Ergodic Convergence Rates of Markov Processes

—Eigenvalues, Inequalities and Ergodic Theory

(Volume I)

Mu-Fa Chen
(Beijing Normal University)

First Edition 2000
Second Edition 2009

Preface to the First Edition

This book is a collection of the papers, during 1993–2000, by the author with cooperators in the field mentioned in the title. More precisely, the topics treated in the book are the first (non-trivial) eigenvalue, spectral gap, Poincarè, logarithmic Sobolev, Nash, Liggett, Liggett-Stroock inequalities, which describe some different types of exponential or algebraic convergence of Markov processes. The relation between these inequalities and three types of traditional ergodicity for Markov processes is also studied.

The papers are arranged according to the order of writing time. Of course, one may read them in a different order. For instance, before going to the details, one may look at the survey articles [10] and [21] for the main results and [11]–[13], [22] for the main ideas.

The book is informal at the present stage. It serves only for communication but not for publication. The purpose of the edition (into a book form) is to save the reader's time in seeking for various journals. Occasionally, there are some additions (unpublished details) or corrections to the original papers.

June 4, 2000. Rome, Italy

Preface to the Second Edition

In the past years, the book has been updated several time by adding some subsequent papers. It is now compiled in a smaller text size with 11pt fonts. Some additional corrections are made and some recent papers are included. Besides, the author's earliest article [03] in the field and two earlier articles on couplings ([01] and [02]) are also in included for the reader's convenience. To keep in a reasonable size, the book is now divided into three volumes.

November 23, 2009. Beijing, China

Contents

Volume I

[01] Coupling for jump processes, Acta Math. Sin. New Ser. 1986, 2:2, 123–136. 1

[02] With S.F. Li, Coupling methods for multidimensional diffusion processes, Ann. of Probab. 1989, 17:1, 151-177 14

[03] Exponential L^2-convergence and L^2-spectral gap for Markov processes, Math. Sin. New Ser. 1991, 7:1, 19–37. 43

[1] With F.Y. Wang, Application of coupling method to the first eigenvalue on manifold, Sci. Sin.(A) 1993, 23:11 (Chinese Edition), 1130-1140; 1994, 37:1 (English Edition), 1-14. 65

[2] Optimal Markovian couplings and applications, Acta Math. Sin. New Ser. 1994, 10:3, 260-275. 82

[3] Optimal couplings and application to Riemannian geometry, Prob. Theory and Math. Stat., Vol. 1, Eds. B. Grigelionis et al, 1994 VPS/TEV, 121-142. 102

[4] On the ergodic region of Schlögl's model, in Proceedings of International Conference on Dirichlet Forms and Stochastic Processes, Edited by Z.M. Ma, M. Röckner and J.A. Yan, Walter de Gruyter Publishers, 1995, 87-102. 124

[5] With F.Y. Wang, Estimation of the first eigenvalue of second order elliptic operators, J. Funct. Anal. 1995, 131:2, 345-363. 139

[6] With F.Y. Wang, Estimates of logarithmic Sobolev constant, J. Funct. Anal. 1997, 144:2, 287-300. 156

[7] Estimation of spectral gap for Markov chains, Acta Math. Sin. New Series 1996, 12:4, 337-360. 169

[8] With F.Y. Wang, Estimation of spectral gap for elliptic operators, Trans. Amer. Math. Soc. 1997, 349:3, 1239-1267. 203

[9] With F.Y. Wang, General formula for lower bound of the first eigenvalue on Riemannian manifolds, Sci. Sin. 1997, 27:1, 34–42 (Chinese Edition); 1997, 40:4, 384–394 (English Edition) 237

[10] 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 Statis. **128**, 123–136, Springer-Verlag, 1998. 249

[11] Coupling, spectral gap and related topics (I), Chin. Sci. Bulletin, 1997, 42:14, 1472–1477 (Chinese Edition); 1997, 42:16, 1321–1327 (English Edition). 262

[12] Coupling, spectral gap and related topics (Ⅱ), Chin. Sci. Bulletin, 1997, 42:15, 1585–1591 (Chinese Edition); 1997, 42:17, 1409–1416 (English Edition). 270

[13] Coupling, spectral gap and related topics (Ⅲ), Chin. Sci. Bulletin, 1997, 42:16, 1696–1703 (Chinese Edition); 1997, 42:18, 1497–1505 (English Edition). 279

[14] 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. 289

[15] With F. Y. Wang, Cheeger's inequalities for general symmetric forms and existence criteria for spectral gap. Abstract: Chin. Sci. Bulletin 1998, 43:14 (Chinese Edition), 1475–1477; 43:18 (English Edition), 1516–1519. Ann. Prob. 2000, 28:1, 235–257 305

[16] Analytic proof of dual variational formula for the first eigenvalue in dimension one, Sci. in China (A) 1999, 29:4 (Chinese Edition), 327–336; 42:8 (English Edition), 805–815. 331

[17] Nash inequalities for general symmetric forms, Acta Math. Sin. Eng. Ser. 1999, 15:3, 353–370. 348

Acta Mathematica Sinica, New Series
1986. Vol. 2, No. 2. pp. 123–136

COUPLING FOR JUMP PROCESSES*)

Chen Mufa

Department of Mathematics, Beijing Normal University,
Received March 16, 1985 Revised April 20, 1985

Coupling is probably the most important technique in the subject of interacting particle systems. It is also very useful for other stochastic processes. For discrete time Markov processes, the coupling theory was studied expansively by Dobrushin[8], Griffeath[9], Wasershtein[10] and others (see the conferences in [9]). For continuous time Markov processes, it becomes more complicated. This paper is devoted to discussing the coupling theory for jump Markov processes.

In Section 1 we introduce three basic conditions for a coupling. Then, in Sections 2–4, we discuss the conditions respectively. Finally, Section 5 presents some basic couplings which should be the most useful ones in the subject we study. The main results of the paper are given by Theorems (13), (16), (21), (24), (26), (30), (36) and (37).

In the subsequent paper[6], which is mainly based on this paper, we will give a construction for large classes of Markov processes on product spaces which need not be compact.

§1 Basic Conditions for Coupling

Let $(E_i, \mathscr{E}_i)$ be an arbitrary measurable space and $(X_t^{(i)})_{t\geqslant 0}$ be a Markov process, $i = 1, 2$. A coupling is simply to construct a Markov process $\big(\widetilde{X}_t\big)_{t\geqslant 0}$, of the two processes $(X_t^{(i)})_{t\geqslant 0}$, $i = 1, 2$ on a common probability space with the product state space $(E, \mathscr{E}) = (E_1 \times E_2, \mathscr{E}_1 \times \mathscr{E}_2)$, which has the property:

(1) marginality.

$$\widetilde{P}^{(x_1,x_2)}\big[\widetilde{X}_t \in A_1 \times E_2\big] = P^{x_1}[X_t^{(1)} \in A_1]$$
$$\widetilde{P}^{(x_1,x_2)}\big[\widetilde{X}_t \in E_1 \times A_2\big] = P^{x_2}[X_t^{(2)} \in A_2]$$
$$x_i \in E_i,\ A_i \in \mathscr{E}_i,\ i = 1,\ 2,\ t \geqslant 0.$$

*) Partially supported by the Ministry of Education and the Foundation of Zhongshan University Advanced Research Centre.

By using the transition probability function, one can rewrite (1) as:

$$\begin{aligned}&\widetilde{P}(t,(x_1,x_2),A_1\times E_2)=P_1(t,x_1,A_1)\\ &\widetilde{P}(t,(x_1,x_2),E_1\times A_2)=P_2(t,x_2,A_2)\\ &\qquad x_i\in E_i,\ A_i\in\mathscr{E}_i,\ i=1,\ 2,\ t\geqslant 0.\end{aligned}\tag{2}$$

Throughout the paper, we assume each $(E_i,\mathscr{E}_i)$ is separable. That is, $\{x\}\in\mathscr{E}_i$ for each $x\in E_i$. Also, we restrict ourselves on jump process $P_i(t,x_i,\cdot)$ with totally stable and conservative q-pair $(q_i(x_i),q_i(x_i,\cdot))$, which means that

$$\begin{aligned}&q_i(x_i)=q_i(x_i,E_i)<\infty,\\ &\frac{\mathrm{d}}{\mathrm{d}t}P_i(t,x_i,B_i)\Big|_{t=0}=q_i(x_i,B_i),-q_i(x_i)\delta(x_i,B_i),\qquad x_i\in E_i,\ B_i\in\mathscr{E}_i,\ i=1,\ 2\end{aligned}$$

where $\delta(x,B)=I_B(x)=1$, if $x\in B$; $=0$, if $x\notin B$. We call a q-pair regular if it determines a unique jump process $P(t,x,\cdot)$.[1] Thus, a coupling for jump processes requires reasonably the following property:

(3) regularity. the q-pair $(\tilde{q}(\tilde{x}),\tilde{q}(\tilde{x},\cdot))$ is regular.

Sometimes, a coupling is used to compare an order relation of two copies of the same jump process with different starting points. In this case, $E_1=E_2=E$, $\mathscr{E}_1=\mathscr{E}_2=\mathscr{E}$ and E is endowed with a semi-order "$\leqslant$". One wants to know whether the process $(X_t)_{t\geqslant 0}$ has

(4) order-preservation.

$$x_1\leqslant x_2\Longrightarrow\widetilde{P}^{(x_1,x_2)}\big[X_t^{(1)}\leqslant X_t^{(2)}\big]=1,\qquad t\geqslant 0,\ (x_1,x_2)\in\widetilde{E}.$$

A function f on E is said monotone, if

$$x_1\leqslant x_2\Longrightarrow f(x_1)\leqslant f(x_2),\qquad (x_1,x_2)\in\widetilde{E}.\tag{5}$$

Now, if (2)—(4) are satisfied, then for each nonnegative monotone function f, we have

$$x_1\leqslant x_2\Longrightarrow P_t^{(1)}f(x_1)\leqslant P_t^{(2)}f(x_2),\qquad (x_1,x_2)\in\widetilde{E},\ t\geqslant 0.\tag{6}$$

where

$$P_t^{(i)}f(x)=\int P_i(t,x,\mathrm{d}y)f(y),\qquad i=1,\ 2.$$

The conditions (2), (3) and (4) are usually needed for a coupling. However, these conditions are indeed not explicit, they depend on the unknown process $\widetilde{P}(t,\tilde{x},\cdot)$. The explicit condition should be described by the given q-pairs

[1] It is also called a q-process.

$(q_i(x_i), q_i(x_i,\cdot))\,(i=1,2)$ only, and this point is just what we are going to do in the next three sections.

§2. Marginality

Let $\widetilde{P}(t,\tilde{x},\tilde{A})$ be a jump process with q-pair $\big(\tilde{q}(\tilde{x}),\tilde{q}(\tilde{x},\cdot)\big)$, then by the conservative assumption, one can see that

$$\lim_{t\downarrow 0}\frac{\widetilde{P}(t,\tilde{x},\tilde{A})-\delta(\tilde{x},\tilde{A})}{t}=\tilde{q}(\tilde{x},\tilde{A})-\tilde{q}(\tilde{x})I_{\tilde{A}}(\tilde{x}),\qquad \tilde{x}\in\widetilde{E},\ \tilde{A}\in\widetilde{\mathscr{E}}.$$

From condition (2), it follows that

$$\begin{aligned}q_1(x_1,A_1)-q_1(x_1)I_{A_1}(x_1)&=\lim_{t\downarrow 0}\frac{P_1(t,x_1,A_1)-\delta(x_1,A_1)}{t}\\&=\lim_{t\downarrow 0}\frac{\widetilde{P}(t,(x_1,x_2),A_1\times E_2)-\delta(x_1,A_1)}{t}\\&=\tilde{q}(x_1,x_2,A_1\times E_2)-\tilde{q}(x_1,x_2)I_{A_1}(x_1),\\&\qquad (x_1,x_2)\in\widetilde{E},\ A_1\in\mathscr{E}_1.\end{aligned}$$

Hence, by the monotone class theorem, we get

$$\begin{aligned}&\int q_1(x_1,\mathrm{d}y_1)f(y_1)-q_1(x_1)f(x_1)\\&=\int\tilde{q}(x_1,x_2;\mathrm{d}y_1,\mathrm{d}y_2)f(y_1)-\tilde{q}(x_1,x_2)f(x_1),\qquad (x_1,x_2)\in\widetilde{E},\ f\in{}_b\mathscr{E}_1,\end{aligned}$$

where ${}_b\mathscr{E}_1$ is the set of all bounded $\mathscr{E}_1$-measurable functions. Regarding ${}_b\mathscr{E}_1$ as a bivariable function, and using the following operators

$$\Omega_i g_i(x_i)=\int q_i(x_i,\mathrm{d}y_i)(g_i(y_i)-g_i(x_i)),\qquad g_i\in{}_b\mathscr{E}_i,\ i=1,\ 2$$

$$\widetilde{\Omega}f(x_1,x_2)=\int\tilde{q}(x_1,x_2;\mathrm{d}y_1,\mathrm{d}y_2)(f(y_1,y_2)-f(x_1,x_2)),\quad (x_1,x_2)\in\widetilde{E},\ f\in{}_b\mathscr{E},$$

one can rewrite the above equality as

$$\begin{aligned}&\widetilde{\Omega}f(\cdot,x_2)=\Omega_1 f\quad\text{independent of } x_2,\quad f\in{}_b\mathscr{E}_1;\\&\widetilde{\Omega}f(x_1,\cdot)=\Omega_2 f\quad\text{independent of } x_1,\quad f\in{}_b\mathscr{E}_2.\end{aligned}\tag{7}$$

In other words, we have proven

(8) Lemma. (2)$\Longrightarrow$(7)

Next, we prove that (7) $\Longrightarrow$ (2).

It is known that q-pair $(q(x),q(x,\cdot))$ on a separable measurable state space $(E,\mathscr{E})$ determines uniquely the minimal jump process $P^{\min}(t,x,\cdot)$. If we define

$$P^{\min}(\lambda,x,\cdot)=\int_0^\infty e^{-\lambda t}P^{\min}(t,x,\cdot)\mathrm{d}t,\qquad t\geqslant 0,\ x\in E\tag{9}$$

then $P^{\min}(\lambda,\cdot,A)$ is the minimal solution to the equation

$$f(x)=\int\frac{q(x,\mathrm{d}y)}{\lambda+q(x)}f(y)+\frac{\delta(x,A)}{\lambda+q(x)},\qquad x\in E\tag{10}$$

for each fixed $\lambda>0$ and $A\in\mathscr{E}$. We also call the Laplace transform $P(\lambda,x,\cdot)$ of a jump process $P(t,x,\cdot)$ a jump process.

(11) Lemma. Suppose that (7) holds, then

$$\begin{aligned}\widetilde{P}^{\min}(\lambda,(x_1,x_2),A_1\times E_2)&\leqslant P_1^{\min}(\lambda,x,A_1)\\ \widetilde{P}^{\min}(\lambda,(x_1,x_2),E_1\times A_2)&\leqslant P_2^{\min}(\lambda,x,A_2)\\ \lambda>0,\ x_i\in E_i,\ A_i\in\mathscr{E}_i,\ i=1,\ 2&\end{aligned}$$

where $P_i^{\min}(\lambda,x,\cdot)\,(i=1,2)$ and $\widetilde{P}^{\min}(\lambda,(x_1,x_2),\cdot)$ are the minimal jump processes determined by $(q_i(x_i),q_i(x_i,\cdot))$ and $(\tilde{q}(\tilde{x}),\tilde{q}(\tilde{x},\cdot))$, respectively. In particular, if $(\tilde{q}(\tilde{x}),\tilde{q}(\tilde{x},\cdot))$ is regular, then so are the marginals.

Proof. By the comparison theorem [2; Theorem 6], it suffices to show that $h(x_1,x_2):=P_1^{\min}(\lambda,x_1,A_1)$ satisfies the inequality

$$h(x_1,x_2)\geqslant\int\frac{\tilde{q}(x_1,x_2;\mathrm{d}y_1,\mathrm{d}y_2)}{\lambda+q(x_1,x_2)}h(y_1,y_2)+\frac{\delta(x_1,A_1)}{\lambda+\tilde{q}(x_1,x_2)},\qquad(x_1,x_2)\in\widetilde{E}.\tag{12}$$

This follows from (7) and (10) immediately. □

(13) Theorem. Suppose that $(\tilde{q}(\tilde{x}),\tilde{q}(\tilde{x},\cdot))$ is regular, then (2) $\Longleftrightarrow$ (7).

Proof. Since Lemma (8), it is enough to prove that (7)$\Longrightarrow$(2). By Lemma (11) and the assumption, one can see that

$$\begin{aligned}\widetilde{P}(\lambda,(x_1,x_2),A_1\times E_2)&\leqslant P_1(\lambda,x_1,A_1)\\ \lambda>0,\ x_i\in E_i,\ i=1,\ 2,\ A_1\in\mathscr{E}_1.&\end{aligned}\tag{14}$$

If

$$P(\lambda,(x_1,x_2),A_1\times E_2)<P_1(\lambda,x_1,A_1)\tag{15}$$

for some $\lambda>0$, $(x_1,x_2)\in\widetilde{E}$ and $A_1\in\mathscr{E}$, then

$$\begin{aligned}1&=\lambda\widetilde{P}(\lambda,(x_1,x_2),A_1\times E_2)+\lambda\widetilde{P}(\lambda,(x_1,x_2),A_1^c\times E_2)\\ &<\lambda P_1(\lambda,x_1,A_1)+\lambda P_1(\lambda,x_1,A_1^c)\\ &=\lambda P_1(\lambda,x_1,E_1)\\ &\leqslant 1.\end{aligned}$$

This is impossible. □

§3. Regularity

The uniqueness criteria for general q-process were obtained by Chen and Zheng[7]. In this section, we first present some sufficient conditions for uniqueness which are usually more practical. Then we study the relationship between the regularity of the coupled q-process and the regularities of its marginal q-processes.

(16) Theorem. Suppose that there exist a sequence $\{E_n\} \subset \mathscr{E}$ and an $\varphi \in \mathscr{E}_+$,[2] such that[3]

$$E_n \uparrow E, \quad \text{as } n \uparrow \infty; \qquad \sup_{x \in E_n} \varphi(x) < \infty, \tag{17}$$

$$\lim_{n \to \infty} \inf_{x \notin E_n} \varphi(x) = \infty; \tag{18}$$

and there also exists a $c \in \mathbb{R}$ such that

$$\int q(x, \mathrm{d}y)\varphi(y) \leqslant (c + q(x))\varphi(x), \qquad x \in E \tag{19}$$

then the q-process is unique, i. e., the q-pair $(q(x), q(x, \cdot))$ is regular.

Proof. Without loss of generality, we may assume that $c \geqslant 0$.

(a). Since for each $\lambda > 0$, $\int P^{\min}(\lambda, \cdot, \mathrm{d}y)\varphi(y)$ is the minimal nonnegative solution to the equation

$$f = \int \frac{q(\cdot, \mathrm{d}y)}{\lambda + q(\cdot)} f(y) + \frac{\varphi(\cdot)}{\lambda + q(\cdot)},$$

and by condition (19),

$$\frac{\varphi}{\lambda - c} = \int \frac{q(\cdot, \mathrm{d}y)}{\lambda + q} \frac{\varphi(y)}{\lambda - c} + \frac{\varphi}{\lambda + q}, \qquad \lambda > c$$

it follows from the comparison theorem that

$$\int P^{\min}(\lambda, \cdot, \mathrm{d}y)\varphi(y) \leqslant \frac{\varphi}{\lambda - c} < \infty.$$

(b). Set

$$q_n(x, \mathrm{d}y) = I_{E_n}(x) q(x, \mathrm{d}y), \quad q_n(x) = q_n(x, E), \qquad x \in E,\ n \geqslant 1 \tag{20}$$

then $(q_n(x), q_n(x, \cdot))$ is a regular bounded q-pair for each $n \geqslant 1$. Clearly, the q-pair $(q_n(x), q_n(x, \cdot))$ also satisfies condition (19), therefore, by (a), one can see that

$$\int P_n(\lambda, x, \mathrm{d}y)\varphi(y) \leqslant \frac{\varphi(x)}{\lambda - c} < \infty, \qquad x \in E,\ \lambda > c,\ n \geqslant 1.$$

(c). For $x \in E_n$, we have

$$\begin{aligned} P^{\min}(\lambda, x, E_n) &= \int \frac{q(x, \mathrm{d}y)}{\lambda + q(x)} P^{\min}(\lambda, y, E_n) + \frac{\delta(x, E_n)}{\lambda + q(x)} \\ &= \int \frac{q_n(x, \mathrm{d}y)}{\lambda + q_n(x)} P^{\min}(\lambda, y, E_n) + \frac{\delta(x, E_n)}{\lambda + q_n(x)} \end{aligned}$$

[2] the set of all nonnegative $\mathscr{E}$-measurable functions

[3] For condition (18), the author has a helpful discussion with S. Z. Tang.

and for $x \notin E_n$, we simply have

$$P^{\min}(\lambda, x, E_n) \geqslant 0 = \int \frac{q_n(x, \mathrm{d}y)}{\lambda + q_n(x)} P^{\min}(\lambda, y, E_n) + \frac{\delta(x, E_n)}{\lambda + q_n(x)}.$$

Thus, we always have

$$P^{\min}(\lambda, x, E_n) \geqslant \int \frac{q_n(x, \mathrm{d}y)}{\lambda + q_n(x)} P^{\min}(\lambda, y, E_n) + \frac{\delta(x, E_n)}{\lambda + q_n(x)}, \quad \lambda > 0,\ x \in E,\ n \geqslant 1.$$

Now, the comparison theorem gives us that

$$P^{\min}(\lambda, x, E_n) \geqslant P_n(\lambda, x, E_n), \qquad \lambda > 0,\ x \in E,\ n \geqslant 1.$$

(d). By (b) and (c), we get

$$\begin{aligned} \lambda P^{\min}(\lambda, x, E_n) &\geqslant \lambda P_n(\lambda, x, E_n) \\ &= 1 - \lambda P_n(\lambda, x, E_n^c) \\ &\geqslant 1 - \frac{\lambda \varphi(x)}{(\lambda - c) \inf_{z \notin E_n} \varphi(z)}, \qquad \lambda > c,\ x \in E. \end{aligned}$$

and so

$$\lambda P^{\min}(\lambda, x, E) \geqslant \lim_{n \to \infty} \lambda P^{\min}(\lambda, x, E_n) \geqslant 1, \qquad \lambda > c.$$

This completes our proof. □

(21) Theorem. For the uniqueness of q-processes, each of the following conditions is sufficient:

(i) there exist a constant $c \in \mathbb{R}$ and an $\varphi \in \mathscr{E}$ such that $\varphi \geqslant q$ and[4]

$$\int q(x, \mathrm{d}y)\varphi(y) \leqslant (c + q(x))\varphi(x), \qquad x \in E;$$

(ii) there exists a $\lambda_0 > 0$ such that

$$\int P^{\min}(\lambda_0, x, \mathrm{d}y)\varphi(y) < \infty, \qquad x \in E;$$

(iii) for each $t \geqslant 0$ and $x \in E$,

$$\int P^{\min}(t, x, \mathrm{d}y)\varphi(y) < \infty.$$

[4]Similar but stronger condition was given by Basis [l].

Proof. By the proof (a) of the above theorem, we have (i)$\Longrightarrow$(ii).

Now assume that condition (ii) holds. By the forward Kotmogorov equation[3]:

$$P^{\min}(\lambda,x,A)=\int P^{\min}(\lambda,x,\mathrm{d}y)\int_A\frac{q(y,\mathrm{d}z)}{\lambda+q(z)}+\frac{\delta(x,A)}{\lambda+q(x)}$$

and the monotone class theorem, it follows that

$$\int P^{\min}(\lambda,x,\mathrm{d}y)f(y)=\int P^{\min}(\lambda,x,\mathrm{d}y)\int\frac{q(y,\mathrm{d}z)}{\lambda+q(z)}f(z)+\frac{f(x)}{\lambda+q(x)},$$
$$\lambda>0,\ x\in E,\ f\in\mathscr{E}_+.$$

In particular, taking $\lambda=\lambda_0$, $f=\lambda_0+q$, we obtain

$$\int(\lambda_0+q(y))P^{\min}(\lambda_0,x,\mathrm{d}y)=\int P^{\min}(\lambda_0,x,\mathrm{d}y)q(y)+1,\qquad x\in E.$$

Combining this with (ii), we have

$$\lambda_0P(\lambda_0,x,E)=1,\qquad x\in E.$$

This certainly implies the uniqueness. The last assertion can be proved in a similar way. □

(22) Remark. It is easy to show that the condition (21) (i) implies the assumptions of Theorem (16). To see this, simply take

$$E_n=\{x\in E:q(x)\leqslant n\},\qquad\varphi(x)=q(x),\qquad x\in E.$$

but the converse fails. The following counterexample is due to J. L. Zheng:

Take $E=\{1,2,\cdots\}$ and let $q_1,q_2,\cdots$ be the prime numbers in the natural order. Set

$$q_{i,i+1}=q_i,\quad i\in E;\qquad q_{ij}=0,\quad j\neq i,i+1.$$

This Q-matrix (q_{ij}) satisfies the assumptions of Theorem (16). To this end, we take $c=1$,

$$\varphi_1=1,\qquad\varphi_i=\prod_{k=1}^{i-1}(1+1/q_k),\quad i\geqslant 2.$$

Since $\prod_{n=1}^{\infty}(1+1/q_n)$ and $\sum_{n=1}^{\infty}1/q_n$ are convergent or divergent simultaneously, it follows that $\lim_{n\to\infty}\varphi_n=\infty$. Therefore, the assumptions are satisfied with

$$E_n=\{0,1,2,\cdots,n\},$$

and so the Q-process is unique.

Next we show that the condition (21) (i) fails. Indeed, we will show that the condition (21) (ii) fails also. By the backward Kolmogorov equation, one can easily figure out:

$$P_{ik}(\lambda) = 0,\ k < i; \qquad P_{ii}(\lambda) = \frac{1}{\lambda + q_i};$$
$$P_{ij}(\lambda) = \frac{q_i \cdots q_{j-1}}{(\lambda + q_i) \cdots (\lambda + q_j)}, \qquad j > i,$$

hence

$$\sum_{j=1}^{\infty} P_{1j}(\lambda) q_j = \sum_{j=1}^{\infty} \frac{q_i \cdots q_j}{(\lambda + q_i) \cdots (\lambda + q_j)} =: \sum_{j=1}^{\infty} a_j.$$

Because

$$\lim_{j \to \infty} j \left(\frac{a_j}{a_{j+1}} - 1 \right) = \lambda \lim_{j \to \infty} \frac{j}{a_{j+1}} = 0, \qquad \lambda > 0$$

one can see that the above series is divergent for each $\lambda > 0$.

(23) Remark. We point out here that the Theorem (16) is quite general. In some special case (for example, for generalized birth-death Q-processes), the conditions of (16) are also necessary.

Now we turn to discuss the relationship between tile regularities of a coupled process and its marginal processes. The next result was proved in Lemma (11).

(24) Theorem. If a coupled q-pair $(\tilde{q}(\tilde{x}), \tilde{q}(\tilde{x}, \cdot))$ satisfying (7) is regular, then its marginal q-pairs $(q_i(x_i), q_i(x_i, \cdot))$, $i = 1, 2$ are all regular.

Note that there are many choices of coupled q-pairs satisfying (7), also, the coupled q-pairs are usually more complicated than the given marginal q-pairs, it is certainly more interesting to prove that the regularities of the marginal q-pairs imply the one of a coupled q-pair. Unfortunately, we do not know at the moment how to prove it completely. What we can do now is to present the following result, which is an interesting application of Theorem (16) and quite general:

(25) Theorem. If the marginal q-pair $(q_i(x_i), q_i(x_i, \cdot))\,(i = 1, 2)$ satisfy the assumption of Theorem (16), then every coupled q-pair satisfying (7) is regular.

Proof. For $i = 1, 2$, we use $E_i^{(n)}$, φ_i and c_i to denote, respectively, the subsets, function and constant in the assumptions of Theorem (16), corresponding to the q-pair $(q_i(x_i), q_i(x_i, \cdot))$. Put

$$\widetilde{E}_n = E_1^{(n)} \times E_2^{(n)}, \qquad n \geqslant 1$$
$$\tilde{\varphi}(x_1, x_2) = \varphi_1(x_1) + \varphi_2(x_2), \qquad (x_1, x_2) \in \widetilde{E}.$$

Then $\{\widetilde{E}_n\}_1^{\infty} \subset \widetilde{\mathscr{E}}$ and $\widetilde{E}_n \uparrow \widetilde{E}$. By (7), one can see that

$$\tilde{q}(x_1, x_2) \leqslant q_1(x_1) + q_2(x_2), \qquad (x_1, x_2) \in \widetilde{E} \tag{26}$$

and so

$$\sup_{(x_1,x_2)\in\widetilde{E}_n} \tilde{q}(x_1,x_2) < \infty, \qquad n \geqslant 1.$$

On the other hand,

$$\lim_{n\to\infty} \inf_{(x_1,x_2)\notin\widetilde{E}_n} \tilde{\varphi}(\tilde{x}) \geqslant \left(\lim_{n\to\infty} \inf_{x_1\notin E_n^{(1)}} \varphi_1(x_1)\right) \bigwedge \left(\lim_{n\to\infty} \inf_{x_2\notin E_n^{(2)}} \varphi_2(x_2)\right) = \infty.$$

Finally, using the assumptions:

$$\int q_i(x_i, \mathrm{d}y_i)\varphi_i(y_i) \leqslant (c_i + q_i(x_i))\varphi_i(x_i), \qquad x_i \in E_i,\ i = 1,\ 2$$

and condition (7), it follows that

$$\int \tilde{q}(x_1,x_2;\mathrm{d}y_1,\mathrm{d}y_2)\tilde{\varphi}(y_1,y_2) \leqslant (c_1 \vee c_2 + \tilde{q}(x_1,x_2))\tilde{\varphi}(x_1,x_2), \qquad (x_1,x_2) \in \widetilde{E}.$$

Therefore the q-pair $(\tilde{q}(\tilde{x}), \tilde{q}(\tilde{x},\cdot))$ also satisfies the assumptions of Theorem (16). □

(27) Corollary. If the marginal q-pairs satisfy simultaneously one of the conditions of Theorem (21), then every coupled q-pair satisfying (7) is regular.

§4. Order-Preservation

In this section, we assume that $E_1 = E_2 = E$, $\mathscr{E}_1 = \mathscr{E}_2 = \mathscr{E}$, that E is endowed a semi–order "$\leqslant$", and the subset $\{(x,y) \in \widetilde{E} : x \leqslant y\} =: \widetilde{F}$ is $\widetilde{\mathscr{E}}$-measurable. We also assume that the coupled q-pair is regular.

We can rewrite the condition (4) as follows:

(28) Order-preservation.

$$\widetilde{P}(t,(x_1,x_2),\widetilde{F}) = 1, \qquad t \geqslant 0,\ (x_1,x_2) \in \widetilde{F}.$$

By differentiation, the above condition gives

$$\tilde{q}\big(x_1,x_2;\widetilde{F}^c\big) = 0, \qquad (x_1,x_2) \in \widetilde{F}. \tag{29}$$

Indeed, we have

(30) Theorem. (28)$\Longleftrightarrow$(29).

Proof. We have seen that (28)$\Longrightarrow$(29). Now assume that (29) holds. Note that

$$\widetilde{P}^{(0)}\big(\lambda,(x_1,x_2),\widetilde{F}^c\big) = \frac{\delta\big(x_1,x_2;\widetilde{F}^c\big)}{\lambda + \tilde{q}(x_1,x_2)} = 0, \qquad (x_1,x_2) \in \widetilde{F}.$$

Suppose

$$\widetilde{P}^{(n)}\big(\lambda,(x_1,x_2),\widetilde{F}^c\big) = 0, \qquad (x_1,x_2) \in \widetilde{F},$$

then, by (29), we get

$$\begin{aligned}&\widetilde{P}^{(n+1)}\big(\lambda,(x_1,x_2),\widetilde{F}^c\big)\\&=\int\frac{\tilde{q}(x_1,x_2;\mathrm{d}y_1,\mathrm{d}y_2)}{\lambda+\tilde{q}(x_1,x_2)}\widetilde{P}^{(n)}\big(\lambda,(y_1,y_2),\widetilde{F}^c\big)+\widetilde{P}^{(0)}\big(\lambda,(x_1,x_2),\widetilde{F}^c\big),\\&=\int_{\widetilde{F}}\frac{\tilde{q}(x_1,x_2;\mathrm{d}y_1,\mathrm{d}y_2)}{\lambda+\tilde{q}(x_1,x_2)}\widetilde{P}^{(n)}\big(\lambda,(y_1,y_2),\widetilde{F}^c\big)=0,\qquad(x_1,x_2)\in\widetilde{F}.\end{aligned}$$

Hence, by induction, it follows that

$$\widetilde{P}^{(n)}\big(\lambda,(x_1,x_2),\widetilde{F}^c\big)=0,\qquad(x_1,x_2)\in\widetilde{F},\ n\geqslant 1$$

and so

$$\widetilde{P}\big(\lambda,(x_1,x_2),\widetilde{F}^c\big)=0,\qquad(x_1,x_2)\in\widetilde{F},\ \lambda>0.$$

This finishes the proof. □

§5. Basic couplings

Let μ_1 and μ_2 be two finite measures on $(E,\mathscr{E})$. Denote by $(\mu_1-\mu_2)^+$ the Jordan-Hahn decomposition of $\mu_1-\mu_2$ and define

$$\mu_1\wedge\mu_2=\mu_1-(\mu_1-\mu_2)^+.$$

Clearly, $\mu_1\wedge\mu_2=\mu_2\wedge\mu_1$.

Let $(q_i(x_i),q_i(x_i,\cdot))$ be a given q-pair on $(E_i,\mathscr{E}_i)$, $i=1,2$. It often happens that

$$E_1\subset E_2\quad(\text{reap., }E_2\subset E_1).$$

and

$$E_1\in\mathscr{E}_2\quad(\text{reap., }E_2\in\mathscr{E}_1).$$

In this case, one can naturally extend the q-pair $(q_1(x_1),q_1(x_1,\cdot))$ to $(E_2,\mathscr{E}_2)$ simply by defining

$$q_1(x)=0,\qquad x\in E_2\setminus E_1.$$

Because of this reason, we may and will assume that

$$E_1=E_2=E,\qquad\mathscr{E}_1=\mathscr{E}_2=\mathscr{E}.$$

The simplest coupling is

(31) Independent Coupling.

$$\begin{aligned}\widetilde{\Omega}f(x_1,x_2)&=\int q_1(x_1,\mathrm{d}y_1)(f(y_1,x_2)-f(x_1,x_2))\\&=\int q_2(x_2,\mathrm{d}y_2)(f(x_1,y_2)-f(x_1,x_2))\\&=\big(\Omega_1f(\cdot,x_2)\big)(x_1)+\big(\Omega_2f(x_1,\cdot)\big)(x_2),\qquad(x_1,x_2)\in\widetilde{E},\ f\in{}_b\widetilde{E}.\end{aligned}$$

Perhaps the following coupling is the most useful one:

(32) Basic Coupling.

$$\begin{aligned}\widetilde{\Omega}f(x_1,x_2) = &\int \big(q_1(x_1,\cdot)-q_2(x_2,\cdot)\big)^+(\mathrm{d}y)[f(y,x_2)-f(x_1,x_2)]\\ &+\int \big(q_2(x_2,\cdot)-q_1(x_1,\cdot)\big)^+(\mathrm{d}y)[f(x_1,y)-f(x_1,x_2)]\\ &+\int \big(q_1(x_1,\cdot)\wedge q_2(x_2,\cdot)\big)(\mathrm{d}y)[f(y,y)-f(x_1,x_2)].\end{aligned}$$

For more examples of couplings, one can see [4] and [5].

It is not hard to check, for the basic coupling, that the order-preservation condition (29) becomes

(33) for each $(x_1,x_2)\in\widetilde{F}$,

$$\begin{aligned}&(q_1(x_1,\cdot)-q_2(x_2,\cdot))^+\big(\{y\in E: y\not\leqslant x_2\}\big)=0,\\ &(q_2(x_2,\cdot)-q_1(x_1,\cdot))^+\big(\{y\in E: x_1\not\leqslant y\}\big)=0.\end{aligned}$$

(34) Basic Coupling for q-Processes with Finite Product State Space.

Let S be a finite set. For each $u\in S$, let $(E_u,\mathscr{E}_u)$ be a measurable space as above. Suppose that $(q^\alpha(x),q^\alpha(x,\cdot))$ is a q-pair on $\big(\prod_{u\in S}E_u,\prod_{u\in S}\mathscr{E}_u\big)=:(E,\mathscr{E})$ satisfying that $q^\emptyset(x)=0$ for all $x\in E$ and the measure $q^\alpha(x,\cdot)$ is constrained on

$$\{y\in E: y_u\neq x_u,\ u\in\alpha;\ y_u=x_u,\ u\in S\setminus\alpha\}$$

for each $\alpha\subset S$. Now, set

$$q(x,\cdot)=\sum_{\alpha\in S}q^\alpha(x,\cdot),\qquad q(x)=q(x,E),\ x\in E.$$

Clearly, $(q(x),q(x,\cdot))$ is a q-pair on $(E,\mathscr{E})$. Corresponding to (32), we can define a coupling as follows:

$$\begin{aligned}\widetilde{\Omega}f(x_1,x_2) = &\sum_{\alpha\subset S}\big(q^\alpha(x_1,\cdot)-q^\alpha(x_2,\cdot)\big)^+(\mathrm{d}y_1)[f(y_1,x_2)-f(x_1,x_2)]\\ &+\sum_{\alpha\subset S}\big(q^\alpha(x_2,\cdot)-q^\alpha(x_1,\cdot)\big)^+(\mathrm{d}y_2)[f(x_1,y_2)-f(x_1,x_2)]\\ &+\sum_{\alpha\subset S}\big(q^\alpha(x_1,\cdot)\wedge q^\alpha(x_2,\cdot)\big)(\mathrm{d}y)[f(y,y)-f(x_1,x_2)],\\ &(x_1,x_2)\in\widetilde{E},\ f\in{}_b\widetilde{\mathscr{E}}.\end{aligned}\tag{35}$$

The basic coupling will play an important role in the subsequent paper [6].

In Addition. After the present paper was written, J. L. Zheng and X. G. Zheng proved that the regularity of the marginal q-pairs implies the one of their coupled q-pair for Markov chains under a slight assumption, by using martingale approach. Then the author and J. L. Zheng find a simple proof for general case. We present the proof in the following two theorems.

(36) Theorem. Given q-pair $(q(x), q(x,\cdot))$ and a sequence $\{E_n\} \subset \mathscr{E}$ such that

$$E_n \uparrow E, \qquad \sup_{x\in E_n} q_n(x) < \infty, \quad n \geqslant 1.$$

Define $(q_n(x), q_n(x,\cdot))$ by (20), Then $(q(x), q(x,\cdot))$ is regular iff

$$\lim_{n\to\infty} P_n\big(\lambda, x, E_n^c\big) = 0, \qquad \lambda > 0,\ x \in E.$$

Proof. The sufficiency follows from

$$P^{\min}\big(\lambda, x, E_n\big) \geqslant P_n\big(\lambda, x, E_n\big), \qquad \lambda > 0,\ x \in E,\ n \geqslant 1$$

which we have seen in the proof of Theorem (16). To prove the necessity, note that by the backward Kolmogorov equation, Fatou lemma and the comparison theorem, we have

$$\varliminf_{n\to\infty} P_n\big(\lambda, x, E_n\big) \geqslant P^{\min}\big(\lambda, x, E\big), \qquad \lambda > 0,\ x \in E.$$

Thus, if $(q(x), q(x,\cdot))$ is regular, then

$$\begin{aligned} 1 &\geqslant 1 - \lambda \varliminf_{n\to\infty} P_n\big(\lambda, x, E_n^c\big) \\ &= \lambda \varlimsup_{n\to\infty} P_n\big(\lambda, x, E_n\big) \\ &\geqslant \lambda P(\lambda, x, E) \\ &= 1. \end{aligned}$$

and so the condition is necessary. □

(37) Theorem. If the marginal q-pairs $(q_i(x_i), q_i(x_i,\cdot))$ $(i = 1, 2)$ are regular, then so is each coupled q-pair satisfying (7).

Proof. Take

$$\begin{aligned} E_i^{(n)} &= \{x_i \in E_i : q_i(x_i) \leqslant n\}, \qquad i = 1,\ 2,\ n \geqslant 1, \\ \widetilde{E}^{(n)} &= E_1^{(n)} \times E_2^{(n)}, \qquad n \geqslant 1. \end{aligned}$$

and define $\big(q_i^{(n)}(x_i), q_i^{(n)}(x_i,\cdot)\big)$, $i = 1, 2$ and $\big(\tilde{q}^{(n)}(x_i), \tilde{q}^{(n)}(x,\cdot)\big)$ by (20), respectively. Since

$$\sup_{\tilde{x}\in\widetilde{E}^{(n)}} \tilde{q}(x) \leqslant \sup_{x_1\in E_1^{(n)}} q_1(x_1) + \sup_{x_2\in E_2^{(n)}} q_2(x_1) < \infty$$

and Theorem (36), it suffices to show that

$$\widetilde{P}^{(n)}\big(\lambda; (x_1, x_2), (\widetilde{E}^{(n)})^c\big) \leqslant P_1^{(n)}\big(\lambda, x_1, (E_1^{(n)})^c\big) + P_2^{(n)}\big(\lambda, x_2, (E_2^{(n)})^c\big)$$
$$\lambda > 0,\ (x_1, x_2) \in \widetilde{E},\ n \geqslant 1.$$

where the q-processes are determined, respectively, by the above q-pairs. But this is an easy consequence of the condition (7) plus an application of the comparison theorem. □

Acknowledgments.

The author would like to thank Prof. S. J. Yan and Mr. J. L. Zheng for their helpful comments,

References

[1] Basis V. Ya., Infinite dimensional Markov processes with almost local interaction of components. *Theory Probability Appl.* 21 (1976), 727–740-(In Russian).

[2] Chen, M. F., Minimal nonnegative solution to an operator equation, *Beijing Shifandaxue Xuebao*, 3 (1979), 66–73 (In Chinese).

[3] Chen, M. F., Reversible Markov processes in abstract space, *Chinese Ann. Math.* 1 (1980), 437–451 (In Chinese).

[4] Chen, M. F., Basic couplings for Markov chains, *Beijing Shifandaxue Xuebao*, 4 (1984), 3–10 (In Chinese).

[5] Chen, M. F., Infinite dimensional reaction-diffusion processes. *Acta Mathematica Sinica, New Series*, 1:3 (1985), 261–273.

[6] Chen, M. F., Existence theorem of interacting panicle systems with non-compact state space. *Sci. Sinica (Series A)*, XXIX: 11(1986), 31-39.

[7] Chen, M. F., and Zheng, X. G., Uniqueness criterion for q-processes, *Sci., Sinica*, XXVI (I983), 11–24.

[8] Dobrushim R. L., Markov processes with a large number of locally interacting components, *Problems of Information Transmission*, 7 (1971), 149–164 and 235–241 (In Russian).

[9] Griffeath, D., Coupling methods for Markov processes, Studies in probability and Ergodic Theory, edited by G. C. Rota., *Adv. in Math., Supplementary Studies* 2 (1978).

[10] Waserstein, L. N., Markov processes on countable product spaces describing large systems of automata, *Problems of Information Transmission* 3 (1969), 64–72 (In Russian).

The Annals of Probability
1989, Vol. 17, No. 1, 151–177

COUPLING METHODS FOR MULTIDIMENSIONAL DIFFUSION PROCESSES

By MU-FA CHEN AND SHAO-FU LI

(Beijing Normal University and Henan Teachers' University)

ABSTRACT. In this paper, coupling methods for diffusion processes are studied mainly to obtain upper bound estimates in two different probability metrics. We use the martingale approach and explore the construction of explicit coupling operators which are sometimes optimal. The paper presents some criteria for the success of coupling and for the finiteness of the moments of the coupling times. Rates of convergence in various metrics are also studied.

1. Introduction. Coupling methods have been used widely in the study of interacting particle systems and other fields. There are some rather comprehensive treatments of coupling in the theory of Markov processes; see Griffeath (1978) and Liggett (1985). These papers contain a large number of references. In the diffusion context, we should mention Davies (1986), Lindvall (1983) and Lindvall and Rogers (1986). In the last paper, the authors obtained a successful coupling for a class of multidimensional diffusion processes by a reflection method and the theory of stochastic differential equations. Their method is effective for Brownian motion and for process in which the covariance matrix is almost constant. In particular, Brownian motion has a successful coupling. A geometric generalization of this Brownian coupling has been developed by Kendall (1986a, b) for use in stochastic differential geometry.

Let λ be a metric on $\mathbb{R}^d$. For $p \geqslant 1$, we define a probability metric W_p (often called Wasserstein or Kantorovich–Robinshtein–Wasserstein metric),

$$W_p(P_1, P_2) = \inf_Q \left[\int_{\mathbb{R}^d \times \mathbb{R}^d} \lambda(x, y)^p Q(dx, dy) \right]^{1/p},$$

Key words and phrases. Coupling, coupling operator, probability metric, coupling time, multidimensional diffusion, martingale approach.

Received July 1987; revised March 1988

AMS 1980 *subject classifications.* Primary 60J60, 60J65, 60H10; secondary 60J45, 60J70

where the infimum is taken over all measures Q on $\mathbb{R}^d \times \mathbb{R}^d$ such that for any measurable set $B \subset \mathbb{R}^d$,

$$\begin{aligned} Q(B \times \mathbb{R}^d) &= P_1(B), \\ Q(\mathbb{R}^d \times B) &= P_2(B). \end{aligned} \tag{1.1}$$

Any such Q is called a coupling of P_1 and P_2. Clearly, any coupling will give us an upper bound estimate for W_p. In this paper, we consider only the Euclidean metric on $\mathbb{R}^d$,

$$\rho(x, y) = |x - y| = \left[\sum_{i=1}^{d} (x_i - y_i)^2\right]^{1/2}$$

and the discrete metric

$$d(x, y) = \begin{cases} 0, & x = y \\ 1, & x \neq y. \end{cases}$$

In the latter case, we will use $V(P_1, P_2)$ to distinguish W_1 from the metric ρ. Note that W_p is an analogue of the L^p-metric. It was proved by Dobrushin (1970) that

$$V(P_1, P_2) = \sup_B |P_1(B) - P_2(B)|,$$

which is just half of the total variation norm.

It will become clear later that different couplings are suitable for different metric. For this reason, we may use the terms "W_p-coupling" and "V-coupling", respectively, for the different purposes. The W_p-couplings are often used in the study of interacting particle systems. [See Chen (1986b, 1987b) and the references there.] More recently, the W_2-couplings have also been used in the study of infinite dimensional diffusion processes by J. M. Xu and the first author.

Now, let us consider the V-coupling. Suppose that $\{X_t\}_{t\geqslant 0}$ and $\{Y_t\}_{t\geqslant 0}$ are diffusion processes in $\mathbb{R}^d$ with the same transition function $P(t, \cdot, \cdot)$ and distributions P^x and P^y, respectively. Let $P^{x,y}$ be a coupling probability measure on $\Omega_{2d} = C([0, \infty); \mathbb{R}^d)$. That is, the first and the second d-dimensional (marginal) distributions of $P^{x,y}$ are the same as P^x and P^y, respectively. Define the coupling time as

$$T = \inf\{t \geqslant 0 \colon X(t) = Y(t)\}.$$

If

$$T < \infty, \quad P^{x,y}\text{-a.s.}$$

and

$$P^{x,y}[X(t) = Y(t);\ t \geqslant T] = 1,$$

we call the coupling $P^{x,y}$ successful. Furthermore, if

$$P^{x,y}[T > t] = o(t^{-\alpha}) \quad \text{as} \quad t \to \infty,$$

for some $\alpha > 0$, then we have

$$V(P(t,x,\cdot),P(t,y,\cdot)) = o(t^{-\alpha}).$$

Thus the key point is to construct a successful coupling $P^{x,y}$ of P^x and P^y.

As we did in the case of jump processes [Chen (1986a, 1987a)], we begin our study with the analysis of coupling operators.

Let

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

be an elliptic operator on $\mathbb{R}^d$ (possibly degenerate). Assume that the solution to the martingale problem for L is well-posed [see Stroock and Varadhan (1979)]. Our goal is to find an elliptic operator on $\mathbb{R}^{2d}$ such that the solution to the martingale problem for this operator has the marginal property (1.1). From the infinitesimal character of diffusion processes and (1.1), it is obvious that the coefficients of the operator should be of the form

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

where c^* is the transpose of c. A trivial example is $c \equiv 0$. In this case, the two coordinates are independent and the coupling is usually not useful. But this means that a coupling operator always exists. Indeed, there are a lot of choices.

EXAMPLE 1.2 (d-dimensional Brownian motion). Take

$$\begin{aligned} c_1(x,y) &= I - 2(x-y)(x-y)^*/|x-y|^2, \\ c_2(x,y) &= I - (x-y)(x-y)^*/|x-y|^2, \\ c_3(x,y) &= \left(\left(1 - \frac{\alpha|x_i - y_i|}{\beta + |x_i - y_i|}\right)\delta_{ij}\right), \end{aligned}$$

where $\alpha \in (0,2]$ and $\beta > 0$. All these couplings are successful (see Section 4). The first one was given in Lindvall and Rogers (1986), called coupling by reflection. Actually, if we denote by $L_{xy}\,(x \neq y)$ the hyperplane $\{z \in \mathbb{R}^d : (z, x-y) = 0\}$ which is just the orthogonal complement of $\{x-y\}$, then for each $z \in \mathbb{R}^d$, $c_1(x,y)z$ is the reflection image of z with respect to the hyperplane L_{xy}. On the other hand, $c_2(x,y)z$ is the projection of z onto the subspace L_{xy}. Hence we call the second one "coupling by projection". The last one has an advantage in that the couplings for different components are independent.

The paper is organized as follows: In the next section, we discuss the W_p-couplings ($p = 1, 2$). The remainder of the paper is devoted to the V-coupling which are much more complicated. In Section 3 we study the constructions of the couplings. In Section 4 we present some criteria for the success of couplings and a large number of examples to illustrate these criteria. Our criteria are exact in some cases. In Section 5 we study the rates of convergence of $P^{x,y}[T > t]$ as $t \to \infty$ for successful couplings. The moment of the coupling time T is also studied there.

2. Coupling for W_p-metric $(p = 1, 2)$**.** Let $\Omega = \Omega_{2d} = C([0,\infty); \mathbb{R}^{2d})$ be the space of continuous trajectories from $[0,\infty)$ into $\mathbb{R}^{2d}$. Given $t \geqslant 0$ and $\omega \in \Omega$, let $Z(t,\omega) = Z_t(\omega)$ denote the position of ω in $\mathbb{R}^{2d}$. Define

$$\mathscr{M}_t = \sigma\{Z_s : s \leqslant t\}, \qquad \mathscr{M} = \sigma\Big(\bigcup\nolimits_{t\geqslant 0} \mathscr{M}_t\Big).$$

Let

$$X(t,\omega) = \pi_1 \circ Z(t,\omega) = (\omega_1(t), \cdots, \omega_d(t)),$$
$$Y(t,\omega) = \pi_2 \circ Z(t,\omega) = (\omega_{d+1}, \cdots, \omega_{2d}(t)).$$

That is, $Z(t,\omega) = (X(t,\omega), Y(t,\omega))$. Similarly, we can define $\mathscr{M}_t^{(1)}$, $\mathscr{M}^{(1)}$ and $\mathscr{M}_t^{(2)}$, $\mathscr{M}^{(2)}$. For example,

$$\mathscr{M}_t^{(1)} = \sigma\{X_s : s \leqslant t\}.$$

We often denote the operator

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

by $L(a(x), b(x))$. Now suppose that $L_1(a_1(x), b_1(x))$ and $L_2(a_2(y), b_2(y))$ are given operators, then we can define an operator $L(a(x,y), b(x,y))$ on $\mathbb{R}^{2d}$ as

$$a(x,y) = \begin{pmatrix} a_1(x) & c(x,y) \\ c(x,y)^* & a_2(y) \end{pmatrix}, \qquad b(x,y) = \begin{pmatrix} b_1(x) \\ b_2(y) \end{pmatrix},$$

where $c(x,y)$ is a real valued $d \times d$ matrix such that the matrix $a(x,y)$ is nonnegative definite. Such an operator $L(a,b)$ is called a coupling of L_1 and L_2.

Throughout this paper, the coefficients of all operators are assumed to be locally bounded. Moreover, we assume that the martingale problem for the marginal diffusion processes are well posed. The solutions are denoted by

$$P_1^x \sim L_1, \qquad x \in \mathbb{R}^d;$$
$$P_2^y \sim L_2, \qquad y \in \mathbb{R}^d.$$

LEMMA 2.1. Let $\{P^{x,y} : x, y \in \mathbb{R}^d\}$ be a family of solutions to the martingale problem for the coupling operator $L(a(x,y), b(x,y))$, denoted by $P^{x,y} \sim L(a,b)$. Then

$$P_1^x = P^{x,y} \circ \pi_1^{-1}, \qquad P_2^y = P^{x,y} \circ \pi_2^{-1}, \qquad x, y \in \mathbb{R}^d.$$

In other words, $P^{x,y}$ is a coupling of P_1^x and P_2^y for every $x, y \in \mathbb{R}^d$.

PROOF. Let $f \in C_0^\infty(\mathbb{R}^d)$ and set $F(x,y) = f(x),\ x,y \in \mathbb{R}^d$. Note that $LF(x,y) = L_1 f(x)$ and the operators are locally bounded. We have, for every set $B \in \mathscr{M}_s^{(1)}$ and $s \leqslant t$, that $\pi_1^{-1}B \in \mathscr{M}_s$ and

$$\begin{aligned}
&\int_B \left(f(X_t) - \int_0^t L_1 f(X_u)du\right) d(P^{x,y} \circ \pi_1^{-1}) \\
&\quad = \int_{\pi_1^{-1}B} \left(F(Z_t) - \int_0^t LF(Z_u)du\right) dP^{x,y} \\
&\quad = \int_{\pi_1^{-1}B} \left(F(Z_s) - \int_0^s LF(Z_u)du\right) dP^{x,y} \\
&\quad = \int_B \left(f(X_s) - \int_0^s L_1 f(X_u)du\right) d(P^{x,y} \circ \pi_1^{-1}).
\end{aligned}$$

This shows that $P^{x,y} \circ \pi_1^{-1} \sim L_1$. By the uniqueness assumption, we certainly get

$$P_1^x = P^{x,y} \circ \pi_1^{-1}, \qquad x,y \in \mathbb{R}^d.$$

Similarly, we have the second equality. □

We need the following elementary result.

LEMMA 2.2. Let $V(t)$ be a differentiable function and $B(t)$ be a locally integrable function on $[0,\infty)$. If

$$\frac{d}{dt}V(t) \leqslant -cV(t) + B(t), \qquad \text{a.e. } t$$

for some $c > 0$, then

$$V(t) \leqslant V(0)e^{-ct} + \int_0^t e^{-c(t-s)}B(s)ds, \qquad t \geqslant 0.$$

THEOREM 2.3. Suppose that $a(x,y)$ and $b(x,y)$ are continuous on $\mathbb{R}^{2d}$ and $P^{x,y} \sim L(a,b)$. If there exist constants $C \geqslant 0$ and $c > 0$ such that

$$L\rho^2(x,y) \leqslant C - c\rho^2(x,y), \qquad x,y \in \mathbb{R}^d,$$

then

$$E^{x,y}\rho^2(X_t,Y_t) \leqslant C/c + e^{-ct}\rho^2(x,y), \qquad x,y \in \mathbb{R}^d.$$

In particular, if $C = 0$, then

$$W_2(P_1(t,x,\cdot), P_2(t,y,\dot{)}) \leqslant \rho(x,y)e^{-ct/2} \to 0 \quad \text{as } t \to \infty,$$

where $P_1(t,x,\cdot)$ and $P_2(t,y,\cdot)$ are, respectively, the transition functions of the marginal diffusions. The same conclusion is true if we replace ρ^2, W_2 and $e^{-ct/2}$ by ρ, W_1 and e^{-ct}, respectively.

PROOF. Set

$$\begin{aligned} S_N &= \inf\{t \geqslant 0 : |X_t - Y_t| > N\}, \\ T_R &= \inf\{t \geqslant 0 : |X_t|^2 + |Y_t|^2 > R\}, \\ S &= S_N \wedge T_R. \end{aligned}$$

Since $P^{x,y} \sim L$, we have

$$E^{x,y}\rho^2(X_{t\wedge S}, Y_{t\wedge S}) = \rho^2(x,y) + \int_0^t E^{x,y} L\rho^2(X_{u\wedge S}, Y_{u\wedge S})du$$

and so

$$\begin{aligned} \frac{d}{dt}E^{x,y}\rho^2(X_{t\wedge S}, Y_{t\wedge S}) &= E^{x,y} L\rho^2(X_{t\wedge S}, Y_{t\wedge S}) \\ &\leqslant C - cE^{x,y}\rho^2(X_{t\wedge S}, Y_{t\wedge S}). \end{aligned}$$

By Lemma 2.2 we obtain

$$E^{x,y}\rho^2(X_{t\wedge S}, Y_{t\wedge S}) \leqslant C/c + \rho^2(x,y)e^{-ct}.$$

Now the conclusion follows by passing the limit $R \uparrow \infty$, $N \uparrow \infty$. □

DEFINITION 2.4. Let $a_1(x) = \sigma_1(x)\sigma_1(x)^*$, $a_2(y) = \sigma_2(y)\sigma_2(y)^*$. We call

$$a(x,y) = \begin{pmatrix} a_1(x) & \sigma_1(x)\sigma_2(y)^* \\ \sigma_2(y)\sigma_1(x)^* & a_2(y) \end{pmatrix}, \qquad b(x,y) = \begin{pmatrix} b_1(x) \\ b_2(y) \end{pmatrix}$$

basic coupling of L_1 and L_2.

EXAMPLE 2.5 [Ornstein-Uhlenbeck (O.U.) process].

$$\sigma_1(x) = \sigma_2(x) = I, \qquad b_1(x) = b_2(x) = -\mu x.$$

Using the basic coupling, we obtain from Theorem 2.3

$$\begin{aligned} &W_2(P(t,x,\cdot), P(t,y,\cdot)) \leqslant (E^{x,y}\rho^2(X_t, Y_t))^{1/2} = e^{-\mu t}\rho(x,y), \\ &W_1(P(t,x,\cdot), P(t,y,\cdot)) \leqslant E^{x,y}\rho(X_t, Y_t) = e^{-\mu t}\rho(x,y), \qquad t \geqslant 0,\ x,y \in \mathbb{R}^d. \end{aligned}$$

EXAMPLE 2.6. Take $\sigma_1(x) = \sigma_2(x) = \sigma = \text{constant}$, $b_1(x) = b_2(x) = 0$. Using the basic coupling, we obtain from Theorem 2.3

$$\begin{aligned} &W_2(P(t,x,\cdot), P(t,y,\cdot)) \leqslant (E^{x,y}\rho^2(X_t, Y_t))^{1/2} = |x-y|, \\ &W_1(P(t,x,\cdot), P(t,y,\cdot)) \leqslant E^{x,y}\rho(X_t, Y_t) = |x-y|. \end{aligned}$$

On the other hand, it is known from Givens and Shortt (1984) that the first inequality is an equality in any dimension. Thus, our basic coupling is exact in

this case. For W_1, the coupling is not exact, but as you will see soon it is the best that we can do.

We now introduce some notation which will be used often later.

NOTATION 2.7. Denote by $\langle , \rangle$ the ordinary inner product in $\mathbb{R}^d$. Set

$$\begin{aligned}
A(x,y) &= a_1(x) + a_2(y) - 2c(x,y),\\
B(x,y) &= b_1(x) - b_2(y),\\
\hat{A}(x,y) &= \langle x-y,\ A(x,y)(x-y)\rangle,\\
\overline{A}(x,y) &= \hat{A}(x,y)/|x-y|^2, \qquad x \neq y,\\
\hat{B}(x,y) &= \langle x-y,\ B(x,y)\rangle.
\end{aligned}$$

It is easy to check that $\hat{A}(x,y) \geqslant 0$ for all $x, y \in \mathbb{R}^d$ [since $a(x,y)$ is nonnegative definite] and that for each $f \in C^2([0,\infty))$, we have

$$\begin{aligned}
2Lf(\rho(x,y)) =& \overline{A}(x,y) f''(\rho(x,y))\\
&+ \big[\operatorname{tr} A(x,y) - \overline{A}(x,y) + 2\hat{B}(x,y)\big] \frac{f'(\rho(x,y))}{\rho(x,y)}.
\end{aligned} \tag{2.8}$$

In particular, we have

$$L\rho^2(x,y) = \operatorname{tr} A(x,y) + 2\hat{B}(x,y) \tag{2.9}$$

and

$$L\rho(x,y) = \frac{1}{2\rho(x,y)}\big[\operatorname{tr} A(x,y) - \overline{A}(x,y) + 2\hat{B}(x,y)\big]. \tag{2.10}$$

Now, we turn to discuss how to choose the coupling operators for $W_2(W_1)$-coupling. For simplicity, we consider only the case that $b_1(x) = b_2(y) = 0$.

REMARK 2.11. In view of Theorem 2.3, (2.9) and (2.10), we may say that a coupling operator $a(x,y)$ is W_2- (respectively, W_1-) optimal if $a(x,y)$ is nonnegative definite and $\operatorname{tr} A(x,y)$ [respectively, $\operatorname{tr} A(x,y) - A(x,y)$] achieves the minimum at each point $(x,y) \in \mathbb{R}^{2d}$ [note that these quantities contain $c(x,y)$ which varies]. Clearly, if $\sigma_1(x) = \sigma_2(y) = \sigma =$ constant, then the basic coupling gives us $\operatorname{tr} A = \overline{A} = 0$ and so is optimal. For the general case, let us fix x and y, assume that $a_1(x)$ and $a_2(y)$ are positive definite and take $\sigma_1(x) = \sqrt{a_1(x)}$, $\sigma_2(y) = \sqrt{a_2(y)}$, the positive definite square roots. In this case, we can rewrite $c(x,y)$ as $\sigma_1(x)H^*(x,y)\sigma_2(y)$. Now, $a(x,y)$ is nonnegative definite if and only if H is contractive. That is, $|Hx| \leqslant |x|$ for all $x \in \mathbb{R}^d$. Using the Hilbert-Schmidt (H.S.) norm for metrics, we can easily prove that the optimal choice of $H(x,y)$ does exist since the domain of H is compact and $\operatorname{tr} A$ (respectively, $\operatorname{tr} A - \overline{A}$) is continuous in H with respect to the H.S. norm. But this optimalization problem is generally

quite difficult. For simplicity, now we restrict ourselves to the case that H is an orthogonal matrix. Then, for W_2, the solution is

$$H(x,y) = [\sigma_2(y)a_1(x)\sigma_2(y)]^{-1/2}\sigma_2(y)\sigma_1(x). \tag{2.12}$$

Then, we have

$$\operatorname{tr} A(x,y) = \operatorname{tr}\left[a_1(x) - a_2(y) - 2(\sigma_2(y)a_1(x)\sigma_2(y))^{1/2}\right].$$

(In this case, even without the orthogonal assumption on H, the optimal solution is still the same. For details, see Givens and Shortt [(1984), pages 237–239].) Furthermore, if $\sigma_1 = \sigma_2 = \sigma$ is diagonal, then we have $H(x,y) = (\sigma_{ii}(x)\sigma_{ii}(y)\delta_{ij})$. For W_1, the optimal solution H should satisfy

$$H\sigma_1(I - \bar{u}\bar{u}^*)\sigma_2 = \left[\sigma_2(I - \bar{u}\bar{u}^*)a_1(I - \bar{u}\bar{u}^*)\sigma_2\right]^{1/2}. \tag{2.13}$$

where $\bar{u} = (x-y)/|x-y|$, $H = H(x,y)$, $\sigma_1 = \sigma_1(x)$, $\sigma_2 = \sigma_2(y)$ and so on. This is quite complicated but still useful in some cases. We will return to this formula later.

A typical application of the above coupling is as follows: Suppose that our diffusion with L_1 has a stationary distribution π, the conditions of Theorem 2.3 are satisfied with $C = 0$ for a coupling of L_1 and itself, and $(x,y) \to P^{x,y}$ measurable, then we have

$$W_2(P(t,x,\cdot),\pi) \leqslant e^{-ct/2}\left[\int \pi(dy)|x-y|^2\right]^{1/2}.$$

In fact,

$$\begin{aligned} W_2(P(t,x,\cdot),\pi) &= W_2\left(P(t,x,\cdot),\ \int \pi(dy)P(t,y,\cdot)\right) \\ &\leqslant \left[\int \pi(dy)E^{x,y}|X_t - Y_t|^2\right]^{1/2} \\ &\leqslant e^{-ct/2}\left[\int \pi(dy)|x-y|^2\right]^{1/2}, \end{aligned} \tag{2.14}$$

and similarly for W_1. As for the existence of stationary distribution for diffusions, see Bhattacharya and Ramasubramanian (1982) and their references. Here, we state a simple sufficient condition. The proof is nontrivial but almost the same as the ones in Basis (1980) and Chen (1986b) which go back to Dobrushin (1970). Hence we omit the proof.

THEOREM 2.15. Let $h \in C^2(\mathbb{R}^d)$ be a compact function, i.e., $h \geqslant 0$, $\{x : h(x) \leqslant k\}$ is a compact set for each $k \geqslant 0$. If there are constants $C \geqslant 0$ and $c > 0$ such that

$$L_1 h(x) \leqslant C - ch(x), \qquad x \in \mathbb{R}^d, \tag{2.16}$$

then the diffusion process determined by L_1 has a stationary distribution π with

$$\int \pi(dx)h(x) \leqslant C/(1-c). \tag{2.17}$$

Now, if (2.16) holds with $h = \rho^2$, then combining (2.14) and (2.17), we obtain

$$W_2(P(t,x,\cdot),\pi) \leqslant \text{const.}(1+|x|)e^{-ct/2} \to 0 \quad \text{as } t \to \infty. \tag{2.18}$$

3. Constructions of couplings for V-metric. Starting with this section, we discuss the couplings for V-metric. The following result describes a fundamental property of our basic coupling.

THEOREM 3.1. Let $a_1 = a_2 = \sigma\sigma^*$, $b_1 = b_2 = b$, and σ and b be continuous on $\mathbb{R}^d$. Suppose that for the basic coupling $L(a(x,y), b(x,y))$:

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

the martingale problem is locally well-posed [and hence globally well-posed by Stroock and Varadhan (1979), Corollary 10.1.2]. If we denote the solution by

$$P_{x,y} \sim L(a(x,y), b(x,y)),$$

then we have

$$X_t = Y_t, \qquad t \geqslant T, \qquad P^{x,y}\text{-a.s. on } [T < \infty]. \tag{3.2}$$

PROOF. A similar result was given in Stroock and Varadhan [(1979), Lemma 8.1.3]. Here we present a different proof. By a modification of Theorem 6.1.3 in Stroock and Varadhan (1979) (we allow $T = \infty$), what we need is to show that

$$P^{x,y}[X_t = Y_t,\ t \geqslant 0] = 1, \qquad x \in \mathbb{R}^d.$$

Next, by a truncating argument, we may assume that $a(x,y)$ and $b(x,y)$ are bounded and continuous, so the martingale problem for L is well-posed.

Now take

$$\eta(x) = \begin{cases} C\exp[-1/(1-|x|^2)], & |x| < 1,\ x \in \mathbb{R}^d, \\ 0, & |x| \geqslant 1,\ x \in \mathbb{R}^d, \end{cases}$$

where C is the normalizing constant. Put $\eta_\varepsilon(x) = \varepsilon^{-d}\eta(x/\varepsilon)$, set $\sigma_\varepsilon(x) = (\sigma^\varepsilon_{ij}(x)) : \sigma^\varepsilon_{ij}(x) = (\sigma_{ij} * \eta)(x)$, $b_\varepsilon(x) = (b^\varepsilon_i(x)) : b^\varepsilon_i(x) = (b_i * \eta)(x)$ and so on. Clearly $\sigma^\varepsilon_{ij}(x), b^\varepsilon_i(x) \in C^\infty_b(\mathbb{R}^d)$. Hence there exist constants A_ε and B_ε such that

$$\begin{aligned} &\|\sigma_\varepsilon(x) - \sigma_\varepsilon(y)\| \leqslant A_\varepsilon|x-y|, \\ &|b_\varepsilon(x) - b_\varepsilon(y)| \leqslant B_\varepsilon|x-y|. \end{aligned}$$

On the other hand, corresponding to the function $\rho(x,y) = |x-y|$, we can construct a sequence of functions $\{\varphi_n\}_1^\infty$ such that

$$\varphi_n \in C^2(\mathbb{R}), \quad \varphi_n(x) \uparrow |x|, \qquad |\varphi_n'| \leqslant 1, \quad 0 \leqslant \varphi_n''(x) \leqslant 2/(nx^2).$$

[See Ikeda and Watanabe (1981), pages 168–169.] Now, let $L_\epsilon(\sigma_\varepsilon^*, b_\varepsilon)$ be the basic coupling. Then by (2.8), we have

$$\begin{aligned} 2L_\varepsilon\varphi_n(|x-y|) &= \varphi_n''(|x-y|)\overline{A}_\varepsilon(x,y) \\ &\quad + \frac{\varphi_n'(|x-y|)}{|x-y|}\big(\operatorname{tr} A_\varepsilon(x,y) - \overline{A}_\varepsilon(x,y) + 2\widehat{B}_\varepsilon(x,y)\big), \end{aligned}$$

where

$$\begin{aligned} \overline{A}_\varepsilon(x,y) &= \left|\big(\sigma_\varepsilon(x)-\sigma_\varepsilon(x)\big)^* \frac{x-y}{|x-y|}\right|^2 \leqslant A_\varepsilon^2|x-y|^2, \\ \operatorname{tr} A_\varepsilon(x,y) &\leqslant \|\sigma_\varepsilon(x)-\sigma_\varepsilon(y)\|^2 \leqslant A_\varepsilon^2|x-y|^2, \\ \big|\widehat{B}_\varepsilon(x,y)\big| &= \big|\langle x-y, b_\varepsilon(x)-b_\varepsilon(y)\rangle\big| \leqslant B_\varepsilon|x-y|^2 \end{aligned}$$

and so

$$2L_\varepsilon\varphi_n(|x-y|) \leqslant A_\varepsilon^2\varphi_n''(|x-y|)\,|x-y|^2 + |\varphi_n'(|x-y|)|\,\big(2A_\varepsilon^2 + 2B_\varepsilon\big)|x-y|.$$

Let $P_\varepsilon^{x,x} \sim L_\varepsilon(\sigma_\varepsilon\sigma_\varepsilon^*, b_\varepsilon)$ and

$$S_N = \inf(t \geqslant 0 : |X_t - Y_t] > N).$$

Then

$$\begin{aligned} &E_\varepsilon^{x,x}\varphi_n\big(\big|X_{t\wedge S_N} - X_{t\wedge S_N}\big|\big) \\ &\quad\leqslant E_\varepsilon^{x,x}\int_0^{t\wedge S_N}\Big\{\frac{1}{2}A_\varepsilon^2\varphi_n''\rho^2 + |\varphi_n'|\,\big(A_\varepsilon^2 + B_\varepsilon\big)\rho\Big\}(X_u, Y_u)du \\ &\quad\leqslant A_\varepsilon^2 t/n + \big(A_\varepsilon^2 + B_\varepsilon\big)\,E_\varepsilon^{x,x}\int_0^{t\wedge S_N}|X_u - Y_u|du. \end{aligned}$$

Let $N \uparrow \infty$ and then $n \uparrow \infty$. We obtain

$$E_\varepsilon^{x,x}|X_t - Y_t| \leqslant (A_\varepsilon^2 + B_\varepsilon)\int_0^t E_\varepsilon^{x,x}|X_u - Y_u|du$$

and so

$$E_\varepsilon^{x,x}|X_t - Y_t| = 0, \qquad t \geqslant 0.$$

This shows that

$$P_\varepsilon^{x,x}(X_t = Y_t, t \geqslant 0) = 1.$$

Finally, by Stroock and Varadhan (1979), Theorem 1.4.6, it is easy to prove that $(P_\varepsilon^{x,x} : \varepsilon > 0)$ is tight, and so we can choose a subsequence $\{\varepsilon_m\}_1^\infty$ such that

$$P_{\varepsilon_m}^{x,x} \longrightarrow P_0^{x,x} \quad \text{weakly.}$$

Then $P_0^{x,x} \sim L(a(x,y), b(x,y))$. By the locally well-posed assumption, we indeed have $P_0^{x,x} = P^{x,x}$. Therefore

$$P^{x,x}[X_t = Y_t] \geqslant \limsup_{m\to\infty} P_{\varepsilon_m}^{x,x}[X_t = Y_t] = 1, \qquad t \geqslant 0.$$

This proves our assertion. □

In order to give different types of couplings, we need more preparation. Denote by $\widetilde{P}^{x,y}$ the solution to the martingale problem for the basic coupling constructed by Theorem 3.1. Set

$$Q_\omega = \delta_\omega \bigotimes_{T(\omega)} \widetilde{P}^{X(T(\omega)),Y(T(\omega))} I_{[T(\omega)<\infty]} + \delta_\omega I_{[T(\omega)=\infty]}, \qquad \omega \in \Omega. \tag{3.3}$$

LEMMA 3.4 Under the hypotheses of Theorem 3.1, if $P^{x,y}$ is a solution to the martingale problem for

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

up to time T, then

$$R = P^{x,y} \bigotimes_T Q$$

is a solution to the martingale problem for

(3.5)

$$a(t,x,y) = \begin{pmatrix} \sigma(x)\sigma(x)^* & I_{[0,T)}(t)c(x,y) \\ & \quad + I_{[T,\infty)}(t)\sigma(x)\sigma(y)^* \\ I_{[0,T)}(t)c(x,y)^* & \\ \quad + I_{[T,\infty)}(t)\sigma(y)\sigma(x)^* & \sigma(y)\sigma(y)^* \end{pmatrix}$$

$$b(t,x,y) = \begin{pmatrix} b(x) \\ b(y) \end{pmatrix}.$$

PROOF. Cf. Stroock and Varadhan [(1979), Section 6.1] for details. □

EXAMPLE 3.6 (Classical coupling). In (3.5), take $c(x,y) = 0$. This means that the processes start from two different points, run independently until they first meet each other, then move together.

EXAMPLE 3.7 (Coupling by reflection). Take

$$c(x,y) = \sigma(x)(I - 2\bar{u}\bar{u}^*)\sigma(y)^*,$$

where $\bar{u} = (x-y)/|x-y|$. If σ is constant and $\det\sigma \neq 0$, we can also take

$$c(x,y) = \sigma\sigma^* - 2\bar{u}\bar{u}^*/|\sigma^{-1}\bar{u}|^2.$$

EXAMPLE 3.8 (Coupling by projection). Take

$$c(x,y) = \sigma(x)(I - \bar{u}\bar{u}^*)\sigma(y)^*.$$

For the above examples, we can first construct the couplings up to time T, then applying Lemma 3.4, link them with the basic coupling so that after time T, they will move together. Sometimes, we have to do so (cf. Section 4). However, it is not always the case. Very often, it is enough to construct a coupling up to the time T. This also enables us to consider the more general case that $L_1 \neq L_2$. Such generalization is useful in some cases [see Chen (1986b), for example].

4. Criteria for success. In this and the next sections, we fix a coupling operator $L(a(x,y), b(x,y))$:

$$a(x,y) = \begin{bmatrix} a_1(x) & c(x,y) \\ c(x,y)^* & a_2(y) \end{bmatrix}, \qquad b(x,y) = \begin{bmatrix} b_1(x) \\ b_2(y) \end{bmatrix},$$

and assume that $P^{x,y}$ $(x \neq y)$ is a solution to the martingale problem for L up to time T:

$$T = \inf\{t \geqslant 0 : X_t = Y_t\}.$$

In other words, for each pair $x \neq y$ and for every $f \in C_0^2(\mathbb{R}^{2d})$: $\operatorname{supp}(f) \subset \{(x,y) \in \mathbb{R}^{2d} : 1/n \leqslant |x-y| \leqslant N\}$ for some n, $N > 1$,

$$f(X_t, Y_t) - \int_0^t Lf(X_u, Y_u)du$$

is a $P^{x,y}$-martingale with respect to $\{\mathscr{M}_t\}_{t\geqslant 0}$.

The idea of our criteria discussed below is to compare the process $Z_t = (X_t, Y_t)$ with the radial process $r_t = |X_t - Y_t|$. To do this, set

$$\begin{aligned} S_N &= \inf(t \geqslant 0 : |X_t - Y_t| > N), \qquad N > 1, \\ T_n &= \inf(t \geqslant 0 : |X_t - Y_t| < 1/n), \qquad n > 1,\ T_n \uparrow T \text{ as } n \uparrow \infty, \\ T_{n,N} &= T_n \wedge S_N. \end{aligned}$$

Choose continuous functions γ and γ^*: $(0,\infty) \longrightarrow \mathbb{R}$ such that

$$\begin{aligned} \gamma(r) &\geqslant \sup_{|x-y|=r} \big(\operatorname{tr} A(x,y) - \overline{A}(x,y) + 2\widehat{B}(x,y)\big)/\overline{A}(x,y), \\ \gamma_*(r) &\leqslant \inf_{|x-y|=r} \big(\operatorname{tr} A(x,y) - \overline{A}(x,y) + 2\widehat{B}(x,y)\big)/\overline{A}(x,y) \end{aligned}$$

and define

$$\begin{aligned} C(r) &= \exp\left[\int_1^r \frac{\gamma(u)}{u}du\right], \qquad C_*(r) = \exp\left[\int_1^r \frac{\gamma_*(u)}{u}du\right], \\ f(r) &= \int_1^r C(s)^{-1}ds, \qquad f_*(r) = \int_1^r C_*(s)^{-1}ds, \qquad r > 0. \end{aligned}$$

Next, choose continuous functions α and α^*: $(0,\infty) \longrightarrow [0,\infty)$ such that[1]

$$\alpha(r) \leqslant \inf_{|x-y|=r} \overline{A}(x,y) \leqslant \sup_{|x-y|=r} \overline{A}(x,y) \leqslant \alpha^*(r)$$

[1]Addition to the original paper: A slight different way, which is not comparable with the original one, to define γ and γ_* goes as follow. Remove $/\overline{A}(x,y)$ in the definition of γ (resp. γ_*) and then replace in what follows the original γ (resp. γ_*) by γ/α (resp. γ_*/α^*). All the results remain the same except some computations of γ (resp. γ_*) in examples have to be modified correspondingly. See also the footnote to Example 4.14.

Then, we have

$$\begin{aligned} &f'(r) > 0, \qquad f''(r) + f'(r)\gamma(r)/r = 0, \\ &f'_*(r) > 0, \qquad f''_*(r) + f'_*(r)\gamma_*(r)/r = 0. \end{aligned} \tag{4.1}$$

Define

$$\begin{aligned} g(r) &= \int_r^1 C(s)^{-1} ds \int_s^1 \frac{C(u)}{\alpha(u)} du, \\ g_*(r) &= \int_r^1 C_*(s)^{-1} ds \int_s^1 \frac{C_*(u)}{\alpha^*(u)} du, \end{aligned}$$

as $r \uparrow \infty$, $f(r) \uparrow f(\infty)$, say. Similarly, we can define $f(0)$, $f_*(\infty)$, $f_*(0)$, $g(0)$ and $g_*(0)$.

THEOREM 4.2. Let $\alpha > 0$ on $(0, \infty)$.

(i) If $f(\infty) = \infty$ and $g(0) < \infty$, then the coupling is successful.

(ii) If $f_*(\infty) < \infty$ or $g_*(0) = \infty$, then the coupling is not successful.

(iii) If $\gamma = \gamma_*$ and $\alpha = \alpha*$, then the coupling is successful if and only if $f(\infty) = \infty$ and $g(0) < \infty$.

COROLLARY 4.3. (i) If α is bounded below by a positive number, $f(\infty) = \infty$, $f(0) > -\infty$ and $\liminf_{r \downarrow 0} f'(r) > 0$, then the coupling is successful.

(ii) If $\alpha > 0$ on $(0, \infty)$, $f_*(\infty) < \infty$ or $f_*(0) = -\infty$, then the coupling is not successful.

PROOF. In case (i), it is easy to check that $g(0) < \infty$. As for case (ii), it suffices to note that $f_*(0) = -\infty \Longrightarrow g_*(0) = \infty$. Thus, the corollary follows from Theorem 4.2 directly. □

Case (i) of Corollary 4.3 was obtained by Lindvall and Rogers [(1986), Lemma 1].

PROOF OF THEOREM 4.2. For the sake of completeness and also for certain subsequent uses, we sketch the proof here though the technique is essentially not new [cf. Friedman (1975)].

Set

$$F_{n,N}(\rho) = -\int_{1/n}^{\rho} C(s)^{-1} ds \int_s^N \frac{C(u)}{\alpha(u)} du, \qquad \frac{1}{n} \leqslant \rho \leqslant N,\ n, N > 1. \tag{4.4}$$

Then

$$\begin{aligned} &-\infty < F_{n,N}(\rho) < 0, \qquad F'_{n,N}(\rho) \leqslant 0, \\ &F''_{n,N}(\rho) + F'_{n,N}(\rho)\gamma(\rho)/\rho = 1/\alpha(\rho). \end{aligned} \tag{4.5}$$

Combining this with (2.8), we have

$$2LF_{n,N}(\rho)(\rho(x,y)) \geqslant 1. \tag{4.6}$$

Put $r = |x - y|$. Since $P^{x,y} \sim L(a, b)$, by a truncating argument we have

$$\begin{aligned} &E^{x,y} F_{n,N}\big(|X_{t\wedge T_{n,N}} - Y_{t\wedge T_{n,N}}|\big) - F_{n,N}(r) \\ &\quad = \frac{1}{2} E^{x,y} \int_0^{t\wedge T_{n,N}} 2LF(|X_u - Y_u|)du \\ &\quad \geqslant E^{x,y}(t \wedge T_{n,N}), \end{aligned}$$

and so

$$E^{x,y}(t \wedge T_{n,N}) \leqslant -2F_{n,N}(r).$$

Letting $t \uparrow \infty$, we get

$$E^{x,y}(T_{n,N}) \leqslant -2F_{n,N}(r) < \infty. \tag{4.7}$$

(i) If $g(0) < \infty$, then

$$F_{0,N}(r) \equiv \lim_{n\to\infty} F_{n,N}(r) > -\infty.$$

From (4.7), it follows that

$$E^{x,y}(T \wedge S_N) \leqslant -2F_{0,N}(r) < \infty. \tag{4.8}$$

On the other hand, since $Lf(\rho(x, y)) \leqslant 0$, we have

$$\begin{aligned} &f(1/n)\, P^{x,y}(T < S_n \wedge t) + f(N)\, P^{x,y}(S_N < T_n \wedge t) \\ &\quad + E^{x,y}\big(f(\rho(X_t, Y_t)) : t \leqslant T_{n,N}\big) \leqslant f(r). \end{aligned} \tag{4.9}$$

Hence, by (4.7) we get

$$f(1/n)\, P^{x,y}(T < S_n) + f(N)\, P^{x,y}(S_N < T_n) \leqslant f(r).$$

Thus

$$P^{x,y}(T_n > S_N) \leqslant \frac{f(r) - f(1/n)}{f(N) - f(1/n)}.$$

Noting that $g(0) < \infty \Longrightarrow f(0) > -\infty$, we have

$$P^{x,y}(T > S_N) \geqslant \frac{f(r) - f(0)}{f(N) - f(0)}.$$

Letting $N \uparrow \infty$ and using (4.8), we obtain[2]

$$P^{x,y}(T = \infty) = 0.$$

[2]Addition to the original proof: The condition "$g(0) < \infty$" is used here, which implies that $P^{x,y}(T = S_N = \infty) = 0$. Hence $\{T = \infty\} = \{T = \infty, T > S_N\} \subset \{T < S_N\}$, $P^{x,y}$-a.s. This point can be easily missed, as we did in the earlier version of the paper, but was pointed out to us by L. P. Huang who is especially acknowledged here.

(ii) First, we assume that $f_*(\infty)<\infty$. By (4.7), we have

$$f_*(1/n)\,P^{x,y}(T_n<S_N)+f_*(N)\,P^{x,y}(T_n>S_N)\leqslant f_*(r).$$

Hence

$$P^{x,y}(T_n<S_N)\leqslant\frac{f_*(N)-f_*(r)}{f_*(N)-f_*(1/n)}$$

and so

$$P^{x,y}[T<\infty]\leqslant P^{x,y}(T_n<\infty)\leqslant\frac{f_*(\infty)-f_*(r)}{f_*(N)-f_*(1/n)}<1.$$

Next, we assume that $g_*(0)=\infty$. Set

$$g_*(\rho,N)=\int_\rho^N C_*(s)^{-1}ds\int_s^N\frac{C_*(u)}{\alpha^*(u)}du<\infty,\qquad 0<\rho\leqslant N.$$

For $0<\rho\leqslant N$, set $g_*^{(0)}(\rho,N)=1$ and define

$$g_*^{(m)}(\rho,N)=\int_\rho^N C_*(s)^{-1}ds\int_s^N\frac{C_*(u)}{\alpha^*(u)}g_*^{(m-1)}(u,N)du$$

inductively. Then, it is easy to check that

$$g_*^{(m)}(\rho,N)\leqslant\frac{1}{m!}g_*(\rho,N)^m,\qquad m\geqslant 0.$$

Hence

$$u_N(\rho):=\sum_{m=0}^\infty g_*^{(m)}(\rho,N)$$

is well-defined for all $\rho\in(0,N]$. Moreover,

$$\begin{aligned}&u_N\geqslant 1,\qquad u_N'\leqslant 0,\\ &1+g_*(\rho,N)\leqslant u_N(\rho)\leqslant\exp(g_*(\rho,N)),\\ &u_N(\rho)=\alpha^*(\rho)\bigg(-u_N''(\rho)+\frac{\gamma_*(r)}{r}u_N'(\rho)\bigg)\end{aligned}$$

and so

$$\lim_{\rho\downarrow 0}u_N(\rho)=\infty,\qquad 2Lu_N(\rho(x,y))\leqslant u_N(\rho(x,y)).$$

Finally, fix $x\neq y$ and set $|x-y|=r>0$. Then, by a truncating argument, for every $N>r$, we have

$$\begin{aligned}u_N(r)&\geqslant E^{x,y}\Big[e^{-T_{n,N}\wedge t/2}u_N\Big(\rho\big(X_{T_{n,N}\wedge t},Y_{T_{n,N}\wedge t}\big)\Big)\Big]\\ &\geqslant E^{x,y}\Big[e^{-t/2}u_N\Big(\rho\big(X_{T_{n,N}\wedge t},Y_{T_{n,N}\wedge t}\big)\Big):T_n\leqslant S_N\wedge t\Big]\\ &=u_N(1/n)e^{-t/2}\,P^{x,y}(T_n\leqslant S_N\wedge t),\end{aligned}$$

that is,

$$P^{x,y}(T_n \leqslant S_N \wedge t) \leqslant e^{-t/2} u_N(r)/u_N(1/n).$$

Letting $n \to \infty$ and then $N \to \infty$, we get

$$P^{x,y}(T \leqslant t) = 0, \qquad t > 0.$$

This gives us

$$P^{x,y}(T < \infty) = 0.$$

Equivalently,

$$P^{x,y}(T = \infty) = 1. \quad \square$$

EXAMPLE 4.10 [Classical coupling of Brownian motion (B.M.) in $\mathbb{R}^d$]. We have $\gamma(r) = \gamma_*(r) = d-1$, $\alpha(r) = \alpha^*(r) = 2$. Hence

$$f(r) = \begin{cases} r-1, & d=1, \\ \log r, & d=2, \\ (1-r^{-d+2})/(d-2), & d \geqslant 3. \end{cases}$$

$$g(r) = \begin{cases} \dfrac{1}{4}(1-r^2), & d=1, \\ \frac{1}{4}\left(-\log r - \dfrac{1}{2} + \dfrac{r^2}{2} \right), & d=2, \end{cases}$$

and so the coupling is successful if and only if $d=1$. This result should come as no surprise since Brownian motion does not hit points in $d \geqslant 2$.

EXAMPLE 4.11 (Coupling of B.M. in $\mathbb{R}^d$ by reflection or projection). In both cases, we have $\gamma = 0$ and $\alpha =$ positive constant. Thus, $f(r) = r-1$, $f'(r) = 1 > 0$ and so these couplings are successful.

EXAMPLE 4.12 (Coupling of different diffusions). Take $d=1$, $a_1(x) = a_1 > 0$, $a_2(y) = a_2 > 0$, $b_1(x) = -b_1 x$, $b_2(y) = -b_1 y - b_2$, $b_1, b_2 > 0$. Using the coupling by reflection, we get

$$\begin{aligned} &\alpha(u) = \alpha = \left(\sqrt{a_1} + \sqrt{a_2}\right)^2, \\ &\gamma(u) = \frac{2}{\alpha}(-b_1 u^2 + b_2 u), \\ &f(r) = \exp\left(-\frac{(b_1-b_2)^2}{\alpha b_1} \right) \int_1^r \exp\left[\frac{b_1}{\alpha} \left(u - \frac{b_2}{b_1} \right)^2 \right] du, \\ &f'(r) > 0. \end{aligned}$$

Hence the coupling is successful. If $a_1 \neq a_2$, then the basic coupling is also successful.

REMARK 4.13. Based on the idea of reflection, Lindvall and Rogers (1986) proposed a coupling by taking

$$c(x,y) = \sigma(x)\left(\sigma(y)^* - 2\frac{\sigma(y)^{-1}(x-y)(x-y)^*}{|\sigma(y)^{-1}(x-y)|^2} \right).$$

Under some hypotheses, they proved that this coupling satisfies the conditions of (i) of Corollary 4.3, so is successful. Since the hypotheses of Theorem 4.2 for success are weaker than those given in Corollary 4.3, our criterion is applicable to their case.

EXAMPLE 4.14[3]. Take $\sigma(x) = \sqrt{2ax}$, $b(x) = cx + d$, $x \geqslant 0$, $a > 0$ and $d \geqslant 0$. The diffusion process on $[0, \infty)$ for this operator is well-defined [cf. Ikeda and Watanabe (1981), pages 221–222]. Use the coupling by reflection,

$$c(x, y) = -2a\sqrt{xy}.$$

We have

$$\begin{aligned}
A(x, y) &= \operatorname{tr} A(x, y) = \overline{A}(x, y) = 2a\big(\sqrt{x} + \sqrt{y}\,\big)^2, \\
\alpha(r) &= 2a \inf_{|x-y|=r} \big(\sqrt{x} + \sqrt{y}\,\big)^2, \\
&= 2a \inf_{x \geqslant 0} \big(\sqrt{x} + \sqrt{x + r}\,\big)^2, \\
&= 2ar, \\
C(s) &= \exp\left[\frac{c}{a}(s - 1)\right], \\
f(r) &= \begin{cases} r - 1, & c = 0 \\ \dfrac{a}{c}\left[1 - \exp\left[\dfrac{c}{a}(1 - r)\right]\right], & c \neq 0. \end{cases}
\end{aligned}$$

Thus, $f(\infty) = \infty$ if and only if $c \leqslant 0$. Notice that in one-dimensional case, if $\gamma = 0$, then

$$g(0) < \infty \Longleftrightarrow \lim_{r \to 0} \int_r^1 \frac{s - r}{\alpha(s)} ds < \infty \tag{4.15}$$

In the present case, $\gamma \leqslant 0$ and

$$\lim_{r \to 0} \int_r^1 \frac{s - r}{s} ds = 1 < \infty.$$

By Theorem 4.2, we conclude that the coupling is successful for all $c \leqslant 0$.

Since $\inf_{r>0} \alpha(r) = 0$, Corollary 4.3 is not available for this example.

[3]Correction to the original proof. Because

$$\gamma(u) = \sup_{|x-y|=u} 2\widehat{B}(x, y)/\overline{A}(x, y) = 2 \sup_{|x-y|=u} c\big(\sqrt{x} - \sqrt{y}\,\big)^2 = 0$$

provided $c \leqslant 0$. Hence, the conclusion that $\gamma(u) = cu/a$ is incorrect. Here, the correction is based on the first footnote of the paper. However, one needs only to remove the line concerning with the original γ.

EXAMPLE 4.16 (One-dimensional linear growth model).

$$\alpha(x) = \alpha x + b, \qquad b(x) = cx + d, \qquad a \neq 0.$$

Consider the basic coupling

$$c(x, y) = (ax + b)(ay + b).$$

Then $\gamma(u) = \gamma_*(u) = -2c/a^2$, $\alpha(u) = \alpha^*(u) = a^2u^2$.

It is easy to check either $f(\infty) < \infty$ or $g(0) = \infty$. Hence the coupling is always not successful. Even if $c \leqslant 0$ then $f(\infty) = \infty$ and $f(0) > -\infty$. Hence it is not difficult to prove that

$$P^{x,y}\Big(\lim_{t\to\infty} |X_t - Y_t| = 0\Big) = 1,$$

but we still have

$$P^{x,y}(T = \infty) = 1, \qquad x \neq y.$$

The above example shows that the basic coupling is useless for the V-metric. However, for negative c, the basic coupling is not only effective but also provides an exponential rate for the W_1-metric (Theorem 2.3). Conversely, for B.M., the basic coupling gives us

$$P^{x,y}(X_t - Y_t = x - y) = 1$$

and so is useless for the W_1-metric. But as we have seen in Example 4.11, we still have an effective coupling for the V-metric. Thus, the suitable couplings are different for different metrics. For different models, we even need different metrics.

Now, we return to the third coupling given in Example 1.2.

EXAMPLE 4.17 (B.M. in $\mathbb{R}^d$).

$$c_{ij}(x, y) = \left(1 - \frac{\alpha|x_i - y_i|}{\beta + |x_i - y_i|}\right)\delta_{ij}, \qquad 1 \leqslant i, j \leqslant d.$$

Observe

$$\begin{aligned}
\overline{A}(x, y) &= 2\alpha \sum_{i=1}^{d} \frac{|x_i - y_i|}{\beta + |x_i - y_i|} \Big/ |x - y|^2 \\
&\geqslant \frac{2\alpha d}{\beta + u}\left(\frac{1}{d}\sum_{i=1}^{d} |x_i - y_i|^3 \Big/ |x - y|^2\right) \\
&\geqslant \frac{2\alpha}{\sqrt{d}} \frac{u}{\beta + u}, \qquad \text{if } |x - y| = u,
\end{aligned}$$

$$\begin{aligned}\operatorname{tr} A(x,y) &= \sum_{i=1}^{d} \frac{2\alpha|x_i - y_i|}{\beta + |x_i - y_i|} \\ &= 2\alpha\bigg(d - \beta\sum_{i=1}^{d} \frac{1}{\beta + |x_i - y_i|}\bigg) \\ &\leqslant 2\alpha d\bigg(1 - \frac{\beta}{\beta + u}\bigg) \\ &= \frac{2\alpha d u}{\beta + u}, \qquad \text{if } |x-y| = u.\end{aligned}$$

Thus,

$$\frac{\operatorname{tr} A(x,y)}{\overline{A}(x,y)} - 1 \leqslant d - 1,$$

and so

$$\gamma(u) = d - 1, \qquad \alpha(u) = \frac{2\alpha}{\sqrt{d}}\,\frac{u}{\beta + u}.$$

Our criterion (Theorem 4.2) is available only for $d = 1$.

However, this coupling is successful in any dimension. The reason is that we can use the following simple result to reduce the general case to the case that $d = 1$.

DECOMPOSITION LEMMA 4.18. If a coupling consists of two independent parts, and each part has the property that when they hit they will move together, then

$$T = T_1 \vee T_2,$$

where T_1 and T_2 are, respectively, the coupling times of the two parts. In other words, the coupling is successful if and only if each part is successful.

We have seen that Theorem 4.2 is less and less effective as the dimension increases. The role of Lemma 4.18 is to deduce the higher dimensional case to the lower dimensional case. The idea is that, if the components of the original process are independent, we may construct a coupling in two steps: First, for each component, construct a coupling such that after the marginals of the component meet each other, they move together (cf. Theorem 3.1 and Lemma 3.4). Second, link these individual couplings together independently. Example 4.17 illustrates such a construction. As another application of this idea, let us again consider B.M. in $\mathbb{R}^d$. Take the coupling diffusion coefficient as

$$a(t,x,y) = \begin{pmatrix} I & c(t,x,y) \\ c(t,x,y) & I \end{pmatrix},$$

where

$$c_{ij}(t,x,y) = \Big(-I_{[0,T_i)}(t) + I_{[T_i,\infty)}(t)\Big)\delta_{ij}, 1 \leqslant i,j \leqslant d,$$

and

$$T_i = \inf\{t \geqslant 0 : X_i(t) = Y_i(t)\}, \qquad 1 \leqslant i \leqslant d.$$

This construction works also for the higher dimensional analogue of Example 4.14.

REMARK 4.19. We now would like to know what coupling is the optimal. We now would like to know what coupling is the optimal for the moment. Based on Theorem 4.2, we may say that a coupling is V-optimal if

$$a(x,y) \text{ is nonnegative definite and } \overline{A}(x,y) \neq 0, \tag{4.20}$$

$$\operatorname{tr} A(x,y) - \overline{A}(x,y) \text{ achieves the minimum and} \tag{4.21}$$

$$\overline{A}(x,y) \text{ achieves the maximum.} \tag{4.22}$$

By the Schwarz inequality, we have

$$\operatorname{tr} A(x,y) \geqslant \overline{A}(x,y). \tag{4.23}$$

Thus, a special case of (4.21) is that (4.23) becomes equality. This happens if and only if

$$c(x,y) + c(x,y)^* = a_1(x) + a_2(y) - \lambda(x,y)^2 \frac{(x-y)(x-y)^*}{|x-y|^2}. \tag{4.24}$$

For B.M., any $\lambda(x,y)$ satisfying

$$0 < \lambda(x,y)^2 \leqslant 4.$$

will give us a solution to (4.20) and (4.21). Furthermore, if we assume that $c = c^*$, then the coupling by reflection [i.e., $\lambda(x,y)^2 = 4$] is V-optimal. Next, if $a_1 = a_2 = \sigma^2 =$ constant, then the couplings either by reflection or by projection do satisfy (4.20) and (4.21). If we insist on choosing an orthogonal matrix H mentioned in Remark 2.11, then, for constant $a_1 = a_2 = \sigma^2$, $\det \sigma \neq 0$,

$$H = I - 2\sigma^{-1}(x-y)(x-y)^*\sigma^{-1}/|\sigma^{-1}(x-y)|^2$$

is a solution to (4.20) and (4.21), but this is no longer true when the matrix σ depends on x. Similarly, if we consider projection matrix H, the solution is

$$H = I - \sigma^{-1}(x-y)(x-y)^*\sigma^{-1}/|\sigma^{-1}(x-y)|^2.$$

It is still an open problem to give a general formula for the optimal couplings for V-metric.

5. Rates of convergence in total variation norm. Let us begin this section with an example [Lindvall and Rogers (1986)]. Consider the coupling of B.M. in $\mathbb{R}^d$ by reflection. Using the functional

$$Z_t^\alpha = \exp\left[\alpha(|x-y| - |X_t - Y_t|) - 2\alpha^2 t\right], \qquad \alpha > 0,$$

it is easy to prove that

$$E^{x,y} \exp[-\lambda T] = \exp\left[-\sqrt{\lambda/2}\,|x-y|\right], \qquad \lambda > o$$

[see Williams (1979), pages 85–86, for example]. Hence

$$P^{x,y}[T > t] = 2\sqrt{\frac{2}{\pi}} \int_0^{|x-y|/(2\sqrt{t})} \exp\left[-\frac{u^2}{2}\right] du.$$

Thus

$$\begin{aligned}
&\frac{1}{2}\|P(t,x,\cdot) - P(t,y,\cdot)\|_{\mathrm{Var}} \\
&\quad = V(P(t,x,\cdot), P(t,y,\cdot)) \leqslant E^{x,y}\big[I_{X_t \neq Y_t}\big] \\
&\quad = P^{x,y}[T > t] \leqslant \text{ Const. } |x-y|/\sqrt{t} \to 0, \qquad \text{as } t \uparrow \infty.
\end{aligned}$$

On the other hand, it is known that

$$\frac{1}{2}\|P(t,x,\cdot) - P(t,y,\cdot)\|_{\mathrm{Var}} = \sqrt{\frac{2}{\pi}} \int_0^{|x-y|/(2\sqrt{t})} \exp\left[-\frac{u^2}{2}\right] du;$$

thus, the coupling by reflection is exact for the V-metric. Similarly, one can easily check that the coupling by projection will give us the same rate $1/\sqrt{t}$ up to a constant. This procedure produces some estimates for the rates of convergence in some special cases. Now, we are going to use a different idea. Obviously, if $E(T^m) < \infty$, then

$$t^m V(P(t,x,\cdot), P(t,y,\cdot)) \leqslant t^m P[T > t] \leqslant E[T^m : T > t] \to 0, \qquad t \to \infty.$$

This leads us to study the moments of T.

Recall that

$$F_{n,N}(r) = -\int_{1/n}^{r} C(s)^{-1} ds \int_s^N \frac{C(u)}{\alpha(u)} du, \qquad \frac{1}{n} \leqslant r \leqslant N,\ n, N > 1.$$

and define

$$\begin{aligned}
F(r) &= \lim_{n\to\infty} \lim_{N\to\infty} F_{n,N}(r), \qquad 0 < r < \infty, \\
M_n(s_1,s_2) &= \big(C_*(s_1)C(s_2)\big)^{-1} \int_{1/n}^{s_2} \frac{C(u)}{\alpha^*(u)} du, \qquad s_1, s_2 > 0.
\end{aligned}$$

For the following result, we are again comparing with a radial process.

THEOREM 5.1. Put $r = |x-y|$.
(i) If $F(r) > -\infty$, then $E^{x,y}(T) < \infty$.
(ii) If $T_{n,N} < \infty$, $P^{x,y}$-a.s. and

$$\lim_{\substack{n\to\infty \\ N\to\infty}} \iint_{\substack{1/n \leqslant s_1 \leqslant r \\ r \leqslant s_2 \leqslant N}} [M_n(s_1,s_2) - M_n(s_2,s_1)] ds_1 ds_2 \int_{1/n}^{N} C_*(s)^{-1} ds = \infty,$$

then $E^{x,y}(T) = \infty$, $x \neq y$.

PROOF. (i) From (4.7), we see that

$$E^{x,y}(T_{n,N}) \leqslant -2F_{n,N}(r). \tag{5.2}$$

Let $N \to \infty$ and then $n \to \infty$ to get

$$E^{x,y}(T) \leqslant -2F(r) < \infty.$$

(ii) Set

$$G_n(r) = \int_{1/n}^{r} C(s)^{-1} ds \int_{1/n}^{s} \frac{C(u)}{\alpha^*(u)} du.$$

Since $G_n(\rho) \geqslant 0$, $G_n'(\rho) \geqslant 0$, $G_n''(\rho) + (1/\rho)\,\gamma(\rho)\,G_n'(\rho) = 1/\alpha^*(\rho)$, for $\rho \geqslant 1/n$. We have

$$2LG_n(\rho(x,y)) \leqslant 1.$$

Hence

$$E^{x,y} G_n\big(\rho\big(X_{T_{n,N}}, Y_{T_{n,N}}\big)\big) \leqslant G_n(r) + \frac{1}{2} E^{x,y}(T_{n,N}). \tag{5.3}$$

On the other hand, if we set

$$H_n(r) = \int_{1/n}^{r} C_*(s)^{-1} ds,$$

then

$$H_n(\rho) \geqslant 0, \quad H_n'(\rho) \geqslant 0, \quad H_n''(\rho) + \frac{1}{\rho}\gamma_*(\rho) H_n'(\rho) = 0, \quad \rho \geqslant \frac{1}{n}.$$

Because

$$E^{x,y} H_n\big(\rho\big(X_{T_{n,N}}, Y_{T_{n,N}}\big)\big) = H_n(r) + E^{x,y} \int_0^{T_{n,N}} LH_n\big(\rho\big(X_u, Y_u\big)\big) du \geqslant H_n(r),$$

we get

$$P^{x,y}(S_N < T_n) \geqslant H_n(r)/H_n(N). \tag{5.4}$$

Combining (5.3) with (5.4), we obtain

$$E^{x,y}\big(T_{n,N}\big) \geqslant 2\frac{H_n(r)G_n(N) - H_n(r)G_n(r)}{H_n(N)}.$$

Since $T_{n,N}$ is increasing as $n \to \infty$ or $N \to \infty$, $T_{n,N} \uparrow T$. Thus, the assumption of the theorem implies that

$$E^{x,y}(T) = \infty.$$

EXAMPLE 5.5 (O.U. process).

$$\sigma(x) = I, \qquad b(x) = -x.$$

Using the coupling by reflection, we get

$$\begin{aligned}
\alpha(r) &= \alpha^*(r) = 4,\\
\gamma(r) &= \gamma_*(r) = -r^2/2,\\
C(r) &= \exp\left[\int_0^r \frac{\gamma(u)}{u}du\right] = \exp\left[-\frac{1}{4}(r^2-1)\right],\\
-F_{n,N}(r) &= \int_{1/n}^r \exp\left[\frac{1}{4}(s^2-1)\right]ds \int_s^N \frac{1}{4}\exp\left[-\frac{1}{4}(u^2-1)\right]du\\
&\leqslant \frac{1}{4}\left(\int_0^r \exp\left[\frac{1}{4}s^2\right]ds\right)\left(\int_0^\infty \exp\left[-\frac{1}{4}u^2\right]du\right).
\end{aligned}$$

Hence $-F(r) < \infty$ for all $r \in (0,\infty)$, and so

$$E^{x,y}(T) < \infty.$$

EXAMPLE 5.6 (The coupling of B.M. in $\mathbb{R}^d$ by reflection).

$$E^{x,y}(T) = \infty, \qquad x \neq y.$$

Now we investigate the higher moments of the coupling time T.

THEOREM 5.7. If there exist constants $\beta > 0$, $0 < \alpha < \beta$, and $c = c(\alpha)$ such that

$$(\beta-2)\overline{A}(x,y) + \operatorname{tr} A(x,y) + 2\widehat{B}(x,y) \leqslant 0 \tag{5.8}$$

for all $(x,y) : 0 < p(x,y) < \infty$, and

$$|F(r)| \leqslant c\, r^\alpha, \qquad 0 < r < \infty \tag{5.9}$$

Then

$$E^{x,y}(T^m) < \infty, \qquad m \in [0, \beta/\alpha). \tag{5.10}$$

PROOF. By (5.8), it is easy to prove that

(i) $\sup_{t\geqslant 0} E^{x,y}\big(|X(t\wedge T_{n,N}) - Y(t\wedge T_{n,N})|^\beta\big) \leqslant |x-y|^\beta. \quad n, N \geqslant 1.$

Next, by using an integration by parts formula for martingale theory [Stroock and Varadhan (1979), Theorem 1.2.8] and a truncating argument, we may prove that

(ii) $E^{x,y}(T_{n,N}^{1+m}) \leqslant 2m(1+m)\,E^{x,y}\displaystyle\int_0^{T_{n,N}} |F(|X(s)-Y(s)|)\,s^{m-1}ds.$

This is the main trick of the proof. Now, by (5.9), Hölder inequality and (i), we would have

$$E^{x,y}\big[|F(|X(s)-Y(s)|)|; s \leqslant T_{n,N}\big] \leqslant C|x-y|^{\alpha}\,P^{x,y}\big[T_{n,N} \geqslant s\big]^{(\beta-\alpha)/\beta}.$$

Inserting this into (ii) and letting $N, n \to \infty$, we would obtain

(iii) $E^{x,y}(T^{1+m}) \leqslant 2Cm(1+m)\,|x-y|^{\alpha}\displaystyle\int_0^{\infty} s^{m-1}P^{x,y}[T \geqslant s]^{(\beta-\alpha)/\beta}ds.$

On the other hand, from Theorem 5.1 and 5.9, we see that $E^{x,y}(T) < \infty$. Thus, by using the inequality (iii), we may maximize the number m with property $E^{x,y}(T^m) < \infty$. For more details, refer to the proof of Lemma 7 in Davies (1986). □

EXAMPLE 5.11. Everything is the same as Example 4.12 but for simplicity, we take $a_1 = a_2 = 1$. We know that

$$\gamma(r) = \frac{1}{2}\Big(- b_1 r^2 + \frac{b_2}{2} r).$$

Hence,

$$\begin{aligned} C(r) &= \exp\left[\int_1^r \frac{\gamma(u)}{u}du\right] = \exp\left[-\frac{b_1}{4}(r^2-1) + \frac{b_2}{2}(r-1)\right], \\ -F_{n,N}(r) &= \frac{1}{4}\int_{1/n}^r \exp\left[\frac{b_1}{4}s^2 - s\right]ds \int_s^N \exp\left[-\frac{b_1}{4}u^2 + \frac{b_2}{2}u\right]du \\ |F(r)| &= \frac{1}{4}\int_0^r \exp\left[\frac{b_1}{4}\left(s - \frac{b_2}{b_1}\right)^2\right]ds \int_s^{\infty} \exp\left[-\frac{b_1}{4}\left(u - \frac{b_2}{b_1}\right)^2 u\right]du \end{aligned}$$

and so, for any $0 < \alpha < 1$,

$$\frac{F(r)}{r^{\alpha}} \sim \frac{\int_r^{\infty} \exp[-(b_1/4)(u-b_2/b_1)^2]}{\alpha r^{\alpha-1}\exp\big[-(b_1/4)(r-b_2/b_1)^2\big]} \to 0, \qquad r \to \infty$$

Thus, (5.1) holds for any $0 < \alpha < 1$, and so

$$E(T^m) < \infty \quad \text{for any} \quad m \geqslant 0.$$

Finally, we consider exponential estimates for the rate of convergence.

THEOREM 5.12. Suppose that

(i) there exist constants $C \geqslant 0,\ c > 0$ such that

$$L\rho^2(x,y) \leqslant C - c\rho^2(x,y), \tag{5.13}$$

(ii) there exist $N > N_1 > C/c$ such that

$$F_{0,N}(N_1) = \int_0^N C(s)^{-1}ds \int_s^N \frac{C(u)}{\alpha(u)}du < \infty \tag{5.14}$$

and

$$\frac{N_1^2\int_{N_1}^N C(s)^{-1}ds}{\int_0^N C(s)^{-1}ds} > \frac{C}{c}. \tag{5.15}$$

Then there exists $t_0 > 0$ such that for $t \geqslant t_0$, we have

$$\begin{aligned} P^{x,y}(T > nt) &\leqslant K_1 k^n, \\ E^{x,y}\rho^2\big(X_{nt}, Y_{nt}\big) &\leqslant K_2 k^n, \end{aligned} \tag{5.16}$$

for some constants $K_1,\ K_2 > 0$ and $k \in (0,1)$.

