fixed finite subset of E. Then the following conclusions hold.

(1) The process P_t is ergodic iff the system of inequalities

$$\begin{cases} \sum_j q_{ij} y_j \leqslant -1, & i \notin H \\ \sum_{i\in H} \sum_{j\ne i} q_{ij} y_j < \infty \end{cases} \tag{1.4}$$

has a nonnegative finite solution (y_i).

(2) The process P_t is exponentially ergodic iff for some $\lambda > 0$ with $\lambda < q_i$ for all i, the system of inequalities

$$\begin{cases} \sum_j q_{ij} y_j \leqslant -\lambda y_i - 1, & i \notin H \\ \sum_{i\in H} \sum_{j\ne i} q_{ij} y_j < \infty \end{cases} \tag{1.5}$$

has a nonnegative finite solution (y_i).

(3) The process P_t is strongly ergodic iff the system (1.4) of inequalities has a bounded nonnegative solution (y_i).

The probabilistic meaning of the criteria reads respectively as follows:

$$\max_{i\in H}\mathbb{E}_i\sigma_H<\infty,\quad \max_{i\in H}\mathbb{E}_i e^{\lambda\sigma_H}<\infty\quad\text{and}\quad \sup_{i\in E}\mathbb{E}_i\sigma_H<\infty,$$

where $\sigma_H=\inf\{t\geqslant \text{the first jumping time}: X_t\in H\}$ and λ is the same as in (1.5). The criteria are not completely explicit since they depend on the test sequences (y_i) and in general it is often non-trivial to solve a system of infinite inequalities. Hence, one expects to find out some explicit criteria for some specific processes. Clearly, for this, the first candidate should be the birth-death process. Recall that for a birth-death process with state space $E=\mathbb{Z}_+=\{0,1,2,\cdots\}$, its Q-matrix has the form: $q_{i,i+1}=b_i>0$ for all $i\geqslant 0$, $q_{i,i-1}=a_i>0$ for all $i\geqslant 1$ and $q_{ij}=0$ for all other $i\neq j$. Along this line, it was proved by Tweedie (1981)(see also Anderson (1991) or Chen (1992)) that

$$S:=\sum_{n\geqslant 1}\mu_n\sum_{j\leqslant n-1}\frac{1}{\mu_j b_j}<\infty\Longrightarrow\text{Exponential ergodicity},\tag{1.6}$$

where $\mu_0=1$ and $\mu_n=b_0\cdots b_{n-1}/a_1\cdots a_n$ for all $n\geqslant 1$. Refer to Wang (1980), Yang (1986) or Hou et al (2000) for the probabilistic meaning of S. The condition is explicit since it depends only on the rates a_i and b_i. However, the condition is not necessary. A simple example is as follows. Let $a_i=b_i=i^\gamma$ $(i\geqslant 1)$ and $b_0=1$. Then the process is exponential ergodic iff $\gamma\geqslant 2$ (see Chen (1996)) but $S<\infty$ iff $\gamma>2$. Surprisingly, the condition is correct for strong ergodicity.

Theorem 1.2 [Zhang, Lin and Hou (2000)]. $S<\infty\Longleftrightarrow$ Strong ergodicity.

Refer to Hou et al (2000). With a different proof, the result is extended by Y. H. Zhang (2001) to the single-birth processes with state space $\mathbb{Z}_+$. Here, the term "single birth"means that $q_{i,i+1}>0$ for all $i\geqslant 0$ but $q_{ij}\geqslant 0$ can be arbitrary for $j<i$. Introducing this class of Q-processes is due to the following observation: If the first inequality in (1.4) is replaced by equality, then we get a recursion formula for (y_i) with one parameter only. Hence, there should exist an explicit criterion for the ergodicity (resp. uniqueness, recurrence and strong ergodicity). For (1.5), there is also a recursion formula but now two parameters are involved and so it is unclear whether there exists an explicit criterion or not for the exponential ergodicity.

Note that the criteria are not enough to estimate the convergence rate $\hat\alpha$ or $\hat\beta$ (cf. Chen (2000a)). It is the main reason why we have to come back to study the well-developed theory of Markov chains. For birth-death processes, the estimation of $\hat\alpha$ was studied by Doorn in a book (1981) and in a series of papers (1985, 1987, 1991). He proved, for instance, the following lower bound

$$\hat\alpha\geqslant\inf_{i\geqslant 0}\big\{a_{i+1}+b_i-\sqrt{a_ib_i}-\sqrt{a_{i+1}b_{i+1}}\big\},$$

which is exact when a_i and b_i are constant. The following formula for the lower bounds was implicated in his papers and rediscovered in a different point of view (in the study on spetral gap) by Chen (1996):

$$\hat\alpha=\sup_{v>0}\inf_{i\geqslant 0}\{a_{i+1}+b_i-a_i/v_{i-1}-b_{i+1}v_i\}.$$

Besides, the precise $\hat{\alpha}$ was determined by Doorn for four practical models. The main tool used in Doorn's study is the Karlin-Mcgregor's representation theorem, a specific spectral representation, involving heavy techniques. There is no explicit criterion for $\hat{\alpha} > 0$ ever appeared so far.

2. The first (non-trivial) eigenvalue (spectral gap). The birth-death processes have a nice property—symmetrizability: $\mu_i p_{ij}(t) = \mu_j p_{ji}(t)$ for all i, j and $t \geqslant 0$. Then, the matrix Q can be regarded as a self-adjoint operator on the real L^2-space $L^2(\mu)$ with norm $\|\cdot\|$. In other words, one can use the well-developed L^2-theory. For instance, one can study the L^2-exponential convergence given below. Assuming that $Z = \sum_i \mu_i < \infty$ and then setting $\pi_i = \mu_i/Z$. Then, the convergence means that

$$\|P_t f - \pi(f)\| \leqslant \|f - \pi(f)\| \leqslant e^{-\lambda_1 t} \tag{2.1}$$

for all $t \geqslant 0$, where $\pi(f) = \int f \mathrm{d}\pi$ and λ_1 is the first non-trivial eigenvalue (more precisely, the spectral gap) of $(-Q)$ (cf. Chen (1992, Chapter 9)).

The estimation of λ_1 for birth-death processes was studied by Sullivan (1984), Liggett (1989) and Landim, Sethuraman and Varadhan (1996) (see also Kipnis & Lamdin (1999)). It was used as a comparison tool to handle the convergence rate for some interacting particle systems, which are infinite-dimensional Markov processes. Here we recall three results as follows.

Theorem 2.1 [Sullivan (1984)]. Let c_1 and c_2 be two constants satisfying

$$c_1 \geqslant \sup_{i \geqslant 1} \frac{\sum_{j \geqslant i} \mu_j}{\mu_i}, \qquad c_2 \geqslant \sup_{i \geqslant 1} \frac{\mu_i}{\mu_i a_i}.$$

Then $\lambda_1 \geqslant 1/4c_1^2 c_2$.

Theorem 2.2 [Liggett (1989)]. Let c_1 and c_2 be two constants satisfying

$$c_1 \geqslant \sup_{i \geqslant 1} \frac{\sum_{j \geqslant i} \mu_j}{\mu_i a_i}, \qquad c_2 \geqslant \sup_{i \geqslant 1} \frac{\sum_{j \geqslant i} \mu_j a_j}{\mu_i a_i}.$$

Then $\lambda_1 \geqslant 1/4c_1 c_2$.

Theorem 2.3 [Liggett (1989)]. For bounded a_i and b_i, $\lambda_1 > 0$ iff (μ_i) has an exponential tail.

The reason we are mainly interested in the lower bounds is that on the one hand, they are more useful in practice and on the other hand, the upper bounds are usually easier to obtain from the following classical variational formula.

$$\lambda_1 = \inf\big\{D(f) : \mu(f) = 0, \mu\big(f^2\big) = 1\big\},$$

where

$$D(f) = \frac{1}{2}\sum_{i,j} \mu_i q_{ij} (f_j - f_i)^2, \qquad \mathscr{D}(D) = \{f \in L^2(\mu) : D(f) < \infty\}$$

and $\mu(f) = \int f \mathrm{d}\mu$.

Let us now leave Markov chains for a while and turn to diffusions.

3. One-dimensional diffusions. As a parallel of birth-death process, we now consider an elliptic operator $L = a(x)\mathrm{d}^2/\mathrm{d}x^2 + b(x)\mathrm{d}/\mathrm{d}x$ on the half line $[0, \infty)$ with $a(x) > 0$ everywhere. Again, we are interested in estimation of the principle eigenvalues, which consist of the typical, well-known Sturm-Liouville eigenvalue problem in the spectral theory. Refer to Egorov & Kondratiev (1996) for the present status of the study and references. Here, we mention two results, which are the most general ones we have ever known before.

Theorem 3.1. Let $b(x) \equiv 0$ (which corresponds to the birth-death process with $a_i = b_i$ for all $i \geqslant 1$) and set

$$\delta = \sup_{x>0} x \int_x^\infty a^{-1}.$$

Here we omit the integration variable when it is integrated with respect to the Lebesgue measure. Then, we have

(1) Kac & Krein (1958): $\delta^{-1} \geqslant \lambda_0 \geqslant (4\delta)^{-1}$, here λ_0 is the first eigenvalue corresponding to the Dirichlet boundary $f(0) = 0$.
(2) Kotani & Watanabe (1982): $\delta^{-1} \geqslant \lambda_1 \geqslant (4\delta)^{-1}$.

It is simple matter to rewrite the classical variational formula as (3.1) below. Similarly, we have (3.2) for λ_0.

Poincaré inequalities.

$$\lambda_1: \quad \|f - \pi(f)\|^2 \leqslant \lambda_1^{-1} D(f) \tag{3.1}$$

$$\lambda_0: \quad \|f\|^2 \leqslant \lambda_0^{-1} D(f), \quad f(0) = 0. \tag{3.2}$$

It is interesting that inequality (3.2) is a special but typical case of the weighted Hardy inequality discussed in the next section.

4. Weighted Hardy inequality. The classical Hardy inequality goes back to Hardy (1920):

$$\int_0^\infty \left(\frac{f}{x}\right)^p \leqslant \left(\frac{p}{p-1}\right)^p \int_0^\infty f'^p, \quad f(0) = 0, f' \geqslant 0,$$

where the optimal constant was determined by Landau (1926). After a long period of efforts by analysts, the inequality was finally extended to the following form, called weighted Hardy inequality (Muckenhoupt (1972))

$$\int_0^\infty f^2 \mathrm{d}\nu \leqslant A \int_0^\infty f'^2 \mathrm{d}\lambda, \quad f \in C^1, f(0) = 0, \tag{4.1}$$

where ν and λ be nonnegative Borel measures.

The Hardy-type inequalities play a very important role in the study of harmonic analysis and have been treated in many publications. Refer to the books: Opic & Kufner (1990), Dynkin (1990), Mazya (1985) and the survey article Davies (1999) for more details. We will come back this inequality soon.

We have finished the overview of the study on the exponential convergence (equivalently, the Poincaré inequality) in the different subjects. In order to have a more concrete feeling about the the difficulties of the topic, we now introduce some simple examples.

5. Difficulties. First, consider the birth-death processes with finite state space E.

When $E=\{0,1\}$, the Q-matrix becomes

$$Q=\begin{pmatrix} -b_0 & b_0 \\ a_1 & -a_1 \end{pmatrix}.$$

Then, it is trivial that $\lambda_1 = a_1 + b_0$. The result is nice since either a_1 or b_0 increases, so does λ_1. If we go one more step, $E=\{0,1,2\}$, then we have four parameters b_0, b_1 and a_1, a_2 and

$$\lambda_1 = 2^{-1}\big[a_1+a_2+b_0+b_1-\sqrt{(a_1-a_2+b_0-b_1)^2+4a_1b_1}\big].$$

Now, the role for λ_1 played by the parameters becomes ambiguous. When $E=\{0,1,2,3\}$, we have six parameters: $b_0, b_1, b_2, a_1, a_2, a_3$. Then

$$\lambda_1=\frac{D}{3}-\frac{C}{3\cdot 2^{1/3}}+\frac{2^{1/3}\left(3\,B-D^2\right)}{3\,C},$$

where the quantities D, B and C are not too complicated:

$$\begin{aligned}
&D=a_1+a_2+a_3+b_0+b_1+b_2,\\
&B=a_3\,b_0+a_2\,(a_3+b_0)+a_3\,b_1+b_0\,b_1+b_0\,b_2+b_1\,b_2+a_1\,(a_2+a_3+b_2)\,,\\
&C=\left(A+\sqrt{4(3\,B-D^2)^3+A^2}\right)^{1/3}.
\end{aligned}$$

However, in the last expression, another quantity is involved:

$$\begin{aligned}
A=&-2\,a_1^3-2\,a_2^3-2\,a_3^3+3\,a_3^2\,b_0+3\,a_3\,b_0^2-2\,b_0^3+3\,a_3^2\,b_1-12\,a_3\,b_0\,b_1+3\,b_0^2\,b_1\\
&+3\,a_3\,b_1^2+3\,b_0\,b_1^2-2\,b_1^3-6\,a_3^2\,b_2+6\,a_3\,b_0\,b_2+3\,b_0^2\,b_2+6\,a_3\,b_1\,b_2-12\,b_0\,b_1\,b_2\\
&+3\,b_1^2\,b_2-6\,a_3\,b_2^2+3\,b_0\,b_2^2+3\,b_1\,b_2^2-2\,b_2^3+3\,a_1^2\,(a_2+a_3-2\,b_0-2\,b_1+b_2)\\
&+3\,a_2^2\,[a_3+b_0-2\,(b_1+b_2)]\\
&+3\,a_2\,\big[a_3^2+b_0^2-2\,b_1^2-b_1\,b_2-2\,b_2^2-a_3(4\,b_0-2\,b_1+b_2)+2\,b_0(b_1+b_2)\big]\\
&+3\,a_1\,\big[a_2^2+a_3^2-2\,b_0^2-b_0\,b_1-2\,b_1^2-a_2(4\,a_3-2\,b_0+b_1-2\,b_2)\\
&+2\,b_0\,b_2+2\,b_1\,b_2+b_2^2+2\,a_3(b_0+b_1+b_2)\big].
\end{aligned}$$

Thus, the roles of the parameters are completely mazed! Of course, it is impossible to compute λ_1 explicitly when the size of the matrix is greater than five!

Next, we go to the estimation of λ_1. Consider the infinite state space $E=\{0,1,2,\cdots\}$. Denote by g and $D(g)$, respectively, the eigenfunction of λ_1 and the degree of g when g is polynomial. Three examples of the perturbation of λ_1 and $D(g)$ are listed in Table 1.1.

$b_i\,(i\geqslant 0)$	$a_i\,(i\geqslant 1)$	λ_1	$D(g)$
$i+c\,(c>0)$	$2i$	1	1
$i+1$	$2i+3$	2	2
$i+1$	$2i+\left(4+\sqrt{2}\right)$	3	3

Table 1.1 Three examples of the perturbation of λ_1 and $D(g)$

The first line is the well known linear model, for which $\lambda_1 = 1$, independent of the constant $c > 0$, and g is linear. Next, keeping the same birth rate, $b_i = i + 1$, changes the death rate a_i from $2i$ to $2i + 3$ (resp. $2i + 4 + \sqrt{2}$), which leads to the change of λ_1 from one to two (resp. three). More surprisingly, the eigenfunction g is changed from linear to quadratic (resp. triple). For the other values of a_i between $2i$, $2i + 3$ and $2i + 4 + \sqrt{2}$, λ_1 is unknown since g is non-polynomial. As seen from these examples, the first eigenvalue is very sensitive. Hence, in general, it is very hard to estimate λ_1.

Hopefully, I have presented enough examples to show the difficulties of the topic.

6. Results about λ_1, $\hat{\alpha}$ and λ_0. It is position to state our results. To do so, define

$$\mathscr{W} = \{w : w_i \uparrow\uparrow, \pi(w) \geqslant 0\}, \quad Z = \sum_i \mu_i, \qquad \delta = \sup_{i>0} \sum_{j \leqslant i-1} \frac{1}{\mu_j b_j} \sum_{j \geqslant i} \mu_j,$$

where $\uparrow\uparrow$ means strictly increasing. By suitable modification, we can define $\mathscr{W}'$ and explicit sequences δ_n and δ_n'. Refer to Chen (2001a) for details.

The next result provides a complete answer to the question proposed in Section 1.

Theorem 6.1. For birth-death processes, the following assertions hold.

(1) *Dual variational formulas*:

$$\lambda_1 = \sup_{w \in \mathscr{W}} \inf_{i \geqslant 0} \mu_i b_i (w_{i+1} - w_i) \Big/ \sum_{j \geqslant i+1} \mu_j w_j \qquad [\text{Chen}(1996)] \tag{6.1}$$

$$= \inf_{w \in \mathscr{W}'} \sup_{i \geqslant 0} \mu_i b_i (w_{i+1} - w_i) \Big/ \sum_{j \geqslant i+1} \mu_j w_j \qquad [\text{Chen}(2001a)] \tag{6.2}$$

(2) *Appoximating procedure and explicit bounds*:

$$Z\delta^{-1} \geqslant {\delta_n'}^{-1} \geqslant \lambda_1 \geqslant \delta_n^{-1} \geqslant (4\delta)^{-1} \text{ for all } n \qquad [\text{Chen}(2000b, 2001a)].$$

(3) *Explicit criterion*: $\lambda_1 > 0$ iff $\delta < \infty$ [Miclo (1999), Chen (2000b)].
(4) *Relation*: $\hat{\alpha} = \lambda_1$ [Chen(1991)].

In (6.1), only two notations are used: the sets $\mathscr{W}$ and $\mathscr{W}'$ of test functions (sequences). Clearly, for each test function, (6.1) gives us a lower bound of λ_1. This explains the meaning of "variational". Because of (6.1), it is now easy to obtain some lower estimates of λ_1, and in particular, one obtains all the lower bounds mentioned above. Next, by exchanging the orders of "sup" and "inf", we get (6.2) from (6.1), ignoring a slight modification of $\mathscr{W}$. In other words, (6.1) and (6.2) are dual of one to the other. For the explicit estimates "$\delta^{-1} \geqslant \lambda_0 \geqslant (4\delta)^{-1}$" and in particular for the criterion, one needs to find out a representative test function w among all $w \in \mathscr{W}$. This is certainly not obvious, because the test function w used in the formula is indeed a mimic of the eigenfunction (eigenvector) of λ_1, and in general, the eigenvalues and the corresponding eigenfunctions can be very sensitive, as we have seen from the above examples. Fortunately, there exists such a representative function with a simple form. We will illustrate the function in the context of diffusions in the second to the last paragraph of this section.

In parallel, for diffusions on $[0, \infty]$, define

$$C(x) = \int_0^x b/a, \qquad \delta = \sup_{x>0} \int_0^x e^{-C} \int_x^\infty e^C/a,$$
$$\mathscr{F} = \big\{ f \in C[0,\infty) \cap C^1(0,\infty) : f(0) = 0 \text{ and } f'|_{(0,\infty)} > 0 \big\}.$$

Theorem 6.2 [Chen (1999a, 2000b, 2001a)]. For diffusion on $[0, \infty)$, the following assertions hold.

(1) *Dual variational formulas*:

$$\lambda_0 \geqslant \sup_{f \in \mathscr{F}} \inf_{x>0} e^{C(x)} f'(x) \bigg/ \int_x^\infty f e^C/a \tag{6.3}$$

$$\lambda_0 \leqslant \inf_{f \in \mathscr{F}'} \sup_{x>0} e^{C(x)} f'(x) \bigg/ \int_x^\infty f e^C/a \tag{6.4}$$

Furthermore, the signs of the equality in (6.3) and (6.4) hold if both a and b are continuous on $[0, \infty)$.

(2) *Appoximating procedure and explicit bounds*: A decreasing sequence $\{\delta_n\}$ and an increasing sequence $\{\delta_n'\}$ are constructed explicitly such that

$$\delta^{-1} \geqslant {\delta_n'}^{-1} \geqslant \lambda_0 \geqslant \delta_n^{-1} \geqslant (4\delta)^{-1} \qquad \text{for all } n.$$

(3) *Explicit criterion*: λ_0 (resp. λ_1) > 0 iff $\delta < \infty$.

We mention that the above two results are also based on Chen and Wang (1997a).

To see the power of the dual variational formulas, let us return to the weighted Hardy's inequality.

Theorem 6.3 [Muckenhoupt (1972)]. The optimal constant A in the inequality

$$\int_0^\infty f^2 \mathrm{d}\nu \leqslant A \int_0^\infty {f'}^2 \mathrm{d}\lambda, \quad f \in C^1, f(0) = 0, \tag{6.5}$$

satisfies $B \leqslant A \leqslant 4B$, where

$$B = \sup_{x>0} \nu[x, \infty] \int_x^\infty (\mathrm{d}\lambda_{\text{abs}}/\mathrm{dLeb})^{-1}$$

and $\mathrm{d}\lambda_{\text{abs}}/\mathrm{dLeb}$ is the derivative of the absolutely continuous part of λ with respect to the Lebesgue measure.

By setting $\nu = \pi$ and $\lambda = e^C \mathrm{d}x$, it follows that the criterion in Theorem 6.2 is a consequence of the Muckenhoupt's Theorem. Along this line, the criteria in Theorems 6.1 and 6.2 for a typical class of the processes were also obtained by Bobkov and Götze (1999a, b), in which, the contribution of an earlier paper by Luo (1992) was noted.

We now point out that the explicit estimates "$\delta^{-1} \geqslant \lambda_0 \geqslant (4\delta)^{-1}$" in Theorems 6.2 or 6.3 follow from our variational formulas immediately. Here we consider the lower bound "$(4\delta)^{-1}$" only, the proof for the upper bound "δ^{-1}" is also easy, in terms of (6.4).

Recall that $\delta = \sup_{x>0} \int_0^x e^{-C} \int_x^\infty e^C/a$. Set $\varphi(x) = \int_0^x e^{-C}$. By using the integration by parts formula, it follows that

$$\begin{aligned}\int_x^\infty \frac{\sqrt{\varphi}\, e^C}{a} &= -\int_x^\infty \sqrt{\varphi}\, \mathrm{d}\Big(\int_\bullet^\infty \frac{e^C}{a}\Big)\\ &\leqslant \frac{\delta}{\sqrt{\varphi(x)}} + \frac{\delta}{2}\int_x^\infty \frac{\varphi'}{\varphi^{3/2}}\\ &\leqslant \frac{2\delta}{\sqrt{\varphi(x)}}.\end{aligned}$$

Hence

$$I\big(\sqrt{\varphi}\big)(x) = \frac{e^{-C(x)}}{\big(\sqrt{\varphi}\big)'(x)}\int_x^\infty \frac{\sqrt{\varphi} e^C}{a} \leqslant \frac{e^{-C(x)}\sqrt{\varphi(x)}}{(1/2)e^{-C(x)}} \cdot \frac{2\delta}{\sqrt{\varphi(x)}} = 4\delta.$$

This gives us the required bound by (6.3).

Theorem 6.2 can be immediately applied to the whole line or higher-dimensional situation. For instance, for Laplacian on compact Riemannian manifolds, it was proved by Chen & Wang (1997b) that

$$\lambda_1 \geqslant \sup_{f \in \mathscr{F}} \inf_{r \in (0,D)} I(f)(r)^{-1} =: \xi_1,$$

where $I(f)$ is the same as before but for some specific function $C(x)$. Thanks are given to the coupling technique which reduces the higher dimensional case to dimension one. We now have

$$\delta^{-1} \geqslant {\delta'_n}^{-1} \downarrow \geqslant \xi_1 \geqslant \delta_n^{-1} \uparrow \geqslant (4\delta)^{-1},$$

similar to Theorem 6.2. Refer to Chen (2000b, 2001a) for details. As we mentioned before, the use of the test functions is necessary for producing sharp estimates. Actually, the variational formula enables us to improve a number of best known estimates obtained previously by geometers, but none of them can be deduced from the estimates "$\delta^{-1} \geqslant \xi_1 \geqslant (4\delta)^{-1}$". Besides, the approximating procedure enables us to determine the optimal linear approximation of ξ_1 in K:

$$\xi_1 \geqslant \frac{\pi^2}{D^2} + \frac{K}{2},$$

where D is the diameter of the manifold and K is the lower bound of Ricci curvature (cf., Chen, Scacciatelli and Yao (2001)). We have thus shown the value of our dual variational formulas.

7. Three basic inequalities. Up to now, we have mainly studied the Poincaré inequality, i.e., (7.1) below. Naturally, one may study other inequalities, for instance, the logarithmic Sobolev inequality or the Nash inequality listed below.

$$\textit{Poincaré inequality}: \quad \|f-\pi(f)\|^2 \leqslant \lambda_1^{-1} D(f) \tag{7.1}$$

$$\textit{Logarithmic Sobolev inequality}: \quad \int f^2 \log(|f|/\|f\|) \mathrm{d}\pi \leqslant \sigma^{-1} D(f) \tag{7.2}$$

$$\textit{Nash inequality}: \ \|f-\pi(f)\|^{2+4/\nu} \leqslant \eta^{-1} D(f) \|f\|_1^{4/\nu} \ (\text{for some } \nu > 0). \tag{7.3}$$

Here, to save notation, σ (resp. η) denotes the largest constant so that (7.2) (resp. (7.3)) holds.

The importance of these inequalities is due to the fact that each inequality describes a type of ergodicity. First, (7.1) $\Longleftrightarrow$ (2.1). Next, the logarithmic Sobolev inequality implies (is indeed equivalent to, in the context of diffusions) the decay of the semigroup P_t to π exponentially in relative entropy with rate σ and the Nash inequality is equivalent to

$$\|P_t f - \pi(f)\| \leqslant C\|f\|_1 / t^{\nu/2}.$$

8. Criteria. Recently, the criteria for the last two inequalities as well as for the discrete spectrum (which means that there is no continuous spectrum and moreover, all eigenvalues have finite multiplicity) are obtained by Mao (2000, 2002a, b), based on the weighted Hardy's inequality. On the other hand, the main parts of Theorems 6.1 and 6.2 are extended to a general class of Banach spaces in Chen (2002a, d, e), which unify a large class inequalities and provide a unified criterion in particular. We can now summarize the results in Table 8.1. The table is arranged in such order that the property in the latter line is stranger than the former one, the only exception is that even though the strong ergodicity is often stronger than the logarithmic Sobolev inequality but they are not comparable in general (Chen (2002b)).

Birth-death processes

Transition intensity:

$$\begin{aligned} i \to i+1 \quad & \text{at rate} \quad b_i = q_{i,i+1} > 0 \\ \to i-1 \quad & \text{at rate} \quad a_i = q_{i,i-1} > 0. \end{aligned}$$

Define

$$\mu_0 = 1, \quad \mu_n = \frac{b_0 \cdots b_{n-1}}{a_1 \cdots a_n}, \ n \geqslant 1; \quad \mu[i,k] = \sum_{i \leqslant j \leqslant k} \mu_j.$$

Property	**Criterion**
Uniqueness	$\sum_{n\geqslant 0}\dfrac{1}{\mu_n b_n}\mu[0,n]=\infty \quad (*)$
Recurrence	$\sum_{n\geqslant 0}\dfrac{1}{\mu_n b_n}=\infty$
Ergodicity	$(*)$ & $\mu[0,\infty)<\infty$
Exponential ergodicity L^2-exponential convergence	$(*)$ & $\sup_{n\geqslant 1}\mu[n,\infty)\sum_{j\leqslant n-1}\dfrac{1}{\mu_j b_j}<\infty$
Discrete spectrum	$(*)$ & $\lim_{n\to\infty}\mu[n,\infty)\sum_{0\leqslant j\leqslant n-1}\dfrac{1}{\mu_j b_j}=0$
Logarithmic Sobolev inequality	$(*)$ & $\sup_{n\geqslant 1}\mu[n,\infty)\log[\mu[n,\infty)^{-1}]\sum_{j\leqslant n-1}\dfrac{1}{\mu_j b_j}<\infty$
Strong ergodicity L^1-exponential convergence	$(*)$ & $\sum_{n\geqslant 0}\dfrac{1}{\mu_n b_n}\mu[n+1,\infty)=\sum_{n\geqslant 1}\mu_n\sum_{j\leqslant n-1}\dfrac{1}{\mu_j b_j}<\infty$
Nash inequality	$(*)$ & $\sup_{n\geqslant 1}\mu[n,\infty)^{(\nu-2)/\nu}\sum_{j\leqslant n-1}\dfrac{1}{\mu_j b_j}<\infty$

Table 8.1 Ten criteria for birth-death processes

Here, "$(*)$ & $\cdots$" means that one requires the uniqueness condition in the first line plus the condition "$\cdots$".[1]

Diffusion processes on $[0,\infty)$ with reflecting boundary

Operator:

$$L=a(x)\frac{\mathrm{d}^2}{\mathrm{d}x^2}+b(x)\frac{\mathrm{d}}{\mathrm{d}x}.$$

Define

$$C(x)=\int_0^x b/a,\qquad \mu[x,y]=\int_x^y e^C/a.$$

The reason we have one more criterion here is due to the equivalence of the logarithmic Sobolev inequality and the exponential convergence in entropy. However, this is no longer true in the discrete case. In general, the logarithmic Sobolev inequality is stronger than the exponential convergence in entropy. A criterion for the exponential convergence in entropy for birth-death processes remains open (cf., Zhang and Mao (2000) and Mao and Zhang (2000)). The two equivalences in the tables come from the next diagram.

[1] In the original paper, for the Nash inequality there is an extra condition which is removed by the following paper:
Wang, J. (2009), Criteria for functional inequalities for ergodic birth-death processes, preprint [Acta Math. Sin. 2012, 28:2, 357–370]

Property	**Criterion**
Uniqueness	$\int_0^\infty \mu[0,x]e^{-C(x)} = \infty \quad (*)$
Recurrence	$\int_0^\infty e^{-C(x)} = \infty$
Ergodicity	$(*)$ & $\mu[0,\infty) < \infty$
Exponential ergodicity L^2-exponential convergence	$(*)$ & $\sup_{x>0} \mu[x,\infty) \int_0^x e^{-C} < \infty$
Discrete spectrum	$(*)$ & $\lim_{x\to\infty} \mu[x,\infty) \int_0^x e^{-C} = 0$
Logarithmic Sobolev inequality Exp. convergence in entropy	$(*)$ & $\sup_{x>0} \mu[x,\infty)\log[\mu[x,\infty)^{-1}]\int_0^x e^{-C} < \infty$
Strong ergodicity L^1-exponential convergence	$(*)$ & $\int_0^\infty \mu[x,\infty)e^{-C(x)} < \infty$?
Nash inequality	$(*)$ & $\sup_{x>0} \mu[x,\infty)^{(\nu-2)/\nu}\int_0^x e^{-C} < \infty$

Table 8.2 Eleven criteria for one-dimensional diffusions

9. New picture of ergodic theory.

Theorem 9.1. Let $(E,\mathscr{E})$ be a measurable space with countably generated $\mathscr{E}$. Then, for a Markov processes with state space $(E,\mathscr{E})$, reversible and having transition probability densities with respect to a probability measure π, we have the diagram shown in Figure 9.1.

$$
\begin{array}{ccccc}
 & & \text{Nash inequality} & & \\
 & \swarrow\!\!\swarrow & & \searrow\!\!\searrow & \\
\text{Logarithmic Sobolev inequality} & & & & L^1\text{-exponential convergence} \\
\Downarrow & & & & \| \\
\text{Exponential convergence in entropy} & & & & \pi\text{-a.s. Strong ergodicity} \\
\Downarrow & & & & \Downarrow \\
\text{Poincaré inequality} & & \Longleftrightarrow & & \pi\text{-a.s. Exponential ergodicity} \\
 & & \Downarrow & & \\
 & & L^2\text{-algebraic ergodicity} & & \\
 & & \Downarrow & & \\
 & & \text{Ordinary ergodicity} & &
\end{array}
$$

Figure 9.1 Diagram of nine types of ergodicity

Here are some remarks about Figure 9.1.

(1) The importance of the diagram is obvious. For instance, by using the estimates obtained from the study on Poincaré inequality, based on the advantage on the analytic approach —- the L^2-theory and the equivalence in the diagram, one can estimate exponentially ergodic convergence rates, for which, the known knowledge is still very limited. Actually,

these two convergence rates are often coincided (cf. the proofs given in Appendix). In particular, one obtains a criterion for the exponential ergodicity in dimension one, which has been opened for a long period. Conversely, one obtains immediately some criteria, which are indeed new, for Poincaré inequality to be held from the well-known criteria for the exponential ergodicity. Here, the L^1-exponential convergence means that $\|P_t f - \pi(f)\|_1 \leqslant C\|f - \pi(f)\|_1 e^{-\varepsilon t}$ for some constants $\varepsilon > 0$ and $C\,(\geqslant 1)$ and for all $t \geqslant 0$. Due to the structure of the L^1-space, which is only a Banach but not Hilbert space, there is still very limited known knowledge about the L^1-exponential convergence rate. Based on the probabilistic advantage and the identity in the diagram, from the study on the strong ergodicity, one learns a lot about the L^1-exponential convergence rate.

(2) The L^2-*algebraic ergodicity* means that $\mathrm{Var}(P_t f) \leqslant CV(f)t^{1-q}\,(t > 0)$ holds for some V having the properties: V is homogeneous of degree two (in the sense that $V(cf + d) = c^2 V(f)$ for any constants c and d) and $V(f) < \infty$ for all functions f with finite support (cf. Liggett (1991)). Refer to Chen and Wang (2000), Röckner and Wang (2001) for the study on the L^2-algebraic convergence.

(3) The diagram is complete in the following sense: each single-directed implication can not be replaced by double-directed one. Moreover, the L^1-exponential convergence (resp., the strong ergodicity) and the logarithmic Sobolev inequality (resp., the exponential convergence in entropy) are not comparable.

(4) The reversibility is used in both of the identity and the equivalence. Without the reversibility, the L^2-exponential convergence still implies π-a.s. exponentially ergodic convergence.

(5) An important fact is that the condition "having densities" is used only in the identity of L^1-exponential convergence and π-a.s. strong ergodicity, without this condition, L^1-exponential convergence still implies π-a.s. strong ergodicity, and so the diagram needs only a little change (However, the reversibility is still required here). Thus, it is a natural open problem to remove this "density's condition".

(6) Except the identity and the equivalence, all the implications in the diagram are suitable for general Markov processes, not necessarily reversible, even though the inequalities are mainly valuable in the reversible situation. Clearly, the diagram extends the ergodic theory of Markov processes.

The diagram was presented in Chen (1999c, 2002b), originally stated mainly for Markov chains. Recently, the identity of L^1-exponential convergence and the π-a.s. strong ergodicity is proven by Mao (2002c). A counter-example of diffusion was constructed by Wang (2001) to show that the strong ergodicity does not imply the exponential convergence in entropy. Partial proofs of the diagram are given in Appendix.

10. Go to Banach spaces. To conclude this paper, we indicate an idea to show the reason why we should go to the Banach spaces.

Theorem 10.1 [Varopoulos, N. (1985); Carlen, E. A., Kusuoka, S., Stroock, D. W. (1987); Bakry, D., Coulhon, T., Ledoux, M. and Saloff-

Coste, L. (1995)]. When $\nu > 2$, the Nash inequality

$$\|f - \pi(f)\|^{2+4/\nu} \leqslant C_1 D(f) \|f\|_1^{4/\nu}$$

is equivalent to the Sobolev-type inequality

$$\|f - \pi(f)\|_{\nu/(\nu-2)}^2 \leqslant C_2 D(f),$$

where $\|\cdot\|_p$ is the $L^p(\mu)$-norm.

In view of Theorem 10.1, it is natural to study the inequality

$$\big\|(f - \pi(f))^2\big\|_{\mathbb{B}} \leqslant A D(f)$$

for a general Banach space $(\mathbb{B}, \|\cdot\|_{\mathbb{B}}, \mu)$. It is interesting that even for the general setup, we still have quite satisfactory results. Refer to Bobkov and Götze (1999a, b) and Chen (2002a, d, e) for details.

Appendix: Partial proofs of Theorem 9.1. The detailed proofs and some necessary counterexamples were presented in Chen (1999c, 2002b) for reversible Markov processes, except the identity of the L^1-exponential convergence and π-a.s. strong ergodicity. Note that for discrete state spaces, one can rule out "a.s." used in the diagram. Here, we prove the new identity and introduce some more careful estimates for the general state spaces. The author would like to acknowledge Y. H. Mao for his nice ideas which are included in this appendix. The steps of the proofs are listed as follows.

(a) Nash inequality $\Longrightarrow$ L^1-exponential convergence and π-a.s. Strong ergodicity.
(b) L^1-exponential convergence $\Longleftrightarrow$ π-a.s. Strong ergodicity.
(c) Nash inequality $\Longrightarrow$ Logarithmic Sobolev inequality.
(d) L^2-exponential convergence $\Longrightarrow$ π-a.s. Exponential ergodicity.
(e) Exponential ergodicity $\Longrightarrow$ L^2-exponential convergence.

(a) Nash inequality $\Longrightarrow$ L^1-exponential convergence and π-a.s. Strong ergodicity [Chen (1999b)]. Denote by $\|\cdot\|_{p\to q}$ the operator's norm from $L^p(\pi)$ to $L^q(\pi)$. Note that

$$\begin{aligned}\text{Nash inequality} &\Longleftrightarrow \mathrm{Var}(P_t(f)) = \|P_t f - \pi(f)\|_2^2 \leqslant C^2 \|f\|_1^2 / t^{q-1} \ (q := \nu/2 + 1)\\ &\Longleftrightarrow \|(P_t - \pi) f\|_2 \leqslant C \|f\|_1 / t^{(q-1)/2}.\\ &\Longleftrightarrow \|P_t - \pi\|_{1\to 2} \leqslant C / t^{(q-1)/2}.\end{aligned}$$

Since $\|P_t - \pi\|_{1\to 1} \leqslant \|P_t - \pi\|_{1\to 2}$, we have

$$\text{Nash inequality} \Longrightarrow L^1\text{-algebraic convergence.}$$

Furthermore, because of the semigroup property, the convergence of $\|\cdot\|_{1\to 1}$ must be exponential, we indeed have

$$\text{Nash inequality} \Longrightarrow L^1\text{-exponential convergence.}$$

[Added in Proof. Actually, we have seen that there is a $t_0 > 0$ and $\gamma \in (0,1)$ such that $\|P_t - \pi\|_{1\to 1} \leqslant \gamma$. Given $t \geqslant 0$, express $t = mt_0 + h$ with $m \in \mathbb{N}_+$ and $h \in [0, t_0)$. Then for every f with $\pi(f) = 0$, we have $\pi(P_t f) = 0$ for all t, and furthermore

$$\|P_t f\|_1 = \|P_{mt_0+h} f\|_1 \leqslant \|P_h f\|_1 \gamma^m \leqslant \|f\|_1 \gamma^{(t-1)/t_0} = \gamma^{-1} e^{(t_0^{-1} \log \gamma)\, t}$$

for all t. This gives the required assertion since $\log \gamma < 0$.]

In the symmetric case: $P_t - \pi = (P_t - \pi)^*$, and so

$$\|P_{2t} - \pi\|_{1\to\infty} \leqslant \|P_t - \pi\|_{1\to 2} \|P_t - \pi\|_{2\to\infty} = \|P_t - \pi\|_{1\to 2}^2.$$

Hence, $\|P_t - \pi\|_{1\to\infty} \leqslant C/t^{q-1}$. Thus,

$$\begin{aligned}\operatorname{ess\,sup}_x \|P_t(x,\cdot) - \pi\|_{\mathrm{Var}} &= \operatorname{ess\,sup}_x \sup_{|f|\leqslant 1} |(P_t(x,\cdot) - \pi) f| \\ &\leqslant \operatorname{ess\,sup}_x \sup_{\|f\|_1 \leqslant 1} |(P_t(x,\cdot) - \pi) f| = \sup_{\|f\|_1\leqslant 1} \operatorname{ess\,sup}_x |(P_t(x,\cdot) - \pi) f| \\ &= \|P_t - \pi\|_{1\to\infty} \leqslant C/t^{q-1} \to 0, \qquad \text{as } \ t \to \infty.\end{aligned}$$

This gives us the π-a.s. strong ergodicity.

(b) L^1-exponential convergence $\Longleftrightarrow$ π-a.s. Strong ergodicity [Mao (2002c)]. Since $(L^1)^* = L^\infty \Longrightarrow \|P_t - \pi\|_{1\to 1} = \|P_t^* - \pi\|_{\infty\to\infty}$ and $P_t^*(x,\cdot) \ll \pi$, we have

$$\begin{aligned}\|P_t^* - \pi\|_{\infty\to\infty} &= \operatorname{ess\,sup}_x \sup_{\|f\|_\infty = 1} |(P_t^* - \pi) f(x)| = \operatorname{ess\,sup}_x \sup_{\sup|f|=1} |(P_t^* - \pi) f(x)| \\ &= \operatorname{ess\,sup}_x \|P_t^*(x,\cdot) - \pi\|_{\mathrm{Var}}.\end{aligned}$$

Hence, π-a.s. strong ergodicity is exactly the same as the L^1-exponential convergence. Without condition "$P_t^*(x,\cdot) \ll \pi$", the second equality becomes "$\geqslant$", and so we have in the general reversible case that

$$L^1\text{-exponential convergence} \Longrightarrow \pi\text{-a.s. Strong ergodicity.}$$

(c) Nash inequality $\Longrightarrow$ Logarithmic Sobolev inequality [Chen (1999b)]. Because $\|f\|_1 \leqslant \|f\|_p$ for all $p \geqslant 1$, we have

$$\|\cdot\|_{2\to 2} \leqslant \|\cdot\|_{1\to 2} \leqslant C/t^{(q-1)/2},$$

and so

$$\text{Nash inequality} \Longrightarrow \text{Poincaré inequality} \Longleftrightarrow \lambda_1 > 0.$$
$$\|P_t\|_{p\to 2} \leqslant \|P_t\|_{1\to 2} \leqslant \|P_t - \pi\|_{1\to 2} + \|\pi\|_{1\to 2} < \infty, \qquad p \in (1,2).$$

The assertion now follows from [Bakry (1992); Theorem 3.6 and Proposition 3.9].

The remainder of the Appendix is devoted to the proof of the assertion:

$$L^2\text{-exponential convergence} \iff \pi\text{-a.s. Exponential ergodicity.} \tag{$A1$}$$

Actually, this is done by Chen (2000a). Because, by assumption, the process is reversible and $P_t(x,\cdot) \ll \pi$. Set $p_t(x,y) = \frac{\mathrm{d}P_t(x,\cdot)}{\mathrm{d}\pi}(y)$. Then we have $p_t(x,y) = p_t(y,x)$, $\pi \times \pi$-a.s. (x,y). Hence

$$\int p_s(x,y)^2\pi(\mathrm{d}y) = \int p_s(x,y)p_s(y,x)\pi(\mathrm{d}y) = p_{2s}(x,x) < \infty \quad \text{(Carlen et al (1987)).} \tag{$A2$}$$

This means that $p_t(x,\cdot) \in L^2(\pi)$ for all $t > 0$ and π-a.s. $x \in E$. Thus, by [Chen (2000a); Theorem 1.2] and the remarks right after the theorem, ($A1$) holds.

The proof above is mainly based on the time-discrete analog result by Roberts and Rosenthal (1997). Here, we present a more direct proof of ($A2$) as follows.

(d) L^2-exponential convergence $\Longrightarrow$ π-a.s. Exponential ergodicity [Chen (1991, 1998, 2000a)]. Let $\mu \ll \pi$. Then

$$\begin{aligned}
\|\mu P_t - \pi\|_{\mathrm{Var}} &= \sup_{|f|\leqslant 1} |(\mu P_t - \pi)f| \\
&= \sup_{|f|\leqslant 1} \left|\pi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}P_t f - f\right)\right| \\
&= \sup_{|f|\leqslant 1} \left|\pi\left(fP_t^*\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right) - f\right)\right| \\
&= \sup_{|f|\leqslant 1} \left|\pi\left[f\left(P_t^*\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\right)\right)\right]\right| \\
&\leqslant \left\|P_t^*\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\right)\right\|_1 \\
&\leqslant \left\|\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\right\|_2 e^{-t\,\mathrm{gap}(L^*)} \\
&= \left\|\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\right\|_2 e^{-t\,\mathrm{gap}(L)}.
\end{aligned} \tag{$A3$}$$

We now consider two cases separately.

In the reversible case with $P_t(x,\cdot) \ll \pi$, by ($A2$), we have

$$\begin{aligned}
\|P_t(x,\cdot) - \pi\|_{\mathrm{Var}} &\leqslant \left\|P_{t-s}\left(\frac{\mathrm{d}P_s(x,\cdot)}{\mathrm{d}\pi} - 1\right)\right\|_1 \leqslant \|p_s(x,\cdot) - 1\|_2 e^{-(t-s)\,\mathrm{gap}(L)} \\
&= \left[\sqrt{p_{2s}(x,x) - 1}\, e^{s\,\mathrm{gap}(L)}\right] e^{-t\,\mathrm{gap}(L)}, \quad t \geqslant s.
\end{aligned} \tag{$A4$}$$

Therefore, there exists $C(x) < \infty$ such that

$$\|P_t(x,\cdot) - \pi\|_{\mathrm{Var}} \leqslant C(x)e^{-t\,\mathrm{gap}(L)}, \qquad t \geqslant 0, \quad \pi\text{-a.s. } (x). \tag{$A5$}$$

Denote by ε_1 be the largest ε such that $\|P_t(x,\cdot)-\pi\|_{\mathrm{Var}} \leqslant C(x)e^{-\varepsilon t}$ for all t. Then $\varepsilon_1 \geqslant \mathrm{gap}(L)=\lambda_1$.