PROOF. Recall

$$f(r) = \int_1^r C(x)^{-1}ds, \qquad Lf(\rho(x,y)) \leqslant 0.$$

From (4.9), we see that

$$P^{x,y}(T_n < S_N \wedge t) \geqslant \frac{f(N)-f(r)}{f(N)-f(1/n)} - P^{x,y}(T_{n,N} \geqslant t).$$

Letting $n \to \infty$, we have

$$P^{x,y}(T_n \leqslant S_N \wedge t) \geqslant \frac{f(N)-f(r)}{f(N)-f(0)} - P^{x,y}(T_{0,N} \geqslant t),$$

where $T_{0,N} = T \wedge S_N$, and so

$$P^{x,y}(T_n \leqslant t) \geqslant \frac{f(N)-f(r)}{f(N)-f(0)} - P^{x,y}(T_{0,N} \geqslant t). \tag{5.17}$$

By the condition (5.14) and using (5.2) we get

$$E^{x,y}(T_{0,N}) \leqslant -2F_{0,N}(r).$$

Hence

$$P^{x,y}(T_{0,N} \geqslant t) \leqslant \frac{E^{x,y}(T_{0,N})}{t} \leqslant -\frac{2F_{0,N}(r)}{t}. \tag{5.18}$$

Combining (5.17) with (5.18) we obtain

$$\begin{aligned} P^{x,y}(T \leqslant t) &\geqslant \frac{f(N)-f(r)}{f(N)-f(0)} + \frac{2F_{0,N}(r)}{t} \\ &\geqslant \frac{\int_{N_1}^N C(s)^{-1}ds}{\int_0^N C(s)^{-1}ds} + \frac{2F_{0,N}(r)}{t} \end{aligned} \tag{5.19}$$

for all $(x,y): 0<\rho(x,y)=r<N_1$. Let

$$\alpha=\frac{N_1^2\int_{N_1}^N C(s)^{-1}ds}{\int_0^N C(s)^{-1}ds}-\frac{C}{c}.$$

Then $\alpha>0$ by (5.15). Clearly, we can find $t_1>0$ such that

$$2N_1^2\frac{|F_{0,N}(N_1)|}{t}\leqslant\frac{\alpha}{2},$$
$$\frac{\int_{N_1}^N C(s)^{-1}ds}{\int_0^N C(s)^{-1}ds}-\frac{2F_{0,N}(N_1)}{t}>0,$$

for all $t\geqslant t_1$. Also, we can find $t_2>0$ such that

$$\frac{C/c}{1-e^{-ct}}<\frac{C}{c}+\frac{\alpha}{2}$$

for all $t\geqslant t_2$. Take $t_0=t_1\vee t_2$. Then for all $t\geqslant t_0$, we have

$$N_1^2\left[\frac{\int_{N_1}^N C(s)^{-1}ds}{\int_0^N C(s)^{-1}ds}+\frac{2F_{0,N}(N_1)}{t}\right]>\frac{C/c}{1-e^{-ct}}. \tag{5.20}$$

Now, we fix $t\geqslant t_0$ and let

$$\frac{\int_{N_1}^N C(s)^{-1}ds}{\int_0^N C(s)^{-1}ds}-\frac{2F_{0,N}(N_1)}{t}=1-\delta,\qquad 0<\delta<1.$$

By (5.19), (5.20), (5.13) and Theorem 2.3 we arrive at

(a) $P^{x,y}(T>t)<\delta,\qquad 0<|x-y|<N_1$,
(b) $N_1^2(1-\delta)(1-e^{-ct})>C/c$,
(c) $E^{x,y}\rho^2(X_t,Y_t)\leqslant C/c+e^{-ct}\rho^2(x,y)$.

Let $\tau_n=nt\wedge T$ and $\{P_\omega\}$ be a regular conditional probability distribution $P^{x,y}|_{\mathscr{M}_{\tau_{n-1}}}$. Then $\delta_{(X(\tau_{n-1}),Y(\tau_{n-1}))}\bigotimes_{\tau_{n-1}}P_\cdot$ is the solution to the martingale problem for the coupling operator $L(a,b)$ starting from $(X(\tau_{n-1}),Y(\tau_{n-1}))$. Define

$$I_n=I_{[\tau_n<T]}=I_{[\tau_n=nt]},$$
$$J_n=\rho^2(X(\tau_n),Y(\tau_n))I_n.$$

Then $I_n\leqslant I_{n-1}$ and $I_{n-1}=0$ implies $I_n=0$. Thus, by using (a), we obtain

$$\begin{aligned}E^{x,y}(I_n)&=E^{x,y}\Big[I_{n-1}E^{x,y}\big(I_n\big|\mathscr{M}_{\tau_{n-1}}\big)\Big]\\&=E^{x,y}\Big[I_{n-1}E^{\delta\otimes P_\cdot}\big(I_n\big),\,\rho(X(\tau_{n-1}),Y(\tau_{n-1}))\leqslant N_1\Big]\\&\quad+E^{x,y}\Big[I_{n-1}E^{\delta\otimes P_\cdot}\big(I_n\big),\,\rho(X(\tau_{n-1}),Y(\tau_{n-1}))>N_1\Big]\\&\leqslant\delta E^{x,y}(I_{n-1})+\frac{1}{N_1^2}E^{x,y}(J_{n-1}),\end{aligned}$$

where $\delta \bigotimes P_\cdot = \delta_{(X(\tau_{n-1}),Y(\tau_{n-1}))} \bigotimes_{\tau_{n-1}} P_\cdot$. Next, by using (c), we get

$$E^{x,y}(J_n) = E^{x,y}\Big[I_{n-1}E^{x,y}\big(J_n\big|\mathscr{M}_{\tau_{n-1}}\big)\Big] \leqslant (C/c)\, E^{x,y}(I_{n-1}) + e^{-ct}E^{x,y}(J_{n-1}).$$

Finally, the assertion (b) guarantees that the eigenvalues of the matrix

$$\begin{pmatrix} \delta & 1/N_1^2 \\ C/c & e^{-ct} \end{pmatrix}$$

are less than 1. Let $\lambda_1,\ \lambda_2$ be the eigenvalues and take $k \in [\lambda_1 \vee \lambda_2, 1)$. Then (5.16) holds for some $K_1,\ K_2 > 0$. $\square$

EXAMPLE 5.21 (O.U. process).

$$\begin{aligned} \sigma(x) &= I, \qquad b(x) = -x \\ L\rho^2(x,y) &= 4 - 2\rho^2(x,y), \qquad C = 4, \quad c = 2, \\ C(r) &= \exp[-(r^2-1)/4], \\ -F_{0,N}(r) &= \int_0^r e^{s^2/4}ds \int_s^N \frac{1}{4}e^{-u^2/4}du < \infty, \\ f(r) &= \int_1^r \exp\big[(s^2-1)/4\big]ds. \end{aligned}$$

Take $N_1 > 2 = C/c$. Then $2/N_1^2 < 1$. Since

$$\lim_{N\to\infty} f(N) = \infty$$

for fixed N_1 and

$$\lim_{N\to\infty} \frac{f(N) - f(N_1)}{f(N) - f(0)} = 1,$$

we can choose N large enough such that

$$\frac{f(N) - f(N_1)}{f(N) - f(0)} > \frac{2}{N_1^2}.$$

This implies (5.15) and hence the hypotheses of Theorem 5.12 are satisfied.

Acknowledgments. This work was done when the authors were visiting the Department of Mathematics, University of Edinburgh. The authors would like to thank the Department for its hospitality, especially Professor T. J. Lyons. The first author also acknowledges the support of the S.E.R.C. of the United Kingdom. Finally, thanks are given to a referee for his detailed comments on the earlier version of this paper.

References

BASIS, V. YA. (1980), *Stationarity and ergodicity of Markov interacting processes*, In Multicomponent Random Systems (R. L. Dobrushin and Ya. G. Sinai, eds.) 37-58. Dekker, New York.

BHATTACHARYA, R. N. and RAMASUBRAMANIAN, S. (1982), *Recurrence and ergodicity of diffusions*, J. Multivariate Anal. 1295-122.

CHEN, M. F. (1986a), *Coupling of jump processes*, Acta Math. Sinica 2 123-136.

CHEN, M. F. (1986b), *Jump Processes and Interacting Particle Systems*, Beijing Normal Univ. Press, Beijing (In Chinese).

CHEN, M. F. (1987a), *Coupling of jump processes* II, Chin. Ann. Math. To appear.

CHEN, M. F. (1987b), *Existence theorems for interacting particle systems with non-compact state spaces*, Sci. Sinica (Ser. A) 30, 148–1156.

DAVIES, P. L. (1986), *Rates of convergence to the stationary distribution for k-dimensional diffusion processes*, J. Appl. Probab. 23, 370–384.

DOBRUSHIN, R. L. (1970), *Prescribing a system of random variables by conditional distributions*, Theory Probab. Appl. 15, 458–486.

FRIEDMAN, A. (1975), *Stochastic Differential Equations and Applications* **I**, Academic, New York.

GIVENS, C. R. and SHORTT, R. M. (1984), *A class of Wasserstein metrics for probability distributions*, Michigan Math. J. 31, 231–240.

GRIFFEATH, D. (1978), *Coupling methods for Markov processes*, Adv. in Math. 2, 1–43.

IKEDA, N. and WATANABE, S. (1981), *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Amsterdam.

KENDALL, W. S. (1986a), *Stochastic differential geometry, a coupling property and harmonic maps*, J. London Math. Soc. (2) 33, 554–566.

KENDALL, W. S. (198Gb), *Nonnegative Ricci curvature and the Brownian coupling property*, Stochastics 19, 111–129.

LIGGETT, T. M. (1985), *Interacting Particle Systems*, Springer New York.

LINDVALL, T. (1983), *On coupling of diffusion processes*, J. Appl. Probab. 20, 82–93.

LINDVALL, T. and ROGERS, L. C. 0. (1986), *Coupling of multidimensional diffusions by reflection*, Ann. Probab. 14, 860–872.

STROOCK, D. W. and VARADHAN, S. R. S. (1979), *Multidimensional Diffusion Processes*, Springer New York.

WILLIAMS, D. (1979), *Diffusions, Markov Processes, and Martingales Foundations* **I**, Wiley, New York.

DEPARTMENT OF MATHEMATICS
BEIJING NORMAL UNIVERSITY
BEIJING, 100875
PEOPLE'S REPUBLIC OF CHINA

DEPARTMENT OF MATHEMATICS
HENAN TEACHER'S UNIVERSITY
XINXIANG, HENAN PROVINCE
PEOPLE'S REPUBLIC OF CHINA

Note of the computation of the partial derivatives of ρ and $f \circ \rho$.

Let $\rho(x,y) = |x-y| = \big(\sum_{i=1}^{d}(x_i - y_i)^2\big)^{1/2}$. Then

$$\begin{aligned}
&\frac{\partial}{\partial x_i}\rho(x,y) = \frac{x_i - y_i}{|x-y|}, \qquad \frac{\partial^2}{\partial x_i^2}\rho(x,y) = \frac{1}{|x-y|} - \frac{(x_i-y_i)^2}{|x-y|^3},\\
&\frac{\partial^2}{\partial x_i \partial x_j}\rho(x,y) = -\frac{(x_i-y_i)(x_j-y_j)}{|x-y|^3}, \qquad i \neq j,\\
&\frac{\partial}{\partial y_i}\rho = -\frac{\partial}{\partial x_i}\rho,\\
&\frac{\partial^2}{\partial x_i \partial y_j}\rho = \frac{\partial}{\partial x_i}\Big(-\frac{\partial}{\partial x_j}\rho\Big) = -\frac{\partial^2}{\partial x_i \partial x_j}\rho,\\
&\frac{\partial^2}{\partial y_i \partial y_j}\rho = \frac{\partial}{\partial y_i}\Big(-\frac{\partial}{\partial x_j}\rho\Big) = -\frac{\partial^2}{\partial x_j \partial y_i}\rho = \frac{\partial^2}{\partial x_i \partial x_j}\rho.
\end{aligned}$$

Furthermore, for $f \in C^2$, noting that

$$\begin{aligned}
&\frac{\partial}{\partial x_i} f\circ\rho = \Big(\frac{\partial}{\partial x_i}\rho\Big) f' \circ \rho,\\
&\frac{\partial^2}{\partial x_i \partial x_j} f\circ\rho = \Big(\frac{\partial}{\partial x_i}\rho\Big)\Big(\frac{\partial}{\partial x_j}\rho\Big) f'' \circ \rho + \Big(\frac{\partial^2}{\partial x_i \partial x_j}\rho\Big) f' \circ \rho\\
&\frac{\partial}{\partial y_i} f\circ\rho = \Big(\frac{\partial}{\partial y_i}\rho\Big) f' \circ \rho = -\Big(\frac{\partial}{\partial x_i}\rho\Big) f' \circ \rho = -\frac{\partial}{\partial x_i} f\circ\rho,\\
&\frac{\partial^2}{\partial x_i \partial y_j} f\circ\rho = \frac{\partial}{\partial x_i}\Big(-\frac{\partial}{\partial x_j} f\circ\rho\Big) = -\frac{\partial^2}{\partial x_i \partial x_j} f\circ\rho,\\
&\frac{\partial^2}{\partial y_i \partial y_j} f\circ\rho = \frac{\partial}{\partial y_i}\Big(-\frac{\partial}{\partial x_j} f\circ\rho\Big) = -\frac{\partial^2}{\partial y_i \partial x_j} f\circ\rho = \frac{\partial^2}{\partial x_i \partial x_j} f\circ\rho,
\end{aligned}$$

we obtain

$$\begin{aligned}
&\frac{\partial}{\partial x_i} f\circ\rho(x,y) = \frac{x_i - y_i}{|x-y|} f' \circ \rho(x,y),\\
&\frac{\partial^2}{\partial x_i^2} f\circ\rho(x,y) = \frac{(x_i-y_i)^2}{|x-y|^2} f'' \circ \rho(x,y) + \Big[\frac{1}{|x-y|} - \frac{(x_i-y_i)^2}{|x-y|^3}\Big] f' \circ \rho(x,y),\\
&\frac{\partial^2}{\partial x_i \partial x_j} f\circ\rho(x,y) = \frac{(x_i-y_i)(x_j-y_j)}{|x-y|^2}\Big[f'' \circ \rho(x,y) - \frac{1}{|x-y|} f' \circ \rho(x,y)\Big], \quad i \neq j,\\
&\frac{\partial}{\partial y_i} f\circ\rho = -\frac{\partial}{\partial x_i} f\circ\rho,\\
&\frac{\partial^2}{\partial x_i \partial y_j} f\circ\rho = -\frac{\partial^2}{\partial x_i \partial x_j} f\circ\rho,\\
&\frac{\partial^2}{\partial y_i \partial y_j} f\circ\rho = \frac{\partial^2}{\partial x_i \partial x_j} f\circ\rho.
\end{aligned}$$

Acta Mathematica Sinica, New Series
1991, Vol.7, No.l, pp.19–37

EXPONENTIAL L^2-CONVERGENCE AND L^2-SPECTRAL GAP FOR MARKOV PROCESSES

Mu-Fa Chen[1]

(Department of Mathematics, Beijing Normal University)
Received August 13, 1988

Abstract. This paper deals with the exponential L^2-convergence for jump processes. We introduce some reduction methods and improve some previous results. Then we prove that for birth–death processes, exponential L^2-convergence coincides indeed with exponential ergodicity which is widely studied in the Markov chain theory.

1. Introduction.

Let $(E, \mathscr{E}, \pi)$ be a probability space, $\{P(t)\}_{t\geqslant 0}$ be a positive, strongly continuous, contractive and Markovian semigroup $(P(t)1 = 1)$ on $L^2(\pi)$ with an invariant measure π. Denote by Ω and $\mathscr{D}(\Omega)$ respectively the infinitesimal generator and its domain induced by $\{P(t)\}_{t\geqslant 0}$. We say that $\{P(t)\}$ converges exponentially in the $L^2(\pi)$ norm if there is a positive ε such that for all $f \in L^2(\pi)$,

$$\|P(t)f - \pi(f)\| \leqslant e^{-\varepsilon t}\|f - \pi(f)\|, \tag{1.1}$$

where $\|\cdot\|$ denotes the $L^2(\pi)$ norm and $\pi(f) = \int f \mathrm{d}\pi$.

Since the constant function $1 \in \mathscr{D}(\Omega)$ and $\Omega 1 = 0$, the vector 1 is an eigenvector of Ω with eigenvalue 0. One may seek for the next-to-largest eigenvalue (resp. real part) of the self adjoint (resp. non-self-adjoint) generator Ω. That is, to seek for the infimum of the spectra of $-\Omega$ restricted to the orthogonal complement space $\{f \in L^2(\pi) : \pi(f) = 0\} \cap \mathscr{D}(\Omega)$. This leads us to define

$$\mathrm{gap}(\Omega) = \inf\{-(\Omega f, f) : f \in \mathscr{D}(\Omega),\ \pi(f) = 0,\ \|f\| = 1\}. \tag{1.2}$$

We know more or less that (1.1) and (1.2) are closely linked (see the next section for more details). Exponential convergence in L^2 sense was proved for

Research supported in part by the National Natural Science Foundation of China and FokYing-Tung Educational Foundation.

[1]In the Chinese journals, the name is often written as Chen Mufa

various classes of stochastic Ising models by Holley and Stroock (1976, 1987), by Holley (1984, 1985a, 1985b) and by Aizenman and Holley (1987). Recently, Liggett (1989) proved that the neatest particle system also exhibits an exponential convergence. He also proved that gap(Ω) coincides indeed with the largest value ε in (1.1). Thus, as in the large deviation theory, we have a common rate formula without ergodic assumption. This is especially useful for the study of interacting particle systems.

Motivated by a quantum field's model, Sullivan (1984) studied the spectral gap for jump processes with state space $\mathbb{Z}^+ = \{0, 1, 2, \dots\}$ or $\mathbb{R}^+$. Under some hypotheses, he proved the existence of the spectral gap for certain bounded operators.

Estimation of the bound of a spectrum has attracted considerable attention in various branches of mathematics. Motivated by a well-known paper by Cheeger (1970) on the lower bound of the Laplacian on a compact manifold, recently, Lawler and Sokal (1988) obtained a general version of Cheeger's inequality for jump processes with general state space and bounded operator. In their paper, our readers can find much more references.

The main purpose of this paper is to extend the previous results to unbounded generators. Some elementary facts from Dirichlet form theory enable us to obtain a complete formula for the convergence rate. This is done in the next section. Then for jump processes, we reduce the non-symmetric case to the symmetric one and reduce the unbounded case to the bounded one. In Section 5, we first improve two results due to Liggett (1989) and Sullivan (1984) respectively. Then, we prove that for birth–death processes, exponential convergence coincides indeed with exponential ergodicity which is widely studied in the Markov chain theory. Also we introduce a procedure to estimate the lower bound of spectral gap for birth-death processes. Finally, we apply Van Doorn's results (1985) to present some bounds of spectral gap for general positive recurrent Markov chains.

In the last section (§6) we briefly discuss the largest eigenvalue of Ω for non-positive recurrent Markov processes by using the techniques developed in the first five sections.

2. Some General Results.

Let $(E, \mathscr{E}, \pi)$ be a probability space and $L^2(\pi)$ be the set of all real square integrable functions with respect to π on $(E, \mathscr{E})$. Given a positive, strongly continuous, contractive and Markovian semigroup $\{P(t)\}_{t\geqslant 0}$ on $L^2(\pi)$ $(P(t)1 = 1)$ with an invariant measure π, we denote by Ω its generator with domain $\mathscr{D}(\Omega)$. Define gap(Ω) by (1.2). Similarly, we can define gap$\big(\widetilde{\Omega}\big)$, where $\widetilde{\Omega}$ is the generator in the weak sense. Denote by $\mathscr{D}\big(\widetilde{\Omega}\big)$ the domain of $\widetilde{\Omega}$ in $L^2(\pi)$. Finally, if the limit

$$\lim_{t\downarrow 0}\frac{1}{t}(f - P(t)f, f) = \lim_{t\downarrow 0}\frac{1}{2t}\int \pi(\mathrm{d}x)(P(t)(f - f(x))^2)(x) \geqslant 0 \tag{2.1}$$

exists, we denote it by $D(f)$. Such functions $f \in L^2(\pi)$ with $D(f) < \infty$ constitute the domain $\mathscr{D}(D)$ of D. In the case of $\{P(t)\}_{t\geqslant 0}$ being symmetric on $L^2(\pi)$, as a direct consequence of elementary spectrum theory (cf. Fukushima (1980)), the

limit defined by (2.1) always exists for all $f \in L^2(\pi)$. We also use $D(f,f)$ to denote the limit. The bilinear form

$$D(f,g) = \frac{1}{4}\big[D(f+g,f+g) - D(f-g,f-g)\big]$$

defined on

$$\mathscr{D}(D) = \{f \in L^2(\pi) : D(f,f) < \infty\}$$

is called the Dirichlet form corresponding to the semigroup $\{P(t)\}_{t\geqslant 0}$. Clearly, in this case, $D(f) = D(f,f)$ with the same domain. This explains why we choose the notations $D(f)$ and $\mathscr{D}(D)$.

Now, we define

$$\operatorname{gap}(D) = \inf\{D(f) : f \in \mathscr{D}(D),\ \pi(f) = 0,\ \|f\| = 1\}. \tag{2.2}$$

For the symmetric case, we have

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

Next, following Liggett (1989), we set

$$\sigma(t) = -\sup\{\log \|P(t)f\| : \pi(f) = 0 \text{ and } \|f\| = 1\}.$$

By the contraction and semigroup properties, it is easy to see that $\sigma(\cdot)$ is super-additive and $\sigma(0) = 0$. Hence, the limit

$$\sigma = \lim_{t\downarrow 0} \frac{\sigma(t)}{t} = \inf_{t>0} \frac{\sigma(t)}{t} \tag{2.3}$$

is well defined.

The following result is an extension of Liggett's (1989, Theorem (2.3)) in which $\sigma = \operatorname{gap}(\Omega)$ was proved.

(2.4) **Theorem.** We have

$$\sigma = \operatorname{gap}(D) = \operatorname{gap}\big(\widetilde{\Omega}\big) = \operatorname{gap}(\Omega).$$

Proof. . The proof is essentially due to Liggett (1989). Clearly,

$$\operatorname{gap}(D) \leqslant \operatorname{gap}\big(\widetilde{\Omega}\big) \leqslant \operatorname{gap}(\Omega) \quad \text{on} \quad \mathscr{D}(\Omega),$$

since

$$D(f) = \big(-\widetilde{\Omega}f, f\big) = (-\Omega f, f) \quad \text{on} \quad \mathscr{D}(\Omega).$$

To prove $\sigma \geqslant \operatorname{gap}(\Omega)$, we simply use the fact:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|P(t)f\|^2 &= 2(P(t)f, \Omega P(t)f)\\ &\leqslant -2\operatorname{gap}(\Omega)\,\|P(t)f\|^2,\\ &f \in \mathscr{D}(\Omega),\ \pi(f) = 0 \text{ and } \|f\| = 1,\end{aligned}$$

and the density of $\mathscr{D}(\Omega)$ in $L^2(\pi)$. Finally, let $f \in \mathscr{D}(D)$, then

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

Hence $\text{gap}(D) \geqslant \sigma$. $\square$

At the moment, except for the fact $\mathscr{D}\big(\widetilde{\Omega}\big) \subset \mathscr{D}(D)$, our knowledge about $\mathscr{D}(D)$ is quite limited. However, it will be clear later, whenever we have a little more information about the generator, the domain $\mathscr{D}(D)$ is actually manageable. The following obvious facts will be helpful for our further study.

(2.5) **Lemma.**

(i) $D(f) \geqslant 0$, $f \in \mathscr{D}(D)$;
(ii) $f \in \mathscr{D}(D) \Longrightarrow g = cf + d \in \mathscr{D}(D)$ and $D(g) = c^2 D(f)$ for all $c, d \in \mathbb{R}$;
(iii) $f, g \in \mathscr{D}(D)$ and $f + g \in \mathscr{D}(D) \Longrightarrow D(f+g) \leqslant 2(D(f) + D(g))$.

As an immediate consequence of Theorem (2.4), we have

(2.6) **Corollary.**

(i) If Ω is bounded, then

$$\sigma = \inf\{(-\Omega f, f) : \pi(f) = 0,\ \|f\| = 1\}.$$

(ii) If Ω is self adjoint, then

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

where $D(f, f)$ is the Dirichlet form corresponding to the semigroup $\{P(t)\}_{t\geqslant 0}$ (resp. generator Ω).

Finally, we want to show that the non-symmetric case can often be reduced to a symmetric case.

Let E be a locally compact separable space with Borel field $\mathscr{E}$, π be a probability measure on $(E, \mathscr{E})$ with $\text{supp}(\pi) = E$. Let $D(f, g)$ $(f, g \in \mathscr{D}(D) \subset L^2(\pi))$ be a generalized Dirichlet form (see Kim (1987) for details). Suppose that the semigroup $\{P(t)\}_{t\geqslant 0}$ corresponding to $D(f, g)$ has an invariant probability π. Obviously, by Theorem (2.4), we have

$$\sigma = \inf\{D(f, f) : f \in \mathscr{D}(D),\ \pi(f) = 0,\ \|f\| = 1\}. \tag{2.7}$$

Next, define the dual of D as follows

$$\widehat{D}(f, g) = D(g, f), \qquad f, g \in \mathscr{D}\big(\widehat{D}\big) = \mathscr{D}(D);$$

and set

$$\overline{D} = \frac{1}{2}\big(D + \widehat{D}\big), \qquad \mathscr{D}\big(\overline{D}\big) = \mathscr{D}(D).$$

Then $\overline{D}$ is a symmetric Dirichlet form for which we have

$$\bar{\sigma} = \inf\big\{\overline{D}(f, f) : \pi(f) = 0,\ \|f\| = 1\big\}. \tag{2.8}$$

But

$$\overline{D}(f, f) = \frac{1}{2}\big(D(f, f) + \widehat{D}(f, f)\big) = D(f, f), \qquad f \in \mathscr{D}\big(\overline{D}\big) = \mathscr{D}(D).$$

Thus, we have proved the following result.

(2.9) **Corollary.** $\sigma = \bar{\sigma}$.

(2.10) **Example.** For the Ornstein-Uhlenbeck process in $\mathbb{R}$,

$$\Omega = \frac{1}{2}\left(\frac{\mathrm{d}^2}{\mathrm{d}x^2} - x\frac{\mathrm{d}}{\mathrm{d}x}\right),$$

we have

$$\sigma = \mathrm{gap}(\Omega) = 1/2,$$

since the eigenvalues of Ω are

$$\lambda_n = n/2, \qquad n \geqslant 0,$$

and the associated eigenvectors belong to $\mathscr{D}(\Omega)$. By the independence of components, this conclusion is also correct in the multidimensional case. Moreover, for the infinite dimensional Ornstein-Uhlenbeck process in Wiener space, we still have

$$\sigma = \mathrm{gap}(\Omega) = 1/2.$$

Cf. Stroock (1981) for details.

More examples for diffusion processes can be found from Karlin and Taylor (1981), Chapter 15, Section 13. Also see Holley and Stroock (1987) and Korzeniowski (1987).

3. Spectral Gap for Jump Processes: General Case.

Let (E, ρ) be a separable locally compact space, $P(t, x, \mathrm{d}y)$ be a jump process on $(E, \rho, \mathscr{E})$. That is,

$$\lim_{t\downarrow 0} P(t, x, A) = P(0, x, A) = I_A(x), \qquad x \in E,\ A \in \mathscr{E}. \tag{3.1}$$

Associated with each jump process $P(t, x, \mathrm{d}y)$, we have a q-pair $(q(x), q(x, \mathrm{d}y))$:

$$\frac{\mathrm{d}}{\mathrm{d}t}P(t, x, A)\Big|_{t=0} = q(x, A) - q(x)I_A(x). \tag{3.2}$$

Unless otherwise stated, we assume that the q-pair is regular. That is, the q-pair is conservative:

$$0 \leqslant q(x, A) \leqslant q(x, E) = q(x) < \infty, \qquad x \in E,\ A \in \mathscr{E},$$

and there is precise one jump process $P(t, x, \mathrm{d}y)$ satisfying (3.2). Moreover, assume that π is an invariant measure of $P(t, x, \mathrm{d}y)$.

Under the above conditions, it is known that the semigroup $\{P(t)\}_{t\geqslant 0}$ induced by jump process $P(t, x, \mathrm{d}y)$ satisfies the hypotheses given at the beginning of the previous section (cf. Chen (1987)).

Define

$$\pi_q(\mathrm{d}x,\mathrm{d}y)=\pi(\mathrm{d}x)q(x,\mathrm{d}y)\quad\text{on }\mathscr{E}\times\mathscr{E},$$
$$D^*(f)=\frac{1}{2}\int\pi_q(\mathrm{d}x,\mathrm{d}y)(f(y)-f(x))^2,$$
$$\mathscr{D}(D^*)=\big\{f\in L^2(\pi):D^*(f)<\infty\big\};$$
$$\mathscr{K}=\big\{f\in L^\infty(\pi):\overline{\mathrm{supp}(f)}\text{ is compact}\big\}$$

and

$$\mathscr{K}_L=\{g=cf+d:f\in\mathscr{K},\ c,d\in\mathbb{R}\}.$$

Suppose that

(3.3). $q(x)$ is locally bounded.

Then we have

(3.4) **Lemma.** Under (3.3), $\mathscr{K}_L\subset\mathscr{D}(D)$.

Proof. By the regularity of the q-pair, it follows that (3.2)holds for all indicators I_A, $A\in\mathscr{E}$, and hence for all bounded $\mathscr{E}$-measurable functions. Thus, we have

$$\lim_{t\downarrow0}\int_{E\setminus\{x\}}P(t,x,\mathrm{d}y)f(y)=\int q(x,\mathrm{d}y)f(y),\tag{3.5}$$

On the other hand, since

$$\left|\int_{E\setminus\{x\}}P(t,x,\mathrm{d}y)f(y)\right|\leqslant\frac{1}{t}(1-P(t,x,\{x\}))\sup_y|f(y)|\leqslant q(x)\sup_y|f(y)|,$$

it follows that

$$\begin{aligned}&\int\pi(\mathrm{d}x)f(x)\frac{1}{t}[f(x)-P(t)f(x)]\\&\quad=\int_{\mathrm{supp}(f)}\pi(\mathrm{d}x)f(x)^2\frac{1-P(t,x,\{x\})}{t}-\int_{\mathrm{supp}(f)}\pi(\mathrm{d}x)f(x)\frac{P(t,x,\mathrm{d}y)}{t}f(y)\\&\quad\to\int_{\mathrm{supp}(f)}\pi(\mathrm{d}x)q(x)f(x)^2-\int_{\mathrm{supp}(f)}\pi(\mathrm{d}x)f(x)\int q(x,\mathrm{d}y)f(y)\quad\text{as }t\downarrow0\end{aligned}$$

(cf. Chen (1986)). Note that π is an invariant measure of $\{P(t)\}_{t\geqslant0}$:

$$\int\pi(\mathrm{d}x)q(x)f^2(x)=\int\pi(\mathrm{d}x)\int q(x,\mathrm{d}y)f(y)^2.\tag{3.6}$$

Combining the above facts, we arrive at

$$\left(\frac{f-P(t)f}{t},f\right)\to D(f)=D^*(f)<\infty\quad\text{as }t\downarrow0$$

for $f\in\mathscr{K}$. Now, the conclusion follows from Lemma (2.5). □

This simple result already enables us to get an upper bound for $\mathrm{gap}(D)$.

(3.7) **Theorem.** Under (3.3), we have

$$\operatorname{gap}(D) \leqslant \frac{1}{2} \inf \left\{ \frac{\pi_q(K \times K^c + K^c \times K)}{\pi(K)\pi(K^c)} : 0 < \pi(K) < 1,\ K \text{ is compact} \right\}.$$

Proof. For a compact K, $0 < \pi(K) < 1$, set $f = cK + d$. Choose c and $d \in \mathbb{R}$ such that $\pi(f) = 0$ and $\|f\| = 1$. Compute $D^*(f)$. The assertion follows from Lemma (3.4). $\square$

(3.8) **Definition.** We say that $\mathscr{C} \subset \mathscr{D}(D^*)$ is a core of D^*, if for every $f \in \mathscr{D}(D^*)$, there exists a sequence $\{f_n\}_1^\infty$ such that

$$D_1^*(f_n - f) := D^*(f_n - f) + \|f_n - f\|^2 \to 0 \quad \text{as } n \to \infty.$$

(3.9) **Lemma.** If

$$\int \pi(\mathrm{d}x) q(x) < \infty, \tag{3.10}$$

then $\mathscr{K}_L$ is a core of D^*.

Proof. We need only to show that $\mathscr{K}$ is a core of D^*. Take a sequence of compacts $E_n \uparrow E$ and let $f \in \mathscr{D}(D^*)$. Set $f_n = f I_{E_n}$. Then

$$\begin{aligned}
D_1^*(f_n - f) &= \frac{1}{2} \int \pi_q(\mathrm{d}x, \mathrm{d}y) \big(f(y) - f(x) - f_n(y) + f_n(x)\big)^2 + \|f_n - f\|^2 \\
&\leqslant \int \pi_q(\mathrm{d}x, \mathrm{d}y) \big[(f_n(y) - f_n(x))^2 + (f(x) - f_n(x))^2\big] + \|f_n - f\|^2 \\
&= \int_{\{f(y) > n\}} \pi_q(\mathrm{d}x, \mathrm{d}y) f(y)^2 + \int_{\{f(x) > n\}} \pi_q(\mathrm{d}x, \mathrm{d}y) f(x)^2 + \|f_n - f\|^2 \\
&\leqslant \Big[\sup_x |f(x)|\Big]^2 \Big[\int \pi(\mathrm{d}x) q(x, [f > n]) + \int_{\{f(x) > n\}} \pi(\mathrm{d}x) q(x)\Big] \\
&\qquad + \|f_n - f\|^2 \\
&\to 0 \quad \text{as } n \to \infty. \quad \square
\end{aligned}$$

(3.11) **Theorem.** If $\mathscr{K}$ is a core of D^*, in particular, if (3.10) holds, then

$$\begin{aligned}
\operatorname{gap}(D) &= \frac{1}{2} \inf \left\{ \int \pi_q(\mathrm{d}x, \mathrm{d}y)(f(y) - f(x))^2 : \pi(f) = 0,\ \|f\| = 1 \right\} \\
&= \frac{1}{2} \inf \left\{ \int \pi_q(\mathrm{d}x, \mathrm{d}y)(f(y) - f(x))^2 : f \in \mathscr{K}_L,\ \pi(f) = 0,\ \|f\| = 1 \right\}.
\end{aligned}$$

Proof. First, $D^*(f_n - f) \to 0$ implies that

$$\int \pi_q(\mathrm{d}x, \mathrm{d}y)(f_n(y) - f_n(x))^2$$

is bounded with respect to n. On the other hand,

$$\begin{aligned}
&\left| \int \pi_q(\mathrm{d}x,\mathrm{d}y)(f_n(y)-f_n(x))^2 - \int \pi_q(\mathrm{d}x,\mathrm{d}y)(f(y)-f(x))^2 \right| \\
&\quad \leqslant \frac{1}{2}\int \pi_q(\mathrm{d}x,\mathrm{d}y)(f_n(y)-f_n(x)-f(y)+f(x))^2 \\
&\qquad \times \int \pi_q(\mathrm{d}x,\mathrm{d}y)(f_n(y)-f_n(x)+f(y)-f(x))^2 \\
&\quad \leqslant CD^*(f_n-f) \to 0 \quad \text{as } n\to\infty,
\end{aligned}$$

and so

$$D(f_n) = D^*(f_n) \to D^*(f) \quad \text{as } n\to\infty.$$

Our assertion follows from Theorem (2.4), Lemma (3.4) and (3.9) immediately. □

The next result is roughly a special case of Corollary (2.9). It shows that if π and $q(x,\cdot)$ have a density with respect to a reference measure, we can avoid the condition (3.10).

(3.12) **Theorem.** Let E be countable and $Q=(q_{ij})$ be an irreducible regular Q-matrix. The Markov chain $(P_{ij}(t))$, determined by the Q-matrix $Q=(q_{ij})$ has an invariant probability measure (π_i). Then

$$\begin{aligned}
\mathrm{gap}(D) &= \frac{1}{2}\inf\left\{ \sum_{i,j} \pi_i q_{i,j}(f_j-f_i)^2 : \pi(f)=0,\ \|f\|=1 \right\} \\
&= \frac{1}{2}\inf\left\{ \sum_{i,j} \pi_i q_{i,j}(f_j-f_i)^2 : f\in\mathscr{K},\ \pi(f)=0,\ \|f\|=1 \right\}.
\end{aligned}$$

Proof. By Theorem (3.11), we need only to prove that $\mathscr{K}_L$ is a core D^*. Define

$$\hat{q}_{ij} = \frac{\pi_j q_{ji}}{\pi_i}, \qquad \bar{q}_{ij} = \frac{1}{2}(q_{ij}+\hat{q}_{ij}), \qquad i,j\in E.$$

It is easy to check that $(\hat{q}_{ij})$ is a conservative Q-matrix with stationary measure (π_i), and so is $(\bar{q}_{ij})$. Moreover, $(\bar{q}_{ij})$ is a reversible Q-matrix with respect to the same probability measure (π_i), and so its Dirichlet form is regular (cf. Chen (1989); Theorem (3.10)). That is, $\mathscr{K}$ a core of $\overline{D}$.[2] However,

$$\begin{aligned}
\overline{D}(f,f) &= \frac{1}{2}\sum_{i,j}\pi_i\bar{q}_{ij}(f_j-f_i)^2 \\
&= \frac{1}{4}\sum_{i,j}\pi_i(q_{ij}+\hat{q}_{ij})(f_j-f_i)^2 \\
&= \frac{1}{2}\sum_{i,j}\pi_i q_{ij}(f_j-f_i)^2 \\
&= D^*(f);
\end{aligned}$$

[2]Correction: the regularity of (q_{ij}) implies the one of $(\hat{q}_{ij})$, but it is still open to imply the regularity of $(\bar{q}_{ij})$. Thus, one has to use the last sentence as an assumption. For more careful discussion, see the author's paper "Equivalence of exponential ergodicity and L^2-exponential convergence for Markov chains" collected in this book.

hence claim that $\mathscr{K}$ is also a core of D^*. $\square$

The above result is due to a simple observation[3]:

$$\text{Re spec.}(\Omega) = \text{spec.}\left(\frac{1}{2}(\Omega + \widehat{\Omega})\right)$$

where $\widehat{\Omega}$ is the adjoint operator of Ω.

(3.13) **Example.** Take

$$Q = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix}$$

Then $\text{gap}(D) = 3/2$ and the eigenvalues of Ω are $0,\ -3/2 \pm \sqrt{3}\, i/2$.

(3.14) **Example.** Take $E = \{0, 1, 2, \dots\}$,

$$q_{0i} = b_i,\ i \geqslant 1, \quad q_0 = \sum_{i=1}^{\infty} b_i < \infty,$$
$$q_{i0} = q_i,\ i \geqslant 1, \quad q_{ij} = 0 \text{ otherwise.}$$

Then

$$\pi_i = \mu_i/\rho, \qquad i \geqslant 0,$$

where

$$\mu_0 = 1, \quad \mu_i = b_i/q_i,\ i \geqslant 1; \quad \rho = \sum_{i=1}^{\infty} b_i/q_i + 1.$$

For every $f \in L^2(\pi)$, $\pi(f) = 0$, $\|f\| = 1$, we have

$$\begin{aligned} 1 &= \frac{1}{2}\sum_{i,j} \pi_i \pi_j (f_j - f_i)^2 \\ &\leqslant \sum_{i,j} \pi_i \pi_j [(f_j - f_0)^2 + (f_i - f_0)^2] \\ &= 2\sum_{i \neq 0} \pi_i (f_i - f_0)^2 \\ &\leqslant 2\sum_{i \neq 0} \pi_i q_{i0} (f_i - f_0)^2 \Big/ \inf_{j \geqslant 1} q_j \\ &= \frac{1}{2}\sum_{i,j} \pi_i q_{ij} (f_j - f_i)^2 \cdot 2 \Big/ \inf_{k \geqslant 1} q_k. \end{aligned}$$

Thus,

$$\text{gap}(D) \geqslant \frac{1}{2} \inf_{i \geqslant 1} q_i.$$

Now, we study two comparison theorems to close this section.

[3]Correction: the Re spec. or spec. here should be replaced by gap since Re spec $\neq$ gap in general.

(3.15) **Theorem.** Let $Q = (q_{ij})$ and $\widetilde{Q} = (\tilde{q}_{ij})$ be two Q-matrices as in (3.12). Denote by (π_i) and $(\tilde{\pi}_i)$ their invariant probability measures respectively. Suppose that

$$\tilde{q}_{ij} \geqslant b\, q_{ij}, \qquad i \neq j$$

for some constant $b > 0$ and

$$c \leqslant \tilde{\pi}_i/\pi_i \leqslant c^{-1}, \qquad i \in E$$

for a constant $c \in (c_0, 1]$,

$$c_0 = \frac{1}{3}\left[1 + \left(\frac{3\sqrt{69}+11}{2}\right)^{1/3} - \left(\frac{3\sqrt{69}-11}{2}\right)^{1/3}\right] \approx 0.56984.$$

Then

$$\operatorname{gap}(\widetilde{D}) \geqslant \frac{b}{c}\,[c^3 - (1-c)^2]\operatorname{gap}(D).$$

Proof. Let $f \in \mathscr{K}_L$, $\tilde{\pi}(f) = 0$, $\|f\|_{\tilde{\pi}} = 1$. Then

$$\begin{aligned}
|\pi(f)|^2 &= \left|\sum_j \tilde{\pi}_j(1-\pi_j/\tilde{\pi}_j)f_j\right|^2 \\
&\leqslant \sum_j \tilde{\pi}_j(1-\pi_j/\tilde{\pi}_j)^2 \sum_k \tilde{\pi}_k f_k^2 \\
&= \sum_j \tilde{\pi}_j(1-\pi_j/\tilde{\pi}_j)^2 \\
&\leqslant \sup_j (1-\pi_j/\tilde{\pi}_j)^2 \\
&\leqslant (c^{-1}-1)^2.
\end{aligned}$$

Hence

$$\begin{aligned}
\frac{1}{2}\sum_{i,j}\tilde{\pi}_i\tilde{q}_{ij}(f_j-f_i)^2 &\geqslant \frac{bc}{2}\sum_{i,j}\pi_i q_{ij}(f_j-f_i)^2 \\
&\geqslant bc\operatorname{gap}(D)\,\|f-\pi(f)\|_\pi^2 \\
&= bc\operatorname{gap}(D)\,[\|f\|_\pi^2 - \pi(f)^2] \\
&\geqslant \frac{b}{c}\,[c^3-(1-c)^2]\operatorname{gap}(D). \quad \square
\end{aligned}$$

For Markov chains, a problem—exponential ergodicity has been well studied. It is known that for every irreducible Markov chain $(P_{ij}(t))$, there is an $\alpha \geqslant 0$ such that

$$|P_{ij}(t) - \pi_j| = O(\exp(-\alpha t)) \quad \text{as } t \to \infty, \tag{3.16}$$

where $\pi_j = \lim_{t\to\infty} P_{ij}(t)$. Set

$$\hat{\alpha} = \sup\{\alpha : (3.16) \text{ holds for all } i \text{ and } j\}. \tag{3.17}$$

If $\hat{\alpha} > 0$, the process is called exponentially ergodic.

(3.18) **Theorem.** Let $(P_{ij}(t))$ be an irreducible positive recurrence Markov chain with stationary distribution (π_i) and Q-matrix $Q = (q_{ij})$. Then

$$\operatorname{gap}(D) \leqslant \hat{\alpha}. \tag{3.19}$$

Proof. Fix $i_0, j_0 \in E$ and take

$$f_j = \delta_{jj_0}, \qquad j \in E.$$

Then

$$e^{-2\sigma t}\|f - \pi(f)\| \geqslant \|P(t)f - \pi(f)\|^2 \geqslant \pi_{i_0}|P_{i_0j_0}(t) - \pi_{j_0}|^2.$$

Since i_0 and j_0 are arbitrary, we obtain

$$\operatorname{gap}(D) = \sigma \leqslant \hat{\alpha}. \quad \square$$

For birth-death processes, we will prove in Section 5 that (3.19) is indeed an equality.

4. Reversible Case, an Approximation Theorem.

In view of Theorem (3.12), in some cases, we can reduce the non-symmetric case to a symmetric one. Hence the symmetric case is more important and often easier to handle.

Throughout this section, we assume (3.3).

For a bounded q-pair, some nice results were obtained by Lawler and Sokal (1988). The purpose of this section is to reduce the unbounded case to a bounded one. To do this, take compact sets $E_n \uparrow E$ $(n \geqslant 0)$. Assume that

$$\pi(E_n^c) > 0, \qquad n \geqslant 0. \tag{4.1}$$

Regard $\Delta_n = E_n^c$ as a single point and set

$$\begin{aligned}
\widehat{E}_{n+1} &= E_n \cup \{\Delta_n\}, \qquad \widehat{\mathscr{E}}_{n+1} = \sigma\big(\mathscr{E} \cap (E_n \cup \{\Delta_n\})\big),\\
\hat{q}_{n+1}(x, A) &= q(x, A \cap E_n) + I_A(\Delta_n)q(x, E_n^c), \qquad x \in E_n,\ A \in \mathscr{E}_{n+1},\\
\hat{q}_{n+1}(\Delta_n, A) &= \frac{1}{\pi(E_n^c)} \int_{A \cap E_n} \pi(\mathrm{d}x)q(x, E_n^c), \qquad A \in \mathscr{E}_{n+1},\\
\hat{q}_{n+1}(x) &= \hat{q}_{n+1}\big(x, \widehat{E}_{n+1}\big), \qquad x \in \widehat{E}_{n+1},\ n \geqslant 0.
\end{aligned}$$

It is easy to see that $\big(\hat{q}_{n+1}(x), \hat{q}_{n+1}(x, \mathrm{d}y)\big)$ is a bounded conservative q-pair, and hence is regular. Finally. let

$$\hat{\pi}_{n+1}(A) = \pi(A \cap E_n) + \pi(E_n^c)I_A(\Delta_n), \qquad A \in \widehat{\mathscr{E}}_{n+1}.$$

From the reversibility of $(q(x), q(x, \mathrm{d}y)$ with respect to π, we obtain

$$\int_A \pi(\mathrm{d}x)q(x, B) = \int_B \pi(\mathrm{d}x)q(x, A), \qquad A, B \in \mathscr{E}.$$

For all $A, B \in \widehat{\mathscr{E}}_{n+1}$, we have

$$\begin{aligned}\int_A \hat{\pi}_{n+1}(\mathrm{d}x)\hat{q}_{n+1}(x,B) &= \int_{A\cap E_n} \pi(\mathrm{d}x)[q(x,B\cap E_n) + I_B(\Delta_n)q(x,E_n^c)] \\ &\quad + I_A(\Delta_n)\pi(E_n^c)\hat{q}_{n+1}(\Delta_n,B) \\ &= \int_{A\cap E_n} \pi(\mathrm{d}x)q(x,B\cap E_n) + I_B(\Delta_n)\int_{A\cap E_n}\pi(\mathrm{d}x)q(x,E_n^c)] \\ &\quad + I_A(\Delta_n)\int_{A\cap E_n}\pi(\mathrm{d}x)q(x,E_n^c).\end{aligned}$$

This is symmetric with respect to A and B. Therefore $(\hat{q}_{n+1}(x), \hat{q}_{n+1}(x,\mathrm{d}y))$ is reversible with respect to $\hat{\pi}_{n+1}$.

Next, for $f \in \mathscr{K}_L$, we have

$$\begin{aligned}\int \pi_q(\mathrm{d}x,\mathrm{d}y)(f(y)-f(x))^2 &= \int_{E_n}\pi(\mathrm{d}x)\int_{E_n} q(x,\mathrm{d}y)(f(y)-f(x))^2 \\ &\quad + 2\int_{E_n}\pi(\mathrm{d}x)\int_{E_n^c} q(x,\mathrm{d}y)(c-f(x))^2 \\ &\qquad (\text{if } f = \text{a constant } c \text{ off } E_n) \\ &= \int_{E_n}\pi_{n+1}(\mathrm{d}x)\int_{E_n}\hat{q}_{n+1}(x,\mathrm{d}y)(f(y)-f(x))^2 \\ &\quad + 2\int_{E_n}\hat{\pi}_{n+1}(\mathrm{d}x)\hat{q}_{n+1}(x,\Delta_n)(c-f(x))^2 \\ &= \int_{\widehat{E}_{n+1}}\hat{\pi}_{n+1}(\mathrm{d}x)\int_{\widehat{E}_{n+1}}\hat{q}_{n+1}(x,\mathrm{d}y)(f(y)-f(x))^2.\end{aligned}$$

By Theorem (3.11), we have

$$\begin{aligned}\mathrm{gap}(D) &= \inf\big\{D^*(f) : \pi(f)=0,\ \|f\|=1,\ f=\text{constant off } E_n \text{ for some } n\geqslant 0\big\} \\ &= \lim_{n\to\infty}\downarrow \inf\big\{D^*(f) : \pi(f)=0,\ \|f\|=1,\ f=\text{constant off } E_n\big\} \\ &= \lim_{n\to\infty}\downarrow \inf\big\{D^*(f) : \hat{\pi}_{n+1}(f)=0,\ \hat{\pi}_{n+1}(f^2)=1\big\} \\ &= \lim_{n\to\infty}\downarrow \mathrm{gap}\big(\widehat{D}_{n+1}\big),\end{aligned}$$

where $\lim_{n\to\infty}\downarrow h_n = h$ meas that $h_n \downarrow h$ as $n\to\infty$. Thus, we have proved the following approximation result:

(4.2) **Theorem.** Let $(q(x), q(x,\mathrm{d}y))$ be a regular q-pair which is reversible with respect to π. Take compacts $E_n \uparrow E$ and assume that $\pi(E_n^c) > 0$ for all $n \geqslant 0$. Define $(\hat{q}_{n+1}(x), \hat{q}_{n+1}(x,\mathrm{d}y))$ on $\widehat{E}_{n+1}$ as above. If $\mathscr{K}_L$ is a core of D^*, then

$$\mathrm{gap}(D) = \lim_{n\to\infty}\downarrow \mathrm{gap}\big(\widehat{D}_{n+1}\big).$$

In particular, we have

(4.3) **Corollary.** Let $E = \{0, 1, 2, \dots\}$, $Q = \{q_{ij}\}$ be an irreducible regular Q-matrix which is reversible with respect to (π_i). Take

$$\widehat{Q}_{n+1} = \begin{pmatrix} -q_0 & q_{01} & \cdots & q_{0n} & \sum_{j>n} q_{0j} \\ q_{10} & -q_1 & \cdots & q_{1n} & \sum_{j>n} q_{1j} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ q_{n0} & q_{n1} & \cdots & -q_n & \sum_{j>n} q_{nj} \\ \hat{q}_{n+1,0} & \hat{q}_{n+1,1} & \cdots & \hat{q}_{n+1,n} & -\hat{q}_{n+1} \end{pmatrix}, \tag{4.4}$$

where

$$\hat{q}_{n+1,j} = \pi_j \sum_{k>n} q_{jk} \Big/ \sum_{k>n} \pi_k, \quad j = 0, 1, \dots, n, \qquad \hat{q}_{n+1} = \sum_{j=0}^{n} \hat{q}_{n+1,j}.$$

Then

$$\operatorname{gap}(D) = \lim_{n\to\infty} \downarrow \operatorname{gap}\big(\widehat{D}_{n+1}\big).$$

5. Spectral Gap for Markov Chains.

Again, we need only to consider the reversible ease.

Let $(P_{ij}(t))$ be an irreducible reversible Markov chain with stationary distribution (π_i) and regular Q-matrix $Q = (q_{ij})$. Suppose that $q_{i,i+1} > 0\,(i \in E)$. For the lower bound of the spectral gap, we have

(5.1) **Theorem.** If there exist constants $b, c > 0$ such that

$$\sum_{j>i} \pi_j \leqslant c\,\pi_i q_{i,i+1}, \qquad i \in E,$$
$$\sum_{j>i} \pi_j q_{j,j+1} \leqslant b\,\pi_i q_{i,i+1}, \qquad i \in E,$$

then

$$\operatorname{gap}(D) \geqslant \frac{l}{c(\sqrt{b+1} + \sqrt{b})^2} > \frac{1}{2c(1+2b)}.$$

(5.2) **Theorem.** If there exist constants $b, c > 0$ such that

$$\sum_{j\geqslant i} \pi_j \leqslant c\,\pi_i, \qquad i \in E,$$
$$\pi_{i+1} \leqslant b\,\pi_i q_{i,i+1}, \qquad i \in E,$$

then

$$\operatorname{gap}(D) \geqslant \frac{1}{2bc(1+2c)} > \frac{1}{4bc^2}.$$

The first theorem was proved by Liggett (1989) under two more assumptions: $\mathscr{K}_L$ is a core of the generator Ω and $\sum_i \pi_i q_i < \infty$. The second one was proved by Sullivan (1984) under two more assumptions: $\inf_{i\geqslant 1} q_{i,i+1} > 0$ and $\sup_i q_i < \infty$.

By Theorem (3.12), it is easy to check that Liggett's proof still works for the above theorems. We omit the details here.

The above results are incomparable. For example, consider the birth-death process:

$$q_{i,i+1} = \alpha, \qquad q_{i,i-1} = \beta, \qquad \beta > \alpha > 0.$$

If $\alpha \geqslant 1$, then (5.1) is better than (5.2). If $\alpha < 1$, (5.2) can be better than (5.1).

Actually, Theorems (5.1) and (5.2) are based on comparing the original process with the birth-death process :

$$\tilde{q}_{ij} = \begin{cases} q_{ij}, & \text{if } j = i+1 \\ 0, & \text{other } j \neq i \end{cases}, \qquad \tilde{q}_i = \sum_{j \neq i} \tilde{q}_{ij}, \qquad \tilde{\pi}_i = \pi_i.$$

The main part of the proofs for (5.1) and (5.2) is to show that the lower bounds hold for this birth-death process, and then to apply Theorem (3.5) to deduce our assertions. This induces us to study more carefully the spectral gap for birth-death processes.

Let $Q = (q_{ij})$ be a birth-death Q-matrix : Set

$$\begin{aligned} q_{i,i+1} &= b_i > 0, && i \geqslant 0, \\ q_{i,i-1} &= a_i > 0, && i \geqslant 1, \\ q_i &= -q_{ii} = a_i + b_i, && i \geqslant 0. \end{aligned}$$

Set

$$\mu_0 = 1, \quad \mu_i = \frac{b_0 \cdots b_{i-1}}{a_1 \cdots a_i},\ i \geqslant 1, \qquad \rho = 1 + \sum_{i=1}^{\infty} \mu_i.$$

Then

$$\pi_i = \mu_i / \rho, \qquad i \geqslant 0.$$

The next result is an improvement over Theorem (3.18) in the existing circumstances.

(5.3) **Theorem.** For every positive recurrent birth-death process, the exponential L^2- convergence is equivalent to the exponential ergodicity. In other words,

$$\operatorname{gap}(D) = \hat{\alpha}.$$

Proof. If $\hat{\alpha} = 0$, then by (3.19), $\operatorname{gap}(D) = 0$. Thus, we may and will assume that $\hat{\alpha} > 0$. Set

$$\begin{aligned} H_0(x) &= 1, \\ -xH_0(x) &= -b_0 H_0(x) + b_0 H_l(x), \\ -xH_n(x) &= a_n H_{n-l}(x) - (a_n + b_n) H_n(x) + b_n H_{n+l}(x), \qquad n \geqslant l,\ x \in \mathbb{R}. \end{aligned}$$

Then $H_n(0) = 1$, $n \geqslant 0$. Recall the Karlin and Mcgregor's representation theorem:

$$P_{ij}(t) = \mu_j \int_0^\infty e^{-xt} H_i(x) H_j(x) \mathrm{d}\psi(x), \tag{5.4}$$

where ψ is a (unique) non-decreasing function which is left continuous and

$$\psi(x) = 0 \quad \text{for } x \leqslant 0, \qquad \psi(x) \to 1 \quad \text{as } x \to \infty.$$

Also,

$$\mu_j \int_0^\infty H_i(x) H_j(x) \mathrm{d}\psi(x) = \delta_{ij}.$$

Write

$$\hat{f}(x) = \sum_i \pi_i H_i(x) f_i, \qquad f \in \mathscr{K}.$$

From (5.4), it follows that

$$(f, P(t)f) = \rho \int_0^\infty e^{-xt} \hat{f}(x)^2 \mathrm{d}\psi(x), \qquad f \in \mathscr{K}. \tag{5.5}$$

In particular,

$$(f, f) = \rho \int_0^\infty \hat{f}(x)^2 \mathrm{d}\psi(x), \qquad f \in \mathscr{K}. \tag{5.6}$$

This gives us an isometric imbedding from $L^2(\pi)$ to $L^2([0,\infty], \rho \mathrm{d}\psi)$. Thus, (5.5) and (5.6) hold for $f \in L^2(\pi)$. Moreover, by (5.5), we see that

$$D(f, f) = \rho \int_0^\infty x \hat{f}(x)^2 \mathrm{d}\psi(x).$$

From the exponential ergodicity, by Van Doorn (1985), Theorem 2.1, Theorem 3.1 and Lemma 3.2, the first two points of the spectrum of ψ are

$$x_1 = 0, \qquad \hat{\alpha} = x_2 > x_1$$

(x is called a point of the spectrum of ψ if $\Delta\psi(x) = \psi(x+) - \psi(x-) > 0$). Notice that $\Delta\psi(0) > 0$, and so[4]

$$\hat{f}(0) = \sum_i H_i(0) f_i = \pi(f).$$

---

[4]By the isometry of $L^2(\pi)$ and $L^2([0,\infty), \rho \mathrm{d}\psi)$, for each $f \in L^2(\pi)$, if we set

$$f_i^{(n)} = \begin{cases} f_i, & \text{if } i \leqslant n \\ 0, & \text{if } i > n, \end{cases}$$

then $\|f^{(n)} - f\| \to 0$, and so

$$\hat{f}^{(n)} = \sum_i \pi_i H_i f_i^{(n)} \to \hat{f} \quad \text{in } L^2([0,\infty), \rho \mathrm{d}\psi).$$

In particular, one can choose a subsequence $\{n_k\}$ such that $\hat{f}^{(n_k)}$ converges to $\hat{f}$ almost everywhere respect to $\rho \mathrm{d}\psi$. So we have

$$\hat{f}(0) = \lim_{k\to\infty} \hat{f}^{n_k}(0) = \lim_{k\to\infty} \sum_i \pi_i H_i(0) f_i^{n_k} = \lim_{k\to\infty} \sum_i \pi_i f_i^{n_k} = \pi(f).$$

This footnote was included in the author's book (1992; 1'st ed. but not 2'nd ed.).

On the other hand, from $x_2 = \hat{\alpha}$ and

$$P_{00}(t) - \pi_0 = \int_0^\infty e^{-xt} \mathrm{d}\psi'(x),$$

where $\psi' = \psi - \Delta\psi(0)$, it follows that

$$\psi(x_2-) = \psi(0+).$$

Hence

$$\begin{aligned}
\mathrm{gap}(D) &= \inf\{D(f,f) : \pi(f) = 0,\ \|f\| = 1\} \\
&= \inf\left\{\rho \int_0^\infty x\hat{f}(x)^2 \mathrm{d}\psi(x) : \pi(f) = 0,\ \|f\| = 1\right\} \\
&= \inf\left\{\rho \int_0^\infty x\hat{f}(x)^2 \mathrm{d}\psi(x) : \hat{f}(0) = 0,\ \|f\| = 1\right\} \\
&\geqslant (x_2 - \varepsilon) \inf\left\{\rho \int_0^\infty x\hat{f}(x)^2 \mathrm{d}\psi(x) : \hat{f}(0) = 0,\ \|f\| = 1\right\} \\
&= (x_2 - \varepsilon) \quad \text{(by (5.6))} \\
&= \hat{\alpha} - \varepsilon,
\end{aligned}$$

for all small $\varepsilon > 0$. Therefore $\mathrm{gap}(D) > \hat{\alpha}$. $\square$

Now, we can combine Theorem (5.3) with the previous results (cf. Van Doom (1981)) to give some examples.

(5.7) **Examples.**

(i) Let $b_i = b$ for $i \geqslant O$, $a_i = ai$ for $i = 0, 1, \ldots, s-1$, and $= sa$ for $i = s, s+1, \cdots$, where the parameters satisfy $a/sb =: \rho < 1$. Then there exists a $\bar{\rho} < 1$ such that

$$\begin{aligned}
0 < \mathrm{gap}(D) < b\big(1 - 1/\sqrt{\rho}\big)^2 \quad &\text{if } \rho < \bar{\rho}, \\
\mathrm{gap}(D) = b\big(1 - 1/\sqrt{\rho}\big)^2 \quad &\text{if } \rho \geqslant \bar{\rho}.
\end{aligned}$$

If $s = 1$ and $b < a$, then

$$\mathrm{gap}(D) = \big(\sqrt{a} - \sqrt{b}\big)^2.$$

(ii) Let $b_i = b/(i+1)$ for $i \geqslant 0$ and $a_i = a$ for $i \geqslant 1$. Then

$$\mathrm{gap}(D) = a - \big(\sqrt{b^2 + 4ab} - b\big)/2.$$

(iii) Let $b_i = \alpha + \lambda_1$ for $i \geqslant 0$ and $a_i = \lambda_2 i$ for $i \geqslant 1$, where $\alpha > 0$ and $\lambda_2 > \lambda_1 \geqslant 0$. Then

$$\mathrm{gap}(D) = \lambda_2 - \lambda_1.$$

Because of Theorem (5.3), we can also rely on some sufficient conditions for the exponential ergodicity to estimate the lower bound of $\mathrm{gap}(D)$. Note that in many cases, it is not possible to compute the spectral function ψ. We would like to know some practical methods to estimate $\mathrm{gap}(D)$. Our next result is such a kind of approach without using ψ. The idea is based on Corollary (4.3). In the present case, our approximation Q-matrix (4.4) becomes

$$\widehat{Q}_{n+1} = \begin{pmatrix} -b_0 & b_0 & 0 & & \\ a_1 & -(a_1+b_1) & b_1 & & \mathbf{0} \\ & \ddots & \ddots & \ddots & \\ & & a_n & -(a_n+b_n) & b_n \\ \mathbf{0} & & 0 & \hat{a}_{n+1} & -\hat{a}_{n+1} \end{pmatrix},$$

where $\hat{a}_{n+1} = \pi_n b_n / \hat{\pi}_{n+1}$, $\hat{\pi}_{n+1} = \sum_{j>n} \pi_j$, $n \geqslant 0$. For each fixed n, define

$$\begin{aligned} s_0(x) &= b_0 + x, \qquad x \in \mathbb{R}, \\ s_1(x) &= \begin{cases} a_1 + b_1 + x - a_1 b_0 / s_0(x) & \text{if } s_0(x) \neq 0, \\ 1 & \text{if } s_0(x) = 0, \end{cases} \\ s_i(x) &= \begin{cases} a_i + b_i + x - a_i b_{i-1} / s_{i-1}(x) & \text{if } s_{i-1}(x) \neq 0,\ s_{i-2}(x) \neq 0, \\ a_i + b_i + x & \text{if } s_{i-2}(x) = 0, \\ 1 & \text{if } s_{i-1}(x) = 0, \end{cases} \\ & \qquad 2 \leqslant i \leqslant n,\ x \in \mathbb{R}, \end{aligned}$$

and

$$\hat{s}_{n+1}(x) = \begin{cases} x + \dfrac{\pi_n b_n}{\hat{\pi}_{n+1}} \left(1 - \dfrac{b_n}{s_n(x)}\right) & \text{if } s_n(x) \neq 0,\ s_{n-1}(x) \neq 0, \\ x + \dfrac{\pi_n b_n}{\hat{\pi}_{n+1}} & \text{if } s_{n-1}(x) = 0, \\ 1 & \text{if } s_n(x) = 0, \quad x \in \mathbb{R}. \end{cases}$$

(5.8) **Theorem.** For the above $\widehat{Q}_{n+1}$,

$$\mathrm{gap}\big(\widehat{D}_{n+1}\big) \geqslant \alpha > 0 \tag{5.9}$$

if and only if there is precisely one term of

$$\{s_0(-\alpha),\ \ldots,\ s_n(-\alpha),\ s_{n+1}(-\alpha)\}$$

which is less or equal to zero. Moreover, if the condition holds for all n, then

$$\mathrm{gap}(D) \geqslant \alpha > 0 \tag{5.10}$$

Proof. Denote by $\widetilde{Q}_{n+1}$ the symmetrized matrix of $\widehat{Q}_{n+1}$:

$$\widetilde{Q}_{n+1} = \begin{pmatrix} -b_0 & \frac{\sqrt{\pi_0}\, b_0}{\sqrt{\pi_1}} & 0 & \\ \frac{\sqrt{\pi_1}\, a_1}{\sqrt{\pi_0}} & -(a_1+b_1) & \frac{\sqrt{\pi_0}\, b_1}{\sqrt{\pi_2}} & \mathbf{0} \\ \ddots & \ddots & \ddots & \\ & \frac{\sqrt{\pi_n}\, a_n}{\sqrt{\pi_{n-1}}} & -(a_n+b_n) & \frac{\sqrt{\pi_n}\, b_n}{\sqrt{\hat{\pi}_{n+1}}} \\ \mathbf{0} & 0 & \frac{\sqrt{\hat{\pi}_{n+1}}\, \hat{a}_{n+1}}{\sqrt{\pi_n}} & -\hat{a}_{n+1} \end{pmatrix}$$

$$= \begin{pmatrix} -b_0 & \sqrt{a_1 b_0} & 0 & \\ \sqrt{a_1 b_0} & -(a_1+b_1) & \sqrt{a_2 b_1} & \mathbf{0} \\ \ddots & \ddots & \ddots & \\ & \sqrt{a_n b_{n-1}} & -(a_n+b_n) & \frac{\sqrt{\pi_n}\, b_n}{\sqrt{\hat{\pi}_{n+1}}} \\ \mathbf{0} & 0 & \frac{\sqrt{\pi_n}\, b_n}{\sqrt{\hat{\pi}_{n+1}}} & -\frac{\pi_n\, b_n}{\hat{\pi}_{n+1}} \end{pmatrix}$$

Then $\widehat{Q}_{n+1}$ and $\widetilde{Q}_{n+1}$ have the same eigenvalues which are denoted by

$$0 = \lambda_{n+1,0} > \lambda_{n+1,1} > \cdots > \lambda_{n+1,n+1}.$$

These eigenvalues must be distinct since the matrices are tridiagonal. From the matrix theory, it is known that

$$-\mathrm{gap}\big(D_{n+1}\big) = \lambda_{n+1,1} < -\alpha$$

if and only if there is precisely one non-positive term among

$$\{s_0(-\alpha),\ \ldots\ ,\ s_n(-\alpha),\ \hat{s}_{n+1}(-\alpha)\}.$$

This proves the first assertion. The second follows from the first one plus Corollary (4.3). □

To show that Theorem (5.8) is feasible, let us consider

(5.11) **Example.** Take $b_i = b > 0$, $i \geqslant 0$; $a_i = ia > 0$, $i \geqslant 1$. For a special case that $b = 1$ and $a = 2$. the bound obtained by Theorem (5.1) is 0.3348. But we have seen in Example (5.7) that $\mathrm{gap}(D) = 2$. Now, we use Theorem (5.8) to show that

$$\mathrm{gap}(D) \geqslant \hat{\alpha} = a > 0.$$

To do this, assume that

$$b/a \neq 1, 2, \ldots$$

for simplicity. The exceptional cases can be discussed in a similar way. Now,

$$\pi_i = \left(\frac{b}{a}\right)\frac{1}{\rho\, i!}, \qquad \rho = \exp[b/a].$$

By induction, it is easy to prove that

$$\begin{aligned} s_0(-a) &= b - a, \\ s_i(-a) &= \frac{b(b-(i+1)a)}{b-ia}, \qquad 1 \leqslant i \leqslant n, \end{aligned}$$

and so

$$\hat{s}_{n+1}(-a) = -a + \frac{ab\pi_n}{\hat{\pi}_{n+1}((n+1)a-b)}.$$

Since

$$\begin{aligned} \hat{\pi}_{n+1}((n+1)a-b) &= \sum_{j>n}\left(\frac{b}{a}\right)^{j-n}\frac{n!}{j!}\big[\,(n+1)a-b\,\big] \\ &= \left(n+1-\frac{b}{a}\right)b\sum_{j=0}^{\infty}\left(\frac{b}{a}\right)^{j}\frac{n!}{(n+1+j)!} \\ &< \left(n+1-\frac{b}{a}\right)b\sum_{j=0}^{\infty}\left(\frac{b}{a}\right)^{j}\frac{1}{(n+1)^{j+1}} \\ &= b \end{aligned}$$

for large n, we have

$$\hat{s}_{n+1}(-a) > 0 \quad \text{for large} \quad n.$$

Clearly, among

$$\{s_0(-a), \cdots, s_n(-a), \hat{s}_{n+1}(-a)\}$$

there is precisely one negative term. Hence from Theorem (5.8) we may deduce our assertion.

As we have just seen above, for estimating the decay parameter $\hat{\alpha}$, the tridiagonal property of birth-death Q-matrices is very helpful. On the same idea, Van Doom (1985) obtained the following bounds.

(5.12) **Theorem.** For a birth-death process with rates a_i and b_i, the decay parameter satisfies

$$\hat{\alpha} \geqslant \inf_{i\geqslant 1}\big\{a_i + b_{i-1} - \sqrt{a_{i-1}b_{i-1}} - \sqrt{a_i b_i}\,\big\},$$

$$\hat{\alpha} \geqslant \frac{1}{2}\inf_{i\geqslant 1}\big\{a_i + a_{i+1} + b_i + b_{i-1} - \sqrt{(a_{i+1}+b_i-a_i-b_{i-1})^2 + 16a_ib_i}\,\big\},$$

$$\hat{\alpha} \leqslant \inf_{n,k\geqslant 0}\left\{1+\sum_{i=n+1}^{n+k}\left[1-2\left(\frac{a_ib_i}{(a_i+b_{i-1})(a_{i+1}+b_i)}\right)^{1/2}\right]\right\}\left\{\sum_{i=n}^{n+k}\frac{1}{a_{i+1}+b_i}\right\}^{-1},$$

$$\hat{\alpha} \leqslant \frac{1}{2}\inf_{i\geqslant 1}\big\{a_i + a_{i+1} + b_i + b_{i-1} - \sqrt{(a_{i+1}+b_i-a_i-b_{i-1})^2 + 4a_ib_i}\,\big\}.$$

Moreover, if

$$\varliminf_{i\to\infty}\{a_i+b_i-\sqrt{a_ib_{i-1}}-\sqrt{a_{i+1}b_i}\}>0,$$

then $\mathrm{gap}(D)>0$.

Having worked so much on the birth-death processes, now let us return to the general Markov chains. By comparing a given Markov chain with a birth-death process as we explained before, we obtain

(5.13) **Theorem.** Let $(P_{ij}(t))$ be a Markov chain given at the beginning of this section. Then

$$\begin{aligned}\mathrm{gap}(D)&\geqslant\inf_{i\geqslant1}\{q_{i,i-1}+q_{i-1,i}-\sqrt{q_{i-1,i-2}q_{i-1,i}}-\sqrt{q_{i,i-1}q_{i,i+1}}\},\\ \mathrm{gap}(D)&\geqslant\frac{1}{2}\inf_{i\geqslant1}\big\{q_{i,i-1}+q_{i+1,i}+q_{i-1,i}+q_{i,i+1}\\ &\qquad-\big[(q_{i+1,i}+q_{i,i+1}-q_{i,i-1}-q_{i-1,i})^2+16\,q_{i,i-1}q_{i,i+1}\big]^{1/2}\big\}.\end{aligned}$$

Moreover, if

$$\varliminf_{i\to\infty}\{q_{i,i-1}+q_{i,i+1}-\sqrt{q_{i,i-1}q_{i-1,i}}-\sqrt{q_{i+1,i}q_{i,i+1}}\}>0,$$

then $\mathrm{gap}(D)>0$.

6. Non-Positive Recurrent Case.

For the non-positive recurrent case, a Markov process has no finite measure as its invariant measure. Thus, the vector 1 does not belong to $L^2(\pi)$ and so the largest eigenvalue of Ω on $L^2(\pi)$ is meaningful. Indeed, it determines the convergence rate. However, our previous results work well in this situation with a slight modification. For example, as an analogue of Theorem (2.4), we have

$$\begin{aligned}\sigma_0&=\inf_{t>0}\frac{1}{t}\inf\{-\log\|P(t)f\|:\|f\|=1\}\\ &=\inf\{(-\Omega f,f):f\in\mathscr{D}(\Omega)\text{ and }\|f\|=1\}\\ &=\inf\big\{\big(-\widetilde{\Omega}f,f\big):f\in\mathscr{D}\big(\widetilde{\Omega}\big)\text{ and }\|f\|=1\big\}\\ &=\inf\{D(f,f):f\in\mathscr{D}(D)\text{ and }\|f\|=1\}.\end{aligned}$$

Also, we can often reduce the non-symmetric case to a symmetric one.

For jump processes, we allow our q-pair $(q(x),q(x,\mathrm{d}y))$ to be non-conservative:

$$d(x):=q(x)-q(x,E)\geqslant0,\qquad x\in E.$$

Any jump process $P(t,x,\mathrm{d}y)$ with a q-pair $(q(x),q(x,\mathrm{d}y))$ and an excessive measure π (σ-finite),

$$\pi\geqslant\pi P(t),\qquad t\geqslant0$$

will gives us a strongly continuous and contractive semigroup $\{P(t)\}_{t\geqslant 0}$ on $L^2(\pi)$ (cf. Chen (1987), (11)). Of course, $(D^*, \mathscr{D}(D^*))$ given in Section 3 should be replaced by

$$D^*(f) = \int \pi(\mathrm{d}x) f(x) \Big[f(x) q(x) - \int q(x, \mathrm{d}y) f(y) \Big],$$
$$\mathscr{D}(D^*) = \{ f \in L^2(\pi) : D^*(f) < \infty \}.$$

In the symmetric case,

$$D^*(f) = \frac{1}{2} \int \pi_q(\mathrm{d}x, \mathrm{d}y) \big(f(y) - f(x) \big)^2 + \int \pi_d(\mathrm{d}x) f(x)^2,$$

where $\pi_q(\mathrm{d}x, \mathrm{d}y) = \pi(\mathrm{d}x) q(x, \mathrm{d}y)$ and $\pi_d(\mathrm{d}x) = \pi(\mathrm{d}x) d(x)$.

From now on, we consider the symmetric case only.

It is interesting that $\sigma_0 = \lambda_0(\pi)$ which was introduced by Stroock (1981). Several equivalent descriptions of $\lambda_0(\pi)$ were discussed in Stroock (1981). For a related problem, see Chen and Stroock (1983) in which a simple estimate ($\sigma_0 \leqslant \inf_{i \in E} q_i$) was obtained.

Now, suppose that the jump process satisfying the backward Kolmogorov equations is unique. Then the symmetric jump process corresponds to a regular Dirichlet form :

$$D(f, f) = D^*(f)$$

(see Chen (1989), Theorem (3.10)). Actually, this process is just the minimal one. Then

$$\begin{aligned} \sigma_0 &= \inf\{ D(f, f) : \|f\| = 1 \} \\ &= \inf\{ D(f, f) : f \in \mathscr{K} \text{ and } \|f\| = 1 \}. \end{aligned}$$

In particular, take a compact K such that $\pi(K) > 0$ and set $f = I_K / (\pi(K))^{1/2}$; then

$$D(f, f) = \frac{\pi_q(K \times K^c) + \pi_d(K)}{\pi(K)}.$$

Therefore,

$$\sigma_0 \leqslant \inf_{\pi(K) > 0} \frac{\pi_q(K \times K^c) + \pi_d(K)}{\pi(K)}$$

gives us an upper bound.

We can easily given an approximation theorem for σ_0 as an analogue of Theorem (4.2). Finally, for the birth-death process, we again have

$$\sigma_0 = \hat{\alpha},$$

where $\hat{\alpha}$ is the exponentially ergodic rate (i.e., $P_{ij}(t) = O(\exp(-\hat{\alpha} t))$ for all i, j). Thus,

$$\text{Exponential } L^2\text{-convergence} \Longleftrightarrow \text{Exponential ergodicity}.$$

Finally, Theorem (5.12)remains valid in the present case.

References

[1]. Aizeman, M. and Holley, R., *Rapid convergence to equilibrium of stochastic Ising models in the Dobrushin-Shlosman regime*, IMA volume on Percolation Theory and Ergodic Theory of Infinite Particle Systerns (H. Kesten, Ed.), Springer, New York, 1987, 1–11.

[2]. Cheeger, J., *A lawer bound for the lowest eigenvalue of the Laplacian*, Problems in Analysis: A Symposium in Honor of S.Bochner (Ed. R.C.Gunning), Princeton Univ. Press, Princeton, N.J., 1970, 195–199.

[3]. Chen, M.F., *Jump Processes and Interacting Particle Systems*, Beijing Normal Univ. Press (In Chinese), 1986.

[4]. Chen, M. F., *Comparison theorems for Green functions of Markov chains*, Chin. Ann. of Math., 1987, to appear.

[5]. Chen, M.F., *Dirichlet forms and symmetrizable jump processes*, Chin. Quart. J. of Math., 4 (1989).

[6]. Chen, M. F. and Stroock, D.W., *λ_π-invariant measures*, Lecture Notes in Math. 986, Seminare de Prob. XVII 1981 / 1982, Ed. J. Azema et M. Yor, Springer-Verlag, 1983, 205–220.

[7]. Fukushima, M., *Dirichlet Forms and Markov Processes*, North-Holland, 1980.

[8]. Holley, R., *Convergence in L^2 of stochastic Ising models: jump processes and diffusions* (1984), Proceedings of the Tanigushi Symposium on Stochastic Analysis, Kyoto, 1982.

[9]. Holley, R., *Possible rates of convergence in finite range, attractive spin systems*, In "Particle Systems, Random Media and Large Deviations" (R. Duma, Ed.), Contemporary Mathematics 41 (1985*a*), 215–234.

[10]. Holley, R., *Rapid convergence to equilibrium in one dimensional stochastic Ising models*, Ann. of Prob. 13(1985b), 72–89.

[11]. Holley, R. and Stroock, D., *L^2 theory for the stochastic Ising model*, Zeit. Wahrsch. Verw. Gebiete, 35 (1976), 87–101.

[12]. Holley, R. and Stroock, D., *Logarithmic Sobolev inequalities and stochastic Ising models*, J. Star. Phys., 46 (1987), 1159–1194.

[13]. Karlin, M.and Taylor, H.M., *A Second Course in Stochastic Processes*, Academic Press, 1981.

[14]. Kim, J.H., *Stochastic calculus related to non-symmetric Dirichlet forms*, Osaka J. Math. (1987).

[15]. Korzeniowski, A., *On logarithmic Sobolev constant for ditfusion semigroups*, J. Funt. Anal., 71 (1987), 363–70.

[16]. Lawler, G. F. and Sokal, A. D., *Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality*, Trans. Amer. Math. Soc., 309 (1988), 557–580.

[17]. Liggett, T. M., *Exponential L^2 convergence of attractive reversible nearest particle systems*, Ann. of Prob., 17 (1989), 403–432.

[18]. Stroock, D.W., *The Malliavin calculus, a functional analytic approach*, J. Funt. Anal., 44 (1981), 212–257.

[19]. Stroock, D.W., *On the spectrum of Markov sernigroup and existence of nvariant measures*, Functional Analysis in Markov processes, Proceedings, Ed. M. Fukushima, Springer-Verlag, 1981, 287–307.

[20]. Sullivan, W. G., *The L^2 spectral gap of certain positive recurrent Markov chains and jump processes*, Zeit. Wahrs. Verw. Gebiete, 67 (1984), 387–398.

[21]. Van Doom, E., *Stochastic Monotonicity and Queueing Applications of Birth-Death Proceses*, Lecture Notes in Statistics 4, Springer-Verlag, 1981.

[22]. Van Doom, E., *Conditions for exponential ergodicity and bounds for the decay parameter of a birth-death process*, Adv. Appl. Prob., 17 (1985), 514–530.

Science in China (A) 37:1(1994), 1–14

APPLICATION OF COUPLING METHOD TO THE FIRST EIGENVALUE ON MANIFOLD

Mu-Fa Chen and Feng-Yu Wang

(Beijing Normal University)
Received September 30, 1992; revised June 1, 1993

Abstract. By using a coupling technique, this paper presents some lower bounds of the first eigenvalue λ_1 of an adjoint operator $\Delta+Z$ on compact M. This method is new and the proofs are straightforward. The method not only achieves the same optimal bounds as those obtained by other techniques but also improves some known estimates. Denote by g, d and D the Riemannian metric, dimension and diameter of M respectively. Suppose that $\mathrm{Ric}_M \geqslant -Kg$ for some real number K. Then, in the case of $Z=0$, the lower bound of λ_1 provided by the paper can be summarized as follows.

$$\begin{aligned}\lambda_1 &\geqslant \max\left\{\frac{\pi^2}{D^2}, -\frac{d}{d-1}K, \frac{8}{D^2}-\frac{K}{3}\right\}, \quad \text{if } K\leqslant 0\\ &\geqslant \max\left\{\frac{\pi^2}{D^2}-K, \frac{8}{D^2}-\frac{K}{3}, \frac{8}{D^2}\exp\left[-\frac{D^2K}{8}\right],\right.\\ &\qquad \left.\frac{8}{D^2}\left(1+\frac{D}{3}\sqrt{K(d-1)}\right)\exp\left[-\frac{D}{2}\sqrt{K(d-1)}\right]\right\}, \quad \text{if } K\geqslant 0.\end{aligned}$$

Besides, a method to estimating the bound for general operators is also given. Two examples, even on non-compact space, show that the estimates obtained by this method can be sharp.

1. Introduction.

It is well known that the estimate of the first eigenvalue play a critical role in analysis of manifold (refer to Schoen and Yau (1988), for example). On the other hand, it is also known that the first eigenvalue of $\Delta + Z$ is just the L^2-exponential convergence rate of the corresponding Markov semigroup (cf. Chen

2000 *Mathematics Subject Classification*. 58G32, 58G25.

Key words and phrases. Coupling method, Riemannian manifold, the first eigenvalue.

Research supported in part by the National Natural Science Foundation and State Education Commission of China. The first author is also supported in part by the NSERC operating grant of D. A. Dawson

(1991) or Chen (1992, Chapter 9)). Such convergence is very important in the study of interacting particle systems since it is now used as a tool to describe phase transitions. Refer to Chen (1992), Sections 9.4 and 11.4 for more references.

Throughout this paper, let (M, g) be a d-dimensional compact Riemannian manifold with distance $\rho = \rho_M$ given by the metric g and assume that $\mathrm{Ric}_M \geqslant -Kg$ for some $K \in \mathbb{R}$. Denote by Δ the Laplace-Beltrami operator on M and let Z be a C^1-vector field. For the adjoint property, Z should have the form $Z = \bigtriangledown f$ for some $f \in C^2$. But in what follows, we prefer to use Z only. The main purpose of this paper is using the coupling technique to study the lower bounds of the first eigenvalue λ_1 of operator $\Delta + Z$ on compact M. Here the "first eigenvalue" means the smallest positive eigenvalue λ of $-(\Delta + Z)$. At the end of this section, we introduce a general method for estimating the bound of λ_1. The method works for those operators which can be considered as a generator of a Markov process. Moreover, Theorem 1.8 and two examples below with non-compact state space (an one-dimensional diffusion process and a Markov chain) show that the method does produce optimal bound. In order to compare our results with the known ones, for the reader's convenience, we first recall some previous works.

The first three results (Theorem 1.1 – Theorem 1.3) deal with the Laplacian operator (i.e., $Z = 0$) only.

Theorem 1.1 (Lichnerowicz (1958)). If $K < 0$, Then

$$\lambda_1 \geqslant -\frac{d}{d-1}K. \tag{1.1}$$

This estimate is quite good for large Ricci curvature. It is indeed sharp when $M = S^d\,(d > 1)$ but it is quite poor when $-K$ is small. During 1975 — 1983, Li and Yau made some nice progress on estimating the lower bounds. In particular, they obtained in the case that $K = 0$ the bound: $\pi^2/\big(2D^2\big)$, where $D = \sup \rho(x, y)$ is the diameter of M. The best bound was then obtained by Zhong and Yang and further generalized by Cai as follows:

Theorem 1.2 (Zhong and Yang (1984), Cai (1991)).

$$\lambda_1 + \max\{0, K\} \geqslant \frac{\pi^2}{D^2}. \tag{1.2}$$

The equality in (1.2) holds when $M = S^1$. In the case that $K > 0$, Li and Yau (1980) obtained the following estimate:

$$\lambda_1 \geqslant \left\{D^2(d-1)\exp\left[1 + \sqrt{1 + 4D^2K(d-1)}\,\right]\right\}^{-1}.$$

Recently, this result has been improved as follows:

Theorem 1.3 (Yang (1989) and Jia (1991)). Let $K > 0$. Then

$$\begin{aligned}\lambda_1 &\geqslant \frac{\pi^2}{D^2}\exp\left[-\frac{1}{2}D\sqrt{K(d-1)}\,\right], \quad \text{if } d \geqslant 5\\ &> \frac{\pi^2}{2D^2}\exp\left[-\frac{1}{2}D\sqrt{K((d-1)\vee 2)}\,\right], \quad \text{if } 2 \leqslant d \leqslant 4.\end{aligned} \tag{1.3}$$

To state our estimates, we need some notations. Given a continuous function $\gamma(\xi) : (0,\infty) \to \mathbb{R}$, which will be determined case by case, define

$$C(r) = \exp\left[\frac{1}{4}\int_0^r \gamma(\xi)d\xi\right] \quad \text{and} \quad F(r) = \int_0^r C(s)^{-1}ds \int_s^D C(u)du, \quad r > 0. \tag{1.4}$$

Next, set

$$K(Z) = \sup\big\{ -\mathrm{Ric}_M(X,X)(x) + \langle \nabla_X Z,\, X\rangle(x) \colon\, x \in M,\, X \in \mathscr{X}(M),\, |X(x)| = 1\big\},$$

where $\mathscr{X}(M)$ is the set of all C^∞-vector fields.

Theorem 1.4. In general, we have

$$\lambda_1 \geqslant \frac{4}{F(D)}\,, \tag{1.5}$$

where F is given by (1.4) with $\gamma(\xi) = K(Z)\xi$. More precisely,

$$\lambda_1 \geqslant \frac{8\alpha}{D^2}\bigg(\int_0^1 \Big[e^{\alpha r(2-r)} - e^{\alpha r^2}\Big]\frac{dr}{r}\bigg)^{-1}, \quad \alpha := \frac{D^2K(Z)}{8}.$$

Theorem 1.5. Let $a : (0,\infty) \to [0,\infty)$ be a continuous function such that

$$a(r) \geqslant \sup_{\rho(x,y)=r} \big(Z\rho(\cdot,y)(x) + Z\rho(x,\cdot)(y)\big), \quad r > 0,$$

where $\sup\emptyset = 0$.

(1) If $K < 0$, then (1.5) holds with the choice:

$$\gamma(\xi) = -2\sqrt{-K(d-1)}\tan\big(\sqrt{-K/(d-1)}\,\xi/2\big) + a(\xi).$$

That is,

$$\lambda_1 \geqslant \frac{4}{D^2}\bigg\{\int_0^1 du\int_0^{1-u} ds\left[\frac{\cos(\alpha'(s+u))}{\cos(\alpha' s)}\right]^{d-1}\exp\left[\frac{D}{4}\int_s^{s+u} a(D\xi)d\xi\right]\bigg\}^{-1},$$

$$\alpha' := \frac{D}{2}\sqrt{\frac{|K|}{d-1}}.$$

(2) If $K \geqslant 0$, then (1.5) holds with the choice:

$$\gamma(\xi) = 2\sqrt{K(d-1)}\,\tanh\big(\xi\sqrt{K/(d-1)}/2\big) + a(\xi), \quad K \geqslant 0,\ \xi > 0.$$

That is,

$$\lambda_1 \geqslant \frac{4}{D^2}\bigg\{\int_0^1 du\int_0^{1-u} ds\left[\frac{\cosh(\alpha'(s+u))}{\cosh(\alpha' s)}\right]^{1-d}\exp\left[\frac{D}{4}\int_s^{s+u} a(D\xi)d\xi\right]\bigg\}^{-1}.$$

In contrast the study in geometry where one looks for a uniform bound for a class of manifolds, the explicit dependence of the bound of λ_1 on the Ricci curvature is critical in the study of phase transition of infinite dimensional diffusions, especially in estimating the constant of Logarithmic Sobolev inequality. Refer to Deuschel and Stroock (1989) for details and references. Because of this reason, the above bounds are stated in their complete form. Since these estimates are still involved, we would like to present some simple approximations. The next two results are consequence of Theorem 1.4 and Theorem 1.5 respectively.

Corollary 1.6.

(1) If $K(Z) \leqslant 0$, then

$$\lambda_1 \geqslant \begin{cases} \dfrac{8}{D^2} - \dfrac{K(Z)}{3}, & \text{if } D^2K(Z) \geqslant -24 \\ \dfrac{8}{D^2} - \dfrac{K(Z)}{7 + 2\log D + \log(-K(Z))}, & \text{if } D^2K(Z) < -24. \end{cases}$$

(2) If $K(Z) \geqslant 0$, then $\lambda_1 \geqslant \max\left\{\dfrac{8}{D^2} - \dfrac{K(Z)}{3},\ \dfrac{8}{D^2}\exp\left[-\dfrac{D^2}{8}K(Z)\right]\right\}$.

Corollary 1.7. Let $|Z| \leqslant m < \infty$.

(1) If $K < 0$, then

$$\lambda_1 \geqslant \begin{cases} \left(\dfrac{8}{D^2} - \dfrac{K}{3}\right)\exp\left[-\dfrac{1}{2}Dm\right], & \text{if } D^2K \geqslant -24 \\ \left(\dfrac{8}{D^2} - \dfrac{K}{7 + 2\log D + \log(-K)}\right)\exp\left[-\dfrac{1}{2}Dm\right], & \text{if } D^2K < -24. \end{cases}$$

(2) If $K \geqslant 0$, then

$$\lambda_1 \geqslant \frac{4}{D^2}\left\{\int_0^1 du \int_0^{1-u} ds \exp\left[\frac{1}{2}Dmu + \frac{D}{4}\int_s^{s+u} \big(2\sqrt{K(d-1)}\,\big) \wedge (DK\xi)d\xi\right]\right\}^{-1}.$$

In particular, we have

$$\lambda_1 \geqslant \max\left\{\left(\frac{8}{D^2} - \frac{K}{3}\right)\exp\left[-\frac{1}{2}Dm\right],\ \frac{8}{D^2}\exp\left[-\frac{D}{8}(4m + DK)\right],\right.$$
$$\left.\frac{2\beta^2}{D^2}\left\{2\exp\left[\frac{\beta}{2}\right] - 2 - \beta\right\}^{-1}\right\}, \quad \beta := D\big(\sqrt{K(d-1)} + m\big).$$

Here and in what follows, when $K = 0$ and $m = 0$, the right-hand side is understood as the limit as $K \to 0$ and $m \to 0$.

Another main result of the paper is as follows.

Theorem 1.8. Set

$$H = \sup\big\{\langle \nabla_X Z,\, X\rangle(x) :\, X \in M,\, X \in \mathscr{X}(M),\, |X(x)| = 1\big\} \vee 0.$$

(1) If $d > 1$ and $\dfrac{d}{d-1}K + H < 0$, then $\lambda_1 \geqslant -\dfrac{d}{d-1}K - H$.

(2) If $K(Z) \leqslant 0$, then $\lambda_1 \geqslant \dfrac{\pi^2}{D^2}$.

(3) If $K(Z) > 0$, then $\lambda_1 \geqslant \dfrac{\pi^2}{D^2} - K(Z)$.

It follows from part (2) of Corollary 1.7 that

$$\lambda_1 \geqslant \frac{2\beta^2}{D^2\big(2\exp[\beta/2]-2-\beta\big)} \geqslant \frac{8}{D^2}\left(1+\frac{\beta}{3}\right)\exp\left[-\frac{\beta}{2}\right].$$

From this and part (3) of Theorem 1.8, it is easy to check that the first estimate of (1.3) actually holds for all $d \geqslant 2$. Combining these facts with Corollaries 1.6, 1.7 and Theorem 1.8, we obtain the lower bounds given in the abstract of the paper.

It will be seen in the next section that the technique used to prove Theorem 1.8 is different to those used for Theorem 1.4 and Theorem 1.5. For the remainder of this section, we show that our method actually works for much general operators, even on non-compact space. To state our result, we still need some notation. Let P_1 and P_2 be probability measures on $(E, \mathscr{E})$. A probability measure P on $(E \times E, \mathscr{E} \times \mathscr{E})$ is called a coupling of P_1 and P_2 if

$$P(B \times E) = P_1(B) \quad \text{and} \quad P(E \times B) = P_2(B) \quad \text{for all } B \in \mathscr{E}.$$

Next, define

$$W(P_1, P_2) = \inf_P \int_{E\times E} \rho(x_1, x_2) P(dx_1, dx_2),$$

where ρ is a metric in E, the infimum varies over all coupling P of P_1 and P_2. W is called the minimum L^1-metric (or Kantorovich-metric or Wasserstein-metric and so on) of P_1 and P_2.

The next general result contains one of the key ideas of the paper.

Theorem 1.9. Let $(E, \rho, \mathscr{E})$ be a separable complete metric space with metric ρ. Consider a reversible Markov process (x_t) (or (y_t)) with reversible measure π and having weak generator A. Given an eigenfunction u corresponding λ_1. Suppose that

(1) u is contained in the weak domain of A in the sense that

$$\mathbb{E}^x u(x_t) - u(x) = \int_0^t \mathbb{E}^x Au(x_s) ds.$$

(2) u is Lipschitz continuous with respect to an equivalent metric $\bar{\rho}$ of ρ.

(3) $W(P(x_t), P(y_t)) \leqslant \bar{\rho}(x, y)\exp[-\sigma t]$ for some $\sigma > 0$ and for all $t \geqslant 0$, x and $y \in E$, where $P(x_t)$ (resp., $P(y_t)$) is the distribution of (x_t) (resp., (y_t)) at time t, starting from x (resp., y).

Then, we have $\lambda_1 \geqslant \sigma$.

At the first look, one may think that Theorem 1.9 is useless since the hypotheses are made on the eigenfunction which is usually unknown. However, in the compact case, conditions (1) and (2) are often satisfied. Actually, as we will see in the next section, Theorem 1.8 is an easy consequence of Theorem 1.9. In the non-compact case, condition (1) can be often deduced from the above equation for localized u plus $|\mathbb{E}^x u(x_t)| < \infty$, which is then fulfilled if π has a positive density since $u \in L^2(\pi)$ and so

$$\int \pi(dx)\mathbb{E}^x|u(x_t)| \leqslant \left(\int \pi(dx)\mathbb{E}^x u(x_t)^2\right)^{1/2} = \left(\int \pi(dx)u(x)^2\right)^{1/2} < \infty.$$

Thus, in the non-compact case, the main restriction is condition (2). Of course, if $\bigtriangledown u$ is bounded, then (2) holds. Otherwise, the real value of Theorem 1.9 is still in the compact case. Nevertheless, at least for Markov chains, we can reduce the non-compact case to the compact one by using a localization procedure, as illustrated by Example 1.11 below. In any case, condition (3) is essential. Under some reasonable assumption, this condition implies the exponential ergodicity of the process in the following sense:

$$W(P(x_t), \pi) \leqslant Ce^{-\sigma t},$$

where C is a constant depending on x, π and ρ. This leads to the starting point of our study: comparing the rate σ here with the exponentially L^2-convergent rate λ_1. Note that however the W-metric is not topologically intrinsic, which depends on the metric ρ in the base space and hence σ also depends on ρ. On the other hand, as a $L^2(\pi)$-eigenvalue, λ_1 depends on π rather than ρ, so it is not obvious why σ and λ_1 are comparable. Generally speaking, there is no hope to compute the metric W exactly. But what we need is only the upper estimate, and this is just the place where the coupling technique is employed.

To illustrate the power of our method, we consider a typical example.

Example 1.10. For one-dimensional Ornstein-Uhlenbeck process,

$$A = \frac{1}{2}\left(\frac{d^2}{dx^2} - 2x\frac{d}{dx}\right),$$

we have $\lambda_1 \geqslant 1$.

Proof. It is known that corresponding the eigenvalues $\lambda_n = n$, the eigenfunctions are

$$(-1)^n e^{x^2}\frac{d^n}{dx^n}\big(e^{-x^2}\big).$$

Clearly, conditions (1) and (2) of Theorem 1.9 are fulfilled with respect to $\bar{\rho} = \rho =$ the ordinary metric. As for condition (3), we use the coupling by reflection. The coefficients of the generator $\widetilde{A}$ of the coupling diffusion process are the following (cf. Chen and Li (1989)):

$$a(x,y) = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \qquad b(x,y) = \begin{pmatrix} -x \\ -y \end{pmatrix}.$$

Take $\rho(x,y) = |x-y|$. Then, we have $\widetilde{A}\rho(x,y) \leqslant -\rho(x,y)$ and so

$$\mathbb{E}^{x,y}\rho(x_t, y_t) \leqslant \rho(x,y)e^{-t}.$$

Hence the assertion follows from Theorem 1.9. The same conclusion holds for higher dimensional case. □

The next example illustrates a localization procedure.

Example 1.11. Consider a reversible birth-death process with birth rate $b_i = \beta_0 + \beta_1 i$ and death rate $a_i = \delta_1 i$, $\delta_1 > \beta_1$. Then, we have $\lambda_1 \geqslant \delta_1 - \beta_1$.

Proof. The condition "$\delta_1 > \beta_1$" comes from the fact that only in this situation, the above linear growth birth-death process is positive recurrent (cf. Chen (1992, Example 4.55). The reversible measure is $\pi_i = \mu_i/\mu$:

$$\mu_0 = 1, \quad \mu_i = \frac{\beta_0(\beta_0+\beta_1)\cdots\big(\beta_0+(i-1)\beta_1\big)}{\delta_1^i\, i!}, \quad \mu = \sum_{i=0}^{\infty} \mu_i.$$

Consider the march coupling (cf. Chen (1992, Chapter 5)):

$$\begin{array}{rl}
(x_t, y_t) = (i,j) \to (i+1, j+1) & \text{at rate } b(x_t) \wedge b(y_t) \\
(i,j) \to (i-1, j-1) & \text{at rate } a(x_t) \wedge a(y_t) \\
(i,j) \to (i+1, j) & \text{at rate } (b(x_t) - b(y_t))^+ \\
(i,j) \to (i, j+1) & \text{at rate } (b(y_t) - b(x_t))^+ \\
(i,j) \to (i-1, j) & \text{at rate } (a(x_t) - a(y_t))^+ \\
(i,j) \to (i, j-1) & \text{at rate } (a(y_t) - a(x_t))^+,
\end{array}$$

where $a^+ = \max\{a, 0\}$ and $a \wedge b = \min\{a, b\}$. Take $\rho(i,j) = |i-j|$. Again, denote by $\widetilde{A}$ the generator of the coupling process. Then, a simple computation shows that $\widetilde{A}\rho(i,j) \leqslant -(\delta_1 - \beta_1)\rho(i,j)$ and hence

$$\mathbb{E}^{i,j}\rho(x_t, y_t) \leqslant \rho(i,j)e^{-(\delta_1-\beta_1)t}.$$

From this and Theorem 1.9, one may guess that $\lambda_1 \geqslant \delta_1 - \beta_1$, which is again exact. To complete the proof, one may study the eigenfunction and then apply Theorem 1.9 directly. But we prefer to avoid to do so. Let $N \geqslant 1$. Consider a modified process with finite state space $\{0, 1, \cdots, N\}$. The rates b_i and $a_i\, (i \leqslant N-1)$ are the same as the original one but replacing a_N with

$$\hat{a}_N := \pi_{N-1} b_{N-1} \Big/ \sum_{j>N-1} \pi_j.$$

From Chen(1991 or 1992), it is known that $\lambda_1(N) \downarrow \lambda_1$ as $N \to \infty$, where $\lambda_1(N)$ is the first positive eigenvalue of the finite Markov chain. Hence, it suffices to prove that $\lambda_1(N) \geqslant \delta_1 - \beta_1$. By Theorem 1.9, the assertion follows once we show that

$$\widetilde{A}_N\rho(i,j) \leqslant -(\delta_1 - \beta_1)\rho(i,j)$$

for suitable coupling $\widetilde{A}_N$. To do so, when $0 \leqslant i \leqslant j < N$ or $0 \leqslant j \leqslant i < N$, we adopt the march coupling as above. If $0 \leqslant i < j = N$, we use the coupling by inner reflection:

$$\begin{array}{rl}
(x_t, y_t) = (i,N) \to (i+1, N-1) & \text{at rate } b(x_t) \wedge \hat{a}_N \\
(i,N) \to (i+1, N) & \text{at rate } (b(x_t) - \hat{a}_N)^+ \\
(i,N) \to (i, N-1) & \text{at rate } (\hat{a}_N - b(x_t))^+ \\
(i,N) \to (i-1, N) & \text{at rate } a(x_t).
\end{array}$$

By symmetry, we can write down the coupling rates in the case that $0 \leqslant j < i = N$. Now, if $0 \leqslant i \leqslant j \leqslant N-1$ or $0 \leqslant j \leqslant i \leqslant N-1$, we have

$$\widetilde{A}_N \rho(i,j) \leqslant -(\delta_1 - \beta_1)\rho(i,j)$$

mentioned above. As for the case that $0 \leqslant i < j = N$, the required estimate is

$$\begin{aligned}\big((\beta_0+\beta_1 i) \wedge \hat{a}_N\big)(-2) + \delta_1 i - \big((\beta_0+\beta_1 i) - \hat{a}_N\big)^+ - \big(\hat{a}_N - (\beta_0+\beta_1 i)\big)^+ \\ \leqslant -(\delta_1-\beta_1)(N-i)\end{aligned}$$

for all $0 \leqslant i < N$. Equivalently,

$$(\delta_1 - \beta_1)N \leqslant \beta_0 + \hat{a}_N.$$

Rewrite $\hat{a}_N$ as follows:

$$\begin{aligned}\hat{a}_N &= \frac{b_{N-1}}{\sum_{j>N-1} \mu_j/\mu_{N-1}} \\ &= \frac{N\delta_1}{1 + \sum_{j=1}^{\infty} \frac{1}{\delta_1^j}\left(\frac{\beta_0+N\beta_1}{N+1}\right)\cdots\left(\frac{\big(\beta_0+(N+j-1)\beta_1\big)}{N+j}\right)} \\ &=: \frac{N\delta_1}{S}.\end{aligned}$$

Choose $m \geqslant 1$ such that $(m-1)\beta_1 < \beta_0 \leqslant m\beta_1$ and choose N large enough so that $\beta_1(1+\varepsilon) < \delta_1$, where $\varepsilon = (m-1)/(N+1)$. Then

$$\begin{aligned}S &\leqslant 1 + \sum_{j=1}^{\infty}\left(\frac{\beta_1}{\delta_1}\right)^j \left(1+\frac{m-1}{N+1}\right)\cdots\left(1+\frac{m-1}{N+j}\right) \\ &\leqslant 1 + \sum_{j=1}^{\infty}\left(\frac{\beta_1(1+\varepsilon)}{\delta_1}\right)^j \\ &= \frac{\delta_1}{\delta_1 - \beta_1(1+\varepsilon)}.\end{aligned}$$

Thus,

$$\hat{a}_N \geqslant \big(\delta_1 - \beta_1(1+\varepsilon)\big)N = \big(\delta_1-\beta_1\big)N - \frac{N}{N+1}(m-1)\beta_1 > \big(\delta_1-\beta_1\big)N - \beta_0. \quad \square$$

It should be clear that the localization procedure illustrated above is meaningful for general Markov chains. But in the case of the rates being non-linear, we should use a finer coupling (for instance, the coupling by inner reflection) instead of the march coupling. We may also have to adopt a finer metric instead of the ordinary one. Refer to Chen (1992, Chapter 5). Since this topic goes beyond the main scope of the paper, we should stop here.

2. Proofs.

Let R be the Riemannian curvature tensor and let $\mathbf{C} = \{(x, y) : x \text{ is the cut locus of } y\}$. In what follows, the main coupling for diffusions will be used, called coupling by reflection, is due to Lindvall and Rogers (1986) by using stochastic differential equations and studied by Chen and Li (1989) by using martingale approach in the context of Euclidian space. The coupling was generalized to manifolds by Kendall (1986 a,b) and Cranston (1991). For the present purpose, the coupling process $\{(x_t, y_t)\}_{t\geqslant 0}$ was explained carefully in Cranston (1991). The reader is urged to refer the Cranston's paper if necessary.

Before the coupling time $T := \inf\{t \geqslant 0 : x_t = y_t\}$, we have

$$d\rho(x_t, y_t) = 2\sqrt{2}\, db_t + \left[\int_{x_t}^{y_t} \sum_{i=2}^{d} \Big(|\bigtriangledown_U W^i|^2 - \langle R(W^i, U)U, W^i\rangle \Big)\right] dt + \big[\langle Z(y_t), U\rangle - \langle Z(x_t), U\rangle\big] dt - dL_t, \tag{2.1}$$

where W^i, $i = 2, \cdots, d$ are Jacobi fields along the unique shortest geodesic γ between x_t and y_t, U is the unit tangent vector to γ, b_t is a Brownian motion in $\mathbb{R}$ and L_t is an increasing process with support contained in $\{t \geqslant 0 : (x_t, y_t) \in \mathbf{C}\}$. When $(x_t, y_t) \in \mathbf{C}$, the coefficient of dt is taken to be 0. Equation (2.1) was labeled by (1.7) in Cranston (1991) but in the later case a coefficient $1/2$ in front of the integration is needed since for which the generator is $\frac{1}{2}\Delta + Z$.

The following result is an analog of Chen and Li (1989), Theorem 5.1.

Lemma 2.1. Take

$$C(r) = \exp\left[\frac{1}{8}K(Z)r^2\right]$$

and define the corresponding F by (1.4). We have

$$\mathbb{E}^{x,y}T \leqslant F(D)/4.$$

Proof. a) Let γ_s: $[0, \rho(x_t, y_t)] \to M$ be the shortest geodesic between x_t and y_t. Denote by U its tangent vector. Then

$$\langle Z(y_t), U\rangle - \langle Z(x_t), U\rangle = \int_0^{\rho(x_t,y_t)} \frac{d\langle Z(\gamma_s), U\rangle}{ds} ds = \int_0^{\rho(x_t,y_t)} \langle \bigtriangledown_U Z, U\rangle(\gamma_s) ds.$$

On the other hand, it was proved in Kendall (1986 b), p. 118 that

$$\int_{x_t}^{y_t} \sum_{i=2}^{d} \Big(|\bigtriangledown_U W^i|^2 - \langle R(W^i, U)U, W^i\rangle \Big) \leqslant -\int_0^{\rho(x_t,y_t)} \mathrm{Ric}_M(U, U)(\gamma_s) ds, \quad t < T. \tag{2.2}$$

Combining these facts with (2.1), we see that

$$d\rho(x_t, y_t) \leqslant 2\sqrt{2} db_t + K(Z)\rho(x_t, y_t)dt, \quad t < T. \tag{2.3}$$

b) Note that $F \in C^\infty(\mathbb{R})$. By Itô formula and using a), we get

$$dF(\rho(x_t, y_t)) \leqslant 2\sqrt{2}\, F'(\rho(x_t, y_t))db_t - 4dt, \quad t < T.$$

Hence

$$\mathbb{E}^{x,y} F(\rho(x_{t\wedge T}, y_{t\wedge T})) \leqslant F(\rho(x,y)) - \mathbb{E}^{x,y} \int_0^{t\wedge T} 4\, ds \leqslant F(D) - 4\mathbb{E}^{x,y}(t \wedge T).$$

Letting $t \to \infty$, we obtain $\mathbb{E}^{x,y} T \leqslant F(D)/4$. $\square$

Proof of Theorem 1.4. Let u be an eigenfunction corresponding to λ_1. Then, by the martingale formulation, we have

$$|u(y) - u(x)| \leqslant \mathbb{E}^{x,y}|u(x_{t\wedge T}) - u(y_{t\wedge T})| + \lambda_1 \mathbb{E}^{x,y} \int_0^{t\wedge T} |u(x_s) - u(y_s)|ds. \quad (2.4)$$

Since $T < \infty$, $a.s.$ by Lemma 2.1, letting $t \uparrow \infty$ in (2.4), it follows that

$$|u(y) - u(x)| \leqslant \lambda_1 \mathbb{E}^{x,y} \int_0^{T} |u(x_s) - u(y_s)|ds. \quad (2.5)$$

Choose x_0 and y_0 so that $u(y_0) - u(x_0) = \sup|u(x) - u(y)| > 0$. Without loss of generality, assume that $u(y_0) - u(x_0) = 1$. Then by (2.5) we have

$$1 \leqslant \lambda_1 \mathbb{E}^{x_0,y_0} \int_0^T |u(x_s) - u(y_s)|ds \leqslant \lambda_1 \mathbb{E}^{x_0,y_0} T.$$

Therefore

$$\lambda_1 \geqslant \left(\mathbb{E}^{x_0,y_0} T\right)^{-1}.$$

Now, the first assertion of Theorem 1.4 follows from Lemma 2.1.

To prove the second assertion, note that

$$\begin{aligned} F(D) &= D^2 \int_0^1 ds\, C(Ds)^{-1} \int_s^1 C(Du)du \\ &= D^2 \int_0^1 ds \int_s^1 \exp\left[\frac{1}{4}\int_{Ds}^{Du} \gamma(\xi)d\xi\right]du \\ &= D^2 \int_0^1 ds \int_0^{1-s} \exp\left[\frac{1}{4}\int_{Ds}^{D(u+s)} \gamma(\xi)d\xi\right]du \\ &= D^2 \int_0^1 du \int_0^{1-u} \exp\left[\frac{D}{4}\int_s^{u+s} \gamma(D\xi)d\xi\right]ds \end{aligned}$$

Now, since $\gamma(\xi) = K(Z)\xi$, we have

$$\frac{D}{4}\int_s^{u+s} \gamma(D\xi)d\xi = \alpha u(u+2s), \quad \alpha = \frac{D^2 K(Z)}{8},$$

and so

$$F(D) = D^2 \int_0^1 du \int_0^{1-u} e^{\alpha u(u+2s)} ds = \frac{D^2}{2\alpha} \int_0^1 \left[e^{\alpha u(2-u)} - e^{\alpha u^2} \right] \frac{du}{u}. \quad \square$$

Proof of Corollary 1.6. First, when $K(Z) \geqslant 0$, the estimate $\lambda_1 \geqslant \frac{8}{D^2} e^{-\alpha}$ follows from Theorem 1.4 and the fact that $\alpha u(u+2s) \leqslant \alpha$ for $0 \leqslant s \leqslant 1-u \leqslant 1$ and hence

$$F(D) = D^2 \int_0^1 du \int_0^{1-u} e^{\alpha u(u+2s)} ds \leqslant D^2 \int_0^1 du \int_0^{1-u} e^{\alpha} ds = \frac{D^2}{2} e^{\alpha}.$$

Next, since $1 - \alpha/3 < 2\exp[-\alpha]$ whenever $\alpha \geqslant 3$, it suffices to show that

$$G(\alpha) := \int_0^1 \left(\frac{1}{\alpha} - C'(\alpha) \right) \left[e^{\alpha x(2-x)} - e^{\alpha x^2} \right] \frac{dx}{x} \leqslant 1$$

for all $\alpha < 3$, where

$$C'(\alpha) = \begin{cases} \dfrac{1}{3}, & \text{if } \alpha \geqslant -3 \\ \dfrac{1}{9 + \log(-\alpha)}, & \text{if } \alpha < -3. \end{cases}$$

a) Let $\alpha \leqslant -3$. Set $\beta = -\alpha$ and $y = \beta x$. Then, the required assertion is reduced to

$$\int_0^\beta \left(\frac{1}{\beta} + \frac{1}{9 + \log \beta} \right) e^{-y^2/\beta} \left[1 - e^{-2y(1-y/\beta)} \right] \frac{dy}{y} \leqslant 1$$

for all $\beta \geqslant 3$. Since $e^x \geqslant 1 + x + x^2/2 + x^3/6$, we have

$$\begin{aligned} &\int_0^\beta e^{-y^2/\beta} \left[1 - e^{-2y+2y^2/\beta} \right] \frac{dy}{y} \\ &\quad \leqslant \int_0^1 \left[1 - e^{-2y(1-1/\beta)} \right] \frac{dy}{y} + \left(1 - e^{-\beta/2} \right) \int_1^\beta \frac{dy}{y} \\ &\quad \leqslant 2\left(1 - \frac{1}{\beta} \right) \int_0^1 \left[1 - y\left(1 - \frac{1}{\beta} \right) + \frac{2}{3} y^2 \left(1 - \frac{1}{\beta} \right)^2 \right] dy + \left[1 - e^{-\beta/2} \right] \log \beta \\ &\quad \leqslant \frac{13}{9} - \frac{11}{9\beta} + \left(1 - e^{-\beta/2} \right) \log \beta, \quad \beta \geqslant 3. \end{aligned}$$

But

$$\left(\frac{1}{\beta} + \frac{1}{9 + \log \beta} \right) \left(\frac{13}{9} - \frac{11}{9\beta} + \left(1 - e^{-\beta/2} \right) \log \beta \right) \leqslant 1, \quad \beta \geqslant 3.$$

the required conclusion follows immediately. When $Z = 0$, the estimate for this region of β has already covered by Lichnerowicz's estimate.

b) We now assume that $|\alpha|<3$. Because $G(0)=1$, we need only to show that $G(\alpha)$ achieves 1 its maximum at $\alpha=0$. Since the continuity of $G(\alpha)$, it suffices to prove that $G'(\alpha)>0$ for $\alpha\in(-3,0)$ and $G'(\alpha)<0$ for $\alpha\in(0,3)$. For this, compute $G'(\alpha)$:

$$\begin{aligned}G'(\alpha)&=-\frac{1}{\alpha^2}\int_0^1\Big[e^{\alpha x(2-x)}-e^{\alpha x^2}\Big]\frac{dx}{x}+\left(\frac{1}{\alpha}-\frac{1}{3}\right)\int_0^1\Big[(2-x)e^{\alpha x(2-x)}-xe^{\alpha x^2}\Big]dx\\&=\frac{1}{\alpha^2}\int_0^1 e^{\alpha(1-x^2)}\Big[e^{2\alpha x(x-1)}-1\Big]\frac{dx}{1-x}+\left(\frac{1}{\alpha}-\frac{1}{3}\right)\int_0^1 e^{\alpha(1-x^2)}dx.\end{aligned}$$

Next, note that

$$\begin{aligned}\int_0^1 e^{\alpha(1-x^2)}\big[2\alpha x(x-1)\big]\frac{dx}{1-x}&=1-e^{\alpha},\\\int_0^1 e^{\alpha(1-x^2)}\big[2\alpha x(x-1)\big]^2\frac{dx}{2(1-x)}&=1-e^{\alpha}+\alpha\int_0^1 e^{\alpha(1-x^2)}dx,\\\int_0^1 e^{\alpha(1-x^2)}\big[2\alpha x(x-1)\big]^3\frac{dx}{6(1-x)}&=\frac{1}{3}\big[4-4e^{\alpha}-2\alpha e^{\alpha}\big]+2\alpha\int_0^1 e^{\alpha(1-x^2)}dx.\end{aligned}$$

Combining these facts with $e^x\geqslant 1+x+x^2/2+x^3/6$, we obtain

$$3\alpha^2e^{-\alpha}G'(\alpha)\geqslant 10e^{-\alpha}-2(5+\alpha)+(12-\alpha)\alpha\int_0^1 e^{-\alpha x^2}dx.$$

The right-hand side is positive for all $\alpha\in(-3,0)$ iff

$$\frac{10}{\beta(12+\beta)}\big(e^{\beta}-1\big)+\frac{2}{12+\beta}>\int_0^1 e^{\beta x^2}dx$$

for all $\beta\in(0,3)$. But

$$\begin{aligned}\int_0^1 e^{\beta x^2}dx-\int_0^1 e^{\beta x}dx&\leqslant-\int_0^1 e^{\beta x^2}\left(\beta x(1-x)+\frac{\beta^2x^2(1-x)^2}{2}\right)dx\\&=1-\frac{1}{8}e^{\beta}-\frac{7-2\beta}{8}\int_0^1 e^{\beta x^2}dx.\end{aligned}$$

Hence

$$\int_0^1 e^{\beta x^2}dx\leqslant\frac{1}{15-2\beta}\left[7+\left(\frac{8}{\beta}-1\right)\big(e^{\beta}-1\big)\right].$$

From these remarks, it should be easy to check that $G'(\alpha)>0$ for all $\alpha\in(-3,0)$.