In the φ-irreducible case, without using the reversibility and transition density, from $(A3)$, one can still derive π-a.s. exponential ergodicity (but may have different rates). Refer to Roberts and Tweedie (2001) for a proof in the time-discrete situation (the title of the quoted paper is confused, where the term "L^1-convergence" is used for the π-a.s. exponentially ergodic convergence, rather than the standard meaning of L^1-exponential convergence used in this paper. These two types of convergence are essentially different as shown in Theorem 9.1). In other words, the reversibility and the existence of the transition density are not essential in this implication.

(e) π-a.s. Exponential ergodicity $\Longrightarrow$ L^2-exponential convergence [Chen (2000a), Mao (2002c)]. In the time-discrete case, a similar assertion was proved by Roberts and Rosenthal(1997) and so can be extended to the time-continuous case by using the standard technique [cf., Chen (1992), §4.4]. The proof given below provides more precise estimates. Let the σ-algebra $\mathscr{E}$ be countably generated. By Numemelin and P. Tuominen (1982) or [Numemelin (1984); Theorem 6.14 (iii)], we have in the time-discrete case that

$$\begin{aligned}&\pi\text{-a.s. geometrically ergodic convergence}\\ &\quad\Longleftrightarrow \|\|P^n(\bullet,\cdot)-\pi\|_{\mathrm{Var}}\|_1 \text{ geometric convergence,}\end{aligned}$$

here and in what follows, the L^1-norm is taken with respect to the variable "$\bullet$". This implies in the time-continuous case that

$$\begin{aligned}&\pi\text{-a.s. exponentially ergodic convergence}\\ &\quad\Longleftrightarrow \|\|P_t(\bullet,\cdot)-\pi\|_{\mathrm{Var}}\|_1 \text{ exponential convergence.}\end{aligned}$$

Assume that $\|\|P_t(\bullet,\cdot)-\pi\|_{\mathrm{Var}}\|_1 \leqslant Ce^{-\varepsilon_2 t}$ with largest ε_2.

We now prove that $\|\|P_t(\bullet,\cdot)-\pi\|_{\mathrm{Var}}\|_1 \geqslant \|P_t-\pi\|_{\infty\to 1}$. Let $\|f\|_\infty=1$. Then

$$\begin{aligned}\|(P_t-\pi)f\|_1 &= \int \pi(\mathrm{d}x)\bigg|\int\Big[P_t(x,\mathrm{d}y)-\pi(\mathrm{d}y)\Big]f(y)\bigg|\\ &\leqslant \int \pi(\mathrm{d}x)\sup_{\|g\|_\infty\leqslant 1}\bigg|\int\Big[P_t(x,\mathrm{d}y)-\pi(\mathrm{d}y)\Big]g(y)\bigg|\\ &= \|\|P_t(\bullet,\cdot)-\pi\|_{\mathrm{Var}}\|_1 \qquad (\text{Need } P_t(x,\cdot)\ll\pi \text{ or reversibility!}).\end{aligned}$$

Next, we prove that $\|P_{2t}-\pi\|_{\infty\to 1}=\|P_t-\pi\|_{\infty\to 2}^2$ in the reversible case. We have

$$\begin{aligned}\|(P_t-\pi)f\|_2^2 &= ((P_t-\pi)f,(P_t-\pi)f)\\ &= (f,(P_t-\pi)^2 f)\\ &= (f,(P_{2t}-\pi)f)\\ &\leqslant \|f\|_\infty\|(P_{2t}-\pi)f\|_1\\ &\leqslant \|f\|_\infty^2\|P_{2t}-\pi\|_{\infty\to 1}.\end{aligned}$$

Hence $\|P_{2t}-\pi\|_{\infty\to 1}\geqslant \|P_t-\pi\|_{\infty\to 2}^2$. The inverse inequality is obvious by using the semigroup property and symmetry:

$$\|P_{2t}-\pi\|_{\infty\to 1}\leqslant \|P_t-\pi\|_{\infty\to 2}\|P_t-\pi\|_{2\to 1}=\|P_t-\pi\|_{\infty\to 2}^2.$$

We remark that in general case, without reversibility, we have $\|P_t-\pi\|_{\infty\to 1}\geqslant \|P_t-\pi\|_{\infty\to 2}^2/2$. Actually,

$$\begin{aligned}\|(P_t-\pi)f\|_2^2&\leqslant \int |(P_t-\pi)f|^2\mathrm{d}\pi\\&\leqslant 2\|f\|_\infty\int |(P_t-\pi)f|\mathrm{d}\pi\\&\leqslant 2\|f\|_\infty^2\|P_t-\pi\|_{\infty\to 1},\qquad f\in L^\infty(\pi).\end{aligned}$$

Finally, assume that the process is reversible. We prove that $\lambda_1=\text{gap}(L)\geqslant\varepsilon_2$. We have just proved that for every f with $\pi(f)=0$ and $\|f\|_2=1$, $\|P_tf\|_2^2\leqslant C\|f\|_\infty^2e^{-2\varepsilon_2 t}$. Following [Wang (2000; Lemma 2.2), or Röckner and Wang (2001)], by the spectral representation theorem, we have

$$\begin{aligned}\|P_tf\|_2^2&=\int_0^\infty e^{-2\lambda t}\mathrm{d}(E_\lambda f,f)\\&\geqslant\bigg[\int_0^\infty e^{-2\lambda s}\mathrm{d}(E_\lambda f,f)\bigg]^{t/s}\qquad\text{(by Jensen inequality)}\\&=\|P_sf\|_2^{2t/s},\quad t\geqslant s.\end{aligned}$$

Thus, $\|P_sf\|_2^2\leqslant\Big[C\|f\|_\infty^2\Big]^{s/t}e^{-2\varepsilon_2 s}$. Letting $t\to\infty$, we get

$$\|P_sf\|_2^2\leqslant e^{-2\varepsilon_2 s},\qquad \pi(f)=0,\ \|f\|_2=1,\ f\in L^\infty(\pi).$$

Since $L^\infty(\pi)$ is dense in $L^2(\pi)$, we have

$$\|P_sf\|_2^2\leqslant e^{-2\varepsilon_2 s},\quad s\geqslant 0,\quad \pi(f)=0,\ \|f\|_2=1.$$

Therefore, $\lambda_1\geqslant\varepsilon_2$. $\square$

Remark A1. Note that when $p_{2s}(\cdot,\cdot)\in L^{1/2}(\pi)$ (in particular, when $p_{2s}(x,x)$ is bounded in x) for some $s>0$, from $(A4)$, it follows that there exists a constant C such that $\|\|P_t(\bullet,\cdot)-\pi\|_{\text{Var}}\|_1\leqslant Ce^{-\lambda_1 t}$. Then, we have $\varepsilon_2\geqslant\lambda_1$. Combining this with (e), we indeed have $\lambda_1=\varepsilon_2$.

Remark A2. It is proved by Hwang et al (2002) that under mild condition, in the reversible case, $\lambda_1=\varepsilon_1$. Refer also to Wang (2002) for related estimates.

Final remark. The main body of this paper is an updated version of Chen (2001c), which was written at the beginning stage of the study on seeking explicit criteria. The resulting picture is now quite complete and so the most parts of the original paper has to be changed, except the first section. This paper also refines a part of Chen (2002c).

Acknowledgement. This paper is based on the talks given at "Stochastic analysis on large scale interacting systems", Shonan Village Center, Hayama, Japan (July 17–26, 2002), "Stochastic analysis and statistical mechanics", Yukawa Institute, Kyoto University, Japan (July 29–30, 2002), the "First Sino-German Conference on Stochastic Analysis—A Satellite Conference of ICM 2002", The Sino-German Center, Beijing (August 29–September 3, 2002), and "Stochastic analysis in infinite dimensional spaces", , RIMS, Kyoto University, Japan (November 6–8, 2002). The author is grateful for the kind invitation, financial support and the warm hospitality made by the organization committees: Profs. T. Funaki, H. Osada, N. Yosida, T. Kumagai and their colleagues and students; Profs. S. Albeverio, Z. M. Ma and M. Röckner; Profs. S. Aida, I. Shigekawa, Y. Takahashi and their colleagues.

The materials of the paper are included in a mini-course presented at the "Second Sino-French Colloquium in Probability and Applications", Wuhan, China (April 9–19, 2001), and included in the talks presented at Institute of Mathematics, Academy Sinica, Taipei (April 19–May 1, 2001), and in Department of Mathematics, Kyoto University, Kyoto, Japan (October 17, 2002–January 17, 2003). The author would like to acknowledge Prof. L. M. Wu; Profs. C. R. Hwang, T. S. Chiang, Y. S. Chow and S. J. Sheu; Profs. I. Shigekawa and Y. Takahashi for the financial support and the warm hospitality.

References

Anderson, W. J. (1991), *Continuous-Time Markov Chains*, Springer Series in Statistics, Springer, Berlin.

Bakry, D. (1992), *L'hypercontractivité et son utilisation en théorie des semigroupes*, LNM, **1581** Springer.

Bakry, D., Coulhon, T., Ledoux, M., Saloff-Coste, L. (1995), *Sobolev inequalities in disguise*, Indiana Univ. Math. J. **44**(4), 1033–1074.

Bobkov, S. G. and Götze, F. (1999a), *Discrete isoperimetric and Poincaré inequalities*, Prob. Th. Rel. Fields **114**, 245–277.

Bobkov, S. G. and Götze, F. (1999b), *Exponential integrability and transportation cost related to logarithmic Sobolev inequalities*, J. Funct. Anal. **163**, 1–28.

Carlen, E. A., Kusuoka, S., Stroock, D. W. (1987), *Upper bounds for symmetric Markov transition functions*, Ann. Inst. Henri Poincaré **2**, 245–287.

Chen, M. F. (1991), *Exponential L^2-convergence and L^2-spectral gap for Markov processes*, Acta Math. Sin. New Ser. **7**(1), 19–37.

Chen, M. F. (1992), *From Markov Chains to Non-Equilibrium Particle Systems*, World Scientific, Singapore.

Chen, M. F. (1996), *Estimation of spectral gap for Markov chains*, Acta Math. Sin. New Ser. **12**(4), 337-360.

Chen, M. F. (1998), *Estimate of exponential convergence rate in total variation by spectral gap*, Acta Math. Sin. Ser. (A) 41:1 (Chinese Ed.), 1–6; Acta Math. Sin. New Ser. **14**(1), 9–16.

Chen, M. F. (1999a), *Analytic proof of dual variational formula for the first eigenvalue in dimension one*, Sci. Sin. (A) **42**:8, 805–815.

Chen, M. F. (1999b), *Nash inequalities for general symmetric forms*, Acta Math. Sin. Eng. Ser. **15**:3, 353–370.

Chen, M. F. (1999c), *Eigenvalues, inequalities and ergodic theory* (II), Advances in Math. **28**:6, 481–505.

Chen, M. F. (2000a), *Equivalence of exponential ergodicity and L^2-exponential convergence for Markov chains*, Stoch. Proc. Appl. **87**, 281–297.

Chen, M. F. (2000b), *Explicit bounds of the first eigenvalue*, Sci. in China, Ser. A, **43**(10), 1051–1059.

Chen, M. F. (2001a), *Variational formulas and approximation theorems for the first eigenvalue in dimension one*, Sci. Chin. (A) **44**(4), 409–418.

Chen, M. F. (2001b), *Ergodic Convergence Rates of Markov Processes — Eigenvalues, Inequalities and Ergodic Theory* [*Collection of papers,* 1993—2001], http://www.bnu.edu.cn/~chenmf/main_eng.htm.

Chen, M. F. (2001c), *Explicit criteria for several types of ergodicity*, Chin. J. Appl. Prob. Stat. **17**(2), 1–8.

Chen, M. F. (2002a), *Variational formulas of Poincaré-type inequalities in Banach spaces of functions on the line*, Acta Math. Sin. Eng. Ser. **18**(3), 417–436.

Chen, M. F. (2002b), *A new story of ergodic theory*, in "Applied Probability", 25-34, eds. R. Chan et al., AMS/IP Studies in Advanced Mathematics, **26**, 2002.

Chen, M. F. (2002c), *Ergodic Convergence Rates of Markov Processes — Eigenvalues, Inequalities and Ergodic Theory*, in Proceedings of "ICM 2002", **III**, 41–52, Higher Education Press, Beijing.

Chen, M. F. (2002d), *Variational formulas of Poincaré-type inequalities for birth-death processes*, preprint, submitted to Acta Math. Sin. Eng. Ser.

Chen, M. F. (2002e), *Variational formulas of Poincaré-type inequalities for one-dimensional processes*, to appear in "IMS volume dedicated to Rabi Bhattacharya".

Chen, M. F., Scacciatelli, E. and Yao, L. (2002), *Linear approximation of the first eigenvalue on compact manifolds*, Sci. Sin. (A) **45**(4), 450–461.

Chen, M. F. and Wang, F. Y. (1997a), *Estimation of spectral gap for elliptic operators*, Trans. Amer. Math. Soc., **349**(3), 1239-1267.

Chen, M. F. and Wang, F. Y. (1997b), *General formula for lower bound of the first eigenvalue*, Sci. Sin. **40**(4), 384–394.

Chen, M. F. and Wang, Y. Z. (2000), *Algebraic Convergence of Markov Chains*, to appear in Ann. Appl. Probab.

Davies, E. B. (1999), *A review of Hardy inequality*, Operator Theory: Adv. & Appl. **110**, 55–67.

Van Doorn, E. (1981), *Stochastic Monotonicity and Queuing Applications of Birth-Death Processes*, Lecture Notes in Statistics **4**, Springer.

Van Doorn, E. (1985), *Conditions for exponential ergodicity and bounds for the decay parameter of a birth-death process*, Adv. Appl. Prob. **17**, 514–530.

Van Doorn, E. (1987), *Representations and bounds for zeros of orthogonal polynomials and eigenvalues of sign-symmetric tri-diagonal matrices*, J. Approx. Th. **51**, 254–266.

Van Doorn, E. (1991), *Quasi-stationary distributions and convergence to quasi-stationarity of birth-death processes*, Adv. Appl. Prob. **23**, 683–700.

Dynkin, E. M. (1990), *EMS*, Springer-Valerg (1991), Berlin.

Egorov, Y. and Kondratiev, V. (1996), *On Spectral Theory of Elliptic Operators*, Birkbäuser, Berlin.

Hwang, C. R., Hwang-Ma, S. Y. and Sheu, S. J. (2002), *Accelerating diffusions, preprint.*

Hardy, G. H. (1920), *Note on a theorem of Hilbert*, Math. Zeitschr. **6**, 314–317.

Hou, Z. T. et al (2000), *Birth-death Processes*, Hunan Sci. Press, Hunan.

Kac, I. S. and Krein, M. G. (1958), *Criteria for discreteness of the spectrum of a singular string*, Izv. Vyss. Učebn. Zaved. Mat. **2**, 136–153 (In Russian).

Kipnis, C. and Landim, C. (1999), *Scaling Limits of Interacting Particle Systems*, Springer-Verlag, Berlin.

Kotani, S. and Watanabe, S. (1982), *Krein's spectral theory of strings and generalized diffusion processes*, Lecture Notes in Math. **923**, 235–259.

Landau, E. (1926), *A note on a theorem concerning series of positive term*, J. London Math. Soc. **1**, 38–39.

Landim, C., Sethuraman, S. and Varadhan, S. R. S. (1996), *Spectral gap for zero range dynamics*, Ann. Prob. **24**, 1871–1902.

Liggett, T. M. (1989), *Exponential L^2 convergence of attractive reversible nearest particle systems*, Ann. Prob. **17**, 403–432.

Liggett, T. M. (1991), L^2 *Rates of convergence for attractive reversible nearest particle systems: the critical case*, Ann. Prob. **19**(3), 935–959.
Luo, J. H. (1992), *On discrete analog of Poincaré-type inequalities and density representation*, preprint (unpublished).
Mao, Y. H. (2000), *On empty essential spectrum for Markov processes in dimension one*, preprint.
Mao, Y. H. (2001), *Strong ergodicity for Markov processes by coupling methods*, preprint.
Mao, Y. H. (2002a), *The logarithmic Sobolev inequalities for birth-death process and diffusion process on the line*, Chin. J. Appl. Prob. Statis. **18**(1), 94–100.
Mao, Y. H. (2002b), *Nash inequalities for Markov processes in dimension one*, Acta. Math. Sin. Eng. Ser. **18**(1), 147–156.
Mao, Y. H. (2002c), *In preparation.*
Mao, Y. H. and Zhang, S. Y. (2000), *Comparison of some convergence rates for Markov process* (In Chinese), Acta. Math. Sin. **43**(6), 1019–1026.
Mazya, V. G. (1985), *Sobolev Spaces*, Springer-Valerg.
Meyn, S. P. and Tweedie, R. L. (1993), *Markov Chains and Stochastic Stability*, Springer, London.
Miclo, L. (1999a), *Relations entre isopérimétrie et trou spectral pour les chaînes de Markov finies*, Prob. Th. Rel. Fields **114**, 431–485.
Miclo, L. (1999b), *An example of application of discrete Hardy's inequalities*, Markov Processes Relat. Fields **5**, 319–330.
Muckenhoupt, B. (1972), *Hardy's inequality with weights*, Studia Math. **XLIV**, 31–38.
Nummelin, E. (1984), *General Irreducible Markov Chains and Non-Negative Operators*, Cambridge Univ. Press.
Nummelin, E. and Tuominen, P. (1982), *Geometric ergodicity of Harris recurrent chains with applications to renewal theory*, Stoch. Proc. Appl. **12**, 187-202.
Opic, B. and Kufner, A. (1990), *Hardy-type Inequalities*, Longman, New York.
Roberts, G. O. and Rosenthal, J. S. (1997), *Geometric ergodicity and hybrid Markov chains*, Electron. Comm. Probab. **2**, 13–25.
Roberts, G. O. and Tweedie, R. L. (2001), *Geometric* L^2 *and* L^1 *convergence are equivalent for reversible Markov chains*, J. Appl. Probab. **38**(A), 37–41.
Röckner, M. and Wang, F. Y. (2001), *Weak Poincaré inequalities and* L^2*-convergence rates of Markov semigroups*, J. Funct. Anal. **185**(2), 564–603.
Sullivan, W. G. (1984), *The* L^2 *spectral gap of certain positive recurrent Markov chains and jump processes*, Z. Wahrs. **67**, 387–398.
Tweedie, R. L. (1981), *Criteria for ergodicity, exponential ergodicity and strong ergodicity of Markov processes*, J. Appl. Prob. **18**, 122–130.
Varopoulos, N. (1985), *Hardy-Littlewood theory for semigroups*, J. Funct. Anal. **63**, 240–260.
Wang, F. Y. (2000), *Functional inequalities, symmegroup properties and spectrum estimates*, Infinite Dim. Anal., Quantum Probab. and related Topics, **3**:2, 263–295.
Wang, F. Y. (2001), *Convergence rates of Markov semigroups in probability distances*, preprint.
Wang, F. Y. (2002), *Coupling,convergence rate of Markov processes and weak Poincaré inequalities*, Sci. in China (A), **45**:8, 975–983.
Wang, Z. K. (1980), *Birth-death Processes and Markov Chains*, Science Press, Beijing (In Chinese).
Yang, X. Q. (1986), *Constructions of Time-homogeneous Markov Processes with Denumerable States*, Hunan Sci. Press, Hunan.
Zhang, S. Y. and Mao, Y. H. (2000), *Exponential convergence rate in Boltzman-Shannon entropy*, Sci. Sin. (*A*) **44**(3), 280–285.
Zhang, Y. H. (2001), *Strong ergodicity for single birth processes*, J. Appl. Prob. **38**, 270–277.

Department of Mathematics, Beijing Normal University, Beijing 100875, The People's Republic of China. E-mail: mfchen@bnu.edu.cn
Home page: http://www.bnu.edu.cn/~chenmf/main_eng.htm

Potential Analysis, Vol. 23, No. 4, 2005, 303-322

CAPACITARY CRITERIA FOR POINCARÉ-TYPE INEQUALITIES

Mu-Fa Chen

(Beijing Normal University)
(Received: 30 September 2003; accepted: 10 February 2005)

Abstract. The Poincaré-type inequality is a unification of various inequalities including the F-Sobolev inequalities, Sobolev-type inequalities, logarithmic Sobolev inequalities, and so on. The aim of this paper is to deduce some unified upper and lower bounds of the optimal constants in Poincaré-type inequalities for a large class of normed linear (Banach, Orlicz) spaces in terms of capacity. The lower and upper bounds differ only by a multiplicative constant, and so the capacitary criteria for the inequalities are also established. Both the transient and the ergodic cases are treated. Besides, the explicit lower and upper estimates in dimension one are computed.

1. Introduction. In this section, we recall some necessary notation and state the main results of this paper.

Let E be a locally compact separable metric space with Borel σ-algebra $\mathscr{E}$, μ an everywhere dense Radon measure on E, and $(D, \mathscr{D}(D))$ a regular Dirichlet form on $L^2(\mu) = L^2(E; \mu)$. The starting point of our study is the following result, due to V. G. Maz'ya (1973) [cf. Maz'ya [17] for references] in the typical case and Z. Vondraček [22] in general. Its proof is simplified recently by M. Fukushima and T. Uemura [10].

Theorem 1.0. For a regular transient Dirichlet form $(D, \mathscr{D}(D))$, the optimal constant A in the Poincaré inequality

$$\|f\|^2 = \int_E f^2 \mathrm{d}\mu \leqslant A\, D(f), \qquad f \in \mathscr{D}(D) \cap C_0(E), \tag{1.1}$$

satisfies $B \leqslant A \leqslant 4B$, where $\|\cdot\|$ is the norm in $L^2(\mu)$ and

$$B = \sup_{\text{compact } K} \frac{\mu(K)}{\mathrm{Cap}(K)}. \tag{1.2}$$

2000 *Mathematics Subject Classification.* 60J55, 31C25, 60J35, 47D07.

Key words and phrases. Dirichlet form, isoperimetric constant, logarithmic Sobolev inequality, Poincaré-type inequality, Orlicz space.

Research supported in part NSFC (No. 10121101) and 973 Project.

Recall that

$$\mathrm{Cap}(K) = \inf\big\{D(f) : f \in \mathscr{D}(D) \cap C_0(E), f|_K \geqslant 1\big\},$$

where $C_0(E)$ is the set of continuous functions with compact support. Certainly, in (1.1), one may replace "$\mathscr{D}(D)\cap C_0(E)$" by "$\mathscr{D}(D)$" or by the extended Dirichlet space "$\mathscr{D}_e(D)$", which is the set of $\mathscr{E}$-measurable functions f: $|f| < \infty$, μ-a.e., there exists a sequence $\{f_n\} \subset \mathscr{D}(D)$ such that $D(f_n - f_m) \to 0$ as $n, m \to \infty$ and $\lim_{n\to\infty} f_n = f$, μ-a.e. Refer to the standard books Fukushima, Oshima and Takeda [9], Ma and Röckner [14] for some preliminary facts about the Dirichlet forms theory.

Actually, inequality (1.1) in one-dimensional case was initiated by G. H. Hardy in 1920 and completed by B. Muckenhoupt in 1970 (see also Opic and Kufner [19]), in the context of diffusions (elliptic operators) with explicitly isoperimetric constant B.

The first goal of this paper is to extend (1.2) to the Poincaré-type inequality

$$\big\|f^2\big\|_{\mathbb{B}} \leqslant A_{\mathbb{B}} D(f), \qquad f \in \mathscr{D}(D) \cap C_0(E), \tag{1.3}$$

for a class of normed linear spaces $(\mathbb{B}, \|\cdot\|_{\mathbb{B}}, \mu)$ of real functions on E. To do so, we need the following assumptions on $(\mathbb{B}, \|\cdot\|_{\mathbb{B}}, \mu)$.

(H_1) $I_K \in \mathbb{B}$ for all compact K.
(H_2) If $h \in \mathbb{B}$ and $|f| \leqslant h$, then $f \in \mathbb{B}$.
(H_3) $\|f\|_{\mathbb{B}} = \sup_{g\in\mathscr{G}} \int_E |f| g \mathrm{d}\mu$,

where $\mathscr{G}$, to be specified case by case, is a class of nonnegative $\mathscr{E}$-measurable functions. By using Fatou's lemma and the completeness of $\mathscr{D}_e(D)$, one can also replace "$\mathscr{D}(D) \cap C_0(E)$" by "$\mathscr{D}_e(D)$" in (1.3).

We can now state our first result as follows.

Theorem 1.1. Assume (H_1)–(H_3). For a regular transient Dirichlet form $(D, \mathscr{D}(D))$, the optimal constant $A_{\mathbb{B}}$ in (1.3) satisfies

$$B_{\mathbb{B}} \leqslant A_{\mathbb{B}} \leqslant 4B_{\mathbb{B}}, \tag{1.4}$$

where

$$B_{\mathbb{B}} := \sup_{\text{compact } K} \frac{\|I_K\|_{\mathbb{B}}}{\mathrm{Cap}(K)}. \tag{1.5}$$

When $\mathbb{B} = L^p(\mu)\,(p \geqslant 1)$, Theorem 1.1 was proven by Fukushima and Uemura [10].

Next, we go to the ergodic case. Assume that $\mu(E) < \infty$ and set $\pi = \mu/\mu(E)$. Throughout this paper, we use the simplified notation: $\bar{f} = f - \pi(f)$, where $\pi(f) = \int f \mathrm{d}\pi$. We adopt a splitting technique. Let $E_1 \subset E$ be open with $\pi(E_1) \in (0, 1)$ and write $E_2 = E_1^c \setminus \partial E_1$. Restricting the functions f in (1.1) to each E_i $\big($i.e., $f|_{E_i^c} = 0$, μ-a.e.$\big)$, by Theorem 1.0, we obtain the corresponding constant B_i as follows.

$$B_i = \sup_{\text{compact } K \subset E_i} \frac{\mu(K)}{\mathrm{Cap}(K)}, \qquad i = 1, 2. \tag{1.6}$$

This notation is meaningful because the restriction to an open set of a regular Dirichlet form is again regular (cf. Fukushima et al [9], Theorem 4.4.3). Moreover, since $(D, \mathscr{D}(D))$ is irreducible, its restrictions to E_1 and E_2 must be transient.

Theorem 1.2. Let $\mu(E) < \infty$. Then for a regular, irreducible, and conservative Dirichlet form, the optimal constant $\overline{A}$ in the Poincaré inequality

$$\|\bar{f}^2\| \leqslant \overline{A}D(f), \qquad f \in \mathscr{D}(D) \cap C_0(E), \tag{1.7}$$

satisfies

$$\begin{aligned} &\overline{A} \geqslant \sup_{\text{open } E_1 \text{ and } E_2} \max\big\{B_1\pi(E_1^c),\, B_2\pi(E_2^c)\big\} \geqslant \frac{1}{2} \sup_{\text{open } E_1:\, \pi(E_1)\in(0,1/2]} B_1, \\ &\overline{A} \leqslant 4 \sup_{\text{open } E_1:\, \pi(E_1)\in(0,1/2]} B_1. \end{aligned} \tag{1.8}$$

In particular, $\overline{A} < \infty$ iff $\sup_{\text{open } E_1:\, \pi(E_1)\in(0,1/2]} B_1 < \infty$.

The restriction of $\mathbb{B}$ to E_i gives us $(\mathbb{B}^i, \|\cdot\|_{\mathbb{B}^i}, \mu^i)$:

$$\begin{aligned} &\mathbb{B}^i = \{f|_{E_i} : f \in \mathbb{B}\}, \quad \mu^i = \mu|_{E_i}, \quad \mathscr{G}^i = \{g|_{E_i} : g \in \mathscr{G}\}, \\ &\|f\|_{\mathbb{B}^i} = \sup_{g\in\mathscr{G}^i} \int_{E_i} |f| g \mathrm{d}\mu^i = \sup_{g\in\mathscr{G}} \int_{E_i} |f| g \mathrm{d}\mu, \qquad i = 1, 2. \end{aligned}$$

Correspondingly, we have a restricted Dirichlet form $\big(D, \mathscr{D}_i\big)$ on $L^2(E_i, \mu^i)$, where $\mathscr{D}_i = \{f \in \mathscr{D}(D) :$ the quasi-version of f equals 0 on E_i^c, q.e.$\}$. The corresponding constants given by Theorem 1.1 are denoted by $A_{\mathbb{B}^i}$ and $B_{\mathbb{B}^i}$ $(i = 1, 2)$, respectively.

In the ergodic case, we also use the following assumptions.

(H_4) $\mu(E) < \infty$.

(H_5) $1 \in \mathbb{B}$.

Denote by c_1 a constant such that

$$|\pi(f)| \leqslant c_1 \|f\|_{\mathbb{B}}, \qquad f \in \mathbb{B}. \tag{1.9}$$

For each $G \subset E$, denote by $c_2(G)$ a constant such that

$$|\pi(f I_G)| \leqslant c_2(G) \|f I_G\|_{\mathbb{B}}, \qquad f \in \mathbb{B}. \tag{1.10}$$

Theorem 1.3. Let $(D, \mathscr{D}(D))$ be a regular, irreducible, and conservative Dirichlet form. Assume that (H_2)–(H_5) hold and that

$$\sup_{\text{open } E_1:\, \pi(E_1)\in(0,1/2]} c_2(E_1)\pi(E_1)\|1\|_{\mathbb{B}} < 1.$$

Then the optimal constant $\overline{A}_{\mathbb{B}}$ in the Poincaré-type inequality

$$\big\|\bar{f}^2\big\|_{\mathbb{B}} \leqslant \overline{A}_{\mathbb{B}} D(f), \qquad f \in \mathscr{D}(D) \cap C_0(E), \tag{1.11}$$

satisfies

$$\overline{A}_{\mathbb{B}} \geqslant \underline{\kappa} \sup_{\text{open } E_1:\, \pi(E_1)\in(0,1/2]} A_{\mathbb{B}^1} \geqslant \underline{\kappa} \sup_{\text{open } E_1:\, \pi(E_1)\in(0,1/2]} B_{\mathbb{B}^1}, \tag{1.12}$$

$$\overline{A}_{\mathbb{B}} \leqslant \bar{\kappa} \sup_{\text{open } E_1:\, \pi(E_1)\in(0,1/2]} A_{\mathbb{B}^1} \leqslant 4\bar{\kappa} \sup_{\text{open } E_1:\, \pi(E_1)\in(0,1/2]} B_{\mathbb{B}^1}, \tag{1.13}$$

where

$$\underline{\kappa} = \Big(1 - \sup_{\text{open } E_1:\, \pi(E_1)\in(0,1/2]} \sqrt{c_2(E_1)\pi(E_1)\|1\|_{\mathbb{B}}}\,\Big)^2, \quad \bar{\kappa} = \Big(1 + \sqrt{c_1\|1\|_{\mathbb{B}}}\,\Big)^2.$$

A typical case for which one needs the Banach form of Poincaré-type inequality is the F-Sobolev inequality (cf. Wang [23], Gong and Wang [11]):

$$\int_E f^2 F\big(f^2\big)\mathrm{d}\mu \leqslant A_F D(f), \qquad f \in \mathscr{D}(D)\cap C_0(E). \tag{1.14}$$

Recall that a function Φ: $\mathbb{R}\to\mathbb{R}$ is an N-function if it is nonnegative, continuous, convex, even (i.e., $\Phi(-x)=\Phi(x)$), and satisfies

$$\Phi(x)=0 \text{ iff } x=0, \qquad \lim_{x\to 0}\Phi(x)/x = 0, \qquad \lim_{x\to\infty}\Phi(x)/x=\infty.$$

We will often assume the following growth condition (or Δ_2-condition) for Φ:

$$\sup_{x\gg 1}\Phi(2x)/\Phi(x)<\infty \ \Big(\iff \sup_{x\gg 1} x\Phi'_-(x)/\Phi(x)<\infty\Big),$$

where Φ'_- is the left derivative of Φ. The conditions listed below for F guarantee that the function $\Phi(x) := |x|F(|x|)$, as an N-function, satisfies the above conditions.

Theorem 1.4. Let $F:\mathbb{R}_+\to\mathbb{R}_+$ satisfy the following conditions:

(1) $2F' + xF'' \geqslant 0$ on $[0,\infty)$.
(2) $F\neq 0$ on $(0,\infty)$, $\lim_{x\to 0}F(x)=0$ and $\lim_{x\to\infty}F(x)=\infty$.
(3) $\sup_{x\gg 1} xF'(x)/F(x)<\infty$.

Then Theorem 1.1 is valid for the Orlicz space $\big(\mathbb{B}=L^{\Phi}(\mu), \|\cdot\|_{\mathbb{B}}=\|\cdot\|_{\Phi}\big)$ with N-function $\Phi(x)=|x|F(|x|)$:

$$L^{\Phi}(\mu)=\Big\{f:\int_E \Phi(f)\mathrm{d}\mu<\infty\Big\}, \tag{1.15}$$

$$\|f\|_{\Phi}=\sup\Big\{\int_E |f|g\mathrm{d}\mu : \int \Phi_c(g)\mathrm{d}\mu\leqslant 1\Big\}, \tag{1.16}$$

where Φ_c is the complementary function of Φ. [If we denote by φ_c the inverse function of the left-derivative of Φ, then Φ_c can be expressed as $\Phi_c(y)=\int_0^{|y|}\varphi_c$.] Furthermore the isoperimetric constant is given by

$$B_{\Phi}=\sup_{\text{compact } K}\frac{\alpha_*(K)^{-1}+\mu(K)F(\alpha_*(K))}{\mathrm{Cap}(K)}, \tag{1.17}$$

where $\alpha_*(K)$ is the minimal positive root of the equation: $\alpha^2F'(\alpha)=\mu(K)$.

The corresponding ergodic case of Theorem 1.4 has been treated in [1; Theorems 11 and 12].

A more particular case is that $F=\log$. Then we have, in the ergodic case, the logarithmic Sobolev inequality

$$\int_E f^2\log\big[f^2/\pi(f^2)\big]\mathrm{d}\mu\leqslant A_{\mathrm{Log}}D(f), \qquad f\in\mathscr{D}(D)\cap C_0(E) \tag{1.18}$$

(due to L. Gross (1976), cf. Gross [12] and the references within). By examining the entropy carefully, using different Banach spaces (but not Orliczian) for the upper and lower bounds respectively, we obtain the following result.

Theorem 1.5. Let $(D, \mathscr{D}(D))$ be a regular, irreducible, and conservative Dirichlet form. Assume that (H_2)–(H_5) hold. Then we have

$$\frac{\log 2}{\log(1+2e^2)}\, B_{\text{Log}}(e^2) \leqslant B_{\text{Log}}(1/2) \leqslant A_{\text{Log}} \leqslant 4\, B_{\text{Log}}(e^2), \tag{1.19}$$

where

$$B_{\text{Log}}(\gamma) = \sup_{\substack{\text{open } O:\ \pi(O)\in(0,1/2] \\ \text{compact } K\subset O}} \frac{\mu(K)}{\text{Cap}(K)} \log\left(1+\frac{\gamma}{\pi(K)}\right). \tag{1.20}$$

One may regard Theorem 1.1–1.5 as extensions of the one-dimensional results obtained by S. G. Bobkov and F. Götze [3], Y. H. Mao [15, 16], F. Barthe and C. Roberto [2] and the author [5, 6]. However, the criteria and estimates given in the quoted papers are completely explicit, without using capacity. Even though the capacitary results in dimension one can also be made explicit, as shown in Section 4, the capacitary results here are much more involved; but this may be the price one has to pay for such a general setup (for the higher dimensions, in particular). Nevertheless, we have got the precise formula for the isoperimetric constant $B_{\mathbb{B}}$ (or B_Φ) in the general setup. Of course, it is valuable to work out more explicit expression for the constant in particular situations.

The proofs of Theorems 1.1–1.4 are presented in the next section. The proof of Theorem 1.5 and some related results are given in Section 3. In the last section, the isoperimetric constants in dimension one are computed explicitly.

2. Proofs of Theorems 1.1–1.4.

The key to prove Theorem 1.0 is the following result [cf. Fukushima and Uemura [10], Theorem 2.1]:

Theorem 2.1.

$$\int_0^\infty \text{Cap}(\{x\in E: |f(x)|\geqslant t\})\text{d}\big(t^2\big) \leqslant 4D(f), \qquad f\in\mathscr{D}(D)\cap C_0(E).$$

Having Theorem 2.1 in mind, the proof of Theorem 1.0 and more generally of Theorem 1.1 is quite standard. Here we follow the proof of Theorem 3.1 in Kaimanovich [13].

Proof of Theorem 1.1. Let $f\in\mathscr{D}(D)\cap C_0(E)$ and set $N_t=\{|f|\geqslant t\}$. Since N_t is compact, by (H_1), $I_{N_t}\in\mathbb{B}$. Next, since $|f|\leqslant\|f\|_\infty I_{\text{supp}(f)}$, by (H_1) and (H_2), $f^2\in\mathbb{B}$. Note that

$$\int_0^\infty I_{N_t}\text{d}\big(t^2\big) = 2\int_0^\infty t I_{\{|f|\geqslant t\}}\text{d}t = 2\int_0^{|f|} t\,\text{d}t = f^2 \quad \text{(co-area formula)}.$$

By (H_3), the definition of $B_{\mathbb{B}}$, and Theorem 2.1, we obtain

$$\begin{aligned}\big\|f^2\big\|_{\mathbb{B}} &= \sup_{g\in\mathscr{G}}\int_E f^2 g\mathrm{d}\mu\\ &= \sup_{g\in\mathscr{G}}\int_E\bigg(\int_0^\infty I_{N_t}\mathrm{d}\big(t^2\big)\bigg)g\mathrm{d}\mu\\ &= \sup_{g\in\mathscr{G}}\int_0^\infty\bigg(\int_E I_{N_t}g\mathrm{d}\mu\bigg)\mathrm{d}\big(t^2\big)\\ &\leqslant \int_0^\infty \big\|I_{N_t}\big\|_{\mathbb{B}}\mathrm{d}\big(t^2\big)\\ &\leqslant B_{\mathbb{B}}\int_0^\infty \mathrm{Cap}(N_t)\mathrm{d}\big(t^2\big)\\ &\leqslant 4B_{\mathbb{B}}D(f).\end{aligned}$$

This implies that $A_{\mathbb{B}} \leqslant 4B_{\mathbb{B}}$.

Next, for every compact K and any function $f \in \mathscr{D}(D)\cap C_0(E)$ with $f|_K \geqslant 1$, by (H_2) and (H_3), we have

$$\big\|I_K\big\|_{\mathbb{B}} \leqslant \big\|f^2\big\|_{\mathbb{B}} \leqslant A_{\mathbb{B}}D(f).$$

Thus,

$$\big\|I_K\big\|_{\mathbb{B}} \leqslant A_{\mathbb{B}}\inf\{D(f) : f\in\mathscr{D}(D)\cap C_0(E), f|_K\geqslant 1\} = A_{\mathbb{B}}\mathrm{Cap}(K).$$

Taking supremum over all compact K, it follows that $B_{\mathbb{B}} \leqslant A_{\mathbb{B}}$ and the proof is completed. □

To prove Theorem 1.2, we need the following result.

Lemma 2.2. Let $(D, \mathscr{D}(D))$ be a regular and conservative Dirichlet form, $\mu(E) < \infty$, $f \in \mathscr{D}(D)\cap C(E)$, and c a constant. Define $f^{\pm} = (f-c)^{\pm}$. Then we have $D(f) \geqslant D(f^+) + D(f^-)$.

Proof. Let $P_t(x, \mathrm{d}y)$ be the transition probability function determined by the Dirichlet form and set $\mu_t(\mathrm{d}x, \mathrm{d}y) = \mu(\mathrm{d}x)P_t(x, \mathrm{d}y)$. Then, by the spectral representation theorem, we have

$$\frac{1}{2t}\int \mu_t(\mathrm{d}x, \mathrm{d}y)[g(y)-g(x)]^2 \uparrow D(g) \qquad \text{as } t\downarrow 0 \text{ for all } g\in L^2(\mu). \tag{2.1}$$

Next, $\{f^+ > 0\}$ and $\{f^- > 0\}$ are open sets on which the restricted Dirichlet forms are also regular. Moreover, since $1 \in \mathscr{D}(D)$, we have $f^{\pm} \in \mathscr{D}(D)$; and hence $f^{\pm}$ belong to the corresponding restricted Dirichlet forms, respectively. Furthermore, it is easy to check the following crucial identity:

$$|f(y)-f(x)| = \big|f^+(y)-f^+(x)\big| + \big|f^-(y)-f^-(x)\big|. \tag{2.2}$$

Therefore

$$\begin{aligned}
D(f) &= \lim_{t\downarrow 0}\frac{1}{2t}\int \mu_t(\mathrm{d}x,\mathrm{d}y)\big[f(y)-f(x)\big]^2\\
&= \lim_{t\downarrow 0}\frac{1}{2t}\int \mu_t(\mathrm{d}x,\mathrm{d}y)\big[\big|f^+(y)-f^+(x)\big|+\big|f^-(y)-f^-(x)\big|\big]^2\\
&\geqslant \lim_{t\downarrow 0}\frac{1}{2t}\int \mu_t(\mathrm{d}x,\mathrm{d}y)\big(f^+(y)-f^+(x)\big)^2+\\
&\quad+\lim_{t\downarrow 0}\frac{1}{2t}\int \mu_t(\mathrm{d}x,\mathrm{d}y)\big(f^-(y)-f^-(x)\big)^2\\
&= D\big(f^+\big)+D\big(f^-\big). \quad \square
\end{aligned} \tag{2.3}$$

Proof of Theorem 1.2. The proof below is essentially the same as in Chen and Wang [7] and Chen [4, Theorem 3.1].

For each $\varepsilon > 0$, choose $f_\varepsilon \in \mathscr{D}(D)\cap C_0(E)$ with $\pi(f_\varepsilon) = 0$ and $\pi(f_\varepsilon^2) = 1$ such that $\overline{A}^{-1} + \varepsilon \geqslant D(f_\varepsilon)$. Next, choose c_ε such that $\pi(f_\varepsilon < c_\varepsilon) \leqslant 1/2$ and $\pi(f_\varepsilon > c_\varepsilon) \leqslant 1/2$. Set $f_\varepsilon^\pm = (f_\varepsilon - c_\varepsilon)^\pm$ and $G_\varepsilon^\pm = \{f_\varepsilon^\pm > 0\}$. Then $G_\varepsilon^\pm$ are open sets and Theorem 1.1 is meaningful for the restricted Dirichlet forms on $G_\varepsilon^\pm$. Denote by $A(G)$ the the optimal constant A in (1.1), when the functions are restricted on G. By Lemma 2.2, we obtain

$$\begin{aligned}
1 &\leqslant 1 + c_\varepsilon^2\\
&= \pi\big[\big(f_\varepsilon^+\big)^2 + \big(f_\varepsilon^-\big)^2\big]\\
&\leqslant A\big(G_\varepsilon^+\big)D\big(f_\varepsilon^+\big) + A\big(G_\varepsilon^-\big)D\big(f_\varepsilon^-\big)\\
&\leqslant \big[A\big(G_\varepsilon^+\big)\vee A\big(G_\varepsilon^-\big)\big]\,\big[D\big(f_\varepsilon^+\big)+D\big(f_\varepsilon^-\big)\big]\\
&\leqslant \big[A\big(G_\varepsilon^+\big)\vee A\big(G_\varepsilon^-\big)\big]\,D\big(f_\varepsilon\big)\\
&\leqslant \big[A\big(G_\varepsilon^+\big)\vee A\big(G_\varepsilon^-\big)\big]\big(\overline{A}^{-1}+\varepsilon\big)\\
&\leqslant \big(\overline{A}^{-1}+\varepsilon\big)\sup_{\text{open } O:\pi(O)\in(0,1/2]} A(O).
\end{aligned}$$

Because ε is arbitrary, we obtain a upper bound of $\overline{A}$.

Next, for every $f \in \mathscr{D}(D)$ with $f|_{G^c} = 0$ and $\pi(f^2) = 1$, we have

$$\pi\big(f^2\big) - \pi(f)^2 = 1 - \pi\big(fI_G\big)^2 \geqslant 1 - \pi\big(f^2\big)\pi(G) = 1 - \pi(G) = \pi(G^c).$$

Hence

$$\overline{A} \geqslant \frac{\pi\big(f^2\big)-\pi(f)^2}{D(f)} \geqslant \frac{\pi(G^c)}{D(f)}.$$

This implies that $\overline{A} \geqslant A(G)\pi(G^c)$. By symmetry, we have

$$\overline{A} \geqslant \max\big\{A(E_1)\pi(E_1^c),\ A(E_2)\pi(E_2^c)\big\}.$$

Therefore

$$\begin{aligned}\overline{A} &\geqslant \sup_{\text{open } E_1 \text{ and } E_2} \max\big\{A(E_1)\pi(E_1^c),\ A(E_2)\pi(E_2^c)\big\}\\ &\geqslant \frac{1}{2} \sup_{\text{open } O:\ \pi(O)\in(0,1/2]} A(O).\end{aligned}$$

This gives us a lower bound of $\overline{A}$.

Finally, the assertion of Theorem 1.2 follows from Theorem 1.0. □

To prove Theorem 1.3, we need the following proposition, taken from Chen [6, Proposition 2.4].

Proposition 2.3. Let $(E, \mathscr{E}, \pi)$ be a probability space and $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ a normed linear space, satisfying (H_5) and (H_2), of Borel measurable functions on $(E, \mathscr{E}, \pi)$.

(1) Let c_1 be given by (1.9). Then

$$\big\|\bar{f}^2\big\|_{\mathbb{B}} \leqslant \big(1+\sqrt{c_1\|1\|_{\mathbb{B}}}\,\big)^2\big\|f^2\big\|_{\mathbb{B}}.$$

(2) Let $c_2(G)$ be given by (1.10). If $c_2(G)\pi(G)\|1\|_{\mathbb{B}} < 1$, then for every f with $f|_{G^c} = 0$, we have

$$\big\|f^2\big\|_{\mathbb{B}} \leqslant \big\|\bar{f}^2\big\|_{\mathbb{B}}\big/\big[1-\sqrt{c_2(G)\pi(G)\,\|1\|_{\mathbb{B}}}\,\big]^2.$$

Proof of Theorem 1.3. (a) Let $f \in \mathscr{D}(D) \cap C_0(E)$. Choose c_f such that $E_1 := \{f > c_f\}$ and $E_2 := \{f < c_f\}$ satisfy $\pi(E_1) \leqslant 1/2$ and $\pi(E_2) \leqslant 1/2$. Then E_1 and E_2 are open sets. Define $f_1 = (f - c_f)^+$ and $f_2 = (f - c_f)^-$. By part (1) of Proposition 2.3, it follows that

$$\big\|\bar{f}^2\big\|_{\mathbb{B}} = \big\|\overline{f-c_f}^{\,2}\big\|_{\mathbb{B}} \leqslant \Big(1+\sqrt{c_1\|1\|_{\mathbb{B}}}\Big)^2\big\|(f-c_f)^2\big\|_{\mathbb{B}}.$$

But

$$\big\|(f-c_f)^2\big\|_{\mathbb{B}} = \big\|f_1^2 + f_2^2\big\|_{\mathbb{B}} \leqslant \big\|f_1^2\big\|_{\mathbb{B}} + \big\|f_2^2\big\|_{\mathbb{B}},$$

hence by Theorem 1.1 and Lemma 2.2, we get

$$\begin{aligned}\big\|\bar{f}^2\big\|_{\mathbb{B}} &\leqslant \Big(1+\sqrt{c_1\|1\|_{\mathbb{B}}}\Big)^2 A_{\mathbb{B}^1} D(f_1) + \Big(1+\sqrt{c_1\|1\|_{\mathbb{B}}}\Big)^2 A_{\mathbb{B}^2} D(f_2)\\ &\leqslant \Big(1+\sqrt{c_1\|1\|_{\mathbb{B}}}\Big)^2 \big(A_{\mathbb{B}^1} \vee A_{\mathbb{B}^2}\big)\big(D(f_1)+D(f_2)\big)\\ &\leqslant \Big(1+\sqrt{c_1\|1\|_{\mathbb{B}}}\Big)^2 \big(A_{\mathbb{B}^1} \vee A_{\mathbb{B}^2}\big) D(f).\end{aligned}$$

This means that

$$\overline{A}_{\mathbb{B}} \leqslant \Big(1+\sqrt{c_1\|1\|_{\mathbb{B}}}\Big)^2\big(A_{\mathbb{B}^1} \vee A_{\mathbb{B}^2}\big) \leqslant \Big(1+\sqrt{c_1\|1\|_{\mathbb{B}}}\Big)^2 \sup_{\text{open } E_1:\ \pi(E_1)\leqslant 1/2} A_{\mathbb{B}^1}.$$

(b) Conversely, assume that (1.11) holds. Let $f \in \mathscr{D}(D) \cap C_0(E)$, $f|_{E_1^c} = 0$ for some open E_1 with $\pi(E_1) \leqslant 1/2$. Then, from part (2) of Proposition 2.3 and (1.11), it follows that

$$\|f^2\|_{\mathbb{B}} \leqslant \frac{\|\bar{f}^2\|_{\mathbb{B}}}{\left(1 - \sqrt{c_2(E_1)\pi(E_1)\|1\|_{\mathbb{B}}}\,\right)^2} \leqslant \frac{\overline{A}_{\mathbb{B}}}{\left(1 - \sqrt{c_2(E_1)\pi(E_1)\|1\|_{\mathbb{B}}}\,\right)^2} D(f).$$

This means that

$$\overline{A}_{\mathbb{B}} \geqslant \left(1 - \sqrt{c_2(E_1)\pi(E_1)\|1\|_{\mathbb{B}}}\,\right)^2 A_{\mathbb{B}^1}.$$

Noticing that $\sup_{\text{open } E_1\colon \pi(E_1)\leqslant 1/2} c_2(E_1)\pi(E_1)\|1\|_{\mathbb{B}} < 1$ by assumption, we obtain

$$\overline{A}_{\mathbb{B}} \geqslant \left(1 - \sup_{\text{open } E_1\colon \pi(E_1)\leqslant 1/2} \sqrt{c_2(E_1)\pi(E_1)\|1\|_{\mathbb{B}}}\right)^2 \sup_{\text{open } E_1\colon \pi(E_1)\leqslant 1/2} A_{\mathbb{B}^1}. \qquad \square$$

Proof of Theorem 1.4. By assumptions, $\Phi(x) = |x|F(|x|)$ is an N-function. From M. M. Rao and Z. D. Ren [20, §3.3, Theorem 13 and Proposition 14], it follows that

$$\|I_G\|_\Phi = \inf_{\alpha>0} \frac{1}{\alpha}\big(1 + \mu(G)\Phi(\alpha)\big).$$

The infimum on the right-hand side is achieved at α^*, which is the minimal root of the equation: $\alpha^2 F'(\alpha) = \mu(G)$. Combining this with (1.5), we get (1.17). $\square$

3. Logarithmic Sobolev inequality.

This section is devoted to the logarithmic Sobolev inequality. First, we present a result as an illustration of the application of Theorem 1.3. Then we prove the refinement Theorem 1.5.

Theorem 3.1. Let $(D, \mathscr{D}(D))$ be a regular, irreducible, and conservative Dirichlet form. Assume that (H_2)–(H_5) hold. Next, let $\Phi(x) = |x|\log(1 + |x|)$. Then the optimal A_{Log} in (1.18) satisfies

$$\frac{(\sqrt{2}-1)^2}{5} B_\Phi \leqslant A_{\mathrm{Log}} \leqslant \frac{51 \times 16}{5} B_\Phi, \tag{3.1}$$

where

$$B_\Phi = \sup_{\substack{\text{open } O\colon \pi(O)\in(0,1/2] \\ \text{compact } K\subset O}} \frac{M(\mu(K))}{\mathrm{Cap}(K)}, \tag{3.2}$$

$$M(x) = \frac{1}{2}\big(\sqrt{1+4x} - 1\big) + x\log\left(1 + \frac{1+\sqrt{1+4x}}{2x}\right) \sim x\log x^{-1} \qquad \text{as } x \to 0.$$

Since the proof of Theorem 3.1 is essentially known, we sketch the main steps only for the reader's convenience.

From now on, we fix $\Phi(x) = |x|\log(1 + |x|)$ and define $\Psi(x) = x^2\log(1 + x^2)$. We need an equivalent norm $\|\cdot\|_{(\Phi)}$ of $\|\cdot\|_\Phi$ as follows

$$\|f\|_{(\Phi)} = \inf\left\{\alpha > 0 : \int_E \Phi(f/\alpha)\mathrm{d}\mu \leqslant 1\right\},$$

which is usually easier to compute. The key observation is the following result:

Lemma 3.2. For any f with $f^2 \in L^{\Phi}(\mu)$, we have

$$\frac{4}{5}\big\|f-\pi(f)\big\|_{(\Psi)}^2 \leqslant \mathscr{L}(f) \leqslant \frac{51}{20}\big\|f-\pi(f)\big\|_{(\Psi)}^2,$$

where $\mathscr{L}(f) = \sup_{c\in\mathbb{R}} \mathrm{Ent}\big((f+c)^2\big)$ and $\mathrm{Ent}(f) = \int_{\mathbb{R}} f \log\big(f/\|f\|_{L^1(\pi)}\big)\mathrm{d}\mu$ for $f \geqslant 0$.

This result comes from Bobkov and Götze [3] and Deuschel and Stroock [8, p. 247], which go back to Rothaus [21]. An improvement of the coefficients is made in Chen [5]. Lemma 3.2 leads to the use of the Orlicz space $\mathbb{B} = L^{\Psi}(\mu)$ with norm $\|\cdot\|_{(\Psi)}$ and the following inequalities

$$\|f\|_{\mathbb{B}}^2 \leqslant A'_{\mathbb{B}} D(f), \qquad f \in \mathscr{D}(D)\cap C_0(E), \tag{3.3}$$

$$\big\|\bar f\big\|_{\mathbb{B}}^2 \leqslant \overline{A}'_{\mathbb{B}} D(f), \qquad f \in \mathscr{D}(D)\cap C_0(E), \tag{3.4}$$

as variants of (1.1) and (1.11). In parallel to Proposition 2.3, we have (cf. Chen [6, Proposition 3.4]) the following result.

Proposition 3.3. Everything in the premise is the same as in Proposition 2.3.

(1) Assume that there is a constant c_1' such that $|\pi(f)| \leqslant c_1'\|f\|_{\mathbb{B}}$ for all $f \in \mathbb{B}$. Then

$$\big\|\bar f\big\|_{\mathbb{B}} \leqslant \big(1 + c_1'\|1\|_{\mathbb{B}}\big)\|f\|_{\mathbb{B}}.$$

(2) Next, for a given $G \in \mathscr{E}$, let $c_2'(G)$ be a constant such that $|\pi(fI_G)| \leqslant c_2'(G)\|fI_G\|_{\mathbb{B}}$ for all $f \in \mathbb{B}$. If $c_2'(G)\|1\|_{\mathbb{B}} < 1$, then for every f with $f|_{G^c} = 0$ we have

$$\|f\|_{\mathbb{B}} \leqslant \big\|\bar f\big\|_{\mathbb{B}}\big/\big[1 - c_2'(G)\|1\|_{\mathbb{B}}\big].$$

Denote by $A'_{\mathbb{B}^i}$ the optimal constant in (3.3) when the functions are restricted to E_i, $i = 1, 2$. By using Proposition 3.3 and following the proof of Theorem 1.3, we obtain the following result.

Theorem 3.4. Let $(D, \mathscr{D}(D))$ be a regular, irreducible, and conservative Dirichlet form. Assume that (H_2)–(H_5) hold and that

$$\sup_{\text{open } E_1:\ \pi(E_1)\in(0,1/2]} c_2'(E_1)\|1\|_{\mathbb{B}} < 1.$$

Then the optimal constants $A'_{\mathbb{B}^i}$ and $\overline{A}'_{\mathbb{B}}$ in (3.3) and (3.4), respectively, obey the following relation:

$$\Big(1 - \sup_{\text{open } E_1:\ \pi(E_1)\in(0,1/2]} \sqrt{c_2'(E_1)\|1\|_{\mathbb{B}}}\,\Big)^2 A'_{\mathbb{B}^1} \leqslant \overline{A}'_{\mathbb{B}} \leqslant 4\Big(1 + \sqrt{c_1'\|1\|_{\mathbb{B}}}\,\Big)^2 A'_{\mathbb{B}^1}.$$

Proof of Theorem 3.1.

(a) First, we compute the constants c_1' and $c_2'(E_1)$ used in Theorem 3.4. Actually, this can be done in the same way as in the proof of the last theorem in Chen [5] or [6]:

$$c_1' = \Psi^{-1}(Z^{-1}), \tag{3.5}$$
$$c_2'(E_1) = \Psi^{-1}(Z_1^{-1})Z_1/Z. \tag{3.6}$$

where Ψ^{-1} is the inverse of Ψ and $Z = \mu(E)$, $Z_1 = \mu(E_1)$. As an illustration, we now prove (3.6). Because of the convexity of Φ, we have for $f_1 := fI_{E_1}$ that

$$\begin{aligned}\|f_1\|_{(\Phi)} &= \inf\left\{\alpha > 0 : \frac{1}{Z_1}\int_{E_1}\Phi(|f|/\alpha)\mathrm{d}\pi \leqslant \frac{1}{Z_1}\right\}\\ &\geqslant \inf\left\{\alpha > 0 : \Phi\bigg(\frac{1}{Z_1}\int_{E_1}|f|\mathrm{d}\pi/\alpha\bigg) \leqslant \frac{1}{Z_1}\right\}\\ &= \frac{Z\pi(|f_1|)}{Z_1\Phi^{-1}(Z_1^{-1})}.\end{aligned}$$

Hence

$$\begin{aligned}\|f_1\|_{(\Psi)}^2 &= \|f_1^2\|_{(\Phi)}\\ &\geqslant \frac{Z}{Z_1\Phi^{-1}(Z_1^{-1})}\pi\big(f_1^2\big)\\ &\geqslant \frac{Z^2}{Z_1^2\Phi^{-1}(Z_1^{-1})}\big[\pi(f_1)\big]^2\\ &= \left[\frac{Z\pi(f_1)}{Z_1\Psi^{-1}(Z_1^{-1})}\right]^2.\end{aligned}$$

This means that one can choose $c_2'(E_1)$ as in (3.6).

(b) Next, since $\|1\|_{(\Psi)} = 1/\Psi^{-1}(Z^{-1})$, $Z_1 \leqslant Z/2$, and $\Psi^{-1}(x)/x$ is decreasing in x, it follows that

$$\begin{aligned}\sup_{\text{open } E_1:\, \pi(E_1)\in(0,1/2]} c_2'(E_1)\|1\|_{(\Psi)} &= \sup_{\text{open } E_1:\, \pi(E_1)\in(0,1/2]} \frac{Z_1\Psi^{-1}(Z_1^{-1})}{Z\Psi^{-1}(Z^{-1})}\\ &\leqslant \frac{\Psi^{-1}(2Z^{-1})}{2\Psi^{-1}(Z^{-1})}\\ &< 1,\end{aligned}$$

and so the assumption of Theorem 3.4 holds.

Note that
$$\left[1 - \frac{\Psi^{-1}(2Z^{-1})}{2\Psi^{-1}(Z^{-1})}\right]^2 \geqslant \frac{(\sqrt{2}-1)^2}{2}, \tag{3.7}$$
as proved at the end of Chen [5]. The estimates in (3.1) now follow from (3.5)–(3.7) and the following comparison of the norms: $\|f\|_{(\Phi)} \leqslant \|f\|_{\Phi} \leqslant 2\|f\|_{(\Phi)}$. $\square$

We now turn to prove Theorem 1.5. Since the N-function $\Phi(x) = |x|\log(1+|x|)$ used in Theorem 3.1 is a little different form the function $|x|\log|x|$ used in the entropy, it is natural to examine the entropy more carefully. The starting point is the classical variational formula for the entropy $\mathrm{Ent}(\varphi) = \int_E \varphi \log(\varphi/\pi(\varphi))\mathrm{d}\pi$:

$$\mathrm{Ent}(\varphi) = \sup\bigg\{\int_E \varphi g \mathrm{d}\pi : \int_E e^g \mathrm{d}\pi \leqslant 1\bigg\}, \qquad \varphi \geqslant 0. \tag{3.8}$$

The right-hand side is very much the same as the norm defined by (H_3). However, the only nonnegative function g in the constraint is zero. This leads us to consider the following upper and lower estimates, due to Barthe and Roberto [2].

Lemma 3.5. Let $(X, \mathscr{B}, \pi)$ be a probability space, $G \in \mathscr{B}$, and $\varphi \in \mathscr{B}_+$ with $\varphi|_{G^c} = 0$. Then we have

(1)

$$\begin{aligned}\mathrm{Ent}(\varphi) + 2\int_X \varphi \mathrm{d}\pi &\leqslant \sup\bigg\{\int_X \varphi g \mathrm{d}\pi : \int_X e^g \mathrm{d}\pi \leqslant e^2 + 1,\ g \geqslant 0\bigg\}\\ &= \sup\bigg\{\int_G \varphi g \mathrm{d}\pi : \int_G e^g \mathrm{d}\pi \leqslant e^2 + \pi(G),\ g \geqslant 0\bigg\}, \qquad \varphi \geqslant 0.\end{aligned}$$

(2) If moreover $\pi(G) < 1$, then

$$\mathrm{Ent}(\varphi) \geqslant \sup\bigg\{\int_G \varphi g \mathrm{d}\pi : \int_G e^g \mathrm{d}\pi \leqslant 1,\ g \geqslant 0\bigg\}, \qquad \varphi \geqslant 0.$$

To compute the bounds in Lemma 3.5, we also need the following result ([2; Lemma 6]).

Lemma 3.6. Let $(X, \mathscr{B}, \mu)$ be a finite measure space, $C \geqslant \mu(X)$, and $G \in \mathscr{B}$ with $\mu(G) > 0$. Then

$$\sup\bigg\{\int_X I_G h \mathrm{d}\mu : \int_X e^h \mathrm{d}\mu \leqslant C \text{ and } h \geqslant 0\bigg\} = \mu(G)\log\bigg(1 + \frac{C - \mu(X)}{\mu(G)}\bigg).$$

The two parts in Lemma 3.5 are used, respectively, for the upper and lower estimates given in Theorem 1.5. In view of (3.8), part (2) of the lemma is quite close to the entropy. As will be seen below, part (1) of the lemma leads us to define a norm by (H_3), using

$$\mathscr{G} = \bigg\{g \geqslant 0 : \int_E e^g \mathrm{d}\pi \leqslant e^2 + 1\bigg\}.$$

It corresponds to $\Phi_c(x) = e^{-2}(e^{|x|} - 1)$ and hence $\Phi(x) = |x|\log|x| + |x|$ which is not an N-function, since $\lim_{x\to 0}\Phi(x)/x = -\infty$; and is even not a Young function, since $\Phi \not\geqslant 0$. Thus, we are out of the Orlicz spaces. In contrast with Theorem 3.1, here two different norms are adopted rather than a single one.

Proof of Theorem 1.5. For convenience, we replace the finite measure μ with the probability measure $\pi = \mu/\mu(E)$ in this proof. This makes no change of A_{Log} in (1.18).

(a) We now consider the normed linear space $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$, where the norm $\|\cdot\|_{\mathbb{B}}$ is defined by (H_3) in terms of

$$\mathscr{G} = \left\{ g \geqslant 0 : \int_E e^g \mathrm{d}\pi \leqslant e^2 + 1 \right\}.$$

Following the proof (a) of Theorem 1.3, for a given $f \in \mathscr{D}(D) \cap C_0(E)$, let c_f be a median of f and set $f_1 = (f - c_f)^+$ and $f_2 = (f - c_f)^-$. By Lemma 9 in Rothaus [21], we have

$$\mathrm{Ent}\big(f^2\big) \leqslant \inf_{c \in \mathbb{R}} \big\{ \mathrm{Ent}\big((f-c)^2\big) + 2\|f - c\|^2 \big\}. \tag{3.9}$$

Applying part (1) of Lemma 3.5 with $G = E$, Theorem 1.1, and Lemma 2.2, we obtain

$$\begin{aligned}
\mathrm{Ent}\big(f^2\big) &\leqslant \big\|(f - c_f)^2\big\|_{\mathbb{B}} \quad \text{(by (3.9) and Lemma 3.5)}\\
&= \big\|f_1^2 + f_2^2\big\|_{\mathbb{B}}\\
&\leqslant \big\|f_1^2\big\|_{\mathbb{B}} + \big\|f_2^2\big\|_{\mathbb{B}}\\
&\leqslant 4B_{\mathbb{B}^1} D(f_1) + 4B_{\mathbb{B}^2} D(f_2) \quad \text{(by Theorem 1.1)}\\
&\leqslant 4\big(B_{\mathbb{B}^1} \vee B_{\mathbb{B}^2}\big)\big(D(f_1) + D(f_2)\big)\\
&\leqslant 4\big(B_{\mathbb{B}^1} \vee B_{\mathbb{B}^2}\big) D(f) \quad \text{(by Lemma 2.2)},
\end{aligned}$$

where $B_{\mathbb{B}^i}$ is given by Theorem 1.1. More precisely, by part (1) of Lemma 3.5 with $G = E_i$, we have $\|f_i\|_{\mathbb{B}} = \|f_i\|_{\mathbb{B}^i}$ with respect to the class

$$\mathscr{G}^i = \left\{ g \geqslant 0 : \int_{E_i} e^g \mathrm{d}\pi \leqslant e^2 + \frac{1}{2} \right\}$$

of functions on $E_i := \{f_i > 0\}$, $i = 1, 2$. We have thus proved that

$$A_{\mathrm{Log}} \leqslant 4\big(B_{\mathbb{B}^1} \vee B_{\mathbb{B}^2}\big). \tag{3.10}$$

By Lemma 3.6, we have

$$\|I_K\|_{\mathbb{B}^i} = \pi(K) \log\left(1 + \frac{e^2 + 1/2 - 1/2}{\pi(K)}\right) = \pi(K) \log\left(1 + \frac{e^2}{\pi(K)}\right).$$

Combining this with (3.10), (1.5), and (1.20), we obtain $A_{\mathrm{Log}} \leqslant 4\, B_{\mathrm{Log}}(e^2)$.

(b) To prove the lower bound, assume (1.18). Let E_1 be open with $\pi(E_1) \leqslant 1/2$ and let $f \in \mathscr{D}(D) \cap C_0(E)$ with $f|_{E_1^c} = 0$. Then by part (2) of Lemma 3.5,

$$\mathrm{Ent}(f^2) \geqslant \sup\left\{ \int_{E_1} f^2 g \mathrm{d}\pi : \int_{E_1} e^g \mathrm{d}\pi \leqslant 1,\ g \geqslant 0 \right\}.$$

The right-hand side is the norm of f, denoted by $\|f\|_{\widetilde{\mathbb{B}}^1}$, with respect to a new class $\widetilde{\mathscr{G}}^1 = \big\{g \geqslant 0 : \int_{E_1} e^g \mathrm{d}\pi \leqslant 1\big\}$ of functions on E_1. To compute this norm, we use Lemma 3.6 again,

$$\begin{aligned}\|I_K\|_{\widetilde{\mathbb{B}}^1} &= \sup_{g\in\widetilde{\mathscr{G}}^1} \int_{E_1} I_K g \mathrm{d}\pi \\ &= \pi(K)\log\left(1+\frac{1-\pi(E_1)}{\pi(K)}\right) \\ &\geqslant \pi(K)\log\left(1+\frac{1}{2\pi(K)}\right), \qquad K\subset E_1.\end{aligned}$$

Combining this estimate with (1.18) and applying Theorem 1.1 to $\big(\widetilde{\mathbb{B}}^1, \|\cdot\|_{\widetilde{\mathbb{B}}^1}, \mu^1\big)$, we obtain $A_{\text{Log}} \geqslant B_{\text{Log}}(1/2)$ as required.

The factor $\log 2/\log(1+e^2)$ in the lower estimate of the theorem is due to the fact that

$$\frac{\log(1+e^2/x)}{\log(1+1/2x)} \uparrow \frac{\log(1+e^2)}{\log 2} < 4 \qquad \text{as } x\uparrow 1/2. \quad \square$$

4. Computation of isoperimetric constant in dimension one.

It is known that in general, the optimal constant A in (1.1) is not explicitly computable even in dimension one. However, the next two results show that the isoperimetric constant B in (1.2) in dimension one is computable and coincides with the Muckenhoupt-type bound (cf., [18], [4]).

Corollary 4.1. Consider an ergodic birth–death process with birth rates $b_i\,(i\geqslant 0)$ and death rates $a_i\,(i\geqslant 1)$. Define

$$\mu_0 = 1, \quad \mu_n = \frac{b_0 b_1\cdots b_{n-1}}{a_1 a_2\cdots a_n}, \quad n\geqslant 1.$$

Then the isoperimetric constant $B_{\mathbb{B}}$ in (1.5) with Dirichlet boundary at 0 can be expressed as follows:

$$B_{\mathbb{B}} = \sup_{n\geqslant 1} \|I_{[n,\infty)}\|_{\mathbb{B}} \sum_{i=0}^{n-1}\frac{1}{\mu_i b_i}.$$

Proof. (a) We show that in the definition of $\text{Cap}(K)$, one can replace "$f|_K \geqslant 1$" by "$f|_K = 1$".

Because $1\in\mathscr{D}(D)$, we have $f\wedge 1\in\mathscr{D}(D)\cap C_0(E)$ if so is f. Then the assertion follows from $D(f)\geqslant D(f\wedge 1)$.

(b) Next, let $K_i\,(i=1,2,\ldots,k)$ be disjoint intervals with natural order. Set $K=[\min K_1, \max K_k]$, where $\min K = \min\{i : i\in K\}$ and $\max K = \max\{i : i\in K\}$. We show that

$$\frac{\|I_K\|_{\mathbb{B}}}{\text{Cap}(K)} \geqslant \frac{\|I_{K_1+\cdots+K_k}\|_{\mathbb{B}}}{\text{Cap}(K_1+\cdots+K_k)}.$$

In other words, the ratio for a disconnected compact set is less than or equal to that of the corresponding connected one. For f with $f|_{K_1+\cdots+K_k} = 1$, the restriction of f to the intervals $[\max K_i, \min K_{i+1}]$ may not be a constant. Thus, if we define $\tilde{f} = f$ on K^c and $\tilde{f}|_K = 1$, then $D\big(\tilde{f}\big) \leqslant D(f)$, due to the character of birth–death processes. This means that $\mathrm{Cap}(K) \leqslant \mathrm{Cap}(K_1 + \cdots + K_k)$. In fact, equality holds, because for f with $f|_K = 1$, we must have $f|_{K_1+\cdots+K_k} = 1$ and so the inverse inequality is trivial. Since $K \supset K_1 + \cdots + K_k$ and (H_3), we have $\|I_K\|_{\mathbb{B}} \geqslant \|I_{K_1+\cdots+K_k}\|_{\mathbb{B}}$. This proves the required assertion.

(c) Because of (b), to compute the isoperimetric constant, it suffices to consider the compact sets having the form $K = \{n, n+1, \ldots, m\}$ for $m \geqslant n \geqslant 1$. We now fix such a compact set K and compute $\mathrm{Cap}(K)$.

Given f with $f|_K = 1$ and $\mathrm{supp}(f) = \{1, \ldots, N\}$, $N \geqslant m$, we have

$$D(f) = \sum_{i=0}^{n-1} \mu_i b_i (f_{i+1} - f_i)^2 + \sum_{i=m}^{N} \mu_i b_i (f_{i+1} - f_i)^2, \tag{4.1}$$

where $f_0 = 0$ and $f_{N+1} = 0$. Then

$$\begin{aligned}\frac{\partial D}{\partial f_j} &= -2\mu_j b_j (f_{j+1} - f_j) + 2\mu_{j-1} b_{j-1} (f_j - f_{j-1}) \\ &= -2\mu_j b_j v_j + 2\mu_{j-1} b_{j-1} v_{j-1}, \qquad 1 \leqslant j \leqslant n-1 \ \text{ or } \ m+1 \leqslant j \leqslant N,\end{aligned}$$

where $v_i = f_{i+1} - f_i$. The condition $\partial D / \partial f_j = 0$ gives us

$$v_j = \frac{\mu_{j-1} b_{j-1}}{\mu_j b_j} v_{j-1}, \qquad 1 \leqslant j \leqslant n-1 \ \text{ or } \ m+1 \leqslant j \leqslant N.$$

Hence

$$v_j = \frac{\mu_0 b_0 v_0}{\mu_j b_j}, \quad 0 \leqslant j \leqslant n-1, \ \text{ and } \ v_j = \frac{\mu_m b_m v_m}{\mu_j b_j}, \quad m \leqslant j \leqslant N. \tag{4.2}$$

Therefore

$$\begin{aligned} f_j &= \sum_{i=0}^{j-1} v_i = \mu_0 b_0 v_0 \sum_{i=0}^{j-1} \frac{1}{\mu_i b_i}, \qquad 0 \leqslant j \leqslant n, \\ f_j &= \sum_{i=m}^{j-1} v_i + 1 = \mu_m b_m v_m \sum_{i=m}^{j-1} \frac{1}{\mu_i b_i} + 1, \qquad m \leqslant j \leqslant N. \end{aligned}$$

On the other hand, since $f_n = 1$ and $v_N = f_{N+1} - f_N = -f_N$, we get

$$1 = \mu_0 b_0 v_0 \sum_{i=0}^{n-1} \frac{1}{\mu_i b_i}, \qquad \frac{\mu_m b_m v_m}{\mu_N b_N} = -\mu_m b_m v_m \sum_{i=m}^{N-1} \frac{1}{\mu_i b_i} - 1.$$

Then

$$\mu_0 b_0 v_0 = \bigg(\sum_{i=0}^{n-1} \frac{1}{\mu_i b_i} \bigg)^{-1}, \qquad \mu_m b_m v_m = -\bigg(\sum_{i=m}^{N} \frac{1}{\mu_i b_i} \bigg)^{-1}. \tag{4.3}$$

Inserting (4.2) and (4.3) into (4.1), we obtain

$$\begin{aligned}D(f) &= \sum_{i=0}^{n-1} \mu_i b_i v_i^2 + \sum_{i=m}^{N} \mu_i b_i v_i^2\\ &= (\mu_0 b_0 v_0)^2 \sum_{i=0}^{n-1} \frac{1}{\mu_i b_i} + (\mu_m b_m v_m)^2 \sum_{i=m}^{N} \frac{1}{\mu_i b_i}\\ &= \bigg(\sum_{i=0}^{n-1} \frac{1}{\mu_i b_i}\bigg)^{-1} + \bigg(\sum_{i=m}^{N} \frac{1}{\mu_i b_i}\bigg)^{-1}.\end{aligned}$$

Since the process is recurrent, $\sum_{i=m}^{\infty} 1/\mu_i b_i = \infty$, we have

$$\mathrm{Cap}(K) = \inf\{D(f), f_0 = 0, f \text{ has finite support}, f|_K \geqslant 1\} = \bigg(\sum_{i=0}^{n-1} \frac{1}{\mu_i b_i}\bigg)^{-1},$$

which is independent of m. Therefore

$$B_{\mathbb{B}} = \sup_K \frac{\|I_K\|_{\mathbb{B}}}{\mathrm{Cap}(K)} = \sup_{1\leqslant n\leqslant m} \frac{\|I_{[n,m]}\|_{\mathbb{B}}}{\mathrm{Cap}([n,m])} = \sup_{n\geqslant 1} \|I_{[n,\infty)}\|_{\mathbb{B}} \sum_{i=0}^{n-1} \frac{1}{\mu_i b_i}$$

as required. $\square$

We remark that once we know the solution f that minimizes $D(f)$, the proof (c) above can be done in a different way as illustrated in the next proof.

Corollary 4.2. Consider an ergodic diffusion on $(0, \infty)$ with operator

$$L = a(x)\mathrm{d}^2/\mathrm{d}x^2 + b(x)\mathrm{d}/\mathrm{d}x$$

and reflecting boundary. Suppose that the corresponding Dirichlet form $(D, \mathscr{D}(E))$ is regular, having the core $\mathbb{C}_d[0, \infty)$: the set of all continuous functions with piecewise continuous derivatives and having compact support. Define $C(x) = \int_0^x b/a$ for $x \geqslant 0$. Then for Dirichlet boundary at 0, we have

$$B_{\mathbb{B}} = \sup_{x\geqslant 0} \|I_{[x,\infty)}\|_{\mathbb{B}} \int_0^x e^{-C}.$$

Proof. In view of (b) in the above proof, to compute the isoperimetric constant, we need only consider the compact $K = [n, m]$, $m > n$, $m, n \in \mathbb{R}_+$. Define

$$g(x) = \begin{cases} \int_0^x e^{-C} \big/ \int_0^n e^{-C}, & \text{if } 0 \leqslant x \leqslant n,\\ 1, & \text{if } n \leqslant x \leqslant m,\\ 1 - \int_m^{x\wedge N} e^{-C} \big/ \int_m^N e^{-C}, & \text{if } x \geqslant m.\end{cases}$$

We now show that $\mathrm{Cap}(K)$ can be computed in terms of $g \in \mathbb{C}_d[0, \infty)$. Note that

$$\mathrm{Cap}(K) = \inf\{D(f) : f \in \mathbb{C}_d[0, \infty) : f|_K = 1\}.$$

Next, let $f_1 \in \mathbb{C}_d[0,n]$ with $f_1(0) = f_1(n) = 0$, $f_2 \in \mathbb{C}_d[m,N]$ with $f_2(m) = f_2(N) = 0$, and study the following variational problem with respect to ε_1 and ε_2:

$$H(\varepsilon_1,\varepsilon_2) = \int_0^n (g' + \varepsilon_1 f_1')^2 e^C + \int_m^N (g' + \varepsilon_2 f_2')^2 e^C.$$

If necessary, one may regard $\int_0^n$ as $\int_{0+}^{n-}$ and similarly for $\int_m^N$. Without loss of generality, assume that $f_k' \neq 0$. Otherwise, we can set $\varepsilon_k = 0$. Clearly, H should have a minimum in a bounded region. From $\partial H/\partial \varepsilon_k = 0$, it follows that

$$\begin{aligned}
\varepsilon_1 &= -\frac{\int_0^n g' f_1' e^C}{\int_0^n {f_1'}^2 e^C} = -\frac{\int_0^n f_1'}{\big(\int_0^n {f_1'}^2 e^C\big)\big(\int_0^n e^{-C}\big)} = -\frac{f_1(n) - f_1(0)}{\big(\int_0^n {f_1'}^2 e^C\big)\big(\int_0^n e^{-C}\big)} = 0,\\
\varepsilon_2 &= -\frac{\int_m^N g' f_2' e^C}{\int_m^N {f_2'}^2 e^C} = -\frac{\int_m^N f_2'}{\big(\int_m^N {f_2'}^2 e^C\big)\big(\int_m^N e^{-C}\big)} = -\frac{f_2(N) - f_2(m)}{\big(\int_m^N {f_2'}^2 e^C\big)\big(\int_m^N e^{-C}\big)} = 0.
\end{aligned}$$

More precisely, if f' is discontinuous at $n_1, \ldots, n_k$, then

$$\int_0^n f' = \int_0^{n_1} f' + \cdots + \int_{n_k}^n f' = \big(f(n_1) - f(0)\big) + \cdots + \big(f(n) - f(n_k)\big) = f(n) - f(0) = 0,$$

since f is continuous. Thus, $H(\varepsilon_1,\varepsilon_2)$ attains its minimum

$$D(g) = \bigg(\int_0^n e^{-C}\bigg)^{-1} + \bigg(\int_m^N e^{-C}\bigg)^{-1}$$

at $\varepsilon_1 = \varepsilon_2 = 0$. Moreover, due to the recurrence, we have $\int_m^\infty e^{-C} = \infty$. Collecting these facts, we obtain $\mathrm{Cap}(K) = \big(\int_0^n e^{-C}\big)^{-1}$. The assertion now follows immediately. $\square$

Because of the linear order in the real line, it is easy to write down the explicit estimates of the logarithmic Sobolev constant A_{Log}, in terms of Theorem 1.5 and Corollaries 4.1 and 4.2.

Corollary 4.3. For ergodic birth–death processes, let m satisfy

$$\pi[0,m) := \sum_{j=0}^{m-1} \mu_j/Z \leqslant 1/2 \quad \text{and} \quad \pi(m,\infty) := \sum_{j=m+1}^{\infty} \mu_j/Z \leqslant 1/2,$$

where $Z = \sum_{k=0}^\infty \mu_k$. Then we have

$$\frac{\log 2}{\log(1+2e^2)}\, B_{\mathrm{Log}}(e^2) \leqslant B_{\mathrm{Log}}(1/2) \leqslant A_{\mathrm{Log}} \leqslant 4\, B_{\mathrm{Log}}(e^2), \tag{4.4}$$

where $B_{\mathrm{Log}}(\gamma) = B_+(\gamma) \vee B_-(\gamma)$ and

$$\begin{aligned}
B_+(\gamma) &= \sup_{n>m} \mu[n,\infty) \log\bigg(1 + \frac{\gamma}{\pi[n,\infty)}\bigg) \sum_{j=m}^{n} \frac{1}{\mu_j b_j},\\
B_-(\gamma) &= \sup_{0\leqslant n<m} \mu[0,n] \log\bigg(1 + \frac{\gamma}{\pi[0,n]}\bigg) \sum_{j=n}^{m-1} \frac{1}{\mu_j b_j}.
\end{aligned} \tag{4.5}$$

Proof. Here we prove the upper estimate only since the proof for the lower estimate is similar. Set $E_1 = \{m+1, m+2, \dots\}$ and $E_2 = \{0, \dots, m-1\}$. Following the proof (a) of Theorem 1.5, we obtain (3.10) with respect to E_1 and E_2. Applying Corollary 4.1 to each E_i, we get $B_\pm(e^2)$. We remark that in the application of Corollary 4.1 to E_1, the Dirichlet boundary is setting at m rather than at 0. In other words, we need to consider the inverse order on E_1.

For one-dimensional diffusion, a similar result of Corollary 4.3 was obtained by Barthe and Roberto [2].

Corollary 4.4. Let μ and ν be Borel measures on $\mathbb{R}$ with $\mu(\mathbb{R}) < \infty$ and denote by h the derivative of the the absolutely continuous part of ν with respect to the Lebesgue measure. Next, set $\pi = \mu/\mu(\mathbb{R})$ and let m be the median of π. Then the optimal constant A_{Log} in the inequality

$$\int_{\mathbb{R}} f^2 \log\big(f^2/\pi(f^2)\big)\mathrm{d}\mu \leqslant A_{\text{Log}} \int_{\mathbb{R}} f'^2 \mathrm{d}\nu, \qquad f \in \mathbb{C}_d(\mathbb{R}), \tag{4.6}$$

(cf. Corollary 4.2 for definition of $\mathbb{C}_d(\mathbb{R})$) satisfies

$$\frac{\log 2}{\log(1+2e^2)}\, B_{\text{Log}}(e^2) \leqslant B_{\text{Log}}(1/2) \leqslant A_{\text{Log}} \leqslant 4\, B_{\text{Log}}(e^2), \tag{4.7}$$

with $B_{\text{Log}}(\gamma) = B_+(\gamma) \vee B_-(\gamma)$, where

$$\begin{aligned} B_+(\gamma) &= \sup_{x>m} \mu[x,\infty) \log\left(1 + \frac{\gamma}{\pi[x,\infty)}\right) \int_m^x \frac{1}{h}, \\ B_-(\gamma) &= \sup_{x<m} \mu(-\infty,x] \log\left(1 + \frac{\gamma}{\pi(-\infty,x]}\right) \int_x^m \frac{1}{h}. \end{aligned} \tag{4.8}$$

Actually, Corollaries 4.3 and 4.4 can be further improved by using the variational formulas presented in Chen [5, 6].

Acknowledgements. A large part of this paper is completed during the author's visit to Japan (Oct. 16, 2002–Jan. 17, 2003). The author is grateful for the kind invitations, financial supports and the warm hospitality made by many Japanese probabilists and their universities: I. Shigekawa, Y. Takahashi, T. Kumagai and N. Yosida at Kyoto Univ.; M. Fukushima, S. Kotani and S. Aida at Osaka Univ.; H. Osada, S. Liang at Nagoya Univ.; T. Funaki and S. Kusuoka at Tokyo Univ. A special appreciation is given to the senior N. Ikeda, S. Watanabe, K. Sato and my Chinese friend Mr. Q. P. Liu at Kyoto Univ. Moreover, the author would like to thank M. Fukushima for providing the preprint by him and Uemura [10] before publication.

Finally, the author deeply acknowledges the referees for their kind and crucial comments on an earlier version of the paper.

References

[1]. Barthe, F., Cattiaux, P. and Roberto, C. (2004), *Interpolated inequalities between exponential and Guassian, Orlicz hypercontractivity and isoperimetry*, arXiv:math.PR /0407219.

[2]. Barthe, F. and Roberto, C. (2003), *Sobolev inequalities for probability measures on the real line*, Studia Math. **159**(3), 481–497.

[3]. Bobkov, S. G. and Götze, F. (1999), *Exponential integrability and transportation cost related to logarithmic Sobolev inequalities*, J. Funct. Anal., **163**, 1–28.

[4]. Chen, M. F. (2000), *Explicit bounds of the first eigenvalue*, Sci. in China, Ser A, **43**(10), 1051–1059.

[5]. Chen, M. F. (2002), *Variational formulas of Poincaré-type inequalities in Banach spaces of functions on the line*, Acta Math. Sin. Eng. Ser. **18**(3), 417–436.

[6]. Chen, M. F. (2003), *Variational formulas of Poincaré-type inequalities for birth–death processes*, Acta Math. Sin. Eng. Ser **19**(4), 625–644.

[7]. Chen, M. F. and Wang, F. Y. (1998), *Cheeger's inequalities for general symmetric forms and existence criteria for spectral gap*, Abstract. Chin. Sci. Bulletin **43**(18), 1516–1519. Ann. Prob. 2000, **28**(1), 235–257.

[8]. Deuschel, J. D. and Stroock, D. W. (1989), *Large Deviations*, Academic Press, New York.

[9]. Fukushima, M., Oshima, Y. and Takeda, M. (1994), *Dirichlet forms and symmetric Markov processes*, Walter de Gruyter.

[10]. Fukushima, M. and Uemura, T. (2003), *Capacitary bounds of measures and ultracontractivity of time changed processes*, J. Math. Pure et Appliquees **82**(5), 553-572.

[11]. Gong, F. Z. and Wang, F. Y. (2002), *Functional inequalities for uniformly integrable semigroups and application to essential spectrums*, Forum. Math. **14**, 293–313.

[12]. Gross, L. (1993), *Logarithmic Sobolev inequalities and contractivity of semigroups*, LNM. **1563**.

[13]. Kaimanovich, V. A. (1992), *Dirichlet norms, capacities and generalized isoperimetric inequalities for Markov operators*, Potential Analysis **1**, 61-82.

[14]. Ma, Z. M. and Röckner, M. (1992), *Introduction to the Theory of* (*non-symmetric*) *Dirichlet Forms*, Springer.

[15]. Mao, Y. H. (2002), *The logarithmic Sobolev inequalities for birth–death process and diffusion process on the line*, Chin. J. Appl. Prob. Statis. **18**(1), 94–100.

[16]. Mao, Y. H. (2002), *Nash inequalities for Markov processes in dimension one*, Acta Math. Sin. Eng. Ser. **18**(1), 147–156.

[17]. Maz'ya, V. G. (1985), *Sobolev Spaces*, Springer.

[18]. Muckenhoupt, B. (1972), *Hardy's inequality with weights*, Studia Math. **XLIV**, 31–38.

[19]. Opic, B. and Kufner, A. (1990), *Hardy-type Inequalities*, Longman, New York.

[20]. Rao, M. M. and Ren, Z. D. (1991), *Theory of Orlicz Spaces*, Marcel Dekker, Inc New York.

[21]. Rothaus, O. S. (1985), *Analytic inequalities, isoperimetric inequalities and logarithmic Sobolev inequalities*, J. Funct. Anal. **64**, 296–313.

[22]. Vondraček, Z. (1996), *An estimate for the L^2-norm of a quasi continuous function with respect to a smooth measure*, Arch. Math. **67**, 408-414.

[23]. Wang, F. Y. (2000), *Functional inequalities for empty essential spectrum*, J. Funct. Anal. **170**, 219–245.

School of Mathematical Sciences, Beijing Normal University, Beijing 100875, The People's Republic of China. E-mail: mfchen@bnu.edu.cn
Home page: http://math.bnu.edu.cn/~chenmf/main_eng.htm

Front. Math. China 2007, 2(3), 329–358

Exponential Convergence Rate in Entropy

Mu-Fa Chen

Beijing Normal University, Beijing 100875, China

Abstract

The exponential convergence rate in entropy is studied for symmetric forms, with a special attention to the Markov chain with a state space having two point only. Some upper and lower bounds of the rate are obtained and five examples with precise or qualitatively exact estimates are presented.