As for the case that $\alpha\in(0,3)$, using $e^x\leqslant 1+x+x^2/2\,(x\leqslant 0)$, the proof is similar and even simpler. □

To prove Theorem 1.5, we need a lemma.

Lemma 2.2. Let $K < 0$. Then

$$I := \int_x^y \sum_{i=2}^d \Big(|\bigtriangledown_U W^i|^2 - \langle R(W^i, U)U, W^i\rangle \Big) \leqslant -2\sqrt{-K(d-1)}\tan\left(\frac{1}{2}\sqrt{\frac{-K}{d-1}}\rho(x,y)\right)$$

for $(x, y) \notin \mathbf{C}$.

Proof. Let $\gamma : [0, \rho(x, y)] \to M$ be the unique shortest geodesic from x to y and $\{e^i(s)\}_{i=1}^d$ be an orthornormal frame field which is parallel along γ: $e^1 = U$, $e^i(0) = W^i(x)$, $e^i(\rho(x, y)) = W^i(y)$, $i \geqslant 2$. Take

$$f(s) = \tan\left(\sqrt{\frac{-K}{d-1}}\cdot\frac{\rho}{2}\right)\sin\left(\sqrt{\frac{-K}{d-1}}\,s\right) + \cos\left(\sqrt{\frac{-K}{d-1}}\,s\right), \quad s \leqslant \rho(x,y).$$

Since $(x, y) \notin \mathbf{C}$, $\rho(x, y) < \pi/\sqrt{-K/(d-1)}$ (see [3], p.27–28). Thus, f is well defined on $(0, \rho(x, y)]$. Next, let $V^i(s) = f(s)e^i(s)$. By the first index lemma, we have

$$\begin{aligned}
I &\leqslant \int_x^y \sum_{i=2}^d \Big(|\bigtriangledown_U V^i|^2 - \langle R(V^i, U)U, V^i\rangle \Big) \\
&= \int_x^y \big[(d-1)f'(s)^2 - f(s)^2 \mathrm{Ric}_M(U,U)\big] \\
&\leqslant \int_0^{\rho(x,y)} \big[(d-1)f'(s)^2 + Kf(s)^2\big]ds \\
&= (d-1)\int_0^{\rho(x,y)} \big[f'(s)^2 + f''(s)f(s)\big]ds \\
&= (d-1)\big[(ff')(\rho(x,y)) - (ff')(0)\big] \\
&= -2\sqrt{-K(d-1)}\,\tan\left(\frac{1}{2}\sqrt{\frac{-K}{d-1}}\,\rho(x,y)\right). \quad \square
\end{aligned}$$

Proof of Theorem 1.5. a) Let $K < 0$. Combining Lemma 2.2 with (2.1), we obtain

$$\begin{aligned}
d\rho(x_t, y_t) \leqslant 2\sqrt{2}\,db_t - 2\sqrt{-K(d-1)}\,\tan\big[\sqrt{-K/(d-1)}\,\rho(x_t,y_t)/2\big]dt \\
+ a(\rho(x_t, y_t))dt, \quad t < T.
\end{aligned}$$

Then for the choice of F given in part (1) of Theorem 1.5, we have

$$dF(\rho(x_t, y_t)) \leqslant 2\sqrt{2}F'(\rho(x_t, y_t))db_t - 4dt, \quad t < T.$$

Hence $\mathbb{E}^{x,y}T \leqslant F(D)/4$, as we did in the proof of Lemma 2.1. Finally, the remainder of the proof is the same as those given in the proof of Theorem 1.4.

b) Let $K \geqslant 0$. It follows from Cranston (1991) that

$$\begin{aligned}\int_{x_t}^{y_t} \sum_{i=2}^{d} \Big(|\bigtriangledown_U W^i|^2 - \langle R(W^i, U)U, W^i\rangle \Big) dt \\ \leqslant 2\sqrt{K(d-1)}\, \frac{\cosh\big[\rho(x_t, y_t)\sqrt{K/(d-1)}\,\big] - 1}{\sinh\big[\rho(x_t, y_t)\sqrt{K/(d-1)}\,\big]} \\ = 2\sqrt{K(d-1)}\, \tanh\big[\rho(x_t, y_t)\sqrt{K/(d-1)}\,/2\big], \quad t < T.\end{aligned}$$

Combining this with (2.1), we obtain

$$\begin{aligned}d\rho(x_t, y_t) \leqslant 2\sqrt{2}\, db_t + 2\sqrt{K(d-1)}\, \tanh\big[\rho(x_t, y_t)\sqrt{K/(d-1)}\,/2\big] dt \\ + a(\rho(x_t, y_t))dt, \qquad t < T.\end{aligned}$$

Then, the remainder of the proof is the same as those given in the last paragraph. □

Proof of Corollary 1.7. If $|Z| \leqslant m < \infty$, then

$$|\langle Z(y_t), U\rangle - \langle Z(x_t), U\rangle| \leqslant 2|Z| \leqslant 2m.$$

Next, note that $\tanh r \leqslant r$ and

$$\tanh r \leqslant \frac{e^{2\Theta} - 1}{e^{2\Theta} + 1} < 1, \quad r \in [0, \Theta].$$

Replacing $\gamma(\xi)$ used in part (2) of Theorem 1.5 with

$$\gamma(\xi) = \big(2\sqrt{K(d-1)}\,\big) \wedge (K\xi) + 2m,$$

we obtain the first assertion in part (2) of Corollary 1.7. The last assertion of Corollary 1.7 follows from the facts:

$$\begin{aligned}D^2 \int_0^1 du \int_0^{1-u} ds \exp\left[\frac{1}{2} Dmu + \frac{D}{4} \int_s^{s+u} 2\sqrt{K(d-1)}\, d\xi\right] \\ = \frac{2D^2}{\beta^2}\left\{2\exp\left[\frac{\beta}{2}\right] - 2 - \beta\right\}, \quad \beta := D\big(\sqrt{K(d-1)} + m\big), \\ D^2 \int_0^1 du \int_0^{1-u} ds \exp\left[\frac{1}{2} Dmu + \frac{D}{4} \int_s^{s+u} DK\xi d\xi\right] \\ \leqslant D^2 \exp\left[\frac{1}{2} Dm\right] \int_0^1 du \int_0^{1-u} ds \exp\left[\frac{D^2 K}{8} u(u + 2s)\right],\end{aligned}$$

and the proof of Corollary 1.6 replacing $\alpha = D^2 K(Z)/8$ with $\alpha = D^2 K/8$. We have thus completed the proof of part (2) of Corollary 1.7. The proof of part (1) is similar and even simpler. □

In view of the above proofs, we see that Theorem 1.4 and Theorem 1.5 are usually not available for non-compact space since the diameter of the manifold is involved in the estimates.

Proof of Theorem 1.9. By condition (1),

$$\mathbb{E}^x u(x_t) = u(x) + \mathbb{E}^x \int_0^t Au(x_s)ds = u(x) + \lambda_1 \mathbb{E}^x \int_0^t u(x_s)ds.$$

The same equality holds for the (y_t)-process. Hence

$$|u(x) - u(y)| \leqslant \big|\mathbb{E}^x u(x_t) - \mathbb{E}^y u(y_t)\big| + \lambda_1 \int_0^t \big|\mathbb{E}^x u(x_s) - \mathbb{E}^y u(y_s)\big| ds. \tag{2.6}$$

Since the state space (E, ρ) and hence $(E, \bar{\rho})$ is separable and complete, by Chen (1992, Lemma 5.2), for each $s \geqslant 0$, x and y, one can choose a coupling probability measure $P^{s,x,y}$ of $P(x_s)$ and $P(y_s)$ such that

$$W(P(x_s), P(y_s)) = \int \bar{\rho}(x', y') P^{s,x,y}(dx', dy').$$

Denote by $E^{s,x,y}$ the expectation with respect to $P^{s,x,y}$. Then, by using the monotone class theorem and $|\mathbb{E}^x u(x_t)| < \infty$, we have

$$\mathbb{E}^x u(x_s) = E^{s,x,y} u. \tag{2.7}$$

Without loss of generality, by condition (2), assume that the Lipschitz constant of u with respect to $\bar{\rho}$ equals 1. Then, from condition (3) and (2.7), it follows that

$$\Big|\mathbb{E}^x u(x_s) - \mathbb{E}^y u(y_s)\Big| \leqslant \int \bar{\rho}(x', y') P^{s,x,y}(dx', dy') = W(P(x_s), P(y_s)) \leqslant \bar{\rho}(x, y) e^{-\sigma s}.$$

Combining this with (2.6), we get

$$|u(x) - u(y)| \leqslant \bar{\rho}(x, y) e^{-\sigma t} + \lambda_1 \bar{\rho}(x, y) \int_0^t e^{-\sigma s} ds. \tag{2.8}$$

Finally, choose $\{x^{(n)}, y^{(n)}\}$ such that

$$\frac{|u(x^{(n)}) - u(y^{(n)})|}{\bar{\rho}(x^{(n)}, y^{(n)})} \to 1, \quad \text{as } n \to \infty.$$

Now, the assertion follows from (2.8) by letting $n \to \infty$ and then $t \to \infty$. □

Proof of Theorem 1.8. a) Let $K < 0$. Set $\bar{\rho} = \sin\left(\frac{1}{2}\sqrt{\frac{-K}{d-1}}\,\rho\right)$. Clearly, $\bar{\rho}$ is an equivalent metric of ρ. Since

$$d\rho(x_t, y_t) \leqslant 2\sqrt{2}\, db_t - 2\sqrt{-K(d-1)}\, \tan\big[\sqrt{-K/(d-1)}\, \rho(x_t, y_t)/2\big] dt + H\rho(x_t, y_t) dt, \quad t < T. \tag{2.9}$$

By Itô formula, there exists a martingale M_t such that

$$\begin{aligned} d\bar{\rho}(x_t,y_t) &\leqslant dM_t + K\bar{\rho}(x_t,y_t)dt + \frac{1}{2}\sqrt{\frac{-K}{d-1}}\,H\rho(x_t,y_t)\cos\left(\frac{1}{2}\sqrt{\frac{-K}{d-1}}\,\rho(x_t,y_t)\right)dt \\ &\quad - \frac{-K}{d-1}\bar{\rho}(x_t,y_t)dt \\ &\leqslant dM_t + \left(\frac{d}{d-1}\,K + H\right)\bar{\rho}(x_t,y_t)dt. \end{aligned}$$

Here in the last step, we have used the inequality $x\cos x \leqslant \sin x$ and $H \geqslant 0$. From this we obtain

$$\mathbb{E}^{x,y}\bar{\rho}(x_t,y_t) \leqslant \bar{\rho}(x,y)\exp\left[\left(\frac{d}{d-1}\,K + H\right)t\right]$$

and so the part (1) of Theorem 1.8 follows from Theorem 1.9.

b) The proofs of part (2) and part (3) of Theorem 1.8 are similar. The only point is replacing (2.9) with (2.3) and choosing $\bar{\rho} = \sin\left(\frac{\pi}{2D}\rho\right)$. □

Add in Proof. After this paper was finished, the same problem for the same operator on the manifold with Dirichlet or Neumann boundary condition and for Schrödinger operator has been studied by the second author. The picture of the bounds of the first eigenvalue is now largely extended.

Acknowledgement. The authors would like to thank the referees for their helpful comments on the earlier version of the paper. The paper is submited to Sci. Sin. The authors acknowlege Prof. D. A. Dawson for the support to including this paper in the Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University and University of Ottawa, No.215, 1993.

References

[1]. 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.
[2]. Chavel, I. (1984), *Eigenvalues in Riemannian Geometry*, Academic Press.
[3]. Cheeger, J. and Ebin, D. G. (1975), *Comparison Theorems in Riemannian Geometry*, North-Holland.
[4]. 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.
[5]. Chen, M. F. (1992), *From Markov Chains to Non-Equilibrium Particle Systems*, World Scientific.
[6]. Chen, M. F. and Li, S. F. (1989), *Coupling methods for milti-dimensional diffusion processes*, Ann. of Probab. 17:1, 151–177.
[7]. Cranston, M. (1991), *Gradient estimates on manifolds using coupling*, J. Funct. Anal. 99, 110–124.
[8]. Deuschel, J. -D. and Stroock, D. W. (1989), *Large Deviations*, Academic Press, Boston.
[9]. 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.

[10]. Kendall, W. (1986 a), *Stochastic differential geometry, a coupling property, and harmonic maps*, J. London Math. Soc. 33, 554–566.
[11]. Kendall, W. (1986 b), *Nonnegative Ricci curvature and the Brownian coupling property*, Stochastics 19, 111–129.
[12]. 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.
[13]. Lindvall, T. and Rogers, L. (1986), *Coupling of milti-dimensional diffusions by reflection*, Ann. of Probab. 14, 860–872.
[14]. Schoen, R. and Yau, S. T. (1988), *Differential Geometry (In Chinese)*, Science Press, Beijing, China.
[15]. Yang, H. C. (1989), *Estimate of the first eigenvalue of a compact Riemannian manifold with Ricci curvature bounded below by a negative constant (In Chinese)*, Sci. Sin.(A) 32:7, 698–700.
[16]. Zhong, J. Q. and Yang, H. C. (1984), *Estimates of the first eigenvalue of a compact Riemannian manifolds*, Sci. Sin. 27:12, 1251–1265.

Department of Mathematics, Beijing Normal University, Beijing 100875,
The People's Republic of China.

Acta Math. Sinica (New Series), 10:3, 260-275, 1994

OPTIMAL MARKOVIAN COUPLINGS AND APPLICATIONS

Mu-Fa Chen

(Beijing Normal University)
Received September 11, 1993

Abstract. This paper is devoted to studying a new topic: optimal Markovian couplings, mainly for time-continuous Markov processes. The study emphasizes the analysis of the coupling operators rather than the processes. Some constructions of optimal Markovian couplings for Markov chains and diffusions are presented, which are often unexpected. Then, the results are applied to study the L^2-convergence for Markov chains and for a diffusion on compact manifold. The estimate of the convergent rate provided by this method can be sharp.

1. Introduction. Markovian Couplings

Let us recall the simple definition of couplings.

Definition 1.1. Let P_k be a probability measure on a measurable space $(E_k, \mathscr{E}_k)$, $k = 1, 2$. A probability measure $\widetilde{P}$ on $(E_1 \times E_2, \mathscr{E}_1 \times \mathscr{E}_2)$ is called a **coupling of P_1 and P_2** if it has the following **marginality**:

$$\widetilde{P}(B_1 \times E_2) = P_1(B_1), \quad \widetilde{P}(E_1 \times B_2) = P_2(B_2), \ \ B_k \in \mathscr{E}_k, \ k = 1, 2.$$

Similarly, we can define a coupling process of two stochastic processes in terms of their distributions at each time t.

In the past two decades or more, the coupling methods have attracted a lot of attentions by many authors. Now the methods have a very wide range of applications. Refer to Liggett (1985), Chen (1992) and Lindvall (1992) for more details and references. The optimalization of couplings was started by Griffeath (1975), where the maximal coupling for time-discrete Markov chains was introduced. However, the maximal couplings are usually non-Markovian. Certainly, the non-Markovian couplings (and even the couplings for non-Markov processes) now consist of an important part of the theory, but they are difficult to handle

2000 *Mathematics Subject Classification.* 60J75, 60J25, 58G32, 58G25.

Key words and phrases. Optimal coupling, Markovian coupling, spectral gap.

Research supported in part by NSFC, the State Education Commission of China, the NSERC operating grant of D. A. Dawson and Centro Vito Volterra

when we come to the time-continuous situation. This paper adopts a different point of view. Roughly speaking, in contrast to the maximal coupling which concentrates on discrete metric (or total variation), we insist on the Markovian couplings and study the optimal problem for various metrics. This enables us to analyze mainly the operators (i.e., the formal generators) rather than the coupling processes. It turns out that the ρ-optimal Markovian coupling (abbrev. ρ-OMC) can still attain the global optimum (i.e., without restricting to Markovian) for some refined metric ρ.

The paper is organized as follows. Based on the relation between couplings and probability metrics, the optimality notion is introduced in Section 2. Then, some OMCs are constructed in Sections 3 — 5 respectively for time-continuous, time-discrete Markov chains and diffusions in $\mathbf{R}^d$. The resulting couplings are often unexpected. Finally, the OMCs are applied to study the estimates of the L^2-convergent rate, which is a recent topic of the application of coupling method, proposed by Chen and Wang (1992). We study the problem for Markov chains and as an addition to the last quoted paper, a new lower bound of the spectral gap of Laplacian on compact manifold is also presented (Section 6). Refer to the survey article [Chen (1993)] for the backgrounds of the study and for more information on the applications. In the remainder of this section, we review some necessary notations and results about jump processes and their couplings .

A **jump process** means a sub-Markovian transition function $P(t,x,A)$ ($x \in E, A \in \mathscr{E}$) which satisfies the **jump condition**: $\lim_{t\to 0} P(t,x,\{x\}) = 1$ for all $x \in E$. Throughout this paper, we are interested only in the **totally stable** and **conservative** case. That is, the limits

$$q(x) := \lim_{t\to 0} \frac{1-P(t,x,\{x\})}{t} \quad \text{and} \quad q(x,A) := \lim_{t\to 0} \frac{P(t,x,A\setminus\{x\})}{t}, \quad x \in E,\ A \in \mathscr{E} \tag{1.1}$$

satisfy $q(x,E) = q(x) < \infty$, for all $x \in E$. Then, the transition rate $(q(x), q(x,dy))$ is called a **q-pair**, which is called **regular** if it determines uniquely a jump process satisfying (1.1). When E is countable, traditionally we use the matrices $Q = (q_{ij} : i,j \in E)$ and $P(t) = (p_{ij}(t) : i,j \in E)$ instead of the q-pair and the jump process respectively. Here $q_{ii} = -q_i,\ i \in E$. We also call $P(t)$ a **Markov chain**.

Next, given two jump processes $P_k(t)$ with q-pairs $(q_k(x_k), q_k(x_k,dy_k))$, $k = 1, 2$. Let $\widetilde{P}(t;x_1,x_2;dy_1,dy_2)$ be a coupling jump process of $P_1(t)$ and $P_2(t)$. That is, by Definition 1.1,

$$\begin{aligned} &\widetilde{P}(t;x_1,x_2;A_1\times E_2) = P_1(t,x_1,A_1),\\ &\widetilde{P}(t;x_1,x_2;E_1\times A_2) = P_2(t,x_2,A_2),\\ &\qquad t \geqslant 0,\ x_k \in E_k,\ A_k \in \mathscr{E}_k,\ k = 1, 2. \end{aligned} \tag{1.2}$$

The corresponding q-pair is denoted by $(\tilde{q}(x_1,x_2),\ \tilde{q}(x_1,x_2;\,dy_1,dy_2))$. Let ${}_b\mathscr{E}$ be the set of all bounded $\mathscr{E}$-measurable functions. Define

$$\Omega_1 f(x_1) = \int q_1(x_1,dy_1)[f(y_1) - f(x_1)], \qquad f \in {}_b\mathscr{E}_1.$$

Similarly, we have Ω_2 and $\widetilde{\Omega}$. Because of the one-to-one correspondence between a q-pair and its operator Ω, we will use both according to our convenience. Now, it is not difficult to prove (cf. [1] or [5; Chapter 5]) that the marginality (1.2) implies the following:

$$\begin{aligned} &\widetilde{\Omega}f(x_1,x_2)=\Omega_1 f(x_1), \qquad f\in {}_b\mathscr{E}_1 \\ &\widetilde{\Omega}f(x_1,x_2)=\Omega_2 f(x_2), \qquad f\in {}_b\mathscr{E}_2,\ x_k\in E_k,\ k=1,\,2, \end{aligned} \tag{1.3}$$

where on the left-hand side, $f\in {}_b\mathscr{E}_k\,(k=1,2)$ is regarded as a function in ${}_b(\mathscr{E}_1\times\mathscr{E}_2)$.

Definition 1.2. Any operator $\widetilde{\Omega}$ satisfying (1.3) is called a **coupling operator**.

In practice, it is quite easy to find out some coupling operators. To see this and also for the later use, we recall some coupling operators for Markov chains. In addition to the well-known **classical** or **Doeblin's coupling** $\widetilde{\Omega}_c$ and the **basic** or **Wasserstein's coupling** $\widetilde{\Omega}_b$, we mention two more examples as follows:

Example 1.3 (March coupling[2] $\widetilde{\Omega}_m$**).** Take $E=\{0,1,2,\cdots\}$ and let

$$\begin{aligned} (i_1,\,i_2)\to(i_1+k,\,i_2+k) \quad &\text{at rate} \quad q^{(1)}_{i_1,i_1+k}\wedge q^{(2)}_{i_2,i_2+k} \\ \to(i_1+k,\,i_2) \quad &\text{at rate} \quad \big(q^{(1)}_{i_1,i_1+k}-q^{(2)}_{i_2,i_2+k}\big)^+ \\ \to(i_1,\quad i_2+k) \quad &\text{at rate} \quad \big(q^{(2)}_{i_2,i_2+k}-q^{(1)}_{i_1,i_1+k}\big)^+, \qquad i_1,\,i_2\in E. \end{aligned}$$

here we have used the convention that $q_{ij}=0$ for all $i\in E$ and $j\notin E$.

The key of this coupling is the term $q^{(1)}_{i_1,i_1+k}\wedge q^{(2)}_{i_2,i_2+k}$. Whenever a term $A\wedge B$ appears, we should have the other two terms $(A-B)^+$ and $(B-A)^+$ automatically, due to the marginality. Thus, in what follows, we will write down the term $A\wedge B$ only for simplicity. The word "march" is a Chinese name, which is the command to soldiers to start marching. One of the original purpose to introduce this coupling is for the order-preservation.

Let us now consider a birth-death process with regular Q-matrix: $q_{i,i+1}=b_i$, $i\geqslant 0$; $q_{i,i-1}=a_i$, $i\geqslant 1$. Then for two copies of the process starting from i_1 and i_2 respectively, we have

Example 1.4 (Coupling by inner reflection[3] $\widetilde{\Omega}_{ir}$**).** Again, take $\widetilde{\Omega}_{ir}=\widetilde{\Omega}_c$ if $|i_1-i_2|\leqslant 1$. For $i_2\geqslant i_1+2$, take

$$\begin{aligned} (i_1,\,i_2)\to(i_1+1,\,i_2-1) \quad &\text{at rate} \quad b_{i_1}\wedge a_{i_2} \\ \to(i_1-1,\,i_2) \quad &\text{at rate} \quad a_{i_1} \\ \to(i_1,\,i_2+1) \quad &\text{at rate} \quad b_{i_2}. \end{aligned}$$

By exchanging i_1 and i_2, we can get the expression of $\widetilde{\Omega}_{ir}$ for the case that $i_1\geqslant i_2$.

The next result is a starting point of the present study, which reduces the coupling jump processes to the rather simple coupling operators.

Theorem 1.5[1]**.** The marginal q-pairs are regular iff so is a (equivalently, any) coupling q-pair. Moreover, (1.2) and (1.3) are equivalent.

2. Optimal Markovian Couplings

Let us recall a probability metric. Let $(E, \rho, \mathscr{E})$ be a metric space. The minimum L^1-metric W is defined by:

$$W(P_1, P_2) = \inf_{\widetilde{P}} \int \rho(x_1, x_2)\widetilde{P}(dx_1, dx_2), \tag{2.1}$$

where $\widetilde{P}$ varies over all couplings of P_1 and P_2. This metric has many different names. It plays a critical role in the study of random fields and interacting particle systems. Here, we mention a result due to Dobrushin (1970): W is equivalent to the Lévy-Prohorov metric when ρ is bounded and W equals half of the total variation when ρ is the discrete metric d: $d(x, y) = 0$ if $x = y$ and $d(x, y) = 1$ if $x \neq y$. Refer to [5; Chapter 0 and Chapter 5] for more information about W. In view of (2.1), we see that every coupling provides an upper bound of $W(P_1, P_2)$. Thus, it is natural to introduce the following

Definition 2.1. A coupling $\overline{P}$ of P_1 and P_2 is called **ρ-optimal** if it attains the infimum on the right-hand side of (2.1).

For a complete separable metric space $(E, \rho, \mathscr{E})$, a ρ-optimal coupling (abbrev. ρ-OC) does exist (cf., [5; Lemma 5.2]), but may not be unique. In the special case of ρ being discrete metric, the ρ-OC is just the maximal coupling mentioned before.

Certainly, one may replace the above P_k by $P_k(t, x_k, dy_k)$, $k = 1, 2$ and define a ρ-OC $\overline{P}(t; x_1, x_2; dy_1, dy_2)$. But this definition is usually not useful since it is not practical. We will emphasize the coupling operators. Consider jump processes again. As usual, for a jump process $P(t, x, dy)$, denote by $P(t)$ the corresponding semi-group on ${}_b\mathscr{E}$. We want to find out a coupling process $\overline{P}(t)$ such that for any coupling process $\widetilde{P}(t)$, $\overline{P}(t)\rho(x_1, x_2) \leqslant \widetilde{P}(t)\rho(x_1, x_2)$ for all (x_1, x_2). The next result reduces the comparison of two semi-groups to the one of their operators. From the proof below, it should be clear that under some mild condition, the conclusion also holds for other type of Markov processes.

Lemma 2.2. Let $P_k(t)$ be a regular jump process with state space $(E, \mathscr{E})$ and q-pair $(q_k(x),\ q_k(x_k, dy_k))$, $k = 1, 2$. Suppose that there exist constants C and c such that

$$\Omega_k\rho(x, a) \leqslant C + c\,\rho(x, a), \qquad x \in E,\ k = 1, 2 \tag{2.2}$$

for some fixed $a \in E$. Given two couplings $\overline{P}(t)$ and $\widetilde{P}(t)$ of $P_1(t)$ and $P_2(t)$. If

$$\overline{P}(t)\rho(x_1, x_2) \leqslant \widetilde{P}(t)\rho(x_1, x_2), \qquad t \geqslant 0,\ x_1,\ x_2 \in E,$$

then we have $\overline{\Omega}\,\rho(x_1, x_2) \leqslant \widetilde{\Omega}\,\rho(x_1, x_2), \quad x_1,\ x_2 \in E$.

Proof. a) Without loss of generality, assume that C, $c > 0$. By (2.2), we have

$$P_k(t)\rho(x, a) \leqslant C\big[e^{ct} - 1\big]/c + e^{ct}\rho(x, a), \qquad x \in E \tag{2.3}$$

(cf., Chen [1992; Lemma 4.13]). Hence, for any coupling semi-group $\widetilde{P}(t)$, we have

$$\begin{aligned}\widetilde{P}(t)\rho(x_1,x_2) &\leqslant \widetilde{P}(t)\rho(x_1,a)+\widetilde{P}(t)\rho(x_2,a)\\ &\leqslant 2C\big[e^{ct}-1\big]/c+e^{ct}\big[\rho(x_1,a)+\rho(x_2,a)\big],\\ &\qquad x_1,\ x_2\in E.\end{aligned}\tag{2.4}$$

b) By Theorem 1.5, for every $f\in {}_b\mathscr{E}$, we have

$$\widetilde{P}(t)f(x_1,x_2)-f(x_1,x_2)=\int_0^t\widetilde{\Omega}\widetilde{P}(s)f(x_1,x_2)ds.$$

In particular,

$$\widetilde{P}(t)\rho_n(x_1,x_2)-\rho_n(x_1,x_2)=\int_0^t\widetilde{\Omega}\widetilde{P}(s)\rho_n(x_1,x_2)ds,\tag{2.5}$$

where $\rho_n=\rho\wedge n$. Moreover, by (2.4) and the marginality, we have

$$\begin{aligned}0&\leqslant\int\widetilde{q}(x_1,x_2;dy_1,dy_2)\widetilde{P}(t)\rho_n(y_1,y_2)\\ &\leqslant\int\widetilde{q}(x_1,x_2;dy_1,dy_2)\widetilde{P}(t)\rho(y_1,y_2)\\ &\leqslant\int\widetilde{q}(x_1,x_2;dy_1,dy_2)\widetilde{P}(t)\big(\rho(y_1,a)+\rho(y_2,a)\big)\\ &\leqslant 2C\widetilde{q}(x_1,x_2)\big[e^{ct}-1\big]/c+e^{ct}\int\widetilde{q}(x_1,x_2;dy_1,dy_2)\big(\rho(y_1,a)+\rho(y_2,a)\big)\\ &=2C\widetilde{q}(x_1,x_2)\big[e^{ct}-1\big]/c+e^{ct}\Big[\widetilde{q}(x_1,x_2)\big(\rho(x_1,a)+\rho(x_2,a)\big)\\ &\quad+\Omega_1\rho(x_1,a)+\Omega_2\rho(x_2,a)\Big].\end{aligned}$$

Combining this with (2.2), we see that for fixed x_1 and x_2,

$$\int\widetilde{q}(x_1,x_2;dy_1,dy_2)\widetilde{P}(t)\rho_n(y_1,y_2)$$

is bounded on finite t-interval uniformly in n. By (2.5), it follows that

$$\widetilde{P}(t)\rho(x_1,x_2)-\rho(x_1,x_2)=\int_0^t\widetilde{\Omega}\widetilde{P}(s)\rho(x_1,x_2)ds.$$

Furthermore, $\lim_{t\to0}\widetilde{P}(t)\rho(x_1,x_2)=\rho(x_1,x_2)$ and then by using (2.4) and the dominated convergence theorem, we get

$$\lim_{t\to0}\int\widetilde{q}(x_1,x_2;dy_1,dy_2)\widetilde{P}(t)\rho(y_1,y_2)=\int\widetilde{q}(x_1,x_2;dy_1,dy_2)\rho(y_1,y_2).$$

Therefore,

$$\lim_{t\to0}\frac{\widetilde{P}(t)\rho(x_1,x_2)-\rho(x_1,x_2)}{t}=\lim_{t\to0}\frac{1}{t}\int_0^t\widetilde{\Omega}\widetilde{P}(s)\rho(x_1,x_2)ds=\widetilde{\Omega}\,\rho(x_1,x_2).$$

From this, the required assertion follows immediately. □

The above result leads to the following definition:

Definition 2.3. A coupling operator $\overline{\Omega}$ is called **ρ-optimal** if

$$\overline{\Omega}\,\rho(x_1,x_2)=\inf_{\widetilde{\Omega}}\widetilde{\Omega}\,\rho(x_1,x_2)$$

for all x_1 and x_2, where $\widetilde{\Omega}$ varies over all coupling operators.

For the existence of a ρ-OC (we may omit the phase "Markovian" since we are dealing with operators), the next result is provided by S. Y. Zhang to the author, the proof is omitted here.

Theorem 2.4. For Markov chains, if (2.2) holds, then there does exist a ρ-OC.

3. Time-Continuous Markov Chains

Starting from this section, we construct some OCs. To be precise, we are interested in those coupling having the following properties:

(1) **Marginality**: That is (1.3).
When $\Omega_1=\Omega_2=\Omega$, we require the following
(2) **Normality**: $\widetilde{\Omega}f(x,x)=\Omega g(x)$, where $g(x):=f(x,x)$.
Finally, if $\Omega_1=\Omega_2=\Omega$, it is natural to require the following
(3) **Symmetry**: $\widetilde{\Omega}f(x_1,x_2)=\widetilde{\Omega}f(x_2,x_1)$ for all $f\in{}_b(\mathscr{E}\times\mathscr{E})$, all x_1 and x_2.

In this section, we deal with OC for birth-death processes. To do so, we need one more coupling:

Definition 3.1. Given a birth-death process with birth rates b_i and death rates a_i. The coupling by reflection $\overline{\Omega}_r$ evolves in the following way:

$$\begin{array}{lll}
(i_1,\,i_2)\to(i_1-1,\,i_2+1) & \text{at rate} & a_{i_1}\wedge b_{i_2}\\
\qquad\to(i_1+1,\,i_2) & \text{at rate} & b_{i_1}\\
\qquad\to(i_1,\,i_2-1) & \text{at rate} & a_{i_2},\qquad \text{if } i_2=i_1+1.\\
(i_1,\,i_2)\to(i_1-1,\,i_2+1) & \text{at rate} & a_{i_1}\wedge b_{i_2}\\
\qquad\to(i_1+1,\,i_2-1) & \text{at rate} & b_{i_1}\wedge a_{i_2},\quad \text{if } i_2\geqslant i_1+2.
\end{array}$$

By symmetry, we can write down the rates for the case that $i_1>i_2$.

Intuitively, the reflection in outside direction is quite strange since it makes the components apart by distance 2 but not by 1. For this reason, it is hardly believed that $\overline{\Omega}_r$ could be better than $\widetilde{\Omega}_{ir}$. Here, I would like to acknowledge X. L. Wang for pointed an error on the original computation of part (1) below and E. A. Perkins for a question which leads to the part (2) below.

Theorem 3.2. Let ρ be a translation-invariant metric on $\mathbf{Z}_+$ and set $u_k:=\rho(k+1)-\rho(k)$, $k\geqslant 0$, where $\rho(k)=\rho(0,k)$. Then, for birth-death process,

(1) $\overline{\Omega}_r$ is ρ-optimal whenever u_k is decreasing in k. Moreover, we have for $i_2-i_1=:k\geqslant 1$,

$$\begin{aligned}
&\overline{\Omega}_r\rho(k)\\
&=\begin{cases}(a_{i_1}\wedge b_{i_2})u_2+(a_{i_1}\vee b_{i_2})u_1-(b_{i_1}+a_{i_2})u_0, & \text{if } k=1\\ (a_{i_1}\wedge b_{i_2})u_{k+1}+(a_{i_1}\vee b_{i_2})u_k-(b_{i_1}\vee a_{i_2})u_{k-1}-(b_{i_1}\wedge a_{i_2})u_{k-2}, & \text{if } k\geqslant 2.\end{cases}
\end{aligned}$$

(2) If u_k is increasing in k, then $\widetilde{\Omega}_m$ is ρ-optimal. Moreover,

$$\widetilde{\Omega}_m\rho(k) = \big[(a_{i_1} - a_{i_2})^+ + (b_{i_2} - b_{i_1})^+\big]u_k - \big[(a_{i_2} - a_{i_1})^+ + (b_{i_1} - b_{i_2})^+\big]u_{k-1},$$

provided $i_2 - i_1 =: k \geqslant 1$.

Proof. a) Clearly, any coupling operator $\widetilde{\Omega}$ should have the following form:

$$\begin{aligned}
&\widetilde{\Omega}f(i_1,i_2)\\
&= I_{[i_1\neq i_2]}\Big\{\lambda_1\big[f(i_1-1,i_2-1)-f(i_1,i_2)\big] + \lambda_2\big[f(i_1+1,i_2+1)-f(i_1,i_2)\big]\\
&\qquad + \lambda_3\big[f(i_1+1,i_2)-f(i_1,i_2)\big] + \lambda_4\big[f(i_1-1,i_2)-f(i_1,i_2)\big]\\
&\qquad + \lambda_5\big[f(i_1,i_2+1)-f(i_1,i_2)\big] + \lambda_6\big[f(i_1,i_2-1)-f(i_1,i_2)\big]\\
&\qquad + \lambda_7\big[f(i_1+1,i_2-1)-f(i_1,i_2)\big] + \lambda_8\big[f(i_1-1,i_2+1)-f(i_1,i_2)\big]\Big\}\\
&\quad + I_{[i_1=i_2]}\Big\{b_{i_1}\big[f(i_1+1,i_2+1)-f(i_1,i_2)\big] + a_{i_1}\big[f(i_1-1,i_2-1)-f(i_1,i_2)\big]\Big\},
\end{aligned}$$

where $\lambda_j \geqslant 0$ and

$$\lambda_1 = \lambda_4 = \lambda_8 = 0 \quad \text{if} \quad i_1 = 0, \qquad \lambda_1 = \lambda_6 = \lambda_7 = 0 \quad \text{if} \quad i_2 = 0. \tag{3.1}$$

By the marginality, we have

$$\begin{cases} \lambda_1 + \lambda_4 + \lambda_8 = a_{i_1} \\ \lambda_2 + \lambda_3 + \lambda_7 = b_{i_1} \\ \lambda_1 + \lambda_6 + \lambda_7 = a_{i_2} \\ \lambda_2 + \lambda_5 + \lambda_8 = b_{i_2}. \end{cases} \tag{3.2}$$

b) By symmetry, we may assume that $i_1 < i_2$ and let $k = i_2 - i_1$. If $k \geqslant 2$, then

$$\begin{aligned}
\widetilde{\Omega}\,\rho(k) =& \lambda_3\big[\rho(k-1)-\rho(k)\big] + \lambda_4\big[\rho(k+1)-\rho(k)\big]\\
&+ \lambda_5\big[\rho(k+1)-\rho(k)\big] + \lambda_6\big[\rho(k-1)-\rho(k)\big]\\
&+ \lambda_7\big[\rho(k-2)-\rho(k)\big] + \lambda_8\big[\rho(k+2)-\rho(k)\big].
\end{aligned} \tag{3.3}$$

We now minimize $\widetilde{\Omega}\,\rho(k)$ under the marginality. Since λ_1 and λ_2 disappeared in the expression of $\widetilde{\Omega}\,\rho(k)$, we eliminate them from (3.2) and obtain

$$\lambda_4 + \lambda_8 - \lambda_6 - \lambda_7 = a_{i_1} - a_{i_2}, \qquad \lambda_3 + \lambda_7 - \lambda_5 - \lambda_8 = b_{i_1} - b_{i_2}.$$

Hence

$$\lambda_4 = a_{i_1} - a_{i_2} + \lambda_6 + \lambda_7 - \lambda_8, \qquad \lambda_5 = b_{i_2} - b_{i_1} + \lambda_3 + \lambda_7 - \lambda_8. \tag{3.4}$$

Substituting this into (3.3), we get

$$\begin{aligned}\widetilde{\Omega}\rho(k) &= \big(a_{i_1}-a_{i_2}+b_{i_2}-b_{i_1}\big)\big[\rho(k+1)-\rho(k)\big]+(\lambda_3+\lambda_6)\big[\rho(k+1)+\rho(k-1)-2\rho(k)\big]\\ &\quad+\lambda_7\big[2\rho(k+1)+\rho(k-2)-3\rho(k)\big]+\lambda_8\big[\rho(k+2)-2\rho(k+1)+\rho(k)\big]\\ &= \big(a_{i_1}-a_{i_2}+b_{i_2}-b_{i_1}\big)u_k+(\lambda_3+\lambda_6)(u_k-u_{k-1})\\ &\quad+\lambda_7(2u_k-u_{k-1}-u_{k-2})+\lambda_8(u_{k+1}-u_k). \qquad (3.5)\end{aligned}$$

c) Again, let $k \geqslant 2$. First, consider part (1) of the theorem. By (3.5), the coefficients of $(\lambda_3+\lambda_6)$, λ_7 and λ_8 are all non-positive, so we should make these $\lambda_j' s$ as large as possible. On the other hand, since $2u_k-u_{k-1}-u_{k-2} \leqslant u_k-u_{k-1}$, the contribution made by λ_7 is bigger than those made by $(\lambda_3+\lambda_6)$. Combining this with the marginality, we see that we should handle λ_7 first rather than $\lambda_3+\lambda_6$. By using the marginality again, the largest choice of λ_7 is $b_{i_1} \wedge a_{i_2}$. Then (3.2) gives us $\lambda_1+\lambda_6 = (a_{i_2}-b_{i_1})^+$, so the largest choice of λ_6 is $(a_{i_2}-b_{i_1})^+$ and then $\lambda_1 = 0$. The same argument gives us $\lambda_3 = (b_{i_1}-a_{i_2})^+$ and $\lambda_2 = 0$. Similarly, we choose $\lambda_8 = a_{i_1} \wedge b_{i_2}$ and then $\lambda_4 = (a_{i_1}-b_{i_2})^+$ and $\lambda_5 = (b_{i_2}-a_{i_1})^+$. Obviously, for this choice of λ_j's, (3.1) holds. Next, consider part (2). By (3.5), we should make λ_3, λ_6, λ_7 and λ_8 as small as possible. Note that λ_3, λ_6 and λ_7 can not vanish simultaneously due to the marginality. From which, we see that λ_3, λ_6 and λ_7 are smaller provided λ_1 and λ_2 are bigger and moreover $2u_k-u_{k-1}-u_{k-2} \geqslant u_k-u_{k-1}$. Thus, we should make λ_7 smaller rather than λ_3 and λ_6 if possible. These considerations lead to the following choice:

$$\lambda_1 = a_{i_1} \wedge a_{i_2},\ \ \lambda_2 = b_{i_1} \wedge b_{i_2},\ \ \lambda_7 = 0,\ \ \lambda_3 = (b_{i_1}-b_{i_2})^+,\ \ \lambda_6 = (a_{i_2}-a_{i_1})^+.$$

Similarly, we have $\lambda_8 = 0$, $\lambda_4 = (a_{i_1}-a_{i_2})^+$, $\lambda_5 = (b_{i_2}-b_{i_1})^+$. We have thus proved the theorem in the case that $k \geqslant 2$.

d) Let $k = 1$. Then the argument of the first part of b) leads to the following:

$$\widetilde{\Omega}\,\rho(k) = \big(a_{i_1}-a_{i_2}+b_{i_2}-b_{i_1}\big)u_k+\lambda_8(u_{k+1}-u_k)+(\lambda_3+\lambda_6)(u_k-u_{k-1})+2\lambda_7 u_k.$$

For part (1), we choose λ_3 and λ_6 so that the right-hand side becomes as small as possible. From $\lambda_1+\lambda_6+\lambda_7 = a_{i_2}$ and $\lambda_2+\lambda_3+\lambda_7 = b_{i_1}$, we obtain $\lambda_6 = a_{i_2}$, $\lambda_3 = b_{i_1}$ and $\lambda_1 = \lambda_2 = \lambda_7 = 0$. Furthermore, $\lambda_8 = a_{i_1} \wedge b_{i_2}$ and then $\lambda_4 = (a_{i_1}-b_{i_2})^+$ and $\lambda_5 = (b_{i_2}-a_{i_1})^+$.

For part (2), the optimal solution is the same as in the second part of the proof of c). □

To see that the OCs may not be unique, consider the discrete metric: $u_0 = 1$ and $u_k = 0$ for all $k \geqslant 1$. Then from the above theorem, we obtain

$$\overline{\Omega}_r\rho(k) = \begin{cases} -(b_{i_1}+a_{i_2}), & \text{if } k = 1\\ -(b_{i_1} \wedge a_{i_2}), & \text{if } k = 2\\ 0, & \text{if } k \geqslant 3.\end{cases}$$

Thus, $\widetilde{\Omega}_b$ (and of course, $\widetilde{\Omega}_{ir}$ or $\overline{\Omega}_r$) is ρ-optimal but not $\widetilde{\Omega}_m$. Next, consider the ordinary metric: $u_k \equiv 1$. We obtain

$$\overline{\Omega}_r\rho(k) = (a_{i_1}+b_{i_2})-(b_{i_1}+a_{i_2}), \qquad k := i_2-i_1 \geqslant 1.$$

Therefore, the five couplings mentioned above achieve the same minimum. The next result is much more surprising, it says that the last conclusion actually holds for a large class of metrics.

Theorem 3.3. Let (u_k) be a positive sequence on $\mathbf{Z}_+$ and set $F(k) = \sum_{j<k} u_j$. Define $\rho(m,n) = |F(m)-F(n)|$. Then, every coupling mentioned above is ρ-optimal. Moreover,

$$\widetilde{\Omega}\rho(i,j) = b_j u_j - a_j u_{j-1} - b_i u_i + a_i u_{i-1}, \qquad u_{-1} := 1. \tag{3.6}$$

Proof. a) Given (i,j) : $j-2 \geqslant i \geqslant 1$ and a coupling operator $\widetilde{\Omega}$, we have

$$\begin{aligned}\widetilde{\Omega}\rho(i,j) &= \lambda_1\big[-u_{j-1}+u_{i-1}\big] + \lambda_2\big[u_j-u_i\big] + \lambda_3\big[-u_i\big] + \lambda_4 u_{i-1} \\ &\quad + \lambda_5 u_j + \lambda_6\big[-u_{j-1}\big] + \lambda_7\big[-u_{j-1}-u_i\big] + \lambda_8\big[u_j+u_{i-1}\big] \\ &= (a_i-a_j)u_{i-1} + (b_j-b_i)u_j + \lambda_1\big[-u_{j-1}+u_{i-1}\big] + \lambda_2\big[u_j-u_i\big] \\ &\quad + \lambda_3\big[-u_i+u_j\big] + \lambda_6\big[-u_{j-1}+u_{i-1}\big] + \lambda_7\big[-u_{j-1}-u_i+u_{i-1}+u_j\big].\end{aligned}$$

Here in the last step, we have used (3.4). Collecting the terms together and applying the marginality, we get

$$\begin{aligned}\widetilde{\Omega}\rho(i,j) &= (a_i-a_j)u_{i-1} + (b_j-b_i)u_j + (\lambda_1+\lambda_6+\lambda_7)\big[u_{i-1}-u_{j-1}\big] \\ &\quad + (\lambda_2+\lambda_3+\lambda_7)\big[u_j-u_i\big] \\ &= (a_i-a_j)u_{i-1} + (b_j-b_i)u_j + a_j\big[u_{i-1}-u_{j-1}\big] + b_i\big[u_j-u_i\big] \\ &= a_i u_{i-1} + b_j u_j - a_j u_{j-1} - b_i u_i, \qquad j-2 \geqslant i \geqslant 1. \end{aligned} \tag{3.7}$$

b) Next, let $j-2 \geqslant i = 0$. Then by the marginality, we have $\lambda_1 = \lambda_4 = \lambda_8 = 0$. Hence

$$\begin{aligned}\widetilde{\Omega}\rho(i,j) &= \lambda_2\big[u_j-u_i\big] + \lambda_3\big[-u_i\big] + \lambda_5 u_j + \lambda_6\big[-u_{j-1}\big] + \lambda_7\big[-u_{j-1}-u_i\big] \\ &= (b_j-b_i)u_j + (\lambda_2+\lambda_3+\lambda_7)\big[u_j-u_i\big] - (\lambda_6+\lambda_7)u_{j-1} \\ &= (b_j-b_i)u_j + b_i\big[u_j-u_i\big] - (a_j-a_i)u_{j-1} \\ &= b_j u_j - a_j u_{j-1} - b_0 u_0, \qquad j-2 \geqslant i = 0. \end{aligned} \tag{3.8}$$

Here, we have also used (3.4) and the marginality.

c) Combining (3.7) with (3.8), we get (3.6) in the case of $j \geqslant i+2$. If $j = i+1$, (3.6) still holds whenever $\lambda_7 = 0$. Since the right-hand side of (3.6) is independent of $\lambda_j' s$, we obtain the required assertion. □

4. Time-Discrete Markov Chains

We now consider the well-studied time-discrete case. The definition of ρ-OMC in this case is an easy modification of Definition 2.1 with the restriction on Markovian couplings. It is interesting that even in a simple situation, the OMC is still not so obvious and not known before.

Theorem 4.1. Take $E = \mathbf{Z}$ and let $P = (P_{ij})$ be a random walk on $\mathbf{Z}$ with

$$P_{i,i+1} = p_i > 0, \quad P_{i,i-1} = q_i > 0, \quad P_{ii} = r_i \geqslant 0, \quad p_i + q_i + r_i = 1, \quad i \in E.$$

Suppose that ρ is a translation-invariant metric on $\mathbf{Z}$ having the property: $u_k := \rho(k+1) - \rho(k) \downarrow$ as $0 \leqslant k \uparrow$, where $\rho(k) = \rho(k,0)$. Then, the transition probability of the ρ-OMC is given as follows (the last column denotes the probability of the corresponding jump): If $i_2 - i_1 = 1$, then

$$\begin{aligned}
(i_1, i_2) &\to (i_1 - 1,\, i_2 - 1) && (q_{i_2} - r_{i_1})^+ \wedge \big\{q_{i_1} - \big[(r_{i_2} - p_{i_1})^+ - (r_{i_1} - q_{i_2})^+\big]^+\big\} \\
&\to (i_1 + 1,\, i_2 + 1) && (p_{i_1} - r_{i_2})^+ \wedge \big\{p_{i_2} - \big[(r_{i_1} - q_{i_2})^+ - (r_{i_2} - p_{i_1})^+\big]^+\big\} \\
&\to (i_1 + 1,\, i_2) && p_{i_1} \wedge r_{i_2} \\
&\to (i_1 - 1,\, i_2) && \big[(r_{i_2} - p_{i_1})^+ - (r_{i_1} - q_{i_2})^+\big]^+ \\
&\to (i_1,\, i_2 + 1) && \big[(r_{i_1} - q_{i_2})^+ - (r_{i_2} - p_{i_1})^+\big]^+ \\
&\to (i_1,\, i_2 - 1) && q_{i_2} \wedge r_{i_1} \\
&\to (i_1 + 1,\, i_2 - 1) && \big\{-q_{i_1} + \big[(r_{i_2} - p_{i_1})^+ - (r_{i_1} - q_{i_2})^+\big]^+ + (q_{i_2} - r_{i_1})^+\big\}^+ \\
&\to (i_1 - 1,\, i_2 + 1) && \big\{q_{i_1} - \big[(r_{i_2} - p_{i_1})^+ - (r_{i_1} - q_{i_2})^+\big]^+ - (q_{i_2} - r_{i_1})^+\big\}^+ \\
&\to (i_1,\, i_2) && (r_{i_2} - p_{i_1})^+ \wedge (r_{i_1} - q_{i_2})^+.
\end{aligned}$$

If $i_2 - i_1 \geqslant 2$, then

$$\begin{aligned}
(i_1, i_2) &\to (i_1 - 1,\, i_2 - 1) && \big[(q_{i_2} - p_{i_1})^+ - r_{i_1}\big]^+ \\
&\to (i_1 + 1,\, i_2 + 1) && \big[(p_{i_1} - q_{i_2})^+ - r_{i_2}\big]^+ \\
&\to (i_1 + 1,\, i_2) && (p_{i_1} - q_{i_2})^+ \wedge r_{i_2} \\
&\to (i_1 - 1,\, i_2) && \big\{\big[r_{i_2} - (p_{i_1} - q_{i_2})^+\big]^+ - \big[r_{i_1} - (q_{i_2} - p_{i_1})^+\big]^+\big\}^+ \\
&\to (i_1,\, i_2 + 1) && \big\{\big[r_{i_1} - (q_{i_2} - p_{i_1})^+\big]^+ - \big[r_{i_2} - (p_{i_1} - q_{i_2})^+\big]^+\big\}^+ \\
&\to (i_1,\, i_2 - 1) && (q_{i_2} - p_{i_1})^+ \wedge r_{i_1} \\
&\to (i_1 + 1,\, i_2 - 1) && p_{i_1} \wedge q_{i_2} \\
&\to (i_1 - 1,\, i_2 + 1) && \big\{q_{i_1} - \big[(q_{i_2} - p_{i_1})^+ - r_{i_1}\big]^+\big\} \wedge \big\{p_{i_2} - \big[(p_{i_1} - q_{i_2})^+ - r_{i_2}\big]^+\big\} \\
&\to (i_1,\, i_2) && \big[r_{i_2} - (p_{i_1} - q_{i_2})^+\big]^+ \wedge \big[r_{i_1} - (q_{i_2} - p_{i_1})^+\big]^+.
\end{aligned}$$

By symmetry, we can write down the transition probability of the coupling for the other cases. Moreover, we have for $i_2 - i_1 = 1$,

$$\begin{aligned}
\overline{P}\rho(i_1, i_2) = {} & \big(1 - p_{i_1} \wedge r_{i_2} - q_{i_2} \wedge r_{i_1}\big) u_0 \\
& + \Big\{q_{i_1} + \big[(r_{i_1} - q_{i_2})^+ - (r_{i_2} - p_{i_1})^+\big]^+ - (q_{i_2} - r_{i_1})^+ \\
& + \big\{ - q_{i_1} + \big[(r_{i_2} - p_{i_1})^+ - (r_{i_1} - q_{i_2})^+\big]^+ + (q_{i_2} - r_{i_1})^+\big\}^+\Big\} u_1 \\
& + \big\{q_{i_1} - \big[(r_{i_2} - p_{i_1})^+ - (r_{i_1} - q_{i_2})^+\big]^+ - (q_{i_2} - r_{i_1})^+\big\}^+ u_2,
\end{aligned} \tag{4.1}$$

and for $i_2 - i_1 =: k \geqslant 2$,

$$\begin{aligned}
\overline{P}\rho(i_1, i_2) = {} & \rho(k) - (p_{i_1} \wedge q_{i_2}) u_{k-2} \\
& - \big[(p_{i_1} - q_{i_2})^+ \wedge r_{i_2} + (q_{i_2} - p_{i_1})^+ \wedge r_{i_1} + (p_{i_1} \wedge q_{i_2})\big] u_{k-1} \\
& + \big\{q_{i_1} - \big[(q_{i_2} - p_{i_1})^+ - r_{i_1}\big]^+\big\} \vee \big\{p_{i_2} - \big[(p_{i_1} - q_{i_2})^+ - r_{i_2}\big]^+\big\} u_k \\
& + \big\{q_{i_1} - \big[(q_{i_2} - p_{i_1})^+ - r_{i_1}\big]^+\big\} \wedge \big\{p_{i_2} - \big[(p_{i_1} - q_{i_2})^+ - r_{i_2}\big]^+\big\} u_{k+1}.
\end{aligned} \tag{4.2}$$

From (4.1) and (4.2), we see that when $r_i \equiv 0$, the OMC starting from $|i_1 - i_2| =$ odd will never meet each other even though the original chain can be positive recurrent. Actually, the same conclusion holds for any Markovian coupling. Thus, the term r_i is critical for success. The proof is omitted since it is lengthy but the technique is similar to that used in the last section.

5. Diffusion processes

Consider diffusion processes in $\mathbf{R}^d$ with operator

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

For simplicity, we write $L \sim (a(x), b(x))$. Given two diffusions with operators $L_k \sim (a_k(x), b_k(x))$, $k = 1, 2$ respectively, it is clear that the coefficients of any coupling operator should be of the form

$$a(x,y) = \begin{pmatrix} a_1(x) & c(x,y) \\ c(x,y)^* & a_2(y) \end{pmatrix}, \qquad b(x,y) = \begin{pmatrix} b_1(x) \\ b_2(y) \end{pmatrix}. \tag{5.1}$$

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

Example 5.1 (March coupling (Chen and Li [1989])). Let

$$a_k(x) = \sigma_k(x)\sigma_k(x)^*, \qquad k = 1, 2.$$

Take $c(x,y) = \sigma_1(x)\sigma_2(y)^*$.

Example 5.2 (Coupling by reflection). Let $L_1 = L_2$. Take

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

(Lindvall and Rogers (1986)) or

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

where $\bar{u} = (x-y)/|x-y|$.

We are now ready to study the OMC for diffusion processes. Given a metric $\rho \in C^2(\mathbf{R}^d \times \mathbf{R}^d \setminus \{(x,x) : x \in \mathbf{R}^d\})$, a coupling operator $\overline{L}$ is called ρ**−optimal** if

$$\overline{L}\rho(x,y) = \inf_{\widetilde{L}} \widetilde{L}\rho(x,y), \qquad x \neq y,$$

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

Theorem 5.3. Let $f \in C^2(\mathbf{R}_+;\mathbf{R}_+)$ with $f(0)=0$, $f'>0$ and $f''\leqslant 0$. Set $\rho(x,y)=f(|x-y|)$. Then, the ρ-optimal solution $c(x,y)$ is given as follows.

(1) If $d=1$, then $c(x,y)=-\sqrt{a_1(x)a_2(y)}$ and moreover,

$$\overline{L}f(|x-y|)=\frac{1}{2}\big(\sqrt{a_1(x)}+\sqrt{a_2(y)}\,\big)^2 f''(|x-y|)+\frac{(x-y)(b_1(x)-b_2(y))}{|x-y|}f'(|x-y|).$$

Next, suppose that $a_k(x)=\sigma_k(x)^2$, $k=1,\,2$ are non-degenarated and write

$$c(x,y)=\sigma_1(x)H^*(x,y)\sigma_2(y).$$

(2) If $f''(r)<0$ for all $r>0$, then $H(x,y)=U(\gamma)^{-1}\big[U(\gamma)U(\gamma)^*\big]^{1/2}$, where

$$\gamma=1-\frac{|x-y|f''(|x-y|)}{f'(|x-y|)}\quad\text{and}\quad U(\gamma)=\sigma_1(x)(I-\gamma\bar{u}\bar{u}^*)\sigma_2(y).$$

(3) If $f(r)=r$, then $H(x,y)$ is a solution to the equation:

$$U(1)H=\big(U(1)U(1)^*\big)^{1/2}.$$

(4) In particular, if $f(r)=r$ and $a_k(x)=\varphi_k(x)\sigma^2$ for some positive function φ_k $(k=1,\,2)$, where σ is independent of x and $\det\sigma>0$. Then $H(x,y)=I-2\sigma^{-1}\bar{u}\bar{u}^*\sigma^{-1}/|\sigma^{-1}\bar{u}|^2$. Moreover,

$$\overline{L}f(|x-y|)=\frac{1}{2|x-y|}\Big\{\big(\sqrt{\varphi_1(x)}-\sqrt{\varphi_2(y)}\,\big)^2\big[\operatorname{tr}\sigma^2-|\sigma\bar{u}|^2\big]+2\langle x-y,\,b_1(x)-b_2(y)\rangle\Big\}.$$

Finally, without the condition "$f(r)=r$", part (4) still holds provided the metric $\rho(x,y)=f(|x-y|)$ is replaced by $\rho(x,y)=f(|\sigma^{-1}(x-y)|)$. Moreover,

$$\begin{aligned}\overline{L}\rho(x,y)=\frac{1}{2}\big(\sqrt{\varphi_1(x)}+\sqrt{\varphi_2(y)}\,\big)^2 f''(|\sigma^{-1}(x-y)|)\\ +\Big\{(d-1)\big(\sqrt{\varphi_1(x)}-\sqrt{\varphi_2(y)}\,\big)^2+2\big\langle\sigma^{-1}(x-y),\,\sigma^{-1}(b_1(x)-b_2(y))\big\rangle\Big\}\\ \times\frac{f'(|\sigma^{-1}(x-y)|)}{2|\sigma^{-1}(x-y)|}.\end{aligned}$$

Proof. a) For any coupling operator $\widetilde{L}$ with coefficients given in (5.1), we have

$$2\widetilde{L}f(|x-y|)=\overline{A}(x,y)f''(|x-y|)+\frac{f'(|x-y|)}{|x-y|}\big[\operatorname{tr}A(x,y)-\overline{A}(x,y)+2\hat{B}(x,y)\big],\tag{5.2}$$

where

$$\begin{aligned}A(x,y)&=a_1(x)+a_2(y)-c(x,y)-c(x,y)^*, &\qquad B(x,y)&=b_1(x)-b_2(y)\\ \overline{A}(x,y)&=\langle x-y,\,A(x,y)(x-y)\rangle/|x-y|^2, &\qquad \hat{B}(x,y)&=\langle x-y,\,B(x,y)\rangle.\end{aligned}$$

Note that $\overline{A}(x,y) \geqslant 0$ because $a(x,y)$ is non-negative definite. Since there is nothing to do about the drifts, we assume that $B(x,y) \equiv 0$. Then, (5.2) is reduced to the following:

$$\begin{aligned} 2\widetilde{L}f(|x-y|) &= \overline{A}(x,y)f''(|x-y|) + \frac{f'(|x-y|)}{|x-y|}\big[\operatorname{tr} A(x,y) - \overline{A}(x,y)\big] \\ &= \frac{f'(|x-y|)}{|x-y|}\big[\operatorname{tr} A(x,y) - \gamma\overline{A}(x,y)\big]. \end{aligned} \tag{5.3}$$

We want to choose $c(x,y)$ so that the right-hand side of (5.3) becomes as small as possible.

b) When $d=1$, the assertion (1) follows by a simple computation. Note that in this case, for the ordinary metric $f(r)=r$ (i.e., $\gamma=1$), both the march coupling or the coupling by reflection attain the same minimum $\operatorname{tr} A - \overline{A} = 0$.

c) Let $d \geqslant 2$. Without any confusions, we write $H = H(x,y)$, $\sigma_1 = \sigma_1(x)$, $\sigma_2 = \sigma_2(y)$ and similarly for a_k's. We now prove that $a(x,y)$ is non-negative definite iff H is **contractive**: $|H\alpha| \leqslant |\alpha|$ for all $\alpha \in \mathbf{R}^d$. Actually, for $\alpha,\, \beta \in \mathbf{R}^d$, we have

$$\begin{aligned} (\alpha^*, \beta^*)a(x,y)\begin{pmatrix}\alpha\\ \beta\end{pmatrix} &= \alpha^* a_1 \alpha + \beta^* a_2 \beta + 2\langle H\sigma_1\alpha,\, \sigma_2\beta\rangle \\ &= |\sigma_1\alpha|^2 + |\sigma_2\beta|^2 + 2\langle H\sigma_1\alpha,\, \sigma_2\beta\rangle. \end{aligned}$$

Thus, $a(x,y)$ is non-negative definite iff

$$|\alpha|^2 + |\beta|^2 + 2\langle H\alpha, \beta\rangle \geqslant 0, \quad \alpha,\, \beta \in \mathbf{R}^d.$$

Setting $\beta = -H\alpha$, it follows that $|H\alpha| \leqslant |\alpha|$. This proves the necessity. The sufficiency is easy.

d) Because

$$A(x,y) = a_1 + a_2 - \sigma_1 H^* \sigma_2 - \sigma_2 H \sigma_1, \qquad \operatorname{tr} A(x,y) = \operatorname{tr}(a_1+a_2) - 2\operatorname{tr}(\sigma_1 H^* \sigma_2)$$

and

$$\overline{A}(x,y) = \bar{u}^*(a_1+a_2)\bar{u} - 2\bar{u}^*\sigma_1 H^* \sigma_2 \bar{u}.$$

We have

$$\operatorname{tr} A(x,y) - \gamma\overline{A}(x,y) = \operatorname{tr}(a_1+a_2) - \gamma\bar{u}^*(a_1+a_2)\bar{u} + 2\big[\gamma\bar{u}^*\sigma_1 H^*\sigma_2\bar{u} - \operatorname{tr}(\sigma_1 H^*\sigma_2)\big].$$

We need only to minimize

$$F(H) := \gamma\bar{u}^*\sigma_1 H^*\sigma_2\bar{u} - \operatorname{tr}(\sigma_1 H^*\sigma_2)$$

under the restriction of H being contractive. Since $F(H)$ is linear and the set $\{H : |H\alpha| \leqslant |\alpha|\}$ is compact, it should be clear that the optimal solution always exists. We claim that the optimum can be only attained by an orthogonal matrix

H. An easier way to see this goes as follows. Actually, we are dealing with the convex programming problem:

$$\text{Minimize} \quad F(H), \qquad \text{Subject to} \ \ |H\alpha|^2 \leqslant 1 \quad \text{and} \quad |\alpha|^2 = 1.$$

As usual, by introducing a slack variable, we reduce the inequality constrain to the equality one. Then, in view of the necessary condition for the minimum, it follows that the constrain should be active (i.e., $|H\alpha| = 1$) unless $\gamma\sigma_2\bar{u}\bar{u}^*\sigma_1 = \sigma_2\sigma_1$, which is impossible except $d = 1$ (and $\gamma = 1$). This proves the required conclusion.

Next, consider the problem:

$$\text{Minimize} \ \ F(H), \quad \text{Subject to} \ \ H^*H = I.$$

Define

$$L = \sum_{i,j} h_{ij}\big[\gamma(\sigma_2\bar{u})_i(\sigma_1\bar{u})_j - (\sigma_2\sigma_1)_{ij}\big] + \sum_{i,j}\lambda_{ij}\bigg[\sum_k h_{ki}h_{kj} - \delta_{ij}\bigg],$$

where $(h_{ij}) := H$ and $\Lambda := (\lambda_{ij})$ which is a symmetric matrix. Solving

$$\frac{\partial L}{\partial H} = \gamma\sigma_2\bar{u}\bar{u}^*\sigma_1 - \sigma_2\sigma_1 + 2H\Lambda = 0$$

we obtain

$$2H\Lambda = \sigma_2(I - \gamma\bar{u}\bar{u}^*)\sigma_1 =: U(\gamma)^*. \tag{5.4}$$

Since $H^*H = I$ and Λ is symmetric, we have $4\Lambda^2 = U(\gamma)U(\gamma)^*$ and so $2\Lambda = \big[U(\gamma)U(\gamma)^*\big]^{1/2}$ (The negative solution can be removed since we are interested in minimum). Substituting this into (5.4), we get

$$U(\gamma)H = \big[U(\gamma)U(\gamma)^*\big]^{1/2}. \tag{5.5}$$

Now, this proves not only assertion (3) but also (2) since $\det U(\gamma) \neq 0$ whenever $\gamma > 1$.

e) To prove the last assertion of the theorem, we need only to consider $a_k(x) = \varphi_k(x)I$, $k = 1, 2$. The general case can be reduced to this by replacing the metric $|x|$ with $|\sigma^{-1}x|$ (From the geometric point of view, the ordinary Riemannian metric I is replaced by σ^{-2}). Because $\operatorname{rank}(\bar{u}\bar{u}^*) = 1$ and $\operatorname{tr}(\bar{u}\bar{u}^*) = 1$, without loss of generality, we can choose an orthogonal matrix O so that

$$\bar{u}\bar{u}^* = O\operatorname{diag}[1, 0, \cdots, 0]O^*,$$

where $\operatorname{diag}[\cdots]$ denotes the diagonal matrix with diagonal elements $[\cdots]$. Then, (5.5) becomes

$$O\operatorname{diag}[1-\gamma, 1, \cdots, 1]O^*H = O\operatorname{diag}[\gamma - 1, 1, \cdots, 1]O^*.$$

Thus, if $\gamma > 1$, then

$$H = O\text{diag}\,[-1, 1, \cdots, 1]O^* = I - 2\bar{u}\bar{u}^*.$$

On the other hand, if $\gamma = 1$, ruling out the useless solution $H = I$, we can assume that $H = I + B$ with $B \neq 0$. Then (5.5) is reduced to $(I - \bar{u}\bar{u}^*)B = 0$. Because $\text{rank}(I - \bar{u}\bar{u}^*) = d - 1$, this equation has only solution $B = \bar{u}v^*$ for some $v \in \mathbf{R}^* \setminus \{0\}$. Now, the orthogonality of H gives us $v = -2\bar{u}$ and so the assertion follows.

f) Finally, consider part (4). In this case, (5.5) is reduced to

$$(I - \bar{u}\bar{u}^*)\sigma H\sigma^{-1} = I - \bar{u}\bar{u}^*.$$

Noticing that $\sigma H\sigma^{-1} = I$ iff $H = I$, the proof is similar to the last paragraph replacing H with $\sigma H\sigma^{-1}$. $\square$

We remark that the conclusion of part (4) does not hold when $\gamma > 1$. The comparison of the last two assertions of Theorem 5.3 leads us to use the Riemannian metric $a(x)^{-1}$ instead of the ordinary one. For simplicity, here we write down the one-dimensional case only. The proof is similar and even simpler. Note that the distance $d(x, y)$ given below is no longer translation-invariant except $a(x) \equiv$ constant.

Corollary 5.4. Let $d = 1$, $L_1 = L_2$ and $a(x) > 0$. Define

$$d(x, y) = \int_x^y a(z)^{-1/2} dz, \qquad x \leqslant y.$$

Given f as above. Set $\rho = f \circ d$. Then, the ρ-OC is $c(x, y) = -\sqrt{a(x)a(y)}$. Moreover,

$$\overline{L}\rho(x, y) = 2(f'' \circ d)(x, y) + \left[\frac{4b(y) - a'(y)}{4\sqrt{a(y)}} - \frac{4b(x) - a'(x)}{4\sqrt{a(x)}}\right](f' \circ d)(x, y), \qquad x \leqslant y.$$

6. Spectral Gap for Markov Chains or Laplacian on Manifold

For a reversible Markov process with generator Ω, except the trivial eigenvalue $\lambda_0 = 0$, the next eigenvalue λ_1 of $-\Omega$ is called the **spectral gap** of Ω, denoted by $\text{gap}(\Omega)$. The importance of the spectral gap is that it describes the exponential L^2-convergence:

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

where π is the reversible measure of the process and $\pi f = \int \pi(dx) f(x)$. Actually, it can be proved that $\varepsilon_{\max} = \text{gap}(\Omega)$. (cf., Liggett (1989) and Chen (1991) or [5; Section 9.1]). In this section, we show how to use couplings to obtain some lower bounds of $\text{gap}(\Omega)$. We consider an example from Markov chain and discuss a property related to the algebraic L^2-convergence. We also study the first eigenvalue of Laplacian on manifold, which is a well-known problem in geometry. Other applications of OMCs will be presented in subsequent papers.

Recall that the **coupling time** is defined by

$$T = \inf\{t \geqslant 0 :\ X_t^1 = X_t^2\}.$$

We will use the following two general results.

Theorem 6.1. Let $\{X_t\}_{t\geqslant 0}$ be a reversible Markov process with weak generator Ω on ${}_b\mathscr{E}$. Denote by f the eigenfunction corresponding to λ_1. Set $g(x,y)=f(x)-f(y)$. Suppose that there is a Markovian coupling $\mathbb{P}^{x_1,x_2}$ of the process with operator $\widetilde{\Omega}$ so that

$$g(X_t^1,X_t^2)-\int_0^t \widetilde{\Omega} g(X_s^1,X_s^2)ds$$

is a martingale up to time T under $\mathbb{P}^{x_1,x_2}$ with respect to the natural flow of σ-algebras. If $\sup_{x\neq y}|f(x)-f(y)|<\infty$, then

$$\mathrm{gap}(\Omega)\geqslant 1\Big/\max_{x_1\neq x_2}\mathbb{E}^{x_1,x_2}T.$$

Theorem 6.2. Let (E,ρ) be a metric space and let $\{X_t\}$, Ω and f be the same as in the previous theorem. Suppose that

(1) $\mathbb{E}^x f(X_t)-f(x)=\int_0^t \mathbb{E}^x\Omega f(X_s)ds$.
(2) There is a coupling $\mathbb{P}^{x_1,x_2}$ such that

$$\mathbb{E}^{x_1,x_2}\gamma(X_t^1,X_t^2)\leqslant \gamma(x_1,x_2)\exp[-\alpha t],\qquad t\geqslant 0,\ x_1,\ x_2\in E$$

for some $\alpha>0$, where $\gamma: E\times E\to[0,\infty)$ with $\gamma(x,y)=0$ iff $x=y$.
(3) $\sup_{y\neq x}|f(y)-f(x)|/\gamma(y,x)<\infty$.

Then, we have $\mathrm{gap}(\Omega)\geqslant\alpha$.

Theorem 6.1 is implicated in [8; Proof of Theorem 1.4]. Theorem 6.2 is an improvement to [8; Theorem 1.9], in which γ is required to be an equivalent metric of ρ. The most interesting case is the following: f is Lipschitz with respect to ρ and $\gamma=\bar{\gamma}\circ\rho$ for some $\bar{\gamma}\in C(\mathbf{R}_+)$ with $\bar{\gamma}(r)=0$ iff $r=0$ and $\inf_{r>0}\bar{\gamma}(r)/r>0$. The proofs are similar to [8] and so are omitted here.

Having these preparations in mind, it is not difficult to present some general results for the spectral gap of Markov chains. But to save the space, we discuss here a simple example only to illustrate the power of the approach. Some key points for the general situation will be indicated below. First, by using a localization procedure (cf. [4]), the non-compact case can be reduced to the compact one.

Example 6.3. Consider the birth-death process with finite space

$$E=\{0,1,\cdots,N+1\}$$

and rates $b_i=a_{i+1}=1$, $0\leqslant i\leqslant N$ and $a_0=b_{N+1}=0$. We adopt the coupling by reflection with a natural modification on the boundary.

a) Let $N\geqslant 2$ and solve the equation:

$$\varphi_0=0,\qquad \varphi_3=3\varphi_1-1,\qquad \varphi_{k+2}=2\varphi_k-\varphi_{k-2}-1,\quad 2\leqslant k\leqslant N-1;$$
$$\varphi_{N+1}=2\varphi_N-\varphi_{N-2}-1=\varphi_{N-1}+1.$$

We obtain

$$\varphi_1 = \frac{(-1)^N + 7 + 8N + 2N^2}{8(2+N)}, \qquad \varphi_2 = \frac{2 + 3N + N^2}{2(2+N)}, \qquad \varphi_3 = 3\varphi_1 - 1;$$

$$\varphi_k = c(1) + c(2)k + \big[-c(1) + c(3)k \big](-1)^k - \frac{1}{8}k^2, \qquad 4 \leqslant k \leqslant N-1;$$

$$c(1) = \frac{1}{16}, \qquad c(2) = \frac{1}{8} + \frac{1}{2}\varphi_1 + \frac{1}{4}\varphi_2, \qquad c(3) = \frac{1}{8} - \frac{1}{2}\varphi_1 + \frac{1}{4}\varphi_2;$$

$$\varphi_{N+1} = \frac{1}{16}\big[7 + (-1)^N + 8N + 2\,N^2\big].$$

Since

$$\overline{\Omega}_r \varphi(|i_1 - i_2|) + 1 \leqslant 0,$$

which is what we need to claim that

$$\mathbb{E}^{i_1,i_2} T \leqslant \varphi(|i_1 - i_2|).$$

Now, as an application of Theorem 6.1, we get $\lambda_1 \geqslant \varphi_{N+1}^{-1}$. Comparing this estimate with the exact value

$$\lambda_1 = 4\sin^2(\pi/(2N+4)),$$

we have $\lambda_1 \varphi_{N+1} \approx \pi^2/8$, as $N \to \infty$.

b) Take

$$F(k) = \sin\frac{k\pi}{2N+4} \Big/ \sin\frac{\pi}{2N+4}$$

and define

$$\alpha(1) = 3 - F(k), \qquad \alpha(k) = 2 - \frac{F(k-2) + F(k+2)}{F(k)}, \quad 2 \leqslant k \leqslant N-1$$

$$\alpha(N) = 2 - \frac{F(N-2) + F(N+1)}{F(N)}, \qquad \alpha(N+1) = 1 - \frac{F(N-1)}{F(N+1)}.$$

Then, by some elementary computations, we obtain

$$\alpha(N) \geqslant \alpha(1) = \cdots = \alpha(N-1) = \alpha(N+1) = 4\sin^2\frac{\pi}{2(N+2)} = \lambda_1.$$

In general, Theorem 6.2 gives us $\mathrm{gap}(\Omega) \geqslant \alpha$ whenever

$$\overline{\Omega}_r F(|i_1 - i_2|) \leqslant -\alpha F(|i_1 - i_2|).$$

For this example, the inequality holds with $\alpha = \lambda_1$ and so our estimate is exact!

Next, if the process is not L^2-exponentially convergent, it is natural to ask for a slower convergence:

$$\|P(t)f - \pi f\|^2 \leqslant CV(f)/t^{q-1}, \qquad t > 0,\ f \in L^2(\pi),$$

where C and $q > 1$ are constants and $V : L^2(\pi) \to [0, \infty]$. Such convergence is called **algebraic L^2-convergence**. It turns out for such convergence, $V(f)$ should satisfy

$$V(cf + d) = c^2 V(f) \quad \text{and} \quad V(P(t)f) \leqslant V(f) \tag{6.1}$$

for all constants c and d, $t > 0$ and $f \in L^2(\pi)$ (cf., Liggett (1991)). We now show that the functional V can be obtained by using couplings.

Corollary 6.4. Let (u_k) be a positive sequence and suppose that $(b_k u_k - a_k u_{k-1})$ $(k \geqslant 0)$ is non-increasing, where $u_{-1} = 1$. Then, for birth-death process, the functional V_1:

$$V_1(f)^{1/2} = \sup_{m \neq n} \frac{|f(m) - f(n)|}{\rho(m,n)} = \sup_{n \geqslant 0} \frac{|f(n+1) - f(n)|}{u_n},$$
$$\rho(m,n) = \Big| \sum_{j<m} u_j - \sum_{j<n} u_j \Big|$$

satisfies (6.1).

It was pointed out in Liggett (1991, p.948) that Corollary 6.4 can be proved by using a rather complicated approach. But this corollary is actually an immediate consequence of Theorem 3.3. Furthermore, due to OMC, Theorem 3.2 provides us a much refined choice of V.

Corollary 6.5. Given a positive non-increasing sequence (u_j), set $F(k) = \sum_{j<k} u_j$. If $\overline{\Omega}_r F(k) \leqslant 0$, Then, for birth-death process, the functional V_2:

$$V_2(f)^{1/2} = \sup_{m \neq n} |f(m) - f(n)| / F(|m-n|)$$

satisfies (6.1).

Finally, we study the lower bound of the first eigenvalue on manifold. Let (M, g) be a d-dimensional compact Riemannian manifold with distance $\rho = \rho_M$ deduced by the metric g and assume that $\mathrm{Ric}_M \geqslant -Kg$ for some $K \in \mathbf{R}$. Denote by Δ, λ_1 and D the Laplace-Beltrami operator on M, the first eigenvalue and the diameter of M. Here we consider the hardest case that $K \geqslant 0$.

Theorem 6.6. Let $K \geqslant 0$. Suppose that for some $\alpha > 0$ the differential inequality

$$4\gamma'' + 2\sqrt{K(d-1)}\,\tanh\left(\frac{r}{2}\sqrt{\frac{K}{d-1}}\right)\gamma' + \alpha\,\gamma \leqslant 0, \qquad r \in [0, D] \tag{6.2}$$

has a solution γ having the property $\gamma' \geqslant 0$ on $[0, D]$ and $\inf_{r \in (0,D]} \gamma(r)/r > 0$. Then

$$\lambda_1 \geqslant \alpha.$$

In particular, we have

$$\lambda_1 \geqslant \frac{1}{4} K(d-1) \tanh^2\left(\frac{D}{2}\sqrt{\frac{K}{d-1}}\right) \mathrm{sech}^2\theta,$$

where θ is the (decreasing) limit of θ_n:

$$\theta_1 = \frac{D}{4}\sqrt{K(d-1)}\,\tanh\left(\frac{D}{2}\sqrt{\frac{K}{d-1}}\right), \qquad \theta_n = \theta_1 \tanh\theta_{n-1}, \quad n \geqslant 2.$$

Proof. a) By using the coupling by reflection of Brownian motion on manifold, constructed by Kendall (1986) (See also Cranston (1991)), it was proved in [8] that

$$d\rho(X_t^1, X_t^2) \leqslant 2\sqrt{2}dB_t + 2\sqrt{K(d-1)}\,\tanh\left(\frac{\rho(X_t^1, X_t^2)}{2}\sqrt{\frac{K}{d-1}}\right)dt, \quad t < T, \tag{6.3}$$

where $\{B_t\}$ is the Brownian motion in $\mathbf{R}$. Here we have used the fact that $K \geqslant 0$.

b) By Itô formula, (6.2) and (6.3), there exists a martingale M_t such that

$$\begin{aligned} d(\gamma\circ\rho)(X_t^1, X_t^2) &\leqslant dM_t + \frac{1}{2}\cdot 8\cdot(\gamma''\circ\rho)(X_t^1, X_t^2)dt \\ &\quad + 2\sqrt{K(d-1)}\,\tanh\left(\frac{\rho(X_t^1, X_t^2)}{2}\sqrt{\frac{K}{d-1}}\right)(\gamma'\circ\rho)(X_t^1, X_t^2)dt \\ &\leqslant dM_t - \alpha(\gamma\circ\rho)(X_t^1, X_t^2)dt. \end{aligned}$$

Hence,

$$\mathbb{E}^{x_1,x_2}(\gamma\circ\rho)(X_t^1, X_t^2) \leqslant (\gamma\circ\rho)(x_1, x_2)\exp[-\alpha t].$$

By Theorem 6.2, we obtain $\lambda_1 \geqslant \alpha$.

c) Next, assume that $K(d-1) > 0$ and take

$$\gamma(\rho) = \exp[-c\rho/8]\sinh(c\delta\rho/8), \qquad 0 \leqslant \rho \leqslant D,$$

where

$$\begin{aligned} c &= 2\sqrt{K(d-1)}\,\tanh\left(\frac{D}{2}\sqrt{\frac{K}{d-1}}\right), \\ \delta &= \sqrt{1-\frac{16\alpha}{c^2}}, \\ \alpha &= \frac{K(d-1)}{4}\tanh^2\left(\frac{D}{2}\sqrt{\frac{K}{d-1}}\right)\mathrm{sech}^2\theta_n, \end{aligned}$$

and $n \geqslant 1$ is fixed. Then, it is easy to check that γ is a solution to (6.2). Moreover, $\gamma' \geqslant 0$ on $[0, D]$[1] and $\inf_{\rho\in(0,D]}\gamma(\rho)/\rho > 0$. By b), we have

$$\lambda_1 \geqslant \alpha = \frac{K(d-1)}{4}\,\tanh^2\left(\frac{D}{2}\sqrt{\frac{K}{d-1}}\right)\mathrm{sech}^2\theta_n. \quad \square$$

Acknowledgement. The first two versions of the paper, with much details, were preprinted as Technical Report No.216, 1993 at Carleton Univ. and No. 147, 1993 at Univ. of Rome II. The author acknowledges Prof. L. Accardi and Prof. D. A. Dawson for their hospitality and valuable discussions.

[1]Note added in proof. The idea of Theorem 6.6 is regarding the coefficient of γ' as a constant (i.e., replacing the variable r in the coefficient by constant D), then equation (6.2) is solvable. This not only gives us the function γ used in the proof, but also indicates that Theorem 6.6 is mainly designed for large D. The proof of $\gamma' > 0$ on $(0, D)$ leads the solution of θ. To see this, note that $\gamma' > 0$ iff $\delta > \tanh(c\delta x/8)$. Thus, it suffices that $\delta \geqslant \tanh(c\delta D/8)$. Equivalently, $\theta_1\tanh\theta_n \geqslant \theta_1\tanh(\theta_1\tanh\theta_n)$, or $\theta_n \geqslant \theta_1\tanh\theta_n$. The last assertion holds by definition of θ_n.

References

1. Chen, M. F. (1986a), *Couplings of jump processes*, Acta Math. Sinica, New Series, 2:2, 123-136.
2. Chen, M. F. (1986b), *Jump Processes and Interacting Particle Systems* (*In Chinese*), Beijing Normal Univ. Press.
3. Chen, M. F. (1990), *Ergodic theorems for reaction-diffusion processes*, J. Statis. Phys. 58:5/6, 939-966.
4. 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.
5. Chen, M. F. (1992), *From Markov Chains to Non-Equilibrium Particle Systems*, World Scientific.
6. Chen, M. F. (1993), *Optimal couplings and application to Riemannian geometry*, to appear in Prob. Theory and Math. Stat., Vol.1, Edited by B. Grigelionis et al. 1994 VPS/TEV.
7. Chen, M. F. and Li, S. F. (1989), *Coupling methods for milti-dimensional diffusion processes*, Ann. of Probab. 17:1, 151–177.
8. Chen, M. F. and Wang, F. Y. (1992), *Application of coupling method to the first eigenvalue on manifold*, Sci. Sin.(A), 23:11(1993)(Chinese Edition), 37:1(1994)(English Edition).
9. Cranston, M. (1991), *Gradient estimates on manifolds using coupling*, J. Funct. Anal. 99, 110–124.
10. Dobrushin, R. L. (1970), *Prescribing a system of random variables by conditional distributions*, Theory Prob. Appl., 15, 458-486.
11. Griffeath, D. (1975), *A maximal coupling for Markov chains*, Z. Wahrs. 31, 95–106.
12. Kendall, W. S. (1986), *Nonnegative Ricci curvature and the Brownian coupling property*, Stochastics 19, 111–129.
13. Liggett, T. M. (1985), *Interacting Particle Systems*, Springer-Verlag.
14. Liggett, T. M. (1989), *Exponential L_2 convergence of attractive reversible nearest particle systems*, Ann. Probab., 17, 403-432.
15. Liggett, T. M. (1991), *L_2 rates of convergence for attractive reversible nearest particle systems*, Ann. Probab., 19:3, 935-959.
16. Lindvall, T. (1992), *Lectures on the Coupling Method*, Wiley, New York.
17. Lindvall, T. and Rogers, L. C. G. (1986), *Coupling of multidimensional diffusion processes*, Ann. of Probab. 14:3, 860–872.

Department of Mathematics, Beijing Normal University, Beijing 100875, The People's Republic of China.

Prob. Theory and Math. Stat., Vol. 1, pp. 121–142
B. Grigelionis et al. (Eds)
1994 VPS/TEV

OPTIMAL COUPLINGS AND APPLICATION TO RIEMANNIAN GEOMETRY

Mu-Fa Chen

(Beijing Normal University)

Abstract. The talk begins with some backgrounds of our study: The spectral gap for four classes of reversible Markov processes and the relation between the spectral gap and the phase transitions. Then, we introduce two aspects of our recent progress: 1) The estimates of the spectral gap (or the first non-trivial eigenvalue) of Laplacian on compact Riemannian manifold. 2) Optimal Markovian couplings. These explain the precise meaning of the vague title. The resulting estimates are quite unexpected, not only recover the known sharp estimates but also produce some new ones without using anything from the previous proofs. The optimal estimates come from the optimal couplings, which are often out of our probabilistic intuition. It seems to the author that the study of couplings is renewed but there is still a lot to be done. We emphasize the ideas, including the applications of the coupling technique, in terms of some simple examples. It is hoped that the materials presented here could be helpful not only for experts but also for newcomers.

Part I. Backgrounds. Spectral Gap and Phase Transitions

1. Markov Chains.

Let us explain what spectral gap is by using a simple example. Consider a birth-death process with finite state space $E = \{0, 1, \cdots, N+1\}$ and Q-matrix

$$Q = (q_{ij}) = \begin{pmatrix} -b_0 & b_0 & 0 & \dots & 0 \\ a_1 & -(a_1+b_1) & b_1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & a_N & -(a_N+b_N) & b_N \\ 0 & \dots & 0 & a_{N+1} & -a_{N+1} \end{pmatrix}.$$

2000 *Mathematics Subject Classification*. 60J25, 60K35, 58G32, 58G25.

Key words and phrases. Optimal Markovian coupling, spectral gap, phase transitions, Riemannian geometry.

Research supported in part by the National Natural Science Foundation and the State Education Commission of China.

Clearly, there is ono-to-one correspondence between the Q-matrix and the generator Ω:

$$\Omega f(i) = \sum_j q_{ij}(f_j - f_i).$$

As a generator of a Markov process, we always have $\Omega 1 = 0 = 0 \cdot 1$. This means that the Q-matrix has an eigenvalue 0 with eigenvector 1. Actually, since the state space is compact, the eigenvalues of $(-\Omega)$ are discrete:

$$0 = \lambda_0 < \lambda_1 \leqslant \cdots \leqslant \lambda_{N+1}.$$

Hence, there is a gap between λ_0 and λ_1:

$$\mathrm{gap}(\Omega) := \lambda_1 - \lambda_0 = \lambda_1.$$

Note that in general it is impossible to obtain the precise value of λ_1, so our main interest is to estimate $\mathrm{gap}(\Omega)$. Even in such a simple situation, the problem is still not so easy as it looks like. As far as I know, there are at least eight different approaches to estimate $\mathrm{gap}(\Omega)$. Two of them are presented in Lawler and Sokal (1988), Diaconis and Stroock (1991) respectively. From the titles of these two papers we see that some idea from geometry was used in the study of spectral gap. However, what I am going to talk in the next part is actually in the opposite direction: we adopt a probabilistic approach to obtain some new estimates to geometry.

Certainly, the problem is meaningful for other types of Markov processes. For instance, we can consider

2. Diffusion processes in $\mathbf{R}^d$.

A classical example is the Ornstein-Uhlenbeck process, for which, we have

$$L = \frac{1}{2}\Delta - x \cdot \nabla.$$

The spectrum of L is completely understood: when $d = 1$, $\lambda_n = n$, $n \geqslant 0$ and the corresponding eigenfunctions are:

$$(-1)^n e^{x^2} \frac{d^n}{dx^n}\big(e^{-x^2}\big).$$

Hence $\lambda_1 = 1$ when $d = 1$. Because the components are independent, we actually have $\lambda_1 = 1$ for all $d \geqslant 1$.

Note that in general the spectrum of the generator L of a diffusion process may be continuous since the state space $\mathbf{R}^d$ is non-compact. In that case, $\mathrm{gap}(L) = 0$. We can also study

3. Diffusion processes on Manifold.

The typical case is the Brownian motion on manifold, for which, we have the Laplace-Beltrami operator Δ. The question is $\mathrm{gap}(\Delta) = ?$ I will return to discuss this topic more carefully. Of course, we can also consider the infinite-dimensional case.

4. Interacting Particle Systems.

The state space is $X^{\mathbf{Z}^d}$, where the spin space X can be $\{0,1\}$, $\mathbf{Z}_+$, $\mathbf{R}^d$ or a manifold M. In the study of particle systems, the generator, denoted by Ω_β, often depends on a parameter β — the inverse temperature. In this context, it can be happened that $\mathrm{gap}(\Omega_\beta) = 0$ even for compact state space.

Why the study of spectral gap is important? One reason is as follows: Actually, we are dealing with a reversible Markov process. So we have a reversible probability measure π. Then, we have the real L^2-space $L^2(\pi)$ and a generator Ω of the process $P(t)$. Hence, we can study the L^2-exponential convergence

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

where $\pi f = \int f d\pi$. Now, the relation between the exponential rate ε and the spectral gap can be stated as follows:

Theorem 1.1 (Liggett (1989) and Chen (1991b)). The maximal exponential rate $\varepsilon_{\max} = \mathrm{gap}(\Omega) = \mathrm{gap}(D)$, where

$$\mathrm{gap}(\Omega) = \inf\big\{ -(\Omega f, f) :\ f \in \mathscr{D}(\Omega),\ \pi f = 0,\ \|f\| = 1\big\},$$
$$\mathrm{gap}(D) = \inf\big\{D(f,f) :\ f \in \mathscr{D}(D),\ \pi f = 0,\ \|f\| = 1\big\}$$

and $D(f,f)$ is the Dirichlet form of the process.

Thus, the spectral gap describes the L^2-exponentially convergent rate. Usually, if a system has no phase transitions, then it is L^2-exponentially ergodic. That is, $\mathrm{gap}(\Omega_\beta) > 0$ for all β below the critical temperature β_c. But at the critical temperature, $\mathrm{gap}(\Omega_{\beta_c}) = 0$. Thus, the study of the spectral gap provides a way to describe the phase transitions. Since the study of phase transitions is a hard subject, the available mathematical tools are still quite limited, people think that the study of spectral gap should be helpful since we have the L^2-theory in mind. In the past ten years or more, there are a lot progress on this topic. Actually, my own interest in the study of spectral gap started from this object. Much of the previous results is collected in the book Chen (1992). Refer to Holley and Stroock (1989), Stroock and Zegarlinski (1992a, b) for further information.

One reason I introduced these four classes of Markov processes is that we have studied the spectral gap for all of them by using mainly the same coupling technique. See Chen (1993a), Chen and Wang (1992, 1993b) and Wang (1992b, 1993a, b) for details.

To see the study of spectral gap is a nice topic in mathematics, we now consider the Laplacian on manifold. It turns out that the study on estimating the spectral gap is a well-known subject in the modern geometry, called

Part II. The First Eigenvalues of Laplacian on Manifold

Let (M, g) be a d-dimensional Riemannian manifold with Riemannian metric g. Denote by ρ and D the distance induced by g and the diameter of M respectively. Assume that the Ricci curvature is bounded below: i.e., $\mathrm{Ric}_M \geqslant -Kg$ for some $K \in \mathbf{R}$. As I mentioned above, we are interested in the estimate of λ_1. The study of this topic goes back to the famous paper by M. Kac entitled "Can one hear the shape of a drum?" The idea is to use the geometric quantities d, D and K to estimate the bound of λ_k's. A large part of the books Chavel (1984) and Schoen and Yau (1988) is devoted to this problem. See also Kröger (1992) for more recent information. Here, we only mention some famous results obtained by geometers.

Case 1: $K \leqslant 0$**.**

In 1958, Lichnerowicz proved the following estimate:

$$\lambda_1 \geqslant -\frac{d}{d-1}\,K, \qquad K < 0, \tag{2.1}$$

which is optimal when $M = S^d\,(d \geqslant 2)$. After 22 years, an important progress was made by Li and Yau (1980) who proved that

$$\lambda_1 \geqslant \frac{\pi^2}{2\,D^2}, \qquad K \leqslant 0.$$

It was then improved by Zhong and Yang (1984) as follows:

$$\lambda_1 \geqslant \frac{\pi^2}{D^2}, \qquad K \leqslant 0, \tag{2.2}$$

which is optimal when $M = S^1$. This is a deep result, included in Schoen and Yau's book. Refer to Wu (1993) for further comments. Actually, (2.2) is the one of his two main results, for which Zhong became the first mathematician who awarded the S. S. Chern's prize.

Case 2: $K \geqslant 0$**.**

In their paper quoted above, Li and Yau obtained the following estimate:

$$\lambda_1 \geqslant \frac{1}{D^2(d-1)\exp\left[1+\sqrt{1+4D^2K(d-1)}\right]}, \qquad K \geqslant 0.$$

Comparing this with the above estimates, we see that the difficulty of the problem increases as K increases. Next, Cai (1991) proved that

$$\lambda_1 \geqslant \frac{\pi^2}{D^2} - K, \qquad K \geqslant 0. \tag{2.3}$$

On the other hand, Yang (1989) and Jia (1991) proved that

$$\begin{aligned} \lambda_1 &\geqslant \frac{\pi^2}{D^2} e^{-\alpha/2}, \qquad \text{if } \ d \geqslant 5 \\ &> \frac{\pi^2}{2\,D^2} e^{-\alpha'/2}, \quad \text{if } \ 2 \leqslant d \leqslant 4, \qquad K \geqslant 0, \end{aligned} \tag{2.4}$$

where $\alpha = D\sqrt{K(d-1)}$ and $\alpha' = D\sqrt{K((d-1)\vee 2)}$. The first estimate in (2.4) for all $d \geqslant 2$ is called Yau's conjecture (mentioned in Yang (1989)).

From the above summary, one sees that the picture is quite complete. For such a well developed subject, can we still do something? The answer is as follows:

Theorem 2.1 (Chen and Wang (1992)).

$$\begin{aligned} \lambda_1 &\geqslant \max\left\{\frac{\pi^2}{D^2},\ -\frac{d}{d-1}K,\ \frac{8}{D^2} - \frac{K}{3}\right\}, \qquad\qquad \text{if } \ K \leqslant 0 \\ &\geqslant \max\left\{\frac{\pi^2}{D^2} - K,\ \frac{8}{D^2} - \frac{K}{3},\ \frac{8}{D^2}\exp\left[-\frac{D^2K}{8}\right],\ \frac{8}{D^2}\left(1+\frac{\alpha}{3}\right)e^{-\alpha/2},\right. \\ &\qquad \left.\frac{1}{4}K(d-1)\tanh^2\left(\frac{D}{2}\sqrt{\frac{K}{d-1}}\right)\operatorname{sech}^2\theta\right\}, \quad \text{if } \ K \geqslant 0, \end{aligned}$$

where θ is obtained in the following way: Let $\theta_1 = \frac{D}{4}\sqrt{K(d-1)}\,\tanh\left(\frac{D}{2}\sqrt{\frac{K}{d-1}}\right)$, $\theta_n = \theta_1 \tanh\theta_{n-1}$, $n \geqslant 2$. Then $\theta_n \downarrow \theta$.

Clearly, the estimates (2.1)–(2.3) are covered by our theorem. Next, it is easy to check that

$$\max\left\{\frac{\pi^2}{D^2} - K,\ \frac{8}{D^2}\left(1+\frac{\alpha}{3}\right)e^{-\alpha/2}\right\} \geqslant \frac{\pi^2}{D^2}e^{-\alpha/2}.$$

Hence,

$$\lambda_1 \geqslant \frac{\pi^2}{D^2}e^{-\alpha/2}$$

holds for all $d \geqslant 2$. This is just the Yau's conjecture, it certainly covers (2.4). Thus, we have not only achieved the optimum but also provided some new estimates. Moreover, it is believed that the last estimate of Theorem 2.1, taken from Chen (1993a), is sharp when K goes to infinity.

No doubt, the theorem is deep in geometry. How about its proof? Our probabilistic proof is surprisingly straightforward, without using anything from the previous proofs. To confirm this, I would like to show quickly the main steps of the proof.

Sketch of the Proof.

Step 1. Let f be the eigenfunction corresponding to λ_1. Then, by the forward Kolmogorov equation (or by the martingale formulation), we often have

$$f(x) = \mathbb{E}^x f(X_t) - \mathbb{E}^x \int_0^t \Omega f(X_s)ds = \mathbb{E}^x f(X_t) + \lambda_1 \mathbb{E}^x \int_0^t f(X_s)ds,$$

where (X_t) is the process starting from x. The same equality holds for the (Y_t)-process starting from y. Making the difference of these two equalities, we obtain

$$|f(x) - f(y)| \leqslant \big|\mathbb{E}^x f(X_t) - \mathbb{E}^y f(Y_t)\big| + \lambda_1 \int_0^t \big|\mathbb{E}^x f(X_s) - \mathbb{E}^y f(Y_s)\big| ds. \quad (2.5)$$

Step 2. In order to get a lower bound of λ_1, in view of the right-hand side of (2.5), what we need to do is to estimate the term $\big|\mathbb{E}^x f(X_t) - \mathbb{E}^y f(Y_t)\big|$ only. This is the point where the coupling is adopted. Given a coupling process (X_t, Y_t) starting from (x, y), we have

$$\mathbb{E}^x f(X_t) - \mathbb{E}^y f(Y_t) = \mathbb{E}^{x,y}\big[f(X_t) - f(Y_t)\big].$$

Since f is smooth, it is Lipschitzian with respect to the metric ρ, we have

$$\big|\mathbb{E}^x f(X_t) - \mathbb{E}^y f(Y_t)\big| \leqslant L(f)\, \mathbb{E}^{x,y}\rho(X_t, Y_t), \quad (2.6)$$

where $L(f)$ is the Lipschitz constant of f.
Step 3. The main condition we need is the following:

$$\mathbb{E}^{x,y}\rho(X_t, Y_t) \leqslant \rho(x, y)\, e^{-\beta t}, \quad t \geqslant 0 \quad (2.7)$$

for some $\beta > 0$. Substituting (2.6) and (2.7) into (2.5), it follows that

$$|f(x) - f(y)| \leqslant L(f)\, \rho(x, y)\left[e^{-\beta t} + \lambda_1 \int_0^t e^{-\beta s} ds\right].$$

Letting $t \to \infty$, we obtain $|f(x) - f(y)| \leqslant L(f)\, \rho(x, y)\, \lambda_1/\beta$. Choosing a sequence $(x^{(n)}, y^{(n)})$ so that

$$\frac{|f(x^{(n)}) - f(y^{(n)})|}{\rho(x^{(n)}, y^{(n)})} \to L(f), \quad \text{as } n \to \infty$$

and then letting $n \to \infty$, we get $\lambda_1 \geqslant \beta$. The proof is finished.

To conclude this part, let us mention an application. Recall that the variation form of λ_1 for Brownian motion is the **Poincaré inequality**:

$$\|f - \pi f\|^2 \leqslant \frac{1}{\lambda_1} \int |\nabla f|^2,$$

which is just a special case Theorem 1.1. A related inequality is the **logarithmic Sobolev inequality** (Gross (1976)):

$$\int f^2 \log \frac{f^2}{\|f\|^2} \leqslant \frac{2}{\alpha} \int |\nabla f|^2$$

for some $\alpha > 0$. The last inequality is now well known and has a very wide range of applications. Especially, it played a critical role in the study of Malliavin calculus. Now, how about the relation between these two inequalities?

It was proved by Simon (1976) and Stroock (1984) that in general, we have $\lambda_1 \geqslant \alpha$. Conversely, it was proved by Deuschel and Stroock (1990) that for compact Riemannian manifold, we have

$$\alpha \geqslant \max\left\{\frac{\lambda_1}{d} - K,\ \frac{3\lambda_1 - Kd}{d+2}\right\},$$

which is sharp when $d = 1$ or $M = S^d$ $(d \geqslant 2)$. Thus, our theorem gives at the same time some new estimates for the constant in the logarithmic Sobolev inequality.

We have seen from the above proof, especially (2.7), that the coupling plays an essential role in the study. Hopefully, I do not need to say anymore about the importance of coupling (cf. Part Ⅳ below). It is the position to talk about

Part Ⅲ. Optimal Markovian Couplings

1. Markovian Couplings.

Definition 3.1. Given two Markov processes $P_k(t)$ on $(E_k, \mathscr{E}_k)$, $k = 1, 2$. A **Markovian coupling** is a Markov process $\widetilde{P}(t)$ on the product space $(E_1 \times E_2, \mathscr{E}_1 \times \mathscr{E}_2)$ having the **marginality**:

$$\begin{aligned}&\widetilde{P}(t; x_1, x_2; A_1 \times E_2) = P_1(t, x_1, A_1),\\ &\widetilde{P}(t; x_1, x_2; E_1 \times A_2) = P_2(t, x_2, A_2),\ \ t \geqslant 0,\ x_k \in E_k,\ A_k \in \mathscr{E}_k,\ k = 1, 2.\end{aligned} \tag{MP}$$

Equivalently,

$$\begin{aligned}&\widetilde{P}(t)f(x_1, x_2) = P_1(t)f(x_1),\\ &\widetilde{P}(t)f(x_1, x_2) = P_2(t)f(x_2),\quad t \geqslant 0,\ x_k \in E_k,\ f \in {}_b\mathscr{E}_k,\ k = 1, 2,\end{aligned} \tag{MP}$$

where ${}_b\mathscr{E}$ is the set of all bounded $\mathscr{E}$-measurable functions. Here, on the left-hand side, f is regarded as a bivariate function.

For the remainder of this section, we restrict ourselves to jump processes. To do so, we need some notation. Let $(E, \mathscr{E})$ be a measurable space such that $\{(x, x) : x \in E\} \in \mathscr{E} \times \mathscr{E}$ and $\{x\} \in \mathscr{E}$ for all $x \in E$. It is well-known that for a given sub-Markovian transition function $P(t, x, A)$ $(t \geqslant 0, x \in E, A \in \mathscr{E})$, if it does satisfy the **jump condition**

$$\lim_{t\to 0} P(t, x, \{x\}) = 1, \qquad x \in E, \tag{3.1}$$

then the limits

$$q(x) := \lim_{t\to 0} \frac{1 - P(t, x, \{x\})}{t} \quad \text{and} \quad q(x, A) := \lim_{t\to 0} \frac{P(t, x, A \setminus \{x\})}{t} \tag{3.2}$$

exist for all $x \in E$ and $A \in \mathscr{R}$, where

$$\mathscr{R} = \left\{ A \in \mathscr{E} : \lim_{t \to 0} \sup_{x \in A} \big[1 - P(t, x, \{x\})\big] = 0 \right\}.$$

Moreover, for each $A \in \mathscr{R}$, $q(\cdot)$, $q(\cdot, A) \in \mathscr{E}$, for each $x \in E$, $q(x, \cdot)$ is a finite measure on $(E, \mathscr{R})$ and

$$0 \leqslant q(x, A) \leqslant q(x) \leqslant \infty \quad \text{for all } x \in E \text{ and } A \in \mathscr{R}.$$

The pair $(q(x),\ q(x, A))\,(x \in E,\ A \in \mathscr{R})$ is called a **q-pair**. The q-pair is said to be **totally stable** if $q(x) < \infty$ for all $x \in E$. Then $q(x, \cdot)$ can be uniquely extended to the whole space $\mathscr{E}$ as a finite measure. Next, the q-pair $\big(q(x), q(x, A)\big)$ is called **conservative** if $q(x, E) = q(x) < \infty$ for all $x \in E$. Because of the above facts, we often call the sub-Markovian transition $P(t, x, A)$ satisfying (3.1) a **jump process** or a **q-process**.

When E is countable, conventionally we use the matrices $Q = (q_{ij} : \ i, j \in E)$ and $P(t) = (p_{ij}(t) : i, j \in E)$ instead of the q-pair and the jump process respectively. Here $q_{ii} = -q_i,\ i \in E$. We also call $P(t) = (p_{ij}(t))$ a **Markov chain** or a **Q-process**.

In practice, what we know in advance is the q-pair (also called the transition density or transition rate) $(q(x), q(x, dy))$ but not $P(t, x, dy)$. Hence, our real interest goes to the opposite direction. How does a q-pair determine the properties of $P(t, x, dy)$? A large part of the book (Chen (1992)) is devoted to the theory of jump processes. Here, we would like to mention that the theory now has some very nice application to the quantum physics which was missed in the book. Refer to the survey article by Konstantinov, Maslov and Chebotarev (1990) and references within.

A q-pair is called **regular** if it is totally stable, conservative and it determines uniquely a jump process.

We now return to our main context. As it was did in (Chen [1984, 1986]), we emphasizes the analysis of coupling operators. To illustrate this point, we introduce a simple result as follows. The total stability of coupling q-pairs was left as a hypothesis in the previous publications.

Lemma 3.2. Let $(q_k(x_k), q_k(x_k, dy_k))$ be a regular q-pair, $k = 1, 2$. Then for any coupling jump process $\widetilde{P}(t; x_1, x_2; dy_1, dy_2)$, its q-pair $(\tilde{q}(x_1, x_2), \tilde{q}(x_1, x_2; dy_1, dy_2))$ on $(E_1 \times E_2, \widetilde{\mathscr{R}})$ should satisfy

$$q_1(x_1) \vee q_2(x_2) \leqslant \tilde{q}(x_1, x_2) \leqslant q_1(x_1) + q_2(x_2),$$

where

$$\widetilde{\mathscr{R}} = \left\{ \tilde{A} \in \mathscr{E}_1 \times \mathscr{E}_2 : \lim_{t \to 0} \sup_{(x_1, x_2) \in \tilde{A}} \big[1 - \widetilde{P}(t; x_1, x_2; \{(x_1, x_2)\})\big] = 0 \right\}.$$

In particular, $(\tilde{q}(x_1, x_2),\ \tilde{q}(x_1, x_2; dy_1, dy_2))$ must be totally stable.

Proof. Denote by $P_k(t, x_k, dy_k)$ the jump process determined by

$$(q_k(x_k),\ q_k(x_k, dy_k)), \quad k = 1,\ 2$$

respectively. By the marginality for processes (MP), we have

$$\begin{aligned}\widetilde{P}(t; x_1, x_2; \{x_1\} \times \{x_2\}) &\geqslant \widetilde{P}(t; x_1, x_2; \{x_1\} \times E_2) - \widetilde{P}(t; x_1, x_2; E_1 \times (E_2 \setminus \{x_2\})) \\ &= \widetilde{P}(t; x_1, x_2; \{x_1\} \times E_2) - 1 + \widetilde{P}(t; x_1, x_2; E_1 \times \{x_2\}) \\ &= P_1(t, x_1, \{x_1\}) - 1 + P_2(t, x_2, \{x_2\}).\end{aligned}$$

By the first part of (3.2), this gives us $\tilde{q}(x_1, x_2) \leqslant q_1(x_1) + q_2(x_2)$. On the other hand, since

$$\widetilde{P}(t; x_1, x_2; \{x_1\} \times \{x_2\}) \leqslant \widetilde{P}(t; x_1, x_2; \{x_1\} \times E_2) = P(t, x_1, \{x_1\}),$$

we obtain $\tilde{q}(x_1, x_2) \geqslant q_1(x_1)$. $\square$

Given two regular marginal q-pairs, by Lemma 3.2, any coupling q-pair should be totally stable. It seems to the author that any coupling q-pair should also be conservative. But this is still an open question. (Note added in proof. An affirmative answer to the question has been obtained by Y. H. Zhang).

From now on, assume that all coupling operators considered below are conservative. Then, we have

$$\begin{aligned}\tilde{q}(x_1, x_2) &= \lim_{t \to 0} \frac{1 - \widetilde{P}(t; x_1, x_2; \{x_1\} \times \{x_2\})}{t}, \quad (x_1, x_2) \in E_1 \times E_2 \\ \tilde{q}(x_1, x_2;\ \tilde{A}) &= \lim_{t \to 0} \frac{1 - \widetilde{P}(t; x_1, x_2; \tilde{A})}{t}, \qquad (x_1, x_2) \notin \tilde{A} \in \mathscr{E}_1 \times \mathscr{E}_2.\end{aligned} \tag{3.3}$$

Define

$$\Omega_1 f(x_1) = \int q_1(x_1, dy_1)[f(y_1) - f(x_1)], \quad f \in {}_b\mathscr{E}_1.$$

Similarly, we can define Ω_2. Corresponding to the coupling process $\widetilde{P}(t)$ we also have $\widetilde{\Omega}$. Because of the one-to-one correspondence between a q-pair and its operator Ω, we will use both according to our convenience. Now, since the marginal q-pairs and the coupling q-pair are all conservative, it is not difficult to prove that (MP) implies that

$$\begin{aligned}\widetilde{\Omega} f(x_1, x_2) &= \Omega_1 f(x_1), \quad f \in {}_b\mathscr{E}_1 \\ \widetilde{\Omega} f(x_1, x_2) &= \Omega_2 f(x_2), \quad f \in {}_b\mathscr{E}_2,\ x_k \in E_k,\ k = 1,\ 2.\end{aligned} \tag{MO}$$

Again, on the left-hand side, f is regarded as a bivariate function. Refer to Chen [1986a or 1992, Chapter 5]. Here, "MO" means the marginality for operators.