Keywords Exponential convergence rate, spectral gap, logarithmic Sobolev constant, symmetric forms, Markov chains
MSC 60J25, 37A35

1 Introduction

Let $(E, \mathscr{E}, \pi)$ be a probability space satisfying $\{(x,x): x \in E\} \in \mathscr{E} \times \mathscr{E}$. Denote by $L^2(\pi)$ the usual real L^2-space with norm $\|\cdot\|$. Consider a *symmetric form* $(D, \mathscr{D}(D))$ (not necessarily a Dirichlet form) on $L^2(\pi)$:

$$\begin{aligned} D(f) &= \frac{1}{2}\int_{E\times E} J(\mathrm{d}x, \mathrm{d}y)[f(y) - f(x)]^2, \\ \mathscr{D}(D) &= \{f \in L^2(\pi) : D(f) < \infty\}, \end{aligned} \tag{1}$$

where $J \geqslant 0$ is a symmetric measure, having no charge on the diagonal set $\{(x,x): x \in E\}$. A typical example is as follows. For a nonnegative kernel $q(x, \mathrm{d}y)$ satisfying $q(x, E \setminus \{x\}) < \infty$ for all $x \in E$, reversible with respect to π (i.e., $\pi(\mathrm{d}x)q(x, \mathrm{d}y) = \pi(\mathrm{d}y)q(y, \mathrm{d}x)$), we simply take

$$J(\mathrm{d}x, \mathrm{d}y) = \pi(\mathrm{d}x)q(x, \mathrm{d}y).$$

More especially, for a Q-matrix $(q_{ij} : i, j \in E)$, reversible with respect to $(\pi_i > 0)$ (i.e., $\pi_i q_{ij} = \pi_j q_{ji}$ for all i, j), we take $J_{ij} = \pi_i q_{ij}$ $(j \neq i)$ as the density of the symmetric measure $J(\mathrm{d}x, \mathrm{d}y)$ with respect to the counting measure.

Received March 6, 2007; accepted April 14, 2007.

The exponential convergence rate we are interested in is defined as follows:

$$\alpha_1 = \inf\left\{\frac{D(f,\log f)}{2\,\mathrm{Ent}(f)} : f \geqslant 0,\ 0 < \mathrm{Ent}(f) < \infty\right\}, \tag{2}$$

where

$$\mathrm{Ent}(f) = \pi\left(f\log\frac{f}{\pi(f)}\right), \qquad \pi(f) = \int f\mathrm{d}\pi.$$

Instead of α_1, sometimes we write $\alpha_1(Q)$ if J is determined by a Q-matrix. In particular, when $E = \{0,1\}$ and

$$Q = \begin{pmatrix} -\theta & \theta \\ 1-\theta & \theta - 1\end{pmatrix}, \qquad \theta \in (0,1), \tag{3}$$

we have

$$\pi_0 = 1-\theta, \quad \pi_1 = \theta, \quad J_{01} = J_{10} = \theta(1-\theta).$$

Then we write $\alpha_1(\theta)$ rather than $\alpha_1(Q)$ throughout this paper. Clearly, we have $\alpha_1(\theta) = \alpha_1(1-\theta)$. Even for this simplest situation, the precise value of $\alpha_1(\theta)$ is still unknown and non-trivial. We will come back to this case later.

It is well known that entropy plays a crucial role in many fields such as physics and information theory. The importance of the constant α_1 is due to the fact that it describes the convergence rate of the semigroup $\{P_t\}_{t\geqslant 0}$ determined by the symmetric form $(D, \mathscr{D}(D))$:

$$\mathrm{Ent}(P_t f) \leqslant \mathrm{Ent}(f)e^{-2\alpha_1 t}, \qquad t \geqslant 0,\ f \geqslant 0 \tag{4}$$

(cf. [17; Theorem 1.1]).

Up to now, our knowledge about α_1 is still quite limited. To state a general result, we need two related constants: the spectral gap (or the first non-trivial eigenvalue) λ_1 and the logarithmic Sobolev constant σ defined as follows:

$$\lambda_1 = \inf\left\{\frac{D(f)}{\mathrm{Var}(f)} : f \in L^2(\pi), f \neq \mathrm{const}\right\}, \tag{5}$$

$$\sigma = \inf\left\{\frac{2D(f)}{\mathrm{Ent}(f^2)} : f \geqslant 0,\ 0 < \mathrm{Ent}(f^2) < \infty\right\}, \tag{6}$$

where $\mathrm{Var}(f)$ is the variation of f with respect to π:

$$\mathrm{Var}(f) = \pi\big(f^2\big) - \pi(f)^2.$$

The main known result proved in [17; Theorem 1.2] says that

$$\sigma \leqslant \alpha_1 \leqslant \lambda_1. \tag{7}$$

This provides us some sharp results whenever $\sigma = \lambda_1$, refer to [1]. In the context of diffusions, it is known that $\sigma = \alpha_1$, so we need only to consider the jump type of symmetric forms defined by (1). In the simplest case of (3), it is known that

$$\lambda_1 = 1, \qquad \sigma = \frac{2(1-2\theta)}{\log(\theta^{-1}-1)}. \tag{8}$$

The first result in (8) is easy but the second one is not so, for which there are several different proofs, refer to [12], [2], [6] and [1]. We will discuss them again in §5.2. For α_1, no precise result is known but the following estimate

$$\frac{1}{2}\big(\sqrt{\theta}+\sqrt{1-\theta}\big)^2 \leqslant \alpha_1 \leqslant \lambda_1 \tag{9}$$

given in [14; Proposition 3.1]. Our first goal is to improve (9)(cf. Lemma 5.4 in §5).

Theorem 1.1 For $\alpha_1(\theta),\ \theta\in[0,1/2]$, we have

$$\frac{1}{2}\left[1+\left(\frac{\theta(1-\theta)}{2}\right)^{1/3}+\frac{1-2\theta}{\log(\theta^{-1}-1)}\right] \leqslant \alpha_1 \leqslant \frac{1}{2}\left[1+\frac{2(1-2\theta)}{\log(\theta^{-1}-1)}\right]. \tag{10}$$

It is interesting to note that the upper bound in (10) is simply the mean of λ_1 and σ, and for the lower bound, a half of σ is replaced by $(\theta(1-\theta)/2)^{1/3}$. We are happy for such simple and explicit formulas. Moreover, the sign of the equalities hold if and only if at the end points $\theta=0$ and $1/2$. A numerical computation shows that the error for the upper bound is $\leqslant 0.0262$ and the one for the lower bounds is $\leqslant 0.01$.

As applications of Theorem 1.1, we have the following two results.

Theorem 1.2 For the symmetric form given in (1), we have the following upper estimates.

(1) The exponential convergence rate α_1 in entropy is bounded from above by

$$\begin{aligned}&\inf_{A:\,\pi(A)\in(0,1/2]}\alpha_1\big(\pi(A)\big)\frac{J(A\times A^c)}{\pi(A)\pi(A^c)}\\ &\qquad\leqslant \inf_{A:\,\pi(A)\in(0,1/2]}\left[\frac{1}{2}+\frac{\pi(A^c)-\pi(A)}{\log\big(\pi(A^c)/\pi(A)\big)}\right]\frac{J(A\times A^c)}{\pi(A)\pi(A^c)}.\end{aligned}$$

(2) The logarithmic Sobolev constant σ is bounded from above by

$$2\inf_{A:\,\pi(A)\in(0,1/2]}\frac{\big(\pi(A^c)-\pi(A)\big)J(A\times A^c)}{\pi(A)\pi(A^c)\log\big(\pi(A^c)/\pi(A)\big)}.$$

All of the formulas are symmetric in A and A^c. Hence one may replace $\inf_{A:\,\pi(A)\in(0,1/2]}$ with $\inf_{A:\,\pi(A)\in[1/2,1)}$ or $\inf_{A:\,\pi(A)\in(0,1)}$.

Theorem 1.3 Let (π_i) be a positive probability measure on a finite set E. Suppose that a reversible Q-matrix (q_{ij}) satisfies

$$c_* := \min_{i,j: i\neq j} q_{ij}/\pi_j > 0.$$

Then we have

$$\alpha_1(Q) \geqslant c_*\, \alpha_1(\pi_*) \geqslant \frac{c_*}{2}\left[1 + \left(\frac{\pi_*(1-\pi_*)}{2}\right)^{1/3} + \frac{1-2\pi_*}{\log(1/\pi_* - 1)}\right],$$

where $\pi_* = \min_i \pi_i$. The sign of the first equality holds provided $q_{ij} = \pi_j$ for all $i \neq j$. Similarly, we have

$$\alpha_1(Q) \leqslant c^*\, \alpha_1(\pi_*) \leqslant \frac{c^*}{2}\left[1 + \frac{2(1-2\pi_*)}{\log(1/\pi_* - 1)}\right],$$

where

$$c^* := \max_{i,j: i\neq j} q_{ij}/\pi_j.$$

Part (2) of Theorem 1.2 improves the upper bound

$$2 \inf_{A:\, \pi(A)\in(0,1)} \frac{J(A\times A^c)}{-\pi(A)\log \pi(A)}$$

given in [8; Theorem 1.1]. To see this, first note that

$$(1-x)\log\frac{1}{1-x} \geqslant x \log \frac{1}{x}, \qquad x \in [1/2, 1],$$

Hence

$$\inf_{A:\, \pi(A)\in(0,1)} \frac{J(A\times A^c)}{-\pi(A)\log \pi(A)} = \inf_{A:\, \pi(A)\in(0,1/2]} \frac{J(A\times A^c)}{-\pi(A)\log \pi(A)}.$$

Next note that

$$\frac{1-2x}{x(1-x)\log(1/x-1)} \leqslant \frac{1}{x\log(1/x)} \qquad x\in[0,1/2].$$

From these facts we obtain the required assertion.

The proofs of Theorems 1.2 and 1.3 are given in §2 and §4, respectively. In §3, we study the upper estimate by a different approach (Theorem 3.1). To illustrate the power of the above results, five examples are treated successively in §2-§3 and are summarized as follows.

Proposition 1.4 Five models for the first non-trivial eigenvalue λ_1, the exponential convergence rate α_1 in entropy, and the logarithmic Sobolev constant σ are given in the Table 1. The results for the first two models are precise, for the third one is approximately sharp, and for the last two models are qualitatively exact. Except the first one, the others are all birth–death processes with Q-matrix:

$$q_{i,i+1} = b_i\,(i\geqslant 0), \quad q_{i,i-1} = a_i\,(i \geqslant 1), \quad q_{ij} = 0\,(j\neq i, i\pm 1).$$

Table 1

	$\boldsymbol{\lambda}_1$	$\boldsymbol{\alpha}_1$	$\boldsymbol{\sigma}$
$J_{ij} = \pi_i \pi_j$ $j \neq i$	1	$1/2$	0
$b_i \equiv b$ $a_i \equiv a$ $(a > b)$	$\big(\sqrt{a} - \sqrt{b}\big)^2$	0	0
$b_i \equiv b$ $a_i = i$	1	$1/2$	0
$b_0 = 1$ $b_i = a_i$ $= i^\gamma$ $(\gamma > 1)$	> 0 iff $\gamma \geqslant 2$	> 0 iff $\gamma > 2$	> 0 iff $\gamma > 2$
$b_0 = 1$ $b_i = a_i$ $= i^2 \log^\gamma (i+1)$ $(\gamma \in \mathbb{R})$	> 0 for all $\gamma \in \mathbb{R}$	> 0 iff $\gamma \geqslant 1$	> 0 iff $\gamma \geqslant 1$

The proof of Theorem 1.1 is much more technical and so is delayed to §5.3. Actually, the proof for the lower estimate is computer aided. In §5.3, a general result for the state space with two points is also studied. The monotonicity of $\alpha_1(\theta)$ in θ is proved in §5.1. In §5.2, we discuss the related eigenequations and the difficulty of our problem.

For the background of the study on these topics and much more related results, refer to [10].

2 Explicit estimate for the upper bound

The upper bounds given in Theorem 1.2 depend only on the probability measure π and the symmetric measure J, and so are said to be "explicit".

In this section, we first present some applications of Theorem 1.2, its proof is given at the end of this section.

Corollary 2.1 Let J be L^1-bounded, i.e. there exists an $M < \infty$ such that $J(A \times E) \leqslant M\pi(A)$ for all $A \in \mathscr{E}$. Suppose that E is an infinite set. Then $\sigma = 0$.

Proof. The proof is quite easy. Since

$$J(A \times A^c) \leqslant J(A, E) \leqslant M\pi(A),$$

by part (2) of Theorem 1.2, we have

$$\sigma \leqslant 2M \varliminf_{\pi(A)\to 0} \frac{1}{\pi(A^c)\log\big(\pi(A^c)/\pi(A)\big)} = 0. \qquad \square$$

The similar result is not true for α_1 as shown by the following example.

Example 2.2 Let E be countable but infinite. Given a probability measure $(\pi_i : i \in E)$, let $J_{ij} = \pi_i\pi_j$ for all $i \neq j$. Then $\lambda_1 = 1$, $\sigma = 0$ but $\alpha_1 = 1/2$. Thus, the upper bound given by Theorem 1.2 is sharp for this model.

Proof. This is a basic example for which $\lambda = 1$, $\sigma = 0$, and the process is even strongly ergodic proved in [9; Example 6] and [8; Example 1.2]. Clearly, by Corollary 2.1, we also get $\sigma = 0$. By [14; Example 1.2], we have $\alpha_1 = 1/2$. Therefore we need only to prove the last assertion. First we have

$$J(A\times A^c) = \sum_{i\in A}\sum_{j\notin A}\pi_i\pi_j = \pi(A)\pi(A^c).$$

Next, when $x \uparrow 1/2$,

$$h(x) := \frac{1-2x}{\log(1/x-1)} \uparrow \frac{1}{2} \tag{11}$$

and $h(0) = 0$. Thus, by part (1) of Theorem 1.2, it follows that

$$\alpha_1 \leqslant \inf_{\pi(A)\in(0,1/2]}\left[\frac{1}{2} + \frac{\pi(A^c)-\pi(A)}{\log\big(\pi(A^c)/\pi(A)\big)}\right] = \lim_{\pi(A)\to 0}\left[\frac{1}{2} + h(\pi(A))\right] = \frac{1}{2}. \qquad \square$$

In what followswe apply Theorem 1.2 to birth–death processes. Denote by $b_i\,(i \geqslant 0)$, and $a_i\,(i \geqslant 1)$ the birth rates and death rates, respectively. Let

$$\mu_0 = 1, \quad \mu_i = \frac{b_0\cdots b_{i-1}}{a_1\cdots a_i}, \qquad i \geqslant 1.$$

Suppose that the process is non-explosive and ergodic:

$$\sum_{k=0}^{\infty}\frac{1}{b_k\mu_k}\sum_{i=0}^{k}\mu_i = \infty, \qquad Z := \sum_i \mu_i < \infty.$$

Then we have the stationary distribution $\pi_i = \mu_i/Z$, $i \geqslant 0$. Set

$$\mu[m,n] = \sum_{i=m}^{n}\mu_i.$$

When $A = \{0, 1, \ldots, n\}$ or $\{n+1, n+2, \ldots\}$, we have $J(A\times A^c) = \pi_n b_n$. By part (1) of Theorem 1.2, we obtain immediately the following result.

Corollary 2.3 For birth–death processes, we have

$$\alpha_1 \leqslant \inf_{n\geqslant 1}\left[\frac{1}{2}+\frac{1-2\mu[n,\infty)/Z}{\log((Z/\mu[n,\infty))-1)}\right]\frac{\mu_{n-1}b_{n-1}}{(1-(\mu[n,\infty)/Z))\mu[n,\infty)}.$$

Example 2.4 Let $b_i \equiv b$ and $a_i = i$. Then $\lambda_1 = 1$, $\sigma = 0$ and $\alpha_1 = 1/2$. Corollary 2.3 gives us

$$\alpha_1 \leqslant \left[\frac{1}{2}+\frac{1-2e^{-b}}{\log(e^b-1)}\right]\frac{b}{1-e^{-b}} \to \frac{1}{2} \qquad \text{if } b\to 0.$$

Thus Corollary 2.3 and part (1) of Theorem 1.2 are approximately sharp for small b.

Proof. For this example, it is an earlier result that $\lambda_1 = 1$ proved in [3]. See also [4; Example 9.27]. By [7; Corollary 2.4], we have $\sigma = 0$. This can also be checked by using the criterion for logarithmic Sobolev inequality (see for instance [10; Theorem 1.10]). Next, it was proved in [14; Example 1.3] that $\alpha_1 = 1/2$. Thus, we need only to prove the last assertion. Note that

$$\mu_i = \frac{b^i}{i!}, \qquad Z = e^b.$$

By Corollary 2.3, we have

$$\alpha_1 \leqslant \left[\frac{1}{2}+\frac{1-2\mu[1,\infty)/Z}{\log((Z/\mu[1,\infty))-1)}\right]\frac{\mu_0 b_0}{(1-(\mu[1,\infty)/Z))\mu[1,\infty)}.$$

The assertion now follows by a simple computation. $\square$

Remark 2.5 By (11), it follows that

$$\frac{J(A\times A^c)}{2\pi(A)\pi(A^c)} \leqslant \left[\frac{1}{2}+\frac{\pi(A^c)-\pi(A)}{\log\big(\pi(A^c)/\pi(A)\big)}\right]\frac{J(A\times A^c)}{\pi(A)\pi(A^c)} \leqslant \frac{J(A\times A^c)}{\pi(A)\pi(A^c)}. \tag{12}$$

Thus, the upper bound in part (1) of Theorem 1.2 vanishes if and only if

$$\inf_{A:\,\pi(A)\in(0,1/2]}\frac{J(A\times A^c)}{\pi(A)\pi(A^c)} = 0.$$

If so, by [13] or [11], we have $\lambda_1 = 0$ and hence $\alpha_1 = 0$, $\sigma = 0$. Therefore, from the qualitative point of view, the upper bounds given in Theorem 1.2 is rough. This is natural since we have used only the information from "two points". However, as shown by the above two examples, the upper estimates are still meaningful for the quantitative estimation. Similar to "$\sigma \leqslant \alpha_1 \leqslant \lambda_1$", our upper estimates for these constants also obey the following relation

$$\begin{aligned}&\frac{2(\pi(A^c)-\pi(A))J(A\times A^c)}{\pi(A)\pi(A^c)\log\big(\pi(A^c)/\pi(A)\big)}\\ &\qquad\leqslant \left[\frac{1}{2}+\frac{\pi(A^c)-\pi(A)}{\log\big(\pi(A^c)/\pi(A)\big)}\right]\frac{J(A\times A^c)}{\pi(A)\pi(A^c)} \leqslant \frac{J(A\times A^c)}{\pi(A)\pi(A^c)}.\end{aligned}$$

Example 2.6 Consider a birth–death process with rates $b_0 = 1$ and $b_i = a_i$ $(i \geqslant 1)$ satisfying
$$Z := 1 + \sum_{i \geqslant 1} \frac{1}{a_i} < \infty.$$
Then the upper bound given in part (1) of Theorem 1.2 is bigger or equal to $2/Z$. However, when
$$b_i = a_i = i^\gamma \, (i \geqslant 1, \gamma > 1),$$
we have $\lambda_1 = \alpha_1 = 0$ for all $\gamma \in (1, 2)$.

Proof. Clearly,
$$\pi_i = \frac{1}{a_i Z}, \qquad J_{i,i+1} = \frac{1}{Z} \, (i \geqslant 0)$$
and $J_{ij} = 0$ for other $i \neq j$. Thus, $J(A \times A^c) = 1/Z$ if and only if A has the form $\{0, 1, \ldots, n\}$ or $\{n+1, n+2, \ldots\}$. Otherwise $J(A \times A^c) \geqslant 2/Z$. Next, $(1 - \pi(A))\pi(A) \leqslant 1/4$ and the equality holds if and only if $\pi(A) = 1/2$. Hence for all $A \neq \emptyset$, we have
$$\frac{J(A \times A^c)}{\pi(A)\pi(A^c)} \geqslant \frac{1}{Z/4} = 4/Z.$$
By (12) and part (1) of Theorem 1.2, we obtain the first assertion.

For the particular case, by the remark after [5; Corollary 1.9], we know that $\lambda_1 > 0$ if and only if $\gamma \geqslant 2$. Thus, when $\gamma \in (1, 2)$, we have $\lambda_1 = 0$ and hence $\alpha_1 = 0$. In the next section, we will prove that $\alpha_1 > 0$ if and only if $\gamma > 2$ (Example 3.3). $\square$

Example 2.7 Consider a birth–death process with rates $b_i \equiv b$ and $a_i \equiv a$, $a > b$. Then $\lambda_1 = \left(\sqrt{a} - \sqrt{b}\right)^2$ and $\alpha_1 = 0$. But the upper bound given by Corollary 2.3 is $(a - b)/2$.

Proof. The result about λ_1 is due to [3] and the one about α_1 is due to [14; Example 1.4]. For this example, we have
$$\mu_i = \left(\frac{b}{a}\right)^i, \qquad Z = \frac{a}{a - b},$$
and so
$$\mu[n, \infty) = Z\left(\frac{b}{a}\right)^n.$$
Hence,
$$\begin{aligned}
&\left[\frac{1}{2} + \frac{1 - 2\mu[n,\infty)/Z}{\log((Z/\mu[n,\infty)) - 1)}\right] \frac{\mu_{n-1} b_{n-1}}{(1 - (\mu[n,\infty)/Z))\mu[n,\infty)} \\
&= \left[\frac{1}{2} + \frac{1 - 2(b/a)^n}{\log((a/b)^n - 1)}\right] \frac{(b/a)^{n-1} b}{(1 - (b/a)^n) Z (b/a)^n} \\
&= (a - b)\left[\frac{1}{2} + \frac{1 - 2(b/a)^n}{\log((a/b)^n - 1)}\right] \frac{1}{1 - (b/a)^n}.
\end{aligned}$$

The right-hand side is decreasing in n, and so the assertion follows. $\square$

In the next section, we will study again the last two examples.

Proof of Theorem 1.2. Let $A \in \mathscr{E}$ with $\pi(A) \in [1/2, 1)$. Regarding the partition $\{A\}$ and $\{A^c\}$ of E as "two points", consider the "two points" function

$$f = f_0 1_A + f_1 1_{A^c}, \qquad f_0, f_1 \geqslant 0.$$

Then

$$\begin{aligned} D(f, \log f) &= \int_{A\times A^c} J(\mathrm{d}x, \mathrm{d}y)\big(f(y) - f(x)\big)\big(\log f(y) - \log f(x)\big) \\ &= \frac{J(A\times A^c)}{\pi(A)\pi(A^c)}\,\theta(1-\theta)(f_1 - f_0)(\log f_1 - \log f_0), \end{aligned}$$

where $\theta = \pi(A^c)$. Next, we have

$$\mathrm{Ent}(f) = (1-\theta) f_0 \log \frac{f_0}{(1-\theta)f_0 + \theta f_1} + \theta f_1 \log \frac{f_1}{(1-\theta)f_0 + \theta f_1}.$$

Thus, corresponding to the "two points" situation, for arbitrarily $\varepsilon > 0$, there exists a "two points" function

$$f^{\varepsilon,\theta} = f_0^{\varepsilon,\theta} 1_A + f_1^{\varepsilon,\theta} 1_{A^c}$$

such that

$$\frac{D(f^{\varepsilon,\theta}, \log f^{\varepsilon,\theta})}{2\,\mathrm{Ent}(f^{\varepsilon,\theta})} \leqslant \big(\alpha_1(\theta) + \varepsilon\big)\,\frac{J(A\times A^c)}{\pi(A)\pi(A^c)}.$$

Making infimum with respect to $\pi(A^c) = \theta \in (0, 1/2]$, it follows that

$$\alpha_1 \leqslant \inf_{\pi(A^c)\in(0,1/2]} \big(\alpha_1\big(\pi(A^c)\big) + \varepsilon\big)\,\frac{J(A\times A^c)}{\pi(A)\pi(A^c)}.$$

Letting $\varepsilon \downarrow 0$, it follows that

$$\alpha_1 \leqslant \inf_{\pi(A^c)\in(0,1/2]} \alpha_1\big(\pi(A^c)\big)\,\frac{J(A\times A^c)}{\pi(A)\pi(A^c)}.$$

By Theorem 1.1 and the symmetry of A and A^c, we obtain

$$\begin{aligned} &\inf_{\pi(A^c)\in(0,1/2]} \alpha_1\big(\pi(A^c)\big)\,\frac{J(A\times A^c)}{\pi(A)\pi(A^c)} \\ &\qquad \leqslant \inf_{\pi(A^c)\in(0,1/2]} \left(\frac{1}{2} + \frac{\pi(A^c) - \pi(A)}{\log\big(\pi(A^c)/\pi(A)\big)}\right)\frac{J(A\times A^c)}{\pi(A)\pi(A^c)}. \end{aligned}$$

Part (1) of the theorem now follows.

The proof of part (2) of the theorem is very much the same and so is omitted. $\square$

Finally, we mention that the method using "two points" function can also be used for an upper estimate of λ_1. However, the resulting bound

$$\inf_{A:\,\pi(A)\in(0,1/2]}\frac{J(A\times A^c)}{\pi(A)\pi(A^c)}$$

is the same as in [13] or [11], mentioned in Remark 2.5. This is due to the fact that for the "two points" Q-matrix

$$Q=\begin{pmatrix}-\theta & \theta\\ 1-\theta & \theta-1\end{pmatrix},\qquad \theta\in(0,1),$$

we have $\lambda_1=1$, independent of θ.

3 Upper estimate in terms of exponentially integrable test functions

Similar to the study on spectral gap [11; Theorem 3.3] and logarithmic Sobolev inequality [15] (cf. [10; §4.7]), we use test functions to estimate the upper bound of α_1 in this section. Actually, each test function provides an upper bound. Hence what we are doing here is just making some restriction on the integrability of the test functions.

Let $r\in\mathscr{E}\times\mathscr{E}$ be a nonnegative symmetric function such that r is positive on the support of J, and moreover the measure

$$J^{(1)}(\mathrm{d}x,\mathrm{d}y):=1_{\{r(x,y)>0\}}\frac{J(\mathrm{d}x,\mathrm{d}y)}{r(x,y)}$$

satisfies

$$J^{(1)}(\mathrm{d}x,E)/\pi(\mathrm{d}x)\leqslant 1,\qquad \pi\text{-a.e.}$$

The next result is mainly used in the qualitative study, a direct computation with a carefully chosen test function can often provide more effective quantitative estimate.

Theorem 3.1 Let $\{\varphi_n\}$ be a nonnegative and non-constant sequence satisfying $\varphi_n\in L^1(\pi)$ and

$$\operatorname{ess\,sup}_J|\varphi_n(x)-\varphi_n(y)|^2r(x,y)=:M_n<\infty\qquad\text{for all } n.\tag{13}$$

Then

$$\alpha_1\leqslant\frac{1}{2}\inf_{s>0}\left[s^2\varliminf_{n\to\infty}\frac{M_n}{\log\pi(e^{s\varphi_n})-s\pi(\varphi_n)}\right].$$

Here we adopt the convention that $1/0=\infty$ and $1/\infty=0$. In particular, we have $\alpha_1=0$ provided $\varphi_n=\varphi$ is independent of n and for some $s>0$, $\pi(e^{s\varphi})=\infty$.

Proof. First we consider a fixed φ satisfying the assumptions given in the theorem. Suppose for a moment that φ is also bounded. Let $f=\exp[s\varphi]$ and $h(s)=\pi(f)$. Then

$$\begin{aligned}
D(f,\log f) &= \frac{1}{2}\int J(\mathrm{d}x,\mathrm{d}y)[f(x)-f(y)][\log f(x)-\log f(y)]\\
&= \frac{s}{2}\int J(\mathrm{d}x,\mathrm{d}y)[f(x)-f(y)][\varphi(x)-\varphi(y)]\\
&\leqslant \frac{s^2}{2}\int J(\mathrm{d}x,\mathrm{d}y)|\varphi(x)-\varphi(y)|^2[f(x)\vee f(y)]\\
&\quad \big(\text{since } \big|e^A-e^B\big|\leqslant |A-B|\big(e^A\vee e^B\big)\big)\\
&= \frac{s^2}{2}\int J^{(1)}(\mathrm{d}x,\mathrm{d}y)r(x,y)|\varphi(x)-\varphi(y)|^2[f(x)\vee f(y)]\\
&\leqslant \frac{s^2M}{2}\int J^{(1)}(\mathrm{d}x,\mathrm{d}y)[f(x)\vee f(y)]\\
&\leqslant s^2M\int J^{(1)}(\mathrm{d}x,\mathrm{d}y)f(x).
\end{aligned}$$

We obtain

$$D(f,\log f)\leqslant s^2M\pi(f)=s^2Mh(s). \tag{14}$$

Next, by Jensen's inequality, we have

$$\infty>\log\pi(e^{s\varphi})\geqslant\pi(\varphi).$$

Thus, the assertion is trivial if $\alpha_1=0$. In what follows we assume that $\alpha_1>0$. Now, by the definition of α_1 and (14), we obtain

$$\begin{aligned}
sh' &= s\pi(\varphi f)\\
&= \pi(f\log f)\\
&\leqslant \pi(f)\log\pi(f)+\frac{D(f,\log f)}{2\alpha_1}\\
&\leqslant h\log h+\frac{s^2M}{2\alpha_1}h.
\end{aligned}$$

That is

$$h'\leqslant h\bigg(\frac{sM}{2\alpha_1}+\frac{\log h}{s}\bigg)$$

or

$$(\log h)'\leqslant\frac{sM}{2\alpha_1}+\frac{\log h}{s}.$$

Applying [10; Lemma A.1] to the interval $[\varepsilon,\infty)$ and the functions

$$u(t)=\log h(t),\quad \varphi(t)=\frac{Mt}{2\alpha_1},\quad \psi(t)=\frac{1}{t},\quad G(u)=u,$$

since

$$\int_{\varepsilon}^{t}\psi(\xi)\mathrm{d}\xi=\log t-\log\varepsilon,$$

$$\int_{\varepsilon}^{t}\exp\left[-\int_{\varepsilon}^{s}\psi(\xi)\mathrm{d}\xi\right]\varphi(s)\mathrm{d}s=\frac{M\varepsilon(t-\varepsilon)}{2\alpha_1},$$

we obtain

$$u(t)\leqslant\exp[\log t-\log\varepsilon]\left[u(\varepsilon)+\frac{M\varepsilon(t-\varepsilon)}{2\alpha_1}\right]=\frac{u(\varepsilon)}{\varepsilon}t+\frac{Mt(t-\varepsilon)}{2\alpha_1}.$$

Letting $\varepsilon\to 0$ and noting that

$$\lim_{\varepsilon\to 0}\frac{u(\varepsilon)}{\varepsilon}=\pi(\varphi),$$

it follows that

$$u(t)\leqslant t\left(\frac{tM}{2\alpha_1}+\pi(\varphi)\right).$$

In other words, for bounded φ, we have

$$\log\pi(e^{s\varphi})\leqslant s\left(\frac{sM}{2\alpha_1}+\pi(\varphi)\right). \tag{15}$$

For unbounded φ, replacing φ with $\varphi\wedge m$ which also satisfies the assumptions of the theorem. Hence the last estimate (15) holds if φ is replaced by $\varphi\wedge m$. Now, the conclusion (15) holds for φ by letting $m\to\infty$. In particular, by the assumptions and the fact that $\alpha_1>0$, it follows that $\pi\big(e^{s\varphi}\big)<\infty$ for all $s>0$. This means that all exponential moments of φ are finite.

Finally, for the sequence given in the theorem, we replace φ in (15) with φ_n, at the same time replace M with M_n. With a slight change of the formula and then letting $n\to\infty$, we obtain the required assertion. □

In the remainder of this section, we apply Theorem 3.1 to birth–death processes. Recall that in this setting, $J_{i,i+1}=\pi_i b_i\,(i\geqslant 0)$ and $J_{ij}=0$ for other $i\neq j$. In what follows, let

$$r_{ij}=\begin{cases}(a_i+b_i)\vee(a_j+b_j), & \text{if } j=i+1\\ 0, & \text{otherwise.}\end{cases}$$

Example 3.2 (Continued) For the model given in Example 2.7, we have $\alpha_1=0$.

Proof. Let $\varphi_i^{(n)} = \varphi_i = i$. Then

$$\sum_i \mu_i \varphi_i = \sum_i i \left(\frac{b}{a}\right)^i < \infty,$$
$$\sup_{i \geqslant 1} r_{i,i+1}(\varphi_{i+1} - \varphi_i)^2 = a + b < \infty,$$
$$\sum_i \mu_i e^{s\varphi_i} = \sum_i \left(\frac{b}{a}\right)^i e^{si} = \infty, \qquad s \geqslant \log \frac{a}{b}.$$

The assertion now follows from Theorem 3.1. $\square$

Example 3.3 (Continued) For the birth–death process given in Example 2.6: $b_0 = 1$, $b_i = a_i = i^\gamma$ $(i \geqslant 1, \gamma > 1)$, we have $\alpha_1 > 0$ if and only if $\gamma > 2$.

Proof. From [7; Example 2.6], we know that $\sigma > 0$ if and only if $\gamma > 2$. Thus, if $\gamma > 2$, we certainly have $\alpha_1 > 0$.

When $\gamma \in (1, 2)$, it is proved in Example 2.6 that $\lambda_1 = \alpha_1 = 0$. It remains to handle with the critical case: $\gamma = 2$. But for all $\gamma \in (1, 2]$, we can still apply Theorem 3.1 to the functions $\varphi^{(n)} = \varphi$: $\varphi_0 = 0$ and $\varphi_i = \log i$:

$$\sum_{i \geqslant 0} \mu_i \varphi_i = \sum_{i \geqslant 1} \frac{\log i}{i^\gamma} < \infty,$$
$$\sup_{i \geqslant 1} r_{i,i+1}(\varphi_{i+1} - \varphi_i)^2 \leqslant \sup_{i \geqslant 1} \frac{i^\gamma}{i^2} \leqslant 1,$$
$$\sum_{i \geqslant 0} \mu_i e^{s\varphi_i} = 1 + \sum_{i \geqslant 1} \frac{i^s}{i^\gamma} = \infty, \qquad s \geqslant \gamma - 1. \qquad \square$$

Example 3.4 Consider the birth–death process with rates

$$b_0 = 1, \qquad b_i = a_i = i^2 \log^\gamma(1 + i)\, (i \geqslant 1, \gamma \in \mathbb{R}).$$

Then $\alpha_1 > 0$ if and only if $\gamma \geqslant 1$.

Proof. By [16; Example 3.1] or [7; Example 2.7], we know that $\sigma > 0$ if and only if $\gamma \geqslant 1$. If so, we have $\alpha_1 > 0$. From the last example, we also know that $\alpha_1 = 0$ if $\gamma = 0$ and so does whenever $\gamma \leqslant 0$ (this will be also proved in the next paragraph). Now, the main case we need the further study is that $\gamma \in (0, 1)$.

To apply Theorem 3.1, let $\gamma < 1$. Take $\varphi^{(n)} = \varphi$:

$$\varphi_i = \log^{1-\gamma/2}(i + 1)\, (i \geqslant 0).$$

Then

$$\sum_{i\geqslant 1}\mu_i\varphi_i=\sum_{i\geqslant 1}\frac{\log^{1-\gamma/2}(i+1)}{i^2\log^{\gamma}(i+1)}=\sum_{i\geqslant 1}\frac{1}{i^2\log^{3\gamma/2-1}(i+1)}<\infty,\qquad \gamma\in\mathbb{R},$$

$$\sup_{i\geqslant 1} r_{i,i+1}(\varphi_{i+1}-\varphi_i)^2\sim 1,\qquad \gamma\in\mathbb{R},$$

$$\begin{aligned}\sum_{i\geqslant 1}\mu_i e^{s\varphi_i}&=\sum_{i\geqslant 1}\frac{\exp[s\log^{1-\gamma/2}(i+1)]}{i^2\log^{\gamma}(i+1)}\\&=\sum_{i\geqslant 1}\exp\big[-2\log i+s\log^{1-\gamma/2}(i+1)-\gamma\log\log(i+1)\big].\end{aligned}$$

The last sum can be ∞ if and only if $\gamma\leqslant 0$ and so does with $s>2$. In other words, with this φ Theorem 3.1 is suitable for $\gamma\leqslant 0$. The last sum must be finite when $\gamma>0$. Thus, in order the last sum to be infinity, one has to increase the increasing order of φ. However, this is impossible since condition (13) controls the increasing order of φ , which is at most of $\log^{1-\gamma/2} i$. This suggests us to relax the assumption (13), using a sequence $\{\varphi^{(n)}\}$ instead of a single φ.

We now turn to prove the assertion for the case of $\gamma\in(0,1)$. From the discussion above, it is naturally to choose the test functions $\varphi_i^{(n)}=\log(i\wedge n+1)$. Then

$$M_n\sim\log^{\gamma} n,$$

$$\sum_{i\geqslant 1}\mu_i\varphi_i^{(n)}=\sum_{i\geqslant 1}\frac{\log(i\wedge n+1)}{i^2\log^{\gamma}(i+1)}\leqslant\sum_{i\geqslant 1}\frac{\log(i+1)}{i^2\log^{\gamma}(i+1)}<\infty\ \text{(independent of } n),$$

$$\sum_{i\geqslant 1}\mu_i e^{s\varphi_i^{(n)}}=\sum_{i=1}^{n}\frac{(i+1)^s}{i^2\log^{\gamma}(i+1)}+(n+1)^s\sum_{i\geqslant n+1}\frac{1}{i^2\log^{\gamma}(i+1)}\geqslant n$$

for large enough s, say $s=4$ for instance. Therefore, the leading order of

$$\frac{4M_n}{\log\pi(e^{4\varphi_n})-s\pi(\varphi_n)}$$

is controlled by

$$\frac{\log^{\gamma} n}{\log n}\to 0,\qquad n\to\infty.$$

This implies the required assertion. $\square$

4 Lower estimate

This section is devoted to the proof of Theorem 1.3.

Consider a reversible Markov chain with a finite state space E, having reversible measure $(\pi_i > 0)$ and Q-matrix $(q_{ij} : i, j \in E)$. Then we have the generator

$$\Omega f(i) = \sum_{j \in E} q_{ij}(f_j - f_i).$$

It corresponds to a Dirichlet form as follows:

$$D(f) = \frac{1}{2} \sum_{i,j \in E} \pi_i q_{ij} (f_j - f_i)^2.$$

Proof of Theorem 1.3. Except the increasing property of $\alpha_1(\theta)$ in $\theta \in (0, 1/2]$, the proof is very much the same as that of [2; Theorem A.1].

By a comparison of the Dirichlet forms, it suffices to consider the case that $q_{ij} = \pi_j$ for all $i \neq j$. Next, by [1; Theorem 6.5], there is a positive, non-constant solution f to the equation

$$-\Omega f - f\Omega \log \frac{f}{\pi(f)} = 2\alpha_1(Q) f \log \frac{f}{\pi(f)},$$

where

$$\pi(f) = \int_E f \mathrm{d}\pi.$$

Because $q_{ij} = \pi_j$ for all $i \neq j$, we can rewrite the equation as follows

$$(2\alpha_1(Q) - 1) f \log f = \big[1 - \pi(\log f) - 2\alpha_1(Q) \log \pi(f)\big] f - \pi(f).$$

Since the function $x \log x$ is convex on $[0, \infty)$ and a straight line can intersect the graph of $x \log x$ in at most two points, it follows that there are at most two values of f. Denote by x_0 and x_1 these values. Set

$$E_0 = \{i : f_i = x_0\}, \qquad E_1 = \{i : f_i = x_1\}$$

and let $\theta = \pi(E_0)$. Without loss of generality, assume that $\theta \in (0, 1/2]$. We thus decompose the state E into two parts E_0 and E_1, on each of them, f is a constant. In other words, we have reduced the problem to the special case that the state space consists of two points only. In view of Theorem 1.1, for the first assertion, what remainder now is to show that

$$\inf_{\theta \in (\pi_*, 1/2]} \alpha_1(\theta) = \alpha_1(\pi_*).$$

This is a consequence of Proposition 5.1 in the next section.

The proof for the last assertion of the theorem is again an application of Theorem 1.1. $\square$

5 The property and estimation of $\alpha_1(\theta)$

This section consists of three subsections. First, we study the monotonicity of $\alpha_1(\theta)$ in θ. Then we discuss the eigenequations corresponding to λ_1, σ and α_1, respectively. Finally, we prove Theorem 1.1. A description of $\alpha_1(\theta)$ is given by Proposition 5.5.

5.1 The monotonicity of $\alpha_1(\theta)$ in θ

Proposition 5.1 The rate $\alpha_1(\theta)$ is increasing in $\theta \in (0, 1/2]$.

Proof. 1) Recall that for the Q-matrix

$$Q = \begin{pmatrix} -\theta & \theta \\ 1-\theta & \theta-1 \end{pmatrix}, \qquad \theta \in (0, 1/2],$$

we have $\pi_0 = 1-\theta,\ \pi_1 = \theta$ and

$$\begin{aligned} D(f, \log f) &= \theta(1-\theta)(f_1 - f_0)(\log f_1 - \log f_0), \\ \mathrm{Ent}(f) &= \pi_0 f_0 \log \frac{f_0}{\pi_0 f_0 + \pi_1 f_1} + \pi_1 f_1 \log \frac{f_1}{\pi_0 f_0 + \pi_1 f_1}. \end{aligned}$$

Set

$$f_1 = \frac{x}{\theta}, \qquad f_0 = \frac{1-x}{1-\theta}.$$

Then $\pi(f) = 1$,

$$\begin{aligned} D(f, \log f) &= (1-\theta)\theta(f_1 - f_0)(\log f_1 - \log f_0) \\ &= (1-\theta)\theta\left(\frac{x}{\theta} - \frac{1-x}{1-\theta}\right)\log\left(\frac{x}{\theta}\Big/\frac{1-x}{1-\theta}\right) \\ &= (x-\theta)\log\frac{x(1-\theta)}{\theta(1-x)}, \\ \mathrm{Ent}(f) = \pi_0 f_0 \log f_0 + \pi_1 f_1 \log f_1 &= (1-x)\log\frac{1-x}{1-\theta} + x\log\frac{x}{\theta}. \end{aligned}$$

Hence,

$$\begin{aligned} \frac{D(f, \log f)}{\mathrm{Ent}(f)} &= \left[(x-\theta)\log\frac{x(1-\theta)}{\theta(1-x)}\right]\Big/\left[(1-x)\log\frac{1-x}{1-\theta} + x\log\frac{x}{\theta}\right] \\ &= 1 - \left[(1-\theta)\log\frac{1-x}{1-\theta} + \theta\log\frac{x}{\theta}\right]\Big/\left[(1-x)\log\frac{1-x}{1-\theta} + x\log\frac{x}{\theta}\right] \\ &=: 1 + h(x, \theta). \end{aligned}$$

We obtain

$$2\alpha_1 = 1 + \inf_{x \in (0,1)} h(x, \theta). \tag{16}$$

Thus, we need to show that $\inf_{x\in(0,1)} h(x,\theta)$ is increasing in $\theta \in (0, 1/2]$.

2) We now prove that

$$\inf_{x\in(0,1)} h(x,\theta) = \min_{x\in(\theta,1-\theta)} h(x,\theta), \qquad \theta \in (0, 1/2]. \tag{17}$$

Since

$$\lim_{x\to\theta} h(x,\theta) = 1 = \lim_{x\to 1-\theta} h(x,\theta),$$

it suffices to show that $h(x,\theta) \geqslant 1$ for all $x \in (0,\theta)$ or $x \in (1-\theta, 1)$. Note that the first two derivatives in x of the function

$$j_\theta(x) := x\log\frac{x}{\theta} + (1-x)\log\frac{1-x}{1-\theta} \tag{18}$$

are as follows

$$\log\frac{x(1-\theta)}{\theta(1-x)}, \qquad \frac{1}{x(1-x)}.$$

It follows that the function j_θ is convex on $[0,1]$ with minimum 0 at θ. In other words, the function is positive on $[0,1]$ with an exception at the point $x=\theta$, at which the function has value zero. Thus, to prove (17), it is enough to show that

$$-\theta\log\frac{x}{\theta} - (1-\theta)\log\frac{1-x}{1-\theta} \geqslant x\log\frac{x}{\theta} + (1-x)\log\frac{1-x}{1-\theta}, \tag{19}$$

for $x \in (0,\theta)$ or $x \in (1-\theta, 1)$. To do so, define

$$g(x,\theta) = (x+\theta-2)\log\frac{1-x}{1-\theta} - (x+\theta)\log\frac{x}{\theta}, \qquad x\in(0,1),\ \theta\in(0,1/2).$$

In what follows, we will use the following simple result repeatedly.

Lemma 5.2 Let $\xi \in C^1[p,q]$.

(1) If $\xi' > 0$ on (p,q), then $\xi > \xi(p)$ on $(p,q]$.

(2) If $\xi' < 0$ on (p,q), then $\xi > \xi(q)$ on $[p,q)$.

a) First, we prove that $g(\cdot,\theta) > 0$ on $(0,\theta)$. Note that $g(\theta,\theta)=0$,

$$\begin{aligned}
\partial_x g(x,\theta) &= \frac{1}{x(x-1)}\bigg[-x+\theta+x(1-x)\bigg(\log\frac{x}{\theta}-\log\frac{1-x}{1-\theta}\bigg)\bigg],\\
g_1(x,\theta) &:= -x+\theta+x(1-x)\bigg(\log\frac{x}{\theta}-\log\frac{1-x}{1-\theta}\bigg), \quad g_1(\theta,\theta)=0,\\
\partial_x g_1(x,\theta) &= -(1-2x)\bigg[\log\frac{1-x}{1-\theta}-\log\frac{x}{\theta}\bigg] = -(1-2x)\log\frac{\theta(1-x)}{x(1-\theta)}.
\end{aligned}$$

The last one is negative because of $x < \theta \leqslant 1/2$. The assertion now follows by using part (2) of Lemma 5.2 twice. First we have $g_1(\cdot,\theta) > 0$ and then $g(\cdot,\theta) > 0$ on $(0,\theta)$.

b) Next, we prove that $g(\cdot,\theta) > 0$ on $(1-\theta, 1)$. Note that $g(1-\theta,\theta) = 0$,

$$-g_1(1-\theta,\theta) = 1-2\theta-2\theta(1-\theta)\log\frac{1-\theta}{\theta} \geqslant 0,$$

and $-\partial_x g_1(\cdot,\theta) > 0$ on $(1-\theta, 1)$. The assertion follows by using part (1) of Lemma 5.2 twice. We have thus proved (19).

3) To complete the proof of the proposition, we show that the following condition

$$\partial_\theta h(x,\theta) > 0 \qquad \text{for all } (x,\theta) : \theta \in (0, 1/2) \text{ and } x \in (\theta, 1-\theta) \tag{20}$$

is sufficient. Actually, for θ_1, θ_2 with $0 < \theta_1 < \theta_2 < 1/2$, let $x_2 \in [\theta_2, 1-\theta_2]$ attain the minimum $h(x_2, \theta_2) = \min_{x\in(\theta_2,1-\theta_2)} h(x,\theta_2)$. Then

$$h(x_2,\theta_2) = h(x_2,\theta_1) + \int_{\theta_1}^{\theta_2} \partial_\theta h(x_2,u)\mathrm{d}u > h(x_2,\theta_1).$$

Here the last inequality is due to the fact that for (x_2, u) with

$$u < \theta_2 \leqslant x_2 \leqslant 1-\theta_2 < 1-u,$$

we have $\partial_\theta h(x_2, u) > 0$ by (20). Noting that by assumption, we have $x_2 \in (\theta_1, 1-\theta_1)$. Therefore

$$\min_{x\in(\theta_2,1-\theta_2)} h(x,\theta_2) = h(x_2,\theta_2) > h(x_2,\theta_1) \geqslant \min_{x\in(\theta_1,1-\theta_1)} h(x,\theta_1)$$

as required. A technical point in (20) is that we use the constraint "$x \in (\theta, 1-\theta)$" but not "$x \in (0,1)$".

4) Finally, we prove (20). To do so, we need some computations. First, we have

$$\begin{aligned}
h(x,\theta) &= -\bigg[(1-\theta)\log\frac{1-x}{1-\theta} + \theta\log\frac{x}{\theta}\bigg]\bigg/\bigg[(1-x)\log\frac{1-x}{1-\theta} + x\log\frac{x}{\theta}\bigg],\\
\partial_\theta h(x,\theta) &= \bigg\{(1-\theta)\theta\Big[\log\frac{1-x}{1-\theta} - \log\frac{x}{\theta}\Big]\Big[(1-x)\log\frac{1-x}{1-\theta} + x\log\frac{x}{\theta}\Big]\\
&\qquad - (x-\theta)\Big(1-\theta)\log\frac{1-x}{1-\theta} + \theta\log\frac{x}{\theta}\Big)\bigg\}\\
&\qquad \bigg/(1-\theta)\theta\Big((1-x)\log\frac{1-x}{1-\theta} + x\log\frac{x}{\theta}\Big)^2.
\end{aligned}$$

Next, define

$$
\begin{aligned}
h_1(x,\theta) &= (1-\theta)\theta\left(\log\frac{1-x}{1-\theta}-\log\frac{x}{\theta}\right)\left((1-x)\log\frac{1-x}{1-\theta}+x\log\frac{x}{\theta}\right)\\
&\quad-(x-\theta)\left((1-\theta)\log\frac{1-x}{1-\theta}+\theta\log\frac{x}{\theta}\right),\qquad \theta<x<1-\theta.\\
m(z,\theta) &:= h_1(\theta+z,\theta)\\
&= -z\left((1-\theta)\log\frac{1-\theta-z}{1-\theta}+\theta\log\frac{\theta+z}{\theta}\right)\\
&\quad+(1-\theta)\theta\left(\log\frac{1-\theta-z}{1-\theta}-\log\frac{\theta+z}{\theta}\right)\\
&\quad\times\left[(1-\theta-z)\log\frac{1-\theta-z}{1-\theta}+(\theta+z)\log\frac{\theta+z}{\theta}\right],\qquad 0<z<1-2\theta.
\end{aligned}
$$

Then we have $h_1(\theta,\theta)=0$ and $m(0,\theta)=0$. Moreover,

$$
\begin{aligned}
m_1(z,\theta) &:= \partial_z m(z,\theta)\\
&= \frac{z^2}{(1-\theta-z)(\theta+z)}-(1-\theta)\log\frac{1-y-z}{1-y}-\theta\log\frac{\theta+z}{\theta}\\
&\quad-(1-\theta)\theta\left(\log\frac{1-y-z}{1-y}-\log\frac{\theta+z}{\theta}\right)^2\\
&\quad-\frac{(1-\theta)\theta\left((1-\theta-z)\log\frac{1-y-z}{1-y}+(\theta+z)\log\frac{\theta+z}{\theta}\right)}{(1-\theta-z)(\theta+z)},\\
m_1(0,\theta) &= 0.
\end{aligned}
$$

$$
\begin{aligned}
\partial_z m_1(z,\theta) &= \bigg[(1-\theta)\theta\left(\theta^2-2(1-z)\theta-(2-z)z\right)\log\frac{\theta+z}{\theta}\\
&\quad+(1-\theta)\theta\left(1-\theta^2-2z\theta-z^2\right)\log\frac{1-\theta-z}{1-\theta}\\
&\quad-z\left(3\theta^2-(3-4z)\theta-(2-z)z\right)\bigg](1-\theta-z)^{-2}(\theta+z)^{-2}\\
&=: m_2(z,\theta)(1-\theta-z)^{-2}(\theta+z)^{-2}.
\end{aligned}
$$

Then we have $m_2(0,\theta)=0$ and

$$
\begin{aligned}
m_3(z,\theta) &:= \partial_z m_2(z,\theta)\\
&= (4-8\theta-3z)z-2(1-\theta)\theta(\theta+z)\log\frac{1-\theta-z}{1-\theta}\\
&\quad-2(1-\theta)\theta(1-\theta-z)\log\frac{\theta+z}{\theta}.
\end{aligned}
$$

By using the inequalities

$$
\left(1+\frac{1}{x}\right)^x<e\quad\text{for } x>0
$$

and
$$\left(1-\frac{1}{x}\right)^{-x} > e \quad \text{for } x > 1,$$
it follows that
$$\theta \log \frac{\theta+z}{\theta} = \theta \log\left(1+\frac{z}{\theta}\right) \leqslant z,$$
$$(\theta-1)\log\frac{1-\theta-z}{1-\theta} = (\theta-1)\log\left(1-\frac{z}{1-\theta}\right) \geqslant z$$
for $\theta \in (0,1/2)$ and $z \in (0, 1-2\theta)$. Therefore
$$m_3(z,\theta) \geqslant z\big(4-8\theta-3z+2\theta(\theta+z)-2(1-\theta)(1-\theta-z)\big) = z(2-4\theta-z) > 0$$
for all $z \in (0,1-2\theta)$. By using part (1) of Lemma 5.2 three times, we obtain successively that $m_2(\cdot,\theta) > 0$, $m_1(\cdot,\theta) > 0$, and $m(\cdot,\theta) > 0$ on $(0,1-2\theta)$. Hence $h_1(x,\theta) > 0$ for $x \in (\theta, 1-\theta)$ and so is $\partial_\theta h(x,\theta)$. $\square$

5.2 About the eigenequations

We now make some comments about the eigenequations corresponding to λ_1, σ and α_1, and show the difficulty in the study of α_1.

For simplicity, consider a finite state space E. Recall that for a given Q-matrix, $(q_{ij} : i, j \in E)$, we have a linear operator Ω. It is well known that for λ_1, we have a linear eigenequation
$$-\Omega f = \lambda_1 f, \qquad f \neq \text{constant}.$$
In particular, for
$$Q = \begin{pmatrix} -\theta & \theta \\ 1-\theta & \theta-1 \end{pmatrix},$$
the eigenequation becomes
$$\begin{cases} \theta(f_0-f_1) = \lambda_1 f_0 \\ (1-\theta)(f_1-f_0) = \lambda_1 f_1. \end{cases}$$
By a division of these two equations, we obtain an equation which is independent of λ_1:
$$\frac{f_1}{f_0} = -\frac{1-\theta}{\theta}.$$
Inserting this into the previous equation, we get $\lambda_1 = 1$.

For the logarithmic Sobolev constant σ, we have a non-linear equation
$$-\Omega f = \sigma f \log f, \qquad f > 0,\ \pi(f^2) = 1.$$

For "two points" matrix, the equation becomes

$$\begin{cases} \theta(f_0 - f_1) & = \sigma f_0 \log f_0 \\ (1-\theta)(f_1 - f_0) & = \sigma f_1 \log f_1. \end{cases}$$

Again, by a division of these two equations, we obtain an equation which is independent of σ:

$$\frac{f_1 \log f_1}{f_0 \log f_0} = -\frac{1-\theta}{\theta}.$$

In general, it is quite hard to solve a non-linear equation. However, by using $\log f_0 / \log f_1 = -1$ to eliminate the non-linear term, and using the constraint $\pi(f^2) = 1$, we obtain the required solution.

For the exponential convergence rate α_1, we also have a non-linear eigen-equation (as used in §4)

$$-\Omega f - f \Omega \log f = 2\alpha_1 f \log f, \qquad f > 0,\ \pi(f) = 1.$$

Note that one more non-linear term $f \Omega \log f$ appears and the left-hand side is not linear with respect to Ω. For "two points" matrix, the equation becomes

$$\begin{cases} \theta(f_0 - f_1 + f_0 \log(f_0/f_1)) & = 2\alpha_1 f_0 \log f_0 \\ (1-\theta)(f_1 - f_0 + f_1 \log(f_1/f_0)) & = 2\alpha_1 f_1 \log f_1. \end{cases}$$

By a division of these two equations, we obtain

$$\frac{(1-\theta)(f_1 - f_0 + f_1 \log(f_1/f_0))}{\theta(f_0 - f_1 + f_0 \log(f_0/f_1))} = \frac{f_1 \log f_1}{f_0 \log f_0}.$$

Equivalently,

$$\frac{(1-\theta)(1 - f_0/f_1 + \log(f_1/f_0))}{\theta(1 - f_1/f_0 + \log(f_0/f_1))} = \frac{\log f_1}{\log f_0}.$$

In view of the constraint $\pi(f) = 1$, replacing f_i with $f_i/\pi(f)$, the right-hand side can be written into a more symmetric form

$$\frac{\log(\theta + (1-\theta) f_0/f_1)}{\log(1 - \theta + \theta f_1/f_0)}.$$

Set $z = f_1/f_0$. Then the equation becomes

$$\frac{(1-\theta)(1 - 1/z + \log z)}{\theta(1 - z - \log z)} = \frac{\log(1-\theta+\theta z) - \log z}{\log(1-\theta+\theta z)}.$$

Thus,

$$1 - \frac{\log z}{\log(1-\theta+\theta z)} = \frac{(1-\theta)(1-1/z+\log z)}{\theta(1-z-\log z)} = \frac{1-\theta}{\theta}\left[-1 + \frac{z + 1/z - 2}{\log z + z - 1}\right],$$

or

$$\theta\frac{\log z}{\log(1-\theta+\theta z)}+(1-\theta)\frac{z+1/z-2}{\log z+z-1}=1.$$

This equation has a trivial solution $z=1$ (corresponding to the constant function $f=1$), which is the exceptional case of $\theta=1/2$ (in that case, $\lambda_1=\sigma=\alpha_1=1$). Since $\theta\in(0,1/2)$, the non-trivial solution $z>1$ (cf. (17)). Making a change of the variable $z=1+x$, the equation becomes

$$\theta\frac{\log(1+x)}{\log(1+\theta x)}+(1-\theta)\frac{x^2}{(1+x)(x+\log(1+x))}=1,\quad x\in(0,\,\theta^{-1}-1),\ \theta\in(0,1/2).$$

The numerical solution shows that the root x of this equation is decreasing in θ. In view of the expression of the equation, it seems easier to fix x and solve the equation in θ. However, when $x=e-1$, the equation becomes

$$\frac{\theta}{\log\big(1+(e-1)\theta\big)}+(1-\theta)\big(1-e^{-1}\big)^2=1.$$

Even for such nearly simplest equation, we still do not know how to find the solution $\theta\approx 0.267361$. Due to these facts, we are unable to find the explicit solution of α_1. Instead, we look for analytic upper and lower bounds with an expression as simple as possible.

5.3 Proof of Theorem 1.1

First, we state an extension of Theorem 1.1. For the Q-matrix

$$Q=\begin{pmatrix}-b & b\\ a & -a\end{pmatrix},\qquad a,b>0,$$

we have the following result.

Theorem 5.3 The constant $\alpha_1(Q)$ satisfies

$$\frac{1}{2}\left[a+b+\left(\frac{ab(a+b)}{2}\right)^{1/3}+\frac{a-b}{\log a-\log b}\right]\leqslant\alpha_1\leqslant\frac{1}{2}\left[a+b+\frac{2(a-b)}{\log a-\log b}\right].$$

In this case, we have

$$\lambda_1=a+b,\qquad \sigma=\frac{2(a-b)}{\log a-\log b}.$$

Proof of Theorem 5.3. Since

$$\begin{pmatrix}-b & b\\ a & -a\end{pmatrix}=(a+b)\begin{pmatrix}-b/(a+b) & b/(a+b)\\ a/(a+b) & -a/(a+b)\end{pmatrix}$$

and the symmetry, we need only consider the Q-matrix given in (3). $\square$

Next, we show that (10) is an improvement of (9). This is obvious for the upper estimate and so we need only to study the lower estimate.

Lemma 5.4 We have

$$\frac{1}{2}\left[1+\left(\frac{\theta(1-\theta)}{2}\right)^{1/3}+\frac{1-2\theta}{\log(\theta^{-1}-1)}\right]\geqslant\frac{1}{2}\big(\sqrt{\theta}+\sqrt{1-\theta}\big)^2,\qquad\theta\in[0,1/2].$$

The sign of the equality holds if and only if at the end points $\theta=0,1/2$.

Proof. It suffices to show that

$$\frac{1-2\theta}{\log(\theta^{-1}-1)}\geqslant\left(\frac{\theta(1-\theta)}{2}\right)^{1/3}\geqslant\big(\theta(1-\theta)\big)^{1/2},\qquad\theta\in(0,1/2),$$

and that the sign of the equality holds if and only if at the end points $\theta=0,1/2$. It is clear that the equalities hold at the end points. To prove the last inequality, it suffices that

$$\frac{1}{4}>\theta(1-\theta),\qquad\theta\in(0,1/2).$$

This is again obvious. To prove the first inequality, it suffices that

$$\frac{2^{1/3}(1-2\theta)}{(\theta(1-\theta))^{1/3}}>\log(\theta^{-1}-1),\qquad\theta\in(0,1/2).$$

Let

$$h(\theta)=\frac{2^{1/3}(1-2\theta)}{(\theta(1-\theta))^{1/3}}-\log(\theta^{-1}-1).$$

Since $h(1/2)=0$, by part (2) of Lemma 5.2, it is enough to show that on $(0,1/2)$, we have $h'<0$:

$$h'(\theta)=-2^{1/3}\frac{1+2\theta-2\theta^2-3(\theta(1-\theta)/2)^{1/3}}{3(\theta(1-\theta))^{4/3}},$$

In other words, we need to show that

$$1+2\theta-2\theta^2-3(\theta(1-\theta)/2)^{1/3}>0,\qquad\theta\in(0,1/2).$$

Making a change of the variable $z=(\theta(1-\theta)/2)^{1/3}$, the left-hand side becomes

$$1+4z^3-3z=(z+1)(2z-1)^2,$$

which is certainly positive on $(-1,1/2)$. $\square$

The remainder of this section is devoted to the proof of Theorem 1.1. The proof is similar to the one given in [6] but is more complicated. First, we need some preparation. Define a function $f_a(x,\theta)$ for $x\in(0,1)$ and $\theta\in(0,1/2)$ as follows

$$f_a(x,\theta)=(ax+\theta)\log\frac{x(1-\theta)}{\theta(1-x)}+(a+1)\log\frac{1-x}{1-\theta},\tag{21}$$

where $a = a(\theta) \geqslant 0$ is a parameter to be specified later. Then

$$\partial_x f_a(x,\theta) = \frac{\theta - x}{x(1-x)} + a \log \frac{x(1-\theta)}{\theta(1-x)},$$
$$\partial^2_{xx} f_a(x,\theta) = \frac{ax(1-x) - \theta + 2\theta x - x^2}{x^2(1-x)^2}.$$

Clearly, we have

$$f_a(\theta,\theta) = 0, \qquad f_a(1-\theta,\theta) = (a-1)(1-2\theta)\log\frac{1-\theta}{\theta}; \tag{22}$$
$$\partial_x f_a(\theta,\theta) = 0, \qquad \partial_x f_a(1-\theta,\theta) = \frac{2\theta-1}{\theta(1-\theta)} + 2a\log\frac{1-\theta}{\theta}; \tag{23}$$
$$\partial^2_{xx} f_a(\theta,\theta) = \frac{a-1}{\theta(1-\theta)}, \quad \partial^2_{xx} f_a(1-\theta,\theta) = \frac{1}{\theta(1-\theta)}\left[a + 3 - \frac{1}{\theta(1-\theta)}\right], \tag{24}$$

and

$$\partial_x f_a\left(\frac{1}{2},\ \theta\right) = 2(2\theta-1) + a\log\left(\frac{1}{\theta} - 1\right), \qquad \partial^2_{xx} f_a\left(\frac{1}{2},\ \theta\right) = \frac{a-1}{4\theta^2(1-\theta)^2}. \tag{25}$$

When

$$a \in \big(2\sqrt{\theta(1-\theta)},\ 1\big], \tag{26}$$

there exist two roots of $\partial^2_{xx} f_a(x,\theta)$:

$$x_1 = \frac{2\theta + a - \sqrt{a^2 - 4\theta(1-\theta)}}{2(1+a)} \geqslant \theta, \quad x_2 = \frac{2\theta + a + \sqrt{a^2 - 4\theta(1-\theta)}}{2(1+a)} \leqslant \frac{1}{2}.$$

Here the sign of the equalities holds if and only if $a = a(\theta) = 1$. If

$$\lim_{\theta\to 0} a(\theta) = 0,$$

then in the extreme case that $\theta = 0$, we have $x_1 = x_2 = 0$. If

$$\lim_{\theta\to 1/2} a(\theta) = 1,$$

then in the extreme case that $\theta = 1/2$, we have $x_1 = x_2 = 1/2$. In the last case, the function $f_a(\cdot,\theta)$ has uniquely the maximum 0 attained at $x_1 = x_2 = \theta = 1/2$. Note that if we take $a(\theta) = 2\sqrt{\theta(1-\theta)}$, then we get again $x_1 = x_2$. In this case, the proof would become much simpler and we then obtain the lower bound in (9).

Set

$$n(\theta) = \frac{2(1-2\theta)}{\log(\theta^{-1}-1)}.$$

Then $n(\theta)$ is increasing in θ,

$$\lim_{\theta\to 0} n(\theta) = 0, \qquad \lim_{\theta\to 1/2} n(\theta) = 1.$$

For $\theta \in (0, 1/2)$ and $a(\theta) \in \big(2\sqrt{\theta(1-\theta)},\ n(\theta)\big]$, we have seen that

$$\theta < x_1 < x_2 < \frac{1}{2} < 1-\theta, \tag{27}$$

and

$$\partial^2_{xx} f_a(\theta,\theta) < 0, \qquad \partial^2_{xx} f_a((x_1+x_2)/2, \theta) > 0, \qquad \partial^2_{xx} f_a(1-\theta,\theta) < 0. \tag{28}$$

It follows that the function $f_a(\cdot,\theta)$ has the following properties.

- It is concave on (θ, x_1), convex on (x_1, x_2) and then concave again on $(x_2, 1-\theta)$.
- It starts from the local maximum $f_a(\theta,\theta) = 0$ at θ, next decreases to the minimum at a point $x_* \in (x_1, x_2)$, then increases to a local maximum at $x^* \in (x_2, 1-\theta)$ and finally decreases again. In particular, the function $f_a(\cdot,\theta)$ has negative value on $(0,\theta)\cup(\theta, x_3)\cup(1-\theta, 1)$, where $x_3 \in (x_*, x^*)$.
- About x^* (usually not explicitly known) which is determined by $\partial_x f_a(x^*, \theta) = 0$, there are two cases.

  (a) If $\partial_x f_a(x_2,\theta) > 0$, then we indeed have $x^* \in (x_2, 1/2]$ since $a(\theta) \leqslant n(\theta)$, hence $\partial_x f_a(1/2,\ \theta) \leqslant 0$ and furthermore $\partial_x f_a(\cdot,\ \theta) < 0$ on $(1/2, 1)$. Otherwise, we have the case

  (b) $\partial_x f_a(x_2,\theta) \leqslant 0$. Since x_2 is the maximal point of $\partial_x f_a(\cdot,\theta)$ on $(x_1, 1-\theta)$, it follows that $\partial_x f_a(\cdot,\theta) \leqslant 0$ on $(x_1, 1)$ and then on $(\theta, 1)$. This means that $f_a(\cdot,\theta)$ is non-increasing on $(\theta, 1)$ with $f_a(\theta,\theta) = 0$.

To have an impression about $f_a(\cdot,\theta)$ and its first two derivatives, we fix $a = (\theta(1-\theta)/2)^{1/3} + (1-2\theta)/\log(\theta^{-1}-1)$ and $\theta = 0.3$. Then $x_1 \approx 0.339$, $x_2 \approx 0.455$ and the pictures are shown by Figures 1–4.

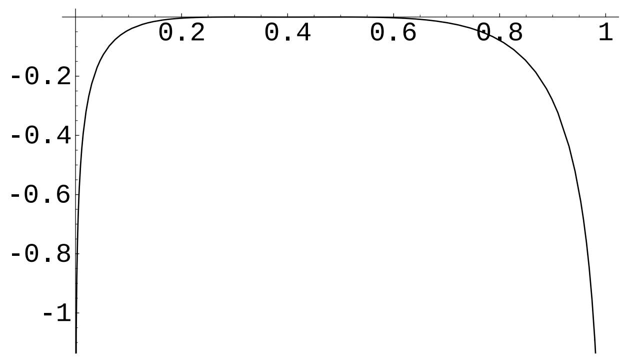


Figure 1: The picture of $f_a(\cdot, 0.3)$.

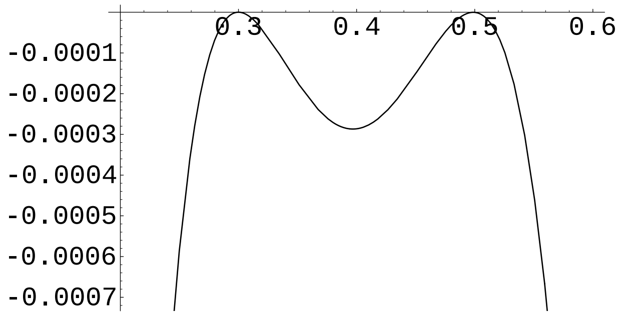


Figure 2: The picture of $f_a(\cdot, 0.3)$ on the smaller interval $[0.2, 0.6]$.

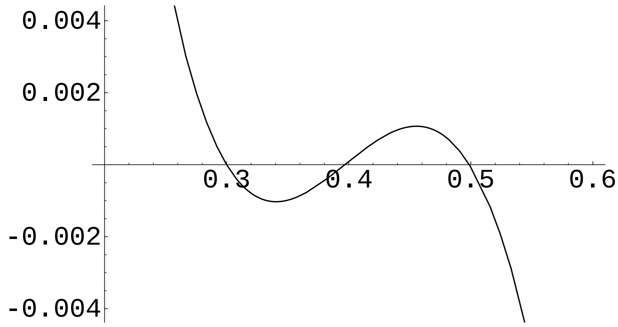


Figure 3: The picture of $\partial_x f_a(\cdot, 0.3)$ on $[0.2, 0.6]$.

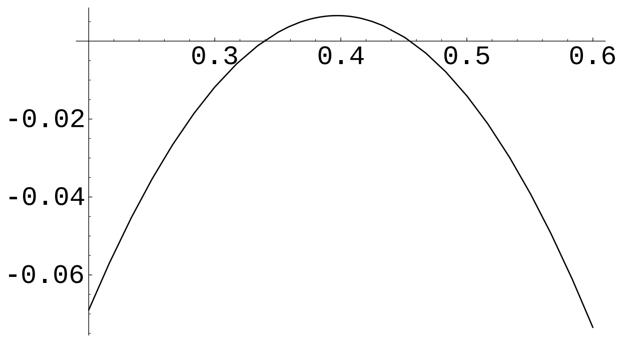


Figure 4: The picture of $\partial^2_{xx} f_a(\cdot, 0.3)$ on $[0.2, 0.6]$.

We can now state a description of $\alpha_1(\theta)$.

Proposition 5.5 The exponential rate $\alpha_1(\theta)$ in entropy is given by

$$2\alpha_1(\theta) - 1 = \frac{x^* - \theta}{x^*(1-x^*)} \Bigg/ \log \frac{x^*(1-\theta)}{\theta(1-x^*)}, \qquad \text{(Figure 5)}$$

where x^* is the solution to the equation

$$\frac{x-\theta}{1-x}\left[1+\left(\log\frac{1-x}{1-\theta}\right)\Big/\left(x\log\frac{x(1-\theta)}{\theta(1-x)}\right)\right] \\ +(1-\theta)\log\frac{1-x}{1-\theta}+\theta\log\frac{x}{\theta}=0, \qquad x\in(\theta,1-\theta). \tag{29}$$

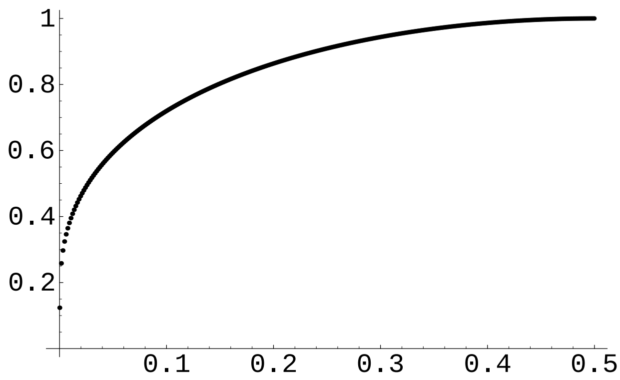


Figure 5: The graph of $2\alpha_1(\theta)-1$.

Proof. Set $a^*=2\alpha_1(\theta)-1$. Noting that the function j_θ defined by (18) is positive on $[0,1]\setminus\{\theta\}$, by (16), (17) and (21), it follows that (x^*,a^*) is the solution to the following equations:

$$\begin{cases} f_a(x,\theta)=0, \\ \partial_x f_a(x,\theta)=0, \quad x\in(\theta,1-\theta). \end{cases}$$

That is

$$\begin{cases} (ax+\theta)\log\frac{x(1-\theta)}{\theta(1-x)}+(a+1)\log\frac{1-x}{1-\theta}=0, \\ \frac{\theta-x}{x(1-x)}+a\log\frac{x(1-\theta)}{\theta(1-x)}=0, \quad x\in(\theta,1-\theta). \end{cases}$$

From the second equation, we obtain

$$a=\frac{x-\theta}{x(1-x)}\Big/\log\frac{x(1-\theta)}{\theta(1-x)}.$$

Inserting this into the first equation, it follows that x^* is the solution to the equation (29). Then we obtain the required assertion. $\square$

Proof of Theorem 1.1 (a) The proof of the upper estimate is now easy. By (16), we need to show that

$$\inf_{x\in(0,1)} h(x,\theta)<n(\theta)=\frac{2(1-2\theta)}{\log(\theta^{-1}-1)}$$

for $\theta\in(0,1/2)$. Equivalently, with $a=n(\theta)$, the inequality

$$f_a(x,\theta)>0$$

holds for some $x \in (\theta, 1-\theta)$ since $f_a < 0$ on $(0,\theta) \cup (1-\theta, 1)$. Noting that the equation

$$n(\theta) = \frac{x-\theta}{x(1-x)} \Big/ \log \frac{x(1-\theta)}{\theta(1-x)}, \qquad x \in (x_2, 1/2]$$

has only one solution $x^* = 1/2$ (independent of θ), it suffices to show that

$$\begin{aligned} f_a\left(\frac{1}{2}, \theta\right) &= \left(\frac{1-2\theta}{\log(1/\theta - 1)} + 1 - \theta\right) \log \frac{1}{2(1-\theta)} \\ &\quad + \left(\frac{1-2\theta}{\log(1/\theta - 1)} + \theta\right) \log \frac{1}{2\theta} \\ &> 0 \end{aligned}$$

for all $\theta \in (0, 1/2)$. Equivalently,

$$(1-2\theta) \log \frac{1}{4\theta(1-\theta)} + \log(\theta^{-1} - 1)\big[\theta \log(\theta^{-1} - 1) - \log(2(1-\theta))\big] > 0.$$

By change of the variable $\theta = 2^{-1} - z$, the condition becomes

$$\begin{aligned} g(z) &:= -2z \log(1-4z^2) - \frac{1}{2} \log \frac{1+2z}{1-2z}\left[\log(1-4z^2) + 2z \log \frac{1+2z}{1-2z}\right] \\ &> 0, \quad z \in (0, 1/2). \end{aligned} \tag{30}$$

The graph of the function g is quite smooth. It starts from 0 at the origin, and then increases to infinity at 1/2 (see Figure 6). The Taylor expansion of g at 0 is as follows.

$$\frac{64}{45}z^7 + \frac{512}{63}z^9 + \frac{18944}{525}z^{11} + \frac{219136}{1485}z^{13} + \frac{2757025792}{4729725}z^{15} + \frac{11337728}{5005}z^{17} + O\left(z^{19}\right).$$

The coefficients in the series above are all positive. Hence, there is nothing to be worried about small z.

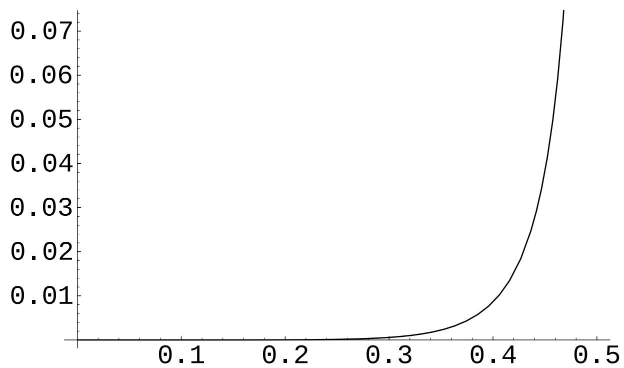


Figure 6: The graph of the function g in (30).

(b) To prove the lower estimate, let

$$a(\theta) = \left(\frac{\theta(1-\theta)}{2}\right)^{1/3} + \frac{1-2\theta}{\log(\theta^{-1}-1)}.$$

Then $a(\theta) \in \left(2\sqrt{\theta(1-\theta)},\ n(\theta)\right)$ for $\theta \in (0, 1/2)$,

$$\lim_{\theta\to 0} a(\theta) = 0, \qquad \lim_{\theta\to 1/2} a(\theta) = 1.$$

Note that $\partial_x f_a(x_2, \cdot)$ has a root at $\theta_0 \approx 0.00118437$ (see Figures 7 and 8).

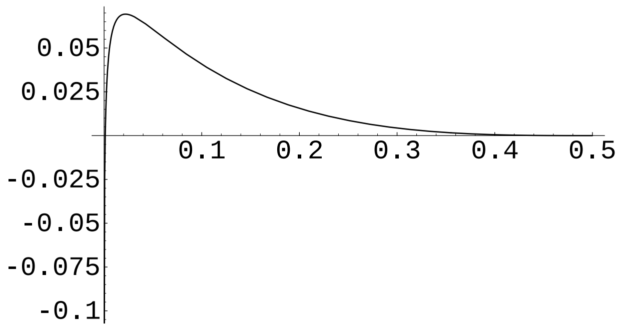


Figure 7: The graph of $\partial_x f_a(x_2, \cdot)$ on $(0, 0.5)$.

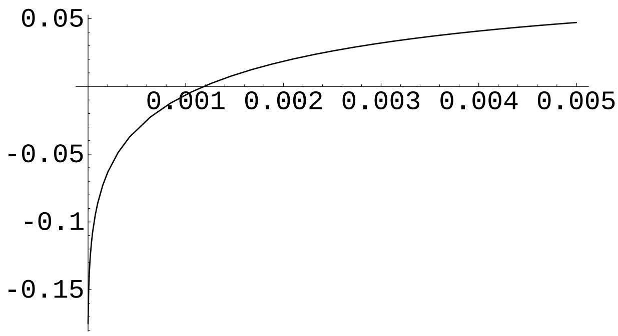


Figure 8: The graph of $\partial_x f_a(x_2, \cdot)$ on $(0, 0.05)$.

Thus, it suffices to show that

$$(a(\theta)x^* + \theta)\log\frac{x^*(1-\theta)}{\theta(1-x^*)} + (a(\theta)+1)\log\frac{1-x^*}{1-\theta} < 0, \qquad \theta \in (\theta_0, 1/2), \tag{31}$$

where x^* is the solution to the equation

$$\begin{aligned} &a(\theta)\log\frac{x(1-\theta)}{\theta(1-x)} = \frac{x-\theta}{x(1-x)}, \\ &x \in \left(\frac{2\theta + a(\theta) + \sqrt{a(\theta)^2 - 4\theta(1-\theta)}}{2(1+a(\theta))},\ \frac{1}{2}\right). \end{aligned} \tag{32}$$

Because x^* satisfies (11), we can simplify (10) as follows.

$$\frac{\theta}{a(\theta)} + x^* \left[1 + \big(a(\theta)+1\big)\frac{1-x^*}{x^*-\theta}\log\frac{1-x^*}{1-\theta}\right] < 0, \qquad \theta \in (\theta_0, 1/2), \qquad (33)$$

This is checked by using computer as shown in Figures 9 and 10.

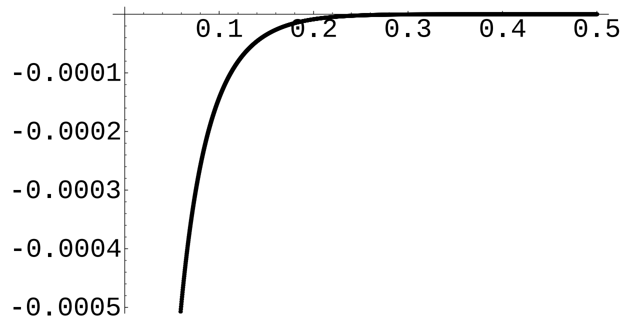


Figure 9: The graph of the function in (33).

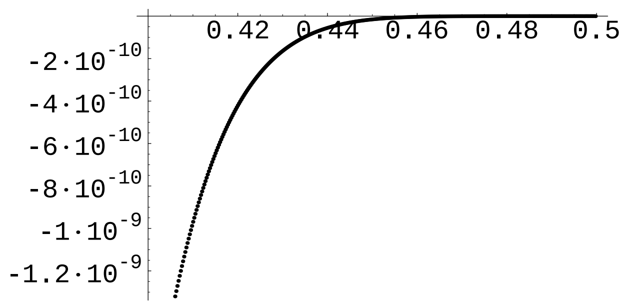


Figure 10: The graph of the function in (33) on $[0.4, 0.5]$.

Since our estimate is sharp at $1/2$, the function tends to zero rapidly as $\theta \to 1/2$.

Alternatively, to avoid solving the non-linear equation (11), one may check directly that

$$k(x,\theta) := (a(\theta)x + \theta)\log\frac{x(1-\theta)}{\theta(1-x)} + (a(\theta)+1)\log\frac{1-x}{1-\theta} < 0$$

for all $\theta \in (\theta_0, 1/2)$ and

$$x \in \left(\frac{2\theta + a(\theta) + \sqrt{a(\theta)^2 - 4\theta(1-\theta)}}{2(1+a(\theta))},\ \frac{1}{2}\right).$$

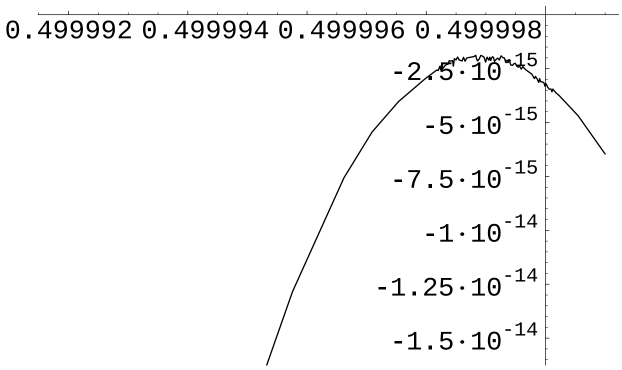


Figure 11: The graph of the function $k(\cdot, 0.48)$ on $[0.49998, 0.500001]$.

As an example, the picture of $k(\cdot, 0.48)$ on $[0.49998, 0.500001]$ is given in Figure 11. The maximum point is $x^* \approx 0.4999989$ very close to $1/2$; and the maximum is approximately -2.0678×10^{-15}, nearly zero but negative. $\square$

Acknowledgements Research supported in part by the Creative Research Group Fund of the National Natural Science Foundation of China (No. 10121101) and by the "985" project from the Ministry of Education in China.

References

[1] Bobkov S G, Tetali P. Modified logarithmic Sobolev inequalities in discrete settings, J Theor Prob 2006, 19(2), 289–336.

[2] Diaconis P, Saloff-Coste L. Logarithmic Sobolev inequalities for finite Markov chains, Ann Appl Prob 1996, 6(3), 695–750.

[3] Chen M F. Exponential L^2-convergence and L^2-spectral gap for Markov processes. Acta Math Sin New Ser 1991, 7(1), 19–37.

[4] Chen M F. From Markov Chains to Non-Equilibrium Particle Systems, World Scientific, Singapore, 1992; Second Edition: 2004.

[5] Chen M F. Estimation of spectral gap for Markov chains. Acta Math Sin New Ser 1996, 12(4), 337–360.

[6] Chen M F. Coupling, spectral gap and related topics (III), Chin Sci Bull 1997, 42(18), 1497–1505.

[7] Chen M F. Nash inequalities for general symmetric forms, Acta Math Sin Eng Ser 1999, 15(3), 353–370.

[8] Chen M F. Logarithmic Sobolev inequality for symmetric forms. Sci Sin (A) 2000, 43(6), 601–608.

[9] Chen M F. A new story of ergodic theory, in "Applied Probability", eds. R. Chan et al, AMS/IP Studies in Advanced Mathematics 26, 2002, 25–34.

[10] Chen M F. Eigenvalues, Inequalities, and Ergodic Theory, Springer, London, 2005.

[11] Chen M F, Wang F Y. Cheeger's inequalities for general symmetric forms and existence criteria for spectral gap. Abstract: Chin Sci Bull 1998, 43(18) 1516–1519; Ann Prob 2000, 28(1), 235–257.

[12] Higuchi Y, Yosshida N. Analytic conditions and phase transition for Ising models, Lecture Notes (in Japanese), 1995.

[13] Lawler G F, Sokal A D. Bounds on the L^2 spectrum for Markov chain and Markov processes: a generalization of Cheeger's inequality, Trans Amer Math Soc 1988, 309, 557–580.

[14] Mao Y H, Zhang S Y. Comparison of some convergence rates for Markov process (in Chinese), Acta Math Sin 2000, 43(6), 1019–1026.

[15] Wang F Y. Functional inequalities for empty essential spectrum, J Funct Anal 2000, 170, 219–245.

[16] Wang F Y. Sobolev type inequalities for general symmetric forms, Proc Amer Math Soc 2001, 128(12), 3675–3682.

[17] Zhang S Y, Mao Y H. Exponential convergence rate in Boltzman–Shannon entropy, Sci Sin (A) 2000, 44(3), 280–285.

Acta Math. Sin. New Ser.
2008, 24(5), pp. 705–736

SPECTRAL GAP AND LOGARITHMIC SOBOLEV CONSTANT FOR CONTINUOUS SPIN SYSTEMS

Mu-Fa Chen

(School of Mathematical Sciences, Beijing
Normal University, Beijing 100875, P.R. China)
E-mail: mfchen@bnu.edu.cn
Received June 8, 2007; accepted September 15, 2007.

Abstract. The aim of this paper is to study the spectral gap and the logarithmic Sobolev constant for continuous spin systems. A simple but general result for estimating the spectral gap of finite dimensional systems is given by Theorem 1.1, in terms of the spectral gap for one-dimensional marginals. The study of the topic provides us a chance, and it is indeed another aim of the paper, to justify the power of the results obtained previously. The exact order in dimension one (Proposition 1.4), and then the precise leading order and the explicit positive regions of the spectral gap and the logarithmic Sobolev constant for two typical infinite-dimensional models are presented (Theorems 6.2 and 6.3). Since we are interested in explicit estimates, the computations become quite involved. A long section (Section 4) is devoted to the study of the spectral gap in dimension one.

1. Introduction. The local Poincaré inequalities (equivalently, spectral gaps) and logarithmic Sobolev inequalities for unbounded continuous spin systems have recently obtained a lot of attention by many authors [1]–[11]. For the present status of the study and further references, the readers may refer to the comprehensive survey article [7]. In the most of the publications, the authors consider mainly the perturbation regime with convex phase at infinity. More recently, the non-convex phase is treated for a class of spin systems based on a criterion for the weighted Hardy inequalities.

The main purpose of this paper is to propose a general formula for the local spectral gaps of continuous spin systems. Let us start from finite dimensions. Let

2000 *Mathematics Subject Classification.* 60K35.

Key words and phrases. Spectral gap, logarithmic Sobolev constant, spin system, principle eigenvalue.

Research supported in part by the Creative Research Group Fund of the National Natural Science Foundation of China (No. 10121101) and by the "985" project from the Ministry of Education in China.

Typeset by $\mathcal{AMS}$-TEX

$U \in C^\infty(\mathbb{R}^n)$ satisfy $Z := \int_{\mathbb{R}^n} e^{-U}\mathrm{d}x < \infty$ and set $\mathrm{d}\mu_U = e^{-U}\mathrm{d}x/Z$. Throughout the paper, we use a particular notation $x_{\setminus i} := (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \in \mathbb{R}^{n-1}$, obtained from $x := (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ by removing the ith component. Clearly, the conditional distribution of x_i given $x_{\setminus i}$ under μ_U is as follows:

$$\mu_U^{x_{\setminus i}}(\mathrm{d}x_i) = e^{-U}\mathrm{d}x_i/Z(x_{\setminus i}), \tag{1.1}$$

where $Z(x_{\setminus i}) = \int_{\mathbb{R}} e^{-U(x)}\mathrm{d}x_i$. The measure $\mu_U^{x_{\setminus i}}$ is the invariant probability measure of the one-dimensional diffusion process, corresponding to the operator $L_i^{x_{\setminus i}} = \mathrm{d}^2/\mathrm{d}x_i^2 - \partial_i U\,\mathrm{d}/\mathrm{d}x_i$.

Let $L = \Delta - \langle \nabla U, \nabla \rangle$. Recall that the spectral gap $\lambda_1(L) = \lambda_1(U)$ is the largest constant κ in the following Poincaré inequality

$$\kappa \mathrm{Var}_{\mu_U}(f) \leqslant \int_{\mathbb{R}^n} |\nabla f|^2 \mathrm{d}\mu_U =: D(f), \qquad f \in C_0^\infty(\mathbb{R}^n), \tag{1.2}$$

where $\mathrm{Var}_{\mu_U}(f)$ is the variation of f with respect to μ_U and $C_0^\infty(\mathbb{R}^n)$ is the set of smooth functions with compact supports.

Denote by $\lambda_1^{x_{\setminus i}} = \lambda_1\big(L_i^{x_{\setminus i}}\big)$ the spectral gap of the one-dimensional operator $L_i^{x_{\setminus i}}$:

$$\lambda_1^{x_{\setminus i}} \mathrm{Var}_{\mu_U^{x_{\setminus i}}}(f) \leqslant \int_{\mathbb{R}} f'^2 \mathrm{d}\mu_U^{x_{\setminus i}}, \qquad f \in C_0^\infty(\mathbb{R}). \tag{1.3}$$

Then, we can state our variational formula for the lower bounds of $\lambda_1(U)$ as follows.

Theorem 1.1. Define

$$\big(\widetilde{\mathrm{Hess}}(U)\big)_{ij} = \begin{cases} \lambda_1^{x_{\setminus i}}, & \text{if } i = j \\ \partial_{ij} U, & \text{if } i \neq j, \end{cases}$$

where $(\mathrm{Hess}(U))_{ij} = \partial_{ij} U := \partial^2 U/\partial_{x_i}\partial_{x_j}$. Then we have

$$\begin{aligned} \lambda_1(U) &\geqslant \inf_{x \in \mathbb{R}^n} \lambda_{\min}\big(\widetilde{\mathrm{Hess}}(U)(x)\big) \\ &\geqslant \inf_{x \in \mathbb{R}^n} \sup_w \min_{1 \leqslant i \leqslant n} \Big(\lambda_1^{x_{\setminus i}} - \sum_{j: j \neq i} |\partial_{ij} U(x)| w_j / w_i\Big), \end{aligned} \tag{1.4}$$

where $w = (w_i)_{i=1}^n$ varies over all positive sequences.