Definition 3.3. Any operator $\widetilde{\Omega}$ satisfying (MO) is called a **coupling operator**.

Before moving further, we recall some coupling operators for Markov chains. In the following examples, f is a bounded function on $E_1 \times E_2$.

Independent coupling $\widetilde{\Omega}_0$.

$$\widetilde{\Omega}_0 f(i_1, i_2) = (\Omega_1 f(\cdot, i_2))(i_1) + (\Omega_2 f(i_1, \cdot))(i_2), \quad i_k \in E_k,\ k = 1, 2.$$

This coupling is trivial but it does show that a coupling operator always exists.

To simplify our notation, in what follows, instead of writing down a Q-matrix or its operator, we will use tables. For instance, a birth-death Q-matrix can be expressed as follows:

$$\begin{array}{rll} i \to i+1 & \text{at rate} & b_i = q_{i,i+1} \\ \to i-1 & \text{at rate} & a_i = q_{i,i-1}. \end{array}$$

Classical coupling $\widetilde{\Omega}_c$. Take $E_1 = E_2 = E$ and let the two marginal Q-matrices be the same $Q = (q_{ij})$. The coupling process evolves as follows: If $i_1 \neq i_2$, then

$$\begin{array}{rll} (i_1, i_2) \to (j_1, i_2) & \text{at rate} & q_{i_1 j_1} \\ \to (i_1, j_2) & \text{at rate} & q_{i_2 j_2}. \end{array}$$

Otherwise,

$$(i, i) \ \to \ (j, j) \ \text{ at rate } \ q_{ij}.$$

Each coupling has its own character. The classical coupling means that the marginals evolve independently until they meet. Then, they move together. A nice way to interpret this coupling is to use a Chinese idiom: fall in love at first sight. That is, the boy and girl had independent paths of their lives before the first time they met each other. Once they met, they are in love at once and will have the same path of their lives forever. When the marginal Q-matrices are the same, all couplings considered below will have the property listed in the last line and hence we will not mention again.

Basic coupling $\widetilde{\Omega}_b$.

$$\begin{array}{rll} (i_1, i_2) \to (j, j) & \text{at rate} & q^{(1)}_{i_1 j} \wedge q^{(2)}_{i_2 j} \\ \to (j, i_2) & \text{at rate} & \big(q^{(1)}_{i_1 j} - q^{(2)}_{i_2 j}\big)^+ \\ \to (i_1, j) & \text{at rate} & \big(q^{(2)}_{i_2 j} - q^{(1)}_{i_1 j}\big)^+, \quad i_1, i_2 \in E. \end{array}$$

The basic coupling means that the components jump to the same place with the biggest possible rate. This explains where the term $q^{(1)}_{i_1 j} \wedge q^{(2)}_{i_2 j}$ comes from, which is the biggest one to guarantee the marginality. This term is the key of the coupling. Note that whenever we have a term $A \wedge B$, we should have the other two terms $(A - B)^+$ and $(B - A)^+$ automatically, again, due to the marginality. Thus, in what follows, we will write down the term $A \wedge B$ only for simplicity.

March coupling $\widetilde{\Omega}_m$. Take $E = \{0, 1, 2, \cdots\}$ and let

$$(i_1, i_2) \ \to \ (i_1 + k, i_2 + k) \ \text{ at rate } \ q^{(1)}_{i_1, i_1 + k} \wedge q^{(2)}_{i_2, i_2 + k},$$

here we have used the convention that $q_{ij} = 0$ for all $i \in E$ and $j \notin E$.

The word "march" is a Chinese name, which is the command to soldiers to start marching. Thus, this coupling means that at each step, the components maintain the same length of jumps with the biggest possible rate.

In the time-discrete case, the classical coupling and the basic coupling are due to Doeblin (1938) and Wasserstein (1969) respectively. The march coupling is due to Chen (1986b). The original purpose for the last coupling is mainly to keep the order-preservation (cf. Part Ⅳ below).

Let us now consider a birth-death process with regular Q-matrix:

$$q_{i,i+1} = b_i, \quad i \geqslant 0; \quad q_{i,i-1} = a_i, \quad i \geqslant 1.$$

Then for two copies of the process starting from i_1 and i_2 respectively, we have

Modified march coupling $\widetilde{\Omega}_{cm}$ (Chen (1990)). Take $\widetilde{\Omega}_{cm} = \widetilde{\Omega}_c$ if $|i_1 - i_2| \leqslant 1$ and $\widetilde{\Omega}_{cm} = \widetilde{\Omega}_m$ if $|i_1 - i_2| \geqslant 2$.

Coupling by inner reflection $\widetilde{\Omega}_{ir}$ (Chen (1990)). Again, take $\widetilde{\Omega}_{ir} = \widetilde{\Omega}_c$ if $|i_1 - i_2| \leqslant 1$. For $i_2 \geqslant i_1 + 2$, take

$$\begin{array}{lll} (i_1,\, i_2) \to (i_1 + 1,\, i_2 - 1) & \text{at rate} & b_{i_1} \wedge a_{i_2} \\ \to (i_1 - 1,\, i_2) & \text{at rate} & a_{i_1} \\ \to (i_1,\, i_2 + 1) & \text{at rate} & b_{i_2}. \end{array}$$

By exchanging i_1 and i_2, we can get the expression of $\widetilde{\Omega}_{ir}$ for the case that $i_1 \geqslant i_2$.

This coupling lets the components move to the closed place (not necessarily the same place as required by the basic coupling) with the biggest possible rate.

From these examples one sees that there are many choices of coupling operator $\widetilde{\Omega}$. Indeed, there are infinite many choices! Thus, in order to use the coupling technique, a basic problem we should study is the regularity of coupling operators. For which, fortunately, we have a complete answer (Chen [1986a or 1992, Chapter 5]).

Theorem 3.4. If the given two marginal q-pairs are regular, then any coupling q-pair (resp., operator) is regular. Conversely, if a coupling q-pair is regular then so are its two marginals. Moreover, (MP) and (MO) are equivalent.

Clearly, Theorem 3.4 simplifies greatly our study on couplings for general jump processes since the marginality (MP) of a coupling process is reduced to the rather simpler marginality (MO) of the corresponding operator.

2. Optimal Markovian Couplings.

Since there are infinite many Markovian couplings, I asked myself several times in the past years: Does there exist an optimal one? Now, let me explain the way how I obtained a reasonable notion for optimal Markovian couplings. The first time we touched this problem was in Chen and Li (1989). It was proved there for Brownian motion, the coupling by reflection (introduced first by Lindvall and

Rogers (1986) in terms of stochastic differential equations) is optimal with respect to the total variation and moreover, for different probability metrics, the effective couplings can be different. At the second time, in Chen (1990), it was proved that for birth-death processes, we have an order as follows:

$$\widetilde{\Omega}_{ir} \succ \widetilde{\Omega}_b \succ \widetilde{\Omega}_c \succ \widetilde{\Omega}_{cm} \succ \widetilde{\Omega}_m,$$

where $A \succ B$ means that A is better than B in some sense. However, only in the last summer, it became clear to the author how to optimalize couplings.

To study optimal couplings, we need one more preparation. As was mentioned several times in the previous publications (Chen [1989a, 1989b, 1992] and Chen and Li (1989)) that it should be helpful to keep in mind the relation between couplings and the probability metrics. It will be clear soon, this is actually one of the key ideas of the study. So far as I know, there are more than 16 different probability metrics, including the total variation, the Lévy-Prohorov metric for the weak convergence and so on. But we often concern with another metric W. Let $(E, \rho, \mathscr{E})$ be a metric space. The **minimum L^1-metric** W is defined by:

$$W(P_1, P_2) = \inf_{\widetilde{P}} \int \rho(x_1, x_2)\widetilde{P}(dx_1, dx_2), \tag{3.3}$$

where $\widetilde{P}$ varies over all couplings of P_1 and P_2. This metric has many different names. It plays an important role in the study of random fields and interacting particle systems. Here, we mention a result due to Dobrushin (1970), which says that W is equivalent to the Lévy-Prohorov metric when ρ is bounded and W equals half of the total variation when ρ is the discrete metric d: $d(x, y) = 0$ if $x = y$ and $d(x, y) = 1$ if $x \neq y$. Refer to Chen [1992, Chapter 0 and Chapter 5] for more information about W.

In view of (3.3), we see that any coupling provides an upper bound of $W(P_1, P_2)$. Thus, it is very natural to introduce the following notion.

Definition 3.5. A coupling $\overline{P}$ of P_1 and P_2 is called **ρ-optimal** if

$$\int \rho(x_1, x_2)\overline{P}(dx_1, dx_2) = W(P_1, P_2).$$

Now, it is natural to define the optimal coupling for time-discrete Markov processes without restricted to the Markovian class. In the special case of ρ being the discrete metric (or equivalently, restricted to the total variation), it is just the **maximal coupling**, started by Griffeath (1975). However, it is well known that the maximal couplings are usually non-Markovian. Even though the maximal couplings as well as other non-Markovian couplings now consist of an important part of the theory and have been widely studied in the literature (refer to Lindvall (1992) and references therein). They are difficult to handle especially when we come to the time-continuous situation. Moreover, it will be clear soon that in the context of diffusions, to deal with the optimal Markovian coupling in terms of their operators, the discrete metric will lost its meaning. Thus, our optimal

Markovian couplings are essentially different from the maximal ones. It should be also pointed out that the sharp estimates introduced in Part II are obtained from the exponential rate in the W-metric with respect to some much more refined metric ρ rather than the discrete one.

Replacing P_k and $\widetilde{P}$ with $P_k(t)$ and $\widetilde{P}(t)$ respectively and then going to the operators, it is not far away to arrive at the following notion (cf. Chen (1993a) for details):

Definition 3.6. A coupling operator $\overline{\Omega}$ is called **ρ-optimal** if

$$\overline{\Omega}\,\rho(x_1, x_2) = \inf_{\widetilde{\Omega}} \widetilde{\Omega}\,\rho(x_1, x_2)$$

for all x_1 and x_2, where $\widetilde{\Omega}$ varies over all coupling operators.

To see the notion is useful, let me introduce one more coupling.

Coupling by reflection. Given a birth-death process with birth rates b_i and death rates a_i. The coupling evolves in the following way: If $i_2 = i_1 + 1$, then

$$\begin{array}{rll} (i_1,\, i_2) \to (i_1 - 1,\, i_2 + 1) & \text{at rate} & a_{i_1} \wedge b_{i_2} \\ \to (i_1 + 1,\, i_2) & \text{at rate} & b_{i_1} \\ \to (i_1,\, i_2 - 1) & \text{at rate} & a_{i_2}. \end{array}$$

If $i_2 \geqslant i_1 + 2$, then

$$\begin{array}{rll} (i_1,\, i_2) \to (i_1 - 1,\, i_2 + 1) & \text{at rate} & a_{i_1} \wedge b_{i_2} \\ \to (i_1 + 1,\, i_2 - 1) & \text{at rate} & b_{i_1} \wedge a_{i_2}. \end{array}$$

By symmetry, we can write down the rates for the other case that $i_1 > i_2$.

Intuitively, the reflection in outside direction is quite strange since it makes the components apart by distance 2 but not by 1. For this reason, even though the coupling came to my attention years ago, I never believed that it could be better than the coupling by inner reflection. But the next result changes my mind.

Theorem 3.7 (Chen (1993a)). For birth-death processes, the coupling by reflection is ρ-optimal for any translation-invariant metric ρ on $\mathbf{Z}_+$ having the property:

$$u_k := \rho(0, k+1) - \rho(0, k), \quad k \geqslant 0$$

is non-increasing in k.

To see that the optimal coupling depends heavily on the metric ρ, note that the above metric ρ can be rewritten as

$$\rho(i, j) = \sum_{k<|i-j|} u_k$$

for some positive non-increasing sequence (u_k). In this way, for any positive sequence (u_k), we can introduce another metric as follows:

$$\tilde{\rho}(i, j) = \left| \sum_{k<i} u_k - \sum_{k<j} u_k \right|.$$

Because $(u_k > 0)$ is arbitrary, this class of metrics is still quite large. Now, among the couplings listed above, which one is $\tilde{\rho}$-optimal coupling?

Theorem 3.8 (Chen (1993a)). For birth-death processes, every coupling mentioned above except the trivial one is $\tilde{\rho}$-optimal.

This result is again quite surprising, far away from our probabilistic intuition. Thus, our optimality does produce some unexpected results!

3. Couplings of diffusion processes.

We now turn to study the couplings for diffusion processes in $\mathbf{R}^d$ with second differential operator

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

For simplicity, we write $L \sim (a(x), b(x))$. Given two diffusions with operators

$$L_k \sim (a_k(x), b_k(x)), \quad k = 1,\, 2$$

respectively, it is clear that the coefficients of any coupling operator $\widetilde{L}$ should be of the form

$$a(x,y) = \begin{pmatrix} a_1(x) & c(x,y) \\ c(x,y)^* & a_2(y) \end{pmatrix}, \quad b(x,y) = \begin{pmatrix} b_1(x) \\ b_2(y) \end{pmatrix}.$$

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

As analog of Markov chains, we have the following examples:

Classical coupling. $c(x,y) \equiv 0$.

March coupling (Chen and Li (1989)). Let $a_k(x) = \sigma_k(x)\sigma_k(x)^*$, $k = 1,\, 2$. Take $c(x,y) = \sigma_1(x)\sigma_2(y)^*$.

Coupling by reflection. Let $L_1 = L_2$. Take

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

(Lindvall and Rogers (1986), Chen and Li (1989)) or

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

where $\bar{u} = (x-y)/|x-y|$.

We are now ready to study the optimal couplings for diffusion processes. Given a metric $\rho \in C^2(\mathbf{R}^d \times \mathbf{R}^d \setminus \{(x,x) : x \in \mathbf{R}^d\})$, a coupling operator $\overline{L}$ is called ρ**−optimal** if

$$\overline{L}\rho(x,y) = \inf_{\widetilde{L}} \rho(x,y), \quad x \neq y,$$

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

Theorem 3.9 (Chen (1993a)). Let $f \in C^2(\mathbf{R}_+;\mathbf{R}_+)$ with $f(0)=0$, $f'>0$ and $f'' \leqslant 0$. Set $\rho(x,y)=f(|x-y|)$. Then, the ρ-optimal solution $c(x,y)$ is given as follows.

(1) If $d=1$, then $c(x,y)=-\sqrt{a_1(x)a_2(y)}$ and moreover,

$$\overline{L}f(|x-y|)=\frac{1}{2}\big(\sqrt{a_1(x)}+\sqrt{a_2(y)}\,\big)^2 f''(|x-y|)+\frac{(x-y)(b_1(x)-b_2(y))}{|x-y|}f'(|x-y|).$$

Next, suppose that $a_k=\sigma_k^2$ $(k=1,\,2)$ is non-degenarated and write

$$c(x,y)=\sigma_1(x)H^*(x,y)\sigma_2(y).$$

(2) If $f''(r)<0$ for all $r>0$, then $H(x,y)=U(\gamma)^{-1}\big[U(\gamma)U(\gamma)^*\big]^{1/2}$, where

$$\gamma=1-\frac{|x-y|f''(|x-y|)}{f'(|x-y|)} \quad \text{and} \quad U(\gamma)=\sigma_1(x)(I-\gamma\bar{u}\bar{u}^*)\sigma_2(y).$$

(3) If $f(r)=r$, then $H(x,y)$ is a solution to the equation:

$$U(1)H=\big(U(1)U(1)^*\big)^{1/2}.$$

(4) In particular, if $f(r)=r$ and $a_k(x)=\varphi_k(x)\sigma^2$ for some positive function φ_k $(k=1,\,2)$, where σ is independent of x and $\det\sigma>0$. Then $H(x,y)=I-2\sigma^{-1}\bar{u}\bar{u}^*\sigma^{-1}/|\sigma^{-1}\bar{u}|^2$. Moreover,

$$\begin{aligned}&\overline{L}f(|x-y|)\\&=\frac{1}{2|x-y|}\Big\{\big(\sqrt{\varphi_1(x)}-\sqrt{\varphi_2(y)}\,\big)^2\big[\operatorname{tr}\sigma^2-|\sigma\bar{u}|^2\big]+2\langle x-y,\,b_1(x)-b_2(y)\rangle\Big\}.\end{aligned}$$

Finally, without the condition "$f(r)=r$", part (4) still holds provided the metric $\rho(x,y)=f(|x-y|)$ is replaced by $\rho(x,y)=f(|\sigma^{-1}(x-y)|)$. Furthermore,

$$\begin{aligned}\overline{L}\rho(x,y)=\frac{1}{2}\big(\sqrt{\varphi_1(x)}+\sqrt{\varphi_2(y)}\,\big)^2 f''(|\sigma^{-1}(x-y)|)\\+\Big\{(d-1)\big(\sqrt{\varphi_1(x)}-\sqrt{\varphi_2(y)}\,\big)^2+2\big\langle\sigma^{-1}(x-y),\,\sigma^{-1}(b_1(x)-b_2(y))\big\rangle\Big\}\\\times\frac{f'(|\sigma^{-1}(x-y)|)}{2|\sigma^{-1}(x-y)|}.\end{aligned}$$

Note that in the last assertion of the theorem, we have replaced the ordinary Riemannian metric I with the new one σ^{-2}. This idea is useful in other cases (see Chen (1993a) and Chen and Wang (1993b)). The above theorem can be used to improve the previous results on success of couplings and the gradient estimates (Chen and Li (1989), Cranston (1992)). See also the next part. As a generalization of the Euclidean case, the coupling by reflection for Brownian motion on Riemannian manifold was constructed by Kendall (1986). See also Cranston (1991).

Part IV. Applications of Coupling Method

It should be helpful for the readers, especially for the newcomers, to survey some applications of couplings. Of course, the applications discussed below can not be complete and depend on the personal test. One may refer to Liggett (1985) and Lindvall (1992) for additional information. Again, we emphasize the main ideas by using couple simple examples. In particular, throughout this part, the metric is taken to be $\rho(x,y) = |x-y|$. That is, $f(r) = r$. In view of Theorem 3.9, this metric may not be optimal since $f'' = 0$. Thus, in practice, an additional work is often needed in order to figure out an effective metric ρ.

1. Spectral gap. Exponential L^2-convergence.

Consider the Ornstein-Uhlenbeck process in $\mathbf{R}^d$. By Theorem 3.9, we have $\overline{L}\rho(x,y) \leqslant -\rho(x,y)$ and so

$$\mathbb{E}^{x,y}\rho(X_t, Y_t) \leqslant \rho(x,y)e^{-t}. \tag{4.1}$$

As we mentioned before, this gives us $\lambda_1 \geqslant 1$, which is indeed exact! Refer to Chen and Wang (1993b) for general results and much more examples.

2. Algebraic L^2-convergence. Lipschitz contractivity.

If the process is not exponential L^2-convergence, one may look for a slower convergence:

$$\|P(t)f - \pi f\| \leqslant V(f)/t^\nu, \quad t > 0,\ f \in L^2(\pi)$$

for some $V : L^2(\pi) \to [0,\infty]$ and $\nu > 0$. Such convergence is called **algebraic** or **geometric L^2-convergence**. It turns out that in this context, the following **Lipschitz contractivity** plays a critical role (cf. Liggett (1991) and Chen (1993a)):

$$L(P(t)f) \leqslant L(f), \quad t \geqslant 0, \tag{4.2}$$

where $L(f)$ is the Lipschitz constant of f. The coupling method provides a natural tool to deduce the property (4.2). For instance, for Brownian motion in $\mathbf{R}^d$, since $\overline{L}\rho(x,y) \leqslant 0$, we have $\mathbb{E}^{x,y}\rho(X_t, Y_t) \leqslant \rho(x,y)$. In other words, (4.2) holds. Even though the proof is extremely simple and very natural. It is indeed enough for us to improve some previous results (see Chen (1993a)).

3. Ergodicity.

The coupling method is often used to study the ergodicity of Markov processes. For instance, for Ornstein-Uhlenbeck process, from (4.1), it follows that

$$W(P(t,x,\cdot),\pi) \leqslant C(x)e^{-t}, \quad t \geqslant 0, \tag{4.3}$$

where π is the stationary distribution of the process and W is the minimum L^1-metric. The estimate (4.3) simply means that the process is exponentially ergodic

with respect to the minimum L^1-metric. See Chen [1992, Chapter 14] and Chen (1993b) for details.

Recall that the coupling time T is defined by $T = \inf\{t \geqslant 0 : X_t = Y_t\}$. Starting from time T, we can adopt the march coupling so that the two components will move together. Then, we have

$$\|P(t,x,\cdot) - P(t,y,\cdot)\|_{\mathrm{Var}} \leqslant 2\,\mathbb{E}^{x,y} I_{[X_t \neq Y_t]} = 2\,\mathbb{P}^{x,y}[T > t]. \tag{4.4}$$

If $\mathbb{P}^{x,y}[T > t] \to 0$ as $t \to \infty$, then the existence of a stationary distribution plus (4.4) gives us the ergodicity with respect to the total variation. See Lindvall (1992) for details and references on this topic. Actually, for Brownian motion, as pointed out in Chen and Li (1989), the coupling by reflection provides the sharp estimate for the total variation.

4. Gradient estimate.

Recall that for every suitable function f, we have

$$f(x) - f(y) = \mathbb{E}^{x,y}\big[f\big(X_{t\wedge T}\big) - f\big(Y_{t\wedge T}\big)\big] - \mathbb{E}^{x,y}\int_0^{t\wedge T} \big[Lf\big(X_s\big) - Lf\big(Y_s\big)\big]ds.$$

Thus, if f is L-harmonic, i.e., $Lf = 0$, then we have

$$f(x) - f(y) = \mathbb{E}^{x,y}\big[f\big(X_{t\wedge T}\big) - f\big(Y_{t\wedge T}\big)\big].$$

Hence

$$|f(x) - f(y)| \leqslant 2\,\|f\|_\infty\,\mathbb{P}^{x,y}[T > t].$$

Letting $t \to \infty$, we obtain

$$|f(x) - f(y)| \leqslant 2\,\|f\|_\infty\,\mathbb{P}^{x,y}[T = \infty].$$

Now, if f is bounded and $\mathbb{P}^{x,y}[T = \infty] = 0$, then $f = \text{const.}$ Otherwise, if $\mathbb{P}^{x,y}[T = \infty] \leqslant \text{const.}\,\rho(x,y)$, then we get

$$\|\nabla f\|_\infty \leqslant \text{const.}\,\|f\|_\infty,$$

which is the gradient estimate we are looking for (cf. Cranston (1991, 1992) and Wang (1992a, 1993c, d)). For Brownian motion in $\mathbf{R}^d$, the optimal coupling gives us $\mathbb{P}^{x,y}[T < \infty] = 1$, and so $f = \text{const.}$ We have thus proved a well-known result: every bounded harmonic function should be constant.

5. Construction of reaction-diffusion processes.

The state space is $\mathbf{Z}_+^{\mathbf{Z}^d}$, which is not locally compact. Since the state space is quite poor, the usual technique of constructing the Markov processes is not suitable. Our construction goes as follows: Take a sequence (Λ_n) of finite subsets of $\mathbf{Z}^d$ instead of $\mathbf{Z}^d$, we obtain a sequence of Markov chains $P_n(t,x,\cdot)$ $(n \geqslant 1)$

with state space $\mathbf{Z}_+^{\Lambda_n}$. Then, prove that $(P_n(t) : n \geqslant 1)$ is a Cauchy sequence in the minimum L^1-metric W with respect to the metric p:

$$p(x,y) = \sum_{u \in \mathbf{Z}^d} k_u |x_u - y_u|, \qquad x = (x_u : u \in \mathbf{Z}^d), \quad y = (y_u : u \in \mathbf{Z}^d) \in \mathbf{Z}_+^{\mathbf{Z}^d},$$

where (k_u) is a positive sequence on $\mathbf{Z}^d$. To do so, we adopt the coupling approach. This probability metric W, which is stronger than the weak convergence since the metric p on the state space is unbounded, enables us to prove not only the existence of a limit $P(t,x,\cdot)$ of the sequence $(P_n(t,x,\cdot))$ for fixed t and x but also the Chapman-Kolmogorov equation of $P(t,x,\cdot)$. Furthermore, the coupling method is used to study the ergodicity of the infinte-dimensional process. Refer to Chen [1992, Part Ⅳ and 1993b] for details.

6. Construction of diffusion processes on Sierpinski carpet.

Since the state space is irregular, the traditional construction is again not suitable. Actually, the construction of diffusion processes on higher dimensional $(d \geqslant 3)$ Sierpinski carpet was opened for several years. It has been solved very recently by Barlow and Bass (1993). The main tool to overcome the difficulty is again the coupling method.

7. Comparison results.

The stochastic order occupies a critical position in the study of probability theory as the usual order-relation is an fundamental structure in mathematics.

Definition 4.1. Let $\mathscr{M}$ be the set of all bounded monotone increasing functions in $\mathbf{R}^d$ with respect to the ordinary semi-order "$\leqslant$". Given μ_1, $\mu_2 \in \mathscr{P}(\mathbf{R}^d)$, we say that $\mu_1 \prec \mu_2$ if for all $f \in \mathscr{M}$, $\mu_1 f \leqslant \mu_2 f$. Given two processes $P_1(t)$ and $P_2(t)$ in $\mathbf{R}^d$, we say that $P_1(t) \prec P_2(t)$ if for all $f \in \mathscr{M}$, $P_1(t)f(x_1) \leqslant P_2(t)f(x_2)$ whenever $x_1 \leqslant x_2$. If in addition $P_1(t) = P_2(t)$, we call $P_1(t)$ monotone.

The coupling method provides a natural way to study the order-preserving property. Refer to Chen [1992, Chapter 5] for the study on jump processes. Here is an example for diffusions.

Example 4.2. Consider two diffusions in $\mathbf{R}$ with

$$a_1(x) = a_2(x) = a(x), \qquad b_1(x) \leqslant b_2(x). \tag{4.5}$$

Then, we have $P_1(t) \prec P_2(t)$.

The conclusion was proved in Ikeda and Watanabe [1981, Section 6.1] by using stochastic differential equation. The same proof with a slight modification works if we adopt the march coupling.

Actually, a criterion for the order-preservation for multidimensional diffusion processes is now presented in Chen and Wang (1993a). From which, we see that the condition (4.5) is not only sufficient but also necessary. A related topic, the preservation of positive correlations for diffusions, is also solved in the same paper.

To illustrate an application of the study, let me introduce a simple example.

Example 4.3. Let μ^λ be the Poisson measure on $\mathbf{Z}_+$ with parameter λ:

$$\mu^\lambda(k) = \frac{\lambda^k}{k!}e^{-\lambda}, \quad k \geqslant 0.$$

Then, we have $\mu^\lambda \prec \mu^{\lambda'}$ whenever $\lambda \leqslant \lambda'$.

I have seen in some books, one proved such kind of result by constructing a coupling measure $\tilde{\mu}$ so that $\tilde{\mu}\{(x,y) : x \leqslant y\} = 1$. Of course, such a proof is lengthy. So I now introduce a very short proof based on the coupling argument.

Consider a birth-death process with rate

$$a(k) \equiv 1, \qquad b^\lambda(k) = \frac{\mu^\lambda(k+1)}{\mu^\lambda(k)} = \frac{\lambda}{k+1} \uparrow \text{ as } \lambda \uparrow .$$

Denote by $P^\lambda(t)$ the corresponding process. It should be clear that

$$P^\lambda(t) \prec P^{\lambda'}(t) \quad \text{whenever} \quad \lambda \leqslant \lambda'$$

(cf. Chen(1992, Theorem 5.26)). Then, by ergodic theorem,

$$\mu^\lambda f = \lim_{t\to\infty} P^\lambda(t)f \leqslant \lim_{t\to\infty} P^{\lambda'}(t)f = \mu^{\lambda'} f$$

for all $f \in \mathscr{M}$. Clearly, the technique by using stochastic processes (goes back to Holley (1974)) provides an intrinsic insight of the order-preservation for probability measures.

An aspect of the applications of coupling method is to compare a rather complicated process with a simpler one. To get some impression, we introduce an example which was used by Chen and Lu (1990) in the study on large deviations for Markov chains.

Example 4.4. Consider a single birth Q-matrix $Q = (q_{ij})$, which means that

$$q_{i,i+1} > 0, \quad \text{and} \quad q_{ij} = 0 \quad \text{for all} \quad j > i+1,$$

and a birth-death Q-matrix $\overline{Q} = (\bar{q}_{ij})$ with $\bar{q}_{i,i-1} = \sum_{j<i} q_{ij}$. If $\bar{q}_{i,i+1} \geqslant q_{i,i+1}$ for all $i \geqslant 0$, then $P(t) \prec \overline{P}(t)$.

The conclusion can be easily deduced by the following coupling:

$$\begin{array}{lll}
(i_1,\, i_2) \to (i_1-k,\, i_2-1) & \text{at rate} & q_{i_1,i_1-k} \wedge q_{i_2,i_2-k} \\
\to (i_1-k,\, i_2) & \text{at rate} & (q_{i_1,i_1-k} - q_{i_2,i_2-k})^+ \\
\to (i_1,\, i_2-1) & \text{at rate} & (q_{i_2,i_2-k} - q_{i_1,i_1-k})^+ \\
\to (i_1+1,\, i_2+1) & \text{at rate} & q_{i_1,i_1+1} \wedge \bar{q}_{i_2,i_2+1} \\
\to (i_1+1,\, i_2) & \text{at rate} & (q_{i_1,i_1+1} - \bar{q}_{i_2,i_2+1})^+ \\
\to (i_1,\, i_2+1) & \text{at rate} & (\bar{q}_{i_2,i_2+1} - q_{i_1,i_1+1})^+,
\end{array}$$

here we have used the convention: $q_{ij} = 0$ if $j < 0$. Refer to Chen [1992, Theorem 8.24] for details. This example illustrates the flexibility in the application of couplings.

Acknowledgements. Most part of the materials in the paper was talked at several universities or institutes in Canada during January–February, 1993 and in Italy during May–June, 1993. It was also talked at the Six International Vilnuis Conference on Probability and Mathematical Statistics (June 28–July 4, 1993) and at the International Conference on Dirichlet Forms and Stochastic Processes (October 25–31, 1993). A preliminary version of the paper was presented at the National Conference on Probability Theory, June, 1992. The author would like to thank the following mathematicians for their hospitality and financial supports: Prof. J. D. Chen and Prof. G. Q. Zhang at Beijing U., Prof. D. A. Dawson, Dr. S. Feng and Dr. Y. D. Wu at Carleton U., Prof. G. O'Brien, Prof. N. Madras and Dr. J. M. Sun at York U., Prof. D. McDonald and Dr. K. Qian at U. of Ottawa, Prof. M. Barlow, Prof. E. A. Perkins, Dr. S. J. Luo and Dr. L. W. Zhang at U. of British Columbia, Prof. E. Scacciatelli and Prof. A. Pellegrinotti at U. of Roma I, Prof. V. Capasso and Prof. Y. G. Lu at U. of Bari, Prof. L. Accardi at U. of Roma II, Prof. C. Boldrighini at U. of Camerino, Prof. B. Grigelionis at Akademijios, Lithuania, Prof. L. Stettner and Prof. J. Zabczyk at Polish Academy of Sciences.

References

Barlow, M. T. and Bass, R (1993), *Coupling and Harnack inequality for Sierpinski carpet*, In preparation.

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.

Chavel, I. (1984), *Eigenvalues in Riemannian Geometry*, Academic Press.

Chen, M. F. (1984), *Couplings of Markov chains* (*In Chinese*), J. Beijing Normal Univ., 4, 3-10.

Chen, M. F. (1986a), *Couplings of jump processes*, Acta Math. Sinica, New Series, 2:2, 123-136.

Chen, M. F. (1986b), *Jump Processes and Interacting Particle Systems* (*In Chinese*), Beijing Normal Univ. Press.

Chen, M. F. (1989a), *Probability metrics and coupling methods*, Pitman Research Notes in Math., 200, 55-72.

Chen, M. F. (1989b), *A survey on random fields* (*In Chinese*), Advances in Math., 18:3, 294-322.

Chen, M. F. (1990), *Ergodic theorems for reaction-diffusion processes*, J. Statis. Phys. 58:5/6, 939-966.

Chen, M. F. (1991a), *On coupling of jump processes*, Chin. Ann. Math. 12(B)4, 385-399.

Chen, M. F. (1991b), *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.

Chen, M. F. (1993a), *Optimal Markovian couplings and applications*, Technical Report, No.215 (1993), Carleton Univ. and No.147(1993), C. V. Volterra, Univ. of Roma II. Acta Math. Sin. New Ser. 10:3, 260-275 (1994).

Chen, M. F. (1993b), *On ergodic region of Schlögl's model*, Carr Reports in Math.&Phys. No.13 (1993).

Chen, M. F. and Li, S. F. (1989), *Coupling methods for multi-dimensional diffusion processes*, Ann. of Probab. 17:1, 151–177.

Chen, M. F. and Lu, Y. G. (1990), *Large deviations for Markov chains*, Acta Sci. Sin. 10:2, 217–222.

Chen, M. F. and Wang, F. Y. (1992), *Application of coupling method to the first eigenvalue on manifold*, Sci. Sin.(A), 23:11(1993)(Chinese Edition), 37:1(1994)(English Edition).

Chen, M. F. and Wang, F. Y. (1993a), *On order-preservation and positive correlations for multidimensional diffusion processes*, Prob. Th. Rel. Fields 95, 421–428.

Chen, M. F. and Wang, F. Y. (1993b), *Estimation of the first eigenvalue of second order elliptic operators*, In preparation.

Cranston, M. (1991), *Gradient estimates on manifolds using coupling*, J. Funct. Anal. 99, 110–124.

Cranston, M. (1992), *A probabilistic approach to gradient estimates*, Canad. Math. Bull. 35, 46–55.

Deuschel, J.-D. and Stroock, D. W. (1990), *Hypercontractivity and spectral gap of symmetric diffusion with applications to the stochastic Ising models*, J. Funct. Anal. 92, 30–48.

Diaconis and Stroock (1991), *Geometric bounds for eigenvalues of Markov chains*, Ann. Appl. Prob. 1:1, 36–61.

Doeblin, W. (1938), *Exposé de la théorie des chaines simples constantes de Markov à un nombre dini d'etats*, Rev. Math. Union Interbalkanique 2, 77–105.

Dobrushin, R. L. (1970), *Prescribing a system of random variables by conditional distributions*, Theory Prob. Appl., 15, 458-486.

Griffeath, D.(1975), *A maximal coupling for Markov chains*, Z. Wahrs. 31, 95–106.

Griffeath, D.(1978), *Coupling methods for Markov processes*, in "Studies in Probability and Ergodic Theory, Adv. Math., Supplementary Studies", Academic Press 2, 1–43.

Gross, L. (1976), *Logarithmic Sobolev inequalities*, Amer. J. Math. 97, 1061–1083.

Holley, R. (1974), *Recent results on the stochastic Ising model*, Rocky Mountain J. Math. 4:3, 479–496.

Holley, R. and Stroock, D. W. (1989), *Uniform and L^2 convergence in one dimensional stochastic Ising models*, Comm. Math. Phys. 123, 85–93.

Ikeda, N. and Watanabe, S. (1981), *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Kodansha, Tokyo.

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.

Kendall, W. (1986), *Nonnegative Ricci curvature and the Brownian coupling property*, Stochastics 19, 111–129.

Konstantinov, A. A., Maslov, V. P. and Chebotarev, A. M. (1990), *Probability representations of solutions of the Cauchy problem for quantum mechanical solutions*, Russian Math. Surveys 45:6, 1–26.

Kröger, P. (1992), *On the spectral gap for compact manifolds*, J. Diff. Geom. 36, 315–330.

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.

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.

Lichnerowicz, A., *Geometrie des Groupes des Transformationes*, Dunod, Paris, 1958.

Liggett, T. M.(1985), *Interacting Particle Systems*, Springer-Verlag.

Liggett, T. M. (1989), *Exponential L_2 convergence of attractive reversible nearest particle systems*, Ann. Probab., 17, 403-432.

Liggett, T. M. (1991), *L_2 rates of convergence for attractive reversible nearest particle systems*, Ann. Probab., 19:3, 935-959.

Lindvall, T. (1992), *Lectures on the Coupling Method*, Wiley, New York.

Lindvall, T. and Rogers, L. C. G. (1986), *Coupling of multidimensional diffusion processes*, Ann. of Probab. 14:3, 860–872.

Schoen, R. and Yau, S. T. (1988), *Differential Geometry* (*In Chinese*), Science Press, Beijing, China.

Simon, B. (1976), *A remark on Nelson's best hypercontractive estimate*, Proc. amer. Math. Soc. 55, 376–378.

Stroock, D. W.(1984), *An Introduction to the Theory of Large Deviations*, Springer.

Stroock, D. W. and Zegarlinski, B. (1992 a), The equivalence of the logarithmic Sobolev inequality and the Dobrushin-Shlosman mixing condition, Comm. Math. Phys. 144:2, 303-323.

Stroock, D. W. and Zegarlinski, B. (1992 b), The logarithmic Sobolev inequality for discrete spin systems on a lattice. Comm. Math. Phys. 149:1, 175-193.

Wang, F. Y. (1992a), *Gradient estimates for generalized harmonic functions on manifolds*, preprint.

Wang, F. Y. (1992b), *Ergodicity for infinite dimensional diffusion processes on manifolds*, to appear in Sci. Sin..

Wang, F. Y. (1993a), *Application of coupling method to the Neumann eigenvalue problem*, to appear in Prob. Th. Rel. Fields.

Wang, F. Y. (1993b), *Estimates of the first Dirichlet eigenvalue*, preprint.

Wang, F. Y. (1993c), *Gradient estimates of a class of functions on Riemannian manifolds*, preprint.

Wang, F. Y. (1993d), *Gradient estimates in* $\mathbf{R}^d$, preprint.

Wasserstein, L. N. (1969), *Markov processes on a countable product space, describing large systems of automata* (*In Russian*), Problem Peredachi Informastii 5, 64–73.

Wu, H. H. (editor) (1993), *Contemporary geometry, J. Q. Zhong Memorial Volume*, Plenum Publ. Co., New York.

Yang, H. C. (1989), *Estimate of the first eigenvalue of a compact Riemannian manifold with Ricci curvature bounded below by a negative constant* (*In Chinese*), Sci. Sin.(A) 32:7, 698–700.

Zhong, J. Q. and Yang, H. C. (1984), *Estimates of the first eigenvalue of a compact Riemannian manifolds*, Sci. Sin. 27:12, 1251–1265.

Department of Mathematics, Beijing Normal University, Beijing 100875, The People's Republic of China.

Proceedings of Inter. Conf. on Dirichlet Forms and Stochastic Processes
Editors: Z.M. Ma, M. Röckner and J.A. Yan
Walter de Gruyter Publishers, 1995, 87-102.

On Ergodic Region of Schlögl's Model

*Mu-Fa Chen**

Abstract. One challenging problem in the context of reaction-diffusions is to prove the ergodicity or non-ergodicity for the Schlögl's models. As an application of the recent progress on optimal Markovian couplings, this paper improves considerably the ergodic region of the second Schlögl model. The model is simplified based on an observation between the microscopic reaction-diffusion processes and the corresponding macroscopic reaction-diffusion equations. The resulting bound for ergodicity provides us an explicit picture and it is hopped that this would be helpful both for understanding the models and for the further study.

1991 Mathematics Subject Classification: 58G32, 58G25.

1. Introduction

This paper deals with the reaction-diffusion processes on $S = \mathbf{Z}^d$ with state space $E = \mathbf{Z}_+^S$ and formal generator

$$\begin{aligned}\Omega f(x) = &\sum_{u\in S}\Big\{b(x_u)\big[f(x+e_u)-f(x)\big]+a(x_u)\big[f(x-e_u)-f(x)\big]\Big\}\\ &+\sum_{u,v\in S}x_u p(u,v)\big[f(x-e_u+e_v)-f(x)\big],\\ &x=(x_u:u\in S)\in E,\end{aligned}\tag{1.1}$$

where e_u is the unit vector in $\mathbf{Z}_+^S$ having value 1 at u and 0 elsewhere, $(p(u,v))$ is a translation invariant transition probability on S with $p(u,u)\equiv 0$. Usually, $a_k = a(k)$ and $b_k = b(k)$ are polynomials:

$$b_k=\sum_{j=0}^{m_0}\beta_j k^{(j)},\qquad a_k=\sum_{j=1}^{m_0+1}\delta_j k^{(j)},$$

where m_0 is a fixed integer, $\beta_j, \delta_j \geqslant 0$, $\beta_0, \delta_1, \delta_{m_0+1} > 0$ and $k^{(j)} = k(k-1)\cdots(k-j+1)$. The operator consists of two parts. The second sum in (1.1) describes the diffusion of the system between u and v and the first sum corresponds to the reaction in u. The processes were constructed by Chen (1985) in a more general

setup (see [4]). Most part of the paper is concentrated on two special cases, where the birth rate b_k and the death rate a_k are as follows:

1.1. **Schlögl's first model**

$$b_k = \beta_0 + \beta_1 k, \qquad a_k = \delta_1 k + \delta_2 k(k-1), \qquad \beta_0,\, \beta_1,\, \delta_1,\, \delta_2 > 0.$$

1.2. **Schlögl's second model**

$$b_k = \beta_0 + \beta_2 k(k-1), \qquad a_k = \delta_1 k + \delta_3 k(k-1)(k-2), \qquad \beta_0,\, \beta_2,\, \delta_1,\, \delta_3 > 0.$$

The Schlögl's models (introduced in 1972) are typical ones in non-equilibrium statistical physics. They have received a lot of attentions by many authors in the past two decades. The readers may refer to [4; Part 4] for an exploration of the current status of the study and for more complete references. However, it seems to the author that the story is still quite a distance to be finished. Especially, we do not know at the moment whether the models exhibit phase transitions or not. It is one of the main open problem in this context, an ergodic conjecture is now made in [9; Conjecture 1.3] for a closed related model.

Before moving further, let us mention some related results. First, the finite dimensional processes (i.e., $|S| < \infty$) are always (exponentially) ergodic[11]. It was proved in [2] that the above Schlögl's models are ergodic under the conditions

$$\beta_1 < \delta_1 \tag{1.2}$$

and

$$\delta_1 > \beta_2 + \frac{3}{4}\delta_3 + \frac{\beta_2^2}{3\delta_3}\ (\geqslant 2\beta_2) \tag{1.3}$$

respectively. Then, the results were improved as follows:

$$\text{Neuhauser (1990): All } \beta_k \text{ and } \delta_k\ (k \geqslant 1) \text{ are large enough} \Longrightarrow \text{ergodicity} \tag{1.4}$$

$$\text{Chen (1990): Fixed } \beta_k \text{ and } \delta_k\ (k \geqslant 1),\ \beta_0 \text{ is large enough} \Longrightarrow \text{ergodicity} \tag{1.5}$$

Note that β_0 does not appear in (1.2)–(1.4). In the reversible case, that is,

$$(k+1)b_k/a_{k+1} = \beta_k/\delta_{k+1} \text{ is independent of } k \tag{1.6}$$

([4; Theorem 14.20]), it was proved by Ding, Durrett and Liggett (1991) that the processes are always ergodic (see also [6]). Except in this lower dimension of parameters, the processes are irreversible. From physical point of view, the reversible case is less interesting since the Schlögl's models came from non-equilibrium systems rather the equilibrium ones. In [3], some explicit sufficient conditions for ergodicity were presented for the first Schlögl model but not for the second one since the latter is too complicated to handle.

The study of this paper is based on the following two aspects of new progress: First, the coupling technique is now understood much deeply. Recall that the results (1.2)–(1.6) were all proved by reducing the infinite dimensional case to the one-dimensional one (and so is comparable with a birth-death process), choosing a translation invariant distance on $\mathbf{Z}_+$ and using different couplings. The main

difference of the proofs can be summarized as follows:

Results	Tools
(1.2) and (1.3)	ordinary distance, march coupling
(1.4)	finer distance, march coupling
(1.5)	refined distance, coupling by inner reflection.

See the next section for details about the couplings mentioned here. Recently, it has been proved in [5] that the optimal Markovian coupling for birth-death processes is not the coupling by inner reflection but the coupling by reflection within the class of translation invariant distances on $\mathbf{Z}_+$. Furthermore, the optimality of couplings depend heavily on the distance. These ideas lead us to improve the previous works on the ergodicity of the processes. We adopt either the new coupling (Theorem 2.1) or a new (non-translation invariant) distance (Theorem 5.1). The resulting ergodic region is considerably enlarged. We believe that the idea should also be valuable for many other situations.

The second recent progress is on the relation between the processes and the corresponding macroscopic reaction-diffusion equation:

$$\begin{cases} \dfrac{\partial f}{\partial t} = \dfrac{1}{2}\Delta f + \displaystyle\sum_{j=0}^{m_0} \beta_j f^j - \sum_{j=1}^{m_0+1} \delta_j f^j \\ f(0,r) = g(r), \end{cases} \tag{1.7}$$

where g is a non-negative bounded $C^2(\mathbf{R}^d)$-function with bounded first derivative. Recall that a non-negative, spatially homogeneous solution $f_0(t)$ to Eq.(1.7) is called **asymptotically stable** if there exists a $\delta > 0$ such that for any solution $f(t,r)$ to Eq.(1.7), whenever $|f(0,r)-f_0(0)| < \delta$, we have $\lim_{t\to\infty} |f(t,r)-f_0(t)| = 0$. Next, let $\lambda_1 > \lambda_2 > \cdots > \lambda_k$ $(k \leqslant m_0+1)$ denote the non-negative roots of the equation:

$$\sum_{j=0}^{m_0} \beta_j \lambda^j - \sum_{j=1}^{m_0+1} \delta_j \lambda^j = 0, \tag{1.8}$$

where λ_j has multiplicity m_j $(1 \leqslant j \leqslant k)$. The following result is due to X. J. Xu ([4; Theorem 16.3]).

Theorem 1.3 The solution $f(t,r) \equiv \lambda_i$ to Eq.(1.7) is asymptotically stable iff m_i is odd and $\sum_{j\leqslant i-1} m_j$ is even.

Briefly, the connection of the two subjects is as follows. For every $r \in \mathbf{R}^d$, set $[r/\varepsilon] = ([r_1/\varepsilon], \cdots, [r_d/\varepsilon]) \in \mathbf{Z}^d$. Denote by $(X_u(t))$ for a moment the rescaling process corresponding to the formal generator (1.1) with a factor ε^{-2} in front of the second sum. Let μ^λ be the product of Poisson distribution with parameter λ and denote by $\mathbb{E}^\varepsilon_{\mu^\lambda}$ the expectation of the rescaling process starting from μ^λ. Then,

$$f(t,r) := \lim_{\varepsilon\to 0} \mathbb{E}^\varepsilon_{\mu^\lambda} X_{[r/\varepsilon]}(t) \tag{1.9}$$

satisfies (1.7) with $g(r) \equiv \lambda$ (cf. Boldrighini et al [1] or [4; Chapter 16]). Thus, an asymptotically stable root λ_j means that for every λ, close enough to λ_j, we have $f(t, r) \to \lambda_j$ as $t \to \infty$.

As mentioned in [4; page 520], all known ergodic or non-ergodic results are consistent with Theorem 1.3. For instance, in the reversible case, there is only one non-negative root, which is asymptotically stable by Theorem 1.3. Accordingly, the model has no phase transitions. Next, consider the first Schlögl model with $\beta_0 = 0$. Then, Eq. (1.8) has two roots $\lambda_1 = (\beta_1 - \delta_1)/\delta_2$ and $\lambda_2 = 0$. It is easy to see that λ_1 is asymptotically stable but not λ_2. This conclusion is reasonable since there is a phase transition whenever β_1 is large enough [4; Theorem 15.8]. However, if $\beta_0 > 0$, then there is only one non-negative root and hence asymptotically stable. From this, one may conjecture that there would be no phase transition for the first Schlögl model and there would exist phase transition for the second Schlögl model since for the latter one not every solution being asymptotically stable. Hence, the second model is more interesting. However, in these two different contexts the objects are actually quite different. There is a scaling factor ε^{-2} ($\varepsilon \downarrow 0$) in front of the diffusion rate $x(u)p(u, v)$ in the study of hydrodynamics in order to obtain the Laplacian in the equation. Thus, in order to regard Eq. (1.7) as an approximation of the particle systems, as indicated by (1.9), the diffusion rate should be large. Alternatively, if we fix the diffusion rate to be 1, then the reaction rates a_k and b_k should be replaced by $\varepsilon^2 a_k$ and $\varepsilon^2 b_k$ respectively. From this point of view, (1.4) and (1.5) are also consistent with Theorem 1.3.

Next, note that for the second Schlögl model, the role played by each of the parameters β_k and δ_k is not clear at all. It seems too hard and may not be necessary to consider the whole parameters. Based on the above observation and to keep the physical meaning (see Section 3 for details), we fix $\beta_2 = 6\alpha$ ($\alpha > 0$), $\delta_1 = 9\alpha$ and $\delta_3 = \alpha$. Then, when $\beta_0 \in (0, 4\alpha)$, there are three roots $\lambda_1 > \lambda_2 > \lambda_3 \geqslant 0$. By Theorem 1.3, λ_1 and λ_3 are asymptotically stable but not λ_2. When $\beta_0 = 4\alpha$, we have $\lambda_2 = 1$ with $m_1 = 2$ and $\lambda_1 = 4$, λ_1 is asymptotically stable but not λ_2. As for $\beta_0 > 4\alpha$, there is only one non-negative root which is certainly asymptotically stable. Hence, we guess that the ergodic region should be located in $(4\alpha, \infty)$ for sufficient small α. Of course, the assertion is true in the reversible case, for which, we have $\beta_0 = 36\alpha$. On the other hand, as mentioned in Durrett and Neuhauser[8] that the reaction-diffusion equations are usually the end of the study of hydrodynamical limits of the reaction-diffusion processes. But we (also [8]) are in the opposite direction, i.e., using the reaction-diffusion equation to investigate the microscopic processes. The main point used in [8] to prove some kind of phase transitions for the reaction-diffusion processes with absorbing state $x_u \equiv 0$ is to look for the critical value at which the speed of the traveling wave solution to (1.7) changes its sign. Let us mention, without details, that in our present situation, this critical value is $\beta_0 = 2\alpha$. From this point of view, the phase transitions would be appeared in $(0, 2\alpha)$. Based on these considerations, we propose a typical non-trivial case, for which we have more precise picture as shown below.

Theorem 1.4 Consider the second Schlögl model with $\beta_0 = 2\alpha$, $\beta_2 = 6\alpha$, $\delta_1 = 9\alpha$ and $\delta_3 = \alpha$. Then, the processes are exponentially ergodic for all $\alpha \geqslant 0.7303$.

We remark that (1.3) does not work for the present situation and the ergodic region provided by the previous method[3] is $\alpha \geqslant 31.788$ (Proposition 2.3). When $\alpha = 0$, the reaction-diffusion processes are just the well-known zero-range processes for which there exist many invariant measures and so are non-ergodic. Intuitively, the ergodicity of a reaction-diffusion process is mainly controlled by the reaction part. However, the exponential rate ($\approx 4\alpha$) of the reaction part goes to zero as $\alpha \to 0$. Therefore, there may exist a critical value α_c so that the processes would not be ergodic for all $\alpha < \alpha_c$.

Two main general results of the paper are Theorem 2.1 and Theorem 5.1. The bound given in Theorem 1.4 is obtained with the help of a computer. Certainly, a rough bound can be derived by hand. We prefer the numerical bound not only for showing the power of the method but also for understanding the model. The proof of Theorem 1.4 is given at the end of the paper, based on Theorem 5.1. A weak bound (3.013), based on Theorem 2.1, is presented in Section 4. Roughly speaking, the proof given in Section 4 uses a translation invariant distance but a rather finer coupling and in Section 5, we adopt a particular distance but use a simple coupling. The reason for the specific choice of β_k's and δ_k's is explained in Section 3. In the next section, some necessary preparations are introduced and the main steps of the proof are sketched. Finally, one may jump from here to the last section for a quick glance at the paper.

2. Preliminaries. The First General Result

In this section, we first recall some couplings which will be used throughout the paper. Then, we introduce the general procedure for proving the ergodicity by using the coupling method.

Given two regular birth-death processes with the same birth rate $q_{i,i+1} := b_i$ and death rate $q_{i,i-1} := a_i$, starting from i_1 and i_2 respectively. The **classical coupling** evolves as follows:

$$\begin{aligned} i_1 \neq i_2, \quad (i_1, i_2) &\to (j_1, i_2) \quad \text{at rate} \quad q_{i_1 j_1} \\ &\to (i_1, j_2) \quad \text{at rate} \quad q_{i_2 j_2}. \end{aligned}$$

Otherwise,

$$(i, i) \to (j, j) \text{ at rate } q_{ij}.$$

All couplings considered below will have the property listed in the last line and hence we will not mention again. The **march coupling** evolves as follows: If $i_1 \neq i_2$, then

$$\begin{aligned} (i_1, i_2) &\to (i_1 + k, i_2 + k) \quad \text{at rate} \quad q_{i_1,i_1+k} \wedge q_{i_2,i_2+k} \\ &\to (i_1 + k, i_2) \quad \text{at rate} \quad [q_{i_1,i_1+k} - q_{i_2,i_2+k}]^+ \\ &\to (i_1, i_2 + k) \quad \text{at rate} \quad [q_{i_2,i_2+k} - q_{i_1,i_1+k}]^+, \end{aligned}$$

here we have used the convention that $q_{ij} = 0$ for all $i \in \mathbf{Z}_+$ $j \notin \mathbf{Z}_+$. The key of this coupling is the first line. Whenever we have a term $A \wedge B$, we should also have

the other two terms $(A-B)^+$ and $(B-A)^+$ automatically, due to the marginality for Markovian couplings. Thus, in what follows, we will write down the term $A\wedge B$ only for simplicity. Next, the **coupling by inner reflection** is defined as follows. If $i_2 = i_1 + 1$, then we adopt the classical coupling. If $i_2 \geqslant i_1 + 2$, take

$$(i_1,\, i_2) \to (i_1 + 1,\, i_2 - 1) \text{ at rate } b_{i_1} \wedge a_{i_2}.$$

By exchanging i_1 and i_2, we can get the rates of the coupling for the case that $i_1 \geqslant i_2$. Finally, the **coupling by reflection** evolves in the following way: If $i_2 = i_1 + 1$, then

$$(i_1,\, i_2) \to (i_1 - 1,\, i_2 + 1) \text{ at rate } a_{i_1} \wedge b_{i_2}.$$

If $i_2 \geqslant i_1 + 2$, then

$$\begin{aligned} (i_1,\, i_2) &\to (i_1 - 1,\, i_2 + 1) \quad \text{at rate} \quad a_{i_1} \wedge b_{i_2} \\ &\to (i_1 + 1,\, i_2 - 1) \quad \text{at rate} \quad b_{i_1} \wedge a_{i_2}. \end{aligned}$$

By symmetry, we can write down the rates for the case that $i_1 > i_2$.

It was proved in [5] that the coupling by reflection is ρ-optimal for any translation-invariant distance ρ which has the property that

$$u_k := \rho(k+1, 0) - \rho(k, 0)$$

is decreasing in k. In this and Section 4, we consider this type of distances only.

We now return to the infinite dimensional case. For the diffusion part, throughout this paper, we adopt the march coupling:

$$\begin{aligned} (x,\, y) &\to (x - e_u + e_v,\, y - e_u + e_v) \quad \text{at rate} \quad (x_u \wedge y_u)\, p(u,v) \\ &\to (x - e_u + e_v,\, y) \quad \text{at rate} \quad (x_u - y_u)^+ p(u,v) \\ &\to (x,\, y - e_u + e_v) \quad \text{at rate} \quad (y_u - x_u)^+ p(u,v). \end{aligned}$$

As for the reaction part, each component is a birth-death process, we can simply use one of the couplings listed above. Then, couple the different components independently. Because of the construction, without any confusion, we will use the same names of couplings for the reaction-diffusion processes as that for the birth-death processes. Let $\overline{\Omega}_r$ be the coupling operator of our reaction-diffusion processes by reflection. Given a positive sequence (u_k) with $u_0 = 1$, which will be determined later, define $F(k) = \sum_{j<k} u_j$. Assume that $x \leqslant y$ (i.e., $x_u \leqslant y_u$ for all $u \in S$) and write $k = y_u - x_u \geqslant 0$. Then, we have

$$\begin{aligned} \overline{\Omega}_r F(k) = &\Big\{ - b(x_u)u_{k-1} - a(y_u)u_{k-1} + \big[a(x_u) \wedge b(y_u)\big](u_k + u_{k+1}) \\ &\quad + \big[a(x_u) - b(y_u)\big]^+ u_k + \big[b(y_u) - a(x_u)\big]^+ u_k \Big\} I_{k=1} \\ &- \Big\{ \big[b(x_u) \wedge a(y_u)\big](u_{k-2} + u_{k-1}) + \big[b(x_u) - a(y_u)\big]^+ u_{k-1} \\ &\quad + \big[a(y_u) - b(x_u)\big]^+ u_{k-1} - \big[a(x_u) \wedge b(y_u)\big](u_k + u_{k+1}) \\ &\quad - \big[a(x_u) - b(y_u)\big]^+ u_k - \big[b(y_u) - a(x_u)\big]^+ u_k \Big\} I_{k \geqslant 2} \\ &+ \sum_v (y_v - x_v) p(v,u) u_k - k \sum_v p(u,v) u_{k-1}, \end{aligned}$$

where $I_{k=1}$ is the indicator of the set $\big\{(x,y):\ y_u-x_u=1\big\}$. Collecting terms, we obtain

$$\begin{aligned}\overline{\Omega}_r F(k)=&\Big\{\big[a(x_u)\wedge b(y_u)\big]u_{k+1}+\big[a(x_u)\vee b(y_u)\big]u_k-\big[b(x_u)+a(y_u)+k\big]u_{k-1}\Big\}I_{k=1}\\&+\Big\{\big[a(x_u)\wedge b(y_u)\big]u_{k+1}+\big[a(x_u)\vee b(y_u)\big]u_k\\&\quad-\big[b(x_u)\vee a(y_u)+k\big]u_{k-1}-\big[b(x_u)\wedge a(y_u)\big]u_{k-2}\Big\}I_{k\geqslant 2}\\&+\sum_v(y_v-x_v)p(v,u)u_k.\end{aligned}\tag{2.1}$$

What we need to do is to find out an $\varepsilon>0$ and a positive decreasing sequence (u_k) with $u_0=1$ so that

$$\begin{aligned}\overline{\Omega}_r F(k)&\leqslant -\varepsilon F(k)-k+\sum_v(y_v-x_v)p(v,u)u_k\\&\leqslant -\varepsilon F(k)-k+\sum_v(y_v-x_v)p(v,u).\end{aligned}\tag{2.2}$$

Here, in the last step, we have used the fact that $u_k\leqslant u_0=1$. This is the main place we have to lost a bit, but it enables us to reduce the infinite dimensional case to the one-dimensional one. Actually, by using the translation-invariance and the order-preserving property of the coupling, it follows from (2.2) that for all translation invariant x (i.e., $x_u\equiv$some $m\in\mathbf{Z}_+$) and y with $x\leqslant y$,

$$\overline{\mathbb{E}}^{(x,y)}F(Y_u(t)-X_u(t))\leqslant\overline{\mathbb{E}}^{(x,y)}F\big(Y_u(1)-X_u(1)\big)e^{-\varepsilon(t-1)},\quad t\geqslant 1,\ u\in S.\tag{2.3}$$

where $X(t)=(X_u(t):u\in S)$ and $Y(t)=(Y_u(t):u\in S)$ are the processes starting from x and y respectively. Now, (2.3) plus the monotonicity, the translation-invariance and the finiteness of the moments of the process gives us the ergodicity. Refer to [4; Chapter 14] for details.

Set $u_{-1}=1$ and define

$$\begin{aligned}r(i,k)=&\big[b_i\vee a_{i+k}+k\big]u_{k-1}+\big[b_i\wedge a_{i+k}\big]u_{k-2}\\&-\big[a_i\vee b_{i+k}\big]u_k-\big[a_i\wedge b_{i+k}\big]u_{k+1},\qquad i\geqslant 0,\ k\geqslant 1.\end{aligned}$$

Combining (2.1) with (2.2), what we need is the following condition.

$$\inf_{i\geqslant 0}\frac{r(i,k)-k}{F(k)}\geqslant\varepsilon,\qquad k\geqslant 1.\tag{2.4}$$

To simplify the condition (2.4), we introduce the differential operator $\Delta_k f(i)=f(i+k)-f(i)$ and set $\Delta=\Delta_1$. Then, some elementary computations give us

$$\begin{aligned}r(i,k)=&\big[\Delta_k a(i)-\Delta_k b(i)+k\big]u_{k-1}+\big[a_i+b_{i+k}\big](u_{k-1}-u_k)\\&+\big[b_i\wedge a_{i+k}\big](u_{k-2}-u_{k-1})+\big[a_i\wedge b_{i+k}\big](u_k-u_{k+1}),\ \ i\geqslant 0,\ k\geqslant 1.\end{aligned}\tag{2.5}$$

In particular,

$$r(0,k)=\big[b_0+a_k+k\big]u_{k-1}+\big[b_0\wedge a_k\big](u_{k-2}-u_{k-1})-b_k u_k,\quad k\geqslant 1.\tag{2.6}$$

Since the degree of a_k is higher than that of b_k, for each $k \geqslant 1$, there exists the minimal integer $\underline{i}_k$, independent of the sequence (u_k), so that $\Delta_k a(i) - \Delta_k b(i)$ is increasing in i for all $i \geqslant \underline{i}_k$. Thus, there exists an integer $i_k^* \leqslant \underline{i}_k$, depending on (u_k), so that

$$r(i_k^*, k) = \min_{0 \leqslant i \leqslant \underline{i}_k} r(i,k) = \inf_{i \geqslant 0} r(i,k), \qquad k \geqslant 1.$$

Similarly, there exists uniquely a $\underline{k}$ (independent of (u_k)) so that for each $k \geqslant \underline{k}$, $\Delta_k a(i) - \Delta_k b(i)$ is increasing in $i(\geqslant 0)$. Furthermore, there is uniquely a k^* (depending on (u_k)) $\leqslant \underline{k}$ so that $r(i,k)$ is increasing in $i(\geqslant 0)$ for all $k \geqslant k^*$. Hence, the condition (2.4) can be rewritten as follows.

$$\begin{cases} \dfrac{r(i_k^*, k) - k}{F(k)} \geqslant \varepsilon, & 1 \leqslant k \leqslant k^* - 1, \\ \dfrac{r(0,k) - k}{F(k)} \geqslant \varepsilon, & k \geqslant k^*. \end{cases} \tag{2.7}$$

The above consideration leads to the following construction of (u_k). Fix $\varepsilon > 0$. Let $1 = u_{-1} = u_0 \geqslant \cdots \geqslant u_{k^*} > 0$ (depending on ε) be a solution to the inequality

$$\begin{cases} \dfrac{r(i_k^*, k) - k}{F(k)} \geqslant \varepsilon, & 1 \leqslant k \leqslant k^* - 1, \\ \dfrac{r(0,k^*) - k^*}{F(k^*)} \geqslant \varepsilon. \end{cases} \tag{2.8}$$

Then, take

$$u_k := u_k(\varepsilon) = u_{k-1} \wedge \frac{\big[b_0 \vee a_k + k\big] u_{k-1} + \big[b_0 \wedge a_k\big] u_{k-2} - k - \varepsilon F(k)}{b_k} \vee 0, \qquad k \geqslant k^* + 1. \tag{2.9}$$

Now, we can summarize the above discussions as follows.

Theorem 2.1 The reaction-diffusion processes are ergodic if for some $\varepsilon > 0$, $u_k > 0$ for all k.

We have seen that it is not trivial at all to figure out the sequence (u_k) (unlike the sequence $(\tilde{u}_k)$ given below) since $i_k^*\,(1 \leqslant k \leqslant k^* - 1)$ and k^* all depend on (u_k). Before moving further, let us recall the u-criterion presented in [3]. Define

$$\begin{aligned} &\tilde{u}_0 = 1, \qquad \tilde{u}_1 = \inf_{i \geqslant 0} \frac{b_i + a_{i+1} - \varepsilon}{a_i + b_{i+1}} \vee 0, \\ &\tilde{u}_k = \left\{ \inf_{i \geqslant 0} \frac{\big[b_i \vee a_{i+k} + k\big] \tilde{u}_{k-1} + \big[b_i \wedge a_{i+k}\big] \tilde{u}_{k-2} - k - \varepsilon \sum_{j=0}^{k-1} \tilde{u}_j}{a_i + b_{i+k}} \right\} \vee 0, \qquad k \geqslant 2. \end{aligned} \tag{2.10}$$

Then, one of main results in [3] says that for fixed β_k and δ_k $(k \geqslant 1)$, we have $\tilde{u}_k > 0$ for all k whenever β_0 is big and ε is small. The conclusion also holds in

the case of Theorem 1.4 for large enough α. The sequence $(\tilde{u}_k)$ comes from

$$\begin{aligned}
&\Delta a(i) - \Delta b(i) + \big[a_i + b_{i+1}\big](1-\tilde{u}_1) \geqslant \varepsilon, \quad i \geqslant 0,\\
&\Big\{\big[\Delta_k a(i) - \Delta_k b(i) + k\big]\tilde{u}_{k-1} + \big[b_i \wedge a_{i+k}\big](\tilde{u}_{k-2} - \tilde{u}_{k-1})\\
&\quad + \big[a_i + b_{i+k}\big](\tilde{u}_{k-1} - \tilde{u}_k) - k\Big\}\big/\big(1 + \tilde{u}_1 + \cdots + \tilde{u}_{k-1}\big) \geqslant \varepsilon,\\
&\qquad\qquad i \geqslant 0,\ k \geqslant 2, \qquad\qquad (2.11)
\end{aligned}$$

which is an analog of (2.4), but replacing the coupling by reflection with the coupling by inner reflection. From this, it is easy to check that sequence $(u_k = \tilde{u}_k)$ defined by (2.10) should satisfy (2.11) and hence (2.8). We have thus proved that a positive sequence (u_k) does exist whenever α is large enough. Moreover, Theorem 2.1 improves the previous u-criterion.

Next, let us consider the first Schlögl model. Because

$$\Delta_k a(i) - \Delta_k b(i) = (\delta_1 - \beta_1)k + \delta_2 k[k + 2ki - 1], \qquad k \geqslant 1$$

is increasing for all $i \geqslant 0$, the conditions (2.4) and (2.11) are actually the same:

$$\frac{r(0,k) - k}{F(k)} \geqslant \varepsilon, \qquad k \geqslant 1. \qquad (2.12)$$

We have thus obtained the following conclusion.

Proposition 2.2 For the first Schlögl model, the condition (2.4) is reduced to (2.12). In other words, the coupling by reflection and the coupling by inner reflection produce the same condition (2.12) for the ergodicity.

Finally, by applying Theorem 2.1 to the sequence $(\tilde{u}_k)$ defined by (2.10) with $\varepsilon \leqslant 10^{-5}$, we obtain the following result.

Proposition 2.3 Under the assumption of Theorem 1.4, the second Schlögl model is ergodic for all $\alpha \geqslant 31.788$.

Comparing Proposition 2.3 with Theorem 1.4, we see that the coupling by reflection does improve the ergodic region provided by the coupling by inner reflection for the second Schlögl model. The reason is that $i_k^* \neq 0$ $(k < k^*)$ for the second model. However, these two couplings coincide each other starting from k^* (but not 1 as that for the first model). Furthermore, when k is bigger than some k_1 $(\geqslant k^*)$, these two couplings also coincide with the march coupling (The reason will be explained at the end of Section 4). The last two conclusions came with no surprise since the optimal couplings depend heavily on the metric ρ and the optimal choice of ρ is determined by the rates of the processes, we will return to this point in the last section.

3. A Simplification of the Second Schlögl Model

In view of (2.8) and (2.9), it is too complicated in general to find out the sequence (u_k) for the second model to get some sharp estimates for the ergodicity. On the other hand, it seems not necessary to consider the whole four parameters β_0, β_2, δ_1 and δ_3. In this section, we show how to simplify the model with the help of the corresponding reaction-diffusion equation. The main point is that, to keep the essential meaning of the model, we should choose the parameters so that the equation

$$\beta_0 + \beta_2\lambda^2 - \delta_1\lambda - \delta_3\lambda^3 = 0 \tag{3.1}$$

contains a non-asymptotically stable root. Of course, we can take $\alpha = \delta_3 = 1$. Let $\lambda = x + \beta_2/3$. Then, (3.1) is reduced to

$$x^3 + px^2 + q = 0, \tag{3.2}$$

where

$$p = \delta_1 - \frac{1}{3}\beta_2^2, \qquad q = -\beta_0 + \frac{1}{3}\beta_2\delta_1 - \frac{2}{27}\beta_2^3. \tag{3.3}$$

When $q^2/4 + p^3/27 > 0$, there is only one real root, which is necessarily positive and asymptotically stable. Hence, the only interesting case is that $q^2/4 + p^3/27 \leqslant 0$. Solving the equation $q^2/4 + p^3/27 = 0$ in variable β_0, we obtain

$$\beta_0^{(1)} = \frac{-\beta_2(2\beta_2^2 - 9\delta_1) - 2(\beta_2^2 - 3\delta_1)^{3/2}}{27}, \quad \beta_0^{(2)} = \frac{-\beta_2(2\beta_2^2 - 9\delta_1) + 2(\beta_2^2 - 3\delta_1)^{3/2}}{27}.$$

It turns out that $q^2/4 + p^3/27 \leqslant 0$ iff $\beta_2^2 \geqslant 3\delta_1$ and $\beta_0^{(1)} \leqslant \beta_0 \leqslant \beta_0^{(2)}$. This rules out the region provided by (1.3). Recall that for the model, β_0 varies from 0 to ∞. So, it is natural to take $\beta_0^{(1)} = 0$. That is,

$$\left(\frac{\beta_2^2}{\delta_1} - 3\right)^{3/2} = \frac{\beta_2}{\sqrt{\delta_1}}\left(\frac{9}{2} - \frac{\beta_2^2}{\delta_1}\right).$$

The only solution to this equation is

$$\beta_2 = 2\sqrt{\delta_1}. \tag{3.4}$$

Then,

$$\beta_0^{(2)} = \frac{2}{27}\delta_1^{3/2}\left[\left(\frac{\beta_2^2}{\delta_1} - 3\right)^{3/2} - \frac{\beta_2}{\sqrt{\delta_1}}\left(\frac{\beta_2^2}{\delta_1} - \frac{9}{2}\right)\right] = \frac{4}{27}{\delta_1}^{3/2}.$$

Therefore, for all $0 < \beta_0 < \frac{4}{27}\delta_1^{3/2}$, we have three non-negative roots:

$$\lambda = \frac{2}{3}\sqrt{\delta_1}\left[1 + \cos\left(\frac{\varphi + 2k\pi}{3}\right)\right], \qquad k = 0, 1, 2 \tag{3.5}$$

where

$$\cos\varphi = \frac{27}{2}\beta_0\delta_1^{-3/2} - 1.$$

Thus, the number of the parameters is reduced from 4 to 2. Our specific choice that $\delta_1 = 9$ is not essential but for simplicity to make β_2 being an integer and δ_1 being different from δ_3.

Now, fix $\delta_3 = 1$, $\delta_1 = 9$ and $\beta_2 = 6$. Then, $q^2/4 + p^3/27 > 0$ iff $\beta_0 > 4$. If so, there is only one non-negative root. Next, $q^2/4 + p^3/27 < 0$ iff $\beta_0 < 4$. In that case, we have three non-negative roots given by (3.5). Finally, when $\beta_0 = 4$, we have $\lambda_1 = 2$ with multiplicity 2 and a single root $\lambda_1 = 4$. Thus, as mentioned before Theorem 1.4, for every $\beta_0 \in (0, 4]$ there is precise one non-asymptotically stable root but there is no such root for all $\beta_0 \in (4, \infty)$. We have thus arrived at the desired position. In our particular situation ($\beta_0 = 2$), the three roots are $2 - \sqrt{3},\ 2,\ 2 + \sqrt{3}$.

4. A Bound Provided by Theorem 2.1

In this section, we prove that under the hypotheses of Theorem 1.4, the processes are ergodic for all $\alpha \geqslant 3.013$. Having the optimal coupling for a large class of distances in mind, as we discussed in Section 2, the next step is to figure out a suitable distance (i.e., a sequence (u_k)). Which is the main goal of this section.

Recall that $\beta_0 = 2\alpha$, $\beta_2 = 6\alpha$, $\delta_1 = 9\alpha$ and $\delta_3 = \alpha$. Then

$$b_k = 2\alpha(1 + 3k(k-1)), \qquad a_k = \alpha k(9 + (k-1)(k-2)).$$

Hence

$$\Delta_k a(i) - \Delta_k b(i) = \alpha k(17 - 18i + 3i^2 - 9k + 3ik + k^2)$$

and furthermore

$$\Delta(\Delta_k a - \Delta_k b)(i) = 3\alpha k(2\,i - 5 + k).$$

It follows that $\underline{k} = 5$, $\underline{i}_1 = \underline{i}_2 = 2$ and $\underline{i}_3 = \underline{i}_4 = 1$ ($\underline{k}$ and $\underline{i}$'s are defined below (2.6)). Next, since

$$a_i - b_{i+k} \leqslant a_i - b(i+1) = 5i - 9i^2 + i^3 - \beta_0 \leqslant 0, \qquad i \leqslant 2,\ k \geqslant 1$$

and

$$b_i - a_{i+k} \leqslant b_i - a_{i+2} = -18 - 17i + 3i^2 - i^3 + \beta_0 \leqslant -18 + \beta_0, \qquad i \leqslant 2,\ k \geqslant 2.$$

We have

$$\begin{aligned} a_i \wedge b_{i+k} &= a_i, \qquad i \leqslant 2,\ k \geqslant 1. \\ b_i \wedge a_{i+k} &= b_i, \qquad i \leqslant 2,\ k \geqslant 2,\ \ \beta_0 \leqslant 18. \end{aligned}$$

Therefore,

$$\begin{aligned} r(i,1) &= \Delta a(i) - \Delta b(i) + 1 + [a_i + b_{i+1}](1 - u_1) + a_i(u_1 - u_2). \\ r(i,k) &= [\Delta_k a_i - \Delta_k b(i) + k]u_{k-1} + [a_i + b_{i+k}](u_{k-1} - u_k) \\ &\quad + b_i(u_{k-2} - u_{k-1}) + a_i(u_k - u_{k+1}), \qquad i \leqslant 2,\ k \geqslant 2,\ \ \beta_0 \leqslant 18. \end{aligned}$$

Now, we are going to choose a positive decreasing sequence (u_k) for small $\varepsilon > 0$ so that the following quantities are all non-negative:

$$\begin{aligned}
r(0,1) - r(2,1) &= 18\alpha(-2 + 2u_1 + u_2)\\
r(1,1) - r(2,1) &= 3\alpha(-9 + 8u_1 + 3u_2)\\
r(0,2) - r(1,2) &= 3\alpha(-5u_1 + 8u_2 + 3u_3)\\
r(2,2) - r(1,2) &= 3\alpha(4 + 9u_1 - 12u_2 - 3u_3)\\
r(0,3) - r(1,3) &= 9\alpha(-3u_2 + 4u_3 + u_4)\\
r(1,4) - r(0,4) &= 3\alpha(15u_3 - 16u_4 - 3u_5).
\end{aligned}$$

In other words, $i_1^* = 2$, $i_2^* = i_3^* = 1$, $i_4^* = 0$ and so $k^* = 4$. To do so, solve the linear equations

$$\begin{cases}
r(2,1) - 1 = \varepsilon F(1)\\
r(1,2) - 2 = \varepsilon F(2)\\
r(1,3) - 3 = \varepsilon F(3)\\
r(0,4) - 4 = \varepsilon F(4).
\end{cases}$$

The solution provided by Mathematica is as follows.

$$\begin{aligned}
u_1 &= \big[3366 - 3645\varepsilon + 765\varepsilon^2 + 425312\alpha - 213893\alpha\varepsilon + 12398506\alpha^2\big]/w,\\
u_2 &= \big[72 - 126\varepsilon + 63\varepsilon^2 - 9\varepsilon^3 + 16604\alpha - 22854\alpha\varepsilon + 6493\varepsilon^2\alpha + 821272\alpha^2\\
&\quad - 673474\alpha^2\varepsilon + 9725428\alpha^3\big]/(\alpha w),\\
u_3 &= \big[444 - 796\varepsilon + 417\varepsilon^2 - 65\varepsilon^3 + 37038\alpha - 52527\alpha\varepsilon + 14562\alpha\varepsilon^2 + 928412\alpha^2\\
&\quad - 876731\alpha^2\varepsilon + 7453594\alpha^3\big]/(\alpha w),\\
u_4 &= \big[24 - 50\varepsilon + 35\varepsilon^2 - 10\varepsilon^3 + \varepsilon^4 + 2364\alpha - 4276\alpha\varepsilon + 2208\alpha\varepsilon^2 - 350\alpha\varepsilon^3 + 80664\alpha^2\\
&\quad - 124934\alpha^2\varepsilon + 37691\alpha^2\varepsilon^2 + 1177860\alpha^3 - 1314558\alpha^3\varepsilon + 6306304\alpha^4\big]/(\alpha^2 w),
\end{aligned}$$

where

$$w = 2\left(648 - 486\varepsilon + 81\varepsilon^2 + 182934\alpha - 69165\alpha\varepsilon + 6874478\alpha^2\right).$$

Then, define u_k for all $k \geqslant 5$ by (2.9).

Take $\varepsilon \leqslant 10^{-5}$. Then for all $\alpha \geqslant 3.013$, we do have a required solution (u_k) (At this step, we use both True Basic and Mathematica).

To complete the proof of the assertion, we should point out a technical point. Suppose that we have already had $u_{k_1} = u_{k_1+1} = \underline{u}$ for some $k_1 \geqslant k^*$. Then, in order for $u_{k_1+2} = \underline{u}$, by (2.9), it suffices that

$$\begin{aligned}
&\frac{\big[b_0 \vee a_{k_1+2} + k_1 + 2\big]\underline{u} + \big[b_0 \wedge a_{k_1+2}\big]\underline{u} - (k_1+2) - \varepsilon F(k_1+2)}{b_{k_1+2}}\\
&= \frac{\big[b_0 + a_{k_1+2} + k_1 + 2\big]\underline{u} - (k_1+2) - \varepsilon F(k_1+2)}{b_{k_1+2}}\\
&\geqslant \underline{u}.
\end{aligned}$$

Note that $F(k) \leqslant k$. It is enough that

$$\underline{u} \geqslant \frac{1}{(k_1-1)(k_1-4)}. \tag{4.1}$$

Thus, by induction, if (4.1) holds, we indeed have $u_k \equiv \underline{u}$ for all $k \geqslant k_1$. In other words, starting from k_1, the coupling by reflection and the coupling by inner reflection all coincide with the march coupling. Hence, the computation of (u_k) can be stopped at k_1+1 whenever (4.1) holds. In our particular case, $k_1 = 13$.

5. The Second General Result and the Proof of Theorem 1.4

In contrast to the distance $\rho(k,\ell) = \sum_{j<|k-\ell|} u_j$ used above, we consider in this section the following distance

$$\rho(k,\ell) = \left|\sum_{j<k} u_j - \sum_{j<\ell} u_j\right|, \qquad k,\, \ell \in \mathbf{Z}_+,$$

where (u_k) is a positive sequence on $\mathbf{Z}_+$. The restriction to this sort of distances is due to the fact that for the special birth-death processes contained in the reaction part, the exponentially convergent rate can be estimated by the coupling argument sharply in terms of this kind of distances. Of course, ρ is non-translation invariant unless $u_k \equiv$constant.

Theorem 5.1 Let (u_k) be a positive sequence on $\mathbf{Z}_+$ with $u_0 = 1$ and $\bar{u} := \sup_k u_k < \infty$. Set $u^* = \sup_{j\geqslant i}(u_j - u_i)$. Suppose that there exists an $\varepsilon > 0$ such that

$$b_{k+1}u_{k+1} - (b_k + a_{k+1} + k + 1 - \varepsilon)u_k + (a_k + k)u_{k-1} + \bar{u} + ku^* \leqslant 0, \quad k \geqslant 0, \tag{5.1}$$

where $a_0 = 0$ and $u_{-1} = 1$. Then the reaction-diffusion processes are ergodic.

Proof. It was proved in [5] that for birth-death processes, the four couplings mentioned in Section 2 are all ρ-optimal. Thus, we now adopt the simplest classical coupling. Denote by $\widetilde{\Omega}_c$ the coupling operator of the reaction-diffusion processes. Fix $x \leqslant y$ and $u \in S$, write $x_u = i \leqslant j = y_u$. We have

$$\begin{aligned}
\widetilde{\Omega}_c\rho(i,j) &= \big\{ -b_iu_i + a_iu_{i-1} + b_ju_j - a_ju_{j-1}\big\}I_{j-i\geqslant 1} - (j-i)u_{j-1} \\
&\quad - i(u_{j-1} - u_{i-1}) + \sum_v (y_v - x_v)p(v,u)u_j + \sum_v x_v p(v,u)(u_j - u_i) \\
&= \big\{b_ju_j - b_iu_i - (a_j + j)u_{j-1} + (a_i + i)u_{i-1}\big\}I_{j-i\geqslant 1} \\
&\quad + \sum_v (y_v - x_v)p(v,u)u_j + \sum_v x_v p(v,u)(u_j - u_i).
\end{aligned} \tag{5.2}$$

The last term on the right-hand side appears since ρ is not translation invariant. Now, by (5.1), we have

$$\begin{aligned}
&\big\{b_j u_j - b_i u_i - (a_j + j)u_{j-1} + (a_i + i)u_{i-1}\big\} I_{j-i\geqslant 1} \\
&\quad = \sum_{\ell=i}^{j-1} \big\{ \big(b_{\ell+1}u_{\ell+1} - b_\ell u_\ell\big) - \big[(a_{\ell+1} + \ell + 1)u_\ell - (a_\ell + \ell)u_{\ell-1}\big]\big\} \\
&\quad \leqslant -\varepsilon \sum_{\ell=i}^{j-1} u_\ell - (j-i)\bar{u} - (j-i)iu^* \\
&\quad \leqslant -\varepsilon\rho(i,j) - (j-i)\bar{u} - iu^*. \qquad (5.3)
\end{aligned}$$

On the other hand, by the order-preserving of the coupling and the translation invariance of the processes, for every translation invariant x and y with $x \leqslant y$, we have

$$\begin{aligned}
&\sum_v \mathbb{E}^{x,y}\Big[\big(Y_v(t) - X_v(t)\big)p(v,u)u_{Y_u(t)}\Big] + \sum_v \mathbb{E}^{x,y}\Big[X_v(t)p(v,u)\big(u_{Y_u(t)} - u_{X_u(t)}\big)\Big] \\
&\quad \leqslant \bar{u}\mathbb{E}^{x,y}\big(Y_u(t) - X_u(t)\big) + u^*\mathbb{E}^{x,y}X_u(t). \qquad (5.4)
\end{aligned}$$

Combining (5.2), (5.3) with (5.4), we arrive at

$$\mathbb{E}^{x,y}\widetilde{\Omega}_c\rho\big(X_u(t), Y_u(t)\big) \leqslant -\varepsilon\mathbb{E}^{x,y}\rho\big(X_u(t), Y_u(t)\big), \qquad t \geqslant 0.$$

The remainder of the proof is exactly the same as we did before (cf. Section 2). □

Proof of Theorem 1.4. Take $\varepsilon \leqslant 10^{-5}$, $u_0 = 1$, $u_1 = u_2 = 3/2 + \varepsilon$ (tricky!),

$$u_{k+1} = \frac{(a_{k+1}+b_k+k+1-\varepsilon)u_k - (a_k + k)u_{k-1} - (k+1)u_1 + k}{b_{k+1}}, \quad k \geqslant 2. \qquad (5.5)$$

Define $k_1 = \inf\{k \geqslant 2 : u_{k+1} \geqslant u_k\}$. When $\alpha = 0.7303$, some numerical computation gives us $k_1 = 15$ (k_1 can be smaller if α is bigger). Due to the same reason as explained at the end of the last section, the computation can be stopped here.

Next, since the sequence (u_k) satisfies (5.1), the conclusion of Theorem 1.4 follows from Theorem 5.1 with $\bar{u} = u_1$ and $u^* = \bar{u} - 1$. □

One may think that Theorem 5.1 does not improve too much the bound. But this is an incorrect impression. When we fix $\alpha = 1$ but leave β_0 to be freedom, the ergodic regions provided by the previous method and Theorem 2.1 are $[8.37, \infty)$ and $[6.062, \infty)$ respectively. However, Theorem 5.1 may works for all $\beta_0 > 0$, at least it works well with respect to the same (u_k) given by (5.5) when $\beta_0 \geqslant 10^{-6}$.

Finally, one may improve further the bound 0.7303 by using the coupling by reflection and a refined distance of the form $f \circ \rho$ for some f with $f(0) = 0$, $f' > 0$ and $f'' \leqslant 0$. However, if you write down the first three terms from (5.5),

$$\begin{aligned}
u_3 &= \frac{87}{76} - \frac{1}{38\alpha}, \qquad u_4 = \frac{-4 - 209\alpha + 2382\alpha^2}{2812\alpha^2}, \\
u_5 &= \frac{-20 - 1465\alpha - 43779\alpha^2 + 233238\alpha^3}{343064\alpha^3}, \qquad \text{(setting } \varepsilon = 0\text{)},
\end{aligned}$$

we see that in order for u_3, u_4 or $u_5 > 0$, one requires $\alpha > 0.023,\ 0.1039$ or 0.218 respectively. Comparing these with our bound 0.7303, it follows that there is only a small room left.

Acknowledgement. The first version of the paper was done while the author visited Dept. of Math., Univ. of Roma "La Sapienza" during May–June, 1993. The author would like to acknowledge the warm hospitality and the financial support of Dept., especially Professor E. Scacciatelli and his colleagues and moreover for their valuable discussions. Thanks are also given to the organizers for their effort for the conference.

References

[1] Boldrighini, C., DeMasi, A., Pellegrinotti, A. and Presutti, E. (1987), *Collective phenomena in interacting particle systems.* Stoch. Proc. Appl. **25**, 137-152.

[2] Chen, M. F. (1986), *Jump Processes and Particle Systems* (In Chinese). Beijing Normal Univ. Press.

[3] Chen, M. F. (1990), *Ergodic theorems for reaction-diffusion processes.* J. Statis. Phys. **58**:5/6, 939-966.

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

[5] Chen, M. F. (1993), *Optimal Markovian couplings and applications.* Technical Report, Carleton University, No: 216 and C. V. Volterra, Univ. Roma II, No: 147, Acta Math. Sin.New Ser. **10**:3(1994).

[6] Chen, M. F., Ding, W. D. and Zhu, D. J. (1994), *Ergodicity of reversible reaction-diffusion processes with general reaction rates.* Acta Math. Sin. New Ser. **10**:1.

[7] Ding, W. D., Durrett, R. and Liggett, T. M. (1990), *Ergodicity of reversible reaction diffusion processes.* Prob. Th. Rel. Fields. **85** (1), 1-26.

[8] Durrett, R. and Neuhauser, C. (1992), *Particle systems and reaction-diffusion equations.* Technical Report 92-1, Math. Sci. Inst., Cornell Univ.

[9] Durrett, R. and Levin, S. (1993), *The importance of being discrete* (*and spatial*). preprint.

[10] Neuhauser, C. (1990), *An ergodic theorem for Schlögl models with small migration.* Prob. Th. Rel. Fields, **85**:1, 27–32.

[11] Yan, S. J. and Chen, M. F. (1986), *Multidimensional Q-processes.* Chin. Ann. Math. **7(B)**:1, 90–110.

J. Funct. Anal. 131:2 (1995), 345-363.

ESTIMATION OF THE FIRST EIGENVALUE OF SECOND ORDER ELLIPTIC OPERATORS

Mu-Fa Chen and Feng-Yu Wang

(Beijing Normal University)
Received October 3, 1993

Abstract. This note studies the first non-trivial eigenvalue of second order self-adjoint elliptic operators in $\mathbf{R}^d$. Some lower bounds of the eigenvalue are obtained by using a probabilistic approach and some geometric consideration. In one-dimensional case, an analytic proof is also presented. The resulting bounds can be sharp.

1. Main Results and Examples

Consider the operator in $\mathbf{R}^d$:

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

where $a(x) := (a_{ij}(x))$ is positive definite, $a_{ij} \in C^2(\mathbf{R}^d)$ and

$$b_i(x) = \sum_j a_{ij}(x)\frac{\partial}{\partial x_j}V(x) + \sum_j \frac{\partial}{\partial x_j}a_{ij}(x)$$

for some $V \in C^2(\mathbf{R}^d)$ with $Z := \int \exp V(x)dx < \infty$. The specific form of b_i implies that L is symmetric with respect to a finite measure. To see this, one may express L as

$$\Delta_g + \nabla_g\left(V + \frac{1}{2}\log\det a\right)$$

in terms of the Laplace-Beltrami operator Δ_g and gradient ∇_g with respect to the Riemannian metric $(g_{ij}(x)) = (a_{ij}(x))^{-1}$. Obviously, the operator L has a trivial eigenvalue $\lambda_0 = 0$. We are interested in the next one λ_1, which is also called the

2000 *Mathematics Subject Classification.* 35P15, 60H30.
Key words and phrases. Optimal coupling, elliptic operator, spectral gap.
Research supported in part by NSFC and the State Education Commission of China.

spectral gap of L. More precisely, let $\pi(dx) = Z^{-1}\exp V(x)dx$ and denote by $\mathscr{D}(L)$ the domain of L in $L^2(\pi)$. Then,

$$\text{gap}\,(L) = \inf\big\{-(f, Lf) :\ f \in \mathscr{D}(L),\ \pi f = 0 \text{ and } \|f\| = 1\big\},$$

where $\pi f = \int f d\pi$ and $\|\cdot\|$ is the L^2-norm in $L^2(\pi)$. The importance of the spectral gap is that it describes the exponential L^2-convergence:

$$\|P(t)f - \pi f\| \leqslant \|f - \pi f\| \exp[-\varepsilon t]$$

for all $t \geqslant 0$ and $f \in L^2(\pi)$, where $\{P(t)\}_{t\geqslant 0}$ is the semigroup determined by L. Actually, it can be proved that $\varepsilon_{\max} = \text{gap}\,(L)$ even for general reversible Markov processes (cf.[16] and [3] or [4; Section 9.1]). Moreover, the L^2-convergence is now used as a tool to describe the phase transitions. The readers are urged to refer to [6] for more details about the background of the study. The spectrum theory is a classical topic in analysis, there is a large number of publications, see for instance the books [2], [20], [21], [26] and references therein. Most of them deal with Dirichlet eigenvalues with (compact) regular domain and the asymptotic behavior of the distributions of the eigenvalues. It is a pity that we are unable to find out from the literature some general results on the estimation of the spectral gap, except some particular examples, which will be mentioned later. The proof of Theorem 1.3 given in Section 3 is somewhat related to the traditional variational method, but as far as we know, the technique is still new in the context of diffusions.

The main tool to study the spectral gap in this paper is the coupling approach. The coupling theory is now an active research subject. Here, we mention within the context of diffusions only a few of the references: [7]–[10], [14], [15], [17]–[19] and [25], from which one can find out some constructions of couplings as well as various applications. Recently, the coupling method has also been used in [8] to study the first eigenvalue of Laplacian on manifolds. Actually, it was illustrated there that the method works for general Markov processes, not necessarily diffusions. For the reader's convenience, we recall briefly the main idea from [8] for the present context. Let $(X_t)_{t\geqslant 0}$ be a reversible diffusion in a regular domain $\mathscr{G} \subset \mathbf{R}^d$ with weak generator $\overline{L}$. Denote by $\mathbb{E}^{(x,y)}$ the expectation of a Markovian coupling diffusion (X_t, Y_t), starting from $(x, y) \in \mathscr{G}^2$. Now, our preliminary result can be summarized as follows.

Theorem 1.0. Let g be the eigenfunction of $\overline{L}$ corresponding to λ_1. Suppose that g is in the weak domain of $\overline{L}$ in the sense:

$$\frac{d}{dt}\mathbb{E}^x g(X_t) = \mathbb{E}^x \overline{L} g(X_t)$$

for all $t \geqslant 0$ and $x \in \mathscr{G}$.

(1) If $\sup_{x\neq y} |g(x) - g(y)| < \infty$, then $\lambda_1 \geqslant 1/\sup_{x,y} \mathbb{E}^{(x,y)} T$, where $T := \inf\{t \geqslant 0 : X_t = Y_t\}$.

(2) If g is Lipschitz with respect to a distance ρ in $\mathscr{G}$ and there exist $\varepsilon > 0$ and $f \in C^1(\mathbf{R}_+)$ with $\inf_{r>0} f(r)/r > 0$ such that

$$\mathbb{E}^{(x,y)} f \circ \rho(X_t, Y_t) \leqslant f \circ \rho(x, y) e^{-\varepsilon t}, \qquad t \geqslant 0,\ \ x, y \in \mathscr{G}, \tag{1.0}$$

then $\lambda_1 \geqslant \varepsilon$.

The proof of Theorem 1.0 is omitted here since some analogs were either proved or discussed several times before (see [8; Proof of Theorem 1.4 and Theorem 1.7], [25; Lemmas 2.4 and 2.5], [5; Theorems 6.1 and 6.2] and [6]). Here, we make some remarks only. The first assumption on g of Theorem 1.0 enables us to use the martingale formulation[23]. Note that in order to apply Theorem 1.0, some prior knowledge of the eigenfunction g is required. This is a problem especially when we deal with the whole space $\mathscr{G} = \mathbf{R}^d$. To overcome the difficulty, we adopt a localization procedure: Study the Neumann eigenvalue problem instead of the original one (resp., study the reflecting diffusions instead of the original diffusion), then a limiting procedure provides us with a global bound. Now, for a compact regular domain $\mathscr{G}$, the hypotheses on g of Theorem 1.0 are fulfilled (For instance, by a standard approximation argument, one can even assume that the coefficients of $\overline{L}$ are smooth and so does g). Thus, the key to apply Theorem 1.0 is the exponential estimate (1.0) or the estimate of the moment of T, which is just the place where the coupling technique is employed.

Having these ideas in mind, by using the Riemannian metric

$$(g_{ij}(x)) = (a_{ij}(x))^{-1},$$

one may regard the present situation as a special case of what we treated before[8],[25]. Unfortunately, this idea is not always practical, especially for higher dimension, since the Riemannian distance is usually not explicit and the lower bound of the Ricci curvature is not easy to be computed. Nevertheless, the idea is still meaningful in some special cases, as illustrated by Theorem 1.2 below.

Our next trick is using a comparison (i.e., condition (1.4) below) to reduce the matrix $a(x)$ to the rather simple one $\tilde{a}(x) = \alpha(x)\sigma^2$, where $\alpha(x)$ is a positive function on $\mathbf{R}^d$ and σ is a positive definite constant matrix. There are two reasons to do so. First, up to now, all of our sharp estimates come from optimal couplings, which depend heavily on the distances[5]. The different classes of distances lead to different optimal couplings and hence to different estimates of the spectral gap. We now use the Euclidean distance $|\sigma^{-1}(x-y)|$ instead of the previous Riemannian one. Here, the factor σ^{-1} comes from the fact proved in [5] that the coupling by reflection, constructed in [19] (also [7]) and used for Theorem 1.1 below, is optimal within the class of distances $f(|\sigma^{-1}(x-y)|)$ rather than the class of $f(|x-y|)$, where $f(0) = 0$, $f' > 0$ and $f'' \leqslant 0$. The second reason is more implicit, the reflecting diffusion depends on the boundary and hence on the metric (See Step 4 of the proof in the next section).

To state our results, we need some notation. Given $\alpha(x)$ and σ as above, let $\beta : \mathbf{R}_+ \to \mathbf{R}_+$ and $\gamma : \mathbf{R}_+ \to \mathbf{R}$ satisfy

$$\beta(r) \leqslant \inf_{|\sigma^{-1}(x-y)|=r} \big(\sqrt{\alpha(x)} + \sqrt{\alpha(y)}\,\big)^2, \qquad r > 0,$$

$$\gamma(r) \geqslant \sup_{|\sigma^{-1}(x-y)|=r} \Big[(d-1)\big(\sqrt{\alpha(x)} - \sqrt{\alpha(y)}\,\big)^2 + \big\langle x-y,\, \hat{b}(x) - \hat{b}(y)\big\rangle\Big], \quad r > 0,$$

where $\langle\,,\rangle$ is the ordinary inner product in $\mathbf{R}^d$ and

$$\hat{b}(x) = \alpha(x)\nabla V(x) + \nabla\alpha(x).$$

In particular, $\hat{b}(x) = \nabla V(x)$ when $\alpha(x) \equiv 1$. Define

$$C(r) = \exp\left[\int_{0+}^{r} \frac{\gamma(s)}{s\beta(s)} ds\right], \quad r > 0; \qquad \delta = \left\{\int_{0+}^{\infty} ds C(s)^{-1} \int_{s}^{\infty} \frac{C(u)}{\beta(u)} du\right\}^{-1}. \tag{1.1}$$

Next, for each $n \geqslant 1$, let $\delta_n > 0$ be a constant so that the inequality

$$\beta(r) f''(r) + \gamma(r) f'(r)/r + \delta_n f(r) \leqslant 0, \qquad r \in (0,\, 2n) \tag{1.2}$$

has a solution f satisfying $f' \geqslant 0$ and $\inf_{r\in(0,\,2n)} f(r)/r > 0$. The quantities $\beta(r)$, $\gamma(r)$, δ and δ_n arise natually from the coupling menntioned in the last paragragh. Roughly speaking, the constant δ is used to control the first mement of T: $\delta^{-1} \geqslant \mathbb{E}^{(x,y)} T$ and the condition (1.2) implies that (1.0) holds in the region $\{(x, y) : |\sigma^{-1}(x - y)| < 2n\}$ for $\varepsilon = \delta_n$ and for the function f given by (1.2).

Theorem 1.1. Suppose that

$$\int_{\mathbf{R}^d} \mathsf{tr}\ a(x)\, \pi(dx) < \infty. \tag{1.3}$$

Let $\alpha(x) \in C^2(\mathbf{R}^d)$ be a positive function and let σ be a positive definite matrix such that

$$\lambda_{\max}\big(\alpha(x)\sigma^2 - a(x)\big) \leqslant 0, \qquad x \in \mathbf{R}^d, \tag{1.4}$$

where $\lambda_{\max}(M)$ denotes the maximal eigenvalue of M. Define δ_n and δ as above and set $\delta_\infty = \overline{\lim}_{n\to\infty} \delta_n$. Then, we have

$$\mathsf{gap}\,(L) \geqslant \max\{\delta_\infty,\, \delta\}. \tag{1.5}$$

When $d \geqslant 2$, the existence of the above $\alpha(x)$ and σ is obvious whenever $a(x)$ is strictly positive definite. The simplest choice is as follows:

$$\alpha = \inf_x \lambda_{\min}(a(x)) > 0 \quad \text{ and } \quad \sigma = I.$$

The condition (1.3) is used for the regularity of the corresponding Dirichlet form. The hypotheses of the theorem may be weakened but some regularity of the coefficients $a(x)$ and $b(x)$ is needed since we require the eigenfunction to be either continuous or locally Lipschitz.

The comparison technique used above (i.e., (1.4)) works in general. But in what follows, to simplify our notation, we will often omit such a generalization procedure. Thus, all the results and examples listed below can be actually extended to a much larger class of operators.

Theorem 1.2. The hypotheses for $a(x)$ and $V(x)$ are the same as above. Suppose additionally that $a(x) = \mathsf{diag}\,(a_{ii}(x))$ and $a_{ii}(x) \equiv a_{ii}(x_i)$, $1 \leqslant i \leqslant d$. Set

$$d(x, y) = \sqrt{\sum_{i=1}^{d} \left(\int_{x_i}^{y_i} \frac{1}{\sqrt{a_{ii}(z)}} dz\right)^2},$$

$$h_i(x) = \sqrt{a_{ii}(x)}\, \frac{\partial}{\partial x_i} V(x) + \frac{a'_{ii}(x)}{2\sqrt{a_{ii}(x)}}, \qquad 1 \leqslant i \leqslant d$$

and $D = \sup_{x,y} d(x,y) \leqslant \infty$. Define

$$\beta_1(r) = 4, \qquad \gamma_1(r) \geqslant \sup_{d(x,y)=r} \sum_{i=1}^{d} \big[h_i(y) - h_i(x)\big] \int_{x_i}^{y_i} \frac{dz}{\sqrt{a_{ii}(z)}}, \qquad r \in (0, D).$$

$$C_1(r) = \exp\left[\int_{0+}^{r} \frac{\gamma_1(s)}{4s} ds\right], \quad r \in (0, D);$$

$$\delta^{(1)} = \left\{\int_{0+}^{D} ds C_1(s)^{-1} \int_{s}^{D} \frac{C_1(u)}{4} du\right\}^{-1}.$$

Let $\delta^{(2)} \geqslant 0$ be a constant so that the inequality

$$4f''(r) + \gamma_1(r) f'(r)/r + \delta^{(2)} f(r) \leqslant 0, \qquad r \in (0, D) \tag{1.6}$$

has a solution f satisfying $f' \geqslant 0$ and $\inf_{r\in(0,\,D)} f(r)/r > 0$. Then, we have $\mathsf{gap}\,(L) \geqslant \max\{\delta^{(1)}, \delta^{(2)}\}$.

Note that except $a(x) \equiv$constant, the distance $d(x,y)$ defined above is not translation-invariant and so is essentially different from $|x-y|$ or $|\sigma^{-1}(x-y)|$. The main tool to prove Theorem 1.2 is an optimal Markovian coupling, even through its construction can be regarded as a special Riemannian case of [14] (also [9]), the explicit form given in the Appendix is new.

Finally, consider the one-dimensional case. We use an analytic approach instead of couplings to do the same job. The conditions given below for $\mathsf{gap}\,(L) > 0$ are not only sufficient but sometimes also necessary. The technique goes back to [16] (also [24]) in the context of Markov chains. The presentation given below is parallel to [3] or [4; Chapter 9].

Theorem 1.3. Let $d = 1$ and $a(x)$, $V(x) \in C^1(\mathbf{R})$. Set $\pi(x) = Z^{-1} \exp V(x)$ and $\pi(x,y) = \int_x^y \pi(dz)$, $x \leqslant y$.

(1) Fix $x_0 \in \mathbf{R}$. Suppose that there exist constants c_1, c_2, d_1 and d_2 such that

$$\pi(x,\infty) \leqslant c_1 a(x)\pi(x), \ \ x \geqslant x_0, \qquad \pi(-\infty, x) \leqslant c_2 a(x)\pi(x), \ \ x < x_0,$$

$$\int_x^\infty a(y)\pi(dy) \leqslant d_1 a(x)\pi(x), \ \ x \geqslant x_0, \quad \int_{-\infty}^x a(y)\pi(dy) \leqslant d_2 a(x)\pi(x), \ \ x < x_0.$$

Then,

$$\mathsf{gap}\,(L) \geqslant 1\big/ \max\big\{4c_1[d_1 + 2c_1\pi(-\infty, x_0)],\, 4c_2[d_2 + 2c_2\pi(x_0,\infty)]\big\}.$$

(2) Suppose that (1.3) holds and $\underline{a} := \inf a(x) > 0$. If for fixed x_0, there exist constants c_1 and c_2 so that

$$\pi(x,\infty) \leqslant c_1\pi(x), \ \ x \geqslant x_0, \qquad \pi(-\infty, x) \leqslant c_2\pi(x), \ \ x < x_0.$$

Then,

$$\mathsf{gap}\,(L) \geqslant \underline{a}\big/ \max\big\{4c_1^2[1 + 2\pi(-\infty, x_0)],\, 4c_2^2[1 + 2\pi(x_0,\infty)]\big\}.$$

(3) If $0 < \underline{a} \leqslant a(x) \leqslant \bar{a} < \infty$ and $M := \sup_x\{\mathsf{sgn}(x)\, V'(x)\} < \infty$. Then, the conditions in part (1) (or part (2)) are necessary for $\mathsf{gap}\,(L) > 0$.

Before moving further, we discuss the condition "$M < \infty$" used in part (3) of Theorem 1.3. Note that the condition "$\pi(x)$ being monotone decreasing on $[0, +\infty)$ and increasing on $(-\infty, 0]$" is equivalent to "$M \leqslant 0$". In this case, the condition becomes trivial. On the other hand, if $M = \infty$, then the process may often be irreversible.

Of course, the above technique can be also used to study the problem for a regular domain in $\mathbf{R}^d$. For instance, when $d = 1$, we have the following results.

Corollary 1.4. Consider the diffusion on $[p, q]$ ($[p, q)$ if $q = \infty$ and so on) with Neumann boundary condition.

(1) The conclusions of Theorems 1.1 and 1.2 hold if the quantities: ∞ in (1.1), δ_n and $2n$ in (1.2), D in Theorem 1.2 are replaced by $q - p$, δ_∞, $q - p$ and $D = \sup_{x,y\in[p,q]} d(x, y)$ respectively.

(2) $\mathrm{gap}\,(L) \geqslant -\inf_{\alpha>0} \sup_{x\in[p,q]} \big\{a(x)\alpha^2 + \big(a'(x) + b(x)\big)\alpha + b'(x)\big\}$.

Corollary 1.5. Consider the same diffusion as above. Then, Theorem 1.3 needs only the change: Fix $x_0 \in [p, q]$ and replace respectively $-\infty$ and $+\infty$ with p and q everywhere.

The following examples illustrate the power of the results.

Example 1.6. Take $a(x) \equiv \sigma^2$ and $V(x) = \langle x, bx\rangle/2 + \langle v, x\rangle$ for some matrix $b = (b_{ij})$ with $\lambda_{\max}(b) < 0$ and $v \in \mathbf{R}^d$. Then, we have $\nabla V(x) = bx + v$ and so

$$\big\langle x-y,\, \nabla V(x)-\nabla V(y)\big\rangle = \big\langle \sigma^{-1}(x-y),\, (\sigma b\sigma)\sigma^{-1}(x-y)\big\rangle \leqslant \lambda_{\max}(\sigma b\sigma)|\sigma^{-1}(x-y)|^2.$$

Take $f(r) = r$, we see that (1.2) holds with $\delta_n = -\lambda_{\max}(\sigma b\sigma)$. Therefore,

$$\mathrm{gap}\,(L) \geqslant -\lambda_{\max}(\sigma b\sigma). \tag{1.7}$$

To see that this estimate may be sharp, consider $\sigma = \mathrm{diag}\,(\sigma_{ii})$ and $(b_{ij}) = \mathrm{diag}\,(b_{ii})$. Then, the components of the diffusions are independent. In this case, it is known that $\mathrm{gap}\,(L)$ is just the minimum of the spectral gap of these marginal diffusions. Hence, (1.7) is actually an equality.

Next, consider the special case that $d = 1$, $a(x) \equiv 1$ and $V(x) = -x^2/2$. Then, $\mathrm{gap}\,(L) = 1$. We now take $x_0 = 0$. Recall the Gautschi's estimate[12][1]:

$$\frac{1}{2}\Big[(x^p + 2)^{1/p} - x\Big] < e^{x^p}\int_x^\infty e^{-y^p}dy \leqslant C_p\left[\left(x^p + \frac{1}{C_p}\right)^{1/p} - x\right], \qquad x \geqslant 0,$$
$$C_p = \Gamma\big(1 + 1/p\big)^{p/(p-1)}, \quad p > 1; \qquad C_2 = \pi/4.$$

We have $c_1 = c_2 = \sqrt{\pi/2}$. Hence, Theorem 1.3 gives us $\mathrm{gap}\,(L) \geqslant 1/(4\pi)$.

[1]Here is the related Conte's inequality:

$$x(1 + x/24 + x^2/12)e^{-3x^2/4} < e^{-x^2}\int_0^x e^{y^2} \leqslant \frac{\pi^2}{8x}\big(1 - e^{-x^2}\big).$$

Example 1.7. Take $d = 1$, $a(x) = x$ and $b(x) = -(x - b_0)$, $x \in [0, \infty)$ for some $b_0 > 0$. By Theorem 1.1 or part (2) of Corollary 1.4, we obtain gap $(L) \geqslant 1$, which is again exact. We remark that Theorem 1.2 is available with the choice $f(r) = r$ and $\delta^{(2)} = 1/2$ whenever $b_0 \geqslant 1/2$. Then it gives us gap $(L) \geqslant 1/2$.

It is well known that under some ordinary conditions in physics, there are only three solvable cases for the Sturm-Liouville eigenvalue problem (cf. [20; §1, §2, §9]). Two of them are covered by the above examples. For the third case, $\lambda_1 = 0$ and so we have nothing to do.

Example 1.8. Take $d = 1$, $a(x) = (1 + x^2)^2$ and $V(x) = -(v + 2)x^2/2$, $v > -2$. Set $\bar{v} = v$ if $v \geqslant -3/2$, $= -(v + 3)^2/[3(v + 2)]$ if $v \in (-2, -3/2)$. We claim that

$$\text{gap}\,(L) \geqslant \begin{cases} \max\{1,\, v\}, & \text{if } \ v \geqslant 0 \\ \max\{1 + v,\, [\pi^2\bar{v}^2/16]\,\text{sech}^2\theta\}, & \text{if } \ v \in (-2, 0), \end{cases}$$

where θ is the decreasing limit of θ_n: $\theta_1 = -\bar{v}\pi^2/8$, $\theta_n = \theta_1 \tanh \theta_{n-1}$, $n \geqslant 2$. To show this, we note a general observation. If we set $h(x) = \sqrt{a(x)}\,V'(x) + a'(x)/\big[2\sqrt{a(x)}\big]$. Then, $h(y) - h(x) = \int_x^y h'(z)dz \leqslant -\delta d(x, y)$ for all $x < y$ iff $h'(x) \leqslant -\delta a(x)^{-1/2}$. Equivalently,

$$2a(x)\big[a''(x) + a'(x)V'(x) + 2[a(x)V''(x) + \delta]\big] \leqslant a'(x)^2. \tag{1.9}$$

In the present case, (1.9) holds iff $\delta \leqslant \bar{v}$. Moreover, $D = \pi$. Now, it is easy to check that (1.6) holds case by case for f and $\delta^{(2)}$ given below:

(1) When $v > 0$, $f(r) = r$ and $\delta^{(2)} = v$.
(2) When $v \geqslant 0$, $f(r) = \sin(r/2)$ and $\delta^{(2)} = 1$.
(3) When $v \in (-1, 0)$, $f(r) = \sin(r/2)$ and $\delta^{(2)} = 1 + v$.
(4) When $v \in (-2, 0)$, $f(r) = 2\exp[-cr/8]\,\sinh(c\kappa r/8)$ and $\delta^{(2)} = [\pi^2\bar{v}^2/16]\,\text{sech}^2\theta$, where $c = -\pi\bar{v}$ and $\kappa = \sqrt{1 - 16\delta^{(2)}/c^2}$.

We can also use $\delta^{(1)}$ to produce some lower bounds. This example illustrates the use of Theorem 1.2 in the non-linear case and it is quite close to [8] and [5]. But for the specific situation, the comparison (1.4) (choose $\alpha(x)\sigma^2 \equiv 1$) yields even better estimate: gap $(L) \geqslant v + 2$.

Example 1.9. Take $d = 1$, $a(x) \equiv 1$ and $V(x) = -x^4$. Then, $b(x) = -4x^3$. Note that

$$(x - y)\big(b(x) - b(y)\big) = -4(x - y)^2(x^2 + xy + y^2) \leqslant -(x - y)^4.$$

We have $\gamma(r) = -r^4$ and so $C(r) = \exp[-r^4/16]$. Furthermore, by (1.8), we obtain

$$\begin{aligned} \delta^{-1} &= \frac{1}{4}\int_0^\infty e^{r^4/16}dr \int_r^\infty e^{-s^4/16}ds \\ &= \int_0^\infty e^{r^4}dr \int_r^\infty e^{-s^4}ds \\ &= \int_0^1 \int_r^\infty + \int_1^\infty \int_r^\infty \end{aligned}$$

$$\leqslant \int_0^\infty e^{-s^4} ds + \int_1^\infty \frac{dr}{4r^3}$$
$$\leqslant \Gamma\left(\frac{5}{4}\right) + \frac{1}{8}.$$

Therefore, $\mathrm{gap}\,(L) \geqslant 1/[\Gamma(5/4) + 1/8] \approx 0.9695$. By (1.8), we see that the lower bound provided by Theorem 1.3 is $\mathrm{gap}\,(L) \geqslant 1/[8\Gamma(5/4)^2] \approx 0.1521$.

Example 1.10. Consider the domain $[0, \infty)$ and take $a(x) \equiv 1$ and $V(x) = -bx\,(b > 0)$. Take $x_0 = 0$, then $b(x) = -b$ and $c_1 = 1/b$. By Corollary 1.5, we get $\mathtt{gap}\,(L) \geqslant b^2/4$. This estimate is optimal since $b^2/4$ is an eigenvalue with eigenfunction $g(x) := e^{bx/2}(1 - bx/2)$. However, there is a small problem: $g \notin L^2(\pi)$. To avoid this, simply use $b - \varepsilon$ instead of b in the expression of the above g to define a new function g_ε, compute

$$D(g_\varepsilon, g_\varepsilon)/\|g_\varepsilon - \pi g_\varepsilon\|^2 \geqslant \mathrm{gap}\,(L)$$

and then pass the limit $\varepsilon \to 0$. Finally, we mention that part (2) of Corollary 1.4 give us the same bound but part (1) of Corollary 1.4 is ineffective for this example.

2. Proofs of Theorem 1.1, Theorem 1.2 and Corollary 1.4

a) The proof of Theorem 1.1 consists of four steps.

Step 1. Corresponding to the operator L, we have a Dirichlet form $(D, \mathscr{D}(D))$, which is the Friedrichs extension of

$$D(f, g) = \int \big\langle \nabla f(x),\, a(x)\nabla g(x)\big\rangle \pi(dx), \quad f,\, g \in C_0^\infty(\mathbf{R}^d).$$

By [11; Theorem 1.6.3] and condition (1.3), the semigroup determined by the Dirichlet form is recurrent and so is conservative. Since the Dirichlet form is regular, i.e., $C_0^\infty(\mathbf{R}^d)$ is dense in $\mathscr{D}(D)$ with respect to the norm $(D(f, f) + \|f\|^2)^{1/2}$, we have

$$\begin{aligned}\mathrm{gap}\,(L) = \mathrm{gap}\,(D) &:= \inf\{D(f, f) :\ f \in \mathscr{D}(D),\ \pi f = 0 \text{ and } \|f\| = 1\} \\ &= \inf\big\{D(f, f) : f = \gamma_1 g + \gamma_2 \text{ for some constants } \gamma_1,\, \gamma_2 \text{ and } g \in C_0^\infty(\mathbf{R}^d), \\ &\qquad \pi f = 0 \text{ and } \|f\| = 1\big\}. \end{aligned} \tag{2.1}$$

Actually, the conclusion holds in general, not necessarily diffusions, see [3] or [4; Theorem 9.1]. Similarly, for the operator $\widetilde{L}$ with coefficients

$$\tilde{a}(x) = \alpha(x)\sigma^2 \quad \text{and} \quad \tilde{b}(x) = \alpha(x)\sigma^2\nabla V(x) + \sigma^2\nabla\alpha(x), \tag{2.2}$$

we have $\mathrm{gap}\,(\widetilde{L}) = \mathrm{gap}\,(\widetilde{D})$. Because both L and $\widetilde{L}$ are symmetric with respect to the same measure π, it follows from (1.4) that $\mathrm{gap}\,(L) \geqslant \mathrm{gap}\,(\widetilde{L})$. Thus, we need only to study $\mathrm{gap}\,(L)$ for the operator L having coefficients given by (2.2).

Step 2. If $\sigma \neq I$, replacing the original Riemannian metric I by the new one σ^{-2}, the matrix $\tilde{a}(x) = \alpha(x)\sigma^2$ becomes diagonal $\alpha(x)I$. At the same time,

$\tilde{b}(x)$ and the distance $|x-y|$ are replaced by $\sigma^{-1}\tilde{b}$ and $|\sigma^{-1}(x-y)|$ respectively. Thus, without loss of generality, from now on, assume that our operator L has coefficients:

$$a(x) = \alpha(x)I \quad \text{and} \quad b(x) = \alpha(x)[\nabla V(x) + \nabla \log \alpha(x)]. \tag{2.3}$$

Step 3. For each $n \geqslant 1$, let B_n be the open ball $\{x : |x| < n\}$ and denote by $\overline{B}_n$ its closure. Next, let N be the unit inward-pointing radial vector field on ∂B_n. In B_n, we have a diffusion process with coefficients (2.3) and with the reflecting boundary. The reflecting diffusion is also reversible with Dirichlet form

$$\begin{aligned} D_n(f,f) &= \int_{B_n} \alpha |\nabla f|^2 d\pi_n, \quad f \in \mathscr{D}(D_n), \\ \mathscr{D}(D_n) &\supset \{f \in C_0^\infty(\mathbf{R}^d) : \ Nf|_{\partial B_n} = 0\}, \end{aligned}$$

where $d\pi_n = d\pi/\pi(\overline{B}_n)$ with support $\overline{B}_n$. We now prove that

$$\overline{\lim_{n\to\infty}} \operatorname{gap}(D_n) \leqslant \operatorname{gap}(D), \tag{2.4}$$

where $\operatorname{gap}(D) = \operatorname{gap}(L)$ is the spectral gap corresponding to (2.3).

Given $\varepsilon \in (0, 1/6)$, choose $f = \gamma_1 g + \gamma_2$ for some constants γ_1 , γ_2 and $g \in C_0^\infty(\mathbf{R}^d)$ so that $\pi f = 0$, $\pi f^2 = 1$ and

$$\int_{\overline{B}_n} \alpha |\nabla f|^2 d\pi \leqslant \operatorname{gap}(D) + \varepsilon, \quad \int_{\overline{B}_n^c} \big(\alpha |\nabla f|^2 + \alpha f^2 + f^2 + 1\big) d\pi \leqslant \varepsilon \tag{2.5}$$

for large enough n. We can do so because of (2.1) and the fact that f being constant out of the support of g. Let G be non-negative, smooth, bounded above by 1, equal to 1 on $(-\infty, 0]$ and zero on $[1, \infty)$. Then, $k_1 := \sup |G'| < \infty$. Take $G_n(x) = G(|x| - n)$. Then $G_n \in C^\infty(\mathbf{R}^d)$. Set $f_n = G_n f + k_2$ for some k_2 so that $\pi_{n+1} f_n = 0$. Clearly, $Nf_n|_{\partial B_{n+1}} = 0$. Next,

$$\begin{aligned} |k_2| &= \pi\big(\overline{B}_{n+1}\big)^{-1} \left| \int_{\overline{B}_{n+1}} f G_n d\pi \right| \\ &= \pi\big(\overline{B}_{n+1}\big)^{-1} \left| \int_{\overline{B}_{n+1} \setminus \overline{B}_n} f G_n d\pi - \int_{\overline{B}_n^c} f d\pi \right| \\ &\leqslant 2\, \pi\big(\overline{B}_n\big)^{-1/2} \left[\int_{\overline{B}_n^c} f^2 d\pi \right]^{1/2} \\ &\leqslant 2 \sqrt{\frac{\varepsilon}{1-\varepsilon}}. \end{aligned}$$

$$\int_{\overline{B}_{n+1}} f_n^2 d\pi \geqslant \int_{\overline{B}_n} f^2 d\pi - k_2^2 \pi(\overline{B}_{n+1}) \geqslant 1 - \varepsilon - \frac{4\varepsilon}{1-\varepsilon} > 1 - 6\varepsilon.$$

$$\begin{aligned}\int_{\overline{B}_{n+1}} \alpha|\nabla f_n|^2 d\pi &= \int_{\overline{B}_{n+1}} \alpha|G_n\nabla f + f\nabla G_n|^2 d\pi \\ &\leqslant \int_{\overline{B}_n} \alpha|\nabla f|^2 d\pi + 2\int_{\overline{B}_n^c} \alpha\big(|\nabla f|^2 + k_1^2 f^2\big) d\pi \\ &\leqslant \mathrm{gap}\,(D) + \varepsilon + 2\varepsilon(1+k_1^2).\end{aligned}$$

Hence,

$$\mathrm{gap}\,(D_{n+1}) \leqslant \frac{\int_{\overline{B}_{n+1}} \alpha|\nabla f_n|^2 d\pi}{\int_{\overline{B}_{n+1}} |f_n|^2 d\pi} < \frac{\mathrm{gap}\,(D) + \varepsilon + 2\varepsilon(1+k_1^2)}{1-6\varepsilon}.$$

Since ε is arbitrary, we obtain $\overline{\lim}_{n\to\infty} \mathrm{gap}\,(D_n) \leqslant \mathrm{gap}\,(D)$. We have thus completed the proof of (2.4).

Actually, when $d = 1$, it can be proved that $\mathrm{gap}\,(D_n) \downarrow \mathrm{gap}\,(D)$ as $n \to \infty$.

Step 4. We now need only to estimate $\mathrm{gap}\,(D_n)$. At this step, we adopt the coupling technique. The operator L_n of the reflecting diffusion in $\overline{B}_n$ equals $I_{B_n}L$ plus N times a measure induced by an increasing process with support on the boundary ∂B_n[13] or [22]. The coupling process used here is simply a modification of the coupling by reflection. Recall that the operator $\overline{L} = \overline{L}^{(x,y)}$ of the coupling by reflection starting from (x, y) has coefficients (see [19] and [7] for details):

$$a(x,y) = \begin{pmatrix} \alpha(x)I, & \sqrt{\alpha(x)}(I - 2\bar{u}\bar{u}^*)\sqrt{\alpha(y)} \\ \sqrt{\alpha(x)}(I - 2\bar{u}\bar{u}^*)\sqrt{\alpha(y)}, & \alpha(y)I \end{pmatrix}, \qquad b(x,y) = \begin{pmatrix} b(x) \\ b(y) \end{pmatrix},$$

where $\bar{u} = \bar{u}(x,y) = (x-y)/|x-y|$. We have

$$\begin{aligned}\overline{L}f(|x-y|) =& \big(\sqrt{\alpha(x)} + \sqrt{\alpha(y)}\big)^2 f''(|x-y|) + \big\{(d-1)\big(\sqrt{\alpha(x)} - \sqrt{\alpha(y)}\big)^2 \\ &+ \langle x-y, \hat{b}(x) - \hat{b}(y)\rangle\big\}\frac{f'(|x-y|)}{|x-y|}. \end{aligned} \tag{2.6}$$

Let $L_n^{(x)}$ be the original operator starting from x. Then, our coupling operator equals $I_{B_n^2}\overline{L}^{(x,y)}$ plus the boundary operator $L_n^{(x)} + L_n^{(x)}$. Since the state space $\overline{B}_n$ is compact, the coefficients of L_n are bounded. Moreover,

$$\int_0^t I_{\partial(B_n\times B_n)}\big(X_s, Y_s\big) ds \leqslant \int_0^t I_{\partial B_n}\big(X_s\big) ds + \int_0^t I_{\partial B_n}\big(Y_s\big) ds = 0.$$

Thus, for every $f \in C^2(\mathbf{R}^d)$, we have

$$\begin{aligned}\mathbb{E}_n^{(x,y)} f(|X_{t\wedge T} - Y_{t\wedge T}|) =& f(|x-y|) + \mathbb{E}_n^{(x,y)} \int_0^{t\wedge T} \overline{L}f(|X_s - Y_s|) ds \\ &+ \mathbb{E}_n^{(x,y)} \int_0^{t\wedge T} \Big[N^{(x)} f(|X_s - Y_s|) dL_s^{(x)} + N^{(y)} f(|X_s - Y_s|) dL_s^{(y)}\Big],\end{aligned}$$

where $L_s^{(x)}$ is the increasing process with support contained in $\{t \geqslant 0 : X_t \in \partial B_n\}$ and

$$N^{(x)} = -\frac{1}{n}\sqrt{\alpha(x)}\sum_i x_i \partial/\partial x_i \quad \text{(cf. [22] or [13])}.$$

Furthermore, due to the fact that $f' \geqslant 0$ and the specific boundary, we have

$$\frac{n}{\sqrt{\alpha(x)}}N^{(x)}f(|x-y|) = -\frac{f'(|x-y|)}{|x-y|}\sum_i x_i(x_i-y_i) \leqslant -\frac{f'(|x-y|)}{2|x-y|}(|x|^2-|y|^2)$$
$$\leqslant 0, \qquad \text{if } x \in \partial B_n \text{ and } y \in \overline{B}_n.$$

The same conclusion holds for the $N^{(y)}$'s term. Therefore,

$$\mathbb{E}_n^{(x,y)} f(|X_{t\wedge T} - Y_{t\wedge T}|) \leqslant f(|x-y|) + \mathbb{E}_n^{(x,y)} \int_0^{t\wedge T} \overline{L} f(|X_s - Y_s|)ds, \qquad (2.7)$$

and the boundary operator disappeared. Applying (2.7) to the function f given in (1.2) and using (2.6), we obtain

$$\mathbb{E}_n^{(x,y)} f(|X_t - Y_t|) \leqslant f(|x-y|) - \delta_n \mathbb{E}_n^{(x,y)} \int_0^t f(|X_s - Y_s|)ds,$$

here we have used the usual convention that starting from T, the two components will move as one. Hence, we have

$$\mathbb{E}_n^{(x,y)} f(|X_t - Y_t|) \leqslant f(|x-y|)e^{-\delta_n t}$$

for all $t \geqslant 0$ and $(x,y) \in \overline{B}_n^2$. By part (2) of Theorem 1.0, we obtain $\text{gap}\,(D_n) \geqslant \delta_n$. Next, applying (2.7) to the function

$$f(r) = \int_{0+}^r C(s)^{-1}ds \int_s^\infty \frac{C(u)}{\beta(u)}du, \qquad r > 0$$

we obtain

$$\mathbb{E}_n^{(x,y)} f(|X_{t\wedge T} - Y_{t\wedge T}|) \leqslant f(|x-y|) - \mathbb{E}^{(x,y)}(t \wedge T).$$

Hence

$$\mathbb{E}_n^{(x,y)} T \leqslant f(\infty) = \delta^{-1}.$$

By part (1) of Theorem 1.0, we get $\text{gap}\,(D_n) \geqslant \delta$.

b) The main difference in proving Theorem 1.2 is that we adopt the optimal coupling constructed in Appendix instead of the previous one. Other obvious changes are similar to what mentioned at the beginning of Step 2. By using the same procedure, one can complete the proof of part (1) of Corollary 1.4.

c) Finally, we prove part (2) of Corollary 1.4. The key point here is that we adopt a new distance $d(x,y) = |e^{\alpha x} - e^{\alpha y}|$ $(\alpha > 0)$. Set

$$\delta(\alpha) = \sup_{x\in[p,q]} \big\{a(x)\alpha^2 + \big(a'(x) + b(x)\big)\alpha + b'(x)\big\}.$$

Applying (4.1) to $f : f(r) = r$, it follows that $\overline{L}d(x,y) \leqslant \delta(\alpha)d(x,y)$ and so

$$\mathbb{E}^{(x,y)} d(X_t, Y_t) \leqslant d(x,y)e^{\delta(\alpha)t}.$$

Hence, the conclusion follows from part (2) of Theorem 1.0. Certainly, the result can be easily improved by considering more general f as we did in Theorems 1.1 and 1.2. The main difference is that for the special $f : f(r) = r$, even the classical coupling[17],[7] still achieves the same bound.

3. Proof of Theorem 1.3 and Corollary 1.5

To prove Theorem 1.3, we need a simple result.

Lemma 3.1. Let $m \in C(\mathbf{R}_+;\mathbf{R}_+)$. If $\int_x^\infty m(y)dy \leqslant bm(x)$, $x \geqslant 0$. Then for every $\gamma \in [0, 1/b)$, we have

$$\int_x^\infty e^{\gamma y} m(y)dy \leqslant \frac{b}{1-\gamma b} e^{\gamma x} m(x), \qquad x \geqslant 0.$$

Proof. Set $M(x) = \int_x^\infty m(y)dy$. By integration by parts formula, we have

$$\begin{aligned}\int_x^\infty e^{\gamma y} m(y)dy &= -\int_x^\infty e^{\gamma y} dM(y)\\ &\leqslant e^{\gamma x} M(x) + \gamma \int_x^\infty e^{\gamma y} M(y)dy\\ &\leqslant b e^{\gamma x} m(x) + \gamma b \int_x^\infty e^{\gamma y} m(y)dy. \quad \square\end{aligned}$$

a) We now start to prove part (1) of Theorem 1.3. Clearly, by the assumptions, (1.3) holds and $a(x) > 0$ for all x. Without loss of generality, assume that $x_0 = 0$. Note that for every $f \in C^1(\mathbf{R})$ with $\pi f = 0$ and $\|f\| = 1$, we have

$$\begin{aligned}1 &= \frac{1}{2}\iint \pi(dx)\pi(dy)\big[f(x)-f(y)\big]^2\\ &= \iint_{x<y} \pi(dx)\pi(dy)\big[f(x)-f(y)\big]^2\\ &= \iint_{0\leqslant x<y} + \iint_{x<y\leqslant 0} + \iint_{x<0<y}\\ &=: \mathrm{I}_1 + \mathrm{I}_2 + \mathrm{I}_3. \end{aligned} \tag{3.1}$$

For every $\gamma_1 \in [0, 1/(2d_1))$, which will be determined below, we have

$$\begin{aligned}\mathrm{I}_1 &= \iint_{0\leqslant x<y} \pi(dx)\pi(dy)\bigg(\int_x^y f'(z)dz\bigg)^2\\ &\leqslant \iint_{0\leqslant x<y} \pi(dx)\pi(dy)\bigg(\int_x^y f'(z)^2 e^{-2\gamma_1 z}dz\bigg)\bigg(\int_x^y e^{2\gamma_1 z}dz\bigg)\\ &= \int_0^\infty f'(z)^2 e^{-2\gamma_1 z}dz \iint_{0\leqslant x\leqslant z<y} \pi(dx)\pi(dy)\bigg(\int_x^y e^{2\gamma_1 \tilde{z}}d\tilde{z}\bigg)\\ &= \int_0^\infty f'(z)^2 e^{-2\gamma_1 z}dz \bigg[\iiint_{0\leqslant x\leqslant \tilde{z}\leqslant z<y} \pi(dx)\pi(dy)e^{2\gamma_1\tilde{z}}d\tilde{z}\\ &\qquad + \iiint_{0\leqslant x\leqslant z<\tilde{z}<y} \pi(dx)\pi(dy)e^{2\gamma_1\tilde{z}}d\tilde{z}\bigg]\\ &\leqslant \int_0^\infty f'(z)^2 e^{-2\gamma_1 z}dz\bigg[\int_0^z e^{2\gamma_1\tilde{z}}d\tilde{z}\int_z^\infty \pi(dy) + \int_z^\infty e^{2\gamma_1\tilde{z}}d\tilde{z}\int_{\tilde{z}}^\infty \pi(dy)\bigg]. \end{aligned} \tag{3.2}$$

By using the assumptions and applying Lemma 3.1 to $m(y) = a(y)\pi(y)$, we get

$$\begin{aligned}
\mathrm{I}_1 &\leqslant \int_0^\infty f'(z)^2 e^{-2\gamma_1 z} dz \left[c_1 a(z)\pi(z) \int_0^z e^{2\gamma_1 \tilde{z}} d\tilde{z} + c_1 \int_z^\infty e^{2\gamma_1 \tilde{z}} a(\tilde{z})\pi(\tilde{z}) d\tilde{z} \right] \\
&\leqslant c_1 \int_0^\infty f'(z)^2 a(z)\pi(z) \left[\frac{1}{2\gamma_1} \big(e^{2\gamma_1 z} - 1\big) e^{-2\gamma_1 z} + \frac{d_1}{1 - 2\gamma_1 d_1} \right] dz \\
&\leqslant c_1 \left[\frac{1}{2\gamma_1} + \frac{d_1}{1 - 2\gamma_1 d_1} \right] \int_0^\infty f'(z)^2 a(z)\pi(dz).
\end{aligned}$$

Minimizing the right-hand side with respect to γ_1, we obtain

$$\mathrm{I}_1 \leqslant 4c_1 d_1 \int_0^\infty f'(z)^2 a(z)\pi(dz). \tag{3.3}$$

Similarly, we have

$$\mathrm{I}_2 \leqslant 4c_2 d_2 \int_{-\infty}^0 f'(z)^2 a(z)\pi(dz). \tag{3.4}$$

Next,

$$\begin{aligned}
\mathrm{I}_3 &= \iint_{x<0<y} \pi(dx)\pi(dy) \big[f(x) - f(y) \big]^2 \\
&\leqslant 2 \iint_{x<0<y} \pi(dx)\pi(dy) \left[\left(\int_0^x f'(z) dz \right)^2 + \left(\int_0^y f'(z) dz \right)^2 \right].
\end{aligned} \tag{3.5}$$

But for $\gamma_2 \in [0, 1/(2c_1))$,

$$\begin{aligned}
&\iint_{x<0<y} \pi(dx)\pi(dy) \left(\int_0^y f'(z) dz \right)^2 \\
&\qquad \leqslant \pi(-\infty, 0) \int_0^\infty \pi(dy) \int_0^y f'(z)^2 e^{-2\gamma_2 z} dz \int_0^y e^{2\gamma_2 \tilde{z}} d\tilde{z} \\
&\qquad \leqslant \frac{\pi(-\infty, 0)}{2\gamma_2} \int_0^\infty f'(z)^2 e^{-2\gamma_2 z} dz \int_z^\infty \big[e^{2\gamma_2 y} - 1 \big] \pi(dy) \\
&\qquad \leqslant \frac{\pi(-\infty, 0)}{2\gamma_2} \int_0^\infty f'(z)^2 e^{-2\gamma_2 z} dz \int_z^\infty e^{2\gamma_2 y} \pi(dy) \\
&\qquad \leqslant \frac{c_1 \pi(-\infty, 0)}{2\gamma_2 (1 - 2\gamma_2 c_1)} \int_0^\infty f'(z)^2 a(z)\pi(dz).
\end{aligned}$$

Here in the last line, we have used Lemma 3.1 to $m(y) = \pi(y)$. By using the same optimizing procedure, we get

$$\iint_{x<0<y} \pi(dx)\pi(dy) \left(\int_0^y f'(z) dz \right)^2 \leqslant 4c_1^2 \, \pi(-\infty, 0) \int_0^\infty f'(z)^2 a(z)\pi(dz). \tag{3.6}$$

Similarly, we have

$$\iint_{x<0<y} \pi(dx)\pi(dy)\bigg(\int_x^0 f'(z)dz\bigg)^2 \leqslant 4c_2^2\,\pi(0,\infty)\int_{-\infty}^0 f'(z)^2 a(z)\pi(dz). \tag{3.7}$$

Collecting (3.1), (3.3)–(3.7) together, we prove part (1) of Theorem 1.3.

b) The proof of part (2) of Theorem 1.3 is a simple modification of the above one. For instance, starting from (3.2) and applying Lemma 3.1 to $m(y)=\pi(y)$, we get

$$\begin{aligned}\mathrm{I}_1 &\leqslant c_1\int_0^\infty f'(z)^2 e^{-2\gamma_1 z}dz\bigg[\pi(z)\int_0^z e^{2\gamma_1\tilde{z}}d\tilde{z}+\int_z^\infty e^{2\gamma_1\tilde{z}}a(\tilde{z})\pi(\tilde{z})d\tilde{z}\bigg]\\ &\leqslant \frac{c_1}{\underline{a}}\bigg[\frac{1}{2\gamma_1}+\frac{c_1}{1-2\gamma_1 c_1}\bigg]\int_0^\infty f'(z)^2 a(z)\pi(dz).\end{aligned}$$

Hence

$$\mathrm{I}_1 \leqslant \frac{4c_1^2}{\underline{a}}\int_0^\infty f'(z)^2 a(z)\pi(dz).$$

c) As for Corollary 1.5, note that the key of the proofs a) and b) is the asymptotic behavior of the integrals $\int_{-\infty}^x \pi(dy)$ and $\int_{-\infty}^x a(y)\pi(dy)$ $\big($resp., $\int_x^\infty \pi(dy)$ and $\int_x^\infty a(y)\pi(dy)\big)$ as x tends to the "boundary" $-\infty$ (resp., $+\infty$). In the present case, the boundaries $-\infty$ and ∞ are replaced by p and q respectively.

d) Finally, we prove part (3) of Theorem 1.3. Clearly, it suffices to consider the case $x \geqslant x_0$ only. Choose x_0 so that $\pi(-\infty,x_0)=\pi(x_0,\infty)=1/2$. Define G and k_1 as in the last section. Let $x \geqslant x_0$ and set $f(y)=G(y-x)-k_3$, $y\in\mathbf{R}$, where $k_3=\int G(y-x)\pi(dx) \leqslant \pi(x,\infty) \leqslant \pi(x_0,\infty)=1/2$. Without loss of generality, we may assume that $M' := \max\big\{M, \sup_{|z|\leqslant|x_0|}|V'(z)|\big\} > 0$. Then

$$\begin{aligned}\pi(x,x+1) &= \pi(x)\int_x^{x+1}\exp\bigg[\int_x^y V'(\xi)d\xi\bigg]dy\\ &\leqslant \pi(x)\int_x^{x+1} e^{M'(y-x)}dy\\ &= \pi(x)\frac{e^{M'}-1}{M'}.\end{aligned}$$

On the other hand,

$$\begin{aligned}&\int a(z)f'(z)^2\pi(dz) \leqslant \bar{a}\,k_1^2\,\pi(x,x+1),\\ &\int f^2 d\pi \geqslant \int_{x+1}^\infty (1-k_3)^2 d\pi \geqslant \frac{1}{4}\pi(x+1,\infty)=\frac{1}{4}\big[\pi(x,\infty)-\pi(x,x+1)\big].\end{aligned}$$

We obtain

$$\begin{aligned}\pi(x,\infty) &\leqslant 4\int f^2 d\pi + \pi(x,x+1)\\ &\leqslant \frac{4}{\operatorname{gap}(L)}\int a(z)f'(z)^2\pi(dz) + \pi(x,x+1)\\ &\leqslant \left[\frac{4\bar{a}\,k_1^2}{\operatorname{gap}(L)}+1\right]\pi(x,x+1)\\ &\leqslant \left[\frac{4\bar{a}\,k_1^2}{\operatorname{gap}(L)}+1\right]\frac{e^{M'}-1}{M'}\pi(x).\end{aligned}$$

4. Appendix

Consider diffusion processes in $\mathbf{R}^d$, starting from x and y respectively, with operators $L^{(x)} \sim (a(x), b(x))$ and $L^{(y)} \sim (a(y), b(y))$. The coefficients of any coupling operator $\widetilde{L}$ should be of the form

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

where $a(x,y)$ is non-negative definite. Note that the only freedom here is the choice of $c(x,y)$. Given a metric $\rho \in C^2(\mathbf{R}^d \times \mathbf{R}^d \setminus \{(x,x) : x \in \mathbf{R}^d\})$, a coupling operator $\overline{L}$ is called $\rho-$**optimal** if $\overline{L}\rho(x,y) = \inf_{\widetilde{L}} \widetilde{L}\rho(x,y)$ for all $x \neq y$, where $\widetilde{L}$ varies over all coupling operators. Refer to [5] for some ρ-optimal solutions. The main purpose of this section is to prove an additional optimal solution (Theorem 4.2). To do so, we need the following result, which can be checked by some computations.

Lemma 4.1. Let $\mathscr{G}$ be an open domain of $\mathbf{R}^d \times \mathbf{R}^d$. Given $\varphi \in C^2(\mathscr{G})$. Then, for any coupling operator $\widetilde{L}$, we have

$$\begin{aligned}\widetilde{L}f\circ\varphi(x,y) =& \Big[\big\langle \nabla_x\varphi,\, a(x)\nabla_x\varphi\big\rangle + \big\langle \nabla_y\varphi,\, a(y)\nabla_y\varphi\big\rangle\\ &+ 2\big\langle \nabla_x\varphi,\, c(x,y)\nabla_y\varphi\big\rangle\Big] f''\circ\varphi(x,y)\\ &+ \left[L^{(x)}\varphi(x,y) + L^{(y)}\varphi(x,y) + 2\sum_{i,j} c_{ij}(x,y)\frac{\partial^2\varphi}{\partial x_i\partial y_j}\right] f'\circ\varphi(x,y),\\ &(x,y)\in\mathscr{G},\end{aligned} \tag{4.1}$$

where ∇_x is the gradient with respect to the variable x, $L^{(x)}$ acts on $\varphi(\cdot, y)$ for fixed y and similar for other notations.

Theorem 4.2. Let $a(x) = \operatorname{diag}(a_{ii}(x))$ with $a_{ii}(x) = a_{ii}(x_i) > 0$, $1 \leqslant i \leqslant d$. Define

$$d(x,y) = \sqrt{\sum_{i=1}^d \left(\int_{x_i}^{y_i} \frac{dz}{\sqrt{a_{ii}(z)}}\right)^2} \quad \text{and} \quad \bar{u}_i = \frac{1}{d(x,y)}\int_{x_i}^{y_i}\frac{dz}{\sqrt{a_{ii}(z)}}, \quad x \neq y.$$

Given $f \in C^2(\mathbf{R}_+;\mathbf{R}_+)$ with $f(0)=0$, $f'>0$ and $f''\leqslant 0$. Set $\rho = f\circ d$. Then, the ρ-optimal solution $c(x,y)$ is given by

$$c(x,y) = a(x)^{1/2}(I-2\bar{u}\bar{u}^*)a(y)^{1/2}.$$

Furthermore,

$$\overline{L}f\circ d(x,y) = 4f''\circ d(x,y) - \sum_{i=1}^d \bar{u}_i\bigg(\frac{2b_i(y)-a'_{ii}(y)}{2\sqrt{a_{ii}(y)}} - \frac{2b_i(x)-a'_{ii}(x)}{2\sqrt{a_{ii}(x)}}\bigg)f'\circ d(x,y). \tag{4.2}$$

Proof. a) First, by some computations, we obtain

$$\begin{aligned}
&\frac{\partial d(x,y)}{\partial x_i} = -\frac{\bar{u}_i}{\sqrt{a_{ii}(x)}}, \qquad \frac{\partial d(x,y)}{\partial y_i} = \frac{\bar{u}_i}{\sqrt{a_{ii}(y)}},\\
&\frac{\partial^2 d(x,y)}{\partial^2 x_i} = \frac{1}{a_{ii}(x)d(x,y)} - \frac{\bar{u}_i^2}{a_{ii}(x)d(x,y)} + \frac{a'_{ii}(x)\bar{u}_i}{2a_{ii}(x)^{3/2}},\\
&\frac{\partial^2 d(x,y)}{\partial^2 y_i} = \frac{1}{a_{ii}(y)d(x,y)} - \frac{\bar{u}_i^2}{a_{ii}(y)d(x,y)} - \frac{a'_{ii}(y)\bar{u}_i}{2a_{ii}(y)^{3/2}},\\
&\frac{\partial^2 d(x,y)}{\partial x_i\partial y_j} = \frac{1}{\sqrt{a_{ii}(x)a_{jj}(y)}\,d(x,y)}\big[\bar{u}_i\bar{u}_j-\delta_{ij}\big], \qquad x\neq y.
\end{aligned} \tag{4.3}$$

Rewrite $c(x,y) = a(x)^{1/2}H(x,y)^*a(y)^{1/2}$ for some $H = H(x,y)$. Substituting (4.3) into (4.1), we get

$$\begin{aligned}
\widetilde{L}f\circ d(x,y) = {}& 2\big[1-\langle \bar{u},\, H\bar{u}\rangle\big]f''\circ d(x,y) + \frac{2}{d(x,y)}\bigg[d-1-\operatorname{tr}H+\langle \bar{u},\, H\bar{u}\rangle\\
&+\sum_i\bigg(\frac{a'_{ii}(x)-2b_i(x)}{2\sqrt{a_{ii}(x)}} - \frac{a'_{ii}(y)-2b_i(y)}{2\sqrt{a_{ii}(y)}}\bigg)\bar{u}_i\bigg]f'\circ d(x,y), \quad x\neq y.
\end{aligned}$$

b) Note that we may forget the term $\sum_i(\cdots)$ in the last line for a moment since it is irrelevant to H. Thus, we need only to minimize

$$F(H) := 2\big[1-\langle \bar{u},\, H\bar{u}\rangle\big]f''\circ d(x,y) + \frac{2}{d(x,y)}\big[d-1-\operatorname{tr}H+\langle \bar{u},\, H\bar{u}\rangle\big].$$

Next, set $A = I+I-2H = 2(I-H)$, $\overline{A} = \langle \bar{u},\, A\bar{u}\rangle$. Then,

$$F(H) = \overline{A}f''\circ d(x,y) + \frac{\operatorname{tr}A-\overline{A}}{d(x,y)}f'\circ d(x,y).$$

Therefore, the proof of [5; Theorem 5.3] gives us the required conclusion. □

Acknowledgement. The authors are greatly indebted to Prof. L. Gross for his advice and encouragement, to Prof. M. Fukushima for introducing us their new book and to a referee, whose detailed comments helped to improve the quality of the paper.

References

[1] Barlow, M. T. and Bass, R. (1993), *Coupling and Hanack inequalities for Sierpinski carpets*, Bull. AMS. New Ser. 29:2, 208–212.

[2] Bérard, P. H. (1986), *Spectral Geometry , Direct and Inverse Problems*, LNM. Vol. 1207, Springer-Verlag.

[3] 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.

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

[5] Chen, M. F. (1993a), *Optimal Markovian couplings and applications*, Technical Report, No.215 (1993), Carleton Univ. and No.147(1993), C. V. Volterra, Univ. of Roma II, Acta Math. Sin. New Ser. 10:3, 260-275 (1994).

[6] Chen, M. F. (1993b), *Optimal couplings and application to Riemannian geometry*, to appear in Prob. Theory and Math. Stat., Vol.1, Edited by B. Grigelionis et al. 1994 VPS/TEV.

[7] Chen, M. F. and Li, S. F. (1989), *Coupling methods for multi-dimensional diffusion processes*, Ann. of Probab. 17:1, 151–177.

[8] Chen, M. F. and Wang, F. Y. (1992), *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.

[9] Cranston, M. (1991), *Gradient estimates on manifolds using coupling*, J. Funct. Anal. 99, 110–124.

[10] Cranston, M. (1993), *A probabilistic approach to Martin boundaries for manifolds with ends*, Probab. Th. Rel. Fields, 96:3, 319–334.

[11] Fukushima, M., Oshima, Y. and Takeda, M. (1994), *Dirichlet Forms and Symmetric Markov Processes*, Walter de Gruyter & Co.

[12] Gautschi, W. (1959), Some elementary inequalities relating to the gamma and incomplete gammafunction. J. Math. and Phys. 38, 77–81.

[13] Ikeda, N. and Watanabe, S. (1981), *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Kodansha, Tokyo.

[14] Kendall, W. (1986), *Nonnegative Ricci curvature and the Brownian coupling property*, Stochastics 19, 111–129.

[15] Kendall, W. (1989), *Coupled Brownian motion and partial domain monotonicity for the Neumann heat kernel*, J. Funct. Anal. 86, 226–236.

[16] Liggett, T. M. (1989), *Exponential L_2 convergence of attractive reversible nearest particle systems*, Ann. Probab., 17, 403-432.

[17] Lindvall, T. (1983), *On coupling for diffusion processes*, J. Appl. Probab. 20, 82–93.

[18] Lindvall, T. (1992), *Lectures on the Coupling Method*, Wiley, New York.

[19] Lindvall, T. and Rogers, L. C. G. (1986), *Coupling of multidimensional diffusion processes*, Ann. of Probab. 14:3, 860–872.

[20] Nikiforov, A. F. and Urarov, V. B. (1988), *Special Functions of Mathematical Physics*, Birkhäuser.

[21] Schechter, M. (1986), *Spectra of Partial Differential Operators*, North-Holland.

[22] Stroock, D. W. and Varadhan, S. R. S. (1971), *Diffusion Processes with boundary conditions*, Comm. Pure Appl. Math. 24, 147–225.

[23] Stroock, D. W. and Varadhan, S. R. S. (1979), *Multidimensional Diffusion Processes*, Springer-Verlag, New York.

[24] Sullivan, W. G. (1984), *The L^2 spectral gap of certain positive recurrent Markov chains and jump processes*, Z. Wahrs. 67, 387–398.

[25] Wang, F. Y. (1994), *Application of coupling method to the Neumann eigenvalue problem*, Prob. Th. Rel. Fields 98, 299–306.

[26] Weinberger, H. F. (1974), *Variational Methods for Eigenvalue Approximation*, Soc. Indus. Appl. Math., Philadelphia.

Department of Mathematics, Beijing Normal University, Beijing 100875, The People's Republic of China.

J. Funct. Anal. 144(1997)2, 287-300.

ESTIMATES OF LOGARITHMIC SOBOLEV CONSTANT — AN IMPROVEMENT OF BAKRY–EMERY CRITERION

Mu-Fa Chen and Feng-Yu Wang

(Beijing Normal University)
Received September 12, 1994; accepted January 27, 1996

Abstract. This paper is mainly devoted to estimate the logarithmic Sobolev (abbrev. L.S.) constant for diffusion operators on manifold or in $\mathbb{R}^d$. In most cases, we study the lower bounds but a generalization to [9; Theorem 1] for the upper bound is also presented (Theorem 1.5). Based on a simple observation (due to [5]) of the comparison between the L.S. constants for different potentials, the powerful Bakry-Emery criterion for the L.S. inequality is improved considerably in the paper, especially for the manifolds with non-positive sectional curvatures (Theorem 1.3 (1)). In terms of our notation: $\beta(r) = \inf_{\rho(x,p)\geqslant r} \inf_{X\in T_x(M),\, \|X\|=1}(\mathrm{Ricc}-\mathrm{Hess}_V)(X,X)$, where $\rho(x,p)$ is the distance between x and an arbitrarily fixed point $p \in M$, the improvement can be roughly stated as follows. The condition "$\inf_{r\geqslant 0}\beta(r) > 0$" for which the criterion is available is now replaced by "$\sup_{r\geqslant 0}\beta(r) > 0$".

1. Main Results and Examples

Let (M, g) be a d-dimensional, connected, complete Riemannian manifold and let Ω be a compact and convex regular domain of M. Suppose that $\mathrm{Ricci} \geqslant Kg$ on M for some constant $K \in \mathbb{R}$. Next, let $L = \Delta + \nabla V$, $V \in C^2(\Omega)$. Consider the reflecting L-diffusion process with reversible measure $d\mu = e^V d\lambda/Z$, where λ is the Riemannian volume element and $Z = \int_\Omega e^V d\lambda$ (cf. [10]). Since Ω is compact, the following **logarithmic Sobolev inequality** (Gross [7])

$$\int_\Omega f^2 \log f^2 d\mu \leqslant \frac{2}{\alpha}\int_\Omega \|\nabla f\|^2 d\mu \tag{1.1}$$

holds for some constant $\alpha > 0$ and for all $f \in C^1(\Omega)$ with $\mu(f^2) := \int_\Omega f^2 d\mu = 1$. The largest constant α, denoted by $\alpha_\Omega(V)$, is called the **L.S. constant**. The inequality has a very wide range of applications. Refer to the survey article [8] for the history and the current status of the study on the topic.

2000 *Mathematics Subject Classification.* 35P15, 60j60.

Key words and phrases. Logarithmic Sobolev constant, Bakry-Emery criterion, diffusion process..

Research supported in part by NSFC and the State Education Commission of China.

One powerful method to deduce the inequality is the **Bakry-Emery** (abbrev. B.-E.) **criterion**[2] which has been reexamined and improved by many authors (refer to [1] and [4] for details and references therein). For instance, Deuschel and Stroock [5; Remark 1.20] mentioned the following comparison between the L.S. constants for different potentials V and U:

$$\alpha_\Omega(V) \geqslant \alpha_\Omega(U)\exp[-\mathrm{osc}_\Omega(V-U)], \tag{1.2}$$

where $\mathrm{osc}_\Omega(V) = \sup_\Omega V - \inf_\Omega V$ (The negative sign in the exponential was missed in [5; (1.21)]). This is a starting point of the paper. To check (1.2), simply use the identity

$$\int_\Omega f\log\frac{f}{\mu(f)}d\mu = \inf\left\{\int_\Omega (f\log f - f\log t - f + t)d\mu :\ t\in(0,\infty)\right\}$$

for all strictly positive and smooth f and note that the integrand on the right-hand side is non-negative for all $t\in(0,\infty)$. At the first look, (1.2) seems quite rough but it does yield sharp estimates as we will see in Corollary 1.6 and examples below. On the other hand, it was proved in [5] and [11] that

$$\alpha_\Omega(V) \geqslant K_\Omega(V) + d^{-1}\lambda_1(0)e^{-\mathrm{osc}_\Omega(V)}, \tag{1.3}$$

where

$$K_\Omega(V) = \inf\big\{(\mathrm{Ric} - \mathrm{Hess}_V)(X,X) :\ X\in T_xM,\ \|X\|=1,\ x\in\Omega\big\}$$

and $\lambda_1(V)$ is the **spectral gap** (= the first non-trivial eigenvalue) of the reflecting L-diffusion on Ω (see [10] for some detailed estimates of $\lambda_1(V)$). Actually, $\lambda_1(V)$ is the largest constant λ for which the **Poincaré inequality**

$$\int_\Omega \big(f-\mu(f)\big)^2 d\mu \leqslant \frac{1}{\lambda}\int_\Omega \|\nabla f\|^2 d\mu, \qquad f\in C^1(\Omega)$$

holds. A well-known fact is that $\lambda_1(V) \geqslant \alpha_\Omega(V)$. When $K>0$, the estimate (1.3) can be sharp in the free boundary situation[5], but they are ineffective for sufficient small K. Thus, we will concentrate on the case of small K (especially, $K\leqslant 0$).

Let ρ be the Riemannian distance induced by g. Fixed $p\in\Omega$ and set $D = \sup_\Omega \rho(x,p)$. Denote by $C(p)$ the cut locus of p. Define

$$\widetilde{\Omega} = \{x\in M : \text{there exists } y\in\Omega \text{ such that } x \text{ belongs to the shortest geodesic from } p \text{ to } y\}.$$

Now, as an addition to [5] and [11], we have the following result.

Theorem 1.1. Suppose that $\Omega \cap C(p) = \emptyset$ and the sectional curvatures of $\widetilde{\Omega}$ are bounded above by a constant $k \in \mathbb{R}$. Then

$$\alpha_\Omega(V) \geqslant \sup_{\beta>0} \Big(\alpha_\beta + d^{-1}e^{-\beta D^2}\lambda_1(0)\Big) e^{-\mathrm{osc}_\Omega\left(V+\beta\rho(\cdot,p)^2\right)},$$

where

$$\alpha_\beta = \begin{cases} K + 2\beta, & \text{if } k \leqslant 0 \\ K + 2\sqrt{k}D \operatorname{ctan}(\sqrt{k}D)\beta, & \text{if } k > 0 \text{ and } 2\sqrt{k}D < \pi. \end{cases}$$

The proof of the theorem is based on the Hessian comparison theorem (see (2.1) and (2.2) in the next section). From which the restriction "$2\sqrt{k}\,D < \pi$" in the last line arises. The next result is a simple consequence of Theorem 1.1.

Corollary 1.2. Under the assumptions of Theorem 1.1, we have

$$\alpha_\Omega(V) \geqslant \begin{cases} e^{-\mathrm{osc}_\Omega(V)}\Big\{\dfrac{2}{D^2}\exp\Big[-1+\dfrac{KD^2}{2}\Big] + \dfrac{\lambda_1(0)}{d}\exp\big[-2+KD^2\big]\Big\}, \\ \qquad\qquad \text{if } \ k \leqslant 0 \\ e^{-\mathrm{osc}_\Omega(V)}\Big\{\dfrac{2\sqrt{k}}{D\tan\left(\sqrt{k}D\right)}\exp\Big[-1+\dfrac{KD\tan\left(\sqrt{k}D\right)}{2\sqrt{k}}\Big] \\ +\dfrac{\lambda_1(0)}{d}\exp\Big[-2+\dfrac{KD\tan\left(\sqrt{k}D\right)}{\sqrt{k}}\Big]\Big\}, \\ \qquad\qquad \text{if } 0 < k \leqslant \dfrac{\pi^2}{4D^2} \text{ and } \dfrac{\sqrt{k}}{\tan\left(\sqrt{k}D\right)} > \dfrac{KD}{2}. \end{cases}$$

Next, we go to the free boundary case. We consider the non-compact manifold only since in the compact case the same topic was treated in [5] and [11]. Again, we will use the comparison (1.2) which also holds in the present situation. However, the potential now becomes more essential, without it, $\alpha(L)$ can be vanished. Hence, to produce a good estimate, the potential U has to be carefully designed especially for unbounded manifold (see also the remark right after the proof of Theorem 1.3).

Consider the operator having the form $L = \Delta + \nabla V$ and assume that its Dirichlet form is regular. Replacing Ω in (1.1) by the whole space M, we obtain the L.S. inequality for L and then we have the constants $\alpha(V) := \alpha_M(V)$ and $K(V) := K_M(V)$. Next, define

$$K(V,x) = \inf\{(\mathrm{Ric} - \mathrm{Hess}_V)(X,X) : \ X \in T_xM, \ \|X\| = 1\}, \qquad x \in M.$$

Clearly, $K(V) = \inf_x K(V,x)$. Note that in the most interesting (non-compact) cases, $\mathrm{osc}(V) = \infty$ and so the criterion (1.3) becomes $\alpha(V) \geqslant K(V)$. Fix $p \in M$ and let $\beta(r) = \inf_{\rho(x,p)\geqslant r} K(V,x)$. Obviously, $\beta(r)$ is increasing in r. Moreover, $\beta(0) = \inf_{r\geqslant 0}\beta(r) = K(V)$. For fixed $k \geqslant 0$, define $f(r) = r$ if $k = 0$ and

$f(r) = \sin\big(\sqrt{k}\,r\big)/\sqrt{k}$ if $k > 0$. Set $\tilde{\beta}(r) = \inf_{u: f(u) \in [r, \pi/(2\sqrt{k}))} \beta(u)/f'(u)$. Here and in what follows, $1/\sqrt{k}$ is understood as ∞ when $k = 0$. Note that $\tilde{\beta}(r) = \beta(r)$ when $k = 0$ since β is an increasing function. Finally, for fixed $a \in \big[0, \pi/\big(2\sqrt{k}\,\big)\big)$, define

$$\gamma(r) = \frac{1}{f(r)} \int_0^{f(r)} \tilde{\beta}(u)du,\ r < \frac{\pi}{2\sqrt{k}} \quad \text{and}$$
$$F_a(r) = \int_0^{r\wedge a} ds \int_0^{f(s)} \big[\gamma(a) - \tilde{\beta}(u)\big]du, \qquad r \geqslant 0. \tag{1.4}$$

We can now state the main result of the paper.

Theorem 1.3. Suppose that the sectional curvatures of M are bounded above by a constant $k \in \mathbb{R}$.

(1) Let $k = 0$. If $M \cap C(p) = \emptyset$ and $\sup_{r\geqslant 0} \beta(r) > 0$, then we have

$$\alpha(V) \geqslant \frac{2}{a_0^2} \exp\left[1 - \int_0^{a_0} r\beta(r)dr\right] > 0, \tag{1.5}$$

where $a_0 > 0$ is the unique solution to the equation $\int_0^a \beta(r)dr = 2/a$.

(2) Let $k > 0$. If $C(p) \cap B\big(p, \pi/\big(2\sqrt{k}\,\big)\big) = \emptyset$ and $\gamma(a) > 0$ for some $a \in \big(0, \pi/\big(2\sqrt{k}\,\big)\big)$, then we have $\alpha(V) \geqslant f'(a)\gamma(a)\exp[-F_a(a)] > 0$.

When $k = 0$, the B.-E. criterion requires that $\inf_{r\geqslant 0} \beta(r) > 0$. From this, one sees that the criterion is now improved considerably by Theorem 1.3 (1). Actually, as we will prove in the next section (see (2.5)), the lower bound given in (1.5) always dominates $\beta(0)$. Besides, note that the L.S. inequality is based on a kind of (uniform) ergodicity, which requires a limiting behavior of the potential when $\rho(x, p) \to \infty$. From this point of view, our condition "$(\lim_{r\to\infty} \beta(r) =) \sup_{r\geqslant 0} \beta(r) > 0$" seems reasonable.

We now turn to study the multi-dimensional diffusion processes. Let

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

where $a(x) = (a_{ij}(x))$ is positive definite, $a_{ij} \in C^2(\mathbb{R}^d)$ and

$$b_i(x) = \sum_{j=1}^d a_{ij}(x) \frac{\partial}{\partial x_j} V(x) + \sum_{j=1}^d \frac{\partial}{\partial x_j} a_{ij}(x)$$

for some $V \in C^2(\mathbb{R}^d)$ with $Z := \int e^V dx < \infty$. The specific form of b_i implies that the L-diffusion process is reversible with respect to $d\mu = Z^{-1}e^V dx$ (see [3]). In the present context, the L.S. inequality becomes

$$\int_{\mathbb{R}^d} f^2 \log f^2 d\mu \leqslant \frac{2}{\alpha(L)} \int_{\mathbb{R}^d} \langle a\nabla f, \nabla f\rangle d\mu \tag{1.6}$$

for all bounded $f \in C^2$ with $\mu(f^2) = 1$, where $\langle \cdot, \cdot \rangle$ is the Euclidean inner product. Here we have used $\alpha(L)$ rather than $\alpha(V)$ to denote the L.S. constant which now depends on the whole coefficients of L, not only on the potential V. Certainly, by using the Riemannian metric $g = a(x)^{-1}$, one can regard the present situation as a special case of what treated above. However, in general, both the Riemannian distance and the Ricci curvature are too complex to be computed. To avoid doing so, we adopt the idea of [3] to simplify the operator by a comparison argument (see the proof of Corollary 1.4 for details). In this way, we obtain the following simple consequence of Theorem 1.3 (1).

Corollary 1.4. Suppose that $a(x) \geqslant \delta\sigma^2$ for some $\delta > 0$ and a positive definite constant matrix σ. Let $\lambda_V(x)$ be the largest eigenvalue of the matrix $\sigma(\partial^2 V(x)/\partial x_i \partial x_j)\sigma$ and let $\bar{\beta}(r) = \inf_{|\sigma^{-1}(x-p)| \geqslant r} \{-\lambda_V(x)\}$ for fixed $p \in \mathbb{R}^d$. If $\sup_{r \geqslant 0} \bar{\beta}(r) > 0$, then we have

$$\alpha(L) \geqslant \frac{2\delta}{a_0^2} \exp\left[1 - \int_0^{a_0} r\bar{\beta}(r)dr\right] > 0,$$

where $a_0 > 0$ is the unique solution to the equation $\int_0^a \bar{\beta}(r)dr = 2/a$.

Finally, we go to study the upper bound of the L.S. constant. As was mentioned above, the spectral gap already provides a upper bound for $\alpha(L)$. A different approach is provided by the following result which is a generalization to [9; Theorem 1].

Theorem 1.5. Suppose that $a(x) \leqslant \nu(x)\bar{a}(x)$ for some non-negative $\nu \in C(\mathbb{R}^d)$ and a matrix $\bar{a}(x)$ with continuous components and having the property: there exist constants $\bar{\nu}_1$, $\bar{\nu}_2 > 0$ such that $\bar{\nu}_1 I \leqslant \bar{a}(x) \leqslant \bar{\nu}_2 I$. Let $\gamma_n = \inf_{|x| \geqslant n}[-V(x)]$. If $\gamma_n > 0$ for large n, $\lim_{n\to\infty} \gamma_n^{-1} \log n = 0$ and there exists a constant C such that $|V(x)|/\|\nabla V(x)\| \leqslant C|x|$ for large $|x|$, then we have

$$\alpha(L) \leqslant \frac{1}{2} \overline{\lim_{|x|\to\infty}} \big[- \langle \bar{a}(x)\nabla V(x), \nabla V(x)\rangle \nu(x)/V(x)\big].$$

We mention that by some slight modifications, Theorem 1.5 can be also extended to a class of manifolds whose volume grows no more faster than a polynomial of the diameter. Combining Corollary 1.4 with Theorem 1.5, we can get the exact value of $\alpha(L)$ for some particular operators, as illustrated below.

Corollary 1.6. Suppose that $\delta\sigma^2 \leqslant a(x) \leqslant \nu(x)\sigma^2$ for some constant $\delta > 0$, positive definite matrix σ and $\nu(x) \in C(\mathbb{R}^d)$ with $\lim_{|x|\to\infty} \nu(x) = \delta$. Take $V(x) = -b|x|^2/2$, $b > 0$. Then we have $\delta b \lambda_{\min}(\sigma)^2 \leqslant \alpha(L) \leqslant \delta b \lambda_{\max}(\sigma)^2$.

The lower bound here coincides with the one of $\lambda_1(L)$ given in [3]. Once σ has a unique eigenvalue, we obtain the exact $\alpha(L)$ for a large class of $a(x)$. To conclude this section, we discuss some examples.

Example 1.7. Consider the domain $[0, \infty)$ and take $a(x) \equiv 1$, $V(x) = -bx$ ($b > 0$). By Theorem 1.5, we have $\alpha(L) = 0$. This means that for the operator with constant diffusion coefficient, the L.S. inequality holds only when the potential V decays faster than linear. However, for this example, we have $\lambda_1(L) = b^2/4$ (see [3]).

Example 1.8[9]**.** Take $M = (0,\infty)$, $a(x) = x$ and $b(x) = -(x-b)$, $b > 0$. Applying Theorem 1.3 (1) to $g(d/dx, d/dx) = x^{-1}$, we get $\alpha(L) \geqslant 1/2$ whenever $b \geqslant 1/2$. In view of Theorem 1.5, this estimate is exact when $b \geqslant 1/2$.

It is interesting that for this example the Riemannian and the Euclidian metrics provide us respectively the sharp estimates of $\alpha(L)$ and $\lambda_1(L)$ ($= 1$ for all $b > 0$[3]), but not conversely.

Example 1.9. Take $\Omega = [a,b] \subset \mathbb{R}$. By setting $\beta = 0$ in Theorem 1.1, we obtain

$$\alpha_\Omega(V) \geqslant \lambda_1(0) e^{-\operatorname{osc}_\Omega(V)} = \frac{\pi^2}{(b-a)^2} e^{-\operatorname{osc}_\Omega(V)}. \qquad \text{(cf. [10])}$$

In particular, $\alpha_\Omega(0) = \lambda_1(0) = \pi^2/(b-a)^2$ since $\alpha_\Omega(0) \leqslant \lambda_1(0)$.

The next two examples illustrate that Corollary 1.4 does improve the B.-E. criterion.

Example 1.10. Take $d = 1$, $a(x) = (1+x^2)^2$ and $V(x) = -vx^2/2$, $v > 0$. Applying Corollary 1.4 to $\delta = 1$, $\lambda_V(x) \equiv -v$ and $\bar\beta(r) \equiv v$, we obtain $\alpha(L) \geqslant v$.

On the other hand, let $g(d/dx, d/dx) = (1+x^2)^{-2}$, then

$$L = \Delta_g + \nabla_g[-vx^2/2 + \log(1+x^2)] =: \Delta_g + \nabla_g \overline{V}.$$

We have

$$\operatorname{Hess}_{\overline{V}}\left((1+x^2)\frac{d}{dx}, (1+x^2)\frac{d}{dx}\right) = \left[(1+x^2)\frac{d}{dx}\right]^2 \overline{V} = -v(1+x^2)(1+3x^2) + 2(1+x^2).$$

Hence, the B.-E. criterion gives us

$$\alpha(L) \geqslant \inf_x \left[v(1+x^2)(1+3x^2) - 2(1+x^2)\right] = v - 2$$

provided $v \geqslant 1/2$, otherwise, the infimum is negative. Therefore, the criterion is available only if $v > 2$.

In contrast to Example 1.8, here the Euclidian metric produces a better estimate for $\alpha(L)$ rather than the Riemannian one.

Example 1.11. Take $a(x) \equiv I$ and $V(x) = -|x|^4 + v|x|^2$ ($v \geqslant 0$). We have

$$\partial^2 V/\partial x_i \partial x_j = -8x_i x_j + (2v - 4|x|^2)\delta_{ij}.$$

That is,

$$(\partial^2 V/\partial x_i \partial x_j) = -8xx^* + (2v - 4|x|^2)I.$$

For $p = 0$ we have $\beta(r) = 4r^2 - 2v$ if $d \geqslant 2$ and $\beta(r) = 12r^2 - 2v$ if $d = 1$. By Theorem 1.3 or Corollary 1.4, we get

$$\alpha(L) \geqslant \begin{cases} \dfrac{8}{3v + \sqrt{3(3v^2+8)}} \exp\left[-\dfrac{3v^2 + 4 + v\sqrt{3(3v^2+8)}}{8}\right], & \text{if } d \geqslant 2 \\ \dfrac{8}{v + \sqrt{v^2+8}} \exp\left[-\dfrac{v^2 + 4 + v\sqrt{v^2+8}}{8}\right], & \text{if } d = 1. \end{cases}$$

In particular, when $v = 0$, we have $\alpha(L) \geqslant 2\sqrt{2/3}\, e^{-1/2} > 0.99$ if $d \geqslant 2$ and $\alpha(L) \geqslant 2\sqrt{2}\, e^{-1/2} > 1.71$ if $d = 1$, which are better than the lower bound of the spectral gap given in [3]. When $d = 1$, the test function $f(x) = x$ gives us $\alpha(L) \leqslant \lambda_1(L) < 2.96$. However, the B.-E. criterion is not available for this example since $\beta(0) = -2v \leqslant 0$.

Example 1.12. Take $d = 1$, $a(x) \equiv 1$ and $V(x) = -x^2/2 + 2\sin x$. Then $\beta(r) \equiv -1$ and so Theorem 1.3 is not suitable. However, applying (1.2) to $V(x) = -x^2/2$ and $V(x) - U(x) = 2\sin x$, we have $\alpha(L) \geqslant e^{-2}$. This means that the condition "$\sup_{r\geqslant 0}\beta(r)$" is still not necessary for the L. S. inequality and a bounded perturbation should be carefully treated before applying Theorem 1.3.

2. Proofs

Proof of Theorem 1.1. Set $\rho(x) = \rho(x, p)$. For $x \in \Omega$, let $\gamma:\ [0, \rho(x)] \to \widetilde{\Omega}$ be the unique shortest geodesic from p to x. Let M_k be a simply connected d-dimensional manifold with constant sectional curvature k. Choose $\tilde{p}$ and $\tilde{x} \in M_k$ such that $\tilde{\rho}(\tilde{p}, \tilde{x}) = \rho(x)$. By assumption, either $k \leqslant 0$ or $k > 0$ but still $2\sqrt{k}\rho(x) < \pi$, we have $\tilde{x} \notin C(\tilde{p})$. For $X \in T_xM$ with $\|X\| = 1$, take $\widetilde{X} \in T_{\tilde{x}}M_k$ so that $\|\widetilde{X}\| = 1$ and $X\rho(x) = \widetilde{X}\tilde{\rho}(\tilde{p}, \cdot)(\tilde{x})$. By Hessian comparison theorem[6], [12], we have

$$\mathrm{Hess}_\rho(X, X) \geqslant \mathrm{Hess}_{\tilde{\rho}(\tilde{p},\cdot)}(\widetilde{X}, \widetilde{X}) = (f'/f)(\rho(x))\big(1 - (X\rho(x))^2\big), \tag{2.1}$$

where

$$f(r) = \begin{cases} r, & \text{if } k = 0 \\ \sinh\big(\sqrt{-k}\, r\big)/\sqrt{-k}, & \text{if } k < 0 \\ \sin\big(\sqrt{k}\, r\big)/\sqrt{k}, & \text{if } k \in \big(0, \pi/\big(2\sqrt{k}\,\big)\big). \end{cases} \tag{2.2}$$

For $x \in \Omega$ and $X \in T_xM$ with $\|X\| = 1$, since $(X\rho)^2 \leqslant \|X\|^2 = 1$, by (2.1), we have

$$\begin{aligned} \mathrm{Hess}_{\rho^2}(X, X) &= 2\rho\mathrm{Hess}_\rho(X, X) + 2(X\rho)^2 \\ &\geqslant \begin{cases} 2, & \text{if } k \leqslant 0 \\ 2\sqrt{k}D\,\mathrm{ctan}\big(\sqrt{k}D\big), & \text{if } k > 0. \end{cases} \end{aligned}$$

Therefore $K_\Omega\big(-\beta\rho^2\big) \geqslant \alpha_\beta$. By (1.3), we get

$$\alpha_\Omega(-\beta\rho^2) \geqslant \alpha_\beta + d^{-1}\lambda_1(0)e^{-\mathrm{osc}_\Omega(-\beta\rho^2)} = \alpha_\beta + d^{-1}\lambda_1(0)e^{-\beta D^2}.$$

Now, Theorem 1.1 follows from (1.2). □

Proof of Corollary 1.2. Note that $\mathrm{osc}_\Omega(V + \beta\rho^2) \leqslant \mathrm{osc}_\Omega(V) + \beta D^2$, by Theorem 1.1, we have

$$\alpha_\Omega(V) \geqslant e^{-\mathrm{osc}_\Omega(V)} \sup_{\beta>0} e^{-\beta D^2}\Big[\alpha_\beta + d^{-1}\lambda_1(0)e^{-\beta D^2}\Big].$$

Then, the desired estimates are obtained by choosing

$$\beta = \begin{cases} \dfrac{1}{D^2} - \dfrac{K}{2}, & \text{if } k \leqslant 0 \\ \dfrac{1}{D^2} - \dfrac{K\tan\left(\sqrt{k}D\right)}{2\sqrt{k}D}, & \text{if } k > 0. \end{cases}$$

Here we have used the condition that $\sqrt{k}/\tan\left(\sqrt{k}D\right) > KD/2$. $\square$

Proof of Theorem 1.3. (1) First, we prove part (1) of Theorem 1.3.

a) Let $\sup_{r\geqslant 0}\beta(r) > 0$. Then, we have $\beta(0) > -\infty$. Since $k = 0$ and $f(r) = r$, from (1.4), it follows that $\gamma(r) = \frac{1}{r}\int_0^r \beta(s)ds$, $r > 0$, $\gamma(0) = \beta(0)$ and

$$C_a(r) = [\gamma(a) - \beta(r)]I_{[r\leqslant a]}, \quad a \geqslant 0, \qquad F_a(r) = \int_0^r ds \int_0^s C_a(u)du, \qquad r \geqslant 0.$$

Note that $\beta(r)$ is increasing in r and so is $\gamma(r)$. Next, let $G(a) = \gamma(a)\exp[-F_a(a)]$ for simplicity. We will prove the following two assertions:

$$\alpha(V) \geqslant \sup_{a\geqslant 0} G(a). \tag{2.3}$$

and

$$\sup_{a\geqslant 0} G(a) = G(a_0) \tag{2.4}$$

where $a_0 > 0$ is determined uniquely by the equation $\int_0^a \beta(r)dr = 2/a$. These assertions certainly imply the statement of Theorem 1.3: $\alpha(V) \geqslant G(a_0)$. We now prove the second assertion. Note that

$$\begin{aligned} F_a(a) &= \int_0^a dr \int_0^r [\gamma(a) - \beta(s)]ds \\ &= \frac{a^2}{2}\gamma(a) - \int_0^a dr \int_0^r \beta(s)ds \\ &= \frac{a^2}{2}\gamma(a) - \int_0^a \gamma(r)d\left(\frac{r^2}{2}\right) \\ &= \frac{1}{2}\int_0^a r^2\gamma'(r)dr. \end{aligned}$$

Hence $G'(a) = \gamma'(a)\left[1 - a^2\gamma(a)/2\right]\exp[-F_a(a)]$. Because $\gamma' \geqslant 0$ and the uniqueness of a_0, we have $G' \geqslant 0$ on $[0, a_0]$ and $G' \leqslant 0$ on $[a_0, \infty)$. Thus, the global maximum of G is achieved at a_0. This proves (2.4). Next, since $a_0^2\gamma(a_0) = 2$, we have

$$\begin{aligned} F_{a_0}(a_0) &= \frac{a_0^2}{2}\gamma(a_0) - \int_0^{a_0} dr \int_0^r \beta(s)ds \\ &= 1 - \int_0^{a_0} (a_0 - s)\beta(s)ds \\ &= 1 - a_0^2\gamma(a_0) + \int_0^{a_0} r\beta(r)dr \\ &= -1 + \int_0^{a_0} r\beta(r)dr. \end{aligned}$$

Thus, $G(a_0)$ coincides with the lower bound given in (1.5). Moreover,

$$G(a_0) = \sup_{a\geqslant 0} G(a) \geqslant G(0) = \beta(0), \tag{2.5}$$

which was mentioned in the last section.

b) We now begin to prove (2.3). Since we always have $\alpha(V) \geqslant 0$, (2.3) is meaningful iff $\sup_{a\geqslant 0}\gamma(a) > 0$ (equivalently, $\sup_{r\geqslant 0}\beta(r) > 0$). Thus, by a), we need only to show that $\alpha(V) \geqslant G(a_0)$. But the proof given below makes no difference if we replace a_0 with any fixed $a > 0$. Because

$$F_a'(r) = \int_0^r C_a(u)du = r\bigg\{\frac{1}{a}\int_0^a \beta(u)du - \frac{1}{r}\int_0^r \beta(u)du\bigg\}, \qquad r < a,$$

we see that $F_a'(r) \geqslant 0$ if $r < a$ and $F_a'(r) = 0$ if $r \geqslant a$. Hence,

$$\operatorname{osc}(F_a) = \sup F_a - \inf F_a = \sup F_a = F_a(a).$$

c) Next, since $C_a(a) = \gamma(a) - \beta(a) \leqslant 0$, C_a may not be continuous at a. For this, we need a modification of C_a. Let $\varepsilon \in (0, a)$ and define

$$C_a^\varepsilon(r) = \begin{cases} C_a(r) - C_a(a)\dfrac{\varepsilon - r}{\varepsilon}, & \text{if } \ r \in [0, \varepsilon] \\ C_a(a)\bigg(1 - \dfrac{r-a}{\varepsilon}\bigg), & \text{if } \ r \in [a, a+\varepsilon] \\ C_a(r), & \text{otherwise,} \end{cases}$$

$$F_a^\varepsilon(r) = \int_0^r ds \int_0^s C_a^\varepsilon(u)du.$$

Then $C_a^\varepsilon \in C(\mathbb{R}_+)$ and $F_a^\varepsilon \in C^2(\mathbb{R}_+)$. Moreover, it is not difficult to check that $\big(F_a^\varepsilon\big)' \geqslant 0$, $\big(F_a^\varepsilon\big)'(r) = 0$ for all $r \geqslant a + \varepsilon$ and $C_a^\varepsilon(r) - \frac{1}{r}\int_0^r C_a^\varepsilon(u)du \leqslant 0$ (Note that $\int_0^a C_a(r)dr = 0$). Hence $\operatorname{osc}(F_a^\varepsilon) = \sup F_a^\varepsilon = F_a^\varepsilon(a+\varepsilon) \to F_a(a)$ as $\varepsilon \to 0$.

d) Take $V_\varepsilon(x) = F_a^\varepsilon(\rho(x))$, where $\rho(x) = \rho(p, x)$. Then $\operatorname{osc}(V_\varepsilon) = F_a^\varepsilon(a+\varepsilon)$. On the other hand, for $x \in M$ and $X \in T_xM$ with $\|X\| = 1$, by (2.1), we have

$$\begin{aligned} \operatorname{Hess}_{V_\varepsilon}(X, X) &= (F_a^\varepsilon)'(\rho)\operatorname{Hess}_\rho(X, X) + (F_a^\varepsilon)''(\rho)(X\rho)^2 \\ &\geqslant \frac{1}{\rho}\int_0^\rho C_a^\varepsilon(u)du + \bigg[C_a^\varepsilon(\rho) - \frac{1}{\rho}\int_0^\rho C_a^\varepsilon(u)du\bigg](X\rho)^2 \\ &\geqslant C_a^\varepsilon(\rho). \end{aligned}$$

Here in the last step, we have used the fact that $(X\rho)^2 \leqslant \|X\|^2 = 1$ and $C_a^\varepsilon(\rho) - \frac{1}{\rho}\int_0^\rho C_a^\varepsilon(u)du \leqslant 0$. Therefore,

$$\inf_{x\in M} K(V - V_\varepsilon, x) \geqslant \inf_{r\geqslant 0}\{C_a^\varepsilon(r) + \beta(r)\} \geqslant \gamma(a).$$

By the B.-E. criterion and (1.2) we obtain $\alpha(V) \geqslant \gamma(a)\exp[-F_a^\varepsilon(a+\varepsilon)]$. Then (2.3) follows by letting $\varepsilon \to 0$.

(2) The proof of part (2) of Theorem 1.3 is similar. Recall that the functions γ and F_a are given by (1.4) with $f(r) = \sin\big(\sqrt{k}\,r\big)/\sqrt{k}$. By using the smoothing approximation as in the proof c) above, we may and will assume that F_a is a C^2-function. Next, for $\rho(x) < a$, we have

$$F_a''(\rho) = f'(\rho)\big[\gamma(a) - \tilde{\beta} \circ f(\rho)\big] \geqslant f'(a)\gamma(a) - \beta(\rho).$$

Thus, as we did in proof d),

$$\mathrm{Hess}_{F_a(\rho)}(X, X) = F_a'(\rho)\mathrm{Hess}_\rho(X, X) + F_a''(\rho)(X\rho)^2 \geqslant f'(a)\gamma(a) - \beta(\rho).$$

Therefore, $K(V - F_a(\rho), x) \geqslant f'(a)\gamma(a)$ for $\rho(x) < a$. On the other hand, since $F_a(\rho) = F_a(a)$ for all $\rho > a$, we have

$$K(V - F_a(\rho), x) = K(V, x) \geqslant \beta(a) \geqslant f'(a)\tilde{\beta} \circ f(a) \geqslant f'(a)\gamma(a)$$

for $\rho(x) \geqslant a$. Now, the desired conclusion follows from the B.-E. criterion and (1.2). $\square$

In view of the proofs of Theorems 1.1 and 1.3, one may expect some further improvement. For instance, one may take $-k$ into account when $k < 0$. In part (1) of Theorem 1.3, one may use $h \circ \rho$ instead of ρ for some suitable function h. However, on the one hand, we restrict ourselves to general and computable estimation. Based on this and also from the geometric point of view, our perturbing potentials are more or less natural. On the other hand, we have tried several different potentials, including the above suggestions, but none of them ever produces a better estimate.

Proof of Corollary 1.4. Consider the operator

$$\overline{L} = \sum_{i,j=1}^{d} (\sigma^2)_{ij}\big[\frac{\partial^2}{\partial x_i \partial x_j} + \frac{\partial V}{\partial x_j}\frac{\partial}{\partial x_i}\big]$$

in $\mathbb{R}^d$. By (1.6), we have

$$\alpha(L) \geqslant \alpha(\delta\overline{L}) = \delta\alpha(\overline{L}). \tag{2.6}$$

On the other hand, under the Riemannian metric $g(\partial/\partial x_i, \partial/\partial x_j) = (\sigma^2)_{ij}^{-1}$, we have $\overline{L} = \Delta_g + \nabla_g V$ (see [3]). For $x \in \mathbb{R}^d$ and $X \in T_x\mathbb{R}^d$ with $g(X, X) = 1$, there exists $c \in \mathbb{R}^d$ such that $X = \sum_i^d c_i \partial/\partial x_i$ and $c^*(\sigma^{-1})^2 c = 1$. Then

$$\mathrm{Hess}_V(X, X) = \sum_{i,j=1}^{d} c_i c_j \frac{\partial^2 V}{\partial x_i \partial x_j} = (\sigma^{-1}c)^*\left[\sigma\left(\frac{\partial^2 V}{\partial x_i \partial x_j}\right)\sigma\right](\sigma^{-1}c) \leqslant \lambda_V(x).$$

Hence $K(V, x) = -\lambda_V(x)$ and so Corollary 1.4 follows from (2.6) and Theorem 1.3 (1). $\square$

Proof of Theorem 1.5. a) As usual, one uses the Riemannian metric

$$g(\partial/\partial x_i, \partial/\partial x_j) = \bar{a}(x)^{-1}$$

instead of the Euclidean one I. Note that the induced Riemannian distance is indeed equivalent to the Euclidean one since $\bar{\nu}_1 I \leqslant \bar{a}(x) \leqslant \bar{\nu}_2 I$. Thus, without loss of generality, we may and will assume that $\bar{a}(x) \equiv I$.

b) Given $g \in C^1(\mathbb{R}^d)$ with compact support, let $f = ge^{u/2}$, where $u = -V + \log Z$. By (1.6), we have

$$\begin{aligned}\int g^2 \log\big(g^2 e^u\big)dx - \bigg(\int g^2 dx\bigg) \log\bigg(\int g^2 dx\bigg) \\ \leqslant \frac{2}{\alpha(L)} \int \nu\Big(\|\nabla g\|^2 + \frac{1}{4} g^2 \|\nabla V\|^2 + g\|\nabla g\|\,\|\nabla V\|\Big)dx.\end{aligned}$$

Equivalently,

$$\begin{aligned}\int g^2 \log g^2 dx - \bigg(\int g^2 dx\bigg) \log\bigg(\int g^2 dx\bigg) - \frac{2}{\alpha(L)} \int \nu\big(g\|\nabla g\|\,\|\nabla V\| + \|\nabla g\|^2\big)dx \\ \leqslant \int u g^2 \bigg(\frac{\|\nabla V\|^2 \nu}{2\alpha(L)u} - 1\bigg)dx. \qquad (2.7)\end{aligned}$$

c) To prove the assertion, it suffices to construct a sequence $g_n \in C^1(\mathbb{R}^d)$ with compact support such that $\int u g_n^2 = 1$ and moreover the left side of (2.7) goes to zero as $n \to \infty$. To see this, assume that

$$\frac{1}{2} \varlimsup_{|x|\to\infty} \big[- \|\nabla V(x)\|^2 \nu(x)/V(x)\big] =: A < \infty.$$

Then in the limit (2.7) yields $0 \leqslant \alpha(L)^{-1} A - 1$. The construction given below is a slight modification from [9]. Choose a non-negative $h \in C^1(\mathbb{R})$ with support $[0,1]$, $\int_0^1 h(s)^2 ds = 1$ and $\inf\{h(s) : \ s \in [0.1, 0.9]\} = 1$. Define

$$\ell_n = \int_{\{n \leqslant |x| \leqslant 2n\}} h\bigg(\frac{|x| - n}{n}\bigg)dx, \qquad g_n(x) = \frac{1}{\sqrt{\ell_n u(x)}} h\bigg(\frac{|x| - n}{n}\bigg).$$

Then g_n is well defined for large n and has support $\{x : n \leqslant |x| \leqslant 2n\}$.

d) Let $\bar{\gamma}_n = \gamma_n + \log Z$, then for large n and $|x| \geqslant n$ we have

$$\|g_n\|_\infty \leqslant \frac{1}{\sqrt{\ell_n \bar{\gamma}_n}} \|h\|_\infty, \qquad \|\nabla g_n\| \leqslant \frac{\|\nabla h\|_\infty}{n\sqrt{\ell_n \bar{\gamma}_n}} + \frac{\|\nabla u\|\,\|h\|_\infty}{2u\sqrt{\ell_n \bar{\gamma}_n}}. \qquad (2.8)$$

On the other hand,

$$\|\nabla u(x)\|\nu(x) \leqslant 3AC|x| \qquad \text{and} \qquad - \frac{\|\nabla u(x)\|^2 \nu(x)}{V(x)} \leqslant 3A \qquad (2.9)$$

for large $|x|$. So for large n,

$$\begin{aligned}\|\nabla u\|\,\|\nabla g_n\|\nu g_n &\leqslant \left\{\|h\|_\infty\|\nabla h\|_\infty\frac{3AC|x|}{n\ell_n\bar\gamma_n}+\frac{3A\|h\|_\infty^2}{2\ell_n\bar\gamma_n}\right\}I_{\{n\leqslant|x|\leqslant 2n\}}\\ &\leqslant\frac{C_1}{\ell_n\bar\gamma_n}I_{\{n\leqslant|x|\leqslant 2n\}}\end{aligned}$$

for some constant $C_1>0$. Note that

$$\ell_n\geqslant\int_{\{1.1n\leqslant|x|\leqslant 1.9n\}}dx\geqslant C_2\int_{\{n\leqslant|x|\leqslant 2n\}}dx$$

for some constant $C_2>0$ (Here is the main place in which the restriction on the growth of the volume is required). We obtain

$$\lim_{n\to\infty}\int\|\nabla u\|\,\|\nabla g_n\|\nu g_n dx\leqslant\lim_{n\to\infty}\frac{C_1}{\bar\gamma_n C_2}=0. \tag{2.10}$$

Next, by the second inequality of (2.9) and the assumption, we have

$$\nu(x)\leqslant\frac{3A}{-V}\left(\frac{|V|}{\|\nabla u\|}\right)^2\leqslant\frac{3AC^2|x|^2}{-V}\leqslant\frac{4AC^2|x|^2}{\bar\gamma_n}$$

for $|x|\in[n,2n]$ and large n. Moreover,

$$\frac{\nu\|\nabla u\|^2}{u^2\ell_n\bar\gamma_n}=\frac{\|\nabla u\|^2\nu}{|u|}\cdot\frac{1}{|u|\ell_n\bar\gamma_n}\leqslant\frac{3A}{\ell_n\bar\gamma_n^2}.$$

Combining these two estimates with the second inequality of (2.8), we obtain

$$\lim_{n\to\infty}\int\nu\|\nabla g_n\|^2dx=0. \tag{2.11}$$

e) Since $g_n^2\leqslant\|h\|_\infty(\ell_n\bar\gamma_n)^{-1}I_{\{n\leqslant|x|\leqslant 2n\}}$, we have

$$\int g_n^2\leqslant\|h\|_\infty/(C_2\bar\gamma_n)\to 0\quad\text{ as }\quad n\to\infty$$

and so

$$\lim_{n\to\infty}\left(\int g_n^2dx\right)\log\left(\int g_n^2dx\right)=0 \tag{2.12}$$

Finally, noticing that $g_n^2\leqslant\|h\|_\infty/(\ell_n\bar\gamma_n)<e^{-1}$ for large n, $|x\log x|$ is increasing in $(0,e^{-1})$ and $\ell_n\leqslant C_3n^d$ for some $C_3>0$, we have

$$\begin{aligned}\left|\int g_n^2\log g_n^2dx\right|&\leqslant\int_{\{n\leqslant|x|\leqslant 2n\}}\frac{\|h\|_\infty}{\ell_n\bar\gamma_n}\left|\log\frac{\|h\|_\infty}{\ell_n\bar\gamma_n}\right|dx\\ &\leqslant\left|\frac{\|h\|_\infty}{C_2\bar\gamma_n}\log\frac{\|h\|_\infty}{\bar\gamma_n}\right|+\frac{\|h\|_\infty}{C_2\bar\gamma_n}\big(d\log n+\log C_3\big)\end{aligned}$$

which goes to zero as $n\to\infty$. Combining this with (2.10)—(2.12), the assertion follows from (1.6) by letting $n\to\infty$. □

Acknowledgements. The authors are greatly indebted to Prof. L. Gross for his advice and encouragement and to a referee, whose comments not only helped to improve the quality of the paper but also led to part (2) of Theorem 1.3.

References

[1] Bakry, D. (1992), *L'hypercontractivité et son utilisation en théorie des semigroupes*, LNM, **1581** Springer.
[2] Bakry, D. and Emery, M. (1984), *Hypercontractivité de semigroups de diffusion*, C. R. A. S. Paris, Seŕie 1, 209(15), 775–778.
[3] Chen, M. F. and Wang, F. Y. (1993), *Estimation of the first eigenvalue of second order elliptic operators*, J. Funct. Anal. 131:2, 345–363.
[4] Deuschel, J.-D. and Stroock, D. W. (1989), *Large Deviations*, Academic Press.
[5] Deuschel, J.-D. and Stroock, D. W. (1990), *Hypercontractivity and spectral gap of symmetric diffusion with applications to the stochastic Ising models*, J. Funct. Anal. 92, 30–48.
[6] Greene, R. E. and Wu, H. (1979), *Function Theory on Manifolds which Posses a Pole*, Springer-Verlag, LNM. vol.699.
[7] Gross, L. (1976), *Logarithmic Sobolev inequalities*, Amer. J. Math. 97, 1061–1083.
[8] Gross, L. (1993), *Logarithmic Sobolev inequalities and contractivity of semigroups*, LNM. **1563**.
[9] Korzeniowski, A. (1987), *On logarithmic Sobolev constant for diffusion semigroups*, J. Funct. Anal. 71, 363–370.
[10] Wang, F. Y. (1994a), *Application of coupling method to the Neumann eigenvalue problem*, Prob. Th. Rel. Fields 98, 299–306.
[11] Wang, F. Y. (1994b), *On estimation of logarithmic Sobolev constant* (*In Chinese*), J. Beijing Normal Univ. 30:4, 448–452.
[12] Wu, H., Sheng, Y. L. and Yu, Y. L. (1989), *Preliminary of Riemannian Geometry*, Beijing Univ. Press.

Department of Mathematics, Beijing Normal University, Beijing 100875, The People's Republic of China.

Acta Math. Sin. New Ser. 12:4 (1996), pp.337–360.

ESTIMATION OF SPECTRAL GAP FOR MARKOV CHAINS

Mu-Fa Chen

(Beijing Normal University)
December 20, 1995

Abstract. The study of the convergent rate (spectral gap) in the L^2-sense is motivated from several different fields: probability, statistics, mathematical physics, computer science and so on and it is now an active research topic. Based on a new approach (the coupling technique) introduced in [7] for the estimate of the convergent rate and as a continuation of [4], [5], [7]—[9], [23] and [24], this paper studies the estimate of the rate for time-continuous Markov chains. Two variational formulas for the rate are presented here for the first time for birth-death processes. For diffusions, similar results are presented in an accompany paper [10]. The new formulas enable us to recover or improve the main known results. The connection between the sharp estimate and the corresponding eigenfunction is explored and illustrated by various examples. A previous result on optimal Markovian couplings[4] is also extended in the paper.

1. Introduction. One Main Result and Examples.

Let E be a countable set. Consider a reversible Markov chain $P(t) = (p_{ij}(t) : i, j \in E)$ with regular and irreducible Q-matrix $Q = (q_{ij})$. The reversible (probability) measure is denoted by (π_i). The purpose of the paper is to study the exponential convergence:

$$\|P(t)f - \pi(f)\| \leqslant \|f - \pi(f)\| \, e^{-\varepsilon t}$$

for all $t \geqslant 0$ and $f \in L^2(\pi)$, where $\|\cdot\|$ denotes the L^2-norm and $\pi(f) = \sum_i f_i \pi_i$. It is known that the maximal exponential rate $\varepsilon_{\max}$ is given by the spectral gap:

$$\text{gap}\,(D) := \inf\{D(f, f) : \pi(f) = 0 \text{ and } \|f\| = 1\}, \tag{1.1}$$

2000 *Mathematics Subject Classification.* 60J27, 60J80.

Key words and phrases. Markov chains, spectral gap, couplings.

Research supported in part by NSFC, Qiu Shi Sci. & Tech. Foundation and the State Education Commission of China

where $D(f,f)$ is the Dirichlet form

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

with domain

$$\mathscr{D}(D)=\{f\in L^2(\pi): D(f,f)<\infty\}.$$

Actually, the spectral gap is nothing but the first (non-trivial) eigenvalue λ_1 of the generator in the L^2-space. Refer to [16] and [2], or [3; Chapter 9]. Clearly, the variational formula (1.1) is very useful for a upper bound but it is much more difficult to handle the lower bound for which various approaches have been developed (see for instance [2], [12], [14]–[19]).

Recall that there is a well-known topic in the study of Markov chains. That is the exponential ergodicity:

$$|p_{ij}(t)-\pi_j|=O(\exp[-\hat{\alpha}t])\quad\text{ as } t\to\infty$$

for some constant $\hat{\alpha}$ (maximal) >0. Refer to van Doorn [21,22] and Zeifman [25] (see also [3; Chapter 9]) for related results and references. It is interesting that these two convergences are indeed coincides each other for birth-death processes and moreover $\lambda_1=\hat{\alpha}^{[2]}$. From this point of view, the study of the spectral gap has much longer history. Due to this relation, on the one hand, we obtain some examples given below for which the spectral gaps are explicitly known and on the other hand, this paper presents a new approach to estimate the rate of the exponential ergodicity for birth-death processes.

The estimate of spectral gap has a very wide range of applications. A fashionable application of the topic is the Markov chains Monte Carlo. Next, the presence or absence of the spectral gap provide us a way to describe the phase transitions (cf. Sokal and Thomas (1988) and Liggett (1989), for example). Among other applications, we mention that the spectral gap is used by Aldous and Brown (1993), Iscoe and McDonald (1994) to study the asymptotics of the exit times, by Deuschel and Stroock (1990) and Chen and Wang (1994) to estimate the logarithmic Sobolev constant and by Jerrum and Sinclair (1989) to study the randomized approximation algorithms. See for instance the survey article [4] for more information about the backgrounds and for more references.

To see the difficulty of the problem, let us look at some examples. Consider the birth-death process with birth rate $b_i=i+1$ and death rate $a_i=2i$. Then $\lambda_1=1$ and the corresponding eigenfunction is linear. We now keep b_i to be the same but add a constant to a_i, i.e., $a_i=2i+\varepsilon\,(\varepsilon>1)$. Then $\lambda_1=2$ and the eigenfunction becomes quadratic. Next, consider the nearly trivial case that the state space consists of three points, $E=\{0,1,2\}$, so we have four parameters b_0,b_1 and a_1,a_2 only. Then

$$\lambda_1=2^{-1}\big[a_1+a_2+b_0+b_1-\sqrt{(a_1-a_2+b_0-b_1)^2+4a_1b_1}\,\big].$$

Even in such a simple situation, the role played by the parameters for λ_1 is still not so obvious. From these, one sees the complexity of the problem and the sensitivity of λ_1.

In order to justify the power of our approach and to compare with the previous results, we now discuss one of the main results of the paper. 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 the following:

$$\pi_i = \frac{\mu_i}{\mu}, \qquad \mu_0 = 1, \quad \mu_i = \frac{b_0 b_1 \cdots b_{i-1}}{a_1 a_2 \cdots a_i},\ i \geqslant 1, \qquad \mu = \sum_i \mu_i.$$

Let $\mathscr{V}$ be the set of all positive sequences $(v_i : i \geqslant 0)$ and define

$$\begin{aligned} R_i(v) &= a_{i+1} + b_i - a_i/v_{i-1} - b_{i+1}v_i \\ &= \Delta a(i) - \Delta b(i) + a_i[1 - v_{i-1}^{-1}] + b_{i+1}[1 - v_i], \\ &\qquad a_0 := 0, \quad v_{-1} := 1, \quad i \geqslant 0, \end{aligned} \tag{1.2}$$

where $\Delta a(i) = a_{i+1} - a_i$. Next, let $\mathscr{W} \subset L^1(\pi)$ be the set of all strictly increasing sequences $(w_i : i \geqslant 1)$ with $\sum_{i\geqslant 1} \mu_i w_i > 0$. Define[1]

$$\begin{aligned} I_i(w) &= b_i\mu_i(w_{i+1} - w_i) \Big/ \sum_{j=i+1}^{\infty} \mu_j w_j, \quad i \geqslant 1, \\ I_0(w) &= b_0\left(1 + w_1 \Big/ \sum_{j=1}^{\infty} \mu_j w_j\right). \end{aligned} \tag{1.3}$$

Theorem 1.1. Consider the birth-death process as above. We have

$$\operatorname{gap}(D) = \sup_{v\in\mathscr{V}} \inf_{i\geqslant 0} R_i(v). \tag{1.4}$$

$$\operatorname{gap}(D) = \sup_{w\in\mathscr{W}} \inf_{i\geqslant 0} I_i(w). \tag{1.5}$$

Moreover, the supremum in both (1.4) and (1.5) can be attained.

Clearly, for each test sequence $v \in \mathscr{V}$ (resp. $w \in \mathscr{W}$), one obtains from (1.4) (resp. (1.5)) a lower bound of $\operatorname{gap}(D)$. Thus, (1.4) and (1.5) are dual variational formulas of (1.1). In view of (1.2) and (1.3), one sees that the differential form (1.4) and the summation form (1.5) are quite different but there is indeed a correspondence between (v_i) and (w_i) (Lemma 2.1). As we will see soon each of them has its own advantage.

By using Theorem 1.1, it is rather easy to prove the following corollaries which contain the main known results. For instance, part (1) below is deduced from (1.4) directly by setting $v_i = \sqrt{a_{i+1}/b_{i+1}}$. The other parts will be proved in Section 2.

[1] If $\sum_{j\geqslant 0} \mu_j w_j = 0$, i.e., $w_0 = -\sum_{j\geqslant 1} \mu_j w_j = 0$, then the second formula below can be included into the first one.

Corollary 1.2. (1) (Van Doorn (1985)).

$$\operatorname{gap}(D) \geqslant \inf_{i\geqslant 0}\big\{a_{i+1}+b_i-\sqrt{a_ib_i}-\sqrt{a_{i+1}b_{i+1}}\big\}.$$

(2) (Sullivan (1985)). If $\sum_{j\geqslant i}\mu_j \leqslant c_1\mu_i$ for all $i\geqslant 1$ and $\mu_{i+1}\leqslant c_2\mu_i b_i$ for all $i\geqslant 0$, then

$$\operatorname{gap}(D)\geqslant \big(\sqrt{c_1}-\sqrt{c_1-1}\big)^2/(c_1c_2)\geqslant 1/\big(4c_1^2c_2\big).$$

(3) (Liggett (1989)). If $\sum_{j\geqslant i}\mu_j\leqslant c_1\mu_i a_i$ and $\sum_{j\geqslant i}\mu_j a_j\leqslant c_2\mu_i a_i$ for all $i\geqslant 1$, then

$$\operatorname{gap}(D)\geqslant \big(\sqrt{c_2}-\sqrt{c_2-1}\big)^2/c_1\geqslant 1/(4c_1c_2).$$

(4) If $a_i=b_i$ and $i\sum_{j\geqslant i}1/a_j\leqslant c$ for all $i\geqslant 1$, then $\operatorname{gap}(D)\geqslant 1/(4c)$.

Applying (1.4) to some typical (v_i), the corresponding lower bounds are given as follows.

Corollary 1.3. (1) $v_i=r[1+1/(i+c)]$, $r\geqslant 1$, $c\in[0,\infty]$.

$$\begin{aligned}\operatorname{gap}(D)&\geqslant \inf_{i\geqslant 0}\left\{a_{i+1}+b_i-\frac{a_i}{r}\left[1-\frac{1}{i+c}\right]-b_{i+1}r\left[1+\frac{1}{i+c}\right]\right\}\\ &=\begin{cases}\inf_{i\geqslant 0}\left\{a_{i+1}+b_i-\dfrac{a_i}{r}-b_{i+1}r\right\}, & \text{if } c=\infty \quad (\text{i.e. } v_i\equiv r)\\ \inf_{i\geqslant 0}\left\{a_{i+1}+b_i-\dfrac{1}{i+1}\left[\dfrac{ia_i}{r}+(i+2)b_{i+1}r\right]\right\}, & \text{if } c=1\\ \inf_{i\geqslant 0}\left\{\Delta a(i)-\Delta b(i)+\dfrac{1}{i+c}[a_i-b_{i+1}]\right\}, & \text{if } r=1.\end{cases}\end{aligned}$$

(2) $v_i=1-c_1/(i+c_2)$, $c_2>0$, $c_1\in(0,c_2)$.

$$\operatorname{gap}(D)\geqslant \inf_{i\geqslant 0}\left\{\Delta a(i)-\Delta b(i)-c_1\left[\frac{a_i}{i-1+c_2-c_1}-\frac{b_{i+1}}{i+c_2}\right]\right\}.$$

Most the results in the paper are meaningful for the finite state space $E_n=\{0,1,\cdots,n\}$ with reflection boundary. For instance, for (1.4), we need only to consider the quantities $R_i(v)$ up to $n-1$. Let us return to the case that $n=2$. By choosing

$$v_0=(2b_1)^{-1}\big[a_1-a_2+b_0-b_1+\sqrt{(a_1-a_2+b_0-b_1)^2+4a_1b_1}\,\big],$$

it follows that $R_0(v)=R_1(v)=\lambda_1$ and so by (1.4) our estimate is sharp. The same bound can be achieved by [16] or [25] but not others. Here and in what follows we compare our estimates with those given in [12], [14]—[17], [19] and [25]. Most of the papers, except [16] and [20,21], deal with bounded operators only but some of them allow the state space to be general (i.e., studying jump processes rather than Markov chains). The result of [19] on Markov chains was extended in [2,3] to unbounded operators with different proof.

Before moving further, let us look at four examples which are standard in queue theory and the spectral gaps for the first three of them are explicitly known[21],[22].

b_i	a_i	λ_1	v_i
b	$a\,(a>b)$	$\big(\sqrt{a}-\sqrt{b}\,\big)^2$	$\sqrt{a/b}$
$\beta_0+\beta_1 i$	$\delta_1 i$	$\delta_1-\beta_1$	1
$\dfrac{b}{i+1}$	a	$a-\dfrac{\sqrt{b^2+4ab}-b}{2b}$	$\dfrac{\big(\sqrt{b^2+4ab}+b\big)(i+2)}{2b(i+1)}$
b	$a\,(i\wedge k)$	$\big(\sqrt{ak}-\sqrt{b}\,\big)^2$	$\sqrt{ak/b}$

By part (1) of Corollary 1.3, we see that our estimates are sharp for all these examples, but for the last one, one needs a restriction

$$1<\sqrt{a/b}\leqslant\sqrt{k}/(k-1)\quad(k\geqslant 2).$$

The estimates given in [16], [19], [25] are all sharp for the first example (but not for the others) and [12], [14] and [17] are not suitable for the first example. We now return to the last example and let $k\geqslant 5$. If

$$\sqrt{k}/(k-1)\leqslant\sqrt{a/b}\leqslant\frac{k}{2}\bigg(1+\sqrt{\frac{k-5}{k-1}}\,\bigg),$$

take $v_i\equiv k/(k-1)$. We get $\operatorname{gap}(D)\geqslant a-b/(k-1)$. These two estimates are the same as in [25]. Finally, if

$$\sqrt{a/b}\geqslant\frac{k}{2}\bigg(1+\sqrt{\frac{k-5}{k-1}}\,\bigg),$$

take $v_i\equiv\sqrt{a/b}$. Note that in general, if $R_0\wedge R_1<\inf_{i\geqslant 2}R_i$ and $R_0\neq R_1$ (here we have ignored v from $R(v)$), one may improve the estimate by replacing the original v_0 with $v_0=(2b_1)^{-1}\big\{\sqrt{\Gamma^2+4a_1b_1}+\Gamma\big\}$, where $\Gamma=a_1+b_0-a_2-b_1+b_2v_1$. Then, for this new sequence (v_i), we have

$$\begin{aligned}R_i&=\Delta a(i)-\Delta b(i)+a_i[1-v_{i-1}^{-1}]+b_{i+1}[1-v_i],\qquad i\geqslant 2\\ R_0&=R_1=2^{-1}\big[a_1+a_2+b_0+b_1-b_2v_1-\sqrt{\Gamma^2+4a_1b_1}\,\big].\end{aligned}\tag{1.6}$$

By using (1.6), we obtain for this example in the last situation that

$$\operatorname{gap}(D)\geqslant\frac{b}{2}\big[3r^2-r+2-r\sqrt{r^2-2r+5}\big],\qquad r=\sqrt{a/b}.$$

From the author's knowledge, the precise value of λ_1 are known only for the above typical examples. However, we can now produce infinitely many new examples with sharp estimates. For instance, we have the following simple but non-trivial ones.

b_i	a_i	λ_1	v_i
$i+\beta_0$ $(\beta_0>0)$	$2(i+1)+\beta_0$ $(a_0=0)$	2	$\dfrac{i+2}{i+1}$
$i+2$	i^2	2	$\dfrac{i+1}{i+3}$
$2+(-1)^i$	$2[2+(-1)^i]$	$6-\sqrt{33}$	$\dfrac{\sqrt{33}\pm 1}{4}$

The lower bound of the first two examples come from part (1) and part (2) of Corollary 1.3 respectively. The test sequence used in the last example is $v_i = \big(\sqrt{33}-1\big)/4$ for even i and $= \big(\sqrt{33}+1\big)/4$ for odd i. To see the estimates are sharp, we need the following result for the upper bound. Recall that there is a one-to-one correspondence between a Q-matrix and its operator Ω:

$$\Omega f(i) = \sum_j q_{ij}(f_j - f_i).$$

In the present case,

$$\Omega f(i) = a_i(f_{i-1} - f_i) + b_i(f_{i+1} - f_i).$$

Proposition 1.4.

(1) Let $\lambda > 0$. If the equation $-\Omega g = \lambda g\,(g_0 = -1)$ has a solution (g_i) which is strictly increasing, then $g \in L^1(\pi)$. If either $g \notin L^2(\pi)$ or $g \in L^2(\pi)$ but still $\pi(g) = 0$ (equivalently, $\lim_{n\to\infty} \mu_n b_n(g_{n+1} - g_n) = 0$), then $\lambda \geqslant \lambda_1$.

(2) Let $f \in L^1(\pi)$. Then

$$\lambda_1 \leqslant \begin{cases} \sum_{i\geqslant 0} \pi_i b_i (f_{i+1} - f_i)^2 \big/ \big[\pi(f^2) - \pi(f)^2\big], & \text{if } f \in L^2(\pi) \\ \underline{\lim}_{n\to\infty} \sum_{i=1}^n \mu_i a_i (f_i - f_{i-1})^2 \big/ \sum_{i=0}^n \mu_i f_i^2, & \text{if } f \notin L^2(\pi). \end{cases}$$

For the above examples, the function g required by part (1) of the proposition has the form

$$g_j = -1 + u_0 \sum_{i=0}^{j-1} \prod_{k=0}^{i-1} v_k, \qquad j \geqslant 1$$

for some positive u_0. Actually, here we have used the inverse way. Originally, we fix a strictly increasing function g with $g_0 = -1$ and a constant $\lambda > 0$. Regarding g as an eigenfunction (that is, $-\Omega g = \lambda g$ and $g \not\equiv 0$), from which the relation of the rates a_i and b_i is determined. Then the sequence (v_i) comes from $u_i = g_{i+1} - g_i$ and $v_i = u_{i+1}/u_i$. This explains our original way to construct the examples with sharp estimate. Note that on the one hand, the function v may have less number of parameters than that of g. On the other hand, we need only $\inf_{i\geqslant 0} R_i(v) = \lambda_1$ rather than $R_i(v) \equiv \lambda_1$ for all $i \geqslant 0$. Hence, Theorem 1.1 provides us much more chance to achieve the sharp estimate rather than using the eigenfunction

only. To see this, consider $a_i = b_i = i^2\,(i \geqslant 1)$. As an application of either part (2) of Corollary 1.3 with $v_i = 1 - (2i+4)^{-1}$ or (1.5) with $w_i = \sqrt{i}$, and part (2) of Proposition 1.4 with $f_i = \sqrt{i}$, we get $\lambda_1 = 1/4$. The corresponding eigenfunction is far complex than $\sqrt{i}$. Generally speaking, it is impractical to use the eigenfunction directly since the λ_1 and its eigenfunction g are known or unknown simultaneously. However, it is usually not difficult to find out some approximation of the eigenfunction. This is just what kept in our mind, the test sequences. On the other hand, since the eigenfunction is very sensitive, it is impossible to have a single unified (v_i) or (w_i) for all models. What one can expect is naturally some classification of the test functions as illustrated in Corollary 1.3.

The next result is a weaker but simpler version of (1.5).

Corollary 1.5. (1) For every $w \in \mathscr{W}$ with $w_0 = 0$, we have

$$\operatorname{gap}(D) \geqslant \inf_{i\geqslant 1}\{a_i + b_i - a_i w_{i-1}/w_i - b_i w_{i+1}/w_i\}.$$

(2) For every $w \in \mathscr{W}$, we have

$$\operatorname{gap}(D) \geqslant \inf_{i\geqslant 1}\left\{w_i \Big/ \sum_{j=1}^{i}[a_j\mu_j]^{-1}\sum_{k=j}^{\infty}\mu_k w_k\right\}.$$

For the remainder of this section, we show that we can always get some non-trivial estimates from Theorem 1.1 whenever $\underline{\lim}_{i\to\infty} R_i(v) > 0$ or $\underline{\lim}_{i\to\infty} I_i(w) > 0$. The next result is convenient in practice since the test sequences (v_i) or (w_i) have already carefully designed. Part (1) below is effective if the number $\underline{m}$ is not too much negative and m_i is big for large enough i.

Corollary 1.6. (1) Set $m_i = \Delta a(i) - \Delta b(i)$ and $\underline{m} = \inf_{i\geqslant 0} m_i$. Define

$$\tilde{b}_{i+1} = (i+1)(m_i - \underline{m}) + b_{i+1}\,(i \geqslant 0).$$

For every $N \geqslant 1$, choose

$$w_N \leqslant \Big[\sum_{j=1}^{N} j\Big(\tilde{b}_k^{-1} + \sum_{k=1}^{j-1} a_k\cdots a_{j-1}/\big(\tilde{b}_k\cdots\tilde{b}_j\big)\Big)\Big]^{-1}.$$

Then we have

$$\operatorname{gap}(D) \geqslant \sup_{N\geqslant 2}\min\Big\{w_N + \underline{m},\ m_N - a_N/(N-1),\ \inf_{i\geqslant N+1} m_i\Big\}.$$

(2) Let $\underline{\lim}_{i\to\infty}(a_i - b_i)/i > 0$. Define $x_i = y_i = a_i/(1+b_i)$ if $a_i < 1 + b_i$ and otherwise $x_i = 1$, $y_i = a_i - b_i$. Then

$$\operatorname{gap}(D) \geqslant \inf_{i\geqslant 1}\Big\{y_i\textstyle\prod_{j=1}^{i-1} x_j \Big/ \Big[1 + \sum_{j=1}^{i-1}\prod_{k=1}^{j-1} x_k\Big]\Big\} > 0.$$

Consider again the non-linear example: $a_i = i^2$ and $b_i = i + 2$. Then Corollary 1.6 (1) with $N = 1$ gives us the lower bound 1.00246 which is about half of the exact value $\lambda_1 = 2$ (by Corollary 1.3 (2) with $c_1 = 2$ and $c_2 = 3$). The following comparison result is very useful in practice to simplify some computations.

Proposition 1.7. (1) Fix $N \geqslant 0$. Define successively

$$\tilde{b}_N = b_N, \quad \tilde{b}_i = b_i \vee \{a_i/v_{i-1} + \tilde{b}_{i+1}v_i - a_{i+1} + \alpha_N\}, \quad 0 \leqslant i \leqslant N-1. \quad (1.7)$$

Then

$$\operatorname{gap}(D) \geqslant \left[\inf_{i\geqslant 0} R_i(v)\right] \bigvee \left[\sup_{N\geqslant 1} \frac{b_0 \cdots b_{N-1}}{\tilde{b}_0 \cdots \tilde{b}_{N-1}} \inf_{i\geqslant N} R_i(v)\right]. \quad (1.8)$$

(2) For any regular birth-death process with rates $(\bar{b}_i, \bar{a}_i)$ having the properties $\bar{b}_{i-1}/\bar{a}_i = b_{i-1}/a_i$ and $\bar{a}_i \geqslant a_i$ $(i \geqslant 1)$, we have the same lower bound given in (1.8).

Consider again the periodic example: $b_i = 2 + (-1)^i$ and $a_i = 2b_i$. By using part (2) of Proposition 1.7 and comparing the example with $b_i \equiv 1$, $a_i \equiv 2$ and $b_i \equiv 3$ and $a_i \equiv 6$ respectively, we get

$$\left(\sqrt{2}-1\right)^2 \approx 0.172 \leqslant \lambda_1 \leqslant 3\left(\sqrt{2}-1\right)^2 \approx 0.515.$$

The exact λ_1 is $6 - \sqrt{33} \approx 0.2554$.

Actually, from part (2) of Proposition 1.7 and (1.1), it follows that the above estimates hold not only for a much larger class of birth-death processes but also for those regular, reversible Markov chains with $q_{i,i-1} > 0$ and $q_{i,i+1} > 0$. Finally, we study when $\operatorname{gap}(D) > 0$.

Corollary 1.8. The spectral gap is positive iff one of the following conditions holds.

(1) There exists $v \in \mathscr{V}$ and $N \geqslant 0$ such that $\inf_{i\geqslant N} R_i(v) > 0$.
(2) There exists $w \in L^1(\pi)$ and $N \geqslant 0$ such that w_i is increasing started from N, $w_N > 0$ and $\inf_{i\geqslant N} I_i(w) > 0$.

Corollary 1.9. The spectral gap is positive if one of the lower bounds given in part (1) of Corollary 1.2 or in Corollary 1.3 when "$\inf_{i\geqslant 0}$" is replaced by "$\underline{\lim}_{i\to\infty}$" is positive, or one of the following conditions holds.

(1) $\underline{\lim}_{i\to\infty}(a_i - b_i)/i > 0$.
(2) (Tweedie (1981)) $S := \sum_{n=1}^{\infty}\{1/a_n + \sum_{k=1}^{n} b_k \cdots b_n/[a_k \cdots a_{n+1}]\} < \infty$.
(3) (Van Doorn (1985)) $\underline{\lim}_{i\to\infty}\left(a_i + b_i - \sqrt{a_i b_{i-1}} - \sqrt{a_{i+1}b_i}\right) > 0$.

In case (2), we indeed have $\operatorname{gap}(D) \geqslant S^{-1}$.

We remark that for $a_i = b_i = i^\gamma$ $(i \geqslant 1)$, part (2) of Corollary 1.9 is suitable iff $\gamma > 2$. When $\gamma = 2$, for $v_i = 1 - 1/(2i+2)$ (corresponding to part (2) of Corollary 1.3), we have $\underline{\lim}_{i\to\infty} R_i(v) \geqslant 1/4 = \lambda_1$. Applying part (2) of Proposition 1.4 to the function $f_j = \sqrt{j}$, we get $\lambda_1 = 0$ for all $\gamma \in (1, 2)$.

The remainder of the paper is organized as follows. In the next section, we will quickly prove all the corollaries given in this section. A general result, which contains the main part of Theorem 1.1 and works for general Markov chains, is stated and proved in Section 3. In section 4, we will return to the birth-death processes and complete the proofs of Theorem 1.1 and the propositions given in this section. We will also prove there an accompany result (Theorem 4.3). The optimality of a coupling used in Section 4 is proved in the Appendix.

2. Proofs of the Corollaries.

As we mentioned before, the sequence (v_i) comes from another sequence (u_i): $v_i = u_{i+1}/u_i$. We will use both according to our convenience. In particular, we may and will use $R_i(u) := a_{i+1} + b_i - a_i u_{i-1}/u_i - b_{i+1}u_{i+1}/u_i$ instead of $R_i(v)$. The next result shows the equivalence of (1.4) and (1.5).

Lemma 2.1. (1) Given $w \in \mathscr{W}$, set

$$u_i = \frac{1}{b_i \mu_i} \sum_{j=i+1}^{\infty} \mu_j w_j,$$

$i \geqslant 0$. Then we have $R_i(u) = I_i(w)$ for all $i \geqslant 0$.

(2) Given positive $(u_i : i \geqslant 0)$ such that $\inf_{i\geqslant 0} R_i(u) > 0$, set

$$w_i = a_i u_{i-1} - b_i u_i + c/(\mu - \mu_0), \qquad i \geqslant 1,$$

where $c = \lim_{n\to\infty} b_n \mu_n u_n < \infty$. Then we have $w_{i+1} > w_i$ for all $i \geqslant 1$, $w \in L^1(\pi)$, $\sum_{i\geqslant 1} \mu_i w_i > 0$ and $I_i(w) \geqslant R_i(w)$ for all $i \geqslant 0$.

Proof. a) It follows from the definition of (u_i) that

$$b_{i-1}\mu_{i-1}u_{i-1} - b_i \mu_i u_i = \mu_i w_i.$$

Since $b_{i-1}\mu_{i-1} = a_i \mu_i$, we obtain

$$a_i u_{i-1} - b_i u_i = w_i, \qquad i \geqslant 1. \tag{2.1}$$

Hence $R_i(u) = (w_{i+1} - w_i)/u_i = I_i(w)$ for all $i \geqslant 1$. On the other hand, by (2.1), we have

$$R_0(u) = a_1 + b_0 - b_1 u_1/u_0 = b_0 + (a_1 u_0 - b_1 u_1)/u_0 = b_0 + w_1/u_0 = I_0(w).$$

We have thus proved part (1) of the lemma.

b) For part (2), we first prove the existence of the limit $\lim_{n\to\infty} b_n \mu_n u_n$. To do so, take $w_i = a_i u_{i-1} - b_i u_i + b_1 u_1$ $(i \geqslant 1)$ for a moment. Note that

$$(w_{i+1} - w_i)/u_i = (a_{i+1}u_i - b_{i+1}u_{i+1} - a_i u_{i-1} + b_i u_i)/u_i = R_i(u) > 0, \qquad i \geqslant 1. \tag{2.2}$$

We have $w_i \uparrow$. On the other hand, since

$$\mu_1 w_1 = a_1 \mu_1 u_0 - b_1 \mu_1 u_1 + b_1 \mu_1 u_1 = a_1 \mu_1 u_0 > 0,$$

we see that $w_1 > 0$ and so $w_i > 0$ for all $i \geqslant 1$. Thus,

$$\begin{aligned} 0 &< \sum_{j=1}^{n} \mu_j w_j \\ &= \sum_{j=1}^{n} \Big[b_{j-1}\mu_{j-1}u_{j-1} - b_j \mu_j u_j \Big] + b_1 u_1 \sum_{j=1}^{n} \mu_j \\ &= b_0 \mu_0 u_0 - b_n \mu_n u_n + b_1 u_1 \sum_{j=1}^{n} \mu_j. \end{aligned}$$

Since the left-hand side is increasing in n, it follows that $b_n\mu_n u_n$ must have a finite limit $c \geqslant 0$ as $n \to \infty$.

Next, redefine

$$w_i = a_i u_{i-1} - b_i u_i + c/(\mu - \mu_0), \qquad i \geqslant 1.$$

Then (2.2) remains the same. Moreover,

$$\sum_{j \geqslant i+1} \mu_j w_j = b_i \mu_i u_i - \frac{c}{\mu - \mu_0} \sum_{1 \leqslant j \leqslant i} \mu_j \leqslant b_i \mu_i u_i, \qquad i \geqslant 0. \tag{2.3}$$

This gives us $w \in L^1(\pi)$ and $\sum_{i \geqslant 1} \mu_i w_i > 0$. Now, by (2.2) and (2.3), we get

$$I_i(w) \geqslant b_i \mu_i R_i(u) u_i \Big/ \sum_{j \geqslant i+1} \mu_j w_j \geqslant R_i(u), \qquad i \geqslant 1,$$

and

$$\begin{aligned} I_0(w) &= b_0 \Big[1 + w_1 \Big/ \sum_{j \geqslant 1} \mu_j w_j \Big] \\ &= b_0 + \frac{a_1 u_0 - b_1 u_1 + c/(\mu - \mu_0)}{u_0} \quad \text{(by (2.3))} \\ &\geqslant a_1 + b_0 - b_1 u_1/u_0 = R_0(u). \end{aligned}$$

Therefore, $I_i(w) \geqslant R_i(u)$ for all $i \geqslant 0$. □

Proof of Corollary 1.8. By Theorem 1.1, the conditions are clearly necessary. We now prove the sufficiency.

a) Part (1) of the corollary follows directly from part (1) of Proposition 1.7.

b) To prove part (2), choose a strictly increasing sequence (linear, for instance) $(\overline{w}_i)$ such that $\overline{w}_1 \geqslant 0$, $\overline{w}_i = w_i$ for all $i \geqslant \overline{N}$. Of course, $\sum_{j \geqslant 1} \mu_j \overline{w}_j > 0$. It is now easy to see that $\inf_{i \geqslant 0} I_i(\overline{w}) > 0$ since the original (w_i) is modified locally only. □

To prove Corollary 1.2, we need a simple result. Part (1) below is an extension to [16; Lemma 3.6].