Setting $w_i \equiv 1$ in (1.4), it follows that

$$\lambda_1(U) \geqslant \inf_{x \in \mathbb{R}^n} \min_{1 \leqslant i \leqslant n} \Big(\lambda_1^{x_{\setminus i}} - \sum_{j: j \neq i} |\partial_{ij} U(x)|\Big) \geqslant \min_{1 \leqslant i \leqslant n} \Big[\inf_{x \in \mathbb{R}^n} \lambda_1^{x_{\setminus i}} - \sum_{j: j \neq i} \|\partial_{ij} U\|_\infty\Big].$$

The last lower bound is more or less the estimate given in [5] and [7], goes back to [3].

The supremum over w in (1.4) comes from a variational formula for the principal eigenvalue of a symmetric Q-matrix (cf. §3 for more details). The use of the variational formula is necessary, since the principal eigenvalue is not computable in general for a large scale matrix.

The essential point for which (1.4) is valuable is that we now have quite complete knowledge about the spectral gap in dimensional one. For instance, as a consequence of part (1) of Theorem 3.1 in [12], we have

$$\lambda_1^{x\setminus i} \geqslant \sup_f \inf_{x_i\in\mathbb{R}} \bigg\{\partial_{ii}U(x) - \frac{f''(x_i) - \partial_i U(x) f'(x_i)}{f(x_i)}\bigg\}, \tag{1.5}$$

where f varies over all positive functions in $C^2(\mathbb{R})$. In particular, setting $f = 1$, we get

$$\lambda_1^{x\setminus i} \geqslant \inf_{x_i\in\mathbb{R}} \partial_{ii}U(x). \tag{1.6}$$

When $\partial_{ii}U(x) = u''(x_i)$ for some $u \in C^2(\mathbb{R})$, independent of i, (1.6) leads to the so-called convex phase condition "$\inf_{x\in\mathbb{R}} u''(x) > 0$." Since a local modification of u does not change the positiveness of λ_1, the convex condition can be replaced by $\varliminf_{|x|\to\infty} u''(x) > 0$ (i.e., the convexity at infinity) as proved in [12; Corollary 3.5], see also Theorem 4.1 below. However, the last condition is still not necessary as shown by [12; Example 3.11 (3): $u'(x) = \gamma x(\gamma + \cos x)^{-1}$ for some $\gamma > 1$] and [5; Proposition 4.4] (see also Example 2.5 below). A more careful examination of spectral gap in dimension one is delayed to §4.

It is possible to avoid the use of test functions w and f in (1.4) and (1.5), respectively. To see this, we introduce an explicit lower estimate of $\lambda_1(U)$. For this, we need additional notations. Choose a practical $\eta_i^{x\setminus i} \leqslant \lambda_1^{x\setminus i}$, as bigger as possible, and define

$$\begin{aligned}
s_i(x) &= \eta_i^{x\setminus i} - \textstyle\sum_{j: j\neq i} |\partial_{ij}U(x)|, &\qquad \underline{s}(x) &= \min_{1\leqslant i\leqslant n} s_i(x),\\
q_i(x) &= \eta_i^{x\setminus i} - \underline{s}(x), & d_i(x) &= s_i(x) - \underline{s}(x),\\
h^{(\gamma)}(x) &= \min_{A:\, \emptyset\neq A\subset\{1,2,\ldots,n\}} \frac{1}{|A|}\bigg[\sum_{i\in A}\frac{d_i(x)}{q_i(x)^\gamma} + \sum_{i\in A, j\notin A}\frac{|\partial_{ij}U(x)|}{[q_i(x)\vee q_j(x)]^\gamma}\bigg],\\
&\qquad \gamma \geqslant 0,
\end{aligned} \tag{1.7}$$

where $a \vee b = \max\{a, b\}$ and $|A|$ is the cardinality of the set A.

Theorem 1.2. We have

$$\lambda_1(U) \geqslant \inf_{x\in\mathbb{R}^n} \bigg\{\underline{s}(x) + \frac{h^{(1/2)}(x)^2}{1 + \sqrt{1 - h^{(1)}(x)^2}}\bigg\}. \tag{1.8}$$

A close related topic to the Poincaré inequality is the logarithmic Sobolev inequality with optimal constant $\sigma(U)$:

$$\sigma(U)\, \mathrm{Ent}_\mu\big(f^2\big) \leqslant 2D(f), \qquad f \in \mathscr{D}(D), \tag{1.9}$$

where $\mathrm{Ent}_\mu(f) = \mu(f \log f) - \mu(f)\log \mu(f)$ for $f \geqslant 0$. Correspondingly, we have the conditional marginal inequality for $\mu_U^{x_{\setminus i}}$, given $x_{\setminus i}$, with optimal constant $\sigma^{x_{\setminus i}}$:

$$\sigma^{x_{\setminus i}} \mathrm{Ent}_{\mu_U^{x_{\setminus i}}}(f) \leqslant \int_{\mathbb{R}} f'^2 \mathrm{d}\mu_U^{x_{\setminus i}}, \qquad f \in C_0^\infty(\mathbb{R}). \tag{1.10}$$

We can now state a very recent result due to [8; Theorem 1], which is consistent with Theorem 1.1.

Theorem 1.3. The logarithmic Sobolev constant $\sigma(U) \geqslant \lambda_{\min}(A)$, where the matrix $A = (A_{ij})$ is defined by

$$A_{ij} = \begin{cases} \inf_x \sigma^{x_{\setminus i}}, & \text{if } j = i \\ -\sup_x |\partial_{ij} U(x)|. & \text{if } j \neq i. \end{cases}$$

In view of the above results, it is clear that the one-dimensional case plays a crucial role. In that case, a representative result of the paper is as follows.

Proposition 1.4. In dimensional one, replace U with $u_{\beta_1,\,\beta_2}(x) = x^4 - \beta_1 x^2 + \beta_2 x$ for some constants $\beta_1 \geqslant 0$ and $\beta_2 \in \mathbb{R}$. Then we have

$$\begin{aligned} 4e^{14}\exp\Big[-\frac{1}{4}\beta_1^2 + 2\log(1+\beta_1)\Big] &\geqslant \inf_{\beta_2} \lambda_1(u_{\beta_1,\,\beta_2}) \\ &\geqslant \inf_{\beta_2} \sigma(u_{\beta_1,\,\beta_2}) \\ &\geqslant \frac{\sqrt{\beta_1^2+8}-\beta_1}{\sqrt{e}} \exp\Big[-\frac{1}{8}\beta_1\Big(\beta_1 + \sqrt{\beta_1^2+8}\Big)\Big]. \end{aligned}$$

In particular, $\inf_{\beta_2} \lambda_1(u_{\beta_1,\,\beta_2})$ and $\inf_{\beta_2} \sigma(u_{\beta_1,\,\beta_2})$ have the same order as $\exp[-\beta_1^2/4 + O(\log \beta_1)]$ as $\beta_1 \to \infty$.

The exponent $\beta_1^2/4$ here equals, approximately as $\beta_1 \to \infty$, the square of the variance of a random variable having the distribution with density $\exp[-x^4 + \beta_1 x^2]/Z$ on the real line.

The remainder of the paper is organized as follows. In the next section, we study an alternative variational formula for spectral gap. This is especially meaningful in the context of diffusions. The proofs of Theorems 1.1 and 1.2 are completed in §3. The one-dimensional spectral gap is the main topic in §4. The logarithmic Sobolev constant is studied in §5, in which Proposition 1.4 is proven. Even though the explicit and universal upper and lower estimates, as well as the criteria, for the spectral gap and logarithmic Sobolev constant are all known (cf. [13; Chapter 5, Theorem 7.4] and §4 below), it is still quite a distance to arrive at Proposition 1.4. Actually, we study this model several times (Examples 4.3, 4.6, 4.9, 5.3, and Proposition 4.7) by using different approaches. Thus, a part of the paper is methodological, it takes time and space to make some comparison of different methods. Two typical infinite-dimensional models are treated in the last section.

2. Alternative variational formula for spectral gap. Let $(E, \mathscr{E}, \mu)$ be a probability space and $L^2(\mu)$ be the ordinary L^2-space of real functions. Corresponding to a μ-reversible Markov process with transition probability $P(t, x, \cdot)$, we have a positive, strongly continuous, contractive and self-adjoint semigroup $\{P_t\}_{t \geqslant 0}$ on $L^2(\mu)$ with generator $(L, \mathscr{D}(L))$. Throughout this section, $(\cdot, \cdot)$ and $\|\cdot\|$ denote, respectively, the inner product and the norm in $L^2(\mu)$. By elementary spectral theory, we have

$$\frac{1}{t}(f - P_t f, f) \uparrow \text{ some } D(f, f) =: D(f) \leqslant \infty \text{ as } t \downarrow 0. \tag{2.1}$$

Set $\mathscr{D}(D) = \{f \in L^2(\mu) : D(f) < \infty\}$ and define $D(f, g) = (D(f+g) - D(f-g))/4$ for $f, g \in \mathscr{D}(D)$. Then, $(D, \mathscr{D}(D))$ is a Dirichlet form. Moreover,

$$D(f, g) = -(Lf, g), \qquad f, g \in \mathscr{D}(L). \tag{2.2}$$

The formula in (2.4) below goes back to [14].

Theorem 2.1. The spectral gap $\lambda_1(L)$ is described by the largest constant κ in the following equivalent inequalities.

$$\kappa \mathrm{Var}_\mu(f) \leqslant D(f), \qquad f \in \mathscr{D}(D), \tag{2.3}$$

$$\kappa D(f) \leqslant \|Lf\|^2, \qquad f \in \mathscr{D}(L). \tag{2.4}$$

Proof. Let $\{E_\alpha)\}_{\alpha \geqslant 0}$ be the spectral representation of L. Then $L = -\int_0^\infty \alpha \mathrm{d}E_\alpha$. The optimal constant κ in (2.3) is known to be $\lambda_1 = \lambda_1(L)$. Note that

$$\begin{aligned}
\|Lf\|^2 &= (Lf, Lf) \\
&= (f, L^2 f) \\
&= \left(f, \int_0^\infty \alpha^2 \mathrm{d}E_\alpha f\right) \\
&= \int_0^\infty \alpha^2 \mathrm{d}(E_\alpha f, f) \\
&= \int_{\lambda_1}^\infty \alpha^2 \mathrm{d}(E_\alpha f, f) \geqslant \lambda_1 \int_{\lambda_1}^\infty \alpha \mathrm{d}(E_\alpha f, f) \\
&= \lambda_1 \int_0^\infty \alpha \mathrm{d}(E_\alpha f, f) \\
&= \lambda_1 (f, -Lf) \\
&= \lambda_1 D(f).
\end{aligned}$$

Because the only inequality here cannot be improved, the largest constant κ in (2.4) is also equal to λ_1. $\square$

Remark 2.2. Actually, it is known and is also easy to check that (2.3) is equivalent to the correlation inequality

$$\lambda_1(L)|\mathrm{Cov}_\mu(f,g)| \leqslant \big(D(f)D(g)\big)^{1/2}, \qquad f,g\in \mathscr{D}(D), \tag{2.5}$$

where $\mathrm{Cov}_\mu(f,g)=\mu(fg)-\mu(f)\mu(g)$ and $\mu(f)=\int f\mathrm{d}\mu$. See the comment below Proposition 3.2 for a proof.

Before moving further, let us mention that the above proof also works for the principal eigenvalue. In this case, $L1\neq 0$ and μ can be infinite. Then the principal eigenvalue λ_0 can be described by the following equivalent inequalities.

$$\begin{aligned} &\tilde{\kappa}\|f\|^2 \leqslant D(f), \qquad f\in\mathscr{D}(D),\\ &\tilde{\kappa}D(f)\leqslant \|Lf\|^2, \qquad f\in\mathscr{D}(L). \end{aligned}\tag{2.6}$$

The formula (2.4) is especially useful for diffusion on Riemannian manifolds. Thus, the next result is meaningful for a more general class of diffusion in $\mathbb{R}^n$ by using a suitable Riemannian structure.

Corollary 2.3. Let $L=\Delta-\langle\nabla U,\nabla\rangle$ for some $U\in C^\infty(\mathbb{R}^n)$ with $Z:=\int_{\mathbb{R}^n}e^{-U}\mathrm{d}x<\infty$ and set $\mu(\mathrm{d}x)=e^{-U}\mathrm{d}x/Z$. Then

$$\|Lf\|^2=\int_{\mathbb{R}^n}\bigg[\sum_{i,j}(\partial_{ij}f)^2+\langle \mathrm{Hess}(U)\nabla f,\nabla f\rangle\bigg]\mathrm{d}\mu, \qquad f\in C_0^\infty\big(\mathbb{R}^n\big), \tag{2.7}$$

where $\langle\cdot,\cdot\rangle$ stands the usual inner product in $\mathbb{R}^n$. In particular, we have

$$\lambda_1(U)\geqslant \inf_{x\in\mathbb{R}^n}\lambda_{\min}(\mathrm{Hess}(U)(x)), \tag{2.8}$$

where $\lambda_{\min}(M)$ is the minimal eigenvalue of the matrix M.

Proof. The proof of (2.7) is mainly a use of integration by parts formula. Because $Lf=\sum_i\big(\partial_{ii}f-\partial_iU\partial_if\big)$, we have

$$\langle\nabla f,\nabla Lf\rangle=\sum_j\partial_jf\sum_i\partial_j\big(\partial_{ii}f-\partial_iU\partial_if\big)=\sum_{i,j}\partial_jf\big(\partial_{iij}f-\partial_{ij}f\partial_iU-\partial_if\partial_{ij}U\big).$$

Next,

$$\begin{aligned}\frac1Z\int_{\mathbb{R}^n}\sum_j\partial_jf\sum_i\big(\partial_{iij}f-\partial_{ij}f\partial_iU\big)e^{-U}&=\frac1Z\int_{\mathbb{R}^n}\sum_j\partial_jf\sum_i\partial_i\Big(\partial_{ij}fe^{-U}\Big)\\ &=-\frac1Z\int_{\mathbb{R}^n}\sum_{i,j}\big(\partial_{ij}f\big)^2e^{-U}\\ &=-\int_{\mathbb{R}^n}\sum_{i,j}\big(\partial_{ij}f\big)^2\mathrm{d}\mu, \quad f\in C_0^\infty\big(\mathbb{R}^n\big).\end{aligned}$$

Noting that μ is a probability measure and the diffusion coefficients are constants, the Dirichlet form is regular (cf. [12; condition (4.13)] for instance). Actually, the martingale problem for L is well posed. Thus, $LC_0^\infty(\mathbb{R}^n) \subset C_0^\infty(\mathbb{R}^n) \subset \mathscr{D}(L)$, and so

$$\|Lf\|^2 = \int_{\mathbb{R}^n} Lf \cdot Lf \mathrm{d}\mu = -\int_{\mathbb{R}^n} \langle \nabla Lf, \nabla f\rangle \mathrm{d}\mu, \qquad f \in C_0^\infty(\mathbb{R}^n).$$

Combining these facts together, we get (2.7).

To prove the last assertion, applying Theorem 2.1 and (2.7), we get

$$\begin{aligned}\lambda_1(L) &= \inf_{f\in\mathscr{D}(L),\, f\neq \text{const}} \frac{\|Lf\|^2}{D(f)}\\ &= \inf_{f\in C_0^\infty(\mathbb{R}^n),\, f\neq \text{const}} \frac{\|Lf\|^2}{D(f)}\\ &\geqslant \inf_{f\in C_0^\infty(\mathbb{R}^n),\, f\neq \text{const}} \int_{\mathbb{R}^n} \langle \mathrm{Hess}(U)\nabla f, \nabla f\rangle \mathrm{d}\mu \Big/ D(f)\\ &\geqslant \inf_{x\in\mathbb{R}^n} \lambda_{\min}(\mathrm{Hess}(U)(x)). \quad \square\end{aligned}$$

Remark 2.4. Actually, under the assumption of Corollary 2.3, the Bakry-Emery criterion (cf. [14] or [7; Corollary 1.6]) implies a stronger conclusion:

$$\sigma(U) \geqslant \inf_{x\in\mathbb{R}^n} \lambda_{\min}(\mathrm{Hess}(U)(x)). \tag{2.9}$$

A simple counterexample for which (2.8) and (2.9) are not effective is the following. This example also shows that (1.4) is an improvement of (2.8).

Example 2.5. Consider the two-dimensional case. Let

$$U(x) = x_1^4 + x_2^4 - \beta\big(x_1^2 + x_2^2\big) + 2Jx_1x_2$$

with constants $\beta \geqslant 0$ and $J \in \mathbb{R}$. Then $\inf_{x\in\mathbb{R}^2} \lambda_{\min}(\mathrm{Hess}(U)(x)) \leqslant 0$ and U is not convex at infinity, but $\lambda_1(U) > 0$ in a region of $(\beta, J) \subset \mathbb{R} \times \mathbb{R}_+$.

Proof. First, we have

$$\mathrm{Hess}(U)(x) = \begin{pmatrix} 12x_1^2 - 2\beta & 2J \\ 2J & 12x_2^2 - 2\beta \end{pmatrix}.$$

Because for the matrix

$$A = \begin{pmatrix} c_1 & 2J \\ 2J & c_2 \end{pmatrix},$$

we have $\lambda_{\min}(A) = 2^{-1}\big(c_1 + c_2 - \sqrt{(c_1 - c_2)^2 + 16J^2}\,\big)$. Hence

$$\lambda_{\min}(\mathrm{Hess}(U)(x)) = 2\min_{x_1,x_2} \Big\{3\big(x_1^2 + x_2^2\big) - \beta - \sqrt{9\big(x_1^2 - x_2^2\big)^2 + J^2}\,\Big\}.$$

Setting $x_1 = x_2 = 0$, we get

$$\inf_{x\in\mathbb{R}^2} \lambda_{\min}(\mathrm{Hess}(U)(x)) \leqslant -2(\beta + |J|) \leqslant 0.$$

Next, since

$$\lim_{|x_1|\to\infty} \Big(3x_1^2 - \beta - \sqrt{9x_1^4 + J^2}\,\Big) = \lim_{z\to 0} \frac{3 - \sqrt{9 + J^2 z^2}}{z} - \beta = -\beta,$$

we have

$$\varliminf_{|x|\to\infty} \lambda_{\min}(\mathrm{Hess}(U)(x)) \leqslant \varliminf_{x_2=0,\ |x_1|\to\infty} \lambda_{\min}(\mathrm{Hess}(U)(x)) \leqslant -2\beta \leqslant 0.$$

This means that U is not convex at infinity. The last assertion of the example is the one of the main aims of this paper and it is even true in the higher dimensions (cf. Theorem 6.3 below). $\square$

3. Proofs of Theorems 1.1 and 1.2 and some remarks. As a preparation, we prove a result which is an improvement of (2.8) and [7; Proposition 3.1]. We adopt the notation given in §1.

Proposition 3.1. We have

$$\lambda_1(U) \geqslant \inf_{x\in\mathbb{R}^n} \lambda_{\min}\big(\widetilde{\mathrm{Hess}}(U)(x)\big). \tag{3.1}$$

Proof. First, applying Theorem 2.1 and (2.7) to the ith marginal, we have

$$\int_{\mathbb{R}} \big[(\partial_{ii} f)^2 + (\partial_{ii} U)(\partial_i f)^2\big] \mathrm{d}\mu_U^{x_{\setminus i}} \geqslant \lambda_1^{x_{\setminus i}} \int_{\mathbb{R}} (\partial_i f)^2 \mathrm{d}\mu_U^{x_{\setminus i}}, \qquad f \in C_0^\infty\big(\mathbb{R}^n\big). \tag{3.2}$$

Next, denote by $\mathrm{Hess}_0(U)$ the symmetric matrix obtained from the Hessen matrix $\mathrm{Hess}(U)$ replacing the diagonal elements with zero. Then, by (3.2), we have

$$\begin{aligned}
&\int_{\mathbb{R}^n} \Big[\sum_{i,j} (\partial_{ij} f)^2 + \langle \mathrm{Hess}(U)\nabla f, \nabla f\rangle\Big] \mathrm{d}\mu_U \\
&\geqslant \sum_i \int_{\mathbb{R}^n} \Big\{\int_{\mathbb{R}} \big[(\partial_{ii} f)^2 + (\partial_{ii} U)(\partial_i f)^2\big] \mathrm{d}\mu_U^{x_{\setminus i}}\Big\} \mathrm{d}\mu_U - \sum_i \int_{\mathbb{R}^n} \big[(\partial_{ii} U)(\partial_i f)^2\big] \mathrm{d}\mu_U \\
&\qquad + \int_{\mathbb{R}^n} \langle \mathrm{Hess}(U)\nabla f, \nabla f\rangle \mathrm{d}\mu_U \\
&\geqslant \sum_i \int_{\mathbb{R}^n} \Big\{\lambda_1^{x_{\setminus i}} \int (\partial_i f)^2 \mathrm{d}\mu_U^{x_{\setminus i}}\Big\} \mathrm{d}\mu_U + \int_{\mathbb{R}^n} \langle \mathrm{Hess}_0(U)\nabla f, \nabla f\rangle \Big] \mathrm{d}\mu_U \\
&= \sum_i \int_{\mathbb{R}^n} \lambda_1^{x_{\setminus i}} (\partial_i f)^2 \mathrm{d}\mu_U + \int_{\mathbb{R}^n} \big\langle \mathrm{Hess}_0(U)\nabla f, \nabla f\big\rangle \mathrm{d}\mu_U \\
&= \int_{\mathbb{R}^n} \langle \widetilde{\mathrm{Hess}}(U)\nabla f, \nabla f\rangle \mathrm{d}\mu_U \\
&\geqslant \inf_{x\in\mathbb{R}^n} \lambda_{\min}\big(\widetilde{\mathrm{Hess}}(U)(x)\big) \int_{\mathbb{R}^n} |\nabla f|^2 \mathrm{d}\mu_U, \qquad f \in C_0^\infty\big(\mathbb{R}^n\big).
\end{aligned} \tag{3.3}$$

Now, the required assertion follows from the proof of the last assertion of Corollary 2.3. $\square$

From the proof of Proposition 3.1, it is clear that the only argument where we may lose somewhat is the first inequality of (3.3), since the terms $\sum_{i\neq j}(\partial_{ij}f)^2$ are ignored there. Hence the estimate (3.1) is mainly meaningful if the interactions are not strong. The interacting potentials considered in this paper are rather simple; for general interactions, one needs some "block estimates" which are not touched here, instead of the "single-site estimates" studied in this paper.

The shorthand of (3.1) is that the minimal eigenvalue $\lambda_{\min}\big(\widetilde{\text{Hess}}(U)\big)$ may not be computable in practice. For this, we need the second variational procedure. To do so, let $s=\min_i\big\{\lambda_1^{x\setminus i}-\sum_{j:j\neq i}|\partial_{ij}U|\big\}$ and define

$$q_{ij}=\begin{cases}|\partial_{ij}U|, & \text{if } i\neq j\\ s-\lambda_1^{x\setminus i}, & \text{if } i=j.\end{cases}$$

Then, $Q:=(q_{ij})$, depending on x, is a symmetric Q-matrix, not necessarily conservative $\big($i.e., $\sum_j q_{ij}\leqslant 0\big)$.

Proof of Theorem 1.1. The first estimate in (1.4) follows from Proposition 3.1. Next, by [15; Theorem 1.1], we have

$$\lambda_{\min}(-Q)\geqslant \sup_{w>0}\min_i\big[-Qw/w\big](i), \tag{3.4}$$

where $Qw(i)=\sum_j q_{ij}w_j$. We remark that the sign of the equality in (3.4) holds once Q is irreducible (cf. [15; Proposition 4.1]). Noting that for every symmetric matrix $B=(b_{ij})$ with nonnegative diagonals and any vector w, we have

$$\langle w,Bw\rangle=\sum_i b_{ii}w_i^2+2\sum_{i\neq j}b_{ij}w_iw_j\geqslant\sum_i b_{ii}w_i^2-2\sum_{i\neq j}|b_{ij}w_iw_j|=\big\langle|w|,\widetilde{B}|w|\big\rangle,$$

where $\widetilde{B}=(\tilde{b}_{ij}):\tilde{b}_{ii}=b_{ii},\ \tilde{b}_{ij}=-|b_{ij}|$ for $i\neq j$ and $|w|=(|w_i|)$. Letting w^* be a vector with $\langle w^*,w^*\rangle=1$ such that $\lambda_{\min}(B)=\langle w^*,Bw^*\rangle$, it follows that

$$\lambda_{\min}(B)\geqslant\big\langle|w^*|,\widetilde{B}|w^*|\big\rangle\geqslant\lambda_{\min}\big(\widetilde{B}\big)\langle|w|^*,|w|^*\rangle=\lambda_{\min}\big(\widetilde{B}\big).$$

Based on this fact and as an application of (3.4), we get

$$\begin{aligned}\lambda_{\min}\big(\widetilde{\text{Hess}}(U)(x)\big)&\geqslant\lambda_{\min}\big(\text{diag}(s)-Q\big)\\&=s+\lambda_{\min}(-Q)\\&\geqslant s+\max_{w>0}\min_i\bigg[-s+\lambda_1^{x\setminus i}-\sum_{j:j\neq i}q_{ij}w_j/w_i\bigg]\\&=\max_{w>0}\min_i\bigg[\lambda_1^{x\setminus i}-\sum_{j:j\neq i}q_{ij}w_j/w_i\bigg].\end{aligned}\tag{3.5}$$

Combining this with the first estimate in (1.4), we get the second one in (1.4), and so complete the proof of Theorem 1.1. $\square$

Proof of Theorem 1.2. In the above proof, replacing $\lambda_1^{x\setminus i}$, s and q_{ii} with $\eta_i^{x\setminus i}$, $\underline{s}(x)$ and $q_{ii}(x)$, respectively, but keep q_{ij} $(i \neq j)$ to be the same, we obtain

$$\lambda_{\min}\big(\widetilde{\text{Hess}}(U)(x)\big) \geqslant \underline{s}(x) + \lambda_{\min}(-Q(x)).$$

By Proposition 3.1, it suffices to estimate $\lambda_{\min}(-Q(x))$. Note that $\lambda_{\min}(-Q(x))$ is nothing but the principal (Dirichlet) eigenvalue of $Q(x)$, often denoted by $\lambda_0(Q(x))$. Because $Q(x)$ is symmetric, and so its symmetrizing measure is just the uniform distribution on $\{1, 2, \dots, n\}$. Now the conclusion of Theorem 1.2 follows from [16; Theorem 1.1] plus some computations. $\square$

We conclude this section with some remarks.

Let $d_i = -q_{ii} - \sum_{j\neq i} q_{ij}$. By setting w_i = constant in (3.4), it follows that $\lambda_{\min}(-Q) \geqslant \min_i d_i$. The sign of the equality holds if (d_i) is a constant. Otherwise, this well-known simplest conclusion is usually rough. For instance, take

$$Q = \begin{pmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -3 \end{pmatrix}.$$

Then $\lambda_{\min}(-Q) = 2 - \sqrt{3} > 0$ (the equality of (3.4) is attained at the positive eigenvector $w = \big(2+\sqrt{3}\,, 1+\sqrt{3}\,, 1\big)$ but $\min_i d_i = 0$. This shows that the use of the variational formula (3.4) is necessary to produce sharper lower bounds.

When $\partial_{ij} U \leqslant 0$ for all $i \neq j$, then $\widetilde{\text{Hess}}(U) = \text{diag}(s) - Q$, and so the sign of the first equality in (3.5) holds. In this case, the estimate (3.5) is quite sharp, since so is (3.4). However, for general $\partial_{ij}U$ $(i \neq j)$, the lower bound in (3.5) may be less effective but we do not have a variational formula as (3.4) in such a general situation.

For a given symmetric matrix $B = (b_{ij})$ $\big(\widetilde{\text{Hess}}(U),$ for instance$\big)$, the classical variational formula, which is especially powerful for upper bounds, is as follows.

$$\begin{aligned}\lambda_{\min}(B) &= \inf\Big\{\sum_{i,j} w_i b_{ij} w_j : \sum_i w_i^2 = 1\Big\}\\ &= \inf\bigg\{\sum_i\bigg(b_{ii} + \sum_{j:j\neq i} b_{ij}\bigg)w_i^2 - \frac{1}{2}\sum_{i,j} b_{ij}(w_j - w_i)^2 : \sum_i w_i^2 = 1\bigg\}.\end{aligned} \tag{3.6}$$

For a given symmetrizable Q-matrix (q_{ij}) with symmetric probability measure μ, set

$$D(f) = \frac{1}{2}\sum_{i,j}\mu_i q_{ij}(f_j - f_i)^2 + \sum_i \mu_i d_i f_i^2,$$

where $d_i = -q_{ii} - \sum_{j\ne i} q_{ij}$ as defined before. Then, an alternative formula of (3.6), in terms of the Donsker-Varadhan's theory of large deviations, goes as follows.

$$\begin{aligned}\lambda_{\min}(-Q) &= \inf_f \Big\{D(f) : \sum_i \mu_i f_i^2 = 1\Big\}\\ &= \inf_{\alpha\geqslant 0} \Big\{D\big(\sqrt{\mathrm{d}\alpha/\mathrm{d}\mu}\big) : \sum_i \alpha_i = 1\Big\}\\ &= \inf_{\alpha\geqslant 0} \Big\{I(\alpha) + \sum_i \alpha_i d_i : \sum_i \alpha_i = 1\Big\}\\ &= \inf_{\alpha\geqslant 0} \Big\{ - \inf_{u>0} \sum_{i,j} \alpha_i q_{ij}(u_j - u_i)/u_i + \sum_i \alpha_i d_i : \sum_i \alpha_i = 1\Big\}\\ &= \inf_{\alpha\geqslant 0} \Big\{\frac{1}{2}\sum_{i,j} \big(\sqrt{\alpha_i q_{ij}} - \sqrt{\alpha_j q_{ji}}\big)^2 + \sum_i \alpha_i d_i : \sum_i \alpha_i = 1\Big\},\end{aligned} \tag{3.7}$$

where I is the I-functional in the theory of large deviations. Refer to [17; Proof of Theorem 8.17] for more details. In other words, the large deviation principle provides an alternative description of the classical variational formula, but not (3.4), for which one needs a variational formula for the Dirichlet forms (cf. [15]).

Finally, we remark that the proof of Proposition 3.1 can be also used in the study of other inequalities. The details are omitted here since they are not used subsequently (cf. [7]). The next one is a partial extension of (2.5).

Proposition 3.2. Under the assumption of Corollary 2.3, we have for every invertible, nonnegative and diagonal matrix D, the largest constant κ:

$$\kappa\big|\mathrm{Cov}_{\mu_U}(f,g)\big| \leqslant \bigg(\int |D\nabla f|^2 \mathrm{d}\mu_U \int |D^{-1}\nabla f|^2 \mathrm{d}\mu_U\bigg)^{1/2}, \qquad f, g \in C_0^\infty\big(\mathbb{R}^n\big) \tag{3.8}$$

satisfies

$$\kappa \geqslant \inf_x \hat{\lambda}_{\min}\big(D\,\widetilde{\mathrm{Hess}}(U)\,D^{-1}(x)\big), \tag{3.9}$$

where $\hat{\lambda}_{\min}(M) = \max\{c : M \geqslant c\,\mathrm{Id}\}$.

Before moving further, let us make some remarks about the proof of Proposition 3.2. Note that

$$\int |D\nabla f|^2 \mathrm{d}\mu = \int \langle D^2 \nabla f, \nabla f\rangle \mathrm{d}\mu$$

which is the Dirichlet form corresponding to the diffusion operator with diffusion coefficients D^2 and potential U. Denote by $\lambda_1\big(D^2, U\big)$ the spectral gap of the last operator, then we have

$$\begin{aligned}\big|\mathrm{Cov}_\mu(f,g)\big|^2 &\leqslant \mathrm{Var}_\mu(f)\mathrm{Var}_\mu(g)\\ &\leqslant \frac{1}{\lambda_1\big(D^2,U\big)\,\lambda_1\big(D^{-2},U\big)} \int \langle D^2\nabla f, \nabla f\rangle \mathrm{d}\mu \int \langle D^{-2}\nabla g, \nabla g\rangle \mathrm{d}\mu.\end{aligned} \tag{3.10}$$

Hence, we obtain a lower bound of the optimal constant in (3.8):

$$\kappa \geqslant \sqrt{\lambda_1\big(D^2, U\big)\,\lambda_1\big(D^{-2}, U\big)}. \tag{3.11}$$

The proof is quite natural. Furthermore, by setting D to be the identity matrix, we obtain (2.5) with sharp constant. However, the estimate (3.11) is usually not sharp in the general case. Note that the sign of the last equality in (3.11) holds if f and g are the correspondent eigenfunctions with respect to the operators, but the sign of the first equality in (3.11) holds iff f and g are proportional almost surely (due to the use of the Cauchy-Schwarz inequality). This can happen only if D is trivial: all the diagonals of D are equal.

A better way to study (3.8) is using the semigroup's approach. Write

$$\begin{aligned}
\mathrm{Cov}_\mu(f,g) &= \int (f-\mu(f))g\mathrm{d}\mu \\
&= -\int\bigg(\int_0^\infty \frac{\mathrm{d}}{\mathrm{d}t}P_t f\mathrm{d}t\bigg)g\mathrm{d}\mu \\
&= -\int_0^\infty \bigg(\int gLP_t f\mathrm{d}\mu\bigg)\mathrm{d}t \\
&= \int_0^\infty \bigg(\int \langle \nabla P_t f, \nabla g\rangle \mathrm{d}\mu\bigg)\mathrm{d}t.
\end{aligned}$$

Now, as a good application of the Cauchy-Schwarz inequality, we get

$$\big|\mathrm{Cov}_\mu(f,g)\big| \leqslant \bigg[\int_0^\infty \bigg(\int |D\nabla P_t f|^2\mathrm{d}\mu\bigg)^{1/2}\mathrm{d}t\bigg]\bigg(\int |D^{-1}\nabla g|^2\mathrm{d}\mu\bigg)^{1/2}.$$

The problem is now reduced to study the decay of $\int |D\nabla P_t f|^2\mathrm{d}\mu$ in t (cf. [7]).

Similarly to Proposition 3.1, as checked by Feng Wang in 2002, we have the following result which improves (2.9), but may be weaker than Theorem 1.3.

Proposition 3.3. Under the assumption of Corollary 2.3, we have

$$\sigma(U) \geqslant \inf_x \lambda_{\min}\big(\overline{\mathrm{Hess}}(U)\big), \tag{3.12}$$

where

$$\big(\overline{\mathrm{Hess}}(U)\big)_{ij} = \begin{cases} \zeta^{x_{\setminus i}}, & \text{if } j = i \\ (\mathrm{Hess}(U))_{ij} & \text{if } j \neq i, \end{cases}$$

$\zeta^{x_{\setminus i}}$ is the optimal constant in the inequality

$$\zeta^{x_{\setminus i}} \int f(\partial_i \log f)^2 \mathrm{d}\mu_U^{x_{\setminus i}} \leqslant \int f\Gamma^i(\log f)\mathrm{d}\mu_U^{x_{\setminus i}}, \qquad 0 \leqslant f \in C_0^\infty\big(\mathbb{R}^n\big), \tag{3.13}$$

and

$$\Gamma^i(f) = (\partial_{ii} f)^2 + (\partial_{ii} U)(\partial_i f)^2.$$

4. One-dimensional case. Explicit estimates. The operator now becomes $L = \mathrm{d}^2/\mathrm{d}x^2 - u'(x)\mathrm{d}/\mathrm{d}x$. Write $b(x) = -u'(x)$. Then b must have a real root. Otherwise, without loss of generality, let $u' \geqslant \varepsilon > 0$. Then $-u$ is strictly decreasing, and so

$$\infty > Z := \int_{\mathbb{R}} e^{-u} \geqslant \int_{-\infty}^{0} e^{-u} > e^{-u(0)} \int_{-\infty}^{0} 1 = \infty,$$

which is a contradiction.

Unless otherwise stated, throughout this section, we consider the operator

$$L = a(x)\frac{\mathrm{d}^2}{\mathrm{d}x^2} + b(x)\frac{\mathrm{d}}{\mathrm{d}x}.$$

Assume that $a \in C(\mathbb{R})$, $a > 0$ and $Z = \int_{\mathbb{R}} e^{C(x)}/a(x) < \infty$, where $C(x) = \int_0^x b/a$. Define $\mu(\mathrm{d}x) = (Za(x))^{-1}e^{C(x)}\mathrm{d}x$. Recall that

$$\lambda_1(L) = \inf\{D(f) : f \in C^1(\mathbb{R}),\ \mu(f) = 0,\ \mu(f^2) = 1\},$$

where $D(f) = \int_{\mathbb{R}} a f'^2 \mathrm{d}\mu$.

Let θ be a fixed real root of b. Choose $K = K_\theta \in C(\mathbb{R} \setminus \{\theta\})$ such that K is increasing (i.e., non-decreasing) in x when $|x-\theta|$ increases, $K(\theta \pm 0) > -\infty$, and moreover

$$K(r) \leqslant \inf_{x:\,\pm(x-r)>0} \big[-b(x)/(x-\theta)\big] \qquad \text{for all } \pm(r-\theta) > 0, \tag{4.1}$$

where and in what follows, the notation "$\pm$" means that there are two cases: one takes "$+$" (resp., "$-$") everywhere in the statement. Define

$$F(s) = F^r(s) = \int_\theta^s \frac{u-\theta}{a(u)}\big[K(r) - K(u)\big]\mathrm{d}u, \qquad s,\ r \in \mathbb{R}, \tag{4.2}$$

$$\begin{aligned} \delta_\pm(K) &= \sup_{r:\,\pm(r-\theta)>0} K(r) \inf_{s:\,\pm(r-\theta)>\pm(s-\theta)>0} \frac{(s-\theta)\exp[-F(s)]}{\int_\theta^s \exp[-F(u)]\mathrm{d}u} && (4.3)\\ &\geqslant \sup_{r:\,\pm(r-\theta)>0} K(r)\exp[-F(r)]. && (4.4) \end{aligned}$$

The next result is a modification of [12; Corollary 3.5]. It is specially useful for those b growing at least linear.

Theorem 4.1.

(1) By using the above notations, we have

$$\lambda_1(L) \geqslant \delta_+(K) \wedge \delta_-(K). \tag{4.5}$$

(2) Suppose additionally that K is a piecewise C^1-function, then we have

$$\delta_\pm(K) \geqslant K(r_\pm)\exp\bigg[-\int_\theta^{r_\pm}\bigg(\int_\theta^x \frac{u-\theta}{a(u)}\mathrm{d}u\bigg)\mathrm{d}K(x)\bigg], \tag{4.6}$$

where

$$r_\pm = \pm\infty, \qquad \text{if } \lim_{r\to\pm\infty} K(r)\int_\theta^r \frac{u-\theta}{a(u)}\mathrm{d}u \leqslant 1, \tag{4.7}$$

and otherwise, $r_\pm$ is the unique solution to the equation

$$K(r)\int_\theta^r \frac{u-\theta}{a(u)}\mathrm{d}u = 1, \qquad \pm(r-\theta) > 0. \tag{4.8}$$

Proof. (a) First, consider the half-line (θ,∞). Assume that $K(r_1) > 0$ for some $r_1 \in (\theta,\infty)$. Otherwise, (4.5) becomes trivial. Fix $r = r_1$ and define

$$f_+(x) = \int_\theta^x \mathrm{d}y \exp[-F(y\wedge r_1)], \qquad x \geqslant \theta.$$

Then, we have $f_+ > 0$ on (θ,∞), $f_+(\theta) = 0$, $f'_+(\theta) = 1$ and

$$\begin{aligned} &f'_+(x) = \exp[-F(x\wedge r_1)] > 0,\\ &f''_+(x) = -\frac{x-\theta}{a(x)}\big[K(r_1) - K(x\wedge r_1)\big]f'_+(x) \leqslant 0, \qquad x \geqslant \theta. \end{aligned}$$

Since $a \in C(\mathbb{R})$, $a > 0$, $K \in C(\mathbb{R}\setminus\{\theta\})$ and $K(\theta+)$ is finite, we have $f_+ \in C^2(\theta,\infty)$.

Next, because K is increasing on (θ,∞) and $K(x) \leqslant -b(x)/(x-\theta)$ for all $x > \theta$, we have

$$\begin{aligned} -(af''_+ + bf'_+)(x) &= \big\{(x-\theta)[K(r_1) - K(x\wedge r_1)] - b(x)\big\}f'_+(x)\\ &\geqslant \big\{(x-\theta)K(r_1) - (x-\theta)K(x) - b(x)\big\}f'_+(x)\\ &\geqslant (x-\theta)K(r_1)f'_+(x), \qquad x > \theta. \end{aligned} \tag{4.9}$$

Since $f''_+ \leqslant 0$, f'_+ is decreasing. By the Cauchy mean value theorem, it follows that $(x-\theta)/f_+(x)$ is increasing on (θ,∞). Hence, by (4.9), we obtain

$$\begin{aligned} -\left[\frac{af''_+ + bf'_+}{f_+}\right](x) \geqslant \frac{r_1-\theta}{f_+(r_1)}K(r_1)f'_+(x) = \frac{r_1-\theta}{f_+(r_1)}K(r_1)f'_+(r_1),\\ x \geqslant r_1. \end{aligned} \tag{4.10}$$

Combining (4.9) with (4.10), it follows that

$$\inf_{x>\theta}\left[-\frac{af''_+ + bf'_+}{f_+}\right] \geqslant K(r_1)\inf_{s\in(\theta,r_1)}\frac{(s-\theta)f'_+(s)}{f_+(s)}.$$

By (4.3), we have thus obtained

$$\inf_{x>\theta}\left[-\frac{af''_+ + bf'_+}{f_+}\right] \geqslant \delta_+(K). \tag{4.11}$$

(b) Next, consider the half-line $(-\infty,\theta)$. The proof is parallel to (a). Let $K(r_1)>0$ for some $r_1<\theta$. Fix $r=r_1$ and define

$$f_-(x)=\int_\theta^x \mathrm{d}y\exp[-F(y\vee r_1)],\qquad x\leqslant\theta.$$

Then $f_-<0$ on $(-\infty,\theta)$, $f_-(\theta)=0$, $f_-'>0$, $f_-'(\theta)=1$ and

$$f_-''(x)=-\frac{x-\theta}{a(x)}[K(r_1)-K(x\vee r_1)]f_-'(x)\geqslant 0$$

for all $x\leqslant\theta$. Moreover $f_-\in C^2(-\infty,\theta)$. Then

$$\begin{aligned}-(af_-''+bf_-')(x)&=\big\{(x-\theta)[K(r_1)-K(x\vee r_1)]-b(x)\big\}f_-'(x)\\&\leqslant\big\{(x-\theta)K(r_1)-(x-\theta)K(x)-b(x)\big\}f_+'(x)\\&\leqslant(x-\theta)K(r_1)f_-'(x),\qquad x<\theta.\end{aligned}$$

Since $f_-<0$ and $f_-''\geqslant 0$, we have

$$-\left[\frac{af_-''+bf_-'}{f_-}\right](x)\geqslant\frac{r_1-\theta}{f_-(r_1)}K(r_1)f_-'(x)=\frac{r_1-\theta}{f_-(r_1)}K(r_1)f_-'(r_1),\qquad x\leqslant r_1.$$

Combining the last two inequalities with (4.3), we get

$$\inf_{x<\theta}\left[-\frac{af_-''+bf_-'}{f_-}\right]\geqslant\sup_{r_1<\theta}K(r_1)\inf_{s\in(\theta,r_1)}\frac{(s-\theta)f_-'(s)}{f_-(s)}=\delta_-(K).$$

Finally, let $f=f_+I_{[\theta,\infty)}+f_-I_{(-\infty,\theta)}$. Then $f\in C^2(\mathbb{R})$ and

$$\begin{aligned}\inf_{x\neq\theta}\left[-\frac{af''+bf'}{f}\right](x)&=\left[\inf_{x>\theta}-\frac{af_+''+bf_+'}{f_+}(x)\right]\wedge\left[\inf_{x<\theta}-\frac{af_-''+bf_-'}{f_-}(x)\right]\\&\geqslant\delta_+(K)\wedge\delta_-(K).\end{aligned}$$

The estimate (4.5) now follows from the last assertion of [12; Theorem 3.1].

(c) To prove (4.4), noticing that K is monotone, we may apply the integration by parts formula and rewrite F as follows.

$$\begin{aligned}F(r)&=\int_\theta^r\frac{u-\theta}{a(u)}[K(r)-K(u)]\mathrm{d}u\\&=K(r)\int_\theta^r\frac{u-\theta}{a(u)}\mathrm{d}u-\int_\theta^r K(u)\mathrm{d}\bigg(\int_\theta^u\frac{z-\theta}{a(z)}\mathrm{d}z\bigg)\\&=\int_\theta^r K'(u)\bigg(\int_\theta^u\frac{z-\theta}{a(z)}\mathrm{d}z\bigg)\mathrm{d}u,\qquad r\neq\theta.\end{aligned}\tag{4.12}$$

By the assumption on K, it follows that $F\geqslant 0$, $F(r)$ is increasing in r as $|r-\theta|$ increases. Hence, by (4.2), we have

$$\begin{aligned}\delta_\pm(K)&\geqslant\sup_{r:\,\pm(r-\theta)>0}K(r)\inf_{s:\,\pm(r-\theta)>\pm(s-\theta)>0}\exp[-F(s)]\\&=\sup_{r:\,\pm(r-\theta)>0}K(r)\exp[-F(r)].\end{aligned}$$

The proof of (4.4) is done.