Lemma 2.2. Let $(m_i : i \geqslant 1)$ and $(n_i : i \geqslant 1)$ be non-negative.

(1) If $\sum_{j \geqslant i} m_j n_j \leqslant c_1 m_i$ and $\sum_{j \geqslant i} m_j \leqslant c_2 m_i$ for all $i \geqslant 1$, then

$$\sum_{j \geqslant i} \gamma^{-j} m_j n_j \leqslant \frac{c_1}{1 - c_2(1-\gamma)} \gamma^{-(i-1)} m_i, \qquad i \geqslant 1, \quad \frac{c_2 - 1}{c_2} < \gamma \leqslant 1.$$

(2) If $\sum_{j \geqslant i} m_j \leqslant c/i$ for all $i \geqslant 1$, then

$$\sum_{j \geqslant i} j^\gamma m_j \leqslant c \Big\{ i^{\gamma - 1} + \sum_{j \geqslant i} \frac{1}{j+1} \Big[(j+1)^\gamma - j^\gamma \Big] \Big\}, \qquad i \geqslant 1, \quad \gamma \in [0, 1).$$

Proof. a) Assume that (m_i) has finite support and set $M_i = \sum_{j\geqslant i} m_j n_j$. Then

$$\begin{aligned}\sum_{j\geqslant i}\gamma^{-j}m_jn_j &= \sum_{j\geqslant i}\gamma^{-j}(M_j - M_{j+1})\\ &= \gamma^{-i}M_i + (1-\gamma)\sum_{j\geqslant i}\gamma^{-(j+1)}M_{j+1}\\ &\leqslant c_1\Big[\gamma^{-i+1}m_i + (1-\gamma)\sum_{j\geqslant i}\gamma^{-j}m_j\Big]. \end{aligned}\tag{2.4}$$

In particular, when $n_j \equiv 1$ and $c_1 = c_2$, we get

$$\sum_{j\geqslant i}\gamma^{-j}m_j \leqslant \frac{c_2}{1-c_2(1-\gamma)}\gamma^{-(i-1)}m_i.$$

Inserting this into (2.4), we get the required assertion.

b) Set $M_i = \sum_{j\geqslant i} m_i$. Then

$$\sum_{j\geqslant i} j^\gamma m_j = i^\gamma M_i + \sum_{j\geqslant i}\Big[(j+1)^\gamma - j^\gamma\Big]M_{j+1} \leqslant c\Big\{i^{\gamma-1} + \sum_{j\geqslant i}\frac{1}{j+1}\Big[(j+1)^\gamma - j^\gamma\Big]\Big\}. \ \square$$

Proof of Corollary 1.2. a) The application of Lemma 2.2 goes as follows. Part (1) of the lemma gives us

$$\frac{1}{\gamma^{-i-1}-\gamma^{-i}}\sum_{j\geqslant i+1}\gamma^{-j}m_jn_j \leqslant \frac{c_1}{(\gamma^{-1}-1)[1-c_2(1-\gamma)]}m_{i+1}, \qquad i\geqslant 0.$$

Minimizing the right-hand side with respect to γ, we get $\gamma_0 = \sqrt{(c_2-1)/c_2}$ and hence

$$\frac{1}{\gamma_0^{-i-1}-\gamma_0^{-i}}\sum_{j\geqslant i+1}\gamma_0^{-j}m_jn_j \leqslant \frac{c_1}{\big(\sqrt{c_2}-\sqrt{c_2-1}\,\big)^2}m_{i+1}, \qquad i\geqslant 0. \tag{2.5}$$

b) Applying (2.5) to $m_j = a_j\mu_j$, $n_j = 1/a_j$ and $w_j = \gamma_0^{-j}$, we get

$$\frac{1}{w_{i+1}-w_i}\sum_{j\geqslant i+1}\mu_jw_j \leqslant \frac{c_1}{\big(\sqrt{c_2}-\sqrt{c_2-1}\,\big)^2}a_{i+1}\mu_{i+1}, \qquad i\geqslant 0. \tag{2.6}$$

From this, we have not only $\inf_{i\geqslant 1} I_i(w) \geqslant \big(\sqrt{c_2}-\sqrt{c_2-1}\,\big)^2/c_1$ but also

$$I_0(w) > b_0w_1\Big/\sum_{j\geqslant 1}\mu_jw_j > b_0(w_1-w_0)\Big/\sum_{j\geqslant 1}\mu_jw_j \geqslant \frac{\big(\sqrt{c_2}-\sqrt{c_2-1}\,\big)^2}{c_1}.$$

This completes the proof of part (3) of the corollary.

c) The proof of part (2) is similar but setting $m_j = \mu_j$ and $n_j \equiv 1$.

d) As for part (4) of the corollary, note that

$$\begin{aligned}
&\lim_{i\to\infty}\frac{1}{i^\gamma-(i-1)^\gamma}\left\{i^{\gamma-1}+\sum_{j\geqslant i}\frac{1}{j+1}\big[(j+1)^\gamma-j^\gamma\big]\right\}\\
&=\lim_{i\to\infty}\frac{i^{\gamma-1}}{i^\gamma-(i-1)^\gamma}+\lim_{i\to\infty}\frac{i^{\gamma-1}}{i^\gamma-(i-1)^\gamma}\cdot\frac{1}{i^{\gamma-1}}\sum_{j\geqslant i}(j+1)^{\gamma-2}\frac{(j+1)^\gamma-j^\gamma}{(j+1)^{\gamma-1}}\\
&=\frac{1}{\gamma}+\lim_{i\to\infty}\frac{1}{i^{\gamma-1}}\sum_{j\geqslant i}(j+1)^{\gamma-2}\\
&=\frac{1}{\gamma}+\frac{1}{1-\gamma}=\frac{1}{\gamma(1-\gamma)}.
\end{aligned}$$

The right-hand side achieves the minimum 4 at $\gamma_0=1/2$. Next, we show that

$$f_i:=\frac{1}{\sqrt{i+1}-\sqrt{i}}\left\{\frac{1}{\sqrt{i+1}}+\sum_{j\geqslant i+1}\frac{1}{j+1}\Big[\sqrt{j+1}-\sqrt{j}\Big]\right\}$$

is increasing in i. For this, it suffices that

$$\begin{aligned}
&\Big[\sqrt{i+2}-\sqrt{i}\Big]\sum_{j\geqslant i+2}\frac{1}{j+1}\Big[\sqrt{j+1}-\sqrt{j}\Big]\\
&\qquad\geqslant\sqrt{\frac{i}{i+1}}-\sqrt{\frac{i+1}{i+2}}+\frac{\sqrt{i+1}+\sqrt{i}}{i+2}\Big[\sqrt{i+2}-\sqrt{i+1}\Big]\\
&\qquad=\frac{\sqrt{i}+\sqrt{i+1}\,\big(\sqrt{i(i+2)}-(i+1)\big)}{(i+2)\sqrt{i+1}}.
\end{aligned}$$

Because of $\frac{1}{j+1}\big(\sqrt{j+1}-\sqrt{j}\,\big)>\frac{1}{\sqrt{j+1}}-\frac{1}{\sqrt{j+2}}$, the left-hand side above is greater than $\big(\sqrt{i+2}-\sqrt{i}\,\big)/\sqrt{i+3}\geqslant[(i+2)(i+3)]^{-1/2}$. But the right-hand side is less than $\sqrt{i}\,/\big[(i+2)\sqrt{i+1}\,\big]$. We have thus proved that $f_{i+1}\geqslant f_i$ for all $i\geqslant 0$ and furthermore $\sup_{i\geqslant 0}f_i=4$. Now, applying part (2) of Lemma 2.2 to $m_i=1/a_i$ and $\gamma=1/2$, we obtain

$$\begin{aligned}
\frac{1}{\sqrt{i+1}-\sqrt{i}}\sum_{j\geqslant i+1}\sqrt{j}/a_j&\leqslant\frac{c}{\sqrt{i+1}-\sqrt{i}}\left\{\frac{1}{\sqrt{i+1}}+\sum_{j\geqslant i+1}\frac{1}{j+1}\Big[\sqrt{j+1}-\sqrt{j}\Big]\right\}\\
&\leqslant 4c.
\end{aligned}$$

Hence $\inf_{i\geqslant 1}I_i(w)\geqslant 1/(4c)$. On the other hand,

$$I_0(w)>b_0w_1\Big/\sum_{j\geqslant 1}\mu_jw_j=\frac{1}{\sum_{j\geqslant 1}\sqrt{j}/a_j}\geqslant\frac{1}{4c}.\quad\square$$

For the remainder of the proofs, we need three lemmas. The second one below is quite simple and hence the proof is omitted.

Lemma 2.3. Let $Q = (q_{ij})$ be a regular Q-matrix with stationary distribution (π_i). If there exist a non-negative function h and constants C, $c > 0$ such that

$$\Omega h \leqslant C - ch,$$

then $\pi(h) \leqslant C/c < \infty$.

Proof. Simply use [3; Lemma 4.13 and Lemma 4.10]. □

Lemma 2.4. Given non-negative (m_i) and positive summable (n_i). If

$$\inf_{i \geqslant M} (m_i - m_{i+1})/n_{i+1} =: \delta > 0$$

for some $M \geqslant 0$ then

$$\inf_{i \geqslant M} m_i / \textstyle\sum_{j=i+1}^{\infty} n_j \geqslant \delta.$$

Lemma 2.5. We have

$$\inf_{i \geqslant M} I_i(w) \geqslant \delta$$

provided

$$\Omega w(i) \leqslant -\delta w_i \quad \text{for all } i \geqslant M+1.$$

Here, when $M = 0$, we preassume that $w_0 = 0$.

Proof. When $M > 0$, the conclusion follows from Lemma 2.4 by setting $m_i = b_i\mu_i(w_{i+1} - w_i)$ and $n_i = \mu_i w_i$. The proof also works in the case of $M = 0$ and $w_0 = 0$ since

$$I_0(w) \geqslant b_0 w_1 \Big/ \sum_{j \geqslant 1} \mu_j w_j = b_0(w_1 - w_0) \Big/ \sum_{j \geqslant 1} \mu_j w_j. \quad \square$$

Proof of Corollary 1.5. Part (1) follows directly from Lemma 2.5. To prove part (2), let $\overline{w} \in \mathscr{W}$ and assume that

$$\inf_{i \geqslant 1} \left\{ \overline{w}_i \Big/ \sum_{j=1}^{i} \frac{1}{a_j \mu_j} \sum_{k=j}^{\infty} \mu_k \overline{w}_k \right\} =: \delta > 0.$$

Define

$$w_i = \sum_{j=1}^{i} \frac{1}{a_j \mu_j} \sum_{k=j}^{\infty} \mu_k \overline{w}_k.$$

Then (w_i) satisfies the condition of Lemma 2.5, $w_0 = 0$ and w_i is strictly increasing (since $\sum_{j \geqslant i} \mu_j \overline{w}_j > 0$ for all $i \geqslant 1$). By Lemma 2.3, we have $w \in L^1(\pi)$ and hence $w \in \mathscr{W}$. Now, the assertion follows from part (1). □

Proof of Corollary 1.6. a) Define

$$c_1 = 1, \quad c_{i+1} = \frac{a_1 \cdots a_i}{\tilde{b}_2 \cdots \tilde{b}_{i+1}}, \ \ i \geqslant 1, \qquad u_i = 1 - w_N \sum_{j=1}^{i \wedge N} c_j \sum_{k=1}^{j} \frac{1}{\tilde{b}_k c_k}.$$

Then, we have

$$u_i - u_{i+1} = w_N c_{i+1} \sum_{k=1}^{i+1} \frac{1}{\tilde{b}_k c_k}, \qquad i \leqslant N-1. \tag{2.7}$$

We now prove that

$$u_i \geqslant (i+1)(u_i - u_{i+1}), \qquad 1 \leqslant i \leqslant N-1. \tag{2.8}$$

By (2.7) and the definition of u_i, (2.8) is equivalent to

$$1 - w_N \sum_{j=1}^{i} c_j \sum_{k=1}^{j} \frac{1}{\tilde{b}_k c_k} \geqslant w_N (i+1) c_{i+1} \sum_{k=1}^{i+1} \frac{1}{\tilde{b}_k c_k}, \qquad 1 \leqslant i \leqslant N-1. \tag{2.9}$$

For this, it suffices that $w_N \sum_{j=1}^{N} j c_j \sum_{k=1}^{j} 1/\big(\tilde{b}_k c_k\big) \leqslant 1$. But this follows from the definition of w_N.

Next, for $1 \leqslant i \leqslant N-1$, from (2.8), (2.7) and the definition of (c_i), it follows that

$$\begin{aligned} m_i u_i + b_{i+1}(u_i - u_{i+1}) - a_i(u_{i-1} - u_i) &\geqslant \tilde{b}_{i+1}(u_i - u_{i+1}) - a_i(u_{i-1} - u_i) + \underline{m} u_i \\ &= w_N \tilde{b}_{i+1} c_{i+1} \sum_{k=1}^{i+1} \frac{1}{\tilde{b}_k c_k} \\ &\quad - w_N a_i c_i \sum_{k=1}^{i} \frac{1}{\tilde{b}_k c_k} + \underline{m} u_i \\ &= w_N + \underline{m} u_i. \end{aligned}$$

For $i = 0$, we have

$$m_0 u_0 + b_1(u_0 - u_1) = m_0 + b_1 w_N / \tilde{b}_1 \geqslant w_N + \underline{m} u_0.$$

Therefore,

$$\min_{0 \leqslant i \leqslant N-1} R_i(u) \geqslant \min_{0 \leqslant i \leqslant N-1} \{w_N / u_i + \underline{m}\} \geqslant w_N + \underline{m}. \tag{2.10}$$

For $i = N$, we have

$$m_N u_N + b_{N+1}(u_N - u_{N+1}) - a_N(u_{N-1} - u_N) = m_N u_N - a_N(u_{N-1} - u_N).$$

Thus, by (2.8), we get

$$R_N(u) = m_N - \frac{a_N(u_{N-1} - u_N)}{u_N} \geqslant m_N - \frac{a_N}{N-1}. \tag{2.11}$$

Finally, as for $i > N$, we have $R_i(u) = m_i$. Combining this with (2.10) and (2.11), we obtain the required assertion in part (1).

b) To prove part (2) of the corollary, choose N so that $\inf_{i\geqslant N}(a_i - b_i) \geqslant 1$ and set

$$w_0 = 0, \;\; w_1 = 1, \;\; w_i = 1 + \sum_{j=1}^{i-1}\prod_{k=1}^{j-1} x_k, \qquad i \geqslant 2.$$

Since

$$w_i - w_{i-1} = \prod_{j=1}^{i-1} x_j$$

and $a_i - b_i x_i = y_i$ for all $i \geqslant 1$, we have

$$-\Omega w(i) = (w_i - w_{i-1})(a_i - b_i x_i) = y_i \prod_{j=1}^{i-1} x_j.$$

Noting that $x_i = 1$ for all $i \geqslant N$, we obtain

$$\frac{-\Omega w(i)}{w_i} = \left\{ y_i \prod_{j=1}^{(i-1)\wedge N} x_j \Big/ \left[1 + \sum_{j=1}^{i-1} \prod_{k=1}^{(j-1)\wedge N} x_k \right] \right\}, \qquad i \geqslant 1.$$

Because $\underline{\lim}_{i\to\infty}(a_i - b_i)/i > 0$, by Lemma 2.5, we get

$$\operatorname{gap}(D) \geqslant \inf_{i\geqslant 1}[-\Omega w(i)/w_i] > 0. \quad \square$$

Proof of Corollary 1.9. Because of part (1) of Proposition 1.7 and part (2) or Corollary 1.6, we need only to consider the last two situations.

a) Note that

$$\begin{aligned}
\sum_{n=1}^{\infty}\left\{\frac{1}{a_n} + \sum_{k=1}^{n}\frac{b_k\cdots b_n}{a_k\cdots a_{n+1}}\right\} &= \sum_{n=1}^{\infty}\frac{1}{a_n} + \sum_{n=1}^{\infty}\mu_{n+1}\sum_{k=1}^{n}\frac{1}{\mu_{k-1}b_{k-1}} \\
&= \sum_{n=1}^{\infty}\frac{1}{a_n} + \sum_{k=1}^{\infty}\frac{1}{\mu_{k-1}b_{k-1}}\sum_{n=k}^{\infty}\mu_{n+1} \\
&= \sum_{n=1}^{\infty}\frac{1}{a_n} + \sum_{k=0}^{\infty}\frac{1}{\mu_k b_k}\sum_{n=k+2}^{\infty}\mu_n \\
&= \sum_{k=0}^{\infty}\frac{1}{\mu_k b_k}\sum_{j=k+1}^{\infty}\mu_j.
\end{aligned}$$

Part (2) as well as the last assertion of the corollary follows from part (2) of Corollary 1.5 by setting $w_j \equiv 1\,(j \geqslant 1)$.

b) Assume that

$$a_i + b_i - \sqrt{a_i b_{i-1}} - \sqrt{a_{i+1}b_i} \geqslant \varepsilon > 0 \quad \text{ for all } i \geqslant N.$$

Then,

$$\sqrt{b_i}\big(\sqrt{b_i}-\sqrt{a_{i+1}}\big)\geqslant \varepsilon+\sqrt{a_i}\big(\sqrt{b_{i-1}}-\sqrt{a_i}\big). \tag{2.12}$$

If there is $N_0\geqslant N$ such that $\sqrt{b_{N_0-1}}\geqslant\sqrt{a_{N_0}}$, then it follows from (2.12) that $\sqrt{b_{i-1}}\geqslant\sqrt{a_i}$ for all $i\geqslant N_0$. This is impossible since $\sum_i\mu_i<\infty$. Therefore, $a_n/b_{n-1}>1$ for all $n\geqslant N$. Define

$$w_i=\left(\frac{a_1\cdots a_i}{b_0\cdots b_{i-1}}\right)^{1/2},\quad i\geqslant 1,\qquad w_0=0.$$

Then, w_i is strictly increasing starting from N. Moreover, by assumption,

$$\Omega w(i)\leqslant -\varepsilon w_i\quad \text{ for all } i\geqslant N$$

and so for some C,

$$\Omega w(i)\leqslant C-\varepsilon w_i\quad \text{ for all } i\geqslant 0.$$

From this and Lemma 2.3, it follows that $w\in L^1(\pi)$. Now, part (2) of the corollary also follows from Lemma 2.5. □

3. General Result and Its Proof.

Let $E=\{0,1,2,\cdots,\}$ and $Q=(q_{ij})$ be a regular, irreducible Q-matrix, which is reversible with respect to the distribution (π_i). We introduce two related Q-matrices to deal with the perturbation of the transition rate (q_{ij}) and of the distribution (π_i) respectively. First, let $\overline{Q}=(\bar{q}_{ij})$ be a Q-matrix, reversible with respect to the same (π_i) and satisfy $q_{ij}\geqslant\bar{q}_{ij}$ for all $j<i$ (and hence for all i, j since the reversibility). Next, for a new distribution $(\tilde{\pi}_i)$ which satisfies

$$0<\inf_i \tilde{\pi}_i/\pi_i\leqslant\sup_i\tilde{\pi}_i/\pi_i<\infty, \tag{3.1}$$

we define a reversible Q-matrix (with respect to $(\tilde{\pi}_i)$) as follows:

$$\tilde{q}_{ij}=\bar{q}_{ij}\quad \text{if } i>j\quad\text{and}\quad \tilde{q}_{ij}=\tilde{\pi}_j\bar{q}_{ji}/\tilde{\pi}_i\quad\text{if } i<j.$$

Besides, we need a localizing procedure. Let $n\geqslant 1$ and define a Q-matrix $\hat{Q}_n=(\hat{q}_{ij})$ on $E_n:=\{0,1,2,\ \cdots,n\}$ as follows:

$$\hat{q}_{ij}=\begin{cases}\tilde{q}_{ij}, & \text{if } i,\ j\leqslant n-1\\ \sum_{k\geqslant n}\tilde{q}_{ik}, & \text{if } i\leqslant n-1,\ j=n\\ \tilde{\pi}_j\sum_{k\geqslant n}\tilde{q}_{jk}/\sum_{k\geqslant n}\tilde{\pi}_k, & \text{if } i=n,\ j\leqslant n-1,\end{cases}$$

$$\hat{q}_{nn}=-\sum_{k=0}^{n-1}\hat{q}_{nk}. \tag{3.2}$$

Clearly, $\hat{Q}_n$ is reversible with respect to the distribution $(\tilde{\pi}_0,\cdots,\tilde{\pi}_{n-1},\sum_{k\geqslant n}\tilde{\pi}_k)$.

A *Markovian coupling* of Ω means a coupling operator $\Omega^{\rm coup}$ on the product space E^2 having the *marginality*: $\Omega^{\rm coup}f(i,j) = \Omega f(i)$ (resp. $= \Omega f(j)$) for all $i,\ j$ and for every bounded function f depending on the first (resp. the second) variable only. As usual, we also require that $\Omega^{\rm coup}f(i,i) = \Omega\bar{f}(i)$ for all i and for every bounded (bivariable) function f, where $\bar{f}_i = f(i,i)$. As a typical example, we mention here the *classical coupling* which is meaningful in general and quite simple:

$$\Omega_c^{\rm coup}f(i_1,i_2) = \begin{cases} \big(\Omega f(\cdot,i_2)\big)(i_1) + \big(\Omega f(i_1,\cdot)\big)(i_2), & \text{if } i_1 \neq i_2 \\ \Omega\bar{f}(i_1), & \text{if } i_1 = i_2, \end{cases}$$

Because of [3; Theorem 5.16], we do not need to worry about the regularity of a coupling operator.

We are now at the position to state our general result.

Theorem 3.1. Assume that the Q-matrices Q, $\overline{Q}$ and $\widetilde{Q}$ given above are all regular. For each $n \geqslant 1$, let $\Omega_n^{\rm coup}$ be a coupling of $\hat{Q}_n$.

(1) For each n, let $\varphi : E_n^2 \to [0,\infty)$ be a solution to the inequality

$$\Omega_n^{\rm coup}\varphi(i_1,i_2) + 1 \leqslant 0, \qquad i_1 \neq i_2, \quad i_1,\ i_2 \in E_n \tag{3.3}$$

with $\varphi(i,i) = 0$ for all $i \in E_n$. Then, we have

$$\operatorname{gap}(D) \geqslant \left(\inf_i \frac{\pi_i}{\tilde{\pi}_i} \Big/ \sup_i \frac{\pi_i}{\tilde{\pi}_i}\right) \varlimsup_{n\to\infty} \Big[\max_{i_1\neq i_2,\, i_1,i_2\in E_n} \varphi(i_1,i_2)\Big]^{-1}.$$

(2) Let ρ be a distance in E. If for each n, there exists α_n such that

$$\Omega_n^{\rm coup}\rho(i_1,i_2) \leqslant -\alpha_n\rho(i_1,i_2), \qquad i_1 \neq i_2, \quad i_1,\ i_2 \in E_n, \tag{3.4}$$

then we have $\operatorname{gap}(D) \geqslant \left(\inf_i \frac{\pi_i}{\tilde{\pi}_i} \Big/ \sup_i \frac{\pi_i}{\tilde{\pi}_i}\right) \varlimsup_{n\to\infty} \alpha_n$.

Remark 3.2. A simple but quite effective sufficient condition for the regularity of the Q-matrix $Q = (q_{ij})$ is the following:

$$\sum_i \pi_i q_i < \infty \tag{3.5}$$

(cf. [3; Proposition 6.13]). Clearly, if (3.5) holds, then so is the Q-matrix $\overline{Q}$ defined above. Moreover, under (3.1) and (3.5), we have

$$\begin{aligned} \sum_i \tilde{\pi}_i\tilde{q}_i &= \sum_i \sum_{j\neq i} \tilde{\pi}_i\tilde{q}_{ij} \\ &= 2\sum_i \sum_{j<i} \tilde{\pi}_i\tilde{q}_{ij} \\ &= 2\sum_i \tilde{\pi}_i \sum_{j<i} \bar{q}_{ij} \\ &\leqslant 2\sum_i \tilde{\pi}_i \sum_{j<i} q_{ij} \leqslant \end{aligned}$$

$$\begin{aligned}&\leqslant 2\Big(\sup_k \frac{\tilde\pi_k}{\pi_k}\Big)\sum_i \pi_i \sum_{j<i} q_{ij}\\ &= \Big(\sup_k \frac{\tilde\pi_k}{\pi_k}\Big)\sum_i \pi_i q_i\\ &< \infty.\end{aligned}$$

Hence, the Q-matrix $\widetilde{Q}$ is also regular.

Proof of Theorem 3.1. a) Because

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

Hence, $\mathrm{gap}(\overline{D}) \leqslant \mathrm{gap}\,(D)$ by (1.1).

b) Note that $\sum_i \pi_i(f_i - \pi(f))^2 = \inf_{t\in\mathbf{R}}\sum_i \pi_i(f_i - t)^2$. We have

$$\begin{aligned}\mathrm{gap}\,(\overline{D}) &= \inf_{f\in L^2(\pi)} \frac{\sum_{i>j}\pi_i \bar q_{ij}(f_j - f_i)^2}{\inf_{t\in\mathbf{R}}\sum_i \pi_i(f_i - t)^2}\\ &= \inf_{f\in L^2(\pi)} \frac{\sum_{i>j}\pi_i \tilde q_{ij}(f_j - f_i)^2}{\inf_{t\in\mathbf{R}}\sum_i \pi_i(f_i - t)^2}\\ &\geqslant \frac{\inf_k \pi_k/\tilde\pi_k}{\sup_k \pi_k/\tilde\pi_k}\inf_{f\in L^2(\pi)} \frac{\sum_{i>j}\tilde\pi_i \tilde q_{ij}(f_j - f_i)^2}{\inf_{t\in\mathbf{R}}\sum_i \tilde\pi_i(f_i - t)^2}\\ &= \frac{\inf_k \pi_k/\tilde\pi_k}{\sup_k \pi_k/\tilde\pi_k}\,\mathrm{gap}\,(\widetilde{D}).\end{aligned}$$

Here in the last step we have used the fact that $L^2(\tilde\pi) = L^2(\pi)$ since $\tilde\pi$ and π are equivalent by (3.1). This technique goes back to [11, 9, 24].

c) Next, by [2] or [3; Theorem 9.12], we have $\mathrm{gap}\,(\hat D_n) \downarrow \mathrm{gap}\,(\widetilde{D})$ as $n\to\infty$. It remains to prove that

$$\mathrm{gap}\,(\hat D_n) \geqslant \Big[\max_{i_1\neq i_2,\, i_1,i_2\in E_n}\varphi(i_1,i_2)\Big]^{-1} \tag{3.6}$$

and

$$\mathrm{gap}\,(\hat D_n) \geqslant \alpha_n. \tag{3.7}$$

Denote by $\big(X_t^1, X_t^2\big)$ the coupling process of the original ones with Q-matrix $\hat Q_n$ and let

$$T = \{t\geqslant 0 : X_t^1 = X_t^2\}$$

be the coupling time. Then, the conditions (3.3) and (3.4) give us

$$\mathbb{E}^{i_1,i_2}T \leqslant \varphi(i_1,i_2), \qquad i_1\neq i_2 \tag{3.8}$$

and

$$\mathbb{E}^{i_1,i_2}\rho\big(X_t^1, X_t^2\big) \leqslant \rho(i_1,i_2)e^{-\alpha_n t}, \qquad t\geqslant 0, \quad i_1\neq i_2 \tag{3.9}$$

respectively. Now, the conclusions of Theorem 3.1 can be deduced from the standard argument given in [7] or [4, 5, 23].

To explain the role played by the eigenfunction mentioned in the first section and also for the reader's convenience, here we prove (3.7) under (3.9). Let g be the eigenfunction of $-\Omega$ corresponding to λ_1 (They may be depend on n but we simply ignore it for simplicity). By the forward Kolmogorov equation, we have $\frac{d}{dt}\mathbb{E}^i g(X_t) = \mathbb{E}^i \Omega g(X_t) = -\lambda_1 \mathbb{E}^i g(X_t)$. Hence

$$\mathbb{E}^i g(X_t) = g_i e^{-\lambda_1 t}. \tag{3.10}$$

Next, consider the coupled process (X_t^1, X_t^2) starting from (i_1, i_2). Since the state space is finite, g is Lipschitz with respect to ρ (This indicates the necessity of using the localizing procedure). Denote by c_g the Lipschitz constant of g. By using (3.10) and then (3.9), we obtain

$$e^{-\lambda_1 t}|g(i_1) - g(i_2)| \leqslant \mathbb{E}^{i_1,i_2}\big|g\big(X_t^1\big) - g\big(X_t^2\big)\big| \leqslant c_g \mathbb{E}^{i_1,i_2}\rho\big(X_t^1, X_t^2\big) \leqslant c_g \rho(i_1, i_2) e^{-\alpha_n t}.$$

Note that if g is strictly monotone (it is the case for the birth-death processes, see Lemma 4.2), by taking $\rho(i, j) = |g_i - g_j|$ and using the order-preserving property of the coupling, the above inequalities can be all replaced by equalities with $\alpha_n = \lambda_1$, without taking the absolute value. This explains a way to obtain the sharp estimates. To complete the proof, simply choose (i_1, i_2) so that

$$|g(i_1) - g(i_2)|/\rho(i_1, i_2) = c_g.$$

Actually, the proof is almost the same if we use directly an eigenfunction h of the coupling operator. Then the equalities hold whenever the coupling process is order-preserved and $h(i, j) = \bar{g}_i - \bar{g}_j$ for some strictly increasing function $\bar{g}$, which may not necessarily be an eigenfunction of the original operator. □

We now mention another approximating method which is also meaningful and sometimes even simpler. That is the restriction of $\widetilde{Q}$ to E_n:

$$q_{ij}^{(n)} = \tilde{q}_{ij}, \quad \text{if } i \neq j,\ i, j \in E_n; \qquad q_{ii}^{(n)} = -\sum_{j \neq i,\, j \in E_n} q_{ij}^{(n)}, \quad i \in E_n;$$

$$\pi_i^{(n)} = \pi_i \Big/ \sum_{k \leqslant n} \pi_k, \qquad i \in E_n. \tag{3.11}$$

In general, the Q-matrix $Q_n = (q_{ij}^{(n)})$ can be reducible. The main advantage of this approximation is that if (3.4) holds with $\alpha_n \equiv \alpha$ for a coupling of the $\widetilde{Q}$-processes, then it holds often automatically for the coupling of the Q_n-processes for all n.

Corollary 3.3. Everything is the same as in Theorem 3.1 except the Q-matrix $\hat{Q}_n$ is replaced by (3.11).

Proof. For simplicity, we omit the superscript "$\sim$" of $\widetilde{Q} = (\tilde{q}_{ij})$ in the proof. Since $Q = (q_{ij})$ is regular, by [3; Theorem 9.9], we can choose a function f

so that $f = c$ =constant out off E_m with mean zero and variance 1 such that $D(f,f) \leqslant \text{gap}\,(D) + \varepsilon$. Then, when $n > m$, we have

$$\pi^{(n)}(f) = -c\sum_{i>n}\pi_i \Big/ \sum_{j\leqslant n}\pi_j \quad \text{and} \quad \pi^{(n)}(f^2) = \Big(1 - c^2\sum_{i>n}\pi_i\Big)\Big/\sum_{j\leqslant n}\pi_j.$$

Thus,

$$\pi^{(n)}(f^2) - \pi^{(n)}(f)^2 = \Big[\sum_{j\leqslant n}\pi_j - c^2\sum_{i>n}\pi_i\Big]\Big/\Big(\sum_{j\leqslant n}\pi_j\Big)^2$$

and so

$$\begin{aligned}
&\frac{1}{2}\sum_{i,\,j\leqslant n}\pi_i^{(n)}q_{ij}^{(n)}(f_j - f_i)^2 \Big/ \Big[\pi^{(n)}(f^2) - \pi^{(n)}(f)^2\Big] \\
&\qquad = \frac{1}{2}\sum_{i,\,j\leqslant n}\pi_i q_{ij}(f_j - f_i)^2 \Big/ \Big[\Big(\sum_{k\leqslant n}\pi_k\Big)\Big(\pi^{(n)}(f^2) - \pi^{(n)}(f)^2\Big)\Big] \\
&\qquad \leqslant \big(\text{gap}\,(D) + \varepsilon\big)\Big/\Big[1 - c^2\sum_{i>n}\pi_i \Big/ \sum_{j\leqslant n}\pi_j\Big].
\end{aligned}$$

We get $\overline{\lim_{n\to\infty}}\,\text{gap}\,(D_n) \leqslant \text{gap}\,(D) + \varepsilon$. But ε can be arbitrarily small, we finally obtain $\overline{\lim_{n\to\infty}}\,\text{gap}\,(D_n) \leqslant \text{gap}\,(D)$. $\square$

For the second approximation given by (3.11), we have proved a weaker conclusion that $\overline{\lim_{n\to\infty}}\,\text{gap}\;(D_n) \leqslant \text{gap}\,(D)$ rather than $\text{gap}\,(D_n) \downarrow \text{gap}\,(D)$ for the first approximation. However, within the context of birth-death processes, the last conclusion also holds for the second approximation.

Proposition 3.4. Consider the restriction of (b_i, a_i) to $\{n, n+1, \cdots, m\}$ $(0 \leqslant n < m \leqslant \infty)$ with reflection boundaries and denote by $\text{gap}_{n,m}$ its spectral gap. Then, we have $\text{gap}\,(D) \leqslant \text{gap}_{n,m}$. Moreover, $\text{gap}_{n,m}$ is decreasing as $m \uparrow$ or $n \downarrow$.

Proof. a) Define $\pi_i^{(n,m)} = \pi_i / \sum_{n\leqslant k\leqslant m}\pi_k$. Take f with $\pi^{(n,m)}(f) = 0$ and $\pi^{(n,m)}(f^2) = 1$ such that

$$\frac{1}{2}\sum_{n\leqslant i,\,j\leqslant m}\pi_i^{(n,m)}q_{ij}(f_j - f_i)^2 \leqslant \text{gap}_{n,m} + \varepsilon.$$

Define $\tilde{f} = fI_{[n\leqslant i\leqslant m]} + f_nI_{[i<n]} + f_mI_{[i>m]}$. Then,

$$\pi\big(\tilde{f}\big) = f_n\sum_{i<n}\pi_i + f_m\sum_{i>m}\pi_i, \qquad \pi\big(\tilde{f}^2\big) = \sum_{n\leqslant i\leqslant m}\pi_i + f_n^2\sum_{i<n}\pi_i + f_m^2\sum_{i>m}\pi_i,$$

$$\begin{aligned}\pi(\tilde{f}^2)-\pi(\tilde{f})^2 &= \sum_{n\leqslant i\leqslant m}\pi_i+f_n^2\sum_{i<n}\pi_i\Big(1-\sum_{i<n}\pi_i\Big)+f_m^2\sum_{i>m}\pi_i\Big(1-\sum_{i>m}\pi_i\Big)\\ &\quad -2f_nf_m\sum_{i<n}\pi_i\sum_{i>m}\pi_i\\ &= \sum_{n\leqslant i\leqslant m}\pi_i\Big(1+f_n^2\sum_{i<n}\pi_i+f_m^2\sum_{i>m}\pi_i\Big)+(f_n-f_m)^2\sum_{i<n}\pi_i\sum_{i>m}\pi_i\\ &\geqslant \sum_{n\leqslant i\leqslant m}\pi_i.\end{aligned}$$

Due to the restriction to the birth-death processes,

$$\sum_{i<j<n}\pi_iq_{ij}(\tilde{f}_j-\tilde{f}_i)^2=\pi_{n-1}q_{n-1,n}(\tilde{f}_{n+1}-\tilde{f}_n)^2=0$$

and

$$\sum_{i\leqslant m<j}\pi_iq_{ij}(\tilde{f}_j-\tilde{f}_i)^2=\pi_mq_{m,m+1}(\tilde{f}_{m+1}-\tilde{f}_m)^2=0,$$

we have

$$\sum_{i<j}\pi_iq_{ij}(\tilde{f}_j-\tilde{f}_i)^2\Big/\Big[\pi(\tilde{f}^2)-\pi(\tilde{f})^2\Big]\leqslant\sum_{n\leqslant i<j\leqslant m}\pi_i^{(n,m)}q_{ij}(f_j-f_i)^2\leqslant\mathrm{gap}_{n,m}+\varepsilon.$$

Therefore, $\mathrm{gap}\,(D)\leqslant\mathrm{gap}_{n,m}+\varepsilon$ and then $\mathrm{gap}\,(D)\leqslant\mathrm{gap}_{n,m}$ by letting $\varepsilon\downarrow 0$.

b) To prove the monotonicity of $\mathrm{gap}_{n,m}$, it suffices to show that $\mathrm{gap}_{n,m}\geqslant\mathrm{gap}_{n,m+1}$. This simply follows from the proof a) and even simpler. For instance, the modified function $\tilde{f}$ becomes $fI_{[n\leqslant i\leqslant m]}+f_mI_{[i=m+1]}$. $\square$

Example 3.5. Let $q_{0k}=\beta_k>0$, $q_{k0}=\frac{1}{2}\,(k\geqslant 1)$ and $q_{ij}=0$ for the other cases of $i\neq j$. The operator $-\Omega$ has eigenvalues 0, 1/2 and $\frac{1}{2}+\sum_{k\geqslant 1}\beta_k$ with 1/2 having infinite multiplicity. The eigenfunctions of $\lambda_1=1/2$ are neither unique nor monotone. Hence, this example is quite different from birth-death processes. Due to the fact that $\beta_k>0$, we may choose a strictly increasing sequence $g_k\,(k\geqslant 2)$ so that $\sum_{k\geqslant 2}\beta_kg_k<\infty$. Next, define $g_1<0$ by $\sum_{k\geqslant 1}\beta_kg_k=0$ and set $g_0=0$. Finally, define a distance ρ on $\mathbf{Z}_+$ by $\rho(0,1)=-g_1$ and $\rho(i,i+1)=g_{i+1}-g_i\,(i\geqslant 1)$. Consider the coupling Ω^{coup}: If $i,\,j\geqslant 1$, $(i,j)\to(0,0)$ at rate 1/2. Otherwise, use the classical coupling. Then, it is not difficult to check that

$$\Omega^{\mathrm{coup}}\rho(i,j)\leqslant-\rho(i,j)/2$$

and so by Corollary 3.3, we have $\mathrm{gap}\,(D)\geqslant 1/2=\lambda_1$.

Example 3.6. Consider a Markov chain with state space $\mathbf{Z}_+^2$ and with the following transition intensity: $(i,j)\to(i+1,j)$ or $(i,j)\to(i,j+1)\,(i,\,j\geqslant 0)$ at rate $\beta/4\,(0<\beta<1)$ respectively. $(i,j)\to(i-1,j)\,(i\geqslant 1,\,j\geqslant 0)$ or $(i,j)\to(i,j-1)\,(i\geqslant 0,\,j\geqslant 1)$ at rate 1/4. Since the components are independent, by the Addition Theorem for spectral gap[16] and the first example

given above (1.6), we have $\lambda_1 = \big(\sqrt{\beta}-1\big)^2/4$. It is also easy to obtain the exact lower bound by using the coupling approach. Take the distance ρ on $\mathbf{Z}_+^2$ to be the sum of the ones on $\mathbf{Z}_+$ used in the example just mentioned above. Adopt the classical coupling for the k-th ($k = 1, 2$) components respectively and then sum them together.

4. Application to Birth-death Processes. Proof of Theorem 1.1.

In this section, we apply Theorem 3.1 to the birth-death processes and complete the proof of the results given in Section 1. Since Proposition 1.7 follows directly from Theorem 3.1, our main task is to prove Theorem 1.1 and Proposition 1.4. To do so, we need some preparations.

We will use two different couplings: The classical coupling $\Omega_c^{\rm coup}$ mentioned in the last section and the coupling by reflection $\Omega_r^{\rm coup}$. The second one is specially defined for birth-death processes with rates (b_i, a_i). Again, we have $\Omega_r^{\rm coup} f(i,i) = \Omega \bar{f}(i)$, where Ω is the operator of the marginals and $\bar{f}_i = f(i,i)$. For $i_1 < i_2$, we have

$$\begin{aligned}
&\Omega_r^{\rm coup}\, f(i_1,i_2)\\
&\quad = I_{[i_2=i_1+1]}\Big\{b_{i_1}\big[f(i_1+1,i_2)-f(i_1,i_2)\big] + a_{i_2}\big[f(i_1,i_2-1)-f(i_1,i_2)\big]\\
&\qquad\qquad + (a_{i_1}\wedge b_{i_2})\big[f(i_1-1,i_2+1)-f(i_1,i_2)\big] + \cdots\Big\}\\
&\quad + I_{[i_2-i_1\geqslant 2]}\Big\{(b_{i_1}\wedge a_{i_2})\big[f(i_1+1,i_2-1)-f(i_1,i_2)\big]\\
&\qquad\qquad + (b_{i_2}\wedge a_{i_1})\big[f(i_1-1,i_2+1)-f(i_1,i_2)\big] + \cdots\Big\}.
\end{aligned}$$

Here, we have omitted six terms on the right-hand side. Once a term $A \wedge B$ appears, two other terms $(A-B)^+$ and $(B-A)^+$ should be also included for the independent jump due to the marginality. For instance, because of the term $a_{i_1} \wedge b_{i_2}$, we should also have $(a_{i_1} - b_{i_2})^+\big[f(i_1-1,i_2)-f(i_1,i_2)\big]$ and $(b_{i_2} - a_{i_1})^+\big[f(i_1,i_2+1)-f(i_1,i_2)\big]$. By symmetry, we can write down the rates for the case that $i_1 > i_2$. Of course, these couplings are also meaningful for the localized processes determined by (3.2) or (3.11) respectively.

Given a positive sequence (u_i), we have a distance

$$\rho(i,j) = \left|\sum_{k<j} u_k - \sum_{k<i} u_k\right|$$

on $\mathbf{Z}_+$. Next, for any function $\gamma : [0,\infty) \to [0,\infty)$ with $\gamma(0) = 0$, $\gamma' > 0$ and $\gamma'' \leqslant 0$, we can define a new distance $\gamma \circ \rho$. A typical example of $\gamma(x)$ is $\log(1+rx)$ for some $r > 0$. It will be proved in the Appendix that with respect to this class of distances, the optimal coupling is the coupling by reflection. A critical application of the study on optimal Markovian couplings is that it leads us to classify couplings according to different distances. For instance, if we restrict to the special case of $\gamma(x) \equiv x$, then the last coupling coincides with the classical one and furthermore

$$\Omega_c^{\rm coup}\rho(i_1,i_2) = b_{i_2}u_{i_2} - a_{i_2}u_{i_2-1} - b_{i_1}u_{i_1} + a_{i_1}u_{i_1-1}, \quad a_0 := 0,\ i_2 > i_1 \qquad (4.1)$$

(cf. [5; Theorem 3.3] and (5.1) below).

The next two lemmas characterize the eigenfunction of λ_1.

Lemma 4.1. Let $\lambda > 0$ and $g \not\equiv 0$ be a solution to the equation $\Omega g = -\lambda g$. Then $g_0 \neq 0$ and

$$\pi_n b_n (g_{n+1} - g_n) = -\lambda \textstyle\sum_{i=0}^{n} \pi_i g_i, \qquad n \geqslant 0. \tag{4.2}$$

Proof. a) The formula (4.2) follows from

$$\begin{aligned} -\lambda \sum_{i=0}^{n} \pi_i g_i &= \sum_{i=0}^{n} \pi_i \Omega g(i) \\ &= \sum_{i=0}^{n} \big[\pi_i a_i (g_{i-1} - g_i) + \pi_i b_i (g_{i+1} - g_i)\big] \\ &= \sum_{i=0}^{n} \big[- \pi_i a_i (g_i - g_{i-1}) + \pi_{i+1} a_{i+1} (g_{i+1} - g_i)\big] \\ &= -\pi_0 a_0 (g_0 - g_{-1}) + \pi_{n+1} a_{n+1} (g_{n+1} - g_n) \\ &= \pi_n b_n (g_{n+1} - g_n). \end{aligned}$$

Here the additional term g_{-1} can be ignored since $a_0 = 0$.

b) If $g_0 = 0$, then by induction, it follows from (4.2) that $g_i \equiv 0$. This is a contradiction. □

Lemma 4.2. Let $\lambda_1 > 0$ and g be a solution to the equation $\Omega g = -\lambda_1 g$ with $g_0 < 0$. Then g_i is strictly increasing and $g \in L^1(\pi)$. Moreover,

$$\pi(g) = -\lim_{n\to\infty} \pi_n b_n (g_{n+1} - g_n)/\lambda_1 \leqslant 0.$$

Proof. a) Since $g_0 < 0$, by (4.2), we have $g_1 > g_0$. If g_i is not strictly increasing, then there would exist an $n \geqslant 1$ such that

$$g_0 < g_1 < \cdots < g_{n-1} < g_n \geqslant g_{n+1}. \tag{4.3}$$

We are going to prove that this is impossible.

b) By (4.2), we have

$$g_k < (\text{resp. } =)\, g_{k+1} \Longleftrightarrow \textstyle\sum_{i=0}^{k} \pi_i g_i < (\text{resp. } =)\, 0. \tag{4.4}$$

c) Define

$$\tilde{g}_n = -\sum_{i=0}^{n-1} \pi_i g_i / \pi_n$$

and

$$\tilde{g}_i = g_i I_{[i<n]} + \tilde{g}_n I_{[i\geqslant n]}.$$

Then, it follows from (4.2)–(4.4) that

$$g_n \geqslant \tilde{g}_n = [\pi_{n-1} b_{n-1} (g_n - g_{n-1})]/(\lambda_1 \pi_n) = [a_n (g_n - g_{n-1})]/\lambda_1 > 0 \tag{4.5}$$

and moreover,

$$\textstyle\sum_{i\leqslant n}\pi_i\tilde{g}_i=0. \tag{4.6}$$

d)[2] Since $\sum_{i\geqslant n+1}\pi_i<1$, $\tilde{g}_n>0$ and (4.6), we obtain

$$\sum_i \pi_i\tilde{g}_i^2-\Big(\sum_i\pi_i\tilde{g}_i\Big)^2=\sum_{i\leqslant n}\pi_i\tilde{g}_i^2+\tilde{g}_n^2\sum_{i\geqslant n+1}\pi_i-\tilde{g}_n^2\Big(\sum_{i\geqslant n+1}\pi_i\Big)^2>\sum_{i\leqslant n}\pi_i\tilde{g}_i^2. \tag{4.7}$$

On the other hand,

$$\begin{aligned}-\sum_i\pi_i\big(\tilde{g}\Omega\tilde{g}\big)(i)&=\lambda_1\sum_{i\leqslant n-1}\pi_i\tilde{g}_i^2-\pi_{n-1}b_{n-1}g_{n-1}\big(\tilde{g}_n-g_n\big)-\pi_n a_n\tilde{g}_n\big(g_{n-1}-\tilde{g}_n\big)\\&=\lambda_1\sum_{i\leqslant n-1}\pi_i\tilde{g}_i^2-\pi_n a_n\big[g_{n-1}\big(\tilde{g}_n-g_n\big)+\tilde{g}_n\big(g_{n-1}-\tilde{g}_n\big)\big].\end{aligned} \tag{4.8}$$

We now consider two cases separately,

i) If $g_{n-1}\geqslant\tilde{g}_n\,(>0)$, then it follows from (4.8) that

$$-\sum_i\pi_i\big(\tilde{g}\Omega\tilde{g}\big)(i)\leqslant\lambda_1\sum_{i\leqslant n-1}\pi_i\tilde{g}_i^2.$$

Combining this with (4.7), we obtain

$$\lambda_1\leqslant-\sum_i\pi_i\big(\tilde{g}\Omega\tilde{g}\big)(i)\Big/\Big[\sum_i\pi_i\tilde{g}_i^2-\Big(\sum_i\pi_i\tilde{g}_i\Big)^2\Big]<\lambda_1\sum_{i\leqslant n-1}\pi_i\tilde{g}_i^2\Big/\sum_{i\leqslant n}\pi_i\tilde{g}_i^2<\lambda_1.$$

This is a contradiction.

ii) If $g_{n-1}<\tilde{g}_n$, then it follows from (4.5) that

$$\begin{aligned}&\pi_n\,a_n\big[g_{n-1}\big(\tilde{g}_n-g_n\big)+\tilde{g}_n\big(g_{n-1}-\tilde{g}_n\big)\big]\\&\quad=\pi_n a_n\big[g_{n-1}\big(\tilde{g}_n-\lambda_1\tilde{g}_n/a_n-g_{n-1}\big)+\tilde{g}_n\big(g_{n-1}-\tilde{g}_n\big)\big]\\&\quad=\pi_n\big(\tilde{g}_n-g_{n-1}\big)\big[\lambda_1\tilde{g}_n-a_n\big(\tilde{g}_n-g_{n-1}\big)\big]-\lambda_1\pi_n\tilde{g}_n^2\\&\quad\geqslant\pi_n\big(\tilde{g}_n-g_{n-1}\big)\big[\lambda_1\tilde{g}_n-a_n\big(g_n-g_{n-1}\big)\big]-\lambda_1\pi_n\tilde{g}_n^2\\&\quad=-\lambda_1\pi_n\tilde{g}_n^2.\end{aligned}$$

Combining this with (4.7) and (4.8), we get

$$\begin{aligned}\lambda_1&\leqslant-\sum_i\pi_i\big(\tilde{g}\Omega\tilde{g}\big)(i)\Big/\Big[\sum_i\pi_i\tilde{g}_i^2-\Big(\sum_i\pi_i\tilde{g}_i\Big)^2\Big]\\&<\Big[\lambda_1\sum_{i\leqslant n-1}\pi_i\tilde{g}_i^2+\lambda_1\pi_n\tilde{g}_n^2\Big]\Big/\sum_{i\leqslant n}\pi_i\tilde{g}_i^2\\&=\lambda_1.\end{aligned}$$

[2]This part of the proof is corrected at the end of the paper.

It is also a contradiction.

e) Having the increasing property of g in mind, it is now easy to show that $g \in L^1(\pi)$ by using Lemma 2.3 with $h = g - g_0$. The last assertion then follows from (4.2) $\square$

Proof of Proposition 1.4. The first assertion of part (1) follows from Lemma 2.3.

Define $\tilde{g}_i = g_i I_{[i\leqslant n]} + g_n I_{[i>n]}$. Then, on the one hand,

$$-\sum_i \pi_i\big(\tilde{g}\Omega\tilde{g}\big)(i) = \lambda \sum_{i\leqslant n} \pi_i g_i^2 + \pi_n b_n(g_{n+1} - g_n). \tag{4.9}$$

On the other hand,

$$\begin{aligned}
&\sum_i \pi_i \tilde{g}_i^2 - \Big(\sum_i \pi_i \tilde{g}_i\Big)^2 \\
&= \sum_{i\leqslant n} \pi_i g_i^2 + g_n^2 \sum_{i\geqslant n+1} \pi_i - g_n^2 \Big(\sum_{i\geqslant n+1} \pi_i\Big)^2 - \Big(\sum_{i\leqslant n} \pi_i g_i\Big)^2 - 2g_n \sum_{i\leqslant n} \pi_i g_i \sum_{j\geqslant n+1} \pi_j \\
&\geqslant \sum_{i\leqslant n} \pi_i g_i^2 - \Big(\sum_{i\leqslant n} \pi_i g_i\Big)^2 - 2g_n \sum_{i\leqslant n} \pi_i g_i \sum_{j\geqslant n+1} \pi_j.
\end{aligned} \tag{4.10}$$

Hence

$$\begin{aligned}
\lambda_1 &\leqslant \varliminf_{n\to\infty} \frac{-\sum_i \pi_i\big(\tilde{g}\Omega\tilde{g}\big)(i)}{\sum_i \pi_i \tilde{g}_i^2 - \Big(\sum_i \pi_i \tilde{g}_i\Big)^2} \\
&\leqslant \varliminf_{n\to\infty} \frac{\lambda \sum_{i\leqslant n} \pi_i g_i^2 + \pi_n b_n(g_{n+1} - g_n)}{\sum_{i\leqslant n} \pi_i g_i^2 - \Big(\sum_{i\leqslant n} \pi_i g_i\Big)^2 - 2g_n \sum_{i\leqslant n} \pi_i g_i \sum_{j\geqslant n+1} \pi_j}.
\end{aligned} \tag{4.11}$$

Next, since g_i is strictly increasing, by (4.2), we have $\sum_{i\leqslant n} \pi_i g_i \leqslant 0$. Combining this with the fact that $g \in L^1(\pi)$, we get

$$-g_n \sum_{i\leqslant n} \pi_i g_i \sum_{j\geqslant n+1} \pi_j \leqslant \Big[-\sum_{i\leqslant n} \pi_i g_i\Big] \sum_{j\geqslant n+1} \pi_j g_j \to 0 \quad \text{as } n \to \infty. \tag{4.12}$$

i) If $g \in L^2(\pi)$ and $\pi(g) = 0$, then by (4.2), $\lim_{n\to\infty} \pi_n b_n(g_{n+1} - g_n) = 0$. Therefore, the right-hand side of (4.11) is less than or equal to λ.

ii) If $g \notin L^2(\pi)$, by (4.12), we have the same conclusion as in i). We have thus proved part (1) of the proposition.

The first assertion of part (2) follows directly from (1.1). The second one follows from Proposition 3.4. $\square$

We are now ready to prove the first main result of the paper.

Proof of Theorem 1.1. a) First, we prove that $\operatorname{gap}(D) \geqslant \inf_{i\geqslant 0} R_i(v)$ for every $v \in \mathscr{V}$. For this, we use the classical coupling. By Corollary 3.3 and Proposition 3.4,

in the present context, we can adopt the second approximation. Then, (3.4) holds with $\alpha_n \equiv \alpha$ for all n provided

$$\Omega_c^{\rm coup}\rho(i,j) \leqslant -\alpha\rho(i,j), \qquad i<j. \tag{4.13}$$

Now, we prove that (4.13) holds iff

$$R_i = a_{i+1}+b_i-[a_iu_{i-1}+b_{i+1}u_{i+1}]/u_i \geqslant \alpha, \qquad a_0:=0, \quad i\geqslant 0 \tag{4.14}$$

[3]For this, set $g_i=\sum_{k<i}u_k$ ($g_0:=0$) and $\rho(i,j)=|g_i-g_j|$, then

$$\Omega_c^{\rm coup}\rho(i,j)=\Omega\rho(\cdot,j)(i)+\Omega\rho(i,\cdot)(j)=\Omega(g_j-g_\cdot)(i)+\Omega(g_\cdot-g_i)(j)=\Omega g(j)-\Omega g(i)$$

for $i<j$, and so

$$\Omega_c^{\rm coup}\rho(i,j)=\Omega_c^{\rm coup}\rho(i,i+1)+\cdots+\Omega_c^{\rm coup}\rho(j-1,j), \qquad i<j.$$

Hence

$$\Omega_c^{\rm coup}\rho(i,j)\leqslant -\alpha\rho(i,j) \quad \text{ for all } i<j$$

is equivalent to

$$\Omega_c^{\rm coup}\rho(i,i+1)\leqslant -\alpha\rho(i,i+1).$$

From this and (4.1), we get the equivalence of (4.13) and (4.14).

b)[4]Next, by Lemma 4.2, the eigenfunction g of λ_1 should be strictly increasing, so we can always obtain a positive sequence (v_i) from g by taking

$$v_i=(g_{i+2}-g_{i+1})/(g_{i+1}-g_i) \quad (i\geqslant 0).$$

Thus, the supremum in (1.4) can be attained.

c) Because of Lemma 2.1, (1.5) follows from (1.4) and the supremum in (1.5) can be also attained. □

Similarly, we can prove the following result.

Theorem 4.3. For birth-death processes, if for some function γ,

$$\inf_{i\geqslant 0}\big[-\Omega_r^{\rm coup}\gamma\circ\rho(i,i+k)/\gamma\circ\rho(i,i+k)\big]=:\alpha_k\geqslant\alpha, \qquad k\geqslant 1$$

where $\Omega_r^{\rm coup}\gamma\circ\rho(i,i+k)$ is given by (5.1), then we have $\mathrm{gap}\,(D)\geqslant\alpha$.

In the particular case that $u_i\equiv 1$, we can forget the original (u_i) and reset $u_i=\gamma_{i+1}-\gamma_i$. Then $\gamma_k=\sum_{j<k}u_j$ and α_k used above becomes

$$\alpha_k=\begin{cases}\inf_{i\geqslant 0}\{b_i+a_{i+1}-(a_i\wedge b_{i+1})u_2-(a_i\vee b_{i+1})u_1\}/u_0, & \text{if } k=1\\ \inf_{i\geqslant 0}\{(b_i\vee a_{i+k})u_{k-1}+(b_i\wedge a_{i+k})u_{k-2}-(a_i\wedge b_{i+k})u_{k+1}\\ \qquad -(a_i\vee b_{i+k})u_k\}/\gamma_k, & \text{if } k\geqslant 2.\end{cases}$$

[3]Then $\lambda_1\geqslant\alpha:=\inf_{i\geqslant 0}R_i$ by part (2) of Theorem 3.1, and furthermore the inequality "$\geqslant$" in (1.4) holds.

[4]When $\lambda_1=0$, the equality in (1.4) is trivial since the inequality "$\geqslant$" holds by a), and by part (1) of Lemma 2.1, $\sup_{v\in\mathscr{V}}\inf_{i\geqslant 0}R_i(v)\geqslant 0$. It remains to consider the case that $\lambda_1>0$.

Example 4.4[5]**.** Take $E_N = \{0, 1, \cdots, N\}$, $a_i = 1\,(1 \leqslant i \leqslant N)$ and $b_i = 1\,(0 \leqslant i \leqslant N-1)$. Applying $\gamma(x) = \sin[\pi x/(2N+2)]$ to the last formula, we obtain the exact bound:

$$\alpha_{N-1} \geqslant \alpha_1 = \cdots = \alpha_{N-2} = \alpha_N = \lambda_1 = 4\sin^2[\pi/(2N+2)].$$

However, if we take $\gamma(x) = x$ in the last formula, then we get nothing.

Note that the distance $\gamma \circ \rho$ used above is not the type used in Theorem 1.1 and it is essential different from that deduced from the eigenfunction as explained in Section 1. The eigenfunction for the last example is

$$g_i = \tan\left[\frac{\pi}{2(N+1)}\right]\sin\left[\frac{i\pi}{N+1}\right] - \cos\left[\frac{i\pi}{N+1}\right] \quad (i \geqslant 0).$$

Generally speaking, the use of a rather simple γ enables us to avoid the technical design of ρ but still obtain good enough estimate. It is much more effective than the comparison result, part (1) of Proposition 1.7. Certainly, one needs much more work to use Theorem 4.1 but it can be simplified. To see further examples and some simplification, refer to the application to the reaction-diffusion processes [6], for which the comparison techniques used in Proposition 1.7 are no longer suitable.

As an application of part (1) of Theorem 3.1, we present an alternative proof of the result "$\mathrm{gap}\,(D) \geqslant S^{-1}$" given in Corollary 1.9. It also gives us a probabilistic explanation of the condition $S < \infty$.

Corollary 4.5. $\mathrm{gap}\,(D) \geqslant \left\{\sum_{n=1}^{\infty}\{1/a_n + \sum_{k=1}^{n} b_k \cdots b_n/[a_k \cdots a_{n+1}]\}\right\}^{-1}$.

Proof. a) Define

$$\bar{x}_0 = 0, \qquad \bar{x}_k = \sum_{i=0}^{k-1} \frac{1}{\mu_i b_i} \sum_{j=i+1}^{\infty} \mu_j.$$

Let τ_0 be the hitting time of the process hits 0 starting from i. Then $(\mathbb{E}^i\tau_0 : i \geqslant 1)$ is the minimal solution to the equation (cf. [3; Lemma 4.48 with $\lambda = 0$]):

$$x_k = \sum_{j \neq k,\,0} \frac{q_{kj}}{q_k} x_j + \frac{1}{q_k}, \qquad k \geqslant 1. \tag{4.15}$$

It is easy to check that $(\bar{x}_k)$ satisfies (4.15) and so $\mathbb{E}^i\tau_0 \leqslant \bar{x}_i$ for all $i \geqslant 1$.

b) Consider the classical coupling $\left(X_t^1, X_t^2\right)$ of two copies of the original processes. Let $T_2 = \inf\left\{t \geqslant 0 : X_t^2 = 0\right\}$. By the order-preserving property of the coupling, we have $X_t^1 \leqslant X_t^2$, $\mathbb{P}^{i_1,i_2}$-a.s. for all $i_1 < i_2$. Hence, we obtain $T_2 \geqslant T$, $\mathbb{P}^{i_1,i_2}$-a.s. for all $i_1 < i_2$. Combining this with a) and part (1) of Theorem 3.1, we obtain the required assertion. □

Example 4.6. Consider $a_i = i^2$ and $b_i = i + 2$. We have

$$G_k = \frac{1}{2} + \sum_{s=1}^{k-1} \frac{1}{s!} + \frac{1}{2k!} \uparrow e - \frac{1}{2}.$$

Hence $\bar{x}_1 = e - 1/2$ and so

$$\bar{x}_{k+1} = \sum_{s=0}^{k} F_s(e - 1/2 - G_s) \leqslant 1 + 2\sum_{s=0}^{k} 1/(s+1)^3 < 4.$$

Therefore gap $(D) \geqslant 1/4$. Note that for $a_i \equiv a$ and $b_i \equiv b$,

$$\lim_{i\to\infty} \mathbb{E}^i \tau_0 = \infty$$

and so Corollary 4.5 is not suitable.

5. Appendix. Optimality of the Couplings.

In this section, we fix the sequence (u_i) and let

$$\rho(i,j) = \left| \sum_{k<i} u_k - \sum_{k<j} u_k \right|.$$

We consider the distances of the type $\gamma \circ \rho$ for some $\gamma : [0,\infty) \to [0,\infty)$ with $\gamma(0) = 0$, $\gamma' > 0$ and $\gamma'' \leqslant 0$. For completeness, we will also deal with the case that $\gamma'' \geqslant 0$ (then $\gamma \circ \rho$ is not necessarily a distance). For this, we need another coupling, the *march coupling*:

$$\begin{aligned}\Omega_m^{\rm coup} f(i_1,i_2) = I_{[i_2-i_1\geqslant 1]}\Big\{&(a_{i_1} \wedge a_{i_2})\big[f(i_1-1,i_2-1) - f(i_1,i_2)\big] \\ &+ (b_{i_1} \wedge b_{i_2})\big[f(i_1+1,i_2+1) - f(i_1,i_2)\big] + \cdots\Big\}.\end{aligned}$$

Recall that a coupling $\overline{\Omega}^{\rm coup}$ is called $\gamma \circ \rho$-*optimal* if for every coupling operator $\Omega^{\rm coup}$, we have

$$\overline{\Omega}^{\rm coup} \gamma \circ \rho(i_1,i_2) \leqslant \Omega^{\rm coup} \gamma \circ \rho(i_1,i_2)$$

for all $i_1 \neq i_2$. The next result is an extension of [5; Theorem 3.2 and Theorem 3.3]: By setting $u_i \equiv 1$ or $\gamma(x) = x$, we obtain [5; Theorem 3.2] and [5; Theorem 3.3] respectively. For simplicity, set $a_0 = 0$, define

$$\begin{aligned}&\nabla_i f(x) = f(x+u_i) - f(x) \quad (x \in \mathbf{R},\ i \geqslant 0),\\ &\nabla_j \nabla_i f(x) = \nabla_i \nabla_j f(x) = \nabla_j(\nabla_i f)(x) \quad (i \neq j),\\ &\nabla_i \nabla_i f = 0\end{aligned}$$

and write

$$u_m^n = \sum_{m\leqslant k\leqslant n} u_k.$$

By convention, $u_m^n = 0$ for all $m > n$.

Theorem 5.1. Consider birth-death processes with rates (b_i, a_i).

(1) If $\gamma'' \leqslant 0$, then the coupling by reflection Ω_r^{coup} is $\gamma \circ \rho$-optimal. Moreover,

$$\begin{aligned}\Omega_r^{\mathrm{coup}}\gamma\circ\rho(i,i+k) =& a_i\nabla_{i-1}\gamma(u_i^{i+k-1}) - a_{i+k}\nabla_{i+k-1}\gamma(u_i^{i+k-2})\\ &- b_i\nabla_i\gamma(u_{i+1}^{i+k-1}) + b_{i+k}\nabla_{i+k}\gamma(u_i^{i+k-1})\\ &+ (b_i\wedge a_{i+k})\nabla_i\nabla_{i+k-1}\gamma(u_{i+1}^{i+k-2})\\ &+ (a_i\wedge b_{i+k})\nabla_{i-1}\nabla_{i+k}\gamma(u_i^{i+k-1}),\\ &\quad i\geqslant 0,\quad k\geqslant 1.\end{aligned} \tag{5.1}$$

(2) If $\gamma'' \geqslant 0$, then the march coupling Ω_m^{coup} is $\gamma \circ \rho$-optimal and moreover,

$$\begin{aligned}&\Omega_m^{\mathrm{coup}}\gamma\circ\rho(i,i+k) =\\ &a_i\nabla_{i-1}\gamma(u_i^{i+k-1}) - a_{i+k}\nabla_{i+k-1}\gamma(u_{i-1}^{i+k-2}) - b_i\nabla_i\gamma(u_{i+1}^{i+k}) + b_{i+k}\nabla_{i+k}\gamma(u_i^{i+k-1})\\ &+ b_i\wedge b_{i+k}\nabla_i\nabla_{i+k}\gamma(u_{i+1}^{i+k-1}) + a_{i+k}\wedge a_i\nabla_{i-1}\nabla_{i+k-1}\gamma(u_i^{i+k-2}),\\ &\quad i\geqslant 0,\ k\geqslant 1.\end{aligned} \tag{5.2}$$

(3) If $\gamma'' \equiv 0$, then the above two couplings and the classical coupling are all ρ-optimal and moreover

$$\Omega_c^{\mathrm{coup}}\rho(i,i+k) = a_iu_{i-1} - a_{i+k}u_{i+k-1} - b_iu_i + b_{i+k}u_{i+k}. \tag{5.3}$$

Proof. We prove here part (1) of the theorem only. The proof for the second part is similar[4].

a) Clearly, any coupling operator Ω^{coup} for birth-death processes should have the following form:

$$\begin{aligned}&\Omega^{\mathrm{coup}}f(i_1,i_2)\\ &= I_{[i_1\neq i_2]}\Big\{\lambda_1\big[f(i_1-1,i_2-1)-f(i_1,i_2)\big] + \lambda_2\big[f(i_1+1,i_2+1)-f(i_1,i_2)\big]\\ &\qquad + \lambda_3\big[f(i_1+1,i_2)-f(i_1,i_2)\big] + \lambda_4\big[f(i_1-1,i_2)-f(i_1,i_2)\big]\\ &\qquad + \lambda_5\big[f(i_1,i_2+1)-f(i_1,i_2)\big] + \lambda_6\big[f(i_1,i_2-1)-f(i_1,i_2)\big]\\ &\qquad + \lambda_7\big[f(i_1+1,i_2-1)-f(i_1,i_2)\big] + \lambda_8\big[f(i_1-1,i_2+1)-f(i_1,i_2)\big]\Big\}\\ &\quad + I_{[i_1=i_2]}\Big\{b_{i_1}\big[f(i_1+1,i_2+1)-f(i_1,i_2)\big] + a_{i_1}\big[f(i_1-1,i_2-1)-f(i_1,i_2)\big]\Big\},\end{aligned} \tag{5.4}$$

where $\lambda_j \geqslant 0$ and

$$\lambda_1=\lambda_4=\lambda_8=0 \quad\text{if}\quad i_1=0, \qquad \lambda_1=\lambda_6=\lambda_7=0 \quad\text{if}\quad i_2=0. \tag{5.5}$$

By the marginality, we have

$$\lambda_1+\lambda_4+\lambda_8=a_{i_1},\ \lambda_2+\lambda_3+\lambda_7=b_{i_1},\ \lambda_1+\lambda_6+\lambda_7=a_{i_2},\ \lambda_2+\lambda_5+\lambda_8=b_{i_2}. \tag{5.6}$$

Hence,

$$\lambda_1 = a_{i_2} - \lambda_6 - \lambda_7, \qquad \lambda_2 = b_{i_1} - \lambda_3 - \lambda_7,$$
$$\lambda_4 = a_{i_1} - a_{i_2} + \lambda_6 + \lambda_7 - \lambda_8, \qquad \lambda_5 = b_{i_2} - b_{i_1} + \lambda_3 + \lambda_7 - \lambda_8.$$

Substituting these into (5.4), we get for $i_1 \neq i_2$

$$\begin{aligned}
&\Omega^{\rm coup} f(i_1, i_2) \\
&= a_{i_2}\big[f(i_1 - 1, i_2 - 1) - f(i_1, i_2)\big] + b_{i_1}\big[f(i_1 + 1, i_2 + 1) - f(i_1, i_2)\big] \\
&\quad + (a_{i_1} - a_{i_2})\big[f(i_1 - 1, i_2) - f(i_1, i_2)\big] + (b_{i_2} - b_{i_1})\big[f(i_1, i_2 + 1) - f(i_1, i_2)\big] \\
&\quad + \lambda_3\big[f(i_1 + 1, i_2) + f(i_1, i_2 + 1) - f(i_1 + 1, i_2 + 1) - f(i_1, i_2)\big] \\
&\quad + \lambda_6\big[f(i_1, i_2 - 1) + f(i_1 - 1, i_2) - f(i_1 - 1, i_2 - 1) - f(i_1, i_2)\big] \\
&\quad + \lambda_7\big[f(i_1 + 1, i_2 - 1) + f(i_1 - 1, i_2) + f(i_1, i_2 + 1) \\
&\qquad\quad - f(i_1 - 1, i_2 - 1) - f(i_1 + 1, i_2 + 1) - f(i_1, i_2)\big] \\
&\quad + \lambda_8\big[f(i_1 - 1, i_2 + 1) + f(i_1, i_2) - f(i_1 - 1, i_2) - f(i_1, i_2 + 1)\big]. \qquad (5.7)
\end{aligned}$$

b) We now minimize $\Omega^{\rm coup}\gamma \circ \rho(i_1, i_2)$ under the marginality (5.6). To do so, let $i_1 = i \geqslant 0$, $i_2 = i + k$ for some $k \geqslant 2$. Then, by (5.7), we obtain

$$\begin{aligned}
&\Omega^{\rm coup}\gamma \circ \rho(i, i + k) = \\
&a_i \nabla_{i-1}\gamma(u_i^{i+k-1}) - a_{i+k}\nabla_{i+k-1}\gamma(u_{i-1}^{i+k-2}) - b_i\nabla_i\gamma(u_{i+1}^{i+k}) + b_{i+k}\nabla_{i+k}\gamma(u_i^{i+k-1}) \\
&\quad + \lambda_3\nabla_i\nabla_{i+k}\gamma(u_{i+1}^{i+k-1}) + \lambda_6\nabla_{i-1}\nabla_{i+k-1}\gamma(u_i^{i+k-2}) \\
&\quad + \lambda_7\big[\nabla_i\nabla_{i+k}\gamma(u_{i+1}^{i+k-1}) + \nabla_{i-1}\nabla_{i+k-1}\gamma(u_i^{i+k-2}) + \nabla_i\nabla_{i+k-1}\gamma(u_{i+1}^{i+k-2})\big] \\
&\quad + \lambda_8\nabla_{i-1}\nabla_{i+k}\gamma(u_i^{i+k-1}). \qquad (5.8)
\end{aligned}$$

Here, we have used the convention $a_0 = 0$ and $a_i\nabla_{i-1}\gamma = 0$ whenever $i = 0$. By assumption, $\gamma' > 0$ and $\gamma'' \leqslant 0$. It follows that $\nabla_i\gamma > 0$ and $\nabla_i\nabla_j\gamma \leqslant 0$. Now, following the proof c) of [5; Theorem 3.2], we obtain $\lambda_7 = b_{i_1} \wedge a_{i_2}$, $\lambda_6 = (a_{i_2} - b_{i_1})^+$, $\lambda_3 = (b_{i_1} - a_{i_2})^+$, $\lambda_1 = \lambda_2 = 0$, $\lambda_8 = a_{i_1} \wedge b_{i_2}$, $\lambda_4 = (a_{i_1} - b_{i_2})^+$ and $\lambda_5 = (b_{i_2} - a_{i_1})^+$. Substituting these into (5.8) and collecting the terms, we obtain (5.1) in the case of $k \geqslant 2$.

c) When $k = 1$, everything is the same as in (5.8) except the coefficient of λ_7 which now becomes $\nabla_i\gamma(u_{i-1}) + \nabla_{i+1}\gamma(u_i) > 0$. Now, the proof d) of [5; Theorem 3.2] gives us $\lambda_6 = a_{i_2}$, $\lambda_3 = b_{i_1}$, $\lambda_1 = \lambda_2 = \lambda_7 = 0$, $\lambda_8 = a_{i_1} \wedge b_{i_2}$, $\lambda_4 = (a_{i_1} - b_{i_2})^+$ and $\lambda_5 = (b_{i_2} - a_{i_1})^+$. Therefore, we get

$$\begin{aligned}
\Omega^{\rm coup}\gamma \circ \rho(i, i + 1) = &-(b_i + a_{i+1})\gamma(u_i) + a_i\nabla_{i-1}\gamma(u_i) + b_{i+1}\nabla_{i+1}\gamma(u_i) \\
&+ (a_i \wedge b_{i+1})\nabla_{i-1}\nabla_{i+1}\gamma(u_i), \quad i \geqslant 0. \qquad (5.9)
\end{aligned}$$

Note that (5.9) coincides with (5.1) by the convention: $\nabla_i\nabla_i = 0$ and $u_m^n = 0$ for $m > n$. □

6. Modification (Unpublished).

Improvement of part (4) **of Corollary** 1.2**.**

(4) If $a_i = b_i$ and $i^\delta \sum_{j\geqslant i} 1/a_j \leqslant c_\delta$ for some $\delta \geqslant 1$ and all $i \geqslant 1$, then $\operatorname{gap}(D) \geqslant \max\{(4c_1)^{-1}, (1-\delta^{-1})c_\delta^{-1}\}$.

Improvement of part (2) of Lemma 2.2. If $\sum_{j\geqslant i} m_j \leqslant c_\delta i^{-\delta}$ for some $\delta \geqslant 1$ and all $i \geqslant 1$, then

$$\sum_{j\geqslant i} j^\gamma m_j \leqslant c_\delta\biggl\{i^{\gamma-\delta} + \sum_{j\geqslant i} \frac{1}{(j+1)^\delta}\Bigl[(j+1)^\gamma - j^\gamma\Bigr]\biggr\}, \qquad i \geqslant 1,\ \ \gamma \in [0,\delta).$$

Proof. Set $M_i = \sum_{j\geqslant i} m_i$. Then

$$\begin{aligned}\sum_{j\geqslant i} j^\gamma m_j &= i^\gamma M_i + \sum_{j\geqslant i}\Bigl[(j+1)^\gamma - j^\gamma\Bigr] M_{j+1}\\ &\leqslant c_\delta\biggl\{i^{\gamma-\delta} + \sum_{j\geqslant i} \frac{1}{(j+1)^\delta}\Bigl[(j+1)^\gamma - j^\gamma\Bigr]\biggr\}. \quad \square\end{aligned}$$

Proof of part (4) *of Corollary* 1.2. a) We need to estimate the upper bound of

$$\begin{aligned}&\frac{1}{i^\gamma - (i-1)^\gamma}\biggl\{i^{\gamma-\delta} + \sum_{j\geqslant i} \frac{1}{(j+1)^\delta}\bigl[(j+1)^\gamma - j^\gamma\bigr]\biggr\}\\ &= \frac{i^{\gamma-\delta}}{i^\gamma - (i-1)^\gamma} + \frac{1}{i^\gamma - (i-1)^\gamma}\sum_{j\geqslant i}\frac{(j+1)^\gamma - j^\gamma}{(j+1)^\delta}. \qquad (6.1)\end{aligned}$$

When $\delta = 1$, the original proof shows that $\operatorname{gap}(D) \geqslant (4c_1)^{-1}$. Then the same estimate holds whenever $\delta \geqslant 1$.

b) We are going to improve the estimate when $\delta > 1$. First, we prove that the function in (6.1) is decreasing in i whenever $\delta \geqslant \gamma \geqslant 1$.

i) Note that the function

$$f(x) := \frac{x^{\gamma-\delta}}{x^\gamma - (x-1)^\gamma}$$

is decreasing on $[1,\infty)$ whenever $\delta \geqslant \gamma \geqslant 1$. Actually,

$$\begin{aligned}f'(x) < 0 &\iff (\gamma-\delta)[x^\gamma - (x-1)^\gamma] < \gamma x[x^{\gamma-1} - (x-1)^{\gamma-1}]\\ &\iff \gamma x(x-1)^{\gamma-1} + (\delta-\gamma)(x-1)^\gamma < \delta x^\gamma\\ &\iff \gamma\biggl(1-\frac{1}{x}\biggr)^{\gamma-1} + (\delta-\gamma)\biggl(1-\frac{1}{x}\biggr)^\gamma < \delta.\end{aligned}$$

The last inequality is obvious.

ii) We now consider the second term in (6.1).

$$\frac{1}{i^\gamma-(i-1)^\gamma}\sum_{j\geqslant i}\frac{1}{(j+1)^\delta}\big[(j+1)^\gamma-j^\gamma\big].$$

It is decreasing iff

$$\frac{\sum_{j\geqslant i+1}(j+1)^{-\delta}\big[(j+1)^\gamma-j^\gamma\big]}{\sum_{j\geqslant i}(j+1)^{-\delta}\big[(j+1)^\gamma-j^\gamma\big]}<\frac{(i+1)^\gamma-i^\gamma}{i^\gamma-(i-1)^\gamma}.$$

Since the left-hand side is less than 1, it suffices to show that $(i+1)^\gamma+(i-1)^\gamma\geqslant 2i^\gamma$. However, this is equivalent to

$$\left(1+\frac{1}{i}\right)^\gamma+\left(1-\frac{1}{i}\right)^\gamma\geqslant 2$$

which is then deduced by $(1+x)^\gamma\geqslant 1+\gamma x\,(\gamma\geqslant 1,\ |x|\leqslant 1)$. Hence the second term in (6.1) is also decreasing.

Combining i) and ii) together, we see that for every $\gamma\geqslant 1$, (6.1) attains the maximum at $i=1$:

$$1+\sum_{j\geqslant 1}\frac{1}{(j+1)^\delta}\big[(j+1)^\gamma-j^\gamma\big].$$

Minimizing this with respect to γ, we obtain $\gamma=1$ and then

$$1+\sum_{j\geqslant 1}\frac{1}{(j+1)^\delta}\leqslant 1+\int_1^\infty\frac{\mathrm{d}x}{x^\delta}=1+\frac{1}{\delta-1}.$$

When $\delta=1$, the estimate is trivial and so we obtain the second lower bound. □

We mention that the similar modification holds for part (5) of Corollary 2.5 and part (5) of Corollary 3.3 in the paper "Estimation of spectral gap for elliptic operators" by Mu-Fa Chen and Feng-Yu Wang (Trans. Amer. Math. Soc. 349:2 (1997), 1239–1267).

Correction of proof d) of Lemma 2.1 (*To be published in a subsequent paper*). Define

$$\bar g_i=g_iI_{\{i<n\}}+g_nI_{\{i\geqslant n\}}.$$

Then, we have

$$\sum_i\pi_i\bar g_i^2=\sum_{i\leqslant n-1}\pi_ig_i^2+g_n^2\sum_{i\geqslant n}\pi_i,$$

$$\sum_i\pi_i\bar g_i=\sum_{i\leqslant n-1}\pi_ig_i+g_n\sum_{i\geqslant n}\pi_i=g_n\sum_{i\geqslant n}\pi_i-\pi_n\tilde g_n,$$

$$\sum_i\pi_i\bar g_i^2-\left(\sum_i\pi_i\bar g_i\right)^2=\sum_{i\leqslant n-1}\pi_ig_i^2+g_n^2\sum_{i\geqslant n}\pi_i-\left(g_n\sum_{i\geqslant n}\pi_i-\pi_n\tilde g_n\right)^2,\tag{6.2}$$

$$-\sum_i\pi_i\big(\bar g_i\Omega\bar g_i\big)(i)=\lambda_1\sum_{i\leqslant n-1}\pi_ig_i^2+\pi_na_ng_n(g_n-g_{n-1}).\tag{6.3}$$

We now prove that

$$\pi_n a_n g_n(g_n - g_{n-1}) < \lambda_1 g_n^2 \sum_{i\geqslant n} \pi_i - \lambda_1 \bigg(g_n \sum_{i\geqslant n} \pi_i - \pi_n \tilde{g}_n \bigg)^2. \tag{6.4}$$

Because $\tilde{g}_n = a_n(g_n - g_{n-1})/\lambda_1$, the left-hand side of (6.4) is equal to $\lambda_1 \pi_n g_n \tilde{g}_n$. Moreover, $g_n > 0$. Thus, (6.4) is equivalent to

$$\pi_n \tilde{g}_n / g_n < \sum_{i\geqslant n} \pi_i - \bigg(\sum_{i\geqslant n} \pi_i - \pi_n \tilde{g}_n / g_n \bigg)^2.$$

That is,

$$\bigg(\sum_{i\geqslant n} \pi_i - \pi_n \tilde{g}_n / g_n \bigg)^2 < \sum_{i\geqslant n} \pi_i - \pi_n \tilde{g}_n / g_n.$$

This clearly holds since $0 < \tilde{g}_n \leqslant g_n$,

$$0 < \sum_{i\geqslant n} \pi_i - \pi_n \tilde{g}_n / g_n = \sum_{i\geqslant n+1} \pi_i + \pi_n (1 - \tilde{g}_n / g_n) < 1.$$

Collecting (6.2)–(6.4) together, it follows that

$$\lambda_1 \leqslant \frac{-\sum_i \pi_i \big(\bar{g}_i \Omega \bar{g}_i\big)(i)}{\sum_i \pi_i \bar{g}_i^2 - \big(\sum_i \pi_i \bar{g}_i\big)^2} < \lambda_1$$

which is a contradiction. □

Acknowledgement. The first version of the paper was mainly completed while the author visited the Universities of Rome I and Rome II in the winter of 1994. The author acknowledges the financial support from Univ. of Rome I and Centro V. Volterra at Univ. of Rome II and acknowledges Prof. L. Accardi and Prof. E. Scacciatelli for their hospitality and valuable discussions. The author then turned to study the same topic for elliptic operators with Feng-Yu Wang[10]. From the last study we obtained an analogue of part (2) of Theorem 1.1. The author acknowledges Prof. F. Y. Wang for the fruitful cooperation.

Added in Proof. The author acknowledges Prof. E. A. Doorn for pointing out a different proof of (1.4). That is, combining [2; Theorem 5.3] and [22; Theorem 3.3] with either Theorem 2 of van Doorn's paper in J. Approx. Th.**51**(1987) or [22; (2.30)] and (4.7) of his paper in Adv. Appl. Prob.23(1991). Note that these papers also indicate some application of Theorem 1.1 to other situations.

References

[1] Aldous, D. J. and Brown, M. (1993), *Inequalities for rate events in time-reversible Markov chains*, IMS Lecture Notes-Monograpg Series **22**, Stochastic Inequalities, 1–16.

[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*, Singapore, World Scientific.
[4] Chen, M. F. (1993), *Optimal couplings and application to Riemannian geometry*, in Prob. Theory and Math. Stat., Vol.1, 121–142, Edited by B. Grigelionis et al. 1994 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. (1995), *On ergodic region of Schlögl's model*, In Proc. Intern. Conf. on Dirichlet Forms & Stoch. Proc. Edited by Z. M. Ma, M. Röckner and J. A. Yan, Walter de Gruyter.
[7] 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.
[8] Chen, M. F. and Wang, F. Y. (1995), *Estimation of the first eigenvalue of second order elliptic operators*, J. Funct. Anal. 131:2, 345–363.
[9] Chen, M. F. and Wang, F. Y. (1994), *Estimates of logarithmic Sobolev constant — An improvement of Bakry–Emery criterion*, to appear in J. Funct. Anal.
[10] Chen, M. F. and Wang, F. Y. (1995), *Estimation of spectral gap for elliptic operators*, to appear in Trans. Amer. Math. Soc.
[11] Deuschel, J.-D. and Stroock, D. W. (1990), *Hypercontractivity and spectral gap of symmetric diffusion with applications to the stochastic Ising models*, J. Funct. Anal. 92, 30–48.
[12] Diaconis and Stroock (1991), *Geometric bounds for eigenvalues of Markov chains*, Ann. Appl. Prob. 1:1, 36–61.
[13] Iscoe, I. and McDonald, D. (1994), *Asymptotics of exit times for Markov jump processes* (I), Ann. Prob. 22:1, 372–397.
[14] Jerrum, M. R. and Sinclair, A. J. (1989), *Approximating the permanent*, SIAM J. Comput.18, 1149–1178.
[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] Liggett, T. M. (1989), *Exponential L_2 convergence of attractive reversible nearest particle systems*, Ann. Probab. 17, 403-432.
[17] Sinclair, A. J. and Jerrum, M. R. (1989), *Approximate counting, uniform generation, and rapidly mixing Markov chains*, Inform. and Comput.82, 93–133.
[18] Sokal, A. D. and Thomas, L. E.(1988), *Absence of mass gap for a class of stochastic contour models*, J. Statis. Phys.51:5/6, 907–947.
[19] Sullivan, W. G.(1984), *The L^2 spectral gap of certain positive recurrent Markov chains and jump processes*, Z. Wahrs.67, 387–398.
[20] Tweedie, R. L.(1981), *Criteria for ergodicity, exponential ergodicity and strong ergodicity of Markov Processes*, J. Appl. Prob.18, 122–130.
[21] Van Doorn, E.(1981), *Stochastic Monotonicity and Queueing Applications of Birth-Death Processes*, Lecture Notes in Statistics 4, Springer.
[22] 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.
[23] Wang, F. Y. (1994), *Application of coupling method to the Neumann eigenvalue problem*, Prob. Th. Rel. Fields 98, 299–306.
[24] Wang, F. Y. (1995), *Spectral gap for diffusion processes on non-compact manifolds*, Chin. Sci. Bull. 40:14, 1145–1149.
[25] Zeifman, A. I. (1991), *Some estimates of the rate of convergence for birth and death processes*, J. Appl. Prob. 28, 268–277.

Received December 27, 1995.

Department of Mathematics, Beijing Normal University, Beijing 100875, The People's Republic of China.

Trans. Amer. Math. Soc. 349:3 (1997), 1239–1267.

ESTIMATION OF SPECTRAL GAP FOR ELLIPTIC OPERATORS

Mu-Fa Chen and Feng-Yu Wang

(Beijing Normal University)
October 23, 1995

Abstract. A variational formula for the lower bound of the spectral gap of an elliptic operator is presented in the paper for the first time. The main known results are either recovered or improved. A large number of new examples with sharp estimate are illustrated. Moreover, as an application of the march coupling[4], the Poincaré inequality with respect to the absolute distribution of the process is also studied.

1. Introduction

Consider the operator

$$L = \sum_{i,j=1}^{d} a_{ij}(x)\partial_i\partial_j \ + \ \sum_{i=1}^{d} b_i(x)\partial_i,$$

where $\partial_i = \frac{\partial}{\partial x_i}$, $a(x) := (a_{ij}(x))$ is positive definite, $a_{ij} \in C^2(\mathbf{R}^d)$ and

$$b_i = \sum_{j=1}^{d}(a_{ij}\partial_j V + \partial_j a_{ij})$$

for some $V \in C^2(\mathbf{R}^d)$ with $Z := \int \exp[V(x)]\mathrm{d}x < \infty$. We denote L by $L \sim (a, b)$ or $L \sim (a, V)$ and let $\pi(dx) = Z^{-1}\exp[V(x)]dx$.

Throughout of this paper, we assume that the L-diffusion process is non-explosive so that the corresponding Dirichlet form is regular. Then the first (non-trivial) eigenvalue λ_1 or the spectral gap can be characterized as

$$\mathrm{gap}(L) = \inf\big\{\pi\big(\langle a\nabla f, \nabla f\rangle\big) : f \in \mathscr{D}, \pi(f) = 0, \pi(f^2) = 1\big\}, \tag{1.1}$$

2000 *Mathematics Subject Classification.* 35P15, 60H30.

Key words and phrases. Spectral gap, diffusion process, coupling.

Research supported in part by the National Natural Science Foundation of China and the Foundation of Institution of Higher Education for Doctoral Program

where $\pi(f) = \int f \mathrm{d}\pi$ and $\mathscr{D} = \{f + c : f \in C_0^\infty(\mathbf{R}^d),\ c \in \mathbf{R}\}$. The variational formula (1.1) is particularly useful for a upper bound of $\mathrm{gap}(L)$. But it is much more difficult to handle the lower bound for which many different approaches have been introduced. The readers are urged to refer to [6] for further comments and references.

To show the difficulty of the problem, we mention here three simple examples. Let $d = 1$ and take $a \equiv 1, b(x) = -x$. Then the first eigenvalue is $\lambda_1 = 1$. We now go to the half line $[0, \infty)$ with reflecting boundary and with the same a. Then $\lambda_1 = 2$ or 3 according as $b(x) = -x$ or $-(x + 1)$. Surprisingly, the order of the corresponding eigenfunctions changes from 1, 2 to 3 successively. From these, one sees that the first eigenvalue is very sensitive.

To get some impression about the results obtained in the paper, let us restrict ourselves to the half line $[0, \infty)$. Denote by $\mathscr{F}$ the set of all functions $f \in L^1(\pi)$ with $f' > 0$ on $(0, \infty)$. Define $C(x) = \int_0^x a(y)^{-1} b(y) \mathrm{d}y$. We will use quite often the following mapping $I : \mathscr{F} \to C[0, \infty)$ or its variations.

$$I(f)(x) = \frac{e^{-C(x)}}{f'(x)} \int_x^\infty \frac{f(u)e^{C(u)}}{a(u)} \mathrm{d}u = \frac{e^{-V(x)}}{a(x)f'(x)} \int_x^\infty f(u)e^{V(u)} \mathrm{d}u, \quad x > 0, f \in \mathscr{F}. \tag{1.2}$$

Then, we have

$$\mathrm{gap}_{[0,\infty)} \geqslant \sup_{f \in \mathscr{F}} \inf_{x>0} I(f)(x)^{-1}. \tag{1.3}$$

This is an alternative statement of Theorem 2.1 (2) given below. No doubt, this is a very convenient formula since it is usually quite easy to choose a test function $f \in \mathscr{F}$ to obtain a non-trivial estimate. Moreover, it is proved that equality in (1.3) actually holds in the regular case (cf. Proposition 6.4). This new variational formula is clearly a dual of (1.1). It is remarkable that the two formulas have no common point.

This paper is based on a new probabilistic method, i.e. the coupling approach, introduced by the authors in [5] and further developed in [3], [6], [15] and [16]. For the reader's convenience, let us explain briefly the main ideas of the method. First, we construct some degenerated elliptic operators $\widetilde{L}$ on the product space $\mathbf{R}^d \times \mathbf{R}^d$ so that $\widetilde{L} f_i(x_1, x_2) = Lf(x_i)$ for $i = 1, 2$, all $f \in C_b^2(\mathbf{R}^d)$ and all $x_1 \neq x_2$, where $f_i(x_1, x_2) = f(x_i)$, $i = 1, 2$, $x_1, x_2 \in \mathbf{R}^d$. The operator $\widetilde{L}$ is then called a coupling of L (see [3] or [4] for details). Next, choose a distance $d(x, y)$ in $\mathbf{R}^d$. Our main estimate comes from the following inequality

$$\widetilde{L} d(x, y) \leqslant -\delta d(x, y), \qquad \text{for all } \ x \neq y \tag{1.4}$$

where $\widetilde{L}$ is a coupling operator and $\delta > 0$ is a constant. From this, we deduce that $\mathrm{gap}(L) \geqslant \delta$. Certainly, we have ignored a lot of technical points in this step. Anyhow, from (1.4), one sees that the estimate depends heavily on the choice of both the coupling operator $\widetilde{L}$ and the distance $d(x, y)$. On the other hand, it is known from [3] that the couplings $\widetilde{L}$ can be classified according to different classes of distances and moreover for each class (usually quite large) of distances,

there often (sometimes uniquely) exists an optimal $\widetilde{L}$. Therefore, constructing a "good" distance plays a critical role in the study of estimates of the spectral gap (as well as many applications of the coupling approach), as illustrated in our recent publications.

The second key point of our method is that the eigenfunction of λ_1 has to be Lipschitz with respect to the distance adopted. This once again gives the choice of the distance a serious influence on the effectiveness of the approach, especially for non-compact spaces. From this point of view, our approach seems quite restrictive. For instance, in [6] we were unable to cover completely the one-dimensional case for which we employed an analytic approach, a continuous analog of [13]. However, this serious problem turns out to be helpful. It provides us a way to construct some effective distances. That is, roughly speaking, choosing the distance from the eigenfunction or its approximations. Fortunately, this idea is successful as one will see soon in the next section. This paper should be considered as a critical step in the study of couplings and the idea of the paper should be useful in various applications of the coupling method as well as in the study of related topics.

Since the topic is quite technical as one can imagine, we choose a special way to organize the paper. Starting from the simplest case, i.e. the half line (Section 2), then go to the full line (Section 3) and finally studying the general case (Section 4). In each section, we introduce the results, explain the ideas and present a large number of examples (which should be considered as a critical part of the paper) in illustration of the results. One sees in a gradual way how the ideas move from a special case to the general one. The proofs are shorter than the statements of the results. Having some preparations (Section 5) at hand, the proofs of the results (except one) of Sections 2–4 are given in Sections 6–8 respectively. The equality in (1.3) is explored at the end of Section 6 and Section 7.

This paper is a continuation of [6] but it is nevertheless self-contained. Some ideas come from our previous papers, not only from the study on the estimate of the spectral gap but also from the study of the estimate of Logarithmic Sobolev inequality (see [7], [16], [17] and references therein). Besides, the paper is also an interaction with the study of the same topic for Markov chains and with the study on path space[10,17]. In particular, a result on the Poincaré inequality with respect to the absolute distributions of the process is included in Section 4 and proved in Section 9. Finally, the paper [12], introduced to one of the authors by S. Kotani, is very helpful.

2. The Case of Half Line

Consider a reflecting diffusion on the half line $[x_0, \infty)$ with operator $L \sim (a, b)$. Set $C(x) = \int_{x_0}^{x} b(u)a(u)^{-1}\mathrm{d}u$. Then, the condition "$Z < \infty$" and the well-known Feller's non-explosive criterion can be stated as follows.

$$Z = \int_{x_0}^{\infty} \frac{\mathrm{e}^{C(x)}}{a(x)}\mathrm{d}x < \infty, \qquad \int_{x_0}^{\infty} \mathrm{d}x\mathrm{e}^{-C(x)} \int_{x_0}^{x} \frac{\mathrm{e}^{C(y)}}{a(y)}\mathrm{d}y = \infty. \tag{2.1}$$

The left-end point of the half line is not essential in this section but it will be critical in the next section. To emphasize the half line, we use $\mathrm{gap}_{[x_0,\infty]}$ instead

of gap(L). Recall that the mapping $I(f)$ was defined in (1.2) but in which the function $C(x)$ is replaced by the one just defined here.

Theorem 2.1. Assume that (2.1) holds.

(1) For every function $f \in C^1[x_0,\infty) \cap C^2(x_0,\infty)$ with $f > 0$ on (x_0,∞), we have

$$\begin{aligned} \mathrm{gap}_{[x_0,\infty]} &\geqslant \inf_{x>x_0} \big[(-af' - bf)'/f\big](x) && (2.2)\\ &= \inf_{x>x_0} \big[- b' - \big(af'' + (a'+b)f'\big)/f\big](x). && (2.3) \end{aligned}$$

(2) For every function $f \in C[x_0,\infty) \cap C^1(x_0,\infty) \cap L^1(\pi)$ with $\pi(f) \geqslant 0$ and $f' > 0$ on (x_0,∞), we have

$$\mathrm{gap}_{[x_0,\infty)} \geqslant \inf_{x>x_0} I(f)(x)^{-1}. \tag{2.4}$$

In particular, if moreover $f \in C^2[x_0,\infty)$, then

$$\mathrm{gap}_{[x_0,\infty]} \geqslant c \quad \text{provided} \quad -(af'' + bf') \geqslant cf \quad \text{for some} \quad c > 0. \tag{2.5}$$

Remark 2.2. (1) At the first look, the differentiation form (2.2) and the integration form (2.4) seem quite different but they are indeed equivalent. To see this, let f_2 be given in part (2) such that the right-hand side of (2.4) is positive. Take $f_1 = f_2' I(f_2)$, then $f_1' > 0$ on (x_0,∞). Since

$$f_1'(x) = -\frac{b(x)}{a(x)} e^{-C(x)} \int_x^\infty \frac{f_2(y) e^{C(y)}}{a(y)} \mathrm{d}y - \frac{f_2(x)}{a(x)},$$

we have $-af_1' - bf_1 = f_2$. Hence

$$[-af_1' - bf_1]'/f_1 = f_2'/f_1 = I(f_2)^{-1}. \tag{2.6}$$

Then (2.2) implies (2.4).

Next, let f_1 be given in part (1) such that the right-hand side of (2.2) is positive. Fix $p > x_0$, let $c_1 = f_1(p) e^{C(p)} (\int_{x_0}^p a^{-1} e^C \mathrm{d}x)^{-1}$. Set $f = c_1 - af_1' - bf_1$. Then $f \in C[x_0,\infty) \cap C^1(x_0,\infty)$ and $f' > 0$ on (x_0,∞). Since

$$\int_{x_0}^p \frac{f \mathrm{e}^C}{a} \mathrm{d}x = f_1(x_0) e^{C(x_0)} \geqslant 0,$$

we have $f(x) > 0$ for $x > p$ and

$$0 < \int_{x_0}^\infty \frac{f e^C}{a} \mathrm{d}x = f_1(x_0) e^{C(x_0)} + c_1 Z - \lim_{y\to\infty} f_1(y) e^{C(y)}. \tag{2.7}$$

Hence $c := \lim_{y\to\infty} f_1(y)e^{C(y)} \geqslant 0$ exists and is finite. Now, we set $f_2 = c/Z - af_1' - bf_1$. Then $f_2 \in C[x_0,\infty)\cap C^1(x_0,\infty)$, $f_2' > 0$ on (x_0,∞) and $\pi(f_2) = Z^{-1}\int_{x_0}^{\infty} a^{-1}f_2e^C\mathrm{d}x = Z^{-1}(f_1e^C)(x_0) \geqslant 0$. Finally, it is easy to see that

$$I(f_2)^{-1} \geqslant f_2'/f_1 = [-af_1' - f_1]'/f_1. \tag{2.8}$$

Then (2.4) implies (2.2).

Of course, each of (2.2) and (2.4) has its own advantage. The computation for (2.2) is much easier than (2.4). While, (2.4) is very helpful to see whether the spectral gap is positive or not and to find out an effective test function f. The last differential form (2.5) is deduced from (2.4), it is generally weaker than (2.4) and hence weaker than (2.2). But for specific f, (2.5) is not comparable with (2.2). See also Example 2.12 below.

(2). Next, if the function f is the derivative of the eigenfunction corresponding to the first non-trivial eigenvalue $\lambda_1 = \mathrm{gap}(L)$, then the function $-\big[af' + bf\big]'/f$, given on the right-hand side of (2.2), equals λ_1 identically. Conversely, if the function just mentioned is a constant $\alpha > 0$ and the function

$$g(x) := c_0 + \int_{x_0}^{x} f(y)\mathrm{d}y, \qquad c_0 := -\frac{(af')(x_0)}{\alpha} \tag{2.9}$$

belongs to $L^2(\pi)$ with $f(x_0) = 0$ and $\lim_{x\to\infty} f(x)e^{C(x)} = 0$, then g is indeed an eigenfunction (cf. Lemma 6.2) and so the lower bound α given by (2.2) is sharp. In this way, one may construct many examples for which our estimates are exact. Due to the correspondence explained in (1), a similar conclusion holds for the estimate (2.4).

(3). In general, the idea is to regard functions g of the form

$$c + \int_{x_0}^{x} f(y)\mathrm{d}y \qquad \text{or} \qquad c + \int_{x_0}^{x} f'(y)I(f)(y)\mathrm{d}y. \tag{2.10}$$

as an approximation of the eigenfunction. To examine the effectiveness of the approximation, when $g \in L^2(\pi)$, simply note by (1.1) that

$$\mathrm{gap}_{[x_0,\infty)} \leqslant \frac{1}{\pi\big(g^2\big) - (\pi g)^2}\int_{x_0}^{\infty} ag'^2\mathrm{d}\pi. \tag{2.11}$$

In the case of $g \notin L^2(\pi)$, instead of (2.11), we adopt

$$\mathrm{gap}_{[x_0,\infty)} \leqslant \varliminf_{n\to\infty} \frac{1}{\pi_n\big(g^2\big) - (\pi_n g)^2}\int_{x_0}^{n} ag'^2\mathrm{d}\pi_n, \tag{2.12}$$

where $\pi_n(\mathrm{d}x) = I_{[x_0,n)}(x)\pi(\mathrm{d}x)/\int_{x_0}^{n}\pi(\mathrm{d}y)$ (cf. Lemma 5.1). Furthermore, if $g \in L^1(\pi)\setminus L^2(\pi)$, then (2.12) becomes

$$\mathrm{gap}_{[x_0,\infty)} \leqslant \varliminf_{n\to\infty} \frac{1}{\pi_n\big(g^2\big)}\int_{x_0}^{n} ag'^2\mathrm{d}\pi_n. \tag{2.13}$$

Clearly, for each test function f, we obtain from (2.3) a lower bound for the spectral gap. The correspondence of some elementary functions f and the lower bounds are listed below.

Corollary 2.3.

(1) $f(x)=(c_1+x-x_0)^\delta$, $c_1\geqslant 0, \delta\in\mathbf{R}$.

$$\mathrm{gap}_{[x_0,\infty)}\geqslant\begin{cases}\inf\limits_{x>x_0}\Big[-b'(x)-\dfrac{\delta(\delta-1)a(x)}{(c_1+x-x_0)^2}-\dfrac{\delta(a'+b)(x)}{c_1+x-x_0}\Big]\\ \inf\limits_{x>x_0}\big[-b'(x)\big] \qquad \text{if}\quad \delta=0\\ \inf\limits_{x>x_0}\Big[-b'(x)-\dfrac{(a'+b)(x)}{c_1+x-x_0}\Big] \qquad \text{if}\quad \delta=1.\end{cases}$$

(2) $f(x)=(c_1+c_2(x-x_0))e^{\delta(x-x_0)}$, $c_1,c_2\geqslant 0$, $c_1+c_2>0$ and $\delta\in\mathbf{R}$.

$$\mathrm{gap}_{[x_0,\infty)}\geqslant\inf_{x>x_0}\Big[-b'(x)-\delta^2a(x)-\delta\big(a'+b\big)(x)-\frac{c_2}{c_1+c_2(x-x_0)}\big[2\delta a+a'+b\big](x)\Big].$$

(3) $f(x)=c_1+c_2(x-x_0)+(x-x_0)^2$, $c_2>-2\sqrt{c_1}$ or $c_1=c_2=0$.

$$\mathrm{gap}_{[x_0,\infty)}\geqslant\inf_{x>x_0}\Big[-b'(x)-\frac{2a(x)+\big(a'+b\big)(x)\big(c_2+2(x-x_0)\big)}{c_1+c_2(x-x_0)+(x-x_0)^2}\Big].$$

By Corollary 2.3, it is easy to obtain some explicit estimates.

Corollary 2.4.

(1) If there exist c_1 and $\varepsilon\leqslant 1$ (resp. $\varepsilon\geqslant 1$) such that $(a'+b)(c_1+x-x_0)\leqslant\varepsilon a$ (resp. $(a'+b)(c_1+x-x_0)\geqslant\varepsilon a$), then $\mathrm{gap}_{[x_0,\infty)}\geqslant\inf_{x>x_0}\Big[\dfrac{a(x)(1-\varepsilon)^2}{4(c_1+x-x_0)^2}-b'(x)\Big]$.

(2) If there exists ε_1, $\varepsilon_2\leqslant 0$ such that $a'+b\leqslant\big(\varepsilon_1+\varepsilon_2(x-x_0)\big)a$, then $\mathrm{gap}_{[x_0,\infty)}\geqslant\inf_x\Big[\Big(\dfrac{\varepsilon_1^2}{4}-\varepsilon_2\Big)a(x)-b'(x)\Big]$.

(3) If $a'+b\leqslant\big(\varepsilon_1+\varepsilon_2(x-x_0)\big)a$ for some $\varepsilon_1\leqslant-\sqrt{-\varepsilon_2}<0$, then $\mathrm{gap}_{[x_0,\infty)}\geqslant\inf_x\{-b'(x)-2\varepsilon_2a(x)\}$.

The next result is deduced from (2.4). Sometimes, it is convenient to decompose the function f given in Theorem 2.1 (2) as $f=f_1+c$ for some $f_1\geqslant 0$ and $c\leqslant\pi(f)$.

Corollary 2.5.

(1) Suppose that $\inf_{x>x_0}a(x)/(c_1+x-x_0)^\gamma:=c>0$ for some $c_1>0$ and $\gamma\geqslant 2$. If there exists $\varepsilon\in(-\infty,\gamma-1)$ such that $(c_1+x-x_0)b(x)\leqslant\varepsilon a(x)$ for large enough x (resp. for all $x\in[x_0,\infty)$), then $\mathrm{gap}_{[x_0,\infty)}>0$ (resp. $\mathrm{gap}_{[x_0,\infty)}\geqslant\frac{c}{4}(\gamma-1-\varepsilon)^2c_1^{\gamma-2}$). When $\gamma=2$ and $c_1=0$, the same conclusion holds by removing the term $c_1^{\gamma-2}$.

(2) If there exist some ε_1 and ε_2, either $\varepsilon_2=0$ and $\varepsilon_1<0$ or $\varepsilon_2<0$, such that $b(x)\leqslant(\varepsilon_1+\varepsilon_2(x-x_0))a(x)$ for large enough x, then $\mathrm{gap}_{[x_0,\infty)}>0$. Furthermore if the condition holds for all $x\in[x_0,\infty)$, then

$$\mathrm{gap}_{[x_0,\infty)}\geqslant\max\Big\{\frac{1}{4}(\varepsilon_1\wedge 0)^2-\varepsilon_2,\ \ -\varepsilon_2\Big[1+\int_0^\infty e^{u+\varepsilon_2u^2/(2(\varepsilon_1\vee 0)^2)}\mathrm{d}u\Big]^{-1},\\ \frac{1}{4}\Big[\int_0^\infty e^{\varepsilon_1u+\varepsilon_2u^2/2}\mathrm{d}u\Big]^{-2}\Big\}\inf_x a(x).$$

(3) If $c_1 := \sup_{x>x_0} e^{-C(x)} \int_x^\infty \frac{e^{C(u)}}{a(u)}\mathrm{d}u < \infty$ and
$c_2 := \sup_{x>x_0} e^{-C(x)} \int_x^\infty e^{C(u)}\mathrm{d}u < \infty$, then $\mathrm{gap}_{[x_0,\infty)} \geqslant 1/(4c_1c_2)$.

(4) If $c := \sup_{x>x_0} \frac{a(x)}{e^{C(x)}} \int_x^\infty \frac{e^{C(u)}}{a(u)}\mathrm{d}u < \infty$, then $\mathrm{gap}_{[x_0,\infty)} \geqslant \inf_x a(x)/(4c^2)$.
In particular, if $\lim_{x\to\infty} e^{C(x)}/a(x) = 0$ and
$\overline{\lim}_{x\to\infty} a(x)/[a'(x) - b(x)] < \infty$, then $\mathrm{gap}_{[x_0,\infty)} > 0$.

(5) If $b \equiv 0$, then $\mathrm{gap}_{[x_0,\infty)} \geqslant \left\{4\sup_{x>x_0}(x-x_0)\int_x^\infty a(u)^{-1}\mathrm{d}u\right\}^{-1}$.

Observe that it is usually not difficult to find a test function so that the estimates (2.2) and (2.4) are non-trivial out of a local region. That is, replacing "$x > x_0$" with "$x > N$" for large enough N, we obtain a positive lower bound. For instance, if $a(x) \equiv 1$, then the function $f(x) = \exp[-\varepsilon C(x)]$, ($\varepsilon \in (0,1)$) works for (2.4) out of a local region. Next, if $\inf_{x>N}\big[-b'(x)\big] > 0$, then the function $f(x) = x$ is enough for (2.2) out of a local region. We are now going to show that this is indeed sufficient for a non-trivial estimate since we can always modify the test function so that the infimum over the whole space $[x_0,\infty)$ is positive. Besides, the results given below actually provide us some optimizing methods to improve the resulting estimate.

Corollary 2.6. Given $f \in C^1[x_0,\infty)$ with $\pi(f) \geqslant 0$, $f'(x) > 0$ for large enough x and

$$\overline{\lim_{x\to\infty}}\, I(f)(x) < \infty. \tag{2.14}$$

Then, we have

$$\mathrm{gap}_{[x_0,\infty)} \geqslant \sup_{c>0}\inf_{x>x_0} I(f_1)(x)^{-1} > 0, \tag{2.15}$$

where $f_1(x) = cx/(1+x) + f(x)$.

This corollary is deduced from (2.4) by using f_1 instead of the original f. The additional term $cx/(1+x)$ changes the sign of f' locally but without interfering with the convergence in (2.14). The next corollary is quite convenient in practice since the test function is fixed and it is also very effective if the decay of the drift $b(x)$ is not slower than linear.

To state the result, we need some notations which will be used several times in what follows. Let $K \in C(x_0, D)$ be a non-decreasing function so that $(x - x_0)K(x)/a(x)$ is locally integrable. Define

$$F^r(s) = \int_{x_0}^{s\wedge r} \frac{u-x_0}{a(u)}\big[K(r) - K(u)\big]\mathrm{d}u, \qquad r \in (x_0, D). \tag{2.16}$$

$$\delta(K) = \sup_{r\in(x_0,D)} K(r) \inf_{s\in(x_0,r]} \frac{(s-x_0)\exp[-F^r(s)]}{\int_{x_0}^s \exp[-F^r(u)]\mathrm{d}u}. \tag{2.17}$$

Then, we have

$$\delta(K) \geqslant \sup_{r\in(x_0,D)} K(r)\exp[-F^r(r)] = K\big(r_0\big)\exp\bigg[-1 + \int_{x_0}^{r_0}\frac{(u-x_0)K(u)}{a(u)}\mathrm{d}u\bigg], \tag{2.18}$$

where r_0 is the unique solution to the equation

$$K(r) = \left(\int_{x_0}^{r} \frac{u - x_0}{a(u)} \mathrm{d}u \right)^{-1}, \qquad r \in (x_0, D). \tag{2.19}$$

When $D < \infty$ and (2.19) has no solution in $(0, D)$, we set $r_0 = D$.

Corollary 2.7. Choose a non-decreasing function $K \in C(x_0, \infty)$ such that

$$K(r) \leqslant \inf_{x \geqslant r} \left[-\frac{(a' + b)(x)}{x - x_0} - b'(x) \right] + \sup_y b'(y)$$
$$\left(\text{resp. } K(r) \leqslant \inf_{x > r} \left[-\frac{b(x)}{x - x_0} \right] \quad (r > x_0) \right).$$

Assume that $(x - x_0)K(x)/a(x)$ is locally integrable. Then, we have

$$\mathrm{gap}_{[x_0,\infty)} \geqslant \beta_0 + \delta(K),$$

where $\beta_0 = -\sup_x b'(x)$ (resp. $\beta_0 = 0$).

The following examples illustrate the power of the above results. Here, we consider the half line $[0, \infty)$ only.

Example 2.8. Take $b(x) = -b\,(b > 0)$, $a(x) \equiv a$. By Corollary 2.3 (2) with $\delta = b/(2a)$, we get $\mathrm{gap}_{[0,\infty)} \geqslant b^2/(4a)$ which is sharp (see [6; example 1.10]). Corollary 2.4 (2) or Corollary 2.5 (2) with $\varepsilon_1 = -b/a$ and $\varepsilon_2 = 0$ as well as Corollary 2.5 (5) give us the same bound.

Example 2.9. Take $a(x) \equiv 1$ and $b(x) = -\alpha x^\beta$, $(\alpha > 0, \beta > -1)$. Applying Corollary 2.6 to $f(x) = \exp\left[\varepsilon x^{\beta+1}\right]$ $(\varepsilon < \alpha/(\beta + 1))$, it follows that $\mathrm{gap}_{[0,\infty)} > 0$ whenever $\beta \geqslant 0$. To get some explicit bounds, we apply Corollary 2.7 which is available iff $\beta \geqslant 1$. The linear case $(\beta = 1)$ will be treated in the next example. We now assume that $\beta > 1$. Then, the lower bounds provided by Corollary 2.7 and (2.18) for the two choices of K are

$$2^{\frac{\beta-1}{\beta+1}} [\alpha(\beta + 1)]^{\frac{2}{\beta+1}} \exp\left[-1 + \frac{2}{\beta + 1} \right] \quad \text{and} \quad 2^{\frac{\beta-1}{\beta+1}} \alpha^{\frac{2}{\beta+1}} \exp\left[-1 + \frac{2}{\beta + 1} \right]$$

respectively. Clearly, the first bound is bigger than the second one. However, if we consider $a(x) = (1 + x^2)^2$ and $b(x) = -\alpha x^3$, then the alternative choice of K works for all $\alpha > 0$ but the first choice of K works only for $\alpha > 1$. Therefore, the two choices of K in Corollary 2.7 are not comparable.

Example 2.10. Take $b(x) = -\alpha x\,(\alpha > 0)$, $a(x) \equiv 1$. By Corollary 2.3 (1) with $c_1 = 0$ and $\delta = 1$ (or Corollary 2.4 (2) with $\varepsilon_1 = 0$ and $\varepsilon_2 = -\alpha$, or Corollary 2.7), we get $\mathrm{gap}_{[0,\infty)} \geqslant 2\alpha$. This estimate is sharp since $g(x) = x^2/2 - 1/(2\alpha)$ is an eigenfunction and so Remark 2.2 (2) is suitable. The same bound can be obtained by using (2.4) or (2.5) with $f(x) = x^2 - 1/\alpha$.

Example 2.11. Take $b(x) = -\alpha x - \beta$. Then

$$V(x) = -\bigg(\log a(x) + \int_0^x (\alpha r + \beta) a(r)^{-1} \mathrm{d}r \bigg).$$

If (2.1) holds, by Corollary 2.3 (1) with $\delta = 0$, it follows that $\mathrm{gap}_{[0,\infty)} \geqslant \alpha^+$. This provides us a non-trivial lower bound for a large number of concrete examples since $a(x)$ is quite arbitrary. If we take $a(x) \equiv 1$, then, Corollary 2.4 (3) gives us $\mathrm{gap}_{[0,\infty)} \geqslant 3\alpha$ provided $\beta \geqslant \sqrt{\alpha} > 0$. Moreover, in the case that $\alpha = \beta = 1$, the estimate is indeed sharp by Remark 2.2 (2). This is quite interesting since the change of β from 0 to 1 leads to not only the change of the spectral gap from 2 to 3 but also the change of the eigenfunction from quadratic to cubic. We now consider the particular case that $a(x) = (1+x)^2$ and $\beta = \alpha$. Then $V(x) = -(2+\alpha)\log(1+x)$ and (2.1) holds iff $\alpha > -1$. By Corollary 2.4 (1) or applying (2.3) to the function $f(x) = (1+x)^{(\alpha-1)/2}$ or applying (2.4) to $f(x) = (1+x)^{(\alpha+1)/2}$, we obtain $\mathrm{gap}_{[0,\infty)} \geqslant (\alpha+1)^2/4 \geqslant \alpha^+$. The last equality holds iff $\alpha = 1$. Note that when $\alpha > 1$, even though $g(x) := x + 1$ is in $L^2(\pi)$ and satisfies $ag'' + bg' = -\alpha g$, but this g is still not the eigenfunction of λ_1 since $g'(0) \neq 0$. For general $\alpha > -1$, the function $g(x) := (1+x)^{(\alpha+1)/2}$ satisfies $ag'' + bg' = -(\alpha+1)^2 g/4$ but g is not the eigenfunction of λ_1 since $g \notin L^2(\pi)$. Thus, Remark 2.2 (2) is not suitable for this example. However, applying (2.13) to $g(x) = (1+x)^{(\alpha+1)/2}$, we obtain

$$\mathrm{gap}_{[0,\infty)} \leqslant \varliminf_{n\to\infty} \frac{\int_0^n ag'^2 \mathrm{e}^V \mathrm{d}x}{\int_0^n g^2 \mathrm{e}^V \mathrm{d}x} \leqslant \varliminf_{n\to\infty} \frac{a(n) g'(n)^2 \mathrm{e}^{V(n)}}{g(n)^2 \mathrm{e}^{V(n)}} = \frac{(1+\alpha)^2}{4}.$$

We have thus achieved the exact bound. This example shows that in order to attain the sharp estimate, we do have some freedom of the choice of test functions rather than using the eigenfunction only.

Example 2.12. Take $b(x) \equiv 0$ and $a(x) = (1+x)^\alpha$. Obviously, (2.1) holds iff $\alpha > 1$. By Corollary 2.4 (1) with $c_1 = 1$ and $\varepsilon = \alpha$ or by Corollary 2.5 (1) with $\gamma = \alpha$ and $c_1 = 1$, we get $\mathrm{gap}_{[0,\infty)} \geqslant (\alpha-1)^2/4$ for all $\alpha \geqslant 2$. This is similar to the last example. Next, applying (2.13) to $g(x) = (1+x)^{(\alpha-1)/2}$, we obtain $\mathrm{gap}_{[0,\infty)} = 0$ for all $\alpha \in (1,2)$, which is the same as the lower bound given by Theorem 2.1. Therefore, $\mathrm{gap}_{[0,\infty)} > 0$ iff $\alpha \geqslant 2$ and our estimate is sharp for all $\alpha \leqslant 2$. However, the lower bound $(\alpha-1)^2/4$ is not sharp when $\alpha > 2$. To see this, applying (2.4) to the family $\{f(x) = (1+x)^\varepsilon - (\alpha-1)/(\alpha-1-\varepsilon) : \varepsilon > 0\}$, we get

$$\begin{aligned} \mathrm{gap}_{[0,\infty)} &\geqslant \sup_{\varepsilon\in(0,\alpha-2)} \bigg[(\alpha-1-\varepsilon)(\alpha-2+\varepsilon) \bigg(\frac{\alpha-2+\varepsilon}{\alpha-2} \bigg)^{(\alpha-2)/\varepsilon} \bigg] \\ &\geqslant e(\alpha-1)(\alpha-2). \end{aligned} \tag{2.20}$$

Setting $\varepsilon = 1/2$ and then letting $\alpha \downarrow 2$, the first estimate of (2.20) gives us $\mathrm{gap}_{[0,\infty)} \geqslant 1/4$, which is sharp. We will show in the next section (Example 3.6)

that the principle term $e\alpha^2$ of the lower bound is also exact as $\alpha \to \infty$. Applying (2.2) to the family $\{f(x) = (1+x)^\varepsilon : \varepsilon > 0\}$, we obtain $\mathrm{gap}_{[0,\infty)} \geqslant (\alpha-1)^2/4$. As for (2.5), we get $\mathrm{gap}_{[0,\infty)} \geqslant 1/4$ (independent of α). Replacing f with $f - \pi(f)$, $\pi(f) = (\alpha-1)/(\alpha-1-\varepsilon)$, it follows from (2.5) that

$$\begin{aligned}\mathrm{gap}_{[0,\infty)} &\geqslant \sup_{\varepsilon\in(0,1\wedge(\alpha-2))}\left[(1-\varepsilon)(\alpha-2+\varepsilon)\left(\frac{\pi(f)(\alpha-2+\varepsilon)}{\alpha-2}\right)^{(\alpha-2)/\varepsilon}\right]\\ &\geqslant e^{2-1/(\alpha-1)}(\alpha-2).\end{aligned}$$

All these estimates are exact at $\alpha = 2$. From these, we see that (2.5) is weaker than (2.4) but it is not comparable with (2.2) for the specific functions.

Example 2.13. Take $a(x) = (1+x)^3$ and $b(x) = (1+x)^2$. By Corollary 2.4 (1) or Corollary 2.5 (1), we have $\mathrm{gap}_{[0,\infty)} \geqslant 1/4$. On the other hand, applying (2.4) to $f(x) = \log(1+x) - 1$, we get $\mathrm{gap}_{[0,\infty)} \geqslant \inf_{x>0} \frac{1+x}{\log(1+x)} = e$.

3. The Case of Full Line

Set $C(x) = \int_0^x b(u)a(u)^{-1}\mathrm{d}u$. Then "$Z < \infty$" becomes

$$\int_{-\infty}^{\infty} \frac{\mathrm{e}^{C(x)}}{a(x)}\mathrm{d}x < \infty. \tag{3.1}$$

The process is non-explosive iff

$$\min\left\{\int_0^\infty \mathrm{d}x\mathrm{e}^{-C(x)}\int_0^x \frac{\mathrm{e}^{C(y)}}{a(y)}\mathrm{d}y,\ \int_{-\infty}^0 \mathrm{d}x\mathrm{e}^{-C(x)}\int_x^0 \frac{\mathrm{e}^{C(y)}}{a(y)}\mathrm{d}y\right\} = \infty. \tag{3.2}$$

Intuitively, the idea in this section is to divide the full line into two half lines. However, there are some technical problems. Note that the spectral gap for the full line can not be bigger than the maximum of the ones for the half lines. Thus, the test function f must be connected in some way around the reference point x_0. For instance, in order for the approximating function g of the eigenfunction to be in $C^2(\mathbf{R})$, we require that $f \in C^1(\mathbf{R})$ in the first term below and $f \in C(\mathbf{R})$ with $f(x_0) = 0$ in the second term below. Actually, what we have in mind is taking the reference point x_0 to be the place at which the eigenfunction vanishes, even though the precise place is usually unknown in advance.

As a variation of $I(f)$, define

$$I^-(f)(x) = \frac{\mathrm{e}^{-C(x)}}{f'(x)}\int_x^{-\infty} \frac{f(u)\mathrm{e}^{C(u)}}{a(u)}\mathrm{d}u, \qquad x < x_0.$$

Theorem 3.1. Assume that (3.1) and (3.2) hold. Let $x_0 \in \mathbf{R}$.

(1) For every function $f \in C^2(\mathbf{R})$ with $f(x) > 0$ for all x, we have

$$\mathrm{gap}(L) \geqslant \inf_x \big[(-af' - bf)'/f\big](x) \tag{3.3}$$

$$= \inf_x \big[-b' - [af'' + (a'+b)f']/f\big](x). \tag{3.4}$$

(2) Let $C(x)=\int_{x_0}^{x} a^{-1}b$. For every function $f\in C(\mathbf{R})\cap C^1(\mathbf{R}\setminus\{x_0\})\cap L^1(\pi)$ with $f(x_0)=0$, $f'(x)>0$ for all $x\neq x_0$, we have

$$\operatorname{gap}(L)\geqslant(\delta_1\vee\delta_2)^{-1}, \tag{3.5}$$

where

$$\delta_1=\sup_{x>x_0} I(f)(x),\qquad \delta_2=\sup_{x<x_0} I^-(f)(x). \tag{3.6}$$

In particular, if moreover $f\in C^2(\mathbf{R})$, then

$$\operatorname{gap}(L)\geqslant\inf_{x\neq x_0}\big[-af''-bf'\big](x)/f(x). \tag{3.7}$$

Applying (3.4) to the functions $f(x)=c_1+|x-x_0|^{2m}$ ($c_1>0$, $m\in\mathbb{N}$) and $f(x)=e^{\delta(x-x_0)}$ ($\delta\in\mathbf{R}$), we obtain the following result.

Corollary 3.2.

(1) If there exists $\varepsilon\leqslant -3$ such that $(a'+b)(x-x_0)\leqslant\varepsilon a$, then $\operatorname{gap}(L)\geqslant\inf_{x\neq x_0}\left[\dfrac{a(x)(1-\varepsilon)^2}{4(x-x_0)^2}-b'(x)\right]$.

(2) If there exists $\varepsilon\leqslant 0$ such that $a'+b\leqslant\varepsilon a$, then $\operatorname{gap}(L)\geqslant\inf\limits_x\big[\frac{\varepsilon^2}{4}a(x)-b'(x)\big]$.

Corollary 3.3.

(1) Suppose that $\inf_{x\neq x_0} a(x)/(x-x_0)^2:=c>0$. If there exists $\varepsilon<1$ such that $(x-x_0)b(x)\leqslant\varepsilon a(x)$ for large $|x|$ (resp. for all $x\neq x_0$), then $\operatorname{gap}(L)>0$ $\big($resp. $\operatorname{gap}(L)\geqslant\frac{c}{4}(1-\varepsilon)^2\big)\big)$.

(2) If there exist some ε_1 and ε_2, either $\varepsilon_2=0$ and $\varepsilon_1<0$ or $\varepsilon_2<0$, such that $\operatorname{sgn}(x-x_0)b(x)\leqslant(\varepsilon_1+\varepsilon_2|x-x_0|)a(x)$ for large enough $|x|$, then $\operatorname{gap}(L)>0$. Furthermore if the condition holds for all $x\neq x_0$, then

$$\operatorname{gap}(L)\geqslant\frac{1}{4}\left[\int_0^\infty e^{\varepsilon_1 u+\varepsilon_2 u^2/2}\mathrm{d}u\right]^{-2}\inf_x a(x).$$

In particular, if $\varepsilon_2=0$, then $\operatorname{gap}(L)\geqslant\frac{1}{4}\varepsilon_1^2\inf_x a(x)$.

(3) If

$$c_1:=\sup_{x>x_0}e^{-C(x)}\int_x^\infty\frac{e^{C(u)}}{a(u)}\mathrm{d}u,\qquad c_1^-:=\sup_{x<x_0}e^{-C(x)}\int_{-\infty}^x\frac{e^{C(u)}}{a(u)}\mathrm{d}u$$
$$c_2:=\sup_{x>x_0}e^{-C(x)}\int_x^\infty e^{C(u)}\mathrm{d}u,\qquad c_2^-:=\sup_{x<x_0}e^{-C(x)}\int_{-\infty}^x e^{C(u)}\mathrm{d}u$$

are all finite, then $\operatorname{gap}_{[x_0,\infty)}\geqslant 1/\max\{4c_1c_2,\,4c_1^-c_2^-\}$.

(4) If $c:=\max\left\{\sup\limits_{x>x_0}\dfrac{a(x)}{e^{C(x)}}\displaystyle\int_x^\infty\frac{e^{C(u)}}{a(u)}\mathrm{d}u,\ \sup\limits_{x<x_0}\dfrac{a(x)}{e^{C(x)}}\displaystyle\int_{-\infty}^x\frac{e^{C(u)}}{a(u)}\mathrm{d}u\right\}<\infty$, then $\operatorname{gap}(L)\geqslant\inf\limits_x\dfrac{a(x)}{4c^2}$. In particular, if $\lim_{|x|\to\infty}e^{C(x)}/a(x)=0$ and $\overline{\lim}_{|x|\to\infty}a(x)/[a'(x)-b(x)]<\infty$, then $\operatorname{gap}(L)>0$.

(5) If $b \equiv 0$, then

$$\mathrm{gap}(L) \geqslant \frac{1}{4}\left[\max\left\{\sup_{x>x_0}(x-x_0)\int_x^\infty a(u)^{-1}\mathrm{d}u,\ \sup_{x<x_0}(x_0-x)\int_{-\infty}^x a(u)^{-1}\mathrm{d}u\right\}\right]^{-1}.$$

Parts (3) and (4) of the corollary improve respectively the first two parts of [6; Theorem 1.3] which were proved by using an analytic approach rather than the coupling one. Moreover, the present proofs become very simple.

By adding a new term, $c\tan^{-1}(x)$ or $cx/\sqrt{1+x^2}$ for instance, to the original function f, from Theorem 3.1 (2), we obtain the following result.

Corollary 3.4. Suppose that there exists a function $f \in C(\mathbf{R})\cap C^1(\mathbf{R}\setminus\{x_0\})$ with $f(x_0)=0$, $f'(x)>0$ for all large enough $|x|$ and

$$\max\left\{\overline{\lim_{x\to+\infty}} I(f)(x),\quad \overline{\lim_{x\to-\infty}} I^-(f)(x)\right\} < \infty.$$

Then, we have $\mathrm{gap}(L) > 0$.

Corollary 3.5. [1] Choose $K \in C(\mathbf{R}\setminus\{x_0\})$ such that $K(x)$ is non-decreasing as $|x-x_0|$ increases, moreover, $K(r) \leqslant \inf_{x\geqslant r} b(x)/(x_0-x)$ for all $r \geqslant x_0$ and $K(r) \leqslant \inf_{x\leqslant r} b(x)/(x_0-x)$ for all $r \leqslant x_0$. Assume that $(x-x_0)K(x)/a(x)$ is locally integrable. Define $F^r(s)$ as in (2.16) for $x_0 \leqslant s \leqslant r$ or $r \leqslant s \leqslant x_0$ (in the later case, replacing $s\wedge r$ with $s\vee r$) and then define $\delta(K)$ as in (2.17) with $D=\infty$. Next, define $\delta^-(K)$ in the same way but replacing "$r > x_0$" and "$s \in (x_0, r]$" with "$r < x_0$" and "$s \in [r, x_0)$" respectively. Then, we have $\mathrm{gap}(L) \geqslant \delta(K)\wedge\delta^-(K)$.

We are now ready to mention a nice result due to Kac and Krein [11] and Kotani [12] by using a different approach: Let $b \equiv 0$. Then

$$\frac{1}{4}\delta^{-1} \leqslant \mathrm{gap}_{[0,\infty)} \leqslant \delta^{-1}, \qquad \frac{1}{4}(\delta\vee\delta^-)^{-1} \leqslant \mathrm{gap}(L) \leqslant (\delta\vee\delta^-)^{-1}, \tag{3.8}$$

where $\delta = \sup_{x\geqslant 0} x\int_x^\infty a(u)^{-1}\mathrm{d}u$ and $\delta^- = \sup_{x\leqslant 0} x\int_x^{-\infty} a(u)^{-1}\mathrm{d}u$. Clearly, the lower bounds coincide with Corollary 2.5 (5) and Corollary 3.3 (5) respectively. To illustrate the power of (3.8), it suffices to look at an example with the half-line.

Example 3.6. Consider the Example 2.12 again. Then, by (3.8), we have $\delta^{-1} = 1$ if $\alpha = 2$ and

$$\delta^{-1} = \frac{(\alpha-1)^\alpha}{(\alpha-2)^{\alpha-2}} = (\alpha-1)^2\left(1+\frac{1}{\alpha-2}\right)^{\alpha-2} \sim e\alpha^2 \qquad \text{if } \alpha > 2.$$

Combining this with the lower bound given in Example 2.12, we see that the upper bound here has the correct order as $\alpha\to\infty$ and the lower bound is exact when $\alpha = 2$.

The examples given below not only illustrate the use of the our results but also show some difference between the half line and the full line.

[1] See also the first author's paper "Spectral gap and logarithmic Sobolev constant for continuous spin systems", Theorem 4.1.

Example 3.7. Take $b(x) = -\alpha x - \beta$. If (3.1) and (3.2) hold, then as in Example 2.11, we have $\mathrm{gap}(L) \geqslant \alpha^+$, independent of β. When $\alpha > 0$ and $\beta = 0$, we indeed have $\mathrm{gap}(L) = \alpha$ for every $a(x)$ having the properties: symmetric with respect to the origin, satisfying (3.1) and (3.2) and $\int x^2 \mathrm{d}\pi < \infty$, since then $g(x) = x$ is an eigenfunction of $\lambda_1 = \alpha$. Especially, when $a(x) \equiv 1$, we have $\mathrm{gap}(L) = \alpha$ but not 2α given in Example 2.10.

Example 3.8. Consider the special case of the above example, $b(x) = -\alpha x$ and $a(x) = 1 + x^2$. Then, $C(x) = -\frac{\alpha}{2}\log(1+x^2)$ and (3.1) holds iff $\alpha > -1$. We have just seen that $\mathrm{gap}(L) = \alpha$ for all $\alpha > 1$. This is different from Example 2.12. Next, applying Theorem 3.1 (2) to the test function $f(x) = x(1+x^2)^\varepsilon$, $\varepsilon = (\alpha-1)/4$, we obtain

$$\begin{aligned}\delta_1 = \delta_2 &= \sup_{x>0} \frac{(1+x^2)^{\alpha/2}}{(1+x^2)^\varepsilon + 2\varepsilon x^2(1+x^2)^{\varepsilon-1}} \int_x^\infty \frac{u(1+u^2)^{-\alpha/2+\varepsilon}}{1+u^2}\mathrm{d}u \\ &= \sup_{x>0} \frac{1+x^2}{[1+(1+2\varepsilon)x^2](\alpha-2\varepsilon)} \leqslant \frac{1}{(1+2\varepsilon)(\alpha-2\varepsilon)} = \frac{4}{(1+\alpha)^2}.\end{aligned}$$

finally, applying (2.12), we get

$$\begin{aligned}\mathrm{gap}(L) &\leqslant \lim_{n\to\infty} \frac{\int_{-n}^n a(x) f'(x)^2 \mathrm{e}^{V(x)}\mathrm{d}x}{\int_{-n}^n f(x)^2 \mathrm{e}^{V(x)}\mathrm{d}x} \\ &= \lim_{n\to\infty} \frac{\int_{-n}^n (1+x^2)^{-\alpha/2}[(1+x^2)^\varepsilon + 2\varepsilon x^2(1+x^2)^{-1+\varepsilon}]^2}{\int_{-n}^n x^2(1+x^2)^{-1-\alpha/2+2\varepsilon}\mathrm{d}x} \\ &= \frac{(1+\alpha)^2}{4}.\end{aligned}$$

Therefore, $\mathrm{gap}(L) = (\alpha+1)^2/4$ for all $\alpha \in (-1, 1]$.

Example 3.9. Take $b(x) = -\alpha x^3$ $(\alpha > 0)$ and $a(x) = (1+x^2)^2$. Applying (3.7) to $f(x) = x(1+x^2)^{-1/2}$, we obtain $\mathrm{gap}(L) \geqslant 3$ which is independent of α. On the other hand, by Corollary 3.5 with $x_0 = 0$, $K(r) = \alpha r^2$ and $r_0^2 = \big(\sqrt{2\alpha+1}+1\big)/\alpha$, we obtain $K(r_0) = \sqrt{2\alpha+1}+1$ and

$$\mathrm{gap}(L) \geqslant K(r_0)\big(1+K(r_0)/\alpha\big)^{\alpha/2} \exp\left[-1 - \frac{\alpha K(r_0)}{2\big(\alpha + K(r_0)\big)}\right] \gtrapprox \sqrt{2\alpha+1}\, e^{-1/2}. \tag{3.9}$$

Especially, when $\alpha = 4$, then the first bound equals $16e^{-2} \approx 2.1654$.

Example 3.10. Take $a(x) \equiv 1$ and $b(x) = -x + \cos x$. This is clearly a perturbation of the ordinary O.U.-process. However, when we apply (3.4) to $f(x) \equiv 1$, which gives the exact eigenvalue of the O.U.-process, we get the trivial bound. We now adopt a comparison technique (see also Proposition 4.5). Note that

$$C(u) - C(x) = -u^2/2 + x^2/2 + \sin u - \sin x \leqslant -u^2/2 + x^2/2 + 2.$$

Inserting this into (3.6) with $f(x) = x$, it follows that $\mathrm{gap}(L) \geqslant e^{-2}$. The estimate can be further improved by noticing

$$C(u) - C(x) \leqslant -u^2/2 + x^2/2 + \varepsilon \sin u - \varepsilon \sin x + 2(1-\varepsilon)$$

and using $f(x) = x+\varepsilon\cos x$ instead of $f(x) = x$. Then we obtain $\mathrm{gap}(L) \geqslant (2e)^{-1}$ by setting $\varepsilon = 1/2$.

To conclude this section, we mention some examples for which the eigenfunction $g \in C^2(\mathbf{R}) \cap L^2(\pi)$ but non-linear.

Examples 3.11. Let $a(x) \equiv 1$. Then we $\mathrm{gap}(L) = 1$ for the following choices of $b(x)$.

(1) $g(x) = x(c + x^2)$, $c > 0$. $b(x) = -\dfrac{x}{3}\left[1 + \dfrac{2(9+c)}{3x^2+c}\right]$.

(2) $g(x) = \int_0^x e^{cy^{2n}}\mathrm{d}y$, $n \in \mathbf{Z}_+$, $c \leqslant 0$. $b(x) = -2ncx^{2n-1} - e^{-cx^{2n}}\int_0^x e^{cy^{2n}}\mathrm{d}y$.

(3) $g(x) = cx + \sin x$, $c > 1$. $b(x) = -\dfrac{cx}{c+\cos x}$.

To prove the assertion, simply use (3.4) with $f = g'$ and note that both g and b are odd functions.

4. The General Case

In contrast the cases of the half line or full line, the structure of the eigenfunction of λ_1 in the higher dimensional case is too complex to be understood and it is often not monotone with respect to the ordinary semi-order. Here, a diffusion semigroup P_t is said to be monotone if $P_tf(x) \leqslant P_tf(y)$ holds for all $x \leqslant y$ and all monotone (non-decreasing) continuous functions f. Even in the case that the eigenfunction is monotone, one still requires the process to be monotone which is a quite strong restrictive condition especially for the higher dimensional diffusions (refer to [8] for details). Thus, in general, it is not practical to use the eigenfunction or its approximation as the distance we required and so we should adopt a different strategy. Roughly speaking, our goal is as follows. First, we use the coupling method on some simple distances in $\mathbf{R}^d$ and reduce our problem to the case of the half line. Then, applying the idea given in Section 2 to construct a new distance $f \circ d$ for some suitable function f. Fortunately, in this way, we still obtain good enough estimates for the spectral gap.

Let $\widetilde{L}$ be a coupling operator of L, $d(x, y)$ be a distance which is in C^2 away from the set $\{(x, x) : x \in \mathbf{R}^d\}$ and let $D = \sup_{x,y} d(x, y)$. Then there exist two functions A and B on $\mathbf{R}^d \times \mathbf{R}^d$ such that for each $f \in C^2[0, D)$ (refer to [4]),

$$\widetilde{L}f \circ d(x, y) = A(x, y)f''(d(x, y)) + B(x, y)f'(d(x, y)), \quad x \neq y. \tag{4.1}$$

Note that $\widetilde{L}$ is a degenerated elliptic operator on $\mathbf{R}^d \times \mathbf{R}^d$, we have $A(x, y) \geqslant 0$ for all x and y. One key step of the coupling approach is to find a function $f \in C^2[0, D)$ with $f(0) = 0$, $f' > 0$ and $f'' \leqslant 0$ on $[0, D)$ such that

$$\widetilde{L}f \circ d(x, y) \leqslant -\delta f \circ d(x, y), \qquad x \neq y \tag{4.2}$$

for some constant $\delta > 0$. Next, choose functions α, $\beta \in C(0, D)$ such that

$$\alpha(r) \leqslant \inf_{d(x,y)=r} A(x, y), \qquad \beta(r) \geqslant \sup_{d(x,y)=r} B(x, y). \tag{4.3}$$

Then, for (4.2), it suffices that

$$\alpha(r)f''(r)+\beta(r)f'(r)\leqslant -\delta f(r),\qquad r\in(0,D). \tag{4.4}$$

We have thus reduced (4.2) to (4.4). Denote by λ^* the largest constant δ in (4.2) as f varies. Clearly, λ^* dominates the largest δ in (4.4). The next result is parallel to Theorem 2.1.

Theorem 4.1.

(1) For every function $f\in C^2[0,D)$ with $f(0)=0$, $f'>0$ and $f''\leqslant 0$ on $[0,D)$, we have

$$\lambda^*\geqslant \inf_{r\in(0,D)}\big[(-\alpha f''-\beta f')/f\big](r). \tag{4.5}$$

(2) Define $C(r)=\int_0^r \alpha^{-1}\beta$ and then define $I(f)$ as in (1.2) but replacing $a(u)$ and $[0,\infty)$ with $\alpha(u)$ and $[0,D)$ respectively. For every function $f\in C[0,D)\cap L^1(\pi)$ with $\pi(f)\geqslant 0$ on $(0,D)$ and satisfying

$$f(r)\geqslant -\beta(r)e^{-C(r)}\int_r^D \frac{f(u)e^{C(u)}}{\alpha(u)}\mathrm{d}u, \tag{4.6}$$

we have

$$\lambda^*\geqslant \inf_{r\in(0,D)}\left\{f(r)^{-1}\int_0^r \mathrm{d}s e^{-C(s)}\int_s^D \frac{f(u)e^{C(u)}}{\alpha(u)}\mathrm{d}u\right\}^{-1}, \tag{4.7}$$

In particular, if moreover $f\in C^1(0,D)$, $f(0)=0$ and $f'>0$ on $(0,D)$, then

$$\lambda^*\geqslant \inf_{r\in(0,D)} I(f)(r)^{-1}, \tag{4.8}$$

Theorem 4.1 is also meaningful for diffusion processes on a manifold which will be treated in a separate paper. Next, if there exists a coupling such that $\inf_{x\neq y}A(x,y)>0$ and $\lambda^*>0$, then the L-diffusion process is ergodic. Part (1) of Theorem 4.1 is rather simple but it has the following useful consequence, which is an analog of the alternative choice of Corollary 2.7.

Corollary 4.2. Choose a non-decreasing function $K\in C(0,D)$ such that $K(r)\leqslant \inf_{s\in[r,D)}[-\beta(s)/s]$, $r\in(0,D)$. Assume that $rK(r)/\alpha(r)$ is locally integrable on $(0,D)$. Define $F^r(s)$ as in (2.16) and then define $\delta(K)$ in (2.17) with $x_0=0$. Then, we have $\lambda^*\geqslant\delta(K)$.

Remark 4.3. (1). The condition (4.6) is used for the non-positive property of the second derivative of the function required by (4.2) or (4.4). However, when $A(x,y)$ in (4.1) is indeed a function of $d(x,y)$ only and $\alpha(r)$ is taken to be the common value of $A(x,y)$ when $d(x,y)=r$, we do not need (4.6). In this case, the resulting function $f\circ d$ may not be a distance but this does not interferes our proof.

(2). When $\beta(r) \geqslant 0$ on $(0, D)$, the condition (4.6) is trivial. In the case of $\beta(r) < 0$ on $(0, D)$ and $\lim_{r\to D} f(r)e^{C(r)}/\beta(r) = 0$, by the integration by parts formula, (4.6) can be rewritten as follows.

$$\int_r^D \left(\frac{f}{\beta}\right)'(u)e^{C(u)}\mathrm{d}u \geqslant 0. \tag{4.9}$$

More simply,

$$f'\beta - \beta' f \geqslant 0 \qquad \text{on} \quad (0, D) \tag{4.10}$$

is enough for (4.9).

By virtue of (4.8), part (2) of Theorem 2.1 and its Corollaries 2.6 and 2.7 are available with a slight modification. We omit the details here to save space. The reason why we use λ^* here rather than $\mathrm{gap}(L)$ is the following. Our approach requires that the eigenfunction be Lipschitz with respect to the distance we adopted. In the compact case, this is not a problem. But for the non-compact case, this may not be true. To overcome this difficulty, we adopt a localizing procedure[6], which then yields some technical problem. So, in general, we are still unable to claim that λ^* is indeed a lower bound of $\mathrm{gap}(L)$. However, the conclusion holds for one-dimensional case.

Corollary 4.4. When $d = 1$, Theorem 4.1 and Corollary 4.2 hold if λ^* is replaced by $\mathrm{gap}(L)$.

Before moving further, we mention a simple comparison result which is a direct consequence of (1.1) (refer to [6] and [16]).

Proposition 4.5. (1) Let $\overline{L} \sim (\bar{a}, V)$, if $a(x) - \bar{a}(x) \geqslant 0$ for all x, then

$$\mathrm{gap}(L) \geqslant \mathrm{gap}(\overline{L}). \tag{4.11}$$

(2) Let $\overline{L} \sim (a, \overline{V})$, we have

$$\mathrm{gap}(L) \geqslant \mathrm{gap}(\overline{L}) \exp[-\delta(V - \overline{V})], \tag{4.12}$$

where $\delta(f) = \sup f - \inf f$.

Let us also mention a sufficient condition for the regularity of Dirichlet forms. In general, [9; Theorem 1.6.3] says that the semi-group is recurrent if there exists $\{u_n\} \subset C_0^\infty(\mathbf{R}^d)$ such that $u_n \to 1$ and $\lim_{n\to\infty} \int \langle a\nabla u_n, \nabla u_n\rangle \mathrm{d}\pi = 0$. From this we conclude that the Dirichlet form is regular if there exists $r_n \uparrow \infty$ such that

$$\lim_{n\to\infty} \int_{r_n \leqslant |x| \leqslant r_n+1} \mathrm{tr}\, a(x)\mathrm{d}\pi = 0. \tag{4.13}$$

Actually, choose $h \in C^\infty(\mathbf{R})$ such that $\|h'\|_\infty \leqslant 2$ and $h(r) = 1$ for $r \leqslant 0, h(r) = 0$ for $r \geqslant 1$. Take $u_n(x) = h(|x| - r_n)$, then $u_n \to 1$ and (4.13) implies $\int \langle a\nabla u_n, \nabla u_n\rangle \mathrm{d}\pi \to 0$.

To study the spectral gap of diffusions in $\mathbf{R}^d$, we consider three concrete distances: the Euclidean distance, the L_1-distance and the Riemannian distance induced by a positive definite diagonal matrix which is dominated by $a(x)$. To state the result, we need some notations. Choose positive functions $a_i \in C^2(\mathbf{R}^d)(i \leqslant d)$ such that $a - \mathrm{diag}\{a_1, a_2, \cdots, a_d\} \geqslant 0$ (non-negative definite) and $\inf_{i,x} a_i(x) > 0$. Let $\bar{b}_i = a_i\partial_i V + \partial_i a_i (i \leqslant d)$. Next, set $\alpha_2 = 4$ and choose $\alpha_1, \alpha_3 \in C(\mathbf{R}_+)$ such that

$$0 < \alpha_1(r) \leqslant \inf_{|x-y|=r} \Big\{ \min_i \Big(\sqrt{a_i(x)} - \sqrt{a_i(y)} \Big)^2 + 4 \min_i \sqrt{a_i(x)a_i(y)} \Big\},$$

$$0 < \alpha_3(r) \leqslant \inf_{|x-y|_1=r} \Big\{ \sum_{i=1}^d \Big(\sqrt{a_i(x)} - \sqrt{a_i(y)} \Big)^2 + 4 \min_i \sqrt{a_i(x)a_i(y)} \Big\},$$

where $|\cdot|$ is the ordinary Euclidean norm and $|x-y|_1 = \sum_{i=1}^d |x_i - y_i|$. Next, choose β_j (j=1,2,3) as follows.

(1) Put $\sigma = \sqrt{\mathrm{diag}\{a_1, a_2, \cdots, a_d\}}$ and choose $\beta_1 \in C(0,\infty)$ so that

$$\beta_1(r) \geqslant \sup_{|x-y|=r} |x-y|^{-1} \Big\{ \|\sigma(x) - \sigma(y)\|^2 - |x-y|^{-2} |(\sigma(x) - \sigma(y))(x-y)|^2 + \langle \bar{b}(x) - \bar{b}(y), x-y \rangle \Big\}.$$

(2) If $a_i(x)$ depends on x_i only for all i. Set

$$\rho(x,y) = \Big[\sum_{i=1}^d \Big(\int_{x_i}^{y_i} a_i(r)^{-1/2} \mathrm{d}r \Big)^2 \Big]^{1/2}, \qquad D = \sup_{x,y} \rho(x,y)$$

and $h_i = \sqrt{a_i}\partial_i V + \partial_i \sqrt{a_i}, i \leqslant d$. Choose $\beta_2 \in C(0, D)$ so that

$$\beta_2(r) \geqslant \sup_{\rho(x,y)=r} \rho(x,y)^{-1} \sum_{i=1}^d [h_i(y) - h_i(x)] \int_{x_i}^{y_i} a_i(r)^{-1/2} \mathrm{d}r, \qquad r \in (0, D).$$

(3) If $a_i(x)$ depends on x_i only and $\bar{b}_i(x)$ is non-decreasing in x_k for $k \neq i$. Choose $\beta_3 \in C(0,\infty)$ so that

$$\beta_3(r) \geqslant \sup_{x \geqslant y, |x-y|_1=r} \sum_{i=1}^d [\bar{b}_i(x) - \bar{b}_i(y)], \qquad r > 0.$$

Finally choose non-decreasing functions $K_1, K_3 \in C(0,\infty)$ and $K_2 \in C(0, D)$ so that $K_j(r) \leqslant \inf_{s \geqslant r}[-\beta_j(s)/s]$.

Theorem 4.6. Theorem 4.1 and Corollary 4.2 are valid if the functions α, β, K and λ^* are replaced by α_j, β_j, K_j and $\mathrm{gap}(L)$ respectively for each $j = 1, 2, 3$.

Let $a(x) = \alpha(x)\sigma^2$ for some positive $\alpha \in C^2(\mathbf{R}^d)$ and positive definite matrix σ. To use Theorem 4.6, by Proposition 4.5, one may compare $a(x)$ with a diagonal matrix directly. But, as was pointed out in [3], [6], the result should

be better if we use directly the distance $|\sigma^{-1}(x-y)|$ instead of the Euclidean one. To this end, take the coordinate transformation $y=\sigma^{-1}x$. Then $\partial/\partial x_i = \sum_{k=1}^d(\sigma^{-1})_{ik}\partial/\partial y_k, i\leqslant d$, and the operator $L\sim(a,b)$ becomes

$$L^{(y)}=\alpha(\sigma y)\sum_{k=1}^d\frac{\partial^2}{\partial y_k^2}+\sum_{k=1}^d\bigg(\sum_{i=1}^d b_i(\sigma y)(\sigma^{-1})_{ik}\bigg)\frac{\partial}{\partial y_k}$$

which is in the desired form of Theorem 4.6.

The following result simplifies the form of K_j's given above. It can be considered as an extension of [7; Theorem 1.3] to multidimensional diffusion processes in the context of spectral gap.

Corollary 4.7. Let a_i, $\bar b_i$ and α_j be the same as in Theorem 1.1. Suppose that $a_i(x)$ depends only on x_i for all i. Set $\kappa=\max_i\|\nabla\sqrt{a_i}\|_\infty^2$. Fix a point $p\in\mathbf{R}^d$ and let $\lambda_{\min}(A)$ be the smallest real part of eigenvalues of matrix A. According to the three cases in Theorem 4.6, we define θ_j $(j=1,2,3)$ as follows.

(1) $\theta_1(r)=\inf_{|x-p|\geqslant r}\lambda_{\min}\big(-\partial_j\bar b_i(x)\big),\quad r\geqslant 0.$

(2) $\theta_2(r)=\inf_{\rho(x,p)\geqslant r}\lambda_{\min}\big(-X_iX_j\overline{V}(x)\big)$, where $X_i=\sqrt{a_i(x_i)}\,\partial_i$ and $\overline{V}=V+\log\sqrt{a_1\cdots a_d}$.

(3) If $\bar b_i(x)$ is non-decreasing in x_k for $i\neq k$, let

$$\theta_3(r)=\inf_{|x-p|_1\geqslant r}\bigg[-\max_j\sum_i\partial_j\bar b_i(x)\bigg].$$

Next, define

$$\begin{aligned}\gamma_j(r)&=\frac{1}{r}\int_0^r\theta_j(u)\mathrm{d}u\quad(j=1,2,3),\\ K_1(r)&=\gamma_1(r/2)-(1-d^{-1})\kappa,\\ K_3(r))&=\gamma_3(r/2),\quad r>0,\\ K_2(r)&=\gamma_2(r/2),\quad r\in(0,D).\end{aligned}$$

Then, Corollary 4.2 holds for these K_j's with the same replacements made in Theorem 4.6. In particular, if $\max\{\theta_1(\infty)-\kappa(1-d^{-1}),\ \theta_2(D),\ \theta_3(\infty)\}>0$, then we have $\mathrm{gap}(L)>0$. Here, $\theta_i(\infty)$ $(i=1,3)$ and $\theta_2(D)$ are understood as the limits as $r\to\infty$ and $r\to D$ respectively.

Obviously, when $d=1$, the case of $j=1$ coincides with the case of $j=3$ for both Theorem 4.6 and Corollary 4.7. As for $d>1$, the first may be better than the latter. For example, this is the case for $d=2$, $a=I$ and $V(x)=-\frac{1}{2}x_1^2+x_1x_2-x_2^2$. Conversely, the latter may be better if $\|\nabla a_i\|_\infty$ is large for some $i\leqslant d$. From these and Example 4.8 below, we conclude that the cases of $j=1,2,3$ are not comparable each other.

Example 4.8. Consider Example 3.9 again. We have $V(x)=\alpha/2-\alpha/[2(1+x^2)]-(2+\alpha/2)\log(1+x^2)$ and (4.13) holds. For $r>0$, we have $\inf_x(b(x+r)-$

$b(x)) = -\alpha r^3/4$. Take $K_1(r) = \alpha r^2/4$ and $\alpha_1(r) = 4$, then $F_1^r(r) = \alpha r^4/64$. By Theorem 4.6 with $j = 1$ or 3, we obtain $\mathrm{gap}(L) \geqslant \sqrt{2\alpha}\exp[-1/2]$. This is weaker but close to (3.9).

Next, we have $\overline{V}(x) = -\alpha/[2(1+x^2)] - (1+\alpha/2)\log(1+x^2)$. Let $X(x) = (1+x^2)\frac{\mathrm{d}}{\mathrm{d}x}$, then

$$-X^2\overline{V}(x) = (2+\alpha)(1+x^2) - \frac{2\alpha}{1+x^2} + \alpha \geqslant 2, \qquad x \in \mathbf{R}.$$

By Corollary 4.7 (2) we obtain $\mathrm{gap}(L) \geqslant 2$ which is independent of α.

Example 4.9. The lower bound $\mathrm{gap}(L) \geqslant \alpha^+$ given in Example 3.7 can be also obtained by using Corollary 4.7 with $j = 1$ or 3.

Example 4.10. We now return to Example 3.10. Let $h(r) = -\sup_x[b(x+r) - b(x)]$, $r \geqslant 0$. It is easy to check that $h(2\pi + r) - 2\pi - r = h(r) - r$ and $h(r) = r - 2\sin\frac{r}{2}$ for $r \in [0, 2\pi]$. Since $r^{-1}h(r)$ is increasing, it can be taken as $K_1(r)$. By Theorem 4.6 we have

$$\mathrm{gap}(L) \geqslant \sup_{r\in[0,\pi]} \left(1 - r^{-1}\sin r\right)\exp[\cos r + (r\sin r)/2 - 1].$$

By setting $r = 1.95$, we obtain $\mathrm{gap}(L) \geqslant 0.329$.

Example 4.11. When $a = I$ we take $\alpha = 4$ and then the estimate provided by (4.7) with $f(r) \equiv 1$ coincides with the one obtained by using the first moment of the coupling time (refer to [6]). Note that the test function $f(x) \equiv 1$ can not be allowed for (4.8) since $f' \equiv 0$. But (4.8) does often produce better estimates. Especially, take $d = 1$ and $b(x) = -4x^3$. Choose $\beta(r) = -r^3$. By (4.7) we get the lower bound $[\Gamma(5/4) + 1/8)]^{-1} \approx 0.9695$ which is the same as in [6; Example 1.9]. Nevertheless, applying (4.8) to the test function $f(x) = \log(1+x)$ and noticing Remark 4.3 (1), a numerical computation shows that $\mathrm{gap}(L) \geqslant 2.4395$.

Example 4.12. Take $a = I$ and $b_i(x) = \sum_j b_{ij}x_j$, where (b_{ij}) is symmetric with $b_{ij} \geqslant 0$ for $i \neq j$ and $\sum_i b_{ij} = -1$ for all j. Next, take $g(x) = \sum_i x_i$. We have

$$V(x) = \frac{1}{2}\sum_{i,j=1}^{d} b_{ij}x_ix_j \leqslant \frac{1}{2}\sum_{i=1}^{d}\left(b_{ii}x_i^2 + \frac{1}{2}\sum_{j\neq i} b_{ij}(x_i^2 + x_j^2)\right) = -\frac{1}{2}|x|^2.$$

Thus (4.13) holds and $g \in L^2(\pi)$. Moreover, it is easy to check that $\pi(g) = 0$. Hence $\mathrm{gap}(L) = 1$ which is just the lower bound provided by Corollary 4.7 with $j = 3$.

To conclude this section, we study the Poincaré inequality with respect to the absolute distribution of the process generated by L. This provides a new way to estimate $\mathrm{gap}(L)$ and may be useful in the study of the spectral gap on path space. The idea used here comes from [10] and [17] in which the logarithmic Sobolev inequalities on path space were studied for diffusions over a Riemannian manifold.

Theorem 4.13. Suppose that there exists $\bar{a} > 0$ such that $\langle a(x)u, u\rangle \leqslant \bar{a}|u|^2$ for all $x, u \in \mathbf{R}^d$. Let $a(x) = \sigma(x)\sigma(x)^*$ and set

$$K = \sup_{x\neq y} |x-y|^{-2}\big[\|\sigma(x)-\sigma(y)\|^2 + \langle b(x)-b(y), x-y\rangle\big].$$

We have

$$P_t f^2(x) \leqslant K^{-1}\bar{a}(\exp[2Kt]-1)P_t|\nabla f|^2(x) + (P_t f(x))^2 \tag{4.14}$$

for all $x \in \mathbf{R}^d$, $t \geqslant 0$ and $f \in C^1(\mathbf{R}^d)$ with $P_t f^2(x) < \infty$. When $K = 0$, the coefficient on the right-hand side is understood as the limit as $K \to 0$..

We mention that (4.14) can be sharp. For example, take $L = \Delta$, then $2t$ is the smallest constant so that (4.14) holds. The bounded assumption of a is unnatural, due to the limitation of the present proof, but we do not know how to remove it.

Remark 4.14. The process considered in Theorem 4.13 is not necessarily reversible. Next, the L-diffusion process is ergodic if $K < 0$. Then, by letting $t \to \infty$ in (4.14), we obtain

$$\mathrm{gap}(L) \geqslant -K\bar{a}^{-1}\inf_x \lambda_{\min}(a(x)). \tag{4.15}$$

5. Preparations for the Proofs

Lemma 5.1. Suppose that $D_n \uparrow \mathbf{R}^d$ is a sequence of normal domains and let $\mathrm{gap}(D_n)$ denote the first Neumann eigenvalue of L on D_n, then we have $\mathrm{gap}(L) \geqslant \overline{\lim}_{n\to\infty}\mathrm{gap}(D_n)$. When $d = 1$, we indeed have $\mathrm{gap}(D_n) \downarrow \mathrm{gap}(L)$.

Proof. a) Note that (refer to [1] and [15])

$$\mathrm{gap}(D_n) = \inf\big\{\pi_n(\langle a\nabla f, \nabla f\rangle) : f \in C^1(D_n), \pi_n(f) = 0, \pi_n(f^2) = 1\big\}, \tag{5.1}$$

where $\pi_n(f) = \pi(I_{D_n} f)$. For any $\varepsilon > 0$, choose $f \in C^1(\mathbf{R}^d)$ such that $\pi(f) = 0, \pi(f^2) = 1$ and $\pi(\langle a\nabla f, \nabla f\rangle) \leqslant \mathrm{gap}(L) + \varepsilon$. Then, there exists $n_0 \geqslant 1$ such that

$$\int_{D_n}\left(f - \int_{D_n} f\mathrm{d}\pi\right)^2 \mathrm{d}\pi \geqslant 1-\varepsilon, \qquad n \geqslant n_0.$$

Hence

$$\mathrm{gap}(D_n) \leqslant \frac{\int_{D_n}\langle a\nabla f, \nabla f\rangle \mathrm{d}\pi}{\int_{D_n}(f - \int_{D_n} f\mathrm{d}\pi)^2\mathrm{d}\pi} \leqslant \frac{\mathrm{gap}(L)+\varepsilon}{1-\varepsilon}.$$

b) Next, when $d = 1$, we need only to prove that $\mathrm{gap}_{[p_1,q_1]} \geqslant \mathrm{gap}_{[p_2,q_2]} \geqslant \mathrm{gap}(L)$ for $[p_1, q_1] \subset [p_2, q_2]$. Let u be an eigenfunction with respect to $\mathrm{gap}_{[p_1,q_1]}$, then $u'(p_1) = u'(q_1) = 0$. We extend u to $\mathbf{R}$ by setting $u(r) = u(p_1)$ for $r \leqslant p_1$ and $u(r) = u(q_1)$ for $r \geqslant q_1$. Then $u \in C^1(\mathbf{R})$, by (5.1) we obtain $\mathrm{gap}_{[p_2,q_2]} \leqslant \mathrm{gap}_{[p_1,q_1]}$.

c) If in addition (4.13) holds, there exist non-negative functions $u_n \in C_0^\infty(\mathbf{R})$ such that $u_n \uparrow 1$ and $\int a\, u_n'^2\,\mathrm{d}\pi \to 0$ as $n \to \infty$. By (1.1) together with an approximation argument, we have

$$\mathrm{gap}(L) \leqslant \lim_{n\to\infty}\frac{\int a\big[(uu_n)'\big]^2\mathrm{d}\pi}{\int (uu_n)^2\mathrm{d}\pi - \big(\int uu_n\mathrm{d}\pi\big)^2} \leqslant \mathrm{gap}_{[p_1,q_1]}.$$

d) To avoiding the use of the sufficient condition (4.13), take $V_\varepsilon = V - \varepsilon a$, $Z_\varepsilon = \int \mathrm{e}^{V_\varepsilon}\mathrm{d}x$, $\varepsilon > 0$. Then $V_\varepsilon \uparrow V$ and $Z_\varepsilon \uparrow Z$ as $\varepsilon \downarrow 0$. Let $L_\varepsilon \sim (a, V_\varepsilon)$, then L_ε satisfies (4.13). By (1.1) we have

$$\mathrm{gap}(L) \leqslant \varliminf_{\varepsilon\to 0} \mathrm{gap}(L_\varepsilon) \leqslant \mathrm{gap}_{[p,q]}. \quad \square$$

Suppose that $a(x) = \sigma(x)\sigma(x)^*$ for all x. Let $\widetilde{L} \sim (\tilde{a}, \tilde{b})$ be the operator of the coupling by reflection[4]:

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

where $c(x,y) = \sigma(x)(I - 2\bar{u}\bar{u}^*)\sigma(y)^*$ and $\bar{u} = (x-y)/|x-y|$.

Lemma 5.2. Let $S = \prod_{i=1}^d [p_i, q_i]$ and (x_t, y_t) be the coupling by reflection of the reflecting L-diffusion process on S. If $\sigma_{ij} = \delta_{ij}\sigma_{ii}(x_i)$ and $b_i(x)$ is non-decreasing in x_k for $k \neq i$, then the coupling preserves the ordinary semi-order: $x_0 \geqslant y_0$ implies $P^{x_0,y_0}(x_t \geqslant y_t, t \geqslant 0) = 1$.

Proof. One may compare the conditions of the lemma with the criteria given in [8]. Let $T = \inf\{t \geqslant 0 : x_t = y_t\}$ be the coupling time, then we need only to prove the order-preservation up to time T.

a) For $n \geqslant 1$, choose $C_n \in C(\mathbf{R})$ with $\mathrm{supp}C_n \subset (0, n^{-1})$, $0 \leqslant C_n \leqslant 2n$ and $\int C_n(u)\mathrm{d}u = 1$. Define $\phi_n(r) = \int_0^r \mathrm{d}s \int_0^s C_n(u)\mathrm{d}u$, then $0 \leqslant \phi_n'(r) \leqslant 1$, $0 \leqslant \phi_n''(r) \leqslant 2/(nr^2)$ and $\phi_n(r) \uparrow r^+$. Next, note that $b_i(x)$ is non-decreasing in x_k for $k \neq i$ and $\phi_n'(y_i - x_i) = 0$ for $y_i \leqslant x_i$, we have

$$\begin{aligned}\widetilde{L}\phi_n(y_i - x_i) =& \phi_n'(y_i - x_i)(b_i(y) - b_i(x)) + \phi_n''(y_i - x_i)\Big[\sum_{j=1}^d (\sigma_{jj}(y_j) - \sigma_{jj}(x_j))^2 \\ &+ \frac{4}{|x-y|^2}\sum_{j=1}^d (y_j - x_j)^2 \sigma_{jj}(x_j)\sigma_{jj}(y_j)\Big] \\ \leqslant& N\sum_{j=1}^d (y_i - x_i)^+ + \frac{N}{n\varepsilon^2} \end{aligned} \tag{5.2}$$

for some constant N and all $x, y \in S$ with $|x - y| \geqslant \varepsilon > 0$. Let $L_{i+}^{(1)}$, $L_{i-}^{(1)}$ be the local times of x_t on $\{x_i = q_i\}, \{x_i = p_i\}$ respectively, and let $L_{i+}^{(2)}$, $L_{i-}^{(2)}$ be those of y_t. Note that $\phi_n'(y_i - x_i) = 0$ for $y_i \leqslant x_i$ and

$$I_{[q_i-\varepsilon, q_i]}(x_i) \leqslant I_{[q_i-\varepsilon, q_i]}(y_i), \qquad I_{[p_i, p_i+\varepsilon]}(y_i) \leqslant I_{[p_i, p_i+\varepsilon]}(x_i)$$

for $q_i \geqslant y_i \geqslant x_i \geqslant p_i$. We have

$$\begin{aligned}&\int_{t_1}^{t_2} \phi_n'(y_i(s)) - x_i(s))\mathrm{d}\big(L_{i+}^{(1)}(s) + L_{i-}^{(2)}(s) - L_{i+}^{(2)}(s) - L_{i-}^{(1)}(s)\big) \\ &= \lim_{\varepsilon\to 0} \int_{t_1}^{t_2} \phi_n'\big(y_i(s) - x_i(s)\big)\Big(I_{[q_i-\varepsilon,q_i]}\big(x_i(s)\big) - I_{[q_i-\varepsilon,q_i]}\big(y_i(s)\big) \\ &\qquad + I_{[p_i,p_i+\varepsilon]}\big(y_i(s)\big) - I_{[p_i,p_i+\varepsilon]}\big(x_i(s)\big)\Big)\mathrm{d}s \leqslant 0, \qquad t_1 \leqslant t_2. \end{aligned} \tag{5.3}$$

b) Let $T_\varepsilon = \inf\{t \geqslant 0 : |x_t - y_t| \leqslant \varepsilon\}$, by (5.2) and (5.3) we have

$$\begin{aligned}
&E^{x_0,y_0} \sum_{i=1}^d \Big[\phi_n\big(y_i(t_2 \wedge T_\varepsilon) - x_i(t_2 \wedge T_\varepsilon)\big) - \phi_n\big(y_i(t_1 \wedge T_\varepsilon) - x_i(t_1 \wedge T_\varepsilon)\big)\Big] \\
&\leqslant E^{x_0,y_0} \int_{t_1 \wedge T_\varepsilon}^{t_2 \wedge T_\varepsilon} \sum_{i=1}^d \widetilde{L}\phi_n\big(y_i(t) - x_i(t)\big)\mathrm{d}t \\
&\leqslant (t_2 - t_1)\frac{dN}{n\varepsilon^2} + N\int_{t_1}^{t_2} E^{x_0,y_0} \sum_{i=1}^d \big(y_i(t \wedge T_\varepsilon) - x_i(t \wedge T_\varepsilon)\big)^+ \mathrm{d}t, \qquad t_1 \leqslant t_2.
\end{aligned}$$

By letting $n \to \infty$, we obtain

$$\frac{d}{dt}E^{x_0,y_0} \sum_{i=1}^d \big(y_i(t \wedge T_\varepsilon) - x_i(t \wedge T_\varepsilon)\big)^+ \leqslant NE^{x_0,y_0} \sum_{i=1}^d \big(y_i(t \wedge T_\varepsilon) - x_i(t \wedge T_\varepsilon)\big)^+.$$

Therefore

$$E^{x_0,y_0} \sum_{i=1}^d \big(y_i(t \wedge T_\varepsilon) - x_i(t \wedge T_\varepsilon)\big)^+ = 0$$

which implies $P^{x_0,y_0}\big(y_i(t \wedge T_\varepsilon) \leqslant x_i(t \wedge T_\varepsilon)\big) = 1, \quad t \geqslant 0$. Since $T_\varepsilon \uparrow T$ as $\varepsilon \downarrow 0$, the lemma follows by letting $\varepsilon \to 0$. □

The following result summarizes our approach to estimate $\mathrm{gap}(L)$ by using coupling.

Theorem 5.3. Let $D_n \uparrow \mathbf{R}^d$ be a sequence of normal domains with inward normal vector fields V_n of ∂D_n under the Riemannian metric $(g(\partial_i, \partial_j)) = a^{-1}$. Next, let $d(x, y) : \mathbf{R}^d \times \mathbf{R}^d \to [0, \infty)$ be in C^2 out of $\{(x, x) : x \in \mathbf{R}^d\}$ and having the properties: $d(x, y) = 0$ iff $x = y$, for each n and $x \in D_n$, $V_n d(x, \cdot)(y)|_{\partial D_n} \leqslant 0$ and there exists $c_n > 0$ such that $d(x, y) \geqslant c_n|x - y|$ for $x, y \in D_n$. If there exists a coupling operator $\widetilde{L}$ of L such that $\widetilde{L}d(x, y) \leqslant -\delta d(x, y)$ for some $\delta > 0$ and all $x \neq y$, then $\mathrm{gap}(L) \geqslant \delta$.

Proof. Fix $n \geqslant 1$, let (x_t, y_t) be the reflecting $\widetilde{L}$-diffusion process on $D_n \times D_n$ under the Riemannian metric $\big(g(\partial_i, \partial_j)\big) = a^{-1}$. Let L_t be the local time of the process on $\partial(D_n \times D_n)$, then

$$\begin{aligned}
\mathrm{d}\, d(x_t, y_t) &= \mathrm{d}M_t + \widetilde{L}d(x_t, y_t)\mathrm{d}t + \big(V_n d(\cdot, y_t)(x_t) + V_n(x_t, \cdot)(y_t)\big)\mathrm{d}L_t \\
&\leqslant \mathrm{d}M_t - \delta d(x_t, y_t)\mathrm{d}t
\end{aligned}$$

up to the coupling time T for some martingale M_t. Here, we take $V_n f(x) = 0$ for $x \notin \partial(D_n)$. By [15; Lemma 2.4 and 3; Theorem 6.2] we obtain $\mathrm{gap}(D_n) \geqslant \delta$. Then Theorem 5.3 follows from Lemma 5.1. □

6. Proofs of Theorem 2.1 and its Corollaries

Proof of Theorem 2.1. a) Let $f \in C^1[x_0,\infty)\cap C^2(x_0,\infty)$ with $f>0$ on (x_0,∞) and define $g(x)=\int_{x_0}^x f$. Then g is strictly increasing and so $d(x,y):=|g(x)-g(y)|$ is a distance in $[x_0,\infty)$. Because the process is monotone (see [8]), we simply use the classical coupling: $\widetilde{L}h(x,y)=(Lh(\cdot,y))(x)+(Lh(x,\cdot))(y)$ for all $h\in C^2([x_0,\infty)^2)$ and $x\neq y$. Then

$$\begin{aligned}\widetilde{L}d(x,y)&=\widetilde{L}d(x,\cdot)(y)-\widetilde{L}d(\cdot,y)(x)=Lg(y)-Lg(x)\\&\leqslant -d(x,y)\inf_{y>x}\big[\big(-(ag''+bg')(y)+(ag''+bg')(x)\big)/\big(g(y)-g(x)\big)\big]\\&\leqslant -d(x,y)\inf_{z>x_0}\big[(-af'-bf)'/f\big](z)\qquad x\leqslant y.\end{aligned}$$

Here in the last step, we have used the Mean Value Theorem. Part (1) of Theorem 2.1 then follows from Theorem 5.3.

b) For part (2) of Theorem 2.1, the proof is similar but applying the coupling to the function

$$g(x)=\int_{x_0}^x f'(y)I(f)(y)\mathrm{d}y.$$

To prove (2.5), note that

$$\big(f'e^C\big)'=(af''+bf')e^C/a. \tag{6.1}$$

By assumption, we have $-\big(f'e^C\big)'\geqslant cfe^C/a$. Therefore, $I(f)(x)\leqslant c^{-1}$ and so the conclusion follows from (2.4). It remains to check that $g'>0$ on (x_0,∞). But this holds iff $\pi(f)\geqslant 0$ due to the fact that $f'>0$ on (x_0,∞). □

Proof of Corollary 2.4. The assertions follow from that of Corollary 2.3 correspondingly with the specific parameters given below.

(1) $\delta=(1-\varepsilon)/2$.
(2) $c_1=0$, $c_2=-\varepsilon_2$ and $\delta=-\varepsilon_1/2$.
(3) $c_1=(\varepsilon_2+\varepsilon_1^2)/\varepsilon_2^2$, $c_2=2\varepsilon_1/\varepsilon_2$. □

To prove Corollary 2.5, we need a simple result which is an extension of [6; Lemma 3.1].

Lemma 6.1. Let $m\in C([x_0,\infty);\mathbf{R}_+)$ and $n\in C([x_0,\infty);(0,\infty))$.

(1) If $\int_x^\infty m(y)/n(y)\mathrm{d}y\leqslant c_1\,m(x)$ and $\int_x^\infty m(y)\mathrm{d}y\leqslant c_2\,m(x)$ for all $x\geqslant x_0$. Then for every $\gamma\in[0,1/c_2)$, we have

$$\int_x^\infty e^{\gamma(y-x_0)}\frac{m(y)}{n(y)}\mathrm{d}y\leqslant\frac{c_1}{1-\gamma c_2}e^{\gamma(x-x_0)}m(x),\qquad x\geqslant x_0.$$

(2) If $(x-x_0)\int_x^\infty m(y)dy\leqslant c$ for all $x\geqslant x_0$, then for every $\gamma\in[0,1)$, we have

$$\int_x^\infty (y-x_0)^\gamma m(y)dy\leqslant\frac{c}{1-\gamma}(x-x_0)^{\gamma-1},\qquad x\geqslant x_0.$$

Proof. Here, we prove part (1) only since the proof of part (2) is simpler. Without loss of generality, assume that $m(x)$ has finite support. Set

$$M(x) = \int_x^\infty \frac{m(y)}{n(y)}\mathrm{d}y.$$

Then

$$\begin{aligned}\int_x^\infty e^{\gamma(y-x_0)}\frac{m(y)}{n(y)}\mathrm{d}y &= -\int_x^\infty e^{\gamma(y-x_0)}\mathrm{d}M(y)\\ &\leqslant c_1 e^{\gamma(x-x_0)}m(x) + c_1\gamma\int_x^\infty e^{\gamma(y-x_0)}m(y)\mathrm{d}y.\end{aligned} \tag{6.2}$$

Consider the special case that $n(x)\equiv 1$ and $c_1 = c_2$. Then, (6.2) gives us

$$\int_x^\infty e^{\gamma(y-x_0)}m(y)\mathrm{d}y \leqslant \frac{c_2}{1-\gamma c_2}e^{\gamma(x-x_0)}m(x), \qquad x\geqslant x_0.$$

Inserting this into (6.2), we obtain the required assertion. □

Proof of Corollary 2.5. a) Note that

$$C(u)-C(x) \leqslant \varepsilon\int_x^u \frac{\mathrm{d}y}{c_1+y-x_0} = \varepsilon\log\frac{c_1+u-x_0}{c_1+x-x_0}, \qquad u>x.$$

We have for $f(x) = (c_1+x-x_0)^\delta$,

$$\begin{aligned}I(f)(x) &\leqslant \frac{1}{c\delta(c_1+x-x_0)^{\delta+\varepsilon-1}}\int_x^\infty \frac{(c_1+u-x_0)^{\delta+\varepsilon}}{(c_1+u-x_0)^\gamma}\mathrm{d}u\\ &= \frac{-1}{c\delta(\delta+\varepsilon-\gamma+1)}\cdot\frac{1}{(c_1+x-x_0)^{\gamma-2}}\\ &\leqslant \frac{-1}{c\delta(\delta+\varepsilon-\gamma+1)}\cdot\frac{1}{c_1^{\gamma-2}} \quad (\gamma\geqslant 2).\end{aligned}$$

Setting $\delta = (\gamma-\varepsilon-1)/2$, we prove part (1) of the corollary. Obviously, when $\gamma = 2$, c_1 is allowed to be zero.

b) For part (2), by assumption, we have

$$\begin{aligned}C(u)-C(x) &\leqslant \varepsilon_1(u-x)+\varepsilon_2(u^2-x^2)/2-\varepsilon_2(u-x)x_0\\ &= [\varepsilon_1+\varepsilon_2(x-x_0)](u-x)+\varepsilon_2(u-x)^2/2.\end{aligned} \tag{6.3}$$

Without loss of generality, assume that $\inf_x a(x) = 1$. Consider the test function $f(x) = (c_1-\varepsilon_2(x-x_0))e^{\delta(x-x_0)}$, $\delta>0$. We obtain

$$\begin{aligned}I(f) &\leqslant \frac{1}{-\varepsilon_2+c_1\delta-\varepsilon_2\delta(x-x_0)}\int_x^\infty [c_1-\varepsilon_2(u-x_0)]e^{[\varepsilon_1+\delta+\varepsilon_2(x-x_0)](u-x)+\varepsilon_2(u-x)^2/2}\\ &= \frac{1}{-\varepsilon_2+c_1\delta-\varepsilon_2\delta(x-x_0)}\int_0^\infty [c_1-\varepsilon_2 u-\varepsilon_2(x-x_0)]e^{[\varepsilon_1+\delta+\varepsilon_2(x-x_0)]u+\varepsilon_2u^2/2}\mathrm{d}u\\ &= \frac{1}{-\varepsilon_2+c_1\delta-\varepsilon_2\delta(x-x_0)}\left[(c_1+\varepsilon_1+\delta)\int_0^\infty e^{[\varepsilon_1+\delta+\varepsilon_2(x-x_0)]u+\varepsilon_2u^2/2}\mathrm{d}u+1\right].\end{aligned}$$

If $\varepsilon_1 < 0$, by setting $c_1 = \delta = -\varepsilon_1/2$, we get

$$I(f) \leqslant \frac{1}{-\varepsilon_2 + \varepsilon_1^2/4}.$$

Next, assume that $\varepsilon_1 \geqslant 0$ and set $c_1 = 0$. Since $\varepsilon_2 < 0$, by (6.3), we have

$$\begin{aligned} I(f) &\leqslant \frac{1}{-\varepsilon_2 - \varepsilon_2\delta(x - x_0)} \left[(\varepsilon_1 + \delta) \int_0^\infty e^{[\varepsilon_1+\delta]u + \varepsilon_2 u^2/2} \mathrm{d}u + 1 \right] \\ &\leqslant \frac{1}{-\varepsilon_2} \left[(\varepsilon_1 + \delta) \int_0^\infty e^{[\varepsilon_1+\delta]u + \varepsilon_2 u^2/2} \mathrm{d}u + 1 \right] \\ &= \frac{1}{-\varepsilon_2} \left[1 + \int_0^\infty e^{u + \varepsilon_2 u^2/[2(\varepsilon_1+\delta)^2]} \mathrm{d}u \right]. \end{aligned}$$

Then by letting $\delta \downarrow 0$, we obtain the estimate in the middle of the expression.

The last estimate in the expression simply follows from (6.3) and Lemma 6.1 (1) with the choice $m(x) = e^{C(x)}$, $n(x) \equiv 1$, $f(x) = e^{\gamma(x-x_0)}$ and

$$\gamma = \frac{1}{2} \left[\int_0^\infty e^{\varepsilon_1 u + \varepsilon_2 u^2/2} \mathrm{d}u \right]^{-1}.$$

c) To prove part (3), simply apply Lemma 6.1 (1) to $m(x) = e^{C(x)}$, $n(x) = a(x)$, $f(x) = e^{\gamma(x-x_0)}$ and $\gamma = \frac{1}{2c_2}$.

d) Part (4) also follows from Lemma 6.1 (1) but with $m(x) = e^{C(x)}/a(x)$, $n(x) \equiv 1$, $f(x) = e^{\gamma(x-x_0)}$ and $\gamma = \frac{1}{2c}$. The particular assertion is then deduced by using the Mean Value Theorem.

e) Finally, part (5) follows from Lemma 6.1 (2) by setting $m(x) = 1/a(x)$, $f(x) = \sqrt{x - x_0}$ and $\gamma = 1/2$. □

Proof of Corollary 2.6. Let $x_0 = 0$ for simplicity. By assumption, there exists $N > 0$ so that

$$f'(x) > 0 \quad \text{for all } x \geqslant N \text{ and } \sup_{x \geqslant N} I(f)(x) < \infty.$$

Since $f_1'(x) = c/(1+x)^2 + f'(x)$, we have $f_1'(x) \geqslant f'(x) > 0$ for all $x \geqslant N$. As for $x \leqslant N$, choose c small enough so that $f_1'(x) \geqslant c/(1+N)^2 + \min_{x \leqslant N} f'(x) > 0$. We now fix c. Because f is an increasing function, there exits $M > 0$ such that $f_1(x) \leqslant c + f(x) \leqslant Mf(x)$ for all $x \geqslant N$. Thus, for $x \geqslant N$, we have

$$0 < I(f_1)(x) \leqslant MI(f)(x) < \infty.$$

Finally, for $x \leqslant N$, we have

$$0 < I(f_1)(x) = \frac{e^{-C(x)}}{f_1'(x)} \int_x^N \frac{f_1(u)e^{C(u)}}{a(u)} \mathrm{d}u + \frac{f_1'(N)}{f_1'(x)} e^{C(N)-C(x)} I(f_1)(N).$$

The right-hand side is bounded in $[x_0, N]$ and so the required conclusion follows from (2.4). □

Proof of Corollary 2.7. For a proof of (2.18), refer to [7; Proof a) of Theorem 1.3].

a) Consider the case that $K(r) \leqslant \inf_{x\geqslant r}[-(a'+b)(x)/(x-x_0) - b'(x)] + \sup_y b'(y)$. Fixed $r_1 \in (x_0,\infty)$ so that $K(r_1) > 0$. Otherwise, we have nothing to do. Define

$$f(x) = \int_{x_0}^x \mathrm{d}y \exp\bigg[-\int_{x_0}^y \frac{u-x_0}{a(u)}[K(r_1)-K(u)]I_{\{u\leqslant r_1\}}\mathrm{d}u\bigg], \qquad x > x_0.$$

Since $f'' \leqslant 0$, f' is decreasing and so $f(x) \geqslant (x-x_0)f'(x)$. By (2.3), we have

$$\begin{aligned} -b'(x) - \frac{af''+(a'+b)f'}{f}(x) &= \beta_0 + \frac{-af''-(a'+b)f'+(-b'-\beta_0)f}{f}(x) \\ &\geqslant \beta_0 + \frac{-af''-[a'+b+(x-x_0)(b'+\beta_0)]f'}{f}(x) \\ &\geqslant \beta_0 + K(r_1)(x-x_0)f'(x)/f(x). \end{aligned}$$

Here in the last step, we have used the properties of f just mentioned above. Noticing that $(x-x_0)/f(x)$ is non-decreasing, we have $(x-x_0)f'(x)/f(x) = (x-x_0)f'(r_1)/f(x) \geqslant (r_1-x_0)f'(r_1)/f(r_1)$ for all $x \geqslant r_1$. This completes the proof of the main case.

b) The proof of the alternative case is similar, but use (2.5) instead of (2.3). □

To conclude this section, we discuss when the equality in (1.3) holds. Suppose that we have a C^2-eigenfunction f of $\lambda_1 > 0$. That is, $-af'' - bf' = \lambda_1 f$ with $f'(x_0) = 0$. Then, as we will prove later, f has the following properties: i) $f \in L^1(\pi)$, ii) $f' > 0$ (or < 0) on (x_0,∞) and iii) $\lim_{x\to\infty} f'(x)e^{C(x)} = -\lambda_1\pi(f)Z$. Now, by (6.1) we have $-\big(f'e^C\big)' = \lambda_1 f e^C/a$. Thus

$$f'(x)e^{C(x)} = \lambda_1\int_x^\infty fe^C/a - \lambda_1\pi(f)Z \geqslant \lambda_1\int_x^\infty [f-\pi(f)]e^C/a$$

since $\pi(f) \leqslant 0$ by ii) and iii). Set $\tilde f = f - \pi(f)$. We have $\pi(\tilde f) = 0$ and $\tilde f'(x)e^{C(x)} \geqslant \lambda_1\int_x^\infty \tilde f e^C/a$. Hence $I(\tilde f)(x)^{-1} \geqslant \lambda_1$ for all $x > x_0$. Combining this with part (2) of Theorem 2.1, we conclude that the equality of (1.3) holds.

The remainder of this section is to prove i) – iii) listed above. Where the second one is essential, from which the first one and then the last one follows immediately from the next lemma.

Lemma 6.2. If $Lf = -\lambda f$ on $[p,q] \subset [0,\infty)$ for some $\lambda \neq 0$, then we have

$$-\lambda\int_p^q f\mathrm{d}\pi = \big[\big(f'e^C\big)(q) - \big(f'e^C\big)(p)\big]/Z.$$

Proof. Simply use (6.1). □

Lemma 6.3. Let $Lf = -\lambda f$ for some $f \in C^2[x_0, \infty)$ and for some $\lambda \geqslant 0$. If there exist $\alpha < \beta$ such that $f \equiv 0$ on $[\alpha, \beta]$, then $f \equiv 0$.

Proof. The assertion is indeed a consequence of the maximum principle (pointed out to the authors by Z. D. Huan). A simple probabilistic proof goes as follows. If $f \not\equiv 0$, without loss of generality, assume that $\gamma := \inf\{x \geqslant \beta : f(x) \neq 0\} < \infty$ and there exists $x_n \downarrow \gamma$ so that $f(x_n) > 0$ (one may replace f with $-f$ if necessary). For each $n \geqslant 1$, choose $y_n \in [\gamma, x_n)$ such that $f(y_n) = \min\{f(x) : x \in [\gamma, x_n]\}$. Then $f(y_n) \leqslant 0$ since $f(\gamma) = 0$. Let x_t be the L-diffusion process starting from y_n and set $\tau_n = \inf\{t \geqslant 0 : x_t \in \{x_n, \gamma - n^{-1}\}\}$. Then

$$\mathbb{E}f(x_{t\wedge\tau_n}) = f(y_n) - \lambda\mathbb{E}\int_0^{t\wedge\tau_n} f(x_s)\mathrm{d}s \leqslant f(y_n)[1 - \lambda\mathbb{E}\tau_n].$$

This implies that $\mathbb{E}f(x_{\tau_n}) \leqslant 0$ for large enough n. But for any n with $\gamma - n^{-1} \geqslant \alpha$, we have $\mathbb{E}f(x_{\tau_n}) \geqslant f(x_n)\mathbb{P}[x_{\tau_n} = x_n] > 0$. The contradiction implies the assertion. □

Proposition 6.4. Suppose that $\lambda_1 > 0$ and $Lf = -\lambda_1 f$ for some $f \in C^2[x_0, \infty)$, $f \neq$constant and $f'(x_0) = 0$. Then $f' \neq 0$ on (x_0, ∞) and furthermore $f \in L^1(\pi)$.

Proof. Suppose that there is a $p > x_0$ such that $f'(p) = 0$.

a) We claim that $f \neq$constant on $[x_0, p]$. Otherwise, we have $f = -\lambda_1^{-1}Lf = 0$ on $[x_0, p]$ which implies that $f \equiv 0$ by Lemma 6.3. We now prove that $f(p) \neq 0$. To do so, set $g = fI_{[x_0,p]} + f(p)I_{(p,\infty)}$. If $f(p) = 0$, then $g \in C^2$, $Lg = -\lambda_1 g$ and $g \equiv 0$ on $[p, \infty)$. By Lemma 6.3, we have $g \equiv 0$ and in particular $f \equiv 0$ on $[x_0, p]$. This again implies $f \equiv 0$ on $[x_0, \infty)$ by Lemma 6.3.

b) By using Lemma 6.2, we have

$$\int_{x_0}^p f\mathrm{d}\pi = 0, \qquad \int_{x_0}^p af'^2\mathrm{d}\pi = -\int_{x_0}^p (fLf)\mathrm{d}\pi = \lambda_1\int_{x_0}^p f^2\mathrm{d}\pi. \tag{6.4}$$

Here in the last step, we have used the assumption $Lf = -\lambda_1 f$.

c) Without loss of generality, assume that $f(p) = 1$. Then $\pi(g) = \pi[p, \infty) < 1$. Therefore, by (6.4), we get

$$\begin{aligned}
\lambda_1 &\leqslant \frac{\pi(ag'^2)}{\pi(g^2) - \pi(g)^2} \\
&= \frac{\pi(ag'^2)}{\int_{x_0}^p f^2\mathrm{d}\pi + \pi[p, \infty) - \pi[p, \infty)^2} \\
&= \frac{\lambda_1\int_{x_0}^p f^2\mathrm{d}\pi}{\int_{x_0}^p f^2\mathrm{d}\pi + \pi[p, \infty) - \pi[p, \infty)^2} \\
&< \lambda_1.
\end{aligned}$$

This is a contradiction.

d) Having the increasing property of f in mind, the last assertion of the lemma follows from [2; Theorem 4.14]. □

7. Proofs of Theorem 3.1 and its Corollaries

Proof of Theorem 3.1. Here we prove part (2) only since the proof of part (1) is similar and even simpler. Choose $\delta > 0$ such that

$$\int_{x_0}^{\infty} \frac{f(x)\mathrm{e}^{C(x)}}{a(x)}\mathrm{d}x = \delta \int_{x_0}^{-\infty} \frac{f(x)\mathrm{e}^{C(x)}}{a(x)}\mathrm{d}x.$$

Define

$$g(x) = \begin{cases} \int_{x_0}^{x} f'(y)I(f)(y)\mathrm{d}y, & x \geqslant x_0, \\ \delta \int_{x_0}^{x} f'(y)I^{-}(f)(y)\mathrm{d}y, & x < x_0. \end{cases}$$

Note that $f(x_0) = 0$. We have $g \in C^2(\mathbf{R})$ and $g' > 0$. Next, let $d(x,y) = |g(x) - g(y)|$. Then the proof of Theorem 2.1 gives us

$$\widetilde{L}d(x,y) \leqslant \begin{cases} -\delta_1^{-1}d(x,y), & \text{if } x > y \geqslant x_0, \\ -\delta_2^{-1}d(x,y), & \text{if } x_0 \geqslant x > y. \end{cases}$$

As for $x > x_0 > y$, we have

$$\widetilde{L}d(x,y) = \big[\widetilde{L}g(x) - \widetilde{L}g(x_0)\big] + \big[\widetilde{L}g(x_0) - \widetilde{L}g(y)\big] \leqslant -\big(\delta_1 \vee \delta_2\big)^{-1}d(x,y).$$

The proof is then completed by using Theorem 5.3. □

Having Theorem 3.1 in mind, the proofs of Corollaries 3.2–3.5 are parallel to Corollaries 2.4–2.7 respectively and hence omitted.

To conclude this section, we study the same problem as in the last part of Section 6. Note that the comment before Lemma 6.2 is the same.

Proposition 7.1. Suppose that $\lambda_1 > 0$ and $Lf = -\lambda_1 f$ for some $f \in C^2(\mathbf{R}) \cap L^2(\pi)$, $f \neq$constant and $(1+|f|)f'e^{C}(x) \to 0$ as $x \to \infty$. Then $f' \neq 0$.

Proof. a) Suppose that there is a p so that $f'(p) = 0$. Then, we should have $f(p) \neq 0$. Otherwise, set $g = fI_{[p,\infty)} + f(p)I_{(-\infty,p)}$. Then, $f''(p) = -\big[(bf' + \lambda_1 f)/a\big](p) = 0$. Hence, $g \in C^2(\mathbf{R})$, $Lg = -\lambda_1 g$ and $g \equiv 0$ on $(-\infty, p]$. By Lemma 6.3, we have $g \equiv 0$ and hence $f \equiv 0$ which is impossible.

b) Without loss of generality, assume that $f(p) = 1 = Z$. Note that $f'fe^{C}(x) \to 0$ as $x \to \infty$, we have

$$\int_p^{\infty} af'^2\mathrm{d}\pi = \int_p^{\infty} f'^2 e^{C}(x)\mathrm{d}x = -\int_p^{\infty} f\big(f'e^{C}\big)'\mathrm{d}x = -\int_p^{\infty} (fLf)\mathrm{d}\pi = \lambda_1 \int_p^{\infty} f^2\mathrm{d}\pi.$$

On the other hand, it follows from Lemma 6.2 that $\int_p^{\infty} f\mathrm{d}\pi = 0$ and so $\pi(g) = \pi(-\infty, p) < 1$. Hence

$$\lambda_1 \leqslant \frac{\pi(ag'^2)}{\pi(g^2) - \pi(g)^2} = \frac{\int_p^{\infty} af'^2\mathrm{d}\pi}{\int_p^{\infty} f^2\mathrm{d}\pi + \pi[-\infty, p) - \pi[-\infty, p)^2} < \lambda_1.$$

This is a contradiction. □

8. Proofs of Theorem 4.1, Theorem 4.6 and Their Corollaries

Proof of Theorem 4.1. Part (1) follows directly from (4.1)–(4.4). The conclusion (4.7) follows by replacing the function f in (4.4) with

$$g(r) = \int_0^r e^{-C(s)} ds \int_s^D \frac{f(u)e^{C(u)}}{\alpha(u)} \mathrm{d}u.$$

Then (4.8) follows from (4.7) by using the Mean Value Theorem. □

Proof of Corollary 4.2. We remark that the corollary in the present case is deduced directly from part (1) of Theorem 4.1 by taking

$$f(r) = \int_0^r \mathrm{d}s \exp\bigg[- \int_0^s \frac{u}{\alpha(u)} [K(r_1) - K(u)] I_{\{u \leqslant r_1\}} \mathrm{d}u \bigg], \qquad r \in [0, D). \quad (8.1)$$

The details are very much the same as in the proof of Corollary 2.7. □

Proof of Corollary 4.4. Simply use Lemma 5.1. □

Proof of Theorem 4.6. By Proposition 4.5, we may assume that $a = \mathrm{diag}\{a_1, \cdots, a_d\}$ and then $\bar{b}_i = b_i,\ i \leqslant d$.

a) When $j = 1$, we may assume that $\int_0^1 s|K_1(s)|\mathrm{d}s < \infty$. Let $\widetilde{L}$ be the coupling operator with[14]

$$c(x, y) = \sqrt{a(x)} \bigg(\sqrt{a(y)} - 2 \frac{\sqrt{a(y)^{-1}}(x - y)(x - y)^*}{|\sqrt{a(y)^{-1}}(x - y)|^2} \bigg).$$

Take $d(x, y) = |x - y|$ and choose $\alpha(r) = \alpha_1(r)$ and $K(r) = K_1(r)$ (see [18]). Let $r_1 \in (0, D)$ so that $K_1(r_1) > 0$ and define $f(r)$ as in (8.1) but replacing K with K_1. It follows from Corollary 4.2 that $\lambda^* \geqslant K(r_1) \inf_{s \in (0, r_1)} f'(s)/f(s)$. Next, for $n \geqslant 1$, let $D_n = \prod_{i=1}^d [-n, n]$. Since $a = \mathrm{diag}\{a_1, \cdots, a_d\}$, the normal vector on ∂S_n coincides with that under the Riemannian metric $(g(\partial_i, \partial_j)) = a^{-1}$. Then $d(x, y) := f(|x - y|)$ satisfies the boundary condition given in Theorem 5.3. From this we claim that $\mathrm{gap}(L) \geqslant \lambda^*$ and so the assertion of the theorem in the case of $j = 1$ follows.

b) Take $d(x, y) = \rho(x, y)$ and $c(x, y) = \sqrt{a(x)}\,(I - 2uu^*)\sqrt{a(y)}$, where $u_i = \frac{1}{\rho(x,y)} \int_{x_i}^{y_i} \frac{1}{\sqrt{a_i(r)}} \mathrm{d}r$, $x \neq y$. Then the proof for the case of $j = 2$ is similar to that for $j = 1$ (refer to [6; Theorem 4.2]).

c) To prove the case of $j = 3$, we use the coupling by reflection and take $d(x, y) = |x - y|_1$. For $x_0 \geqslant y_0$, Lemma 5.2 gives $d(x_t, y_t) = \sum_{i=1}^d (x_i(t) - y_i(t))$, P^{x_0, y_0}-a.s. On the other hand, for $d(x, y) = \sum_i (x_i - y_i)$, we have

$$A(x, y) \geqslant \alpha_2(r), \qquad K_3(r) \leqslant - \sup_{x \geqslant y, d(x,y) \geqslant r} B(x, y).$$

Next, let u_n be the first Neumann eigenfunction on D_n, then there exists $x \geqslant y$ such that $u_n(x) \neq u_n(y)$. [15; Lemma 2.1 and the proof of Lemma 2.4] then give $\mathrm{gap}(D_n) \geqslant \delta$. This proves the theorem in the case of $j = 3$. □

Proof of Corollary 4.7. We consider the cases of $j = 1$ and $j = 3$ only since the proof of $j = 2$ is similar. Actually, by [7; Theorem 1.3], the lower bound given for $j = 2$ is also a lower bound of the logarithmic Sobolev constant. To see this, take the Riemannian metric $g(\partial_i, \partial_j) = \delta_{ij} a_i^{-1}$. Then $\{X_i\}$ is a normal orthogonal basis with $\nabla_{X_i} X_j = 0$ for all i, j, the sectional curvature is zero, ρ is the Riemannian distance and $L = \Delta_g + \nabla_g \overline{V}$.

a) For $|x - y| = r$, let $\psi(s) = x + s(y - x),\ s \in [0, 1]$. We have

$$\begin{aligned}
&\|\sigma(x) - \sigma(y)\|^2 - |x - y|^{-2} |(\sigma(x) - \sigma(y))(x - y)|^2 + \langle \bar{b}(x) - \bar{b}(y), x - y \rangle \\
&\quad \leqslant r^2 \kappa (1 - d^{-1}) + \sum_{i,j=1}^{d} (x_i - y_i)(x_j - y_j) \int_0^1 \partial_i \bar{b}_i(\psi(u)) \mathrm{d}u \\
&\quad \leqslant r^2 \kappa (1 - d^{-1}) - r^2 \lambda_{\min} \Big(- \int_0^1 \partial_j \bar{b}_i(\psi(u)) \mathrm{d}u \Big) \\
&\quad \leqslant r^2 \kappa (1 - d^{-1}) - r^2 \int_0^1 \lambda_{\min} \big(- \partial_j \bar{b}_i(\psi(u)) \big) \mathrm{d}u.
\end{aligned}$$

Next, choose $u_0 \in [0, 1]$ such that $|\psi(u_0) - p| = \min_{u \in [0,1]} |\psi(u) - p|$. Then $|\psi(u) - p| \geqslant |\psi(u) - \psi(u_0)| = |u - u_0| r$. Note that θ_1 is non-decreasing, we obtain

$$\int_0^1 \lambda_{\min} \big(- \partial_j \bar{b}_i(\psi(u)) \big) \mathrm{d}u \geqslant \int_0^1 \theta_1(|u - u_0| r) \mathrm{d}u \geqslant \frac{2}{r} \int_0^{r/2} \theta_1(u) \mathrm{d}u = \gamma_1(r/2).$$

Hence we can take $K_3(r) = \gamma_3(r/2)$.

b) Finally, note that

$$\begin{aligned}
\sum_{i=1}^{d} \big(\bar{b}_i(y) - \bar{b}_i(x) \big) &= \sum_{j=1}^{d} (y_j - x_j) \int_0^1 \sum_{i=1}^{d} \partial_j \bar{b}_i(\psi(u)) \mathrm{d}u \\
&\leqslant -|y - x|_1 \int_0^1 \big(- \max_j \sum_{i=1}^{d} \partial_j \bar{b}_i(\psi(u)) \big) \mathrm{d}u \\
&\leqslant -|y - x|_1 \int_0^1 \theta_3(|\psi(u) - p|) \mathrm{d}u \leqslant -|y - x|_1 \gamma_3(r/2),
\end{aligned}$$

we can take $K_3(r) = \gamma_3(r/2)$. □

9. Proof of Theorem 4.13

Lemma 9.1. Let (x_t, y_t) be a coupling of the L-diffusion process. If

$$E^{x,y} |x_t - y_t|^2 \leqslant |x - y|^2 \exp[2ct]$$

for all $t \geqslant 0,\ x, y \in \mathbf{R}^d$ and some $c \in \mathbf{R}$, then we have $|\nabla P_t f|^2 \leqslant \exp[2ct] P_t |\nabla f|^2$ for all $t \geqslant 0$ and $f \in C_0^1(\mathbf{R}^d)$.

Proof. Since $f \in C_0^1(\mathbf{R}^d)$, for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\frac{|f(x) - f(y)|}{|x - y|} \leqslant |\nabla f(x)| + \varepsilon, \qquad |x - y| \in (0, \delta).$$

Let T be the coupling time. We have

$$\begin{aligned}\frac{|P_tf(x)-P_tf(y)|}{|x-y|} &\leqslant E^{x,y}\left\{\frac{|f(x_t)-f(y_t)|}{|x_t-y_t|}\cdot\frac{|x_t-y_t|}{|x-y|}I_{\{T>t\}}\right\}\\ &\leqslant \exp[ct]\left\{E^{x,y}\frac{|f(x_t)-f(y_t)|^2}{|x_t-y_t|^2}I_{\{T>t\}}\right\}^{1/2}\\ &\leqslant \exp[ct]\big\{E^{x,y}(|\nabla f(x_t)|+\varepsilon)^2+\|\nabla f\|_\infty P^{x,y}(|x_t-y_t|\geqslant\delta)\big\}^{1/2}\\ &\leqslant \exp[ct]\big\{P_t|\nabla f|^2(x)+2\varepsilon\|\nabla f\|_\infty\\ &\quad+\varepsilon^2+\|\nabla f\|_\infty\delta^{-2}|x-y|^2\exp[2ct]\big\}^{1/2}.\end{aligned}$$

The assertion now follows by letting $y\to x$ and then $\varepsilon\to 0$. □

Proof of Theorem 4.13. a) Suppose that $f\in C_0^1(\mathbf{R}^d)$. Let $\widetilde{L}$ be the operator of march coupling (see [3] or [4]), i.e., $c(x,y)=\sigma(x)\sigma(y)^*$. Let $h(x,y)=|x-y|^2$, we have

$$\widetilde{L}h(x,y)=2\|\sigma(x)-\sigma(y)\|^2+2\langle b(x)-b(y),x-y\rangle\leqslant 2Kh(x,y),\qquad x,y\in\mathbf{R}^d.$$

Then $E^{x,y}|x_t-y_t|^2\leqslant|x-y|^2\exp[2Kt]$, $t\geqslant 0$. By Lemma 9.1 we have

$$|\nabla P_tf|^2\leqslant\exp[2Kt]P_t|\nabla f|^2,\qquad f\in C_0^1(\mathbf{R}^d).\tag{9.1}$$

For given $t>0$, let $H(r)=P_r(P_{t-r}f)^2$, $r\in[0,t]$. By (9.1) we have

$$\begin{aligned}H'(r)&=P_rL(P_{t-r}f)^2-2P_r(P_{t-r}f)LP_{t-r}f\\ &=2P_r\langle a\nabla P_{t-r}f,\nabla P_{t-r}f\rangle\leqslant 2\bar{a}\exp[2K(t-r)]P_t|\nabla f|^2.\end{aligned}$$

By integrating over r from 0 to t, we obtain (4.14).

b) In general, fix x and t, let $\pi_t=\delta_xP_t$. Next, given $f\in C^1(\mathbf{R}^d)$ with $\pi_t(f)=0$ and $\pi_t(f^2)=1$. Let $B_n=\{y:|y-x|\leqslant n\}$, $n\geqslant 1$. For any $\varepsilon>0$, there exists n_ε such that

$$\bar{a}\int_{B_n^c}(|\nabla f|^2+f^2)\mathrm{d}\pi_t+\pi_t(B_n^c)<\varepsilon$$

for all $n\geqslant n_\varepsilon$. Choose $h\in C^\infty(\mathbf{R})$ such that $0\leqslant h\leqslant 1$, $h(r)=1$ for $r\leqslant 0$ and $h(r)=0$ for $r\geqslant 1$. Let $f_n(y)=f(y)h(|y-x|-n)$. Then $f_n\in C_0^1(\mathbf{R}^d)$ and

$$\begin{aligned}&\bar{a}\int|\nabla f_n|^2\mathrm{d}\pi_t\geqslant\bar{a}\int|\nabla f|^2\mathrm{d}\pi_t-(\bar{a}\|h\|_\infty^2+1)\varepsilon,\\ &\int\left(f_n-\int f_n\mathrm{d}\pi_t\right)^2\mathrm{d}\pi_t\geqslant 1-3\varepsilon,\quad n\geqslant n_\varepsilon.\end{aligned}$$

Combining these with a) and letting $\varepsilon\to 0$, we complete the proof. □

Acknowledgements. The authors are greatly indebted to Prof. R. Durrett for encouragement and to a referee for careful comments on the first version of the paper.

References

[1]. Chavel, I., *Eigenvalues in Riemannian Geometry*, New York: Academic Press, 1984.

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

[3]. Chen, M. F., *Optimal Markovian couplings and applications*, Acta Math. Sin. New Ser. 10:3 (1994), 260–275.

[4]. Chen, M. F. and Li, S. F., *Coupling methods for multidimensional diffusion processes*, Ann. Probab. 17:1(1989), 151–177.

[5]. Chen, M. F. and Wang, F. Y., *Application of coupling method to the first eigenvalue on manifold*, Sci. Sin. (A), 37:1(1994), 1–14.

[6]. Chen, M. F. and Wang, F. Y., *Estimation of the first eigenvalue of the second order elliptic operators*, J. Funct. Anal. 131:2 (1995), 345–363.

[7]. Chen, M. F. and Wang, F. Y., *Estimates of logarithmic Sobolev constant– An improvement of Bakry-Emery criterion*, preprint (1994).

[8]. Chen, M. F. and Wang, F. Y., *On order-preservation and positive correlations for multi-dimensional diffusion processes*, Probab. Theory Relat. Fields, 95 (1993), 421–428.

[9]. Fukushima, M., Oshima, Y. and Takeda, M., *Dirichlet Forms and Symmetric Markov Processes*, Walter de Gruyter & Co., 1994.

[10]. Hsu, E. P., *Logarithmic Sobolev inequalities on path spaces*, preprint (1994).

[11]. Kac, I. S. and Krein, M. G., *Criteria for discreteness of the spectrum of a singular string*, Izv. Vyss. Učebn. Zaved. Mat 2 (1958), 136–153 (In Russian).

[12]. Kotani, S. and Watanabe, S., *Krein's spectral theory of strings and generalized diffusion processes*, Lecture Notes in Math., 923 (1982), 235–259.

[13]. Liggett, T. M., *Exponential L_2 convergence of attractive reversible nearest systems*, Ann. Prob. 17(1989), 403–432.

[14]. Lindvall, T. and Rogers, L. C. G., *Coupling of multidimensional diffusions by reflection*, Ann. Probab. 14 (1986), 860–872.

[15]. Wang, F. Y., *Application of coupling method to the Neumann eigenvalue problem*, Prob. Theory Relat. Fields, 98 (1994), 299–306.

[16]. Wang, F. Y., *Spectral gap for diffusion processes on non-compact manifolds*, Chinese Sci. Bull., 40:14 (1995), 1145–1149.

[17]. Wang, F. Y., *Logarithmic Sobolev inequalities for diffusion processes with application to path space*, preprint (1995).

[18]. Wang, F. Y., *Gradient estimates on* $\mathbf{R}^d$, Canad. Math. Bull., 37:4(1994), 560–570.

Received by the editors December 3, 1995.

Department of Mathematics, Beijing Normal University, Beijing 100875, P.R. China.

Appendix (Addition to the original proof. Jan. 26, 2010).

We study some property of eigenfunction in dimension one when the coefficients of the operator are not necessarily continuous. Let J be an interval of $\mathbb{R}$, finite or infinite, open or closed. The operator is $L = a(x)\frac{\mathrm{d}^2}{\mathrm{d}x^2} + b(x)\frac{\mathrm{d}}{\mathrm{d}x}$. Set $C(x) = \int_\theta^x b/a$ for some reference point $\theta \in J$. Throughout this appendix, we make the following

Hypothesis A.1.

(1) $a > 0$ on J.

(2) b/a and e^C/a are locally integrable with respect to the Lebesgue measure.

In the paper, we have used several times the "eigenfunction" f of $\lambda \in \mathbb{R}$ in the sense that $f \in C^2(J)$ and

$$Lf = -\lambda f \qquad \text{on } J. \tag{A.1}$$

To be distinguished, we call f an a.e.-eigenfunction of λ if f' is absolutely continuous on each compact subinterval of J and $(A.1)$ holds almost everywhere on J. Clearly, if a and b are continuous, then these eigenfunctions are the same.

The next result is standard in ODE, refer to [A1; Theorem 1.2.1 and its proof plus Theorem 2.2.1].

Theorem A.2. Under Hypothesis A.1, for every λ, $\gamma^{(0)}$, $\gamma^{(1)} \in \mathbb{R}$, an a.e.-eigenfunction f of $\lambda \in \mathbb{R}$ always exists and is indeed unique. More precisely, the function f can be obtained by the following successive approximation. Define

$$F^{(1)} = F(\theta) = \begin{pmatrix} \gamma^{(0)} \\ \gamma^{(1)} \end{pmatrix}, \quad F^{(n+1)}(x) = F(\theta) + \int_\theta^x GF^{(n)}, \quad x \in J,\ n \geqslant 1, \tag{A.2}$$

where $G(x) = \begin{pmatrix} 0 & e^{-C} \\ -\lambda e^C/a & 0 \end{pmatrix}$. Then

$$F^{(n)} \to \begin{pmatrix} f \\ e^C f' \end{pmatrix} =: F \qquad \text{as } n \to \infty \tag{A.3}$$

uniformly on each compact subinterval of J. In other words, F is the unique solution to the equation

$$F(x) = F(\theta) + \int_\theta^x GF, \qquad x \in J \tag{A.4}$$

and so it is absolutely continuous on each compact subinterval of J.

Theorem A.2 enables us to improve the last three results in Section 6, replacing the eigenfunction with the a.e.-eigenfunction. For instance, we have the following modificatoin.

Lemma A.3. Let $J = (x_0, \infty) \supset (\alpha, \beta)\,(\alpha < \beta)$. Under Hypothesis A.1, if an a.e.-eigenfunction f of $\lambda \in \mathbb{R}$ satisfies $f|_{(\alpha,\beta)} = 0$, then $f = 0$ on J.

Proof. Take $\theta = (\alpha + \beta)/2$ for instance. By assumption, f vanishes in a neighborhood of θ and so $F(\theta) = 0$. By induction and (A.2), it follows that $F^{(n)} = 0$ for all n. Therefore we have $f = 0$ by (A.3). □

Lemma A.4. Let $J = (x_0, \infty)$, $a > 0$, $\int_J e^C/a < \infty$, and let g be the a.e.-eigenfunction of $\lambda_1 \geqslant 0$ with $g(x_0) = 0$. Then $g \in L^1(\pi)$ and moreover $g'|_J \neq 0$ once $\lambda_1 > 0$.

Proof. The result $g \in L^1(\pi)$ is trivial in the case of $\lambda_1 = 0$ since g is a constant. Next, let $\lambda_1 > 0$. The original proof (Proposition 6.4) for $g'|_J \neq 0$ needs no change. We now assume that $g'|_J > 0$. By Theorem A.2, with $g'(x_0) = 0$ and $g(x_0) = -1$ (based on Lemma 6.2), there exists a dx-zero set U such that

$Lg = -\lambda_1 g$ on U^c. Clearly, $\pi(U) = 0$. Denote by $\{P_t^N\}_{t\geqslant 0}$ the (maximal) symmetric semigroup generated by the restricted operator L^N of L on (x_0, N) having reflection boundaries at x_0 and N. By symmetry, for each $t \geqslant 0$, we have

$$P_t^N \mathbb{1}_U = 0, \qquad \pi\text{-a.e.} \tag{A.5}$$

Denote by H_t the π-zero set in (A.5) and set $H = U \cup_{\text{rational } r\geqslant 0} H_r$. Then $\pi(H) = 0$ and (A.5) holds on H^c for all t by the right-continuity of $t \to P_t^N \tilde{g}$. Let $\tilde{g} = g - g(x_0)$. Then $\tilde{g}(x_0) = 0$ and $\tilde{g}'|_J > 0$. Noting that $L^N = L$ on (x_0, N), we have $L^N \tilde{g} \leqslant \lambda_1 - \lambda_1 \tilde{g}$ on $(x_0, N) \cap U^c$. Let $\pi^N(\mathrm{d}x) = e^C a^{-1} \mathrm{d}x \Big/ \int_{x_0}^N e^C a^{-1}$. Then

$$\pi^N\big(\tilde{g}\big) = \pi^N\big(P_t^N \tilde{g}\big) = \pi^N\bigg(\tilde{g} + \int_0^t P_s^N L^N \tilde{g} \mathrm{d}s\bigg) \leqslant \pi^N\big(\tilde{g}\big) + \pi^N\bigg(\int_0^t \big(\lambda_1 - \lambda_1 P_s^N \tilde{g}\big)\mathrm{d}s\bigg).$$

This gives us

$$t \geqslant \pi^N\bigg(\int_0^t P_s^N \tilde{g} \mathrm{d}s\bigg) = \int_0^N \pi^N\big(P_s^N \tilde{g}\big)\mathrm{d}s = t\pi^N\big(\tilde{g}\big)$$

and so $\pi^N\big(\tilde{g}\big) \leqslant 1$. The required assertion now follows by letting $N \to \infty$. □

Lemma A.5. Everything is the same as in Lemma A.4. We have $I(g)^{-1} = \lambda_1$ and then "=" in (2.4) holds.

Proof. Clearly, we need only to consider the case that $\lambda_1 > 0$. By using Lemma A.4 and [A2; Lemma 2.3] with a slight modification, we have $g' > 0$ and $\pi(g) = 0$. The last property means that $\int_{x_0}^x g e^C a^{-1} = -\int_x^\infty g e^C a^{-1}$. Since x_0 is the reflecting boundary, we have $g'(x_0) = 0$. By (A.4), we obtain

$$\big[e^C g'\big](x) = -\lambda_1 \int_{x_0}^x \frac{e^C}{a} g = \lambda_1 \int_x^\infty \frac{e^C}{a} g, \qquad x \in J.$$

This gives us the first assertion and then the second follows by part (2) of Theorem 2.1. □

Lemma A.5 is an addition to part (2) of Theorem 2.1. For part (1) of the theorem, since we use Cauchy's mean value theorem, some stronger conditions are required.

Lemma A.6. Under the assumptions of Lemma A.4, (2.3) with equality sign holds provided $a, b \in C[x_0, \infty)$ and having finite derivatives in (x_0, ∞), $f \in C^1[x_0, \infty)$ with $f|_{(x_0,\infty)} > 0$ and finite f'' in (x_0, ∞).

Proof. The original proof for "$\geqslant$" in (2.3) needs a little modification only. For "=" in the case of $\lambda_1 > 0$, simply use the eigenfunction as a test function. To see "=" in the case of $\lambda_1 = 0$, using the test function $h(x) = e^{-C(x)} \int_x^\infty e^C f/a$, it follows that the right-hand side of (2.3) is bounded below by $\inf_{x\in J} I(f)^{-1}$ which is nonnegative whenever $\pi(f) \geqslant 0$. □

References

[A1]. Zettl, A. (2005), *Sturm–Liouville Theory*, AMS, Providence, Rhode Island.

[A2]. Chen, M.F. (1998), *Estimate of exponential convergence rate in total variation by spectral gap*, Acta Math. Sin. Ser. (A) 41:1, 1–6; Acta Math. Sin. New Ser. 14:1, 9–16.

Sci. Sin. 40:4 (1997), 384–394

GENERAL FORMULA FOR LOWER BOUND OF THE FIRST EIGENVALUE ON RIEMANNIAN MANIFOLDS*

Mu-Fa Chen and Feng-Yu Wang

(Department of Mathematics, Beijing
Normal University, Beijing 100875, China)
Received January 22, 1996; revised September 6, 1996

Abstract. A general formula for the lower bound of the first eigenvalue on compact Riemannian manifolds is presented in this paper for the first time. The formula improves the main known sharp estimates including Lichnerowicz's estimate and Zhong-Yang's estimate. Moreover, the results are extended to the noncompact manifolds. The study is based on a probabilistic approach (i.e., the coupling method) introduced by the authors previously.

Keywords: The first eigenvalue, coupling method, Riemannian manifold.

Let M be a compact connected Riemannian manifold with boundary ∂M either empty or convex. Let $L = \Delta + \nabla V$ for some $V \in C^2(M)$. Denote by λ_1 the first (non-trivial) eigenvalue of L on M with Neumann boundary condition if $\partial M \neq \emptyset$.

The estimate of λ_1 is a well known topic in differential geometry (refer to the books Bérard [1], Chavel [2] and Schoen-Yau [3]). For recent progress of the study, readers are urged to refer to [4]–[7] in which a new technique (coupling method) was adopted. As a continuation of the above papers, this note presents a general formula for the lower bound of λ_1. The basic idea of the paper comes from a recent work by the authors[1] in which the same topic was treated for elliptic operators in $\mathbf{R}^d$.

1 Main results

Suppose that $\mathrm{Ric}_M \geqslant -K$ for some $K \in \mathbf{R}$. Let d, D and ρ denote respectively the dimension, diameter and Riemannian distance. Let

$$K(V) = \inf\{r : \mathrm{Hess}_V - \mathrm{Ric}_M \leqslant r\}.$$

Project supported in part by the National Natural Science Foundation of China, Qiu Shi Science & Technologies Foundation and the Foundation of Institution of Higher Education for Doctoral Program

[1]Chen, M. F. and Wang, F. Y., Estimation of spectral gap for elliptic operators, 1995, to appear in Trans. of AMS.

Denote by $\mathrm{cut}(x)$ the cut locus of x and define

$$a(r) = \sup\{\langle\nabla\rho(x,\cdot)(y), \nabla V(y)\rangle + \langle\nabla\rho(\cdot,y)(x), \nabla V(x)\rangle : \rho(x,y) = r, y \notin \mathrm{cut}(x)\}$$

for $r \in (0, D]$ and set $a(0) = 0$. Next, set $K^+ = \max\{0, K\}$, $K^- = (-K)^+$ and choose $\gamma \in C[0, D]$ such that

$$\gamma(r) \geqslant \min\Big\{K(V)r,\, 2\sqrt{K^+(d-1)}\tanh\Big[\frac{r}{2}\sqrt{K^+/(d-1)}\Big] - 2\sqrt{K^-(d-1)}\tan\Big[\frac{r}{2}\sqrt{K^-/(d-1)}\Big] + a(r)\Big\}.$$

Finally, define

$$C(r) = \exp\left[\frac{1}{4}\int_0^r \gamma(s)\mathrm{d}s\right], \qquad r \in [0, D].$$

Then the main result of the paper is the following.

Theorem 1. For any $f \in C[0, D]$ with $f > 0$ on $(0, D)$, we have

$$\lambda_1 \geqslant 4 \inf_{r\in(0,D)} f(r)\Big\{\int_0^r C(s)^{-1}\mathrm{d}s \int_s^D C(u)f(u)\mathrm{d}u\Big\}^{-1}. \tag{1.1}$$

Before moving further, let us make some comments on Theorem 1.

a) Let μ be the probability measure deduced from $\mathrm{e}^{V(x)}\mathrm{d}x$. Then the classical variational formula says that for every $f \in C^1(M)$ with $\mu(f) := \int_M f\mathrm{d}\mu = 0$, we have

$$\lambda_1 \leqslant \mu(\|\nabla f\|^2)/\mu(f^2).$$

From this, one may regard Theorem 1 as a dual of the classical variational formula. However, these two formulas have no common point. As far as we know, no analog of (1.1) has ever appeared in the literature. The proof of Theorem 1 is based on the coupling method[4],[6] which is completely different from all the known geometric approaches including Lichnerowicz's argument[8] and Li-Yau's technique (refer to [3]). We do not know at the moment whether (1.1) can be deduced from the previous geometric approaches.

b) We claim that all the estimates given in [4]–[6], which have already covered all the known sharp estimates (c.f. the comments right after the corollaries given below), can be deduced from Theorem 1. First, [4; Theorems 1.4, 1.5] and [6; Theorems 1.4, 1.5] were proved in terms of the first moment of the coupling time but now can be deduced directly from (1.1) by choosing $f \equiv 1$.

c) To cover the other results obtained in [4]–[6], it suffices to show that (1.1) is equivalent to the following differential formula: For any $g \in C^2[0, D]$ with $g(0) = 0$ and $g' > 0$ on $[0, D)$, we have

$$\lambda_1 \geqslant -\sup_{r\in(0,D)} \big\{4g''(r) + \gamma(r)g'(r)\big\}/g(r). \tag{1.2}$$

Actually, the results just mentioned were deduced from (1.2). To prove the equivalence, let f be given in (1.1) and take

$$g(r) = \int_0^r C(s)^{-1}\mathrm{d}s \int_s^D C(u)f(u)\mathrm{d}u.$$

Then (1.2) implies (1.1). On the other hand, for g given in (1.2), if

$$-\sup_{r\in(0,D)} \big\{4g''(r) + \gamma(r)g'(r)\big\}/g(r) =: \delta > 0,$$

then we have

$$\begin{aligned}\frac{1}{4}\int_0^r C(s)^{-1}\mathrm{d}s \int_s^D C(u)g(u)\mathrm{d}u &\leqslant \frac{1}{\delta}\int_0^r C(s)^{-1}\mathrm{d}s \int_s^D \big(-Cg'\big)'(u)\mathrm{d}u \\ &= \frac{1}{\delta}\int_0^r C(s)^{-1}\big[C(s)g'(s) - C(D)g'(D)\big]\mathrm{d}s \\ &\leqslant \frac{1}{\delta}g(r).\end{aligned}$$

Thus (1.1) implies (1.2) by taking $f = g$. Of course, each of (1.1) and (1.2) has its own advantage. The computation for (1.2) is much easier but as we have just shown, for the same test function (i.e., $f \equiv g$), the lower bound given by (1.2) can not be better than that given by (1.1). Since (1.2) was often used in our previous publications, this paper concentrates on (1.1).

d) Let $\bar{\lambda}_1$ be the first mixed eigenvalue of the operator $4\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \gamma(x)\frac{\mathrm{d}}{\mathrm{d}x}$ on $(0, D)$ with Dirichlet condition at 0 and Neumann condition at D and let $f(> 0 \text{ on } (0, D))$ be the corresponding eigenfunction. Then Theorem 1 implies that $\lambda_1 \geqslant \bar{\lambda}_1$. The equality indeed holds in the typical case that $M = S^d$ and $V \equiv 0$.

It should be not surprising that (1.1) can produce a lot of new estimates since the test function f can be quite arbitrary. But it is surprising that the estimates of the first eigenvalue given in [4] and [6] can still be improved, as illustrated by the following corollaries.