(d) The second part of the theorem is to compute $\sup_{r\neq\theta} G(r)$, where $G(r) = K(r)\exp[-F(r)]$. The answer is given by (4.6). To do so, first consider the half line (θ,∞). Because K is a piecewisely C^1, we may assume that $\big(\theta,\infty\big) = \cup_i (c_i, d_i]$, $K \in C^1(c_i, d_i)$ and $K' \geqslant 0$ on (c_i, d_i) for every i. By (4.12), we have for every i,

$$G'(r) = K'(r)\left[1 - K(r)\int_\theta^r \frac{u-\theta}{a(u)}\mathrm{d}u\right]\exp[-F(r)], \qquad r \in (c_i, d_i). \tag{4.13}$$

Let $\lim_{r\to\infty} K(r)\int_\theta^r \frac{u-\theta}{a(u)}\mathrm{d}u \leqslant 1$ and set $\tilde\theta_+ = \inf\big\{r > \theta : K(r) > 0\big\}$. Note that K is increasing, $K > 0$ on $\big(\tilde\theta_+,\infty\big)$, and so $K(r)\int_\theta^r \frac{u-\theta}{a(u)}\mathrm{d}u$ is strictly increasing on $\big(\tilde\theta_+,\infty\big)$, but is less or equal to zero on $\big(\theta,\tilde\theta_+\big)$ when $\theta < \tilde\theta_+$. It follows that $K(r)\int_\theta^r \frac{u-\theta}{a(u)}\mathrm{d}u \leqslant 1$ for all $r \in (\theta,\infty)$. By (4.13), we have $G'(r) \geqslant 0$ on every (c_i, d_i) since so does $K'(r)$. This fact plus the continuity of G implies that $\sup_{r>\theta} G(r) = \lim_{r\to\infty} G(r)$.

Otherwise, we have

$$\lim_{r\to\infty} K(r)\int_\theta^r \frac{u-\theta}{a(u)}\mathrm{d}u > 1 \quad \text{and}$$
$$\lim_{r\to\tilde\theta_+ +} K(r)\begin{cases} = 0 < \Big(\int_\theta^{\tilde\theta_+} \frac{u-\theta}{a(u)}\mathrm{d}u\Big)^{-1} & \text{if } \theta < \tilde\theta_+ \\ < \infty = \Big(\int_\theta^{\theta} \frac{u-\theta}{a(u)}\mathrm{d}u\Big)^{-1} & \text{if } \tilde\theta_+ = \theta. \end{cases}$$

Since $K(r)$ is increasing and $\Big(\int_\theta^r \frac{u-\theta}{a(u)}\mathrm{d}u\Big)^{-1}$ is strictly decreasing, the curves $K(r)$ and $\Big(\int_\theta^r \frac{u-\theta}{a(u)}\mathrm{d}u\Big)^{-1}$ must have uniquely an intersection on $\big(\tilde\theta_+,\infty\big)$, or equivalently on (θ,∞). So we have $\sup_{r>\theta} G(r) = \sup_{r>\tilde\theta_+} G(r) = G(r_+)$, where r_+ is the unique solution to the equation (4.8).

The proof of the assertions on $(-\infty,\theta)$ is parallel. $\square$

The next two examples illustrate the applications of Theorem 4.1, and are treated several times in the paper.

Example 4.2. Let $u(x) = \alpha x^2 + \beta x$ for some constants $\alpha > 0$ and $\beta \in \mathbb{R}$, and $a(x) \equiv 1$. Then we have $\lambda_1(L_{\alpha,\beta}) \geqslant \delta_+(K) \wedge \delta_-(K) = 2\alpha$ which is exact.

Proof. Since $-b(x) = -2\alpha x - \beta$, we have root $\theta = -\beta/(2\alpha)$, and so $-b(x)/(x-\theta) = 2\alpha$. Thus, $K(r) =$ constant 2α. By (4.3), we get $\delta_\pm(K) = 2\alpha$ as claimed. It is easy to check that the estimate is exact, since the corresponding eigenfunction is linear. $\square$

Example 4.3. Let $u(x) = x^4 - \beta_1 x^2 + \beta_2 x$ for some constants $\beta_1, \beta_2 \in \mathbb{R}$ and $a(x) \equiv 1$. Then we have

$$\lambda_1(L_{\beta_1,\beta_2}) \geqslant \delta_+(K) \wedge \delta_-(K) \geqslant \frac{\sqrt{\beta_1^2+2}-\beta_1}{\sqrt{e}}\exp\left[-\frac{1}{2}\beta_1\Big(\beta_1 + \sqrt{\beta_1^2+2}\Big)\right]$$

uniformly in β_2. When $\beta_2 = 0$, we have

$$\lambda_1(L_{\beta_1,\,\beta_2}) \geqslant \delta_+(K) \wedge \delta_-(K) \geqslant \frac{\sqrt{\beta_1^2+8}-\beta_1}{\sqrt{e}} \exp\left[-\frac{1}{8}\beta_1\Big(\beta_1+\sqrt{\beta_1^2+8}\Big)\right].$$

Proof. First, we have $b(x) = -u'(x) = -4x^3 + 2\beta_1 x - \beta_2$. Let θ be a real root of u'. For instance, we may take

$$\theta = \begin{cases} 0 & \text{if } \beta_2 = 0 \\ -\Big(\frac{\beta_2}{4}\Big)^{1/3} & \text{if } \beta_1 = 0 \\ 2\sqrt{\frac{-\beta_1}{6}}\,\sinh\Big(\frac{1}{3}\mathrm{arc\,sinh}\,C\Big) & \text{if } \beta_1 < 0 \\ 2\sqrt{\frac{\beta_1}{6}}\,\mathrm{sgn}\,(C)\cosh\Big(\frac{1}{3}\mathrm{arc\,cosh}\big(\mathrm{sgn}\,(C)\,C\big)\Big) & \text{if } \beta_1 > 0 \text{ and } |C| > 1 \\ 2\sqrt{\frac{\beta_1}{6}}\,\cos\Big(\frac{4}{3}\pi + \frac{1}{3}\mathrm{arc\,cos}\,C\Big) & \text{if } \beta_1 > 0 \text{ and } |C| \leqslant 1, \end{cases}$$

where $C = \beta_2\Big(\frac{3}{2|\beta_1|}\Big)^{3/2}$. The reason we choose $4\pi/3$ rather than 0 or $2\pi/3$ in the last line is for the consistency of the case $\beta_2 = 0$. However, in what follows, we will not use the explicit formula of θ, we are going to work out only the estimate uniform in θ. Because

$$\frac{-b(x)}{x-\theta} = 4(x-\theta)^2 + 12\theta(x-\theta) + 12\theta^2 - 2\beta_1 = 4(x+\theta/2)^2 + 3\theta^2 - 2\beta_1,$$

we obtain

$$\inf_{x>r}\frac{-b(x)}{x-\theta} = \begin{cases} 4(r+\theta/2)^2 + 3\theta^2 - 2\beta_1, & \text{if } r \geqslant -\theta/2 \\ 3\theta^2 - 2\beta_1, & \text{if } r \leqslant -\theta/2, \end{cases} \qquad r \geqslant \theta.$$

$$\inf_{x<r}\frac{-b(x)}{x-\theta} = \begin{cases} 4(r+\theta/2)^2 + 3\theta^2 - 2\beta_1, & \text{if } r \leqslant -\theta/2 \\ 3\theta^2 - 2\beta_1, & \text{if } r \geqslant -\theta/2, \end{cases} \qquad r < \theta.$$

Naturally, one may define $K(r)$ as the right-hand sides, but then the computations for the lower bounds of $\delta_\pm(K)$ become very complicated. Here, we adopt a simplification. Set $r_\theta = r - \theta$. Because

$$\begin{aligned} &4(r+\theta/2)^2 + 3\theta^2 - 2\beta_1 = 12(\theta + r_\theta/2)^2 + r_\theta^2 - 2\beta_1 \geqslant r_\theta^2 - 2\beta_1, \\ &3\theta^2 - 2\beta_1 \geqslant 9\theta^2/4 - 2\beta_1, \end{aligned}$$

when $r \geqslant \theta$ (equivalently, $r_\theta \geqslant 0$), we can choose

$$K(r) = K_\theta(r) = \begin{cases} r_\theta^2 - 2\beta_1, & \text{if } r_\theta > -3\theta/2 \\ 9\theta^2/4 - 2\beta_1, & \text{if } r_\theta < -3\theta/2. \end{cases}$$

By symmetry, one can define $K(r)$ for the case of $r \leqslant \theta$ as follows:

$$K(r) = \begin{cases} r_\theta^2 - 2\beta_1, & \text{if } r_\theta < -3\theta/2 \\ 9\theta^2/4 - 2\beta_1, & \text{if } r_\theta > -3\theta/2. \end{cases}$$

Obviously, K is a continuous piecewise C^1-function.

Suppose that $\theta < 0$ for a moment. We use the notation $G(r)$ defined in the proof (d) of Theorem 4.1. Since $G(r)$ is continuous in r, $G(r)$ is equal to the constant $K(-\theta/2)$ on $(\theta, -\theta/2]$, and $K' > 0$ on (θ, ∞), we have $\sup_{r>\theta} G(r) = \sup_{r\geqslant -\theta/2} G(r)$. Clearly, $\lim_{r\to\infty} K(r)\int_\theta^r (u-\theta)\mathrm{d}u = \infty$ and hence we can ignore (4.7) and handle with (4.8) only. There are two cases.

(a) Let $K(-\theta/2)\int_\theta^{-\theta/2}(u-\theta)\mathrm{d}u < 1$. That is $9\theta^2/4 < \beta_1 + \sqrt{\beta_1^2+2}$. In this case, the solution to (4.8) should satisfy $r_+ - \theta > -3\theta/2$. Solving equation

$$\left(r_\theta^2 - 2\beta_1\right)\int_\theta^r (u-\theta)\mathrm{d}u = 1, \qquad r_\theta > -3\theta/2,$$

we get $(r_+ - \theta)^2 = \beta_1 + \sqrt{\beta_1^2+2}$. Then

$$\begin{aligned} -\frac{1}{2}\int_\theta^{r_+} (x-\theta)^2 K'(x)\mathrm{d}x &= -\frac{1}{2}\int_{-3\theta/2}^{r_+-\theta} x^2 \cdot 2x\mathrm{d}x \\ &= -\frac{1}{4}(r_+ - \theta)^4 + \frac{81}{64}\theta^4 \\ &= -\frac{1}{4}\left(\sqrt{\beta_1^2+2}+\beta_1\right)^2 + \frac{81}{64}\theta^4. \end{aligned}$$

Hence we obtain

$$\begin{aligned} \sup_{r\geqslant -\theta/2} G(r) = G(r_+) &\geqslant \left(\sqrt{\beta_1^2+2}-\beta_1\right)\exp\left[-\frac{1}{4}\left(\sqrt{\beta_1^2+2}+\beta_1\right)^2 + \frac{81}{64}\theta^4\right] \\ &\geqslant \left(\sqrt{\beta_1^2+2}-\beta_1\right)\exp\left[-\frac{1}{4}\left(\sqrt{\beta_1^2+2}+\beta_1\right)^2\right] \\ &= \frac{\sqrt{\beta_1^2+2}-\beta_1}{\sqrt{e}}\exp\left[-\frac{1}{2}\beta_1\left(\beta_1+\sqrt{\beta_1^2+2}\right)\right]. \end{aligned}$$

(b) Let $K(-\theta/2)\int_\theta^{-\theta/2}(u-\theta)\mathrm{d}u \geqslant 1$. Equivalently, $9\theta^2/4 \geqslant \beta_1 + \sqrt{\beta_1^2+2}$. In this case, the solution to (4.8) satisfies $r_+ \in (\theta, -\theta/2)$. Since K is a constant on $(\theta, -\theta/2)$, by (4.13) and (4.12), $G = K$ on $(\theta, -\theta/2]$. Hence

$$\begin{aligned} \sup_{r>\theta} G(r) = G(r_+) = K(-\theta/2) = \frac{9}{4}\theta^2 - 2\beta_1 &\geqslant \sqrt{\beta_1^2+2} - \beta_1 \\ &\geqslant \left(\sqrt{\beta_1^2+2} - \beta_1\right)\exp\left[-\frac{1}{4}\left(\sqrt{\beta_1^2+2}+\beta_1\right)^2\right]. \end{aligned}$$

Combining (a) with (b) and (4.6), we obtain

$$\delta_+(K) \geqslant \frac{\sqrt{\beta_1^2+2}-\beta_1}{\sqrt{e}}\exp\left[-\frac{1}{2}\beta_1\left(\beta_1+\sqrt{\beta_1^2+2}\right)\right].$$

Next, we estimate $\delta_-(K)$. Now, $K(r) = r_\theta^2 - 2\beta_1$ on $(-\infty, \theta)$ since $\theta < 0$. From (4.8), we get the same solution $(r_- - \theta)^2 = \beta_1 + \sqrt{\beta_1^2 + 2}$. But

$$\begin{aligned} -\frac{1}{2}\int_\theta^{r_-} (x-\theta)^2 K'(x)\mathrm{d}x &= -\frac{1}{2}\int_0^{r_- - \theta} x^2 \cdot 2x\mathrm{d}x \\ &= -\frac{1}{4}(r_- - \theta)^4 = -\frac{1}{4}\Big(\sqrt{\beta_1^2+2}+\beta_1\Big)^2. \end{aligned}$$

By (4.6) again, we get

$$\delta_-(K) \geqslant \frac{\sqrt{\beta_1^2+2} - \beta_1}{\sqrt{e}} \exp\bigg[-\frac{1}{2}\beta_1\Big(\beta_1 + \sqrt{\beta_1^2+2}\Big)\bigg].$$

Therefore, we have proved the required lower bound in the case of $\theta < 0$.

By symmetry, the same conclusion holds when $\theta > 0$. The proof for $\theta = 0$ is much simpler as shown below.

When $\beta_2 = 0$, we simply let $\theta = 0$. Then

$$\frac{-b(x)}{x} = 4x^2 - 2\beta_1, \qquad x \neq 0.$$

We choose $K(r) = 4r^2 - 2\beta_1$. Then the equation (4.8) gives us

$$r_\pm^2 = \frac{1}{4}\Big(\beta_1 + \sqrt{\beta_1^2+8}\Big).$$

Because

$$\int_0^r \bigg[\int_0^x u\mathrm{d}u\bigg]\mathrm{d}K(x) = \int_0^r 4x^3\mathrm{d}x = r^4,$$

by (4.6), we obtain the last required assertion. □

We will improve the estimate of Example 4.3 in §5 (Example 5.3) by a different method.

Before moving further, let us make some remarks about the estimate given in Example 4.3. Recall that at the beginning of the proof, in choosing the function $K(r)$, the term $12(\theta + r_\theta/2)^2$ was removed, this simplified greatly the proof since the original quartic equation is reduced to a quadratic one. For this reason one may worry lost too much in the estimation and we want to know the best estimate we can get by part (2) of Theorem 4.1. For this, we use a different trick. Consider the case of $\theta < 0$ only. We use the complete form of K:

$$K(r) = \begin{cases} 4(r+\theta/2)^2 + 3\theta^2 - 2\beta_1, & \text{if } r \geqslant -\theta/2 \\ 3\theta^2 - 2\beta_1, & \text{if } \theta \leqslant r \leqslant -\theta/2 \\ 4(r+\theta/2)^2 + 3\theta^2 - 2\beta_1, & \text{if } r < \theta. \end{cases}$$

(i) Following the proof of Example 4.3, we study first the estimation of $\delta_+(K)$. There are two cases.

(a) Let $K(-\theta/2)\int_\theta^{-\theta/2}(u-\theta)\mathrm{d}u<1$. That is, $3\theta^2<\beta_1+\sqrt{\beta_1^2+8/3}$. The idea is that in looking for a uniform estimate, we may regard r as a parameter rather than θ. In other words, instead of solving equation (4.8)

$$\big(4r_\theta^2+12\theta r_\theta+12\theta^2-2\beta_1\big)\int_\theta^r(u-\theta)\mathrm{d}u=1,\qquad r_\theta>-3\theta/2$$

in r, we solve the equation in θ. Then the equation has two solutions:

$$\theta=\frac{1}{6}\Big(-3r_\theta\pm\sqrt{6\beta_1+6/r_\theta^2-3r_\theta^2}\Big).$$

Since θ is real, r_θ must satisfy

$$r_\theta^2\leqslant\beta_1+\sqrt{\beta_1^2+2}.\tag{4.14}$$

Next, in the "+" case, $\theta<0$ iff

$$r_\theta^2>\Big(\beta_1+\sqrt{\beta_1^2+8}\Big)\Big/4,\tag{4.15}$$

and it is obvious that $r_\theta>-3\theta/2$. In the "−" case, it is automatically that $\theta<0$ and $r_\theta>3\theta/2$ iff

$$r_\theta^2>\Big(3\beta_1+\sqrt{9\beta_1^2+24}\Big)\Big/4,\tag{4.16}$$

To estimate the decay exponent, note that on the one hand, we have

$$\begin{aligned}\theta r_\theta^3&=\frac{1}{6}r_\theta^2\Big(-3r_\theta^2\pm\sqrt{6\beta_1r_\theta^2+6-3r_\theta^4}\Big)\\&=\frac{1}{6}z\Big(-3z\pm\sqrt{6\beta_1z+6-3z^2}\Big),\end{aligned}$$

where $z=r_\theta^2$. On the other hand, we have

$$-\frac{1}{2}\int_{-3\theta/2}^{r_+-\theta}x^2K'(x)\mathrm{d}x=-r_\theta^4-2\theta r_\theta^3-\frac{27}{16}\theta^4.$$

Replacing r_θ^2 with z on the right-hand side plus some computation, we finally get

$$\begin{aligned}-\frac{1}{2}\int_{-3\theta/2}^{r_+-\theta}x^2K'(x)\mathrm{d}x=&-\frac{3}{64}\Big[-2z^2+8\beta_1z+\beta_1^2+8+\frac{2\beta_1}{z}+\frac{1}{z^2}\Big]\\&\pm\frac{\sqrt{3}}{96}\big(9+9\beta_1z-23z^2\big)\sqrt{\frac{2\beta_1}{z}+\frac{2}{z^2}-1}.\end{aligned}$$

To obtain the uniform lower bound, by (4.14) and (4.15), we need to minimize the right-hand side under the constrain

$$\begin{cases}\big(\beta_1+\sqrt{\beta_1^2+8}\big)/4<z\leqslant\beta_1+\sqrt{\beta_1^2+2} & \text{in the "+" case}\\ \big(3\beta_1+\sqrt{9\beta_1^2+24}\big)/4<z\leqslant\beta_1+\sqrt{\beta_1^2+2} & \text{in the "−" case.}\end{cases}$$

A numerical computation shows that the first case is smaller than the second one and its leading term is approximately $-0.8\,\beta_1^2$.

(b) Let $K(-\theta/2)\int_\theta^{-\theta/2}(u-\theta)\mathrm{d}u \geqslant 1$. That is, $3\theta^2 \geqslant \beta_1+\sqrt{\beta_1^2+8/3}$. Then we have the lower bound $\sqrt{\beta_1^2+8/3}-\beta_1$ which is decayed slowly than exponential.

(ii) Next, in the case of $r<\theta$, by assumption, $\theta<0$ and $r_\theta<0$, we have only one solution

$$\theta=\frac{1}{6}\Big(-3r_\theta-\sqrt{6\beta_1+6/r_\theta^2-3r_\theta^2}\,\Big),$$

and furthermore $\theta<0$ iff

$$r_\theta^2<\Big(\beta_1+\sqrt{\beta_1^2+8}\,\Big)\Big/4.$$

To estimate the decay exponent, note that on the one hand, since $r_\theta<0$, we have

$$\begin{aligned}\theta r_\theta^3&=\frac{1}{6}r_\theta^2\Big(-3r_\theta^2+\sqrt{6\beta_1 r_\theta^2+6-3r_\theta^4}\,\Big)\\&=\frac{1}{6}z\Big(-3z+\sqrt{6\beta_1 z+6-3z^2}\,\Big).\end{aligned}$$

On the other hand, we have

$$-\frac{1}{2}\int_0^{r_--\theta}x^2K'(x)\mathrm{d}x=-r_\theta^4-2\theta r_\theta^3.$$

Hence

$$-\frac{1}{2}\int_0^{r_--\theta}x^2K'(x)\mathrm{d}x=-\frac{1}{\sqrt{3}}\,z\sqrt{2\beta_1 z+2-z^2}.$$

To obtain the uniform lower bound, it suffices to minimize the right-hand side under the constrain

$$0<z<\Big(\beta_1+\sqrt{\beta_1^2+8}\,\Big)\Big/4.$$

A numerical computation shows that the resulting estimate is bigger than $-0.8\,\beta_1^2$.

(iii) Finally, we conclude that the estimate on the exponent obtained so far is approximately $-0.8\,\beta_1^2$. Comparing this with our estimate $-\beta_1^2$, it is clear that there is no much room left for an improvement by part (2) of Theorem 4.1.

We now study the general criteria and estimates of $\lambda_1(L)$ and $\lambda_0^{\pm}(\theta)$ (see (4.17) below for definitions) in dimension one. For this, we need more notation.

Fix an arbitrary reference point $\theta\in\mathbb{R}$, not necessarily a root of $b(x)=-u'(x)$. Let $\mathbb{R}_\theta^+=(\theta,\infty)$, $\mathbb{R}_\theta^-=(-\infty,\theta)$, $\overline{\mathbb{R}}_\theta^+=[\theta,\infty)$, and $\overline{\mathbb{R}}_\theta^-=(-\infty,\theta]$. Recall that

$C_\theta(x) = \int_\theta^x b/a$. Define

$$\begin{aligned}
&\varphi_\theta(x) = \int_\theta^x e^{-C}, \qquad \delta_\theta^\pm = \sup_{x\in\mathbb{R}_\theta^\pm} \varphi(x)\int_x^{\pm\infty}\frac{e^C}{a},\\
&\mathscr{F}_{I\theta}^\pm = \Big\{f\in C\big(\overline{\mathbb{R}_\theta^\pm}\big)\cap C^1(\mathbb{R}_\theta^\pm): f(\theta)=0,\ f'\big|_{\mathbb{R}_\theta^\pm}>0\Big\},\\
&\mathscr{F}_{II\theta}^\pm = \Big\{f\in C\big(\overline{\mathbb{R}_\theta^\pm}\big): f(\theta)=0,\ (\pm f)\big|_{\mathbb{R}_\theta^\pm}>0\Big\},\\
&I_\theta^\pm(f)(x) = \frac{e^{-C(x)}}{f'(x)}\int_x^{\pm\infty}\frac{fe^C}{a}, \qquad \pm(x-\theta)\geqslant 0,\ \ f\in\mathscr{F}_{I\theta}^\pm,\\
&II_\theta^\pm(f)(x) = \frac{1}{f(x)}\int_\theta^{\pm\infty}\varphi_\theta(x\wedge\cdot)\,\frac{fe^C}{a} = \frac{1}{f(x)}\int_\theta^x e^{-C(y)}\mathrm{d}y\int_y^{\pm\infty}\frac{fe^C}{a},\\
&\qquad\qquad \pm\,(x-\theta)\geqslant 0,\ \ f\in\mathscr{F}_{II\theta}^\pm,\\
&\lambda_0^\pm(\theta) = \inf\Big\{D(f): f|_{\mathbb{R}\setminus\mathbb{R}_\theta^\pm}=0,\ f\in C\big(\overline{\mathbb{R}_\theta^\pm}\big)\cap C^1(\mathbb{R}_\theta^\pm),\ \mu(f^2)=1\Big\}. \qquad (4.17)
\end{aligned}$$

Theorem 4.4. The comparison of $\lambda_1(L)$ and $\lambda_0^\pm(\theta)$ and their estimates are given as follows.

(1) $\inf_{\theta\in\mathbb{R}}\big[\lambda_0^+(\theta)\vee\lambda_0^-(\theta)\big]\geqslant\lambda_1(L)\geqslant\sup_{\theta\in\mathbb{R}}\big[\lambda_0^+(\theta)\wedge\lambda_0^-(\theta)\big]$. In particular, $\lambda_1(L)=\lambda_0^+(\bar\theta)$, where $\bar\theta$ is the solution to the equation $\lambda_0^+(\theta)=\lambda_0^-(\theta)$, $\theta\in[-\infty,\infty]$.
(2) If m is the medium of μ, then $2\big[\lambda_0^+(m)\wedge\lambda_0^-(m)\big]\geqslant\lambda_1(L)\geqslant\lambda_0^+(m)\wedge\lambda_0^-(m)$.
(3) $\lambda_0^\pm(\theta)\geqslant\sup_{f\in\mathscr{F}_{II\theta}^\pm}\inf_{x\in\mathbb{R}_\theta^\pm} II_\theta^\pm(f)(x)^{-1}\geqslant\sup_{f\in\mathscr{F}_{I\theta}^\pm}\inf_{x\in\mathbb{R}_\theta^\pm} I_\theta^\pm(f)(x)^{-1}$. Moreover, the sign of equalities hold whenever both a and b are continuous.
(4) $\big(\delta_\theta^\pm\big)^{-1}\geqslant\lambda_0^\pm(\theta)\geqslant\big(4\delta_\theta^\pm\big)^{-1}$.

Proof. The first assertion of part (1) is just [18; Theorem 3.3]. The lower bound in part (2) follows from the one of part (1). As remarked above [18; Theorem 3.3], from the proof of [18; Theorem 3.1], it follows that

$$\lambda_1(L)\leqslant\inf_{\theta\in\mathbb{R}}\big[\lambda_0^+(\theta)\mu(\theta,\infty)\big]\wedge\big[\lambda_0^-(\theta)\mu(-\infty,\theta)\big].$$

Hence, the upper bound in part (2) follows immediately. The variational formulas for the lower bounds given in part (3) is a copy of [19; Theorem 1.1]. In which, the corresponding variational formulas for the upper bounds are also presented, but omitted here. Part (4) was proven in [18; Theorem 1.1]. From these quoted papers, one can find some more sharper estimates and further references.

It remains to prove the second assertion of part (1). For this, it suffices to show that $\lambda_0^\pm(\theta)$ is continuous in θ. By symmetry, it is enough to prove that $\lambda_0^+(\theta)$ is continuous in θ. Let $\theta_1<\theta_2<\infty$. Clearly, $\lambda_0^+(\theta_1)<\lambda_0^+(\theta_2)$. Given $\varepsilon\in(0,1)$, choose $f=f_\varepsilon\in C^1(\theta_1,\infty)\cap C[\theta_1,\infty)$ such that $f(\theta_1)=0$, $\int_{\theta_1}^\infty f^2\mathrm{d}\mu=1$ and $A-\varepsilon\leqslant\lambda_0^+(\theta_1)$, where $A=A_\varepsilon=\int_{\theta_1}^\infty f'^2\mathrm{d}\mu$. By the continuity of f, when

$\theta_2 - \theta_1 > 0$ is sufficient small, we have

$$\begin{aligned}\left| f(\theta_2)^2 \int_{\theta_2}^{\infty} \mathrm{d}\mu - 2f(\theta_2) \int_{\theta_2}^{\infty} f \mathrm{d}\mu - \int_{\theta_1}^{\theta_2} f^2 \mathrm{d}\mu \right| \\ &\leqslant f(\theta_2)^2 \int_{\theta_1}^{\infty} \mathrm{d}\mu + 2f(\theta_2) + \int_{\theta_1}^{\theta_2} f^2 \mathrm{d}\mu \\ &< \varepsilon.\end{aligned}$$

Then $\int_{\theta_2}^{\infty} [f - f(\theta_2)]^2 \mathrm{d}\mu > 1 - \varepsilon$ and furthermore

$$\lambda_0^+(\theta_2) \leqslant \int_{\theta_2}^{\infty} f'^2 \mathrm{d}\mu \Big/ \int_{\theta_2}^{\infty} [f - f(\theta_2)]^2 \mathrm{d}\mu \leqslant \frac{A}{1-\varepsilon} \leqslant \frac{\lambda_0^+(\theta_1) + \varepsilon}{1 - \varepsilon}.$$

Since ε can be arbitrarily small, we obtain the required assertion. □

As an illustration of the applications of Theorem 4.4, we discuss Examples 4.2 and 4.3 again.

Example 4.5. Everything in premise is the same as in Example 4.2. We have

$$\frac{2\alpha}{\delta} \geqslant \lambda_1(L_{\alpha,\beta}) \geqslant \frac{\alpha}{4\delta},$$

where

$$\delta = \sup_{x>0} \int_0^x e^{y^2} \mathrm{d}y \int_x^{\infty} e^{-y^2} \mathrm{d}y \approx 0.239405.$$

Proof. First, we have the root $\theta = -\beta/(2\alpha)$ of $u'(x)$, it is also the medium of the measure. Next,

$$C_\theta(x) = -\alpha(x-\theta)^2, \quad \varphi_\theta(x) = \int_0^{x-\theta} e^{\alpha y^2} \mathrm{d}y, \quad \int_x^{\infty} e^{-\alpha(y-\theta)^2} \mathrm{d}y = \int_{x-\theta}^{\infty} e^{-\alpha y^2} \mathrm{d}y.$$

Hence $\delta_\theta^+ = \delta/\alpha$. By symmetry, we also have $\delta_\theta^- = \delta/\alpha$. The assertion now follows from parts (2) and (4) of Theorem 4.4. □

Example 4.6. Everything in premise is the same as in Example 4.3. We have

(1) $\lim_{|\beta_2| \to \infty} \lambda_1(L_{\beta_1, \beta_2}) = \infty$.
(2) For $\beta_1 \geqslant 0$, we have

$$\lambda_1(L_{\beta_1, 0}) \leqslant 4e^{14} \exp\left[-\frac{1}{4}\beta_1^2 + 2\log(1+\beta_1) \right].$$

Proof. By symmetry of $u(x)$ in x, one may assume that $\beta_2 \geqslant 0$. Let θ be a real root of $u'(x)$. Clearly, $\lim_{\beta_2\to\infty}\theta = -\infty$. Moreover, $u(x)-u(\theta) = (x-\theta)^2\big[(x-\theta)^2 + 4\theta(x-\theta) + 6\theta^2 - \beta_1\big]$. Hence

$$\begin{aligned}
&\int_\theta^x e^{u(y)}\mathrm{d}y\int_x^\infty e^{-u(z)}\mathrm{d}z\\
&\quad=\int_\theta^x e^{u(y)-u(\theta)}\mathrm{d}y\int_x^\infty e^{-u(z)+u(\theta)}\mathrm{d}z\\
&\quad=\int_0^{x-\theta} e^{y^2(y^2+4\theta y+6\theta^2-\beta_1)}\mathrm{d}y\int_{x-\theta}^\infty e^{-z^2(z^2+4\theta z+6\theta^2-\beta_1)}\mathrm{d}z\\
&\quad=\int_0^{x-\theta}\mathrm{d}y\int_{x-\theta}^\infty \exp\bigg[-(z^2-y^2)\bigg(z^2+y^2+4\theta\bigg(z+\frac{y^2}{z+y}\bigg)+6\theta^2-\beta_1\bigg)\bigg]\mathrm{d}z.
\end{aligned}\tag{4.18}$$

(a) We now prove the first assertion. It says that the parameter β_2 plays a role for $\lambda_1(L_{\beta_1,\beta_2})$, in contrast with Example 4.5. For $x \geqslant \theta$, by (4.18), we have

$$\begin{aligned}
&\int_\theta^x e^{u(y)}\mathrm{d}y\int_x^\infty e^{-u(z)}\mathrm{d}z\\
&\quad=\int_0^{x-\theta}\mathrm{d}y\int_{x-\theta}^\infty \mathrm{d}z\exp\bigg[-(z^2-y^2)\bigg[(z+2\theta)^2+\bigg(y+\frac{2\theta y}{z+y}\bigg)^2\\
&\qquad\qquad\qquad -4\theta^2\bigg(\frac{y}{z+y}\bigg)^2+2\theta^2-\beta_1\bigg]\bigg]\\
&\quad\leqslant\int_0^{x-\theta}\mathrm{d}y\int_{x-\theta}^\infty \mathrm{d}z\exp\bigg[-(z^2-y^2)\bigg[-4\theta^2\bigg(\frac{y}{z+y}\bigg)^2+2\theta^2-\beta_1\bigg]\bigg].
\end{aligned}$$

Since $z \geqslant y \geqslant 0$, we have $y/(z+y) \leqslant 1/2$. The right-hand side is controlled by

$$\int_0^{x-\theta}\mathrm{d}y\int_{x-\theta}^\infty e^{-(z^2-y^2)(\theta^2-\beta_1)}\mathrm{d}z,\qquad x\geqslant\theta.\tag{4.19}$$

We now use Conte's estimate (cf. [20]):

$$x\bigg(1+\frac{x^2}{12}\bigg)e^{-3x^2/4} < e^{-x^2}\int_0^x e^{y^2} \leqslant \frac{\pi^2}{8x}(1-e^{-x^2}),\qquad x>0$$

and Gautschi's estimate (cf. [21]):

$$\begin{gathered}
\frac12\Big[(x^p+2)^{1/p}-x\Big] < e^{x^p}\int_x^\infty e^{-y^p}dy \leqslant C_p\bigg[\bigg(x^p+\frac{1}{C_p}\bigg)^{1/p}-x\bigg],\qquad x\geqslant 0,\\
C_p := \Gamma\big(1+1/p\big)^{p/(p-1)},\quad p>1;\qquad C_2=\pi/4.
\end{gathered}$$

Thus,

$$\begin{aligned}
\int_0^x e^{cy^2}\mathrm{d}y\int_x^\infty e^{-cz^2}\mathrm{d}z &\leqslant \frac{\pi^2}{8c\sqrt{c}\,x}\big(1-e^{-cx^2}\big)\cdot\frac{\pi}{4}\bigg(\sqrt{cx^2+\frac{4}{\pi}}-\sqrt{c}\,x\bigg)\\
&\leqslant \frac{\pi^2}{8c\sqrt{c}\,x}\sqrt{\frac{\pi}{4}}\,\big(1-e^{-cx^2}\big),\qquad x\geqslant 0.
\end{aligned}$$

Noting that $\big(1-e^{-cx^2}\big)/x \leqslant cx \leqslant c$ for all $x\in(0,1]$ and $\big(1-e^{-cx^2}\big)/x \leqslant 1/x \leqslant 1$ for all $x \geqslant 1$, we obtain

$$\int_0^x e^{cy^2}\mathrm{d}y\int_x^\infty e^{-cz^2}\mathrm{d}z \leqslant \frac{\pi^{5/2}}{16\sqrt{c}}, \qquad x\geqslant 0,\ c\geqslant 1.$$

Therefore

$$\begin{aligned}\delta_\theta^+ &= \sup_{x>\theta}\int_\theta^x e^{u(y)}\mathrm{d}y\int_x^\infty e^{-u(z)}\mathrm{d}z\\ &\leqslant \sup_{x>0}\int_0^x \mathrm{d}y\int_x^\infty e^{-(z^2-y^2)(\theta^2-\beta_1)}\mathrm{d}z\\ &\leqslant \frac{\pi^{5/2}}{16\sqrt{\theta^2-\beta_1}}\to 0 \qquad \text{as}\ \ \theta\to-\infty.\end{aligned}$$

For δ_θ^-, the proof is similar. As an analogue of (4.18), we have

$$\begin{aligned}&\int_x^\theta e^{u(y)}\mathrm{d}y\int_{-\infty}^x e^{-u(z)}\mathrm{d}z\\ &=\int_{x-\theta}^0 \mathrm{d}y\int_{-\infty}^{x-\theta}\mathrm{d}z\exp\bigg[-(z^2-y^2)\bigg[(z+2\theta)^2+\Big(y+\frac{2\theta y}{z+y}\Big)^2\\ &\qquad\qquad -4\theta^2\Big(\frac{y}{z+y}\Big)^2+2\theta^2-\beta_1\bigg]\bigg].\end{aligned}$$

Since $z\leqslant y\leqslant 0$, we have $|y/(z+y)|\leqslant 1/2$, we obtain

$$\int_x^\theta e^{u(y)}\mathrm{d}y\int_{-\infty}^x e^{-u(z)}\mathrm{d}z\leqslant\int_{x-\theta}^0\mathrm{d}y\int_{-\infty}^{x-\theta}e^{-(z^2-y^2)(\theta^2-\beta_1)}\mathrm{d}z, \qquad x\leqslant\theta.$$

We have thus returned to (4.19).

Now, the first assertion follows from parts (1) and (4) of Theorem 4.4.

(b) For the upper bound in part (2), since $\beta_2=0$, we have $\theta=0$. We need to show that

$$\sup_{x>0}\int_0^x e^{y^4-\beta_1y^2}\mathrm{d}y\int_x^\infty e^{-z^4+\beta_1z^2}\mathrm{d}z\geqslant\frac{1}{4e^{14}}\exp\Big[\frac14\beta_1^2-2\log(1+\beta_1)\Big].$$

Since

$$\begin{aligned}&\int_0^x e^{y^4-\beta_1y^2}\mathrm{d}y\int_x^\infty e^{-z^4+\beta_1z^2}\mathrm{d}z\\ &\quad=\frac14\int_{-\beta_1/2}^{x^2-\beta_1/2}\frac{e^{y^2}}{\sqrt{y+\beta_1/2}}\mathrm{d}y\int_{x^2-\beta_1/2}^\infty\frac{e^{-z^2}}{\sqrt{z+\beta_1/2}}\mathrm{d}z\\ &\quad>\frac14\int_{-\beta_1/2}^{x^2-\beta_1/2}\frac{e^{y^2}}{\sqrt{y+\beta_1/2}}\mathrm{d}y\int_{x^2-\beta_1/2}^{\beta_1/2}\frac{e^{-z^2}}{\sqrt{z+\beta_1/2}}\mathrm{d}z,\end{aligned}$$

when $\beta_1 \geqslant 1$, we have

$$\begin{aligned}
&\int_{-\beta_1/2}^{1-\beta_1/2} \frac{e^{y^2}}{\sqrt{y+\beta_1/2}} \mathrm{d}y \int_{1-\beta_1/2}^{\beta_1/2} \frac{e^{-z^2}}{\sqrt{z+\beta_1/2}} \mathrm{d}z \\
&\qquad \geqslant \frac{1}{\beta_1} \int_{-\beta_1/2}^{1-\beta_1/2} e^{y^2} \mathrm{d}y \int_{1-\beta_1/2}^{\beta_1/2} e^{-z^2} \mathrm{d}z.
\end{aligned}$$

It suffices to show that

$$\frac{1}{\beta_1} \int_{-\beta_1/2}^{1-\beta_1/2} e^{y^2} \mathrm{d}y \int_{1-\beta_1/2}^{\beta_1/2} e^{-z^2} \mathrm{d}z \geqslant \frac{1}{e^{14}} \exp\left[\frac{1}{4}\beta_1^2 - 2\log(1+\beta_1)\right],$$

or

$$\int_{-\beta_1/2}^{1-\beta_1/2} e^{y^2} \mathrm{d}y \int_{1-\beta_1/2}^{\beta_1/2} e^{-z^2} \mathrm{d}z \geqslant \exp\left[\frac{1}{4}\beta_1^2 - \log(1+\beta_1) - 14\right].$$

Since

$$\begin{aligned}
&\int_{1-\beta_1/2}^{\beta_1/2} e^{-z^2} \mathrm{d}z \to \int_{-\infty}^{\infty} e^{-z^2} \mathrm{d}z < \infty, \\
&\int_{-\beta_1/2}^{1-\beta_1/2} e^{y^2} \mathrm{d}y = \int_{\beta_1/2-1}^{\beta_1/2} e^{y^2} \mathrm{d}y \geqslant \exp\left[\left(\frac{\beta_1}{2} - 1\right)^2\right] \to \infty, \\
&\frac{\int_{-\beta_1/2}^{1-\beta_1/2} e^{y^2} \mathrm{d}y}{\exp[\beta_1^2/4 - \log\beta_1]} \sim \frac{\exp[\beta_1^2/4] - \exp[(1-\beta_1/2)^2]}{\exp[\beta_1^2/4]} \\
&\qquad \sim 1 - e^{1-\beta_1} \\
&\qquad \sim 1 \quad \text{as } \beta_1 \to \infty,
\end{aligned}$$

it is easy to check first that

$$\log\left[\int_{-\beta_1/2}^{1-\beta_1/2} e^{y^2} \mathrm{d}y \int_{1-\beta_1/2}^{\beta_1/2} e^{-z^2} \mathrm{d}z\right] \geqslant \frac{1}{4}\beta_1^2 - \log(1+\beta_1) - 14$$

for $\beta_1 \geqslant 1$ and then the required assertion for $\beta_1 \geqslant 0$ by using mathematical softwares. $\square$

Before moving further, let us study the lower bounds of $\inf_{\beta_2 \geqslant 0} \lambda_1(L_{\beta_1,\beta_2})$ in terms of $\delta_\theta^\pm$. For this, we return to (4.18). Because

$$4\theta\left(z + \frac{y^2}{z+y}\right) + 6\theta^2 = 6\left[\theta + \frac{1}{3}\left(z + \frac{y^2}{z+y}\right)\right]^2 - \frac{2}{3}\left(z + \frac{y^2}{z+y}\right)^2 \geqslant -\frac{2}{3}\big(z + y/2\big)^2,$$

and so

$$\begin{aligned}
z^2 + y^2 + 4\theta\left(z + \frac{y^2}{z+y}\right) + 6\theta^2 - \beta_1 &\geqslant z^2 + y^2 - \frac{2}{3}\big(z+y/2\big)^2 - \beta_1 \\
&= \frac{1}{6}(2z^2 - 4zy + 5y^2) - \beta_1 \\
&\geqslant \frac{1}{6}(z^2 + y^2) - \beta_1,
\end{aligned}$$

it follows that

$$\begin{aligned}
\int_0^x \mathrm{d}y \int_x^\infty \exp\bigg[-(z^2-y^2)\bigg(z^2+y^2+4\theta\bigg(z+\frac{y^2}{z+y}\bigg)+6\theta^2-\beta_1\bigg)\bigg]\mathrm{d}z \\
&\leqslant \int_0^x \mathrm{d}y \int_x^\infty \mathrm{d}z\, e^{-(z^2-y^2)((z^2+y^2)/6-\beta_1)} \\
&= \int_0^x \mathrm{d}y\, e^{y^4/6-\beta_1 y^2}\int_x^\infty e^{-z^4/6+\beta_1 z^2}\mathrm{d}z. \qquad (4.20)
\end{aligned}$$

Combining (4.18) with (4.20), we obtain

$$\delta_\theta^+ \leqslant \sup_{x>0}\int_0^x \mathrm{d}y\, e^{y^4/6-\beta_1 y^2}\int_x^\infty e^{-z^4/6+\beta_1 z^2}\mathrm{d}z.$$

The same upper bound holds for δ_θ^-. By parts (1) and (4) of Theorem 4.4, we obtain a lower estimate of $\inf_{\beta_2\geqslant 0}\lambda_1(L_{\beta_1,\,\beta_2})$. However, the resulting bound is smaller than those given in Example 4.3.

We mention that the lower bound given in Example 4.3 may still be improved by applying part (3) of Theorem 4.3 to the test functions $f_\pm$ constructed in the proof of Theorem 4.1. This observation is due to [22]. The proof is quite easy. Let for instance

$$-\sup_{x\in(\theta,\infty)}\frac{af_+''+bf_+'}{f_+}(x)\geqslant\delta>0.$$

Then $f_+\leqslant -(af_+''+bf_+)/\delta$. Noting that $\big(e^C f_+'\big)'=e^C(af_+''+bf_+)/a$, we obtain

$$\begin{aligned}
I_\theta^+(f_+)(x) &= \frac{e^{-C(x)}}{f_+'(x)}\int_x^\infty \frac{f_+e^C}{a} \\
&\leqslant \frac{1}{\delta}\frac{e^{-C(x)}}{f_+'(x)}\int_x^\infty\bigg(-\frac{af_+''+bf_+}{a}\bigg)e^C \\
&= \frac{1}{\delta}\frac{e^{-C(x)}}{f_+'(x)}\int_x^\infty\big(-e^C f_+'\big)' \\
&\leqslant \frac{1}{\delta}\frac{e^{-C(x)}}{f_+'(x)}e^{C(x)}f_+'(x) \\
&= \frac{1}{\delta}, \qquad x>\theta.
\end{aligned}$$

Alternatively, one may apply the approximation procedure given in [19] to improve the lower bound. However, all the computations are quite complicated, and so we do not want to go further along this line.

We remark that the process in Example 4.6 (Example 4.3) possesses much stronger ergodic properties.

Proposition 4.7. The processes corresponding to Example 4.3 is not only exponentially ergodic but also strongly ergodic. It has the empty essential spectrum. It satisfies the logarithmic Sobolev inequality but not the Nash (Sobolev) inequality.

Proof. One may use the criteria given in [13; §5.4] to justify these assertions. For the reader's convenience, here we mention three criteria as follows. By the symmetry, we need only to write down the conditions on the half-line $[0,\infty)$.

Logarithmic Sobolev inequality:

$$\sup_{x>0}\bigg(\int_x^\infty e^{-u}\bigg)\bigg(\log\int_x^\infty e^{-u}\bigg)\int_0^x e^u<\infty.$$

Strong ergodicity:

$$\int_0^\infty \mathrm{d}x e^{u(x)}\int_x^\infty e^{-u}<\infty.$$

Nash (*Sobolev*) *inequality*:

$$\sup_{x>0}\bigg(\int_x^\infty e^{-u}\bigg)^{1-2/\nu}\int_0^x e^u<\infty,\qquad \nu>2.$$

The second condition holds since

$$\frac{\int_x^\infty e^{-u}}{x^{-2}e^{-u}}\sim\frac{1}{2x^{-3}+x^{-2}u'}\sim\frac{x^3}{xu'}\to 0,\qquad x\to\infty.$$

However, replacing x^{-2} with x^{-1} at the beginning, the same proof shows that the standard Ornstein-Uhlenbeck process is not strongly ergodic. For the third condition, note that $\int_x^\infty e^{-u}$ and $\int_0^x e^u$ have the leading order e^{-u} and e^u respectively. Hence the leading order of

$$\bigg(\int_x^\infty e^{-u}\bigg)^{1-2/\nu}\int_0^x e^u$$

is $e^{2u/\nu}\to\infty$ as $x\to\infty$. Similarly, one can check the first condition. Alternatively, to see that the logarithmic Sobolev inequality holds, simply use the fact that $\lim_{|x|\to\infty}u''(x)>0$ (see [23]). We will come back to this point in Example 5.3. Finally, the logarithmic Sobolev inequality implies the essential spectrum to be empty. $\square$

Finally, we study a perturbation of $\lambda_1(L)$.

Proposition 4.8. Let $a(x)\equiv 1$ and assume that $\delta_\theta^\pm<\infty$ for some $\theta\in\mathbb{R}$. Next, let h satisfy $\int_{\mathbb{R}}e^{C+h}<\infty$. Define $\delta_\theta^\pm(h)=\sup_{x\in\mathbb{R}_\theta^\pm}\int_\theta^x e^{-C-h}\int_x^{\pm\infty}e^{C+h}$. If there exist constants $K_1^\pm,\dots,K_4^\pm$ such that

$$\pm\int_x^{\pm\infty}e^C\leqslant K_1^\pm e^{C(x)},\qquad \pm(x-\theta)\geqslant 0,\tag{4.21}$$

$$\pm\int_\theta^x e^{-C}\leqslant K_2^\pm e^{-C(x)},\qquad \pm(x-\theta)\geqslant 0,\tag{4.22}$$

$$\pm\int_x^{\pm\infty}e^C\big|e^h-1\big|\leqslant K_3^\pm e^{C(x)},\qquad \pm(x-\theta)\geqslant 0,\tag{4.23}$$

$$\pm\int_\theta^x e^{-C}\big|e^{-h}-1\big|\leqslant K_4^\pm e^{-C(x)},\qquad \pm(x-\theta)\geqslant 0,\tag{4.24}$$

then
$$\delta_\theta^{\pm}(h) \leqslant \delta_\theta^{\pm} + K_2^{\pm}K_3^{\pm} + K_1^{\pm}K_4^{\pm} + K_4^{\pm}K_3^{\pm} < \infty.$$

Proof. Here, we consider $\delta_\theta^{+}(h)$ only. As in [5], we have

$$\begin{aligned}
&\int_\theta^x e^{-C-h}\int_x^\infty e^{C+h}\\
&= \bigg[\int_\theta^x e^{-C} + \int_\theta^x e^{-C}\big(e^{-h}-1\big)\bigg]\cdot\bigg[\int_x^\infty e^{C} + \int_x^\infty e^{C}\big(e^{h}-1\big)\bigg]\\
&= \int_\theta^x e^{-C}\int_x^\infty e^{C} + \int_\theta^x e^{-C}\int_x^\infty e^{C}\big(e^{h}-1\big)\\
&\quad + \int_\theta^x e^{-C}\big(e^{-h}-1\big)\int_x^\infty e^{C} + \int_\theta^x e^{-C}\big(e^{-h}-1\big)\int_x^\infty e^{C}\big(e^{h}-1\big)\\
&\leqslant \delta_\theta^{+} + K_2^{+}K_3^{+} + K_1^{+}K_4^{+} + K_4^{+}K_3^{+} < \infty. \quad \square
\end{aligned}$$

The above result is a revised version of [5; Theorem 3.4], where instead of (4.23) and (4.24), the conditions

(i) $C(x)$ is strictly uniformly convex up to a bounded function
(ii) $\int_{\mathbb{R}}\big(e^{|h|}-1\big) < \infty$

are employed. It is easy to check that these conditions together are stronger than (4.23) and (4.24). Clearly, under (4.21) and (4.22), conditions (4.23) and (4.24) are automatic for bounded h, for which, the condition (ii) here may fail.

Example 4.9. Let $a(x) \equiv 1$ and $C_\beta(x) = -x^4 + \beta x^2$. Then $\lambda_1(L_\beta) > 0$ for all $\beta \in \mathbb{R}$.

Proof. The case of $\beta < 0$ is easy since $-C_\beta$ is convex. Hence we assume that $\beta \geqslant 0$. Then $-C_\beta$ is convex for large enough x and so the conclusion is known. Here we check it by using Proposition 4.8. Take $C(x) = -x^4$ and regard $h(x) = \beta x^2$ as a perturbation of $C(x)$. Clearly, $\int_{\mathbb{R}}\big(e^{|h|}-1\big) = \infty$. Set $\theta = 0$.

First, by Gautschi's estimate, we have

$$e^{-C(x)}\int_x^\infty e^{C} = e^{x^4}\int_x^\infty e^{-y^4}\mathrm{d}y \leqslant C_4\bigg[\bigg(x^4 + \frac{1}{C_4}\bigg)^{1/4} - x\bigg] \leqslant \Gamma\bigg(\frac{5}{4}\bigg) \approx 0.9064$$

for all $x > 0$. Next, we have

$$\begin{aligned}
e^{C(x)}\int_x^\infty e^{-C}\big|e^{-h}-1\big| &= e^{-x^4}\int_x^\infty e^{y^4}\big|e^{-\beta y^2}-1\big|\mathrm{d}y\\
&\leqslant e^{C(x)}\int_x^\infty e^{-C}\\
&= e^{-x^4}\int_x^\infty e^{y^4}\mathrm{d}y\\
&< 0.6, \qquad x > 0.
\end{aligned}$$

Moreover,

$$e^{-C(x)}\int_x^\infty e^C\big|e^{-h}-1\big| < e^{-C(x)}\int_x^\infty e^{C+h} \leqslant e^{\beta(0.7\,\beta+26)}, \qquad x>0.$$

By symmetry, the same estimates hold on $(-\infty, 0]$. Now, by Proposition 4.8 and Theorem 4.4, it follows that the leading order of the lower estimate of $\lambda_1(L_\beta)$ is $\exp[-0.7\,\beta^2]$ which is not far away from the optimal one: $\exp[-\beta^2/4]$. $\square$

5. Logarithmic Sobolev inequality.

We begin this section with a result taken from [23; Corollary 1.4].

Lemma 5.1. Let $L=\Delta-\langle\nabla U,\nabla\rangle$ in $\mathbb{R}^n$ and define $\gamma(r)=\inf_{|x|\geqslant r}\lambda_{\min}(\mathrm{Hess}(U)(x))$. If $\sup_{r\geqslant 0}\gamma(r)>0$, then we have

$$\sigma(L)\geqslant\frac{2e}{a_0^2}\exp\left[-\int_0^{a_0} r\gamma(r)\mathrm{d}r\right]>0,$$

where $a_0>0$ is the unique solution to the equation $\int_0^a\gamma(r)\mathrm{d}r=2/a$.

This lemma says that the logarithmic Sobolev constant is positive whenever so is $\lambda_{\min}(\mathrm{Hess}(U)(x))$ at infinity. Unfortunately, as shown by Example 2.5, our models do not satisfy this condition even in the two-dimensional case. Hence, we justify the power of the estimate provided by the lemma only in dimensional one (compare with the criterion for the inequality, see for instance [13; Theorem 7.4]).

Example 5.2. For Example 4.2, we have $\lambda_1(L_{\alpha,\beta})\geqslant\sigma(L_{\alpha,\beta})\geqslant 2\alpha$ which are exact.

Proof. Because $u(x)=\alpha x^2+\beta x$, we have $u''(x)=2\alpha$ and so

$$\gamma(r)=\inf_{|x|\geqslant r}u''(x)=2\alpha.$$

Next, since $\int_0^a\gamma(r)\mathrm{d}r=2\alpha\,a$. The unique solution to the equation

$$\int_0^a\gamma(r)\mathrm{d}r=\frac{2}{a}$$

is $a_0^2=1/\alpha$. Noticing that $\int_0^a r\gamma(r)\mathrm{d}r=\alpha\,a^2$, by Lemma 5.1, we obtain

$$\sigma(L_{\alpha,\beta})\geqslant\frac{2e}{a_0^2}\exp\big[-\alpha\,a_0^2\big]=2\alpha.$$

This is clearly exact since the well-known fact $\lambda_1(L_{\alpha,\beta})\geqslant\sigma(L_{\alpha,\beta})$ (cf. [13; Theorem 8.7]) and Example 4.2. $\square$

Example 5.3. For Example 4.3, we have

$$\inf_{\beta_2} \lambda_1(L_{\beta_1,\beta_2}) \geqslant \inf_{\beta_2} \sigma(L_{\beta_1,\beta_2}) \geqslant \frac{\sqrt{\beta_1^2+8}-\beta_1}{\sqrt{e}} \exp\left[-\frac{1}{8}\beta_1\Big(\beta_1+\sqrt{\beta_1^2+8}\Big)\right]$$
$$\geqslant \begin{cases} -2\beta_1 + \dfrac{2}{\sqrt{e/2-\beta_1}}, & \text{if } \beta_1 < 0 \\ 2\sqrt{2/e}, & \text{if } \beta_1 = 0 \\ \dfrac{1}{\sqrt{e/8}+\beta_1} \exp\big[-\beta_1^2/4\big], & \text{if } \beta_1 > 0 \end{cases}$$

Proof. Because $u(x) = x^4 - \beta_1 x^2 + \beta_2 x$, we have $u''(x) = 12x^2 - 2\beta_1$ and $\gamma(r) = \inf_{|x|\geqslant r} u''(x) = 12r^2 - 2\beta_1$. Next, since $\int_0^a \gamma(r)\mathrm{d}r = 4a^3 - 2\beta_1 a$, the solution to the equation $\int_0^a \gamma(r)\mathrm{d}r = 2/a$ is as follows

$$a_0^2 = \frac{\beta_1 + \sqrt{\beta_1^2+8}}{4}.$$

Next, since

$$\int_0^a r\gamma(r)\mathrm{d}r = a^2(3a^2-\beta_1),$$

by Lemma 5.1, we obtain

$$\begin{aligned}\sigma(L_{\beta_1,\beta_2}) &\geqslant \frac{2e}{a_0^2}\exp\big[-a_0^2(3a_0^2-\beta_1)\big]\\ &= \frac{\sqrt{\beta_1^2+8}-\beta_1}{\sqrt{e}}\exp\left[-\frac{1}{8}\beta_1\Big(\beta_1+\sqrt{\beta_1^2+8}\Big)\right]. \quad \square\end{aligned}$$

Note that in the case of $\beta_1 < 0$, the Bakry-Emery criterion (cf. (2.9)) is available and gives us the lower bound $-2\beta_1$ which is smaller than the estimate above. Example 5.3 is somehow unexpect since it improves Example 4.3 (In the special case that $\beta_2 = 0$, they are coincided). The reason is due to the fact that only the uniform estimate is treated in Example 4.3 and the linear term of U is ruled out in Lemma 5.1 (but the universal estimates depend on the linear term, cf. [13; Theorem 7.4]). Otherwise, the two methods may not be comparable in view of part (1) of Example 4.6. As mentioned in [23; Example 1.12] that the bounded perturbations should be carefully treated before applying Lemma 5.1.

Proof of Proposition 1.4. Let $\beta_1 \geqslant 0$. Note that

$$\sqrt{1+\frac{8}{\beta_1^2}} \leqslant 1 + \frac{4}{\beta_1^2}.$$

We have $\sqrt{\beta_1^2+8} \leqslant \beta_1 + 4/\beta_1$. Hence

$$\exp\left[-\frac{1}{8}\beta_1\Big(\beta_1+\sqrt{\beta_1^2+8}\Big)\right] \geqslant \frac{1}{\sqrt{e}}\exp\left[-\frac{1}{4}\beta_1^2\right].$$

Similarly, we have

$$\sqrt{\beta_1^2+8}-\beta_1=\frac{8}{\sqrt{\beta_1^2+8}+\beta_1}\geqslant\frac{4}{\beta_1+2\beta_1^{-1}}.$$

By Example 5.3, we obtain $\inf_{\beta_2}\sigma(L_{\beta_1,\beta_2})\geqslant\exp[-\beta_1^2/4-\log(1+\beta_1)]$ for $\beta_1\geqslant 2$. Combining this with Example 4.6, we get the required assertion. $\square$

6. Continuous spin systems.

We begin this section with the ergodicity of our models in the finite dimensions. Consider the particle system on Λ with periodic boundary. Then the generator is

$$L_\Lambda=\Delta+\langle b,\,\nabla\rangle$$

where

$$b_i(x)=-u'(x_i)-2J\sum_{j\in N(i)}(x_i-x_j)$$

for some $u\in C^\infty(\mathbb{R})$, constant J, and $N(i)$ is the nearest neighbors of i. For simplicity, assume that $J\geqslant 0$, but it is not essential in this section. Recall that for the coupling by reflection, the coupling operator $\overline{L}$ has the coefficients

$$a(x,y)=\begin{pmatrix} I & I-2\bar{u}\bar{u}^* \\ I-2\bar{u}\bar{u}^* & I\end{pmatrix},\qquad b(x,y)=\begin{pmatrix} b(x)\\ b(y)\end{pmatrix},$$

where $\bar{u}=\bar{u}(x,y)=(x-y)/|x-y|$. Furthermore, for $f\in C[0,\infty)\cap C^2(0,\infty)$, we have

$$\overline{L}f(|x-y|)=4f''(|x-y|)+\frac{\langle x-y,\,b(x)-b(y)\rangle}{|x-y|}f'(|x-y|),\qquad x\neq y$$

(cf. [13; Theorem 2.30]). To illustrate the idea, we restrict ourselves to the second model.

Theorem 6.1. Let $u(x_i)=x_i^4-\beta x_i^2$ for all $i\in\Lambda$. Then the process is exponentially ergodic for any finite Λ. Moreover, the coupling by reflection (X_t,Y_t) gives us

$$\overline{\mathbb{E}}^{x,y}f(|X_t-Y_t|)\leqslant f(|x-y|)e^{-\varepsilon t},\qquad t\geqslant 0,$$

where

$$f(r)=\int_0^r e^{-C(s)}\mathrm{d}s\int_s^\infty e^C\sqrt{\varphi},\qquad r>0,$$

$$C(r)=-\frac{1}{16|\Lambda|}r^4+\frac{\beta}{4}r^2,\qquad \varphi(r)=\int_0^r e^{-C},$$

$$\varepsilon=\varepsilon(\Lambda,\beta)=4\inf_{r>0}\frac{\sqrt{\varphi(r)}}{f(r)}>0.$$

Proof. Because $u'(x_i) = 4x_i^3 - 2\beta x_i$ and

$$b_i(x) = -4x_i^3 + 2\beta x_i - 2J \sum_{j\in N(i)} (x_i - x_j)$$

for all i. Thus,

$$b_i(x) - b_i(y) = -4\big(x_i^3 - y_i^3\big) + 2\beta(x_i - y_i) - 2J \sum_{j\in N(i)} (x_i - y_i - x_j + y_j).$$

Hence

$$\begin{aligned}\langle x - y,\, b(x) - b(y)\rangle &= -4\sum_i (x_i - y_i)^2\big(x_i^2 + x_i y_i + y_i^2\big) + 2\beta\sum_i (x_i - y_i)^2 \\ &\quad - J\sum_i \sum_{j\in N(i)} (x_i - y_i - x_j + y_j)^2 \\ &\leqslant -\sum_i (x_i - y_i)^4 + 2\beta \sum_i (x_i - y_i)^2 \\ &\leqslant -|\Lambda|^{-1}|x-y|^4 + 2\beta|x-y|^2,\end{aligned}$$

where $|\Lambda|$ is the cardinality of Λ. It follows that

$$\frac{\langle x - y,\, b(x) - b(y)\rangle}{|x-y|} \leqslant -\frac{1}{|\Lambda|}|x-y|^3 + 2\beta|x-y|.$$

If we take $f(r) = r$, then for all $x \neq y$, we have

$$\overline{L} f(|x-y|) = \frac{\langle x - y,\, b(x) - b(y)\rangle}{|x-y|} f'(|x-y|) \leqslant -\bigg(\frac{1}{|\Lambda|}|x-y|^2 - 2\beta\bigg)|x-y|.$$

This is not enough for the exponential convergence except in the case that $\beta < 0$ for which we have $\inf_{r>0}\big(r^2/|\Lambda| - 2\beta\big) = -2\beta > 0$. Due to this reason, we need a much carefully designed f. Define the function f as in the theorem, then we have

$$f'(r) = e^{-C(r)} \int_r^\infty e^C \sqrt{\varphi}, \qquad f'' = -\frac{1}{4}\gamma f' - \sqrt{\varphi}.$$

We obtain

$$4f'' + \gamma f' = -4\sqrt{\varphi} \leqslant -\varepsilon f$$

with

$$\gamma(r) = -\frac{1}{|\Lambda|} r^3 + 2\beta r, \qquad \varepsilon = 4\inf_{r>0} \frac{\sqrt{\varphi(r)}}{f(r)}.$$

By the Cauchy mean value theorem, it follows that

$$\begin{aligned}\inf_{r>0}\frac{\sqrt{\varphi}}{f} &\geqslant \inf_{r>0}\frac{(\sqrt{\varphi})'}{f'} = \frac{1}{2}\inf_{r>0}\varphi^{-1/2}\bigg/\int_r^\infty e^C\sqrt{\varphi} \\ &\geqslant \frac{1}{2}\inf_{r>0}\frac{(\varphi^{-1/2})'}{-e^C\sqrt{\varphi}} = \frac{1}{4}\bigg(\inf_{r>0}\frac{e^{-C}}{\varphi}\bigg)^2 > 0.\end{aligned}$$

Therefore we obtain $\varepsilon > 0$. This proves our second assertion.

The exponential ergodicity is easy to check by using the so called "drift condition" with test function $x \to |x|^2$, but this is not enough to get a convergence rate. We now prove the exponential ergodicity with respect to $f \circ |\cdot|$. Note that here we do not assume that $f \circ |\cdot|$ is a distance. Otherwise, the assertion follows from [17; Theorem 5.23]. We have proved in the last paragraph that $\overline{\mathbb{E}}^{x,y} f(|X_t - Y_t|)$ is continuous in y. Moreover

$$\overline{\mathbb{E}}^{x,\mu_U} f(|X_t - Y_t|) = \int_{\mathbb{R}^{|\Lambda|}} \mu_U(\mathrm{d}y) \overline{\mathbb{E}}^{x,y} f(|X_t - Y_t|) \leqslant e^{-\varepsilon t} \int_{\mathbb{R}^{|\Lambda|}} \mu_U(\mathrm{d}y) f(|x-y|),$$

where μ_U is the probability measure having density e^{-U}/Z_U, corresponding to the potential

$$U(x) = \sum_{i\in\Lambda} u(x_i) + J \sum_{i\in\Lambda} \sum_{j\in N(i)} (x_i - x_j)^2.$$

Because the left-hand side controls the Wasserstein distance, with respect to the cost function $f\circ|\cdot|$, of the laws of the processes starting from x and μ_U respectively, we obtain an exponential ergodicity provided

$$\int_{\mathbb{R}^{|\Lambda|}} \mu_U(\mathrm{d}y) f(|x-y|) < \infty.$$

To check this, noting that

$$-U(x) \leqslant \sum_{i\in\Lambda} \big(-x_i^4 + \beta x_i^2 \big) \leqslant -\frac{1}{|\Lambda|}|x|^4 + \beta |x|^2$$

and $f(|x-y|) \leqslant f(|x|+|y|)$, it suffices to consider the radius part. That is,

$$\int_0^\infty f(r+z) \exp[-z^4/|\Lambda| + \beta z^2] \mathrm{d}z < \infty \qquad \text{for every } r \geqslant 0.$$

This can be done by using a comparison:

$$\begin{aligned}
\frac{f(r+z)\exp[-z^4/|\Lambda| + \beta z^2]}{z^{-2}} &= \frac{f(r+z)}{z^{-2}\exp[z^4/|\Lambda| - \beta z^2]} \\
&\sim \frac{e^{-C(r+z)} \int_{r+z}^\infty e^C \sqrt{\varphi}}{[-2z^{-3} + z^{-2}(4z^3/|\Lambda| - 2\beta z)]\exp[z^4/|\Lambda| - \beta z^2]} \\
&\sim \frac{\int_{r+z}^\infty e^C \sqrt{\varphi}}{z \exp[z^4/|\Lambda| - \beta z^2 + C(r+z)]} \\
&\sim 0 \qquad \text{as } z \to \infty.
\end{aligned}$$

Finally, by [17; Theorem 9.18] and its remark, we also have $\lambda_1(U, \Lambda, \beta) > 0$. $\square$

Theorem 6.1 is meaningful since it works for all finite dimensions. Note that $\varepsilon(\Lambda, \beta) \to 0$ as $|\Lambda| \to \infty$, which is natural since the model exhibits a phase transition. However, this result does not describe an ergodic region in the infinite dimensional situation.

For the remainder of this section, we apply the results obtained in the previous sections to some specific continuous spin systems. Denote by $\langle ij\rangle$ the nearest bonds in $\mathbb{Z}^d$, $d \geqslant 1$. Set $N(i) = \{j : j \text{ is the endpoint of an bond } \langle ij\rangle\}$. Then, $|N(i)| :=$ the cardinality of the set $N(i) = 2d$. Consider the Hamiltonian $H(x) = J\sum_{\langle ij\rangle}(x_i - x_j)^2$, where $J \geqslant 0$ is a constant. For a finite set $\Lambda \subset \mathbb{Z}^d$ (denoted by $\Lambda \Subset \mathbb{Z}^d$) and a point $\omega \in \mathbb{R}^{\mathbb{Z}^d}$, define the finite-dimensional conditional Gibbs distribution $\mu_U^{\Lambda,\omega}$ as follows.

$$\mu_U^{\Lambda,\,\omega}(\mathrm{d}x_\Lambda) = e^{-U_\Lambda^\omega(x_\Lambda)}\,\mathrm{d}x_\Lambda \big/ Z_\Lambda^\omega, \tag{6.1}$$

where $x_\Lambda = (x_i, i\in\Lambda)$, Z_Λ^ω is the normalizing constant and

$$U_\Lambda^\omega(x_\Lambda) = \sum_{i\in\Lambda} u(x_i) + J\sum_{\langle ij\rangle:\, i,j\in\Lambda}(x_i - x_j)^2 + J\sum_{i\in\Lambda,\, j\in N(i)\setminus\Lambda}(x_i-\omega_j)^2 \tag{6.2}$$

for some function $u\in C^\infty(\mathbb{R})$, to be specified latterly. One can rewrite U_Λ^ω as

$$U_\Lambda^\omega(x_\Lambda) = \sum_{i\in\Lambda} u(x_i) + J\sum_{i\in\Lambda}\sum_{j\in N(i)}(x_i - z_j)^2, \tag{6.3}$$

where

$$z_j = \begin{cases} x_j, & \text{if } j\in\Lambda\\ \omega_j, & \text{if } j\notin\Lambda. \end{cases}$$

Correspondingly, we have an operator L_Λ^ω and a Dirichlet form D_Λ^ω as follows.

$$L_\Lambda^\omega = \Delta_\Lambda - \langle\nabla_\Lambda U_\Lambda^\omega,\, \nabla_\Lambda\rangle, \qquad D_\Lambda^\omega(f) = \int_{\mathbb{R}^{|\Lambda|}} |\nabla_\Lambda f|^2 \mathrm{d}\mu_U^{\Lambda,\,\omega}. \tag{6.4}$$

Our purpose in this section is to estimate $\lambda_1\big(L_\Lambda^\omega\big) = \lambda_1\big(U_\Lambda^\omega\big)$. By (1.6), we have the simplest lower bound of the marginal eigenvalues as follows.

$$\lambda_1^{x_{\Lambda\setminus i},\,\omega} \geqslant \inf_{x\in\mathbb{R}} u''(x) + 4dJ, \tag{6.5}$$

where $x_{\Lambda\setminus i} = (x_j, j\in\Lambda\setminus\{i\})$. The function $C(x)$ defined in Section 4 becomes

$$\begin{aligned} C_\Lambda^{x_{\Lambda\setminus i},\,\omega}(x_i) &= -u(x_i) - J\sum_{j\in N(i)}(x_i - z_j)^2\\ &= -u(x_i) - 2dJx_i^2 + 2J\bigg(\sum_{j\in N(i)} z_j\bigg)x_i - J\sum_{j\in N(i)} z_j^2,\\ &\quad i\in\Lambda. \end{aligned} \tag{6.6}$$

The last term can be ignored, since it does not make influence to $\mu_U^{x_{\Lambda\setminus i}}$, and so neither $\lambda_1^{x_{\Lambda\setminus i}}$. The coefficient of the second to the last term varies over whole $\mathbb{R}$ if $J\neq 0$.

We consider two models only: $u(x) = \alpha x^2$ and $u(x) = x^4 - \beta x^2$ for some constants $\alpha > 0$ and $\beta\in\mathbb{R}$, respectively.

Theorem 6.2. Let $u(x) = \alpha x^2$ for some constant $\alpha > 0$ and let $U(x) = \sum_i u(x_i) + H(x)$ with Hamiltonian $H(x) = J\sum_{\langle ij\rangle}(x_i - x_j)^2$. Then we have

$$\inf_{\Lambda \Subset \mathbb{Z}^d} \inf_{\omega \in \mathbb{R}^{\mathbb{Z}^d}} \lambda_1\big(U_\Lambda^\omega\big) \geqslant \inf_{\Lambda \Subset \mathbb{Z}^d} \inf_{\omega \in \mathbb{R}^{\mathbb{Z}^d}} \sigma\big(U_\Lambda^\omega\big) \geqslant 2\alpha. \tag{6.7}$$

Proof. It suffices to prove the second estimate. By Example 5.2 and Theorem 1.3, the proof is very much the same as proving

$$\inf_{\Lambda \Subset \mathbb{Z}^d} \inf_{\omega \in \mathbb{R}^{\mathbb{Z}^d}} \lambda_1\big(U_\Lambda^\omega\big) \geqslant 2\alpha.$$

Hence we prove here the last assertion only. First, we have

$$|\partial_{ij}U(x)| = \begin{cases} 2J, & i, j \in \Lambda,\ |i-j| = 1 \\ 0, & i, j \in \Lambda,\ |i-j| > 1. \end{cases} \tag{6.8}$$

The right-hand side is independent of x, which is the main reason why we were looking for the uniform estimates (with respect to the linear term) in Examples 4.2 and 4.3. By (6.5), we have $\lambda_1^{x_{\Lambda\setminus i},\,\omega} \geqslant 2\alpha + 4dJ$, which is indeed sharp in view of Example 4.2. Combining these facts together and using (1.4) with $w_i \equiv 1$, it follows that

$$\begin{aligned} \lambda_1\big(U_\Lambda^\omega\big) &\geqslant \inf_{x \in \mathbb{R}^{|\Lambda|}} \min_{i \in \Lambda} \Bigg[2\alpha + 4dJ - \sum_{j \in \Lambda:\, |i-j|=1} 2J\Bigg] \\ &= 2\alpha + 4dJ - 2J \max_{i \in \Lambda} \big|\{\langle i,\, j\rangle : j \in \Lambda\}\big| \\ &\geqslant 2\alpha \end{aligned}$$

uniformly in $\omega \in \mathbb{R}^{\mathbb{Z}^d}$ and $\Lambda \Subset \mathbb{Z}^d$. The sign of the last equality holds once Λ contains a point together with all of its neighbors. $\square$

In the last step of the proof, we did not use Theorem 1.2 since the matrix $\big(|\partial_{ij}U(x)| : i, j \in \Lambda\big)$ is very simple. Nevertheless, it provides us a good chance to justify the power of Theorem 1.2. To do so, take $\eta_i^{x_{\Lambda\setminus i}} = 2\alpha + 4dJ = \lambda_1^{x_{\Lambda\setminus i}}$. Then

$$\begin{aligned} &s_i(x) = \eta_i^{x_{\Lambda\setminus i}} - \sum_{j \in \Lambda: j \neq i} |\partial_{ij}U(x)| = 2\alpha + 4dJ - 2J\big|\{\langle i, j\rangle : j \in \Lambda\}\big|, \qquad i \in \Lambda, \\ &\underline{s}(x) = \min_{i \in \Lambda} s_i(x) = 2\alpha. \end{aligned}$$

Since $h^{(\gamma)} \geqslant 0$, Theorem 1.2 already gives us $\lambda_1\big(U_\Lambda^\omega\big) \geqslant \inf_x \underline{s}(x) = 2\alpha$ as expected, without using $h^{(\gamma)}$. To see the role played by $h^{(\gamma)}$, note that

$$\begin{aligned} &q_i(x) = \eta_i^{x_{\Lambda\setminus i}} - \underline{s}(x) = 4dJ, \qquad i \in \Lambda, \\ &d_i(x) = s_i(x) - \underline{s}(x) = 4dJ - 2J\big|\{\langle i, j\rangle : j \in \Lambda\}\big|, \qquad i \in \Lambda. \end{aligned}$$

Note that $d_i(x)$ here depends on i. Thus

$$\begin{aligned}
h^{(\gamma)}(x) &= \min_{A:\,\emptyset\neq A\subset\Lambda}\frac{1}{|A|}\bigg[\sum_{i\in A}\frac{d_i(x)}{q_i(x)^\gamma}+\sum_{i\in A,\,j\in\Lambda\setminus A}\frac{|\partial_{ij}U(x)|}{[q_i(x)\vee q_j(x)]^\gamma}\bigg]\\
&= \frac{2J}{(4dJ)^\gamma}\min_{A:\,\emptyset\neq A\subset\Lambda}\frac{1}{|A|}\sum_{i\in A}\Big[2d-\big|\{\langle i,j\rangle: j\in\Lambda\}\big|+\big|\{\langle i,j\rangle: j\in\Lambda\setminus A\}\big|\Big]\\
&= \frac{2J}{(4dJ)^\gamma}\min_{A:\,\emptyset\neq A\subset\Lambda}\frac{1}{|A|}\sum_{i\in A}\Big[\big|\{\langle i,j\rangle\}\big|-\big|\{\langle i,j\rangle: j\in A\}\big|\Big]\\
&= \frac{2J}{(4dJ)^\gamma}\min_{A:\,\emptyset\neq A\subset\Lambda}\frac{1}{|A|}\sum_{i\in A}\big|\{\langle i,j\rangle: j\notin A\}\big|.\\
&=: \frac{2J}{(4dJ)^\gamma}\min_{A:\,\emptyset\neq A\subset\Lambda}\frac{|\partial A|}{|A|}.
\end{aligned}$$

Clearly, the right-hand side depends reasonably on the geometry of Λ. Roughly speaking, by the isoperimetric principle, the last minimum of the ratio is approximately $|\partial B|/|B|$, where B is the largest ball contained in Λ. Anyhow, for regular Λ (cube for instance),

$$h^{(\gamma)}(x)\leqslant\frac{2J}{(4dJ)^\gamma}\cdot\frac{|\partial\Lambda|}{|\Lambda|}\to 0\qquad\text{as }\Lambda\uparrow\mathbb{Z}^d.$$

Hence for this model, $h^{(\gamma)}$ makes no contribution to $\lambda_1\big(U_\Lambda^\omega\big)$ for the estimate uniformly in Λ.

Theorem 6.3. Let $u(x)=x^4-\beta x^2$ for some constant $\beta\in\mathbb{R}$ and let $U(x)=\sum_i u(x_i)+H(x)$ with Hamiltonian $H(x)=-2J\sum_{\langle ij\rangle}x_ix_j$. Then we have

$$\begin{aligned}
\inf_{\Lambda\Subset\mathbb{Z}^d}\inf_{\omega\in\mathbb{R}^{\mathbb{Z}^d}}\lambda_1\big(U_\Lambda^\omega\big)&\geqslant\inf_{\Lambda\Subset\mathbb{Z}^d}\inf_{\omega\in\mathbb{R}^{\mathbb{Z}^d}}\sigma\big(U_\Lambda^\omega\big)\\
&\geqslant\frac{\sqrt{\beta^2+8}-\beta}{\sqrt{e}}\exp\Big[-\frac{1}{8}\beta\Big(\beta+\sqrt{\beta^2+8}\Big)\Big]-4dJ,
\end{aligned}\tag{6.9}$$

For simplicity, we write $r=2dJ$. The right-hand side is positive if $(\beta,r)\in\mathbb{R}\times\mathbb{R}_+$ is located in the region below the curve in Figure 1 (including the region of $\beta\leqslant 0$ vertically below the shade one.)

Proof. As shown in part (2) of Example 4.6, for zero boundary condition $\omega=0$, we have

$$\lim_{\beta\to\infty}\sigma^{x_{\Lambda\setminus i},\,\omega}\leqslant\lim_{\beta\to\infty}\lambda_1^{x_{\Lambda\setminus i},\,\omega}=0.$$

In other words, due to the double-well potential, the spectral gap and then the logarithmic constant will be absorbed as $\beta\to\infty$. Combining Example 5.3 with Theorem 1.3 and following the last step of the proof Theorem 6.2, we obtain the required lower estimate. □

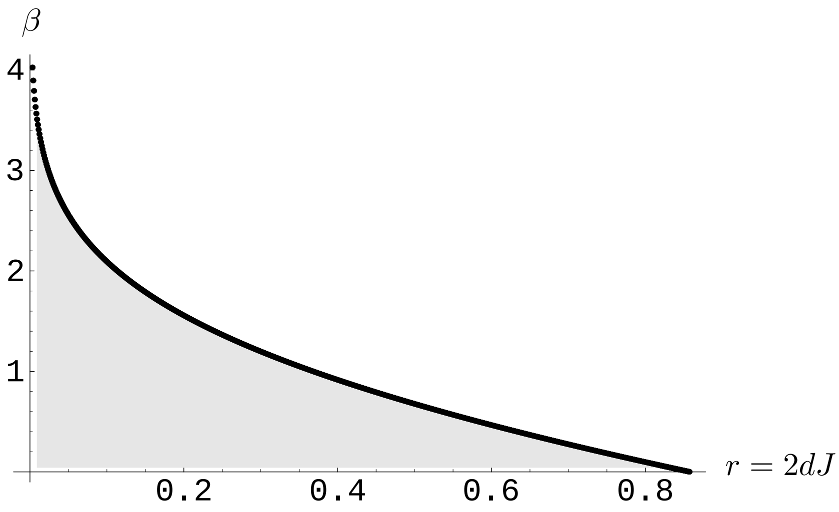


Figure 1.

For the Hamiltonian $H(x) = J\sum_{\langle ij\rangle}(x_i - x_j)^2$ discussed several times before, simply replacing β with $\beta - 2dJ$ in Theorem 6.3, we obtain the following estimate:

$$\begin{aligned}
&\inf_{\Lambda\Subset\mathbb{Z}^d}\ \inf_{\omega\in\mathbb{R}^{\mathbb{Z}^d}} \lambda_1\bigl(U_\Lambda^\omega\bigr)\\
&\qquad\geqslant \inf_{\Lambda\Subset\mathbb{Z}^d}\ \inf_{\omega\in\mathbb{R}^{\mathbb{Z}^d}} \sigma\bigl(U_\Lambda^\omega\bigr)\\
&\qquad\geqslant \frac{\sqrt{(\beta-r)^2+8}-\beta+r}{\sqrt{e}}\exp\left[-\frac{1}{8}(\beta-r)\Bigl(\beta-r+\sqrt{(\beta-r)^2+8}\Bigr)\right]\\
&\qquad\quad - 2r, \qquad (6.10)
\end{aligned}$$

where $r = 2dJ$. The ergodic region is shown in Figure 2.

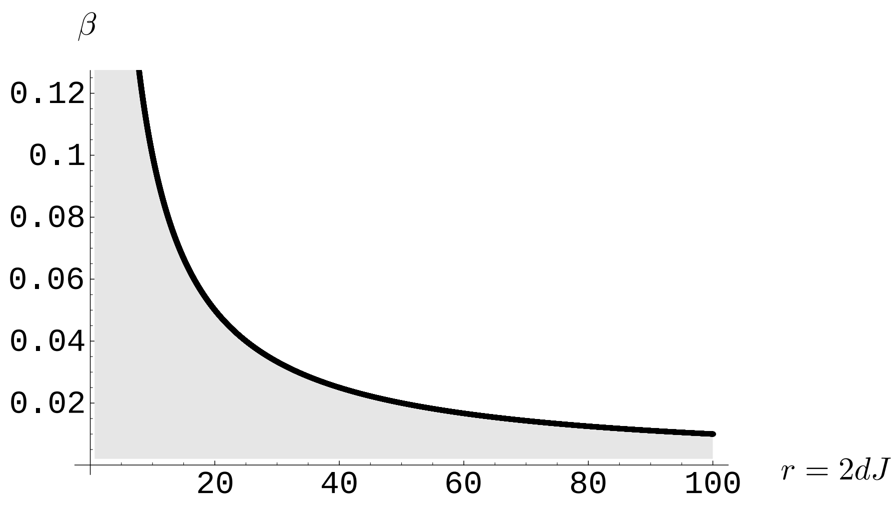


Figure 2.

Remark 6.4. As mentioned below the proof of Proposition 3.1, by considering the interacting terms more carefully, one may improve Theorem 1.1 for stronger interactions. For instance, since the variance of a random variable having the distribution with density $\exp[-x^4+\beta x^2]/Z$ on the real line is asymptotically $\beta/2$ for $\beta \geqslant 0$, and is bounded above by

$$\frac{\Gamma(3/4)}{\Gamma(1/4)} + \beta \Big/ \bigg[2 + \frac{4\Gamma(1/4)}{9(1+\beta)\Gamma(3/4)}\bigg],$$

by using [9; Proposition 5.8], when $\beta \geqslant 0$, the lower bound of $\inf_{\Lambda \Subset \mathbb{Z}^d} \inf_{\omega \in \mathbb{R}^{\mathbb{Z}^d}} \lambda_1\big(U_\Lambda^\omega\big)$ given in Theorem 6.3 can be improved as follows: replacing the interaction term $4dJ$ in (6.9) with

$$4dJ\bigg[\frac{\Gamma(3/4)}{\Gamma(1/4)}+\beta \Big/ \bigg[2+\frac{4\Gamma(1/4)}{9(1+\beta)\Gamma(3/4)}\bigg]\bigg] \frac{\sqrt{\beta^2+8}-\beta}{\sqrt{e}} \exp\bigg[-\frac{1}{8}\beta\Big(\beta+\sqrt{\beta^2+8}\Big)\bigg]. \tag{6.11}$$

Finally, we mention that there is another technique which works even in the irreversible situation (cf. [17; Theorem 14.10]) to handle with the exponentially ergodic region, because the second model (Theorem 6.3) is attractive (stochastic monotone) and has the moments of all orders, plus a use of the translation invariant. However, as known that the logarithmic Sobolev inequality already implies an exponential ergodicity in the entropy and moreover, the usual exponential ergodicity is equivalent to the Poincaré inequality with nearly the same convergence exponent in the present context (cf. [13; Theorem 8.13]), there is almost no room to improve the ergodic region.

Acknowledgements. This paper takes an unusual long period in preparation. The most part of the paper was finished in 2002, but the exact coefficient 1/4 of decay rate $\exp[-\beta^2/4]$ for the second model (Theorem 6.3) was left to be open and so the earlier draft was communicated within a small group only. Recently, the author came back to compute the logarithmic Sobolev constant which leads to the precise coefficient and hence completes the paper. The author would like to acknowledge the organizers for several conferences in which partial results of the paper were presented: I. Shigekawa (the 11th International Research Institute of Mathematical Society of Japan, 2002), L.M. Wu (Chinese and French Workshop on Probability and Applications, 2004), Z.M. Ma and M. Röckner (Second Sino-German Meeting on Stochastic Analysis, 2007). The author is also greatly indebted to the referees for their helpful comments.

References

[1]. Bodineau, T. and Helffer, B. (1999), *The log-Sobolev inequality for unbounded spins systems*, J. Funct. Anal. 166:1, 168-178.
[2]. Bodineau, T. and Helffer, B. (2000), *Correlations, spectral gap and log-Sobolev inequalities for unbounded spins systems*, Diff. Eq. Math. Phys., AMS/IP Stud. Adv. Math. 16, Amer. Math. Soc., Providence, RI, 51–66.

[3]. Deuschel J.-D. and D. W. Stroock (1990), *Hypercontractivity and spectral gap of symmetric diffusions with applications to the stochastic Ising models*, J. Funct. Anal. 92, 30–48.

[4]. Gao, F.Q. and Wu, L.M. (2007),, *Transportation-information inequalities for Gibbs measures*, preprint.

[5]. Gentil, I. and Roberto, C. (2001), *Spectral gaps for spin systems: some non-convex phase examples*, J. Funct. Anal. 180, 66–84.

[6]. Helffer, B. (1999), *Remarks on decay of correlations and Witten Laplacians III. Application to logarithmic Sobolev inequalities*, Ann. Inst. H. Poincaré (B), Prob. Stat. 35:4, 483-508.

[7]. Ledoux, M. (2001), *Logarithmic Sobolev inequalities for unbounded spin systems revised*, "Séminaire de Probabilités" XXXV. LNM 1755, 167–194. Springer.

[8]. Otto, F. and Reznikoff, M.G. (2007), *A new criterion for the logarithmic Sobolev inequality and two applications*, J. Funct. Anal. 243, 121–157.

[9]. Wu, L. M. (2006), *Poincaré and transportation inequalities for Gibbs measures under the Dobrushin uniqueness condition*, Ann. Prob. 34:5, 1960–1989.

[10]. Yoshida, N. (1999), *The log-Sobolev inequality for weakly coupled lattice field*, Prob. Theor. Relat. Fields 115, 1–40.

[11]. Zegarlinski, B. (1996), *The strong decay to equilibrium for the stochastic dynamics of unbounded spin systems on a lattice*, Comm. Math. Phys. 175, 401–432.

[12]. Chen, M. F. and Wang, F. Y. (1997), *Estimation of spectral gap for elliptic operators*, Trans. Amer. Math. Soc. 349, 1239–1267.

[13]. Chen, M. F. (2005), *Eigenvalues, Inequalities, and Ergodic Theory*, Springer, London.

[14]. Bakry, D. and Emery, M. (1985), *Diffusions hypercontractives*, in "Seminaire de probabilit'es" XIX, 1983/84, Springe, Berlin, 177–206.

[15]. Chen, M. F. (2000), *The principal eigenvalue for jump processes*, Acta Math. Sin. Eng. Ser. 16:3, 361–368.

[16]. Chen, M. F. and Wang, F. Y. (1998), *Cheeger's inequalities for general symmetric forms and existence criteria for spectral gap*, Abstract. Chin. Sci. Bull. 43:18, 1516–1519. Ann. Prob. 2000, 28:1, 235–257.

[17]. Chen, M. F. (1992), *From Markov Chains to Non-Equilibrium Particle Systems.*, World Scientific. 2nd ed. 2004.

[18]. Chen, M. F. (2000), *Explicit bounds of the first eigenvalue*, Sci. Chin. Ser. A 43(10), 1051–1059.

[19]. Chen, M. F. (2001), *Variational formulas and approximation theorems for the first eigenvalue in dimension one*, Sci. Chin. Ser. A 44(4), 409–418.

[20]. Conte, J. M., et al (1963/64), *Solution of Problem* 5607, Revue Math. Spéc. 74, 227–230.

[21]. Gautschi, W. (1959), *Some elementary inequalities relating to the gamma and incomplete gammafunction*, J. Math. and Phys. 38, 77–81.

[22]. Chen, M. F. and Wang, F. Y. (1997), *General formula for lower bound of the first eigenvalue on Riemannian manifolds*, Sci. Chin. Ser. A 40:4, 384–394.

[23]. Chen, M. F. and F. Y. Wang (1997), *Estimates of logarithmic Sobolev constant*, J. Funct. Anal. 144:2, 287–300.

Note: Figures 1 and 2 were missed in the publication and the correction appeared in the same journal, Vol. 25, No. 12, pp. 2199-2199.