Corollary 1. In general, we have

$$\lambda_1 \geqslant K(V)\Big\{\exp\Big[\frac{1}{8}K(V)D^2\Big] - 1\Big\}^{-1} \geqslant \frac{8}{D^2} - \frac{1}{2}K(V). \tag{1.3}$$

Next,

$$\lambda_1 \geqslant \frac{\pi^2}{8}K(V)\Big\{\exp\Big[\frac{1}{8}K(V)D^2\Big] - 1\Big\}^{-1} \geqslant \frac{\pi^2}{D^2} - \frac{\pi^2}{16}K(V), \text{ if } K(V) \geqslant 0, \tag{1.4}$$

$$\lambda_1 \geqslant \frac{\pi^2}{D^2} - \Big(1 - \frac{2}{\pi}\Big)K(V), \qquad \text{if } K(V) \leqslant 0. \tag{1.5}$$

Corollary 2. Suppose that $V \equiv 0$. If $K \leqslant 0$, then

$$\lambda_1 \geqslant \frac{\pi^2}{D^2} - \max\Big\{\frac{\pi}{4d}, 1 - \frac{2}{\pi}\Big\}K, \tag{1.6}$$

$$\lambda_1 \geqslant -\frac{dK}{d-1}\Big\{1 - \cos^d\Big[\frac{D}{2}\sqrt{-K/(d-1)}\Big]\Big\}^{-1} \geqslant \frac{8}{D^2} - \frac{K}{2}, \quad d > 1. \tag{1.7}$$

Corollary 3. Suppose that $V \equiv 0$. If $K \geqslant 0$, then

$$\lambda_1 \geqslant \frac{\pi^2}{D^2} - \Big(\frac{\pi}{2} - 1\Big)K, \tag{1.8}$$

$$\lambda_1 \geqslant \frac{\pi^2}{D^2}\sqrt{1 + 2D^2K/\pi^4}\cosh^{1-d}\Big[\frac{D}{2}\sqrt{K/(d-1)}\Big], \quad d > 1. \tag{1.9}$$

Now, we compare the above estimates with some known best ones. Obviously, when $K(V) < 0$, each of (1.3) and (1.5) improves an estimate of [4]:

$$\lambda_1 \geqslant 8/D^2 - K(V)/3.$$

When $K < 0$, each of (1.5) and (1.6) improves Zhong-Yang's estimate[9]: $\lambda_1 \geqslant \pi^2/D^2$. When $K > 0$, each of (1.4) and (1.8) improves Cai's estimate[10]: $\lambda_1 \geqslant \pi^2/D^2 - K$. While (1.9) improves Yang-Jia's estimate[11],[12]:

$$\lambda_1 \geqslant \frac{\pi^2}{D^2}\exp\big[-\frac{1}{2}D\sqrt{K(d-1)}\big]$$

(this estimate was proved in [12] only for $d \geqslant 5$). Finally, since $D\sqrt{-K/(d-1)} \leqslant \pi$ for $K \leqslant 0$ and usually the strict inequality holds, (1.7) improves Lichnerowicz's estimate: $\lambda_1 \geqslant -dK/(d-1)$.

The remainder of the paper is organized as follows. A short proof of Theorem 1 is given at the beginning of the next section. Most of the section is devoted to prove the corollaries. The proofs are technical but contain a nice use of the FKG inequality which is well known in statistical physics. The noncompact case is studied in the last section.

2 Proofs

Proof of Theorem 1. Let (x_t, y_t) be the coupling by reflection of the L-diffusion process (with reflecting boundary if $\partial M \neq \emptyset$)[13],[14]. Then we have

$$\mathrm{d}\rho(x_t, y_t) \leqslant 2\sqrt{2}\mathrm{d}b_t + \gamma(\rho(x_t, y_t))\mathrm{d}t, \tag{2.1}$$

where b_t is a one-dimensional Brownian motion (see [4] and [6]). It should be mentioned that (2.1) was first proved by Kendall[13] for $V \equiv 0$ and $\gamma(r) = Kr$, the present form of γ is due to Cranston[14] and Chen-Wang[4] respectively for the cases $K \geqslant 0$ and $K \leqslant 0$. For $f \in C[0, D]$ with $f > 0$ on $(0, D)$, let δ be the lower bound given in Theorem 1 and define

$$g(r) = \int_0^r C(s)^{-1}\mathrm{d}s\int_s^D C(u)f(u)\mathrm{d}u, \qquad r \in [0, D].$$

By (2.1) and Itô's formula, we have

$$\mathrm{d}g(\rho(x_t,y_t)) \leqslant 2\sqrt{2}g'(\rho(x_t,y_t))\mathrm{d}b_t - \delta g(\rho(x_t,y_t))\mathrm{d}t.$$

Now, Theorem 1 follows from [4; Theorem 1.9]. □

The following elementary inequality will be used frequently in the remainder of this section.

Lemma 1 (FKG inequality). Let $p,q \in [-\infty,\infty]$ with $p<q$, and let $\nu(\mathrm{d}x)$ be a probability measure on (p,q). If $f,g \in C_b(p,q)$ are nondecreasing, then

$$\int_p^q f(x)g(x)\nu(\mathrm{d}x) \geqslant \int_p^q f(x)\nu(\mathrm{d}x)\int_p^q g(x)\nu(\mathrm{d}x).$$

Proof. Simply note that

$$\int_p^q fg\mathrm{d}\nu - \int_p^q f\mathrm{d}\nu\int_p^q g\mathrm{d}\nu = \frac{1}{2}\int_p^q\int_p^q \big[f(x)-f(y)\big]\big[g(x)-g(y)\big]\nu(\mathrm{d}x)\nu(\mathrm{d}y) \geqslant 0. \quad \square$$

Proof of Corollary 1. a) Take $\gamma(r) = K(V)r$, then $C(r) = \exp[\frac{1}{8}K(V)r^2]$. The first estimate of (1.3) follows from Theorem 1 with $f(r)=r$. The second bound of (1.3) is a linear approximation of the first one with respect to $K(V)$. To prove it, let

$$g(r) = r - \Big(\exp\Big[\frac{1}{8}D^2 r\Big] - 1\Big)\Big(\frac{8}{D^2} - \frac{r}{2}\Big), \qquad r \in \mathbf{R}.$$

Then $g(0) = g'(0) = 0$ and

$$g''(r) = \frac{D^4 r}{128}\exp\Big[\frac{1}{8}D^2 r\Big].$$

We have $g'(r) \geqslant 0$ for all r. Hence $g(r) \geqslant 0$ for $r \geqslant 0$ and $g(r) \leqslant 0$ for $r \leqslant 0$. This implies the second estimate of (1.3).

b) Suppose that $K(V) \geqslant 0$. Throughout of this section, set $\beta = \pi/(2D)$. Take $f(r) = \sin(\beta r)$. Applying the FKG inequality to $\nu(\mathrm{d}r) = Z^{-1}r\mathrm{d}r$, where and in what follows, Z is the normalizing constant to make $\nu(\mathrm{d}r)$ to be a probability measure, we obtain

$$\begin{aligned}
&\int_s^D \exp\Big[\frac{1}{8}K(V)r^2\Big]\sin(\beta r)\mathrm{d}r \\
&\quad = \int_s^D \exp\Big[\frac{1}{8}K(V)r^2\Big]\frac{\sin(\beta r)}{r}r\mathrm{d}r \\
&\quad \leqslant \frac{2}{D^2-s^2}\Big(\int_s^D \sin(\beta r)\mathrm{d}r\Big)\int_s^D \exp\Big[\frac{1}{8}K(V)r^2\Big]r\mathrm{d}r \\
&\quad = \frac{8\cos(\beta s)(\exp[\frac{1}{8}K(V)D^2] - \exp[\frac{1}{8}K(V)s^2])}{K(V)\beta(D^2-s^2)}.
\end{aligned}$$

Therefore

$$\begin{aligned}&\int_0^r C(s)^{-1}\mathrm{d}s\int_s^D C(u)f(u)\mathrm{d}u\\&\quad\leqslant\int_0^r\frac{8\cos(\beta s)}{K(V)\beta(D^2-s^2)}\Big\{\exp\Big[\frac{1}{8}(D^2-s^2)\Big]-1\Big\}\mathrm{d}s\\&\quad\leqslant\frac{8}{D^2\beta^2K(V)}\Big\{\exp\Big[\frac{1}{8}D^2\Big]-1\Big\}f(r),\quad r\in[0,D].\end{aligned}$$

By Theorem 1 we obtain the first estimate of (1.4) and then the second one follows from the second inequality of (1.3).

c) Suppose that $K(V)\leqslant 0$. Take $f(r)=\sin(\beta r)$. Then

$$\begin{aligned}I(s)&:=\int_s^D\exp\Big[\frac{1}{8}K(V)r^2\Big]\sin(\beta r)\mathrm{d}r\\&=\frac{\cos(\beta s)}{\beta}\exp\Big[\frac{1}{8}K(V)s^2\Big]+\frac{K(V)}{4\beta}\int_s^D\exp\Big[\frac{1}{8}K(V)r^2\Big]r\cos(\beta r)\mathrm{d}r.\end{aligned}\tag{2.2}$$

Applying the FKG inequality to $\nu(\mathrm{d}r)=Z^{-1}\sin(\beta r)\mathrm{d}r$, we get

$$\begin{aligned}\int_s^D\exp\Big[\frac{1}{8}K(V)r^2\Big]r\cos(\beta r)\mathrm{d}r&\geqslant I(s)\Big(\int_s^D\sin(\beta r)\mathrm{d}r\Big)^{-1}\int_s^D r\cos(\beta r)\mathrm{d}r\\&=\cos^{-1}(\beta s)I(s)\big[D-s\sin(\beta s)-\beta^{-1}\cos(\beta s)\big]\\&\geqslant I(s)D\Big(1-\frac{2}{\pi}\Big)\cos(\beta s).\end{aligned}\tag{2.3}$$

Here we have used the fact that

$$\pi/2-r\sin r-\cos r\geqslant(1-2/\pi)\cos^2 r$$

for all $r\in[0,\pi/2]$. Combining (2.2) with (2.3) we obtain

$$I(s)\leqslant\frac{\cos(\beta s)}{\beta}\exp\Big[\frac{1}{8}K(V)s^2\Big]+\frac{DK(V)}{4\beta}\Big(1-\frac{2}{\pi}\Big)I(s)\cos(\beta s).\tag{2.4}$$

Next, we claim that $g(s):=I(s)\exp[-\frac{1}{8}K(V)s^2]$ is nonincreasing in s. Actually, noticing that $r^{-1}\sin(\beta r)$ is decreasing, we have

$$\begin{aligned}g'(s)&=\frac{-K(V)}{4}s\exp\Big[-\frac{1}{8}K(V)s^2\Big]\int_s^D\exp\Big[\frac{1}{8}K(V)r^2\Big]\sin(\beta r)\mathrm{d}r-\sin(\beta s)\\&\leqslant\frac{-K(V)}{4}\sin(\beta s)\exp\Big[-\frac{1}{8}K(V)s^2\Big]\int_s^D\exp\Big[\frac{1}{8}K(V)r^2\Big]r\mathrm{d}r-\sin(\beta s)\\&=-\sin(\beta s)\exp\Big[\frac{1}{8}K(V)(D^2-s^2)\Big]\\&\leqslant 0.\end{aligned}$$

Applying the FKG inequality to $\nu(\mathrm{d}r) = Z^{-1}\mathrm{d}r$, we get

$$\int_0^r g(s)\cos(\beta s)\mathrm{d}s \geqslant \frac{\sin(\beta r)}{\beta r}\int_0^r g(s)\mathrm{d}s \geqslant \frac{2}{\pi}\int_0^r g(s)\mathrm{d}s.$$

Finally, multiplying the both sides of (2.4) by $\exp\big[-\frac{1}{8}K(V)s^2\big]$ and then making the integration from 0 to r, we obtain

$$\int_0^r g(s)\mathrm{d}s \leqslant \Big[1-\frac{D^2K(V)}{\pi^2}\Big(1-\frac{2}{\pi}\Big)\Big]^{-1}\frac{\sin(\beta r)}{\beta^2}.$$

From this and Theorem 1, (1.5) follows. □

For the case $V \equiv 0$, we choose

$$\gamma(r)=2\sqrt{K^+(d-1)}\tanh\big[\frac{r}{2}\sqrt{K^+/(d-1)}\big]-2\sqrt{K^-(d-1)}\tan\big[\frac{r}{2}\sqrt{K^-/(d-1)}\big].$$

Then

$$C(r) = \begin{cases} \cosh^{d-1}\big[\frac{r}{2}\sqrt{K/(d-1)}\big], & \text{if } K \geqslant 0,\\ \cos^{d-1}\big[\frac{r}{2}\sqrt{-K/(d-1)}\big], & \text{if } K \leqslant 0. \end{cases}$$

From now on, set $\alpha = \frac{1}{2}\sqrt{|K|/(d-1)}$. In what follows, we will often use the following simple result.

Lemma 2. Let $f \in C^1[0,D]$. If there exists $r_0 \in [0,D]$ such that $f' \leqslant 0$ on $[0,r_0]$ and $f' \geqslant 0$ on $[r_0,D]$, then $f \leqslant \max\{f(0),f(D)\}$ on $[0,D]$.

Proof of Corollary 2. a) Since (1.5) holds and $\pi/(4d) < 1-2/\pi$ for all $d>2$, to prove (1.6) we need only to show that $\lambda_1 \geqslant \pi^2/D^2 - \pi K/(4d)$ for $d=2$. To this end, take $f(r)=\sin(\beta r)$. Noticing that $\alpha \leqslant \beta$, we have

$$\begin{aligned} I(s) &:= \int_s^D \cos(\alpha r)\sin(\beta r)\mathrm{d}r\\ &= \frac{1}{\beta}\cos(\alpha s)\cos(\beta s) - \frac{\alpha}{\beta}\int_s^D \sin(\alpha r)\cos(\beta r)\mathrm{d}r\\ &\leqslant \frac{1}{\beta}\cos(\alpha s)\cos(\beta s) - \frac{\alpha^2}{\beta^2}\int_s^D \sin(\beta r)\cos(\beta r)\mathrm{d}r\\ &= \frac{1}{\beta}\cos(\alpha s)\cos(\beta s) - \frac{\alpha^2}{2\beta^3}\cos^2(\beta s). \end{aligned}$$

Next, by Lemma 2 we see that $g(r) := \frac{D}{2}\sin(\beta r) - \int_0^r \cos^2(\beta s)\mathrm{d}s \leqslant 0$ on $[0,D]$. Hence

$$\begin{aligned} \int_0^r I(s)\cos^{-1}(\alpha s)\mathrm{d}s &= \frac{\sin(\beta r)}{\beta^2} - \frac{\alpha^2}{2\beta^3}\int_0^r \cos^{-1}(\alpha s)\cos^2(\beta s)\mathrm{d}s\\ &\leqslant \frac{\sin(\beta r)}{\beta^2}\Big(1-\frac{\alpha^2 D}{4\beta}\Big) = \frac{\sin(\beta r)}{\beta^2}\Big(1-\frac{\alpha^2 D}{4\beta}\Big). \end{aligned}$$

By Theorem 1 we obtain

$$\lambda_1 \geqslant \frac{\pi^2}{D^2}\Big(1-\frac{\alpha^2 D}{4\beta}\Big)^{-1} \geqslant \frac{\pi^2}{D^2}\Big(1+\frac{\alpha^2 D}{4\beta}\Big) = \frac{\pi^2}{D^2} - \frac{\pi}{8}K.$$

b) To prove (1.7), take $f(r)=\sin(\alpha r)$. We have

$$\begin{aligned}&\int_0^r \cos^{1-d}(\alpha s)\mathrm{d}s \int_s^D \cos^{d-1}(\alpha u) f(u)\mathrm{d}u\\ &\quad= \frac{1}{d\alpha}\int_0^r \cos^{1-d}(\alpha s)\big[\cos^d(\alpha s)-\cos^d(\alpha D)\big]\mathrm{d}s\\ &\quad\leqslant \alpha^{-2}d^{-1}\big[1-\cos^d(\alpha D)\big]f(r).\end{aligned}$$

By Theorem 1 we obtain the first estimate. To check the second one, we need only to prove that

$$g(r) := \frac{dr^2}{d-1} - \Big(\frac{8}{D^2}+\frac{r^2}{2}\Big)\big[1-\cos^d(\sigma r)\big] \geqslant 0, \qquad r\geqslant 0,$$

where $\sigma = D/(2\sqrt{d-1})$. Note that

$$\begin{aligned}g'(r) &= \frac{2dr}{d-1} - r\big[1-\cos^d(\sigma r)\big] - \Big(\frac{8}{D^2}+\frac{r^2}{2}\Big)d\sigma\cos^{d-1}(\sigma r)\sin(\sigma r)\\ &\geqslant \frac{2dr}{d-1} - r\big[1-\cos^d(\sigma r)\big] - \Big(\frac{8}{D^2}+\frac{r^2}{2}\Big)d\sigma^2 r\cos^{d-1}(\sigma r) = rh(r),\end{aligned}$$

where

$$h(r) = 2d/(d-1) - 1 + \cos^d(\sigma r) - (8/D^2 + r^2/2)d\sigma^2\cos^{d-1}(\sigma r).$$

We have $h(0)=0$ and

$$\begin{aligned}h'(r) =& -d\sigma\cos^{d-1}(\sigma r)\sin(\sigma r) - rd\sigma^2\cos^{d-1}(\sigma r)\\ &+\Big(\frac{8}{D^2}+\frac{r^2}{2}\Big)d\sigma^3(d-1)\cos^{d-2}(\sigma r)\sin(\sigma r)\\ \geqslant& \sigma d\cos^{d-2}(\sigma r)\sin(\sigma r)\Big(-2+\frac{8}{D^2}(d-1)\sigma^2\Big)\\ =& 0.\end{aligned}$$

Hence

$$g'(r) \geqslant rh(r) \geqslant rh(0) = 0.$$

This implies that $g(r)\geqslant g(0)=0$. □

Proof of Corollary 3. a) Take $f(r)=\sin(\beta r)$. Then

$$\begin{aligned}I(s) :=& \int_s^D C(u)f(u)\mathrm{d}u\\ =& \int_s^D \cosh^{d-1}(\alpha u)\sin(\beta u)\mathrm{d}u\\ =& \frac{1}{\beta}\cosh^{d-1}(\alpha s)\cos(\beta s) + \frac{\alpha}{\beta}(d-1)\int_s^D \cosh^{d-2}(\alpha u)\sinh(\alpha u)\cos(\beta u)\mathrm{d}u\\ \leqslant& \frac{1}{\beta}\cosh^{d-1}(\alpha s)\cos(\beta s) + \frac{\alpha^2}{\beta}(d-1)\int_s^D \cosh^{d-1}(\alpha u)u\cos(\beta u)\mathrm{d}u. \qquad (2.5)\end{aligned}$$

Here in the last step, we have used the fact that $\sinh r \leqslant r\cosh r$ for all $r \geqslant 0$. Noting that $u\cot(\beta u)$ is decreasing while $\cosh^{d-1}(\alpha u)$ is increasing and applying the FKG inequality to $\nu(\mathrm{d}r) = Z^{-1}\sin(\beta r)$, we obtain

$$\begin{aligned}\int_s^D \cosh^{d-1}(\alpha u)u\cos(\beta u)\mathrm{d}u &= \int_s^D \cosh^{d-1}(\alpha u)u\cot(\beta u)\sin(\beta u)\mathrm{d}u\\ &\leqslant I(s)\Big(\int_s^D \sin(\beta u)\mathrm{d}u\Big)^{-1}\int_s^D u\cos(\beta u)\mathrm{d}u\\ &= I(s)\cos^{-1}(\beta s)\big[D - s\sin(\beta s) - \beta^{-1}\cos(\beta s)\big]\\ &\leqslant (D-\beta^{-1})I(s)\\ &= D\Big(1-\frac{2}{\pi}\Big)I(s). \end{aligned} \tag{2.6}$$

Here, we have used the fact that $D - s\sin(\beta s) \leqslant D\cos(\beta s)$ which can be deduced by using Lemma 2. By (2.5) and (2.6) we obtain

$$I(s) \leqslant \frac{1}{\beta}\cosh^{d-1}(\alpha s)\cos(\beta s)\Big[1-\Big(1-\frac{2}{\pi}\Big)\frac{D^2K}{2\pi}\Big]^{-1}.$$

Then (1.8) follows from Theorem 1.

b) Take $f(r) = \cosh^{1-d}(\alpha r)\sin(\beta r)$, we have

$$\begin{aligned}&\int_0^r C(s)^{-1}\mathrm{d}s\int_s^D C(u)f(u)\mathrm{d}u\\ &\quad= \frac{1}{\beta}\int_0^r \cosh^{1-d}(\alpha s)\cos(\beta s)\mathrm{d}s\\ &\quad\leqslant \frac{1}{\beta}\cosh^{d-1}(\alpha D)f(r)\int_0^D \cosh^{1-d}(\alpha s)\cos(\beta s)\mathrm{d}s. \end{aligned} \tag{2.7}$$

To check the last inequality, let $c = \int_0^D \cosh^{1-d}(\alpha s)\cos(\beta s)\mathrm{d}s$ and take

$$g(r) = \int_0^r \cosh^{1-d}(\alpha s)\cos(\beta s)\mathrm{d}s - c\cosh^{d-1}(\alpha D)f(r).$$

Then $g'(r) = \cosh^{1-d}(\alpha r)\cos(\beta r)h(r)$, where

$$h(r) = 1 + c(d-1)\alpha\cosh^{d-1}(\alpha D)\tanh(\alpha r)\tan(\beta r) - c\beta\cosh^{d-1}(\alpha D).$$

Since $h(r)$ is increasing in r and $h(0) < 0,\ h(D) = \infty$, it follows from Lemma 2 that $g(r) \leqslant \max\{g(0), g(D)\} = 0$. This proves the required assertion. By (2.7) and Theorem 1 we have

$$\lambda_1 \geqslant \frac{\pi^2}{D^2}\cosh^{1-d}(\alpha D)\Big\{\beta\int_0^D \cosh^{1-d}(\alpha s)\cos(\beta s)\mathrm{d}s\Big\}^{-1}. \tag{2.8}$$

c) Now we go to estimate the integral in the right-hand side of (2.8). Let $c = D^2K/(2\pi^2)$ and $g(r) = \cosh^{d-1}(\alpha r) - 1 - c\sin^2(\beta r)$. Then $g(0) = 0$ and

$$\begin{aligned} g'(r) &= (d-1)\alpha\cosh^{d-2}(\alpha r)\sinh(\alpha r) - 2c\beta\sin(\beta r)\cos(\beta r) \\ &\geqslant (d-1)\alpha^2 r - 2c\beta^2 r = 0. \end{aligned}$$

Hence $g(r) \geqslant 0$ and so (2.8) implies that

$$\lambda_1 \geqslant \frac{\pi^2}{D^2}\cosh^{1-d}(\alpha D)\left\{\frac{\pi\sqrt{2}}{D\sqrt{K}}\arctan\Big[\frac{D\sqrt{K}}{\pi\sqrt{2}}\Big]\right\}^{-1}. \tag{2.9}$$

Next, let $c = 4/\pi^2$ and $g(r) = \arctan\,\mathrm{r} - r/\sqrt{1+cr^2}$. Then

$$g'(r) = (1+r^2)^{-1} - (1+cr^2)^{-3/2} \geqslant 0$$

if and only if

$$h(r) := c^3r^4 + (3c^2-1)r^2 + 3c - 2 \geqslant 0.$$

Since $h(0) = 3c - 2 < 0$, there exists uniquely $r_0 > 0$ such that $h(r) \leqslant 0$ on $[0, r_0]$ and $h(r) > 0$ for $r > r_0$. By Lemma 2, we have $g(r) \leqslant \max\{g(0), g(\infty)\} = 0$. Therefore $r(\arctan r)^{-1} \geqslant \sqrt{1 + 4r^2/\pi^2}$ and (1.9) then follows from (2.9). □

3 Spectral gap for noncompact manifolds

Let M be a complete connected Riemannian manifold with $D = \infty$. Suppose that $\mathrm{Ric}_M \geqslant -K$ for some $K \geqslant 0$, $K(V) < \infty$ and

$$Z := \int \exp[V(x)]\mathrm{d}x < \infty.$$

Then the L-diffusion process is nonexplosive with reversible measure $\mu(\mathrm{d}x) = Z^{-1}\exp[V(x)]\mathrm{d}x$. The spectral gap of L is characterized as

$$\lambda_1 = \inf\{\mu(\|\nabla f\|^2)/\mu(f^2) : f \in C^1(M) \cap L^2(\mu), \mu(f) = 0\}.$$

Let γ and C be the same as defined in Section 1 but replacing D with ∞.

Theorem 2. If there exists a sequence of convex regular domains $D_n \uparrow M$, then Theorem 1 holds in the present case with D replaced by ∞.

Proof. Let $\lambda_1(n)$ be the first Neumann eigenvalue of L on D_n. Then, the proof of [7; Lemma 1] gives us $\lambda_1 \geqslant \overline{\lim_{n\to\infty}}\lambda_1(n)$. Theorem 2 then follows from Theorem 1. □

By Whitehead theorem (see [15; Theorem 5.14]), the assumption of Theorem 2 holds if the sectional curvatures of M are nonpositive and the cut locus of some point is empty.

For fixed $p \in M$, let

$$\beta(r) = \inf_{\rho(x,p)\geqslant r}\{-\mathrm{Hess}_V(X,X) : X \in T_xM, \|X\| = 1\}.$$

Then we have the following result.

Corollary 4. Under the assumption of Theorem 2, if $\beta(\infty) := \lim_{r\to\infty}\beta(r) > 0$, then $\lambda_1 > 0$. If additionally the sectional curvatures are nonpositive and the cut locus of each point is empty, then Theorem 2 holds with

$$\gamma(r) = 2\sqrt{K(d-1)}\tanh\left[\frac{r}{2}\sqrt{K/(d-1)}\right] - 2\int_0^{r/2}\beta(u)\mathrm{d}u. \tag{3.1}$$

Especially, for this γ we have

$$\lambda_1 \geqslant \frac{8}{a_0^2}\exp\left[-1-\frac{1}{4}\int_0^{a_0}\gamma(u)\mathrm{d}u\right] > 0,$$

where $a_0 > 0$ is the unique solution to the equation $\gamma(a) = -8/a$.

Proof. a) Suppose that $\beta(r_0) > 0$ for some $r_0 > 0$. For every minimal geodesic, the length of the part contained in the geodesic ball $B(p, r_0)$ is not larger than $2r_0$. Let $x, y \in M$ with $\rho(x, y) = r$ and let $l(s) : [0, r] \to M$ be the minimal geodesic from x to y with unit tangent vector field U_s. Then

$$\begin{aligned}&\langle\nabla V(x), \nabla\rho(\cdot, y)(x)\rangle + \langle\nabla V(y), \nabla\rho(x, \cdot)(y)\rangle\\ &\quad= \int_0^r \mathrm{Hess}_V(U_s, U_s)\mathrm{d}s\\ &\quad\leqslant 2r_0[\beta(r_0) - \beta(0)] - r\beta(r_0).\end{aligned} \tag{3.2}$$

Hence we can choose

$$\gamma(r) = 2\sqrt{K(d-1)} + 2r_0[\beta(r_0) - \beta(0)] - r\beta(r_0)$$

and the first assertion of the corollary follows from Theorem 2 with $f(r) = r$.

b) From now on, we assume that the sectional curvatures are nonpositive and the cut locus of each point is empty. Let $l : [0, \rho(x, y)] \to M$ be the minimal geodesic from x to y, take s_0 such that $\rho(p, l(s_0)) = \min_s\{\rho(p, l(s))\}$. We claim that $\rho(p, l(s)) \geqslant |s - s_0|$. Without loss of generality, assume that $s > s_0$. Let $X_1 = \exp^{-1}_{l(s_0)}(p)$ and $X_2 = \exp^{-1}_{l(s_0)}(l(s))$, then

$$\langle X_1, X_2\rangle = -\frac{\mathrm{d}}{\mathrm{d}s}\rho(p, l(s))\Big|_{s=s_0} \leqslant 0.$$

Next, choose $p', q' \in \mathbf{R}^d$ such that $|p'| = \rho(p, l(s_0))$, $|q'| = s - s_0$ and $\langle p', q'\rangle = \langle X_1, X_2\rangle$. Let $I : T_{l(s_0)}M \to \mathbf{R}^d$ be a linear map preserving the inner product and satisfying $I(X_1) = p', I(X_2) = q'$. Finally, let $c(t) : [0, \rho(p, l(s))] \to M$ be the minimal geodesic from p to $l(s)$. Then $\bar{c}(t) := I \circ \exp^{-1}_{l(s_0)} \circ c(t)$ is a curve from p' to q'. By Rauch comparison theorem (see [15; Corollary 1.30]), we have

$$\rho(p, l(s)) \geqslant \text{length of } \bar{c}(t) \geqslant |p' - q'| \geqslant |q'| = s - s_0.$$

Here, we have used the fact $\langle p', q'\rangle \leqslant 0$.

c) Given $r > 0$, let $x, y, l(s)$ and U_s be the same as in a). By (3.2) and b) we obtain

$$\begin{aligned}&\langle\nabla V(x), \nabla\rho(\cdot, y)(x)\rangle + \langle\nabla V(y), \nabla\rho(x, \cdot)(y)\rangle\\ &\qquad\leqslant -\int_0^{s_0}\beta(s)\mathrm{d}s - \int_0^{r-s_0}\beta(s)\mathrm{d}s\\ &\qquad\leqslant -2\int_0^{r/2}\beta(u)\mathrm{d}u,\end{aligned}$$

where the last step is due to the fact that β is nondecreasing. Hence Theorem 2 holds with $\gamma(r)$ given by (3.1).

d) Noting that $\gamma(r)/r$ is decreasing and $-8/r^2$ is increasing, the solution $a_0 > 0$ exists uniquely whenever $\beta(\infty) > 0$. Take

$$f(r) = r\exp\bigg[-\frac{r}{4}\int_0^{r\wedge a_0}\Big(\gamma(u) - \frac{u}{a_0}\gamma(a_0)\Big)\mathrm{d}u\bigg].$$

Then Theorem 2 gives us

$$\begin{aligned}\lambda_1 &\geqslant -\frac{\gamma(a_0)}{a_0}\exp\bigg[-\frac{1}{4}\int_0^{a_0}\Big(\gamma(u) - \frac{u}{a_0}\gamma(a_0)\Big)\mathrm{d}u\bigg]\\ &= \frac{8}{a_0^2}\exp\bigg[-1 - \frac{1}{4}\int_0^{a_0}\gamma(u)\mathrm{d}u\bigg]. \quad \square\end{aligned}$$

References

[1]. Bérard P H, *Spectral Geometry, Direct and Inverse Problem*, LNM, 1207, Springer-Verlag, 1986.

[2]. Chavel I, *Eigenvalues in Riemannian Geometry*, Academic Press, 1984.

[3]. Schoen R, Yau S T, *Differential Geometry* (*in Chinese*), Beijing: Science Press, 1988.

[4]. Chen M F, Wang F Y, *Application of coupling method to the first eigenvalue on manifold*, Sci. Sin. (A), 1994, 37(1): 1–14.

[5]. Chen M F, *Optimal Markovian couplings and applications*, Acta Math. Sin New Ser, 1994, 10(3): 260–275.

[6]. Wang F Y, *Application of coupling method to the Neumann eigenvalue problem*, Prob Th Re Fields, 1994, 98: 299–306.

[7]. Wang F Y, *Spectral gap for diffusion processes on noncompact manifolds*, Chin Sci Bull, 1995, 40(14): 1145–1149.

[8]. Lichnerowicz A, *Géométrie des Groupes des Transformations*, Paris, Dunod, 1958.

[9]. Zhong J Q, Yang H C, *On estimate of the first eigenvalue of a compact Riemannian manifold*, Sci Sin (A), 1984, 12: 1251–1265.

[10]. 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.

[11]. Yang H C, *Estimate of the first eigenvalue on a compact Riemannian manifold with negative lower bound of Ricci curvature* (*in Chinese*), Sci Sin (A), 1989, 32(7): 689–700.

[12]. Jia F, *Estimate of the first eigenvalue on a compact Riemannian manifold with negative lower bound of Ricci curvature* (*in Chinese*), Chin Ann Math, 1991, 12(4): 496–502.

[13]. Kendall W S, *Nonnegative Ricci curvature and the Brownian coupling property*, Stochastics, 1986, 19: 111–129.

[14]. Cranston M, *Gradient estimates on manifolds using coupling*, J Funct Anal, 1991, 99: 110–124.

[15]. Cheeger J, Ebin D G, *Comparison Theorems in Riemannian Geometry*, North-Holland, 1975.

Probability Towards 2000
Edited by L. Accardi and C. Heyde
Lecture Notes in Statistics, Vol. **128**, 123–136
Springer-Verlag, 1998

TRILOGY OF COUPLINGS AND GENERAL FORMULAS FOR LOWER BOUND OF SPECTRAL GAP

Mu-Fa Chen

(Beijing Normal University)
December 31, 1995

Abstract. This paper starts from a nice application of the coupling method to a traditional topic: the estimation of spectral gap (=the first non-trivial eigenvalue). Some new variational formulas for the lower bound of the spectral gap of Laplacian on manifold or elliptic operators in $\mathbb{R}^d$ or Markov chains are reported [10],[15],[16]. The new formulas are especially powerful for the lower bounds, they have no common points with the classical variational formula (which goes back to Lord S. J. W. Rayleigh (1877) or E. Fischer (1905) and is particularly useful for the upper bounds). No analog of the new formulas ever appeared before. The formulas enable us to recover or improve the main known results. This will be illustrated by a comparison of the new results with the known ones in geometry. Next, we will explain the mathematical tool for proving the results. That is, the trilogy of the recent development of the coupling theory: The Markovian coupling, the optimal Markovian coupling and the construction of distances for coupling. Finally, some related results and some problems for the further study are also mentioned. It is hoped that the paper could be readable not only for probabilists but also for geometers and analysts.

Part I. Backgrounds.

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

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

2000 *Mathematics Subject Classification.* 35P15, 60H30, 60J27, 60J80.
Key words and phrases. Diffusions, manifold, Markov chains, spectral gap, couplings.
Research supported in part by NSFC and the State Education Commission of China.

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:

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

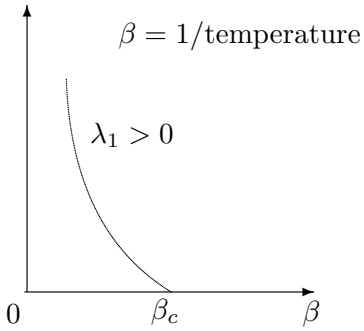
Hence, there is a gap between λ_0 and λ_1:

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

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.



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

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

where $\pi(f) = \int f d\pi$, with the largest rate λ_1 and when the temperature goes to the critical value, the rate will go to zero. This provides a way to describe the phase transitions and it is now an active research field[6],[33],[39],[41],[43],[44],[51].

The next application we would like to mention is the

(2) Computer Science.

a) Complexity of randomized approximation algorithms. The existence of spectral gap is used to prove a randomized approximation algorithm to be polynomial. See Jerrum and Sinclair (1989) for instance.

b) A fashionable application of the topic is the Markov chains Monte Carlo. There are too many publications to be listed here.

(3) Finally but not the last, the spectral gap have been used by Aldous and Brown (1993) and by Iscoe and McDonald (1994) for the asymptotics of the exit times.

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. 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 order of g is changed from linear to quadratic. Next, for finite state space, it is trivial when $E=\{0,1\}$, $\lambda_1 = a_1 + b_0$. 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

$$\lambda_1 = 2^{-1}\big[a_1 + a_2 + b_0 + b_1 - \sqrt{(a_1 - a_2 + b_0 - b_1)^2 + 4a_1b_1}\big].$$

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

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.

One may think all these examples are rather special but the last one of birth-death process and the last two of diffusions are indeed new. Actually, we were unable in [13] to cover by our approach the general case of one-dimensional diffusions, for which an analytic approach was adopted.

Part II. Old Results and New Results.

1. Story of λ_1 in Geometry.

The most well-developed subject of the first eigenvalue λ_1 is the Riemannian geometry. For instance, a large part of each book [2], [4] and [40] is devoted to the problem. About 2000 references are included in [2]. For the latter use, we now review quickly some famous estimates obtained by the geometers.

Let (M, g) be a compact and connected Riemannian manifold with Riemannian metric g. Denote by d and D respectively the dimension and the diameter of M. Assume that $\mathrm{Ricci}_M \geqslant Kg$ for some $K \in \mathbb{R}$. The main aim of the study is to use

the geometric quantities d, D and K to estimate λ_k's of Laplacian Δ. The main lower bounds of λ_1 obtained by different authors are listed in the following table.

Lichnerowicz (1958).	$\underline{\underline{\dfrac{d}{d-1}K}}, \quad K \geqslant 0.$
Li and Yau (1980).	$\dfrac{\pi^2}{2D^2}, \quad K \geqslant 0.$
Zhong and Yang (1984).	$\underline{\underline{\dfrac{\pi^2}{D^2}}}, \quad K \geqslant 0$
Li and Yau (1980).	$\dfrac{1}{D^2(d-1)\exp\left[1+\sqrt{1-4D^2K(d-1)}\right]}, \quad K \leqslant 0.$
Cai (1991).	$\underline{\underline{\dfrac{\pi^2}{D^2}+K}}, \quad K \leqslant 0.$
Yang (1989) and Jia (1991).	$\underline{\underline{\dfrac{\pi^2}{D^2}e^{-\alpha/2}}}, \quad \text{if } d \geqslant 5, \quad K \leqslant 0.$
Yang (1989) and Jia (1991).	$\dfrac{\pi^2}{2D^2}e^{-\alpha'/2}, \quad \text{if } 2 \leqslant d \leqslant 4, \quad K \leqslant 0$

where $\alpha = D\sqrt{-K(d-1)}$ and $\alpha' = D\sqrt{-K((d-1)\vee 2)}$.

The first estimate is very good since it is optimal for the sphere in any dimension $d \geqslant 2$. The third one is optimal when $K = 0$. The last line but one (for all $d \geqslant 2$) is called the Yau's conjecture (mentioned in Yang (1989)). The importance of Li and Yau's work is that it introduced a new approach and made a deep influence to the subsequent study. No doubt, the results are very deep in geometry.

It was only recently that the coupling approach was introduced for the first time to study the estimate of λ_1 and produced the following estimates, some of them are new in geometry.

2. Theorem (Chen and Wang (1993)).

$$\lambda_1 \geqslant \max\left\{\underline{\underline{\frac{\pi^2}{D^2}}}, \ \underline{\underline{\frac{d}{d-1}K}}, \frac{8}{D^2}+\frac{K}{3}\right\}, \qquad \text{if } K \geqslant 0$$

$$\lambda_1 \geqslant \max\left\{\underline{\underline{\frac{\pi^2}{D^2}+K}}, \ \frac{8}{D^2}+\frac{K}{3}, \frac{8}{D^2}\exp\left[\frac{D^2K}{8}\right], \ \frac{8}{D^2}\left(1+\frac{\alpha}{3}\right)e^{-\alpha/2}, \right.$$
$$\left. \underline{\underline{-\frac{K(d-1)}{4}\tanh^2\left(\frac{D}{2}\sqrt{\frac{-K}{d-1}}\right)\operatorname{sech}^2\theta}}\right\}, \qquad \text{if } K \leqslant 0,$$

where θ is decreasing limit of θ_n:

$$\theta_1 = \frac{\alpha}{4}\tanh\left(\frac{D}{2}\sqrt{\frac{-K}{d-1}}\right), \qquad \theta_n = \theta_1\tanh\theta_{n-1}, \quad n \geqslant 2.$$

The last estimate is taken from [8] and it is sharp in some sense.

Clearly, all the sharp estimates in the previous table are included here. Moreover, the last two estimates in the table are also included:

$$\max\left\{\frac{\pi^2}{D^2}+K,\ \frac{8}{D^2}\left(1+\frac{\alpha}{3}\right)e^{-\alpha/2}\right\}\geqslant\frac{\pi^2}{D^2}e^{-\alpha/2}.$$

$$\Longrightarrow \lambda_1\geqslant\frac{\pi^2}{D^2}e^{-\alpha/2},\quad d\geqslant 2$$

$$\Longrightarrow \text{Yau's conjecture}\Longrightarrow \text{Yang \& Jia's estimates.}$$

We have seen that in the past 40 years or so, geometers have made a series of hard efforts to improve the lower bounds step by step. The resulting bounds by different approaches are not comparable. On the other hand, several simple examples were in my mind for which I did not know how to handle by using our approach. Thus, I had a feeling for many years that each approach has its own advantage and there is no best one. I could not imagine, even half a year ago, that we can eventually find out a general formula by using our approach.

3. New Results. Manifolds.

Define $\mathscr{F}=\{f\in C[0,D]: f>0 \text{ on } (0,D)\}$, $C(r)=1$ if $K=0$ and

$$C(r)=\begin{cases}\cos^{d-1}\left[\dfrac{r}{2}\sqrt{\dfrac{K}{d-1}}\right], & \text{if } K>0\\[2ex] \cosh^{d-1}\left[\dfrac{r}{2}\sqrt{\dfrac{-K}{d-1}}\right], & \text{if } K<0.\end{cases}$$

Then, our general formula is given as follows.

Theorem (Chen & Wang (1997)[16]). For Laplacian on M, we have

$$\lambda_1\geqslant 4\sup_{f\in\mathscr{F}}\inf_{r\in(0,D)}f(r)\left[\int_0^r C(s)^{-1}ds\int_s^D C(u)f(u)du\right]^{-1}.$$

Before moving further, let us recall the well-known classical variational formula, the Max-Min formula:[1]

$$\lambda_1=\inf\left\{\mu\big(\|\nabla f\|^2\big)/\mu\big(f^2\big): f\in C^1(M),\mu(f)=0\right\},$$

where μ is the Riemannian measure on M. It is especially useful for the upper bound of λ_1 and is used in almost all of the literature on this topic. But it is much harder to handle the lower bound for which many approaches have been developed in the history but no general formula ever appeared before. Comparing the formula with ours, one sees that there are no common points. To see the power

[1]Historical Note: In [4], it is named the Rayleigh's formula, goes back to Lord S. J. W. Rayleigh (1877). On the other hand, in the book "Inequalities" by E. F. Beckenback & R. Bellman (§25 and §26), it says that the formula goes back to E. Fischer(1905) and generalized by R. Courant(1924) (cf. [17]) and the original Rayleigh's result means $\lambda_0=0$ rather than λ_1. The relation of λ_1 and the L^2-exponential convergence mentioned above was studied in [33] and [5].

of our formula, by setting $f = 1$, one can deduce all the bounds without underlines given in Theorem (1993)(cf. [12]). Of course, it should not be surprising that the new formula can produce a lot of new estimates since the test function f can be quite arbitrary. But it is surprising that the estimates of the first eigenvalue given in [12] and [49] can still be improved, as illustrated by the following corollary.

Corollary (Chen & Wang (1997)[16]).

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

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

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

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

The corollary improves respectively the Zhong & Yang's, the Lichnernowicz's, the Cai's and the Yang & Jia's estimate. For instance, since $D\sqrt{K/(d-1)} \leqslant \pi$ for $K \geqslant 0$ and usually the strict inequality holds, the second one above improves the Lichnerowicz's estimate.

4. New Result. Diffusions in Half Line.

For diffusions in $\mathbb{R}^d$ or for Markov chains, we have a similar story as for geometry but not discussed here. We mention only some general results. Consider the diffusions in half line with operator

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

and reflecting boundary. Define

$$C(x) = \int_0^x \frac{b(y)}{a(y)}dy, \qquad \pi(dx) = \frac{1}{Z}\frac{e^{C(x)}}{a(x)}dx,$$

where Z is a normalizing constant. Set

$$\mathscr{F} = \{f \in L^1(\pi) : \pi(f) \geqslant 0 \text{ and } f'|_{(0,\infty)} > 0\}.$$

Theorem (Chen & Wang (1997)[15]).

$$\lambda_1 \geqslant \sup_{f\in\mathscr{F}}\inf_{x>0}\left[\frac{e^{-C(x)}}{f'(x)}\int_x^\infty \frac{f(u)e^{C(u)}}{a(u)}du\right]^{-1}.$$

Moreover, in the regular case, the equality holds.

5. New Result. Birth-Death Processes.
Recall that

$$\pi_i = \frac{\mu_i}{\mu}, \quad \mu_0 = 1, \quad \mu_i = \frac{b_0 b_1 \cdots b_{i-1}}{a_1 a_2 \cdots a_i}, \quad i \geqslant 1, \quad \mu = \sum_i \mu_i.$$

Let $\mathscr{W} \subset L^1(\pi)$ be the set of all strictly increasing sequences $(w_i : i \geqslant 1)$ with $\sum_{i\geqslant 1} \mu_i w_i > 0$. Define

$$I_i(w) = b_i \mu_i (w_{i+1} - w_i) \Big/ \sum_{j=i+1}^{\infty} \mu_j w_j, \ i \geqslant 1, \quad I_0(w) = b_0 \Big(1 + w_1 \Big/ \sum_{j=1}^{\infty} \mu_j w_j\Big).$$

Theorem (Chen (1996)[10]).

$$\lambda_1 = \sup_{w\in\mathscr{W}} \inf_{i\geqslant 0} I_i(w)$$
$$\lambda_1 = \sup_{(v_i>0)} \inf_{i\geqslant 0} \{a_{i+1} + b_i - a_i/v_{i-1} - b_{i+1} v_i\}.$$

We remark that here the equalities hold. In other words, we have complete dual variational formulas for λ_1. Furthermore, the formulas remain the same if λ_1 is replaced by the exponentially ergodic rate $\hat{\alpha}$ which is a traditional topic in the study of Markov chains [5] or [6]. The first formula above is a summation form which is similar to the integration form used previously. The second one is a differential form. An analog of the last form also works for the cases of manifolds or the diffusions but omitted here.

Part Ⅲ. Trilogy of Couplings.

1. Markovian Couplings.

We now turn to discuss the trilogy of couplings: The Markovian coupling, the optimal Markovian coupling and the construction of distances for couplings. We will also sketch the main proof of our results reported above. Since the story for Markov chains is similar, we concentrate on diffusions. 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.

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

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

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

The last coupling was extended to manifold by W.S. Kendall[1986]. See also M. Cranston[1991]. In the case that $x=y$, the first and the third ones are defined to be the same as the second one.[2] 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

Sketch of the Main Proof.
Here we adopt the analytic language. Given a self-adjoint elliptic operator L, denote by $\{T_t\}_{t\geqslant 0}$ the semigroup determined by L: $T_t=e^{tL}$. Corresponding to $\widetilde{L}$, we have $\{\widetilde{T}_t\}_{t\geqslant 0}$. The coupling simply means that

$$\widetilde{T}_t f(x,y)=T_t f(x) \text{ (resp. } \widetilde{T}_t f(x,y)=T_t f(y)) \tag{1}$$

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 regarded as a bivariate function.
Step 1. Let g be the eigenfunction of $-L$ corresponding to λ_1. We have

$$\frac{d}{dt}T_t g(x)=T_t Lg(x)=-\lambda_1 T_t g(x).$$

Hence

$$T_t g(x)=g(x)e^{-\lambda_1 t}. \tag{2}$$

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

$$\widetilde{T}_t\rho(x,y)\leqslant \rho(x,y)e^{-\alpha t}. \tag{3}$$

This condition is implied by

$$\widetilde{L}\rho(x,y)\leqslant -\alpha\rho(x,y),\quad x\neq y \tag{4}$$

[2]For this reason, the term "basic coupling" was used in [11] for the second coupling. We now use the term "march" rather than "basic" since it is more intrinsic and consistent with the one used for Markov chains by the author years ago.

Setting $g_1(x,y)=g(x)$ and $g_2(x,y)=g(y)$, we obtain

$$\begin{aligned} e^{-\lambda_1 t}|g(x)-g(y)| &= \big|T_t g(x)-T_t g(y)\big| \quad \text{(by (2))} \\ &= \big|\widetilde{T}_t g_1(x,y)-\widetilde{T}_t g_2(x,y)\big| \quad \text{(by (1))} \\ &\leqslant \widetilde{T}_t|g_1-g_2|(x,y) \\ &\leqslant c_g \widetilde{T}_t \rho(x,y) \quad \text{(Lipschitz property)} \\ &\leqslant c_g \rho(x,y) e^{-\alpha t} \quad \text{(by (3))}. \end{aligned}$$

Step 3. Choose $\{(x_n,y_n)\}$ so that

$$\frac{|g(x_n)-g(y_n)|}{\rho(x_n,y_n)} \to c_g.$$

We obtain $\lambda_1 \geqslant \alpha$. $\square$

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 the 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 (4). For this, one needs not only a good coupling but also a good choice of the distance.

2. 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? In what sense of optimality we are talking about?

Definition. 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[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.

Part (2) of the theorem is newly added but it is an analog of the birth-death processes and its proof is similar to that of part (1). Note that in case (2), ρ may not be a distance but the definition of ρ-optimal coupling is still meaningful.

3. 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 (4) 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 the three classes of processes discussed in the paper. Here, we write down the answer for the case of diffusions in half line only.

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

This construction of distances consists of the last part of the trilogy.

Part Ⅳ. Related Results and Problems.

The new results presented in Part Ⅱ are particular ones of [10], [15] and [16]. Actually, in [16], we deal with the operator $\Delta + \nabla V$ for some $V \in C^2(M)$ on manifolds (maybe non-compact) with Neumann boundary or without boundary. In [15], we deal with self-adjoint elliptic operators in $\mathbb{R}^d$. It is not difficult to go to the full line from the half line but in the higher dimensional case one needs more work. Our new formulas are also meaningful for the gradient estimates, Dirichlet eigenvalues, the mixed eigenvalues and much more others. The recent papers, based on or tightly related to the coupling approach, are partially collected in the references.

For a long period, the coupling method has been mainly used for the convergence in the total variation which then deduces the study on success of couplings. The impression in one's mind is often that the coupling method is useful only if it is successful and it can only provide us a rough estimate. However, from what illustrated above, one sees how big change has been made recently. Actually, the study of the spectral gap is only the most recent topic of various applications of the coupling method. One may refer to [6], [7] and [34] for other applications. For instance, the coupling by reflection is a good choice for the present purpose. But when one looks for the order-preserving coupling, the march coupling is clearly better than the previous one. A nice application of a geometric generalization of the march coupling is given in [24] and [56]. Now, what coupling is the best one for the order-preserving? For which, we now have only a partial answer (see [61] for instance). Next, a fundamental problem for the couplings of time-continuous Markov processes is the uniqueness (well-posed) one. For Markov chains (more generally, for jump processes), this problem was solved completely (cf. [6; Theorem 5.16, Theorem 5.17], [7] and [30]). For diffusions, the same conclusion is conjectured to be true but only partial solution is known now. Finally, what is

the optimal coupling for the weak convergence? How to construct "good" couplings for semimartingales? These are only a few of questions we mention here randomly. In conclusion, the theory of couplings is still too young. There is a lot to be done and the subject should have a nice future.

Acknowledgements. Most part of the materials in the paper was or will be talked at several conferences and universities or institutes: The 60th Anniversary Conference of Chinese Mathematical Society (May 1995, Beijing), the 23rd Conference on Stochastic Processes and Their Applications (June 1995, Singapore), the Symposium on Probability Towards the Year 2000 (October, 1995, New York), Stochastic Differential Geometry and Infinite-Dimensional Analysis (April 1996, Hangzhou), Workshop on Interacting Particle Systems and Their Applications (June 1996, Haifa), Cornell University, University of Illinois, Institute of Advanced Mathematics (Hangzhou), Institute of Applied Mathematics (Chinese Academy), University of Biefeled, Bar-Ilan University and Technion-Israel Institute of Technology. The author would like to thank the following mathematicians for their hospitality and financial supports: Prof. Louis H. Y. Chen, Dr. J. H. Lou and Dr. K. P. Choi at Singapore U., Prof. L. Accardi at U. of Roma II, Profs. C. Heyde, K. Sigman and Y. Z. Shao at Columbia U., Profs. R. Durrett, L. Gross and Z. Q. Chen at Cornell U., Prof. D. L. Burkholder at U. of Illinois, Profs. Z. M. Ma and J. A. Yan at Applied Inst. Chin. Acad., Prof. D. Elworthy at Warwick U., Profs. F. Götze and M. Röckner at U. of Bielefeld, Prof. K. J. Hochberg at Bar-Ilan U. and Prof. B. Granovsky at Technion-Israel Inst. of Tech. The author also acknowledges Prof. F. Y. Wang for the fruitful cooperation.

References

[1]. Aldous, D. J. and Brown, M. (1993), *Inequalities for rare events in time-reversible Markov chains*, IMS Lecture Notes-Monograpg Series **22**, Stochastic Inequalities 1–16.

[2]. Bérard, P. H. (1986), *Spectral Geometry: Direct and Inverse Problem*, LNM. vol. 1207, Springer-Verlag.

[3]. 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.

[4]. Chavel, I. (1984), *Eigenvalues in Riemannian Geometry*, Academic Press.

[5]. 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.

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

[7]. 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.

[8]. Chen, M. F.(1994), *Optimal Markovian couplings and applications*, Acta Math. Sin. New Ser.10:3, 260–275.

[9]. Chen, M. F. (1995), *On ergodic region of Schlögl's model*, Dirichlet Forms & Stoch. Proc. Edited by Z. M. Ma, M. Röckner and J. A. Yan, Walter de Gruyter, 1995, 87–102.

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

[11]. Chen, M. F. and Li, S. F. (1989), *Coupling methods for multi-dimensional diffusion processes*, Ann. of Probab. 17:1, 151–177.

[12]. 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.

[13]. Chen, M. F. and Wang, F. Y. (1995), *Estimation of the first eigenvalue of second order elliptic operators*, J. Funct. Anal. 131:2, 345–363.
[14]. Chen, M. F. and Wang, F. Y. (1997), *Estimates of logarithmic Sobolev constant — An improvement of Bakry–Emery criterion*, J. Funct. Anal. 144:2, 287–300.
[15]. Chen, M. F. and Wang, F. Y. (1997), *Estimation of spectral gap for elliptic operators*, Trans. Amer. Math. Soc. 349, 1239-1267.
[16]. Chen, M. F. and Wang, F. Y. (1997), *General formula for lower bound of the first eigenvalue on Riemannian manifolds*, Sci. Sin. 27:1, 34–42 (Chinese Edition); 40:4, 384–394 (English Edition).
[17]. Courant, R. and Hilbert, D. (1953), *Methods of Mathematical Physics*, Interscience Publishers.
[18]. Cranston, M. (1991), *Gradient estimates on manifolds using coupling*, J. Funct. Anal. 99, 110–124.
[19]. Cranston, M. (1992), *A probabilistic approach to gradient estimates*, Canad. Math. Bull. 35, 46–55.
[20]. Deuschel, J.-D. and Stroock, D. W. (1990), *Hypercontractivity and spectral gap of symmetric diffusion with applications to the stochastic Ising models*, J. Funct. Anal. 92, 30–48.
[21]. Diaconis and Stroock (1991), *Geometric bounds for eigenvalues of Markov chains*, Ann. Appl. Prob. 1:1, 36–61.
[22]. Doeblin, W. (1938), *Exposé de la théorie des chaines simples constantes de Markov à un nombre dini d'etats*, Rev. Math. Union Interbalkanique 2, 77–105.
[23]. Fischer, E. (1905), *Über quadratische Formen mit reellen Koeffizienten*, Monatsh. Math. Phys.16, 234–249.
[24]. Hsu, E. P. (1995), *Logarithmic Sobolev inequality on path spaces*, C. R. Acad. Sci. Paris, 320, 1209–1214.
[25]. Iscoe, I. and McDonald, D. (1994), *Asymptotics of exit times for Markov jump processes* (I), Ann. Prob. 22:1, 372–397.
[26]. Jerrum, M. R. and Sinclair, A. J. (1989), *Approximating the permanent*, SIAM J. Comput.18, 1149–1178.
[27]. 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.
[28]. Kendall, W. (1986), *Nonnegative Ricci curvature and the Brownian coupling property*, Stochastics 19, 111–129.
[29]. Kendall, W. S. (1994), *Probability, convex, and harmonic maps II: smoothness via probabilistic gradient inequalities*, J. Funct. Anal.124.
[30]. 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.
[31]. 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.
[32]. Lichnerowicz, A., *Geometrie des Groupes des Transformationes*, Dunod, Paris, 1958.
[33]. Liggett, T. M. (1989), *Exponential L_2 convergence of attractive reversible nearest particle systems*, Ann. Probab., 17, 403-432.
[34]. Lindvall, T. (1992), *Lectures on the Coupling Method*, Wiley, New York.
[35]. Lindvall, T. and Rogers, L. C. G. (1986), *Coupling of multidimensional diffusion processes*, Ann. of Probab. 14:3, 860–872.
[36]. Lu, Y. G. (1994), *An estimate on non-zero eigenvalues of Laplacian in non-linear version*, preprint.
[37]. Lu, Y. G. (1994), *Estimate of the first non-zero eigenvalue of Laplace-de Rahm and the Laplace-Beltrami operators*, preprint.
[38]. Lü, J. S. (1995), *The optimal coupling for single-birth reaction-diffusion processes and its applications* (In Chinese), J. Beijing Normal Univ. 33:1, 10–17.
[39]. Minlos, R. A. and Trisch, A. (1994), *Complete spectral decomposition of the generator for one-dimensional Glauber dynamics* (*In Russian*), Uspekhi Matem. Nauk, 209–210.

[40]. Schoen, R. and Yau, S. T. (1988), *Differential Geometry* (*In Chinese*), Science Press, Beijing, China.
[41]. Schonmann, R. H. and Shlosman, S. B. (1994), *Complete analyticity for 2D Ising completed*, Comm. Math. Phys.170, 453–482.
[42]. Sinclair, A. J. and Jerrum, M. R. (1989), *Approximate counting, uniform generation, and rapidly mixing Markov chains*, Inform. and Comput.82, 93–133.
[43]. Sokal, A. D. and Thomas, L. E.(1988), *Absence of mass gap for a class of stochastic contour models*, J. Statis. Phys.51:5/6, 907–947.
[44]. Stroock, D. W. and Zegarlinski, B. (1992), *The equivalence of the logarithmic Sobolev inequality and the Dobrushin-Shlosman mixing condition*, Comm. Math. Phys. 144:2, 303-323.
[45]. Sullivan, W. G.(1984), *The L^2 spectral gap of certain positive recurrent Markov chains and jump processes*, Z. Wahrs.67, 387–398.
[46]. Wang, F. Y. (1994), *Gradient estimates on* $\mathbf{R}^d$, Canad. Math. Bull. $XX(2)$, 1–11.
[47]. Wang, F. Y. (1994), *Gradient estimates for generalized harmonic function on manifold* (*In Chinese*), Chin. Sci. Bull.39:6, 492–495.
[48]. Wang, F. Y. (1994), *Ergodicity for infinite-dimensional diffusion processes on manifolds*, Sci. Sin. Ser A, 37(2), 137–146.
[49]. Wang, F. Y. (1994), *Application of coupling method to the Neumann eigenvalue problem*, Prob. Th. Rel. Fields 98, 299–306.
[50]. Wang, F. Y. (1994), *Estimate of the first Dirichlet eigenvalue by using the diffusion processes*, Prob. Th. Rel. Fields 101, 363–369.
[51]. Wang, F. Y. (1995), *Uniqueness of Gibbs states and the L^2-convergence of infinite-dimensional reflecting diffusion processes*, Sci. Sin. Ser A, 32:8, 908–917.
[52]. Wang, F. Y. (1994), *On estimates of logarithmic Sobolev constant* (*In Chinese*), J. Beijing Normal Univ. 30:4, 448–452.
[53]. Wang, F. Y. (1996), *Estimates of logarithmic Sobolev constant for finite volume continuous spin systems*, J. Stat. Phys. 84:1/2, 277–293.
[54]. Wang, F. Y. (1994), *A Probabilistic approach to the first Dirichlet eigenvalue on non-compact Riemannian manifold*, Acta Math. Sin. New Ser. 13:1, 116–126.
[55]. Wang, F. Y. (1995), *Spectral gap for diffusion processes on non-compact manifolds*, Chin. Sci. Bull., 40:14, 1145–1149.
[56]. Wang, F. Y. (1996), *Logarithmic Sobolev inequalities for diffusion processes with application to path space*, Chin. J. Appl. Prob. Stat. 12:3, 255–264.
[57]. Wang, F. Y. and Xu, M. P. (1997), *On order-preservation of couplings for multi-dimensional diffusion processes*, Chin. J. of Prob. and Stat. 13:2, 142–148.
[58]. Yang, H. C. (1989), *Estimate of the first eigenvalue of a compact Riemannian manifold with Ricci curvature bounded below by a negative constant* (*In Chinese*), Sci. Sin.(A) 32:7, 698–700.
[59]. Yuan, X. B. (1995), *Gradient estimates and the first mixed eigenvalue*, Master's thesis at Beijing Normal Univ.
[60]. Zhang, Y. H. (1994), *Conservativity of couplings for jump processes* (*In Chinese*), J. Beijing Normal Univ. 30:3, 305–307.
[61]. Zhang, Y. H.(1995), *The construction of order-preserving coupling for one-dimensional Markov chains*, Chin. J. Appl. Prob. Stat. 12:4, 376–382.
[62]. Zhong, J. Q. and Yang, H. C. (1984), *Estimates of the first eigenvalue of a compact Riemannian manifolds*, Sci. Sin. 27:12, 1251–1265.

Department of Mathematics, Beijing Normal University, Beijing 100875, The People's Republic of China.

Chinese Science Bulletin 42(16), 1997, 1321–1326

Coupling, spectral gap and related topics (I)

CHEN Mufa
Department of Mathematics, Beijing Normal University, Beijing 100875, China

Keywords: Diffusions, manifold, Markov chains, spectral gap, couplings.

This is the first one of a series of three papers. They are partially surveys on three aspects: 1) explaining the main ideas of our recent application of the coupling method to the estimation of spectral gap, 2) introducing some more recent progress on the study on some related topics, 3) collecting some open problems for the further study. The technical details are often avoided in order to keep the paper to be readable at the graduate level.

Let us begin with a summery of the papers. There are altogether six parts which are divided into three papers, each of them contains two parts. In part 1, we explain three main steps in our proof for estimating the spectral gap by using the coupling method. In part 2, we explain two key difficulties of the proof and explain how to overcome them. Some ideas are explored here for the first time. In part 3, the above ideas are applied to the Laplacian on Riemannian manifolds. We introduce one formula and four of its corollaries for the lower bound of the spectral gap. The comparison with the known sharp estimates are also discussed. In part 4, we show the application of the above ideas to four typical eigenvalue problems and some new results are reported. In part 5, we discuss six related topics. Some open problems are proposed in Parts 3–5. Certainly, the problems depend on the personal interest, they are either important or interesting enough. Some of them may be rather easy. For these problems, a large number of related references are also collected, but far away from being complete. In order to have a test of the hard mathematics, in the last part, we present a new proof for computing the logarithmic Sobolev constant in the nearly trivial case: 2×2 matrix.

The set of the papers are self-contained but it may be considered as a companion of the survey articles [1] and [2]. The background of the study, an introduction to various couplings and a sketched analytic proof are presented in [1] and [2].

1 Three steps of the proof.

1.1 Choosing a Coupling

Let (b_t) be the standard Brownian motion (abbrev. BM) in $\mathbb{R}^d$ and let (x_t) be the solution to the stochastic differential equation (abbrev. SDE):

$$\mathrm{d}x_t = \sqrt{2}\,\mathrm{d}b_t, \qquad x_0 = x. \tag{1.1}$$

The process corresponds to the operator Δ (half of it corresponds to the BM). Certainly, we can define a process (y_t) in the same way:

$$\mathrm{d}y_t = \sqrt{2}\,\mathrm{d}b_t, \qquad y_0 = y. \tag{1.2}$$

Now, because the processes (x_t) and (y_t) are defined on the same probability space, we obtain a coupling, called the *march coupling* (x_t, y_t)[3]. However, in what follows, we will use another process (y_t) which is defined by

$$\mathrm{d}y_t = \sqrt{2}\, H(x_t, y_t)\mathrm{d}b_t, \qquad y_0 = y, \tag{1.3}$$

where $H(x,y) = I - 2(x-y)(x-y)^*/|x-y|^2$. Note that $H(x,y)$ has no meaning when $x = y$, so the process (y_t) given in (1.3) is meaningful only up to the *coupling time* $T := \inf\{t \geqslant 0 : x_t = y_t\}$. Starting from the time T, we define $y_t = x_t$. We have thus constructed a process (y_t). Clearly, this (y_t) strongly depends on (x_t). Of course, the solutions of Eq.(1.2) and Eq.(1.3) are different, but they do have the same distribution, due to the invariance of orthogonal transform of BM and the fact that $H(x,y)$ is a reflection matrix. The last couple (x_t, y_t) is called the *coupling by reflection*[4],[3].

Intuitively, the construction of (y_t) can be completed in two steps: Let $y \neq x$.

(1) Parallelly transport x_t from x to y along the line (x, y).
(2) Make the mirror reflection of the transported image of x_t in the hyperplane which is perpendicular to the line (x, y) at y.

Then, the mirror image gives us the process (y_t).

For the diffusion (x_t) on manifold M with generator Δ, a process (y_t) can be constructed in a similar way. Roughly speaking, one simply replaces the phrase "the line (x, y)" in the above construction by "the unique shortest geodesic γ between x and y". Certainly, there are some technical details and geometric difficulty (the cutlocus for instance) in the construction[5],[6].

The appearance of the coupling by reflection is a critical step in the development of the coupling theory. For a long period, one knows mainly the classical coupling, it is successful (i.e., $\mathbb{P}[T < \infty] = 1$) for BM in $\mathbb{R}^d$ iff $d = 1$[3]. Thus, one may have an impression that a process having a successful coupling ought to be recurrent. But the coupling by reflection shows that the success can be much weaker than the recurrence since this coupling is successful in any dimension[4],[3]. The key point is that the strong dependence of (y_t) on (x_t) enable us to reduce the higher dimensional case to dimension one.

1.2 Computing the Distance

Throughout the paper, we consider a connected Riemannian manifold M with $\mathrm{Ric}_M \geqslant K$ for some $K \in \mathbb{R}$. In the most cases, we consider here compact M only. Denote by ρ the Riemannian distance on M. For the distance of the coupled process (x_t, y_t), the following formula was proved by Kendall (1986)[5] and Cranston (1991)[6].

$$\mathrm{d}\rho(x_t, y_t) = 2\sqrt{2}\,\mathrm{d}b_t + \bigg[\int_{x_t}^{y_t} \sum_{i=2}^{d} \Big(|\nabla_U W^i|^2 - \langle R(W^i, U)U,\, W^i\rangle\Big)\bigg]\mathrm{d}t - \mathrm{d}L_t, \quad t < T \tag{1.4}$$

where W^i, $i = 2, \cdots, d$ are Jacobi fields along the unique shortest geodesic γ between x_t and y_t, U is the unit tangent vector to γ and the integral in $[\cdots]$

is along γ. (b_t) is a BM in $\mathbb{R}$ and (L_t) is an increasing process with support contained in $\{t \geqslant 0 : (x_t, y_t) \in \mathbf{C}\}$, $\mathbf{C} := \{(x, y) : x \text{ is the cutlocus of } y\}$. When $(x_t, y_t) \in \mathbf{C}$, the coefficient of dt is taken to be 0.

The formula is a finer version of the deterministic situation. The second term on the right-hand side of (1.4) is more or less familiar and comes from the second variation of arclength. The first and the last terms are new in the stochastic case. Since the measure of cutlocus equals zero, the last term is not essential. Next, because the mean of the first term is zero, it will be ignored once we make the expectation as we will see soon in the next step. However, the condition "$t < T$" is critical in order to avoid the singularity at $t = T$. This is the main place for which the present proof is probabilistic.

To estimate $\rho(x_t, y_t)$, we need only to handle the second term on the right-hand side of (1.4). By comparing M with a manifold with constant sectional curvature, Cranston (1991)[6] proved that when $K < 0$ the term is controlled by

$$2\sqrt{-K(d-1)}\,\tanh\left(\frac{\rho_t}{2}\sqrt{\frac{-K}{d-1}}\right), \qquad \rho_t := \rho(x_t, y_t). \tag{1.5}$$

It was then proved by Chen and Wang (1993)[7] that the same conclusion remains true when $K > 0$ and in which case, (1.5) can be rewritten as

$$-2\sqrt{K(d-1)}\,\tan\left(\frac{1}{2}\sqrt{\frac{K}{d-1}}\,\rho_t\right).$$

Set

$$\gamma(r) = 2\sqrt{-K(d-1)}\,\tanh\left(\frac{1}{2}\sqrt{\frac{-K}{d-1}}\,r\right).$$

Then, we obtain

$$\mathrm{d}\rho_t \leqslant 2\sqrt{2}\mathrm{d}b_t + \gamma(\rho_t)\mathrm{d}t - \mathrm{d}L_t \leqslant 2\sqrt{2}\mathrm{d}b_t + \gamma(\rho_t)\mathrm{d}t, \qquad t < T. \tag{1.6}$$

Equivalently, $\rho_{t\wedge T} - \rho_0 \leqslant 2\sqrt{2}\int_0^{t\wedge T}\mathrm{d}b_s + \int_0^{t\wedge T}\gamma(\rho_s)\mathrm{d}s$. Making expectation, we get

$$\widetilde{\mathbb{E}}^{x,y}\rho_{t\wedge T} \leqslant \rho_0 + \widetilde{\mathbb{E}}^{x,y}\int_0^{t\wedge T}\gamma(\rho_s)\mathrm{d}s. \tag{1.7}$$

In order to get an exponential rate, we need the condition

$$\gamma(r) \leqslant -\alpha\, r \qquad \text{for some } \alpha > 0. \tag{1.8}$$

When $K > 0$, since $\tan\theta \geqslant \theta$ on $[0, \pi/2]$, we have $\alpha = K$. Under (1.8), we have

$$\begin{aligned}
\widetilde{\mathbb{E}}^{x,y}\int_0^{t\wedge T}\gamma(\rho_s)\mathrm{d}s &\leqslant -\alpha\widetilde{\mathbb{E}}^{x,y}\int_0^{t\wedge T}\rho_s\mathrm{d}s \\
&= -\alpha\widetilde{\mathbb{E}}^{x,y}\int_0^{t}\rho_{s\wedge T}\mathrm{d}s \\
&= -\alpha\int_0^{t}\widetilde{\mathbb{E}}^{x,y}\rho_{s\wedge T}\mathrm{d}s,
\end{aligned}$$

since $\rho_{t\wedge T}=0$ for all $t\geqslant T$. Combining this with (1.7), we obtain $\widetilde{\mathbb{E}}^{x,y}\rho_{t\wedge T}\leqslant \rho_0 e^{-\alpha t}$. Equivalently,

$$\widetilde{\mathbb{E}}^{x,y}\rho_t\leqslant \rho_0 e^{-\alpha t},\qquad t\geqslant 0. \tag{1.9}$$

This is the key estimate of our method.

1.3 Estimating λ_1

Let g be an eigenfunction of λ_1: $-\Delta g=\lambda_1 g$, $g\neq$const. Then $\mathbb{E}^x g(x_t)=g(x)e^{-\lambda_1 t}$ for all $t\geqslant 0$. This gives us a relation between the λ_1, g and the process (x_t). The same relation holds for (y_t). Note that the coupling property gives us $\widetilde{\mathbb{E}}^{x,y}g(x_t)=\mathbb{E}^x g(x_t)$. By (1.9), we have

$$\begin{aligned}
e^{-\lambda_1 t}|g(x)-g(y)| &= |\mathbb{E}^x g(x_t)-\mathbb{E}^y g(y_t)|\\
&= \big|\widetilde{\mathbb{E}}^{x,y}\big[g(x_t)-g(y_t)\big]\big|\\
&\leqslant L(g)\widetilde{\mathbb{E}}^{x,y}\rho_t\\
&\leqslant L(g)\rho_0 e^{-\alpha t}\\
&= L(g)\rho(x,y)e^{-\alpha t},\qquad t\geqslant 0,
\end{aligned}$$

where $L(g)$ is the Lipschitz constant of g with respect to the distance ρ. Choosing a sequence $(x^{(n)},y^{(n)})$ so that $g(x^{(n)},y^{(n)})/\,\rho(x^{(n)},y^{(n)})\to L(g)$, the last inequality gives us immediately $\lambda_1\geqslant\alpha$ and hence our proof is completed.

The last step is rather simple but may not be so easy to find out. This is indeed a character of various applications of coupling method, once the idea is understood, the proof often becomes quite straightforward.

2 Two Difficulties.

Roughly speaking, we have explained half of the first version of the paper by Chen & Wang (1993)[7]. The problem is that the above arguments are still not enough to obtain the sharp estimate. For instance, when $K>0$, we get the lower bound $\alpha=K$ only as mentioned right after (1.8). The best we can get when $K>0$ is $8/D^2$ rather than the sharp one π^2/D^2, where D is the diameter of the compact manifold M. Even for the bound $8/D^2$, we still need to estimate $\widetilde{\mathbb{E}}^{x,y}T$ for which we are not going to discuss here.

We now return to analyze the proof discussed in the previous part. In the last step, we need the Lipschitz property of g. Since the non-compact case can often be reduced to the compact one[8] and in the latter case, g is smooth and hence the Lipschitz property is automatic. Thus, in the whole proof, the key is the estimate (1.9), for which we require not only a good coupling but also a good distance. This is not surprising since the convergence rate is not a topological concept, it certainly depends heavily on the choice of the distance. There is no reason why the underlying Riemannian distance should be always a correct choice.

2.1 Optimal Markovian Coupling

The first question is how about the coupling used above. Is there an optimal choice? This problem is quite hard and it was actually studied twice before[3],[9] but unsolved. However, the aim for the optimality becomes clear now. That is choosing coupling to make the rate α as bigger as possible, or in a slightly wider

sense, to make $\widetilde{\mathbb{E}}^{x,y}\rho(x_t,y_t)$ as smaller as possible for all $t\geqslant 0$ and for every fixed pair (x,y) and fixed ρ. Because we are dealing with Markovian coupling, we can use the language of coupling operators[3]. Of course, one can translate the discussions here in SDE. Note that under mild assumption, the last statement is equivalent to that $\widetilde{L}\rho(x,y)$ is as smaller as possible for every pair (x,y), $x\neq y$ [10]. This leads to the definition of *ρ-optimal coupling operator* $\overline{L}$:

$$\overline{L}\rho(x,y)=\inf_{\widetilde{L}}\widetilde{L}\rho(x,y),\qquad x\neq y$$

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

Theorem 1[10]**.** Consider the BM in $\mathbb{R}^d$. Then, the coupling by reflection is ρ-optimal for every ρ having the form $\rho(x,y)=f(|x-y|)$, where $f\in C^2[0,\infty)$, $f(0)=0$, $f'>0$ on $(0,\infty)$ and $f''\leqslant 0$.

The role of $f(|x-y|)$ is reducing the higher-dimensional case to dimension one. In order the ρ defined above be a distance, the first two conditions of f are necessary and the third condition guarantees the triangle inequality.

The above theorem overcomes our first difficulty. That is a classification of couplings. The story of Markovian coupling and the optimal Markovian couplings was talked in [1] and [2] and hence is not repeated here. For more recent progress on optimal couplings, refer to [8], [10]–[17]. For other recent progress on the coupling theory, refer to [18]–[23].

The above result tells us that the coupling by reflection is already good enough even for the BM on manifolds. Furthermore, it suggests us to use $f\circ\rho$ instead of the original Riemannian distance ρ. The construction of new distance is the second main difficulty of the study and this consists of the context of the remainder of this part.

2.2 Modification of Riemannian Distance

To illustrate the use of the above idea, assume that $K\geqslant 0$ and take $\bar{\rho}=\sin\frac{\pi\rho}{2D}$. Since $\pi\leqslant D$, $\bar{\rho}$ is a distance. To computer $\mathrm{d}\bar{\rho}_t$, apply the Itô's formula plus a comparison argument,

$$\mathrm{d}\bar{\rho}_t\leqslant\frac{\pi}{2D}\cos\frac{\pi\rho_t}{2D}\cdot 2\sqrt{2}\mathrm{d}b_t-\frac{1}{2}\cdot\frac{\pi^2}{4D^2}\cdot\sin\frac{\pi\rho_t}{2D}\cdot 8\mathrm{d}t,\qquad t<T.$$

The first term is a martingale, denoted by M_t. We then obtain

$$\mathrm{d}\bar{\rho}_t\leqslant dM_t-\frac{\pi^2}{D^2}\bar{\rho}_t\mathrm{d}t$$

for all $t<T$. Repeating the proof given in the last part, we get

$$\widetilde{\mathbb{E}}^{x,y}\bar{\rho}_t\leqslant\bar{\rho}_0\exp\left[-\frac{\pi^2}{D^2}t\right].$$

Thus, we obtain luckily $\lambda_1\geqslant\pi^2/D^2$ which is optimal in the case of zero curvature. By using the same function sin with a slight modifications (which come from some

controlling equations of (1.6) with constant coefficients), we can obtain the other two optimal lower bounds, as shown in the final version of [7; Theorem 1.6]. Finally, it is interesting to remark that $2\theta/\pi \leqslant \sin\theta \leqslant \theta$ on $[0, \pi/2]$ and so the distances $\bar{\rho}$ and ρ used above are actually equivalent. However, the resulting rates are essentially different.

2.3 Redesignated Distances

Is there any other choice of the distance? The question is again easy to state but not so easy to think. Indeed, we did not know for a long time where we can start from. This problem becomes more serious when one goes to the non-compact situation. Intuitively, those distance can not be good if with respect to it the eigenfunction g is too far away from being Lipschitz. As usual, we are taught by simple examples. Consider the diffusion on the half-line $[0, \infty)$ with operator $L = a\,\mathrm{d}^2/\mathrm{d}x^2 - b\,\mathrm{d}/\mathrm{d}x$ for some constants a, $b > 0$. If one adopts the Euclidean distance, then it gives nothing. So what distance should we take? Our goal is to look at the eigenfunction of $\lambda_1 = b^2/4$ (setting $a = 1$ without loss of generality):

$$g(x) = (1 - bx/2)\exp[bx/2] \in L^1(\pi) \setminus L^2(\pi).$$

This suggests us to construct a new distance ρ from the leading part of g: $\rho(x, y) = |\exp[\gamma x] - \exp[\gamma y]|$ for suitable $\gamma > 0$. Surprisingly, it gives us the exact estimate of λ_1 even though the eigenfunction g is still not Lipschitz with respect to this distance[8]. Furthermore, once g being strictly monotone (it is indeed the case of dimension one but the proof is rather technical[24]), we can always take $|g(x)-g(y)|$ as the distance we required. This provides us a way to construct and to classify the distances according to different classes of elementary function g[11],[24].

However, there is still a serious difficulty in the construction of the new distance since the eigenvalue λ_1 and its eigenfunctions g are either known or unknown simultaneously. To see this, consider another example on the half-line with operator $L = a(x)\mathrm{d}^2/\mathrm{d}x^2$. A beautiful estimate due to Kac and Krein (1958)[25] and Kotani (1982)[26] says that

$$\frac{1}{4}\left(\sup_{x>0} x\int_x^\infty \frac{du}{a(u)}\right)^{-1} \leqslant \lambda_1 \leqslant \left(\sup_{x>0} x\int_x^\infty \frac{du}{a(u)}\right)^{-1}.$$

Now, in order to recover this estimate by using our method, according to what discussed above, we have to know some information about the eigenfunction g. Even in such a simple situation, it is still no hope to solve g from $a(x)$ explicitly. What can we do now? Once again, we examine the eigen-equation:

$$\begin{aligned} a(x)g'' = -\lambda_1 g &\Longleftrightarrow g'(s) = \int_s^\infty \frac{\lambda_1 g(u)}{a(u)}du \qquad (\text{since } g'(0) = 0) \\ &\Longleftrightarrow g(x) = g(0) + \int_0^x \mathrm{d}s \int_s^\infty \frac{\lambda_1 g(u)}{a(u)}du. \end{aligned} \tag{2.1}$$

What we have done is just rewriting the differential equation into the corresponding integration equation. Is the last equation helpful? The answer is affirmative. We now move step by step as follows.

(1) Regard $\lambda_1 g$ as a new function f.

(2) Regard the right-hand side of (2.1) as an approximation of the left-hand side g.
(3) Ignore the constant $g(0)$ on the right-hand side since we are interested only in $g(x) - g(y)$.

In other words, these considerations suggest us to take

$$\tilde{g}(x) = \int_0^x \mathrm{d}s \int_s^\infty \frac{f(u)}{a(u)} du \tag{2.2}$$

as an approximation of g (up to a constant) and then take $\rho(x, y) = |\tilde{g}(x) - \tilde{g}(y)|$. The function f used above is called a *test function*. A slight different explanation of the construction goes as follows. Even though the equation (2.1) can not be solved explicitly, but as usual we do have a successive approximation procedure. Thus, one may regard (2.2) as the first step of the approximation and go further step by step. However, the further approximations are not really useful since it becomes on the one hand too complicated and on the other hand it is not as effective as modifying the test function f directly.

Next, we consider the general operator on the half-line: $L = a(x)\mathrm{d}^2/\mathrm{d}x^2 + b(x)\mathrm{d}/\mathrm{d}x$. By standard ODE, it can be reduced to the above simple case. The approximation function now becomes[24]

$$g(r) = \int_0^r e^{-C(s)} \mathrm{d}s \int_s^\infty \frac{f(u)e^{C(u)}}{a(u)} du, \qquad C(r) := \int_0^r \frac{b}{a}. \tag{2.3}$$

We have thus obtained a general construction of the mimic eigenfunctions and furthermore of the required distances. It should be not surprised that the reconstruction of the distances is a powerful tool in many situations. This will be illustrated in the subsequent parts.

2.4 Optimizing the Distances

Before moving further, let us mention that an optimizing method of the distance induced from (2.3) as well as some comparison methods is developed in Chen & Wang (1995)[24]. In short word, the condition "$\widetilde{L}\rho(x, y) \leqslant -\alpha\rho(x, y)$ holds for all large enough $\rho(x, y)$" but not necessarily "for all $x \neq y$" (the latter condition is equivalent to (1.9)) is enough to guarantee a positive lower bound of λ_1.

(*Received March* 29, 1997)

References

1. Chen, M. F., Optimal Markovian couplings and application to Riemannian geometry, in *Prob. Theory and Math. Stat.* (Eds. Grigelionis B et al.), VPS/TEV, 1994, 121..
2. Chen, M. F., Trilogy of couplings and general formulas for lower bound of spectral gap, in *"Proceedings of the Symposium on Probability Towards the Year 2000"*(Eds. Accardi, L. Heyde, C.) 1996.
3. Chen, M. F., Li, S. F., Coupling methods for multi-dimensional diffusion processes, *Ann. of Probab.*, 1989, 17(1): 151.
4. Lindvall, T., Rogers, L. C. G., Coupling of multidimensional diffusion processes, *Ann. of Prob.*, 1986, 14(3): 860.

5. Kendall, W., Nonnegative Ricci curvature and the Brownian coupling property, *Stochastics*, 1986, 19: 111.
6. Cranston, M., Gradient estimates on manifolds using coupling, *J. Funct. Anal.*, 1991, 99: 110.
7. Chen, M. F., Wang, F. Y., Application of coupling method to the first eigenvalue on manifold, *Sci. Sin.*, Ser. A, 1993, 23(11)(Chinese Edition): 1130; 1994, 37(1)(English Edition): 1.
8. Chen, M. F., Wang, F. Y., Estimation of the first eigenvalue of second order elliptic operators, *J. Funct. Anal.*, 1995, 131(2): 345–363.
9. Chen, M. F., Ergodic theorems for reaction-diffusion processes, *J. Stat. Phys.*, 1990, 58 (5/6): 939.
10. Chen, M. F., Optimal Markovian couplings and applications, *Acta Math. Sin. New Ser.*, 1994, 10(3): 260.
11. Chen, M. F., Estimation of spectral gap for Markov chains, *Acta Math. Sin. New Ser.*, 1996, 12(4): 337.
12. Wang, F. Y., Xu, M. P., On order-preservation of couplings for multi-dimensional diffusion processes, *Chin. J. Appl. Prob. Stat.*, 1996.
13. Zhang, S. Y., Measurable coupling of transition probability and probability distance, *Chin. Ann. Math.* (In Chinese), 1995, 16(A:6): 769.
14. Zhang, S. Y., Xu, K., On the existence of the optimal measurable coupling of transition probability, *Acta Math. Sin.* (In Chinese), 1997, 40(1): 5.
15. Zhang, Y. H., Conservativity of couplings for jump processes, *J. Beijing Normal Univ.* (In Chinese), 1994, 30(3): 305.
16. Zhang, Y. H., The construction of order-preserving coupling for one-dimensional Markov chains, *Chin. J. Appl. Prob. Stat.* (In Chinese), 1996, 12(4): 376.
17. Zhang, Y. H., Optimal coupling for a class of distances (In Chinese), *J. Beijing Normal Univ.* (In Chinese), 1996, 32(4): 463.
18. Aldous, D. J., Thorisson, H., Shift-coupling, *Stoch. Proc. Appl.*, 1993, 44: 1.
19. Thorisson, H., Shift-coupling in continuous time, *Prob. Th. Rel. Fields*, 1994, 99: 477.
20. Roberts, G. O., Rosenthal, J. S., Quantitative bounds for convergence rates of continuous time Markov processes, *Electr. J. Prob.*, 1996, 1: 1.
21. Kersting, G., Harmonic coordinates for diffusions in the plane, *Ann. of Prob.*, 1996, 24(3): 1239.
22. Arous, G. B., Cranston, M., Kendall, W. S., Coupling constructions for hypoelliptic diffusions: two examples, *Proc. Symp. Pure Math.*, 1995, 57: 193.
23. Last, G., Coupling with compensators, *Stoch. Proc. Appl.*, 1996, 65: 147.
24. Chen, M. F., Wang, F. Y., Estimation of spectral gap for elliptic operators, *Trans. Amer. Math. Soc.*, 1997, 349: 1209.
25. Kac, I. S., Krein, M. G., Criteria for discreteness of the spectrum of a singular string (In Russian), *Izv. Vyss. Učebn. Zaved. Mat.*, 1958, 2: 136.
26. Kotani, S., Watanabe, S., Krein's spectral theory of strings and generalized diffusion processes, *Lecture Notes in Math.*, 1983, 923: 235.

Acknowledgement Research supported in part by National Natural Science Foundation of China (Grant No. 19631060), Qiu Shi Science and Technology Foundation and the State Education Commission of China.

Chin. Sci. Bulletin, 1997, 42:17, 1409–1416.

Coupling, spectral gap and related topics (Ⅱ)

CHEN Mufa

Department of Mathematics, Beijing Normal University, Beijing 100875, China

Keywords: Diffusions, manifold, Markov chains, spectral gap, couplings.

This is the second one of a series of three papers. The ideas introduced in the last paper are used to study the estimate of spectral gap and four classes of typical eigenvalue problems on manifolds. The comparison with the known optimal estimates are given, some new progress is reported and some open problems are proposed.

3 One formula and four corollaries.

Up to now, we have discussed only the construction of the mimic eigenfunctions g in the case of half-line. But how to go to the whole line and further to $\mathbb{R}^d$ and manifold M? This seems quite difficult. However, the answer is still rather simple once the idea was figured out. As we have seen from Part 1 of the first paper, the coupling method reduces the higher-dimensional case to computing the distance of the coupled process, and then the distance itself consists of a process valued in the half-line $[0,\infty)$. We have thus returned to what treated in the last part.

Recall that

$$\gamma(r) = 2\sqrt{-K(d-1)}\,\tanh\left(\frac{1}{2}\sqrt{\frac{-K}{d-1}}\,r\right)$$

and $\rho_t = \rho(x_t, y_t)$. It is known from (1.6) in the first paper that

$$\mathrm{d}\rho_t \leqslant 2\sqrt{2}\mathrm{d}b_t + \gamma(\rho_t)\mathrm{d}t, \qquad t < T. \tag{3.1}$$

The operator corresponding to (3.1) with equality is $L = 4\mathrm{d}^2/\mathrm{d}x^2 + \gamma(x)\mathrm{d}/\mathrm{d}x$ on $[0, D]$ with absorbing boundary at 0 and reflecting boundary at D. This is indeed simpler than what we discussed in the last part ($a(x) \equiv 4$). Redefine $C(r) = \exp[\frac{1}{4}\int_0^r \gamma(s)\mathrm{d}s]$. Then the approximation function defined by (2.3) becomes

$$g(r) = \int_0^r C(s)^{-1}\mathrm{d}s \int_s^D C(u)f(u)\mathrm{d}u,$$

up to a constant factor. Now the same proof as given in Part 1 of the first paper implies rather easily the following result.

Theorem 2 (General formula)[1].

$$\lambda_1 \geqslant \sup_{f\in\mathscr{F}} \inf_{r\in(0,D)} \frac{4f(r)}{\int_0^r C(s)^{-1}\mathrm{d}s \int_s^D C(u)f(u)\mathrm{d}u}, \tag{3.2}$$

where $\mathscr{F} = \{f \in C[0,D] : f > 0 \text{ on } (0,D)\}$.

Even for the simplest function $f \equiv 1$, (3.2) already provides us a non-trivial lower bound which was obtained in [I; 7] (i.e., ref. [7] in the first paper of the series) by a different proof. Next, set $\beta = \pi/(2D)$ and $\alpha = 2^{-1}\sqrt{|K|/(d-1)}$. Applying the formula to the elementary test functions $\sin(\beta r)$, $\sin(\alpha r)$, $\sin(\beta r)$ and $\cosh^{1-d}(\alpha r)\sin(\beta r)$ successively, we obtain the following corollaries.

Corollary 3[1].

$$\lambda_1 \geqslant \frac{\pi^2}{D^2} + \max\Big\{\frac{\pi}{4d}, 1 - \frac{2}{\pi}\Big\}K, \qquad K \geqslant 0 \tag{3.3}$$

$$\lambda_1 \geqslant \frac{dK}{d-1}\big\{1 - \cos^d(\alpha D)\big\}^{-1}, \qquad d > 1, \quad K \geqslant 0 \tag{3.4}$$

$$\lambda_1 \geqslant \frac{\pi^2}{D^2} + \Big(\frac{\pi}{2} - 1\Big)K, \qquad K \leqslant 0 \tag{3.5}$$

$$\lambda_1 \geqslant \frac{\pi^2}{D^2}\sqrt{1 - 2D^2K/\pi^4}\,\cosh^{1-d}(\alpha D), \qquad d > 1, \quad K \leqslant 0. \tag{3.6}$$

As was mentioned by Chen and Wang (1995)[1] (3.3) improves Zhong-Yang's estimate[2]: $\lambda_1 \geqslant \pi^2/D^2$ $(K \geqslant 0)$. (3.4) improves Lichnerowicz's estimate[3]: $\lambda_1 \geqslant dK/(d-1)$ $(K > 0)$. (3.5) improves Cai's estimate[4]: $\lambda_1 \geqslant \pi^2/D^2 + K$ $(K \leqslant 0)$, while (3.6) improves Yang-Jia's estimate[5],[6]: $\lambda_1 \geqslant \frac{\pi^2}{D^2}\exp[-\alpha D]$ $(K \leqslant 0)$.

We now make two additions.

1) The Lichnerowicz's estimate was partially improved by Bérard, Besson and Gallot (1985)[7] as follows:

$$\lambda_1 \geqslant d\left\{\frac{\int_0^{\pi/2}\cos^{d-1}t\mathrm{d}t}{\int_0^{D/2}\cos^{d-1}t\mathrm{d}t}\right\}^{2/d}, \qquad K = d-1 > 0.$$

In this case, in (3.4), our corollary says that

$$\lambda_1 \geqslant \frac{dK}{d-1}\big\{1 - \cos^d(\alpha D)\big\}^{-1} = \frac{d}{1 - \cos^d(D/2)}.$$

The comparison of these two estimates goes as follows:

$$\frac{d}{1 - \cos^d(D/2)} \geqslant d\,\frac{\int_0^{\pi/2}\cos^{d-1}t\mathrm{d}t}{\int_0^{D/2}\cos^{d-1}t\mathrm{d}t} \geqslant d\left\{\frac{\int_0^{\pi/2}\cos^{d-1}t\mathrm{d}t}{\int_0^{D/2}\cos^{d-1}t\mathrm{d}t}\right\}^{2/d}.$$

The main idea used in the last quoted paper is the isoperimetric inequality plus the Cheeger's inequality (1970)[8]. This is one of the main two tools appearing in the 1980's, the other one is the Li-Yau's gradient estimate method[9]. Since then, both methods have a great number of applications including the discrete situation[10]–[17].

2) The general result obtained in [6] is as follows:

$$\lambda_1 \geqslant \frac{\pi^2}{16(1+c_0)D^2}\,\frac{(d-1)x}{\big(\exp\big[\sqrt{(d-1)x}\,/4\big] - 1\big)^2}, \qquad d \geqslant 3$$

where $x = -KD^2 \geqslant 0$ and $c_0 \in [0, 1)$ depends on the bound of the eigenfunction. Next, (3.6) implies that

$$\lambda_1 \geqslant \frac{\pi^2}{D^2} \cosh^{1-d}\left(\frac{1}{2}\sqrt{\frac{x}{d-1}}\right), \qquad d \geqslant 2.$$

Then, the latter bound is greater than or equal to the former one for all $d \geqslant 2$, even though c_0 is replaced by 0.

Note that the coefficients of the linear terms given in (3.3) and (3.5) are the following $1 - \frac{2}{\pi} \approx 0.36 > \frac{1}{3}$, $\frac{\pi}{2} - 1 \approx 0.57 > \frac{1}{2}$. Due to the error produced from the use of FKG-inequality, these coefficients are not sharp. We conjecture that the first coefficient belongs to $(1 - \pi/2, 3/5)$ and the second one to $(2/5, \pi/2 - 1)$. The next problem may not be of great importance but it has its own interest.

Problem 1. Determine the precise value of these coefficients.

For this problem, since we are looking for the lower bound independent of d, and moreover, function $\gamma(r)$ defined before (3.1) is increasing in d to $-Kr$, one needs only to study (3.1) replacing $\gamma(\rho_t)$ with $-K\rho_t$.

Finally, from the author's knowledge, all the known optimal estimates concerning the geometric quantities d, D and K only, are improved by one of (3.3)–(3.6), except the one given by Chen (1994) (see ref.[10] in part (I)): $\lambda_1 \geqslant \frac{1}{4}K(d-1)\tanh^2(\alpha D)\operatorname{sech}^2\theta$ (the factor $\tanh^2(\alpha D)$ was missed in the paper). However, the last estimate is still covered by the general formula and it may be improved by using some test function, but the work seems rather involved, and hence we did not do it.

4 Four eigenvalue problems.

It is well known that there are mainly four classes of eigenvalue problems[13]:

(i) The closed eigenvalue,
($\ddot{i}$) the Neumann eigenvalue,
($\dddot{i}$) the Dirichlet eigenvalue, and
(iv) the mixed eigenvalue.

For the first two situations, we have $\lambda_0 = 0 < \lambda_1 \leqslant \lambda_2 \leqslant \cdots$ (in compact case for instance), that is, we have a trivial eigenvalue $\lambda_0 = 0$. For the latter two situations, we do not have the trivial one, and then we still denote by λ_1 the first eigenvalue. Thus, the former two and the latter ones are essentially different. The proof discussed above works for the first two cases, even for more general operators $L = \Delta + \nabla V$ for some $V \in C^2(M)$[1,19]. Clearly, what we have done is the estimate of the spectral gap $\lambda_1 - \lambda_0 = \lambda_1$. Thus, in the last two cases, we can also ask the same question about the spectral gap $\lambda_2 - \lambda_1$.

In the Dirichlet case, the spectral gap $\lambda_2 - \lambda_1$ of Laplacian Δ coincides with λ_1 of the Neumann eigenvalue of operator $L = \Delta + 2\nabla \log u_1$, where u_1 is the eigenfunction of the Dirichlet eigenvalue λ_1 for Δ[18]. Therefore, the study of the Dirichlet spectral gap $\lambda_2 - \lambda_1$ of Laplacian Δ can be essentially reduced to what treated above, and hence we also have a general formula for the lower bound of $\lambda_2 - \lambda_1$. Of course, since the function u_1 is not explicitly known, more efforts are needed in order to obtain some explicit lower bound of $\lambda_2 - \lambda_1$. See Wang (1996)[1]

[1] Wang, F. Y., Estimates of the gap between the first two Dirichlet eigenvalues, 1996, preprint.

for details. Nevertheless, due to the advantage of our new approach, much more new results are deduced for this topic.

We would like to make some remarks here for the Dirichlet eigenvalue (called D-problem for short). Similarly, we have N-problem. It is interesting to note that in history the most of the papers in this field are devoted to the D-problem rather than the N-problem. The main reason is that the D-problem is equivalent to the maximum principle (see ref. [20] and the references within). Let $B(p,n)$ be the ball centered at p with radius n. It is well known (go back to Barta (1937)[20]) that

$$\lambda_1 \geqslant \sup_f \inf_{B(p,n)} (-Lf)/f,$$

where f varies over all $C^2(B(p,n))$ functions with $f|_{\partial B(p,n)} = 0$ and $f > 0$ on $B(p,n)$. In other words, we do have a variational formula for the lower bound for the D-problem. Note that the maximum principle is a powerful tool in PDE. It should not be surprised that one can do a lot for the D-problem. However, this formula does not work for the N-problem (or the closed eigenvalue problem). The reason is simply that the eigenfunction g in the Neumann case must cross zero and so is Lg (because the mean of g equals zero). Hence, there is a singularity of $(-Lg)/g$ around the point which makes serious difficulty when the eigenfunction g is replaced by its perturbation f. Traditionally, one transfers the N-problem to the D-problem. This explains the reason why one often thinks that the N-problem is more difficult than the D-problem. It seems that the N-problem is also more difficult than the closed problem. For instance, for the Neumann eigenvalue λ_1 with convex boundary, the best known lower bound is the Lichnerowicz's estimate obtained by Escobar (1990)[21] in the case of $K > 0$, and up to now we have not seen from literature a proof about "$\lambda_1 \geqslant \pi^2/D^2$ for general $K \geqslant 0$. The known estimates of λ_1 for the N-problem in the case of $K < 0$ are all less than the known estimates for the closed eigenvalue[9,15,17,22]. However, as we mentioned above, Theorem 2 and Corollary 3 are all suitable for the Neumann eigenvalue λ_1 with convex boundary. These discussions also show that the use of coupling enables us to avoid the singularity, just as mentioned above. The degeneracy of the coupled process appears at time T only, and before time T, the process is quite regular. This is somehow similar to the D-problem for which the degeneracy appears at the boundary only. In other words, the coupling method plays a substitute role in our proof as the maximum principle played for the D-problem.

For the D-problem, we do not need coupling. Instead, one considers the exit time $\tau_{B(p,n)} := \inf\{t \geqslant 0 : x_t \neq B(p,n)\}$ of the BM[23],[24]. Nevertheless, our approximation of the eigenfunction is still helpful here. For instance, Wang (1996)[2] proves the following result.

Theorem 4. Suppose that p is a pole. Let $\gamma \in C[0,n)$ such that $L\rho(p,x) \geqslant \gamma(\rho(p,x))$ for all $x \neq p$ and set $C(x) = \exp\big[\int_0^x \gamma(u)\mathrm{d}u\big]$. Then

$$\lambda_1(B(p,n)) \geqslant \sup_{f\in C[0,n]} \inf_{r\in[0,n]} \frac{f(r)}{\int_r^n C(s)^{-1}\mathrm{d}s \int_0^s f(u)C(u)\mathrm{d}u}.$$

[2]Wang, F. Y., Positivity of the principle eigenvalue on Riemannian manifolds, 1996, preprint

Certainly, for Schrödinger operator, the problem is similar to the D-problem.

Problem 2. Study the lower bound of λ_1 and $\lambda_2 - \lambda_1$ for the mixed eigenvalue problem.

As an example, as mentioned by Chen and Wang (1995)[1], (3.2) is also the lower bound of λ_1 for a diffusion on $[0, D]$ with Dirichlet boundary at the left-end point and with Neumann boundary at the right-end point.

It is the position to mention the following open problem.

Problem 3. Prove the general formula (3.2) by using geometric-analysis.

This is a valuable work. Once such a proof could exist, one would adopt more geometric tools, avoid some restriction of the probabilistic limitation, and go to more general situation.

Next, can the formula still be improved? For this, we reexamine again the proof given in Part 1 of the first paper: recall our key condition (1.9):

$$\widetilde{\mathbb{E}}^{x,y} f \circ \rho(x_t, y_t) \leqslant f \circ \rho(x, y) \exp[-\lambda_1 t]$$

for all $x \neq y$. This is equivalent to that

$$\widetilde{L} f \circ \rho(x, y) \leqslant -\lambda_1 f \circ \rho(x, y)$$

for all $x \neq y$. In other words,

1) $f \circ \rho$ is super-harmonic of $\widetilde{L} + \lambda_1$.

Before moving further, let us make some remarks on condition 1). Of course, one may express this condition in terms of SDE. Note that the eigenvalue λ_1 of L must be an eigenvalue of the coupling operator $\widetilde{L}$. We often take $f \circ \rho\,(x, y)$ to be $|g(x) - g(y)|$ (in dimension one for instance) for an eigenfuction g of λ_1 of L but this is not necessary, it is stronger than condition 1). Moreover, it is also not completely necessary that $f \circ \rho$ is a distance even though it is in all of our practice.

Next, in the second step of the proof, for computing $\rho(x_t, y_t)$, the original manifold is compared with the following:

2) M has constant sectional curvature.

If one of the above conditions does not hold, then our formula may not be sharp. A particular example is $M = SO(n)$. Refer to Wang (1996)[25]. Thus, there are some possibilities to improve the formula and there indeed may be several different formulas if one makes some restriction on the manifolds or uses more geometric quantities.

Problem 4. Can one extend the formula by including the volume of M[26,27].

Problem 5. Can one relax the condition "$\mathrm{Ric}_M \geqslant K$" by "$\mathrm{Ric}_M(x) \geqslant K(\rho(x))$", where $\rho(x) = \rho(p, x)$? [28,29].

Recall that in higher-dimensional case, there are usually a lot of symmetries.

Problem 6. How to describe and represent the geometric symmetry in the formula?

We will return the last topic in the next part (Theorem 5). It is worthy to work on some restrictive Riemannian manifolds. For instance, the well-known

Selberg's conjecture (1965) "$\lambda_1 \geqslant 1/4$" is in dimension two. A very important particular case is the complex manifolds which have more topological structure and are more close to physics.

Problem 7. Study the estimate of λ_1 for complex manifolds [30]–[34].

It is believed that the coupling method should be useful for this and the next five problems.

Problem 8. Study the estimate of λ_1 for algebraic varieties or sub-manifolds[35,[36].

It is clear that we now have a chance to reexamine the spectral theory and so one may ask many questions. For instance, the non-linear PDEs are very popular now. Our idea works for the following non-linear case: $\Delta f = -\lambda_1 F(f)$, where F is a Lipschitz function (see Lu (1993)[37]). In view of refs. [38] and [39], it seems possible to study the following

Problem 9. Study the estimate of λ_1 for the operator Δ^α for some real $\alpha > 0$.

Up to now, we have discussed only the 0-form. To go to the higher-order differential forms, the known mathematical tools are very limited and so are the results[13]. For instance, the Harnack inequality does not hold in this context. Since our new method does not use the inequality, it gives us a light to

Problem 10. Study the estimate of λ_1 for differential forms. Refer to [13, 40–42] and Lu (1994)[3]

Most of the problems are meaningful for diffusions in $\mathbb{R}^d$ and they become even harder in the latter case, due to the variant coefficient of the second-order term. In the past ten years or so, a large number of papers are devoted to study λ_1 for Markov chains. For which, the geometric tools (the Cheeger's inequality, the isoperimetric inequality, the Harnack inequality, the Nash inequality and so on) have played a critical role. Refer to refs. [16, 43–59] and Pan and Ycart (1995)[4]. Our new method works well also for the discrete situation, as illustrated in Chen refs. [10, 11] in part (I). It is the time to study more carefully the following problem.

Problem 11. Estimate λ_1 for Markov chains on graphs or for finite groups.

Recall that our main estimate comes from

$$\widetilde{\mathbb{E}}^{x,y}\bar{\rho}(x_t, y_t) \leqslant \bar{\rho}(x, y)e^{-\alpha t}$$

for all $t \geqslant 0$ and $x \neq y$ which is usually stronger than what we need for estimating λ_1, since it indeed implies an ergodic property with respect to the distance $\bar{\rho}$ with exponential rate α. For this estimate, the process should be neither reversible nor time-homogeneous. So the same technique is meaningful for the following

Problem 12. Study the exponential convergence rate for irreversible or time-inhomogeneous Markov processes. Refer to [60], Chen et al (1996)[5] *and Granovski*

[3]Lu, Y. G., Estimate of the first non-zero eigenvalue of Laplace-de Rahm and the Laplace-Beltrami operators, 1994, preprint

[4]Pan, Y. Y., Ycart, B., Gaps asymptotiques de générateurs de Markov perturbés, 1995, preprint

[5]Chen, Z. Q., Hu, Y. Z., Qian, Z. M., Zheng, W. A., Estimates on distance between two diffusion semigroups of uniformly elliptic divergence form operators, 1996, preprint

and Zeifman (1996)[6].

(*Received March* 29, 1997)

References

1. Chen, M. F., Wang, F. Y., General formula for lower bound of the first eigenvalue on Riemannian manifolds, *Sci. Sin.*, 1997, 27(1): 34 (Chinese Edition); 1997, 40(4): 384(English Edition).
2. Zhong, J. Q., Yang, H. C., Estimates of the first eigenvalue of a compact Riemannian manifolds, *Sci. Sin.*, 1984, 27(12): 1251.
3. Lichnerowicz, A. , *Geometrie des Groupes des Transformationes*, Dunod, Paris, 1958.
4. Cai, K. R., Estimate on lower bound of the first eigenvalue of a compact Riemannian manifold, *Chin. Ann. of Math.*, 1991, 12(B)(3): 267.
5. Yang, H. C., Estimate of the first eigenvalue of a compact Riemannian manifold with Ricci curvature bounded below by a negative constant, *Sci. Sin.*, (In Chinese), 1989, (A)32(7): 689.
6. Jia, F., Estimate of the first eigenvalue of a compact Riemannian manifold with Ricci curvature bounded below by a negative constant (In Chinese), *Chin. Ann. Math.*, 1991, 12(A)(4): 496.
7. Bérard, P. H., Besson, G., Gallot, S., Sur une inéqualité isopérimétrique qui généralise celle de Paul Lévy-Gromov, *Invent. Math.*, 1985, 80: 295.
8. Cheeger, J., 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, 1970.
9. Li, P., Yau, S. T., Estimates of eigenvalue of a compact Riemannian manifold, *Ann. Math. Soc. Proc. Symp. Pure Math.*, 1980, 36: 205.
10. Bakry, D., Ledoux, M., Lévy-Gromov's isoperimetric inequality for an infinite dimensional diffusion generator, *Invent. Math.*, 1996, 123: 259.
11. Bérard, P. H., Spectral Geometry: Direct and Inverse Problem, *LNM.* vol. 1207, 1986 Springer-Verlag.
12. Bobkov, S., A functional form of the isoperimetric inequality for the Gaussian measure, *J. Funct. Anal.*, 1996, 135(1): 39.
13. Chavel, I., Eigenvalues in Riemannian Geometry, Academic Press, 1984.
14. Ledoux, M., Isoperimetry and Gaussian Analysis, *Ecole d'été de Probabilités de Saint-Flour* 1994, to appear.
15. Li, P., Lecture Notes on Geometric Analysis, Seoul National Univ. Korea 1993.
16. Saloff-Coste, L., Lectures on Finite Markov Chains, Ecole d'été de Probabilités de Saint-Flour 1996, to appear.
17. Yau, S.T., Schoen, R., Differential Geometry, Science Press (in Chinese), Beijing, China 1988.
18. Singer, M., Wong, B., Yau, S. T., Yau, S. S. T., An estimate of the gap of the first two eigenvalues in the Schrödinger operator, *Ann. Scuola Norm. Sup. Pisa*, 1985, Series IV, Volume XII(2): 319.
19. Wang, F. Y., Application of coupling method to the Neumann eigenvalue problem, *Prob. Th. Rel. Fields*, 1994, 98: 299.
20. Berestycki, H., Nirenberg, L., Varadhan, S. R. S., The principal eigenvalue and maximum principle for second-order elliptic operators in general domains, *Comm. Pure and Appl.*, 1994, XLVII: 47.
21. Escobar, J. F., Uniqueness theorems on conformal deformation of metrics, Sobolev inequalities, and an eigenvalue estimate, *Comm. Pure and Appl. Math.*, 1990, XLIII: 857.
22. Chen, R., Neumann eigenvalue estimate on a compact Riemannian manifold, *Proc. Amer. Math. Soc.*, 1990, 108(4): 961.

[6]Granovski, B. L., Zeifman, A. I., The decay function of nonhomogeneous birth-death processes, with application to mean-field models, 1996, preprint

23. Wang, F. Y., Estimate of the first Dirichlet eigenvalue by using the diffusion processes, *Prob. Th. Rel. Fields*, 1994, 101: 363.
24. Wang, F. Y., A Probabilistic approach to the first Dirichlet eigenvalue on non-compact Riemannian manifold, *Acta Math. Sin. New Ser.*, 1997, 13(1): 116.
25. Wang, F. Y., Estimation of the first eigenvalue and the lattice Yang-Mills fields, *Chin. J. Math.* (In Chinese), 1996, 17A(2): 147; Chin. J. Contem. Math. (In English), 1996, 17(2): 119.
26. Buser, P., Colbois, B., Dodziuk, J., Tubes and eigenvalues for negatively curved manifolds, *J. Geom. Anal.*, 1993, 3(1): 1.
27. Tonno, S., The first eigenvalue of the Laplacian on spheres, *Tôhoku Math. J.*, 1979, 31: 179.
28. Li, P., Tam, L. F., Harmonic functions and the structure of complete manifolds, *J. Diff. Geom.*, 1992, 35: 359.
29. Wang, X. H., Bounded harmonic functions on a class of complete Riemannian manifolds, *Acta Math. Sin.* (In Chinese) 1995, 38(2): 171.
30. Baum, H., Eigenvalue estimates for Dirac operators coupled to instantons, *Annals of Global Anal. Geom.*, 1994, 12: 193.
31. Kirchberg, K. D., Compact six-dimensional Káhler spin manifolds of positive scalar curvature with the smallest possible first eigenvalue of the Dirac operator, *Math. Ann.*, 1988, 282: 157.
32. Kirchberg, K. D., The first eigenvalue of the Dirac operator on Kähler manifolds, *J. Geom. Phys.*, 1990, 7(4): 449.
33. Lu, K. P., The $(0,1)$ heat form of B^n and its application, *Acta Math. Sin.* (In Chinese), 1994, 37(2): 160.
34. Lu, K. P., The heat kernel of unitary group and its application, *Acta Math. Sin.* (In Chinese), 1994, 37(6): 744.
35. Li, P., Tian, G., On the heat kernel of the Bergmann metric on algebraic varieties, *J. Amer. Math. Soc.*, 1995, 8(4): 857.
36. Li, P., Schoen, R., L^p and mean value properties of subharmonic functions on Riemannian manifolds, *Acta Math.*, 1984, 153: 279.
37. Lu, Y. G., An estimate on non-zero eigenvalues of Laplacian in non-linear version, 1996.
38. Hoh, W., The martingale problem for a class of pseudo differential operators, *Math. Ann.*, 1994, 300: 121.
39. Hoh, W., Pseudo differential operators with negative definite symbols and the martingale problem, *Stochastics and Stoch. Reports*, 1995, 55: 225.
40. Asada, S., Notes of eigenvalues of Laplacian acting on p-forms, *Hokkaido Math. J.*, 1979, 8: 220.
41. Asada, S. , On the first eigenvalue of the Laplacian acting on p-forms, *Hokkaido Math. J.*, 1980, 9: 112.
42. Tonno, S., The spectrum of the Laplacian for 1-forms, *Proc. Amer. Math. Soc.*, 1974, 45(1): 125.
43. Chung, F. R. K., Spectral Graph Theory, CBMS Lecture Notes 1996, AMS Publ.
44. Chung, F. R. K., Yau, S. T., A Harnack inequality for homogeneous graphs and subgraphs, *Comm. Anal. Geom.*, 1994, 2: 628.
45. Chung, F. R. K., Yau, S. T., Logarithmic Harnack inequality, *Math. Research Letters*, 1997.
46. Chung, F. R. K., Graham, R. L., Yau, S. T., On sampling with Markov chains, *Random Structures and Algorithm*, 1996, 9: 55.
47. Chung, F. R. K., Graham, R. L., Yau, S. T., Eigenvalues and diameters for manifolds and graphs, *Advances*, 1997.
48. Diaconis, P., Saloff-Coste, L., Comparison theorems for reversible Markov chains, *Ann. Appl. Prob.*, 1993, 3(3): 696.
49. Diaconis, P., Saloff-Coste, L., Nash inequality for finite Markov chains, *J. Theor. Prob.*, 1996, 459–510.
50. Diaconis, P., Saloff-Coste, L., Logarithmic Sobolev inequality for finite Markov chains, *Ann. Appl. Prob.*, 1996.

51. Diaconis, P., Stroock , D. W., Geometric bounds for eigenvalues of Markov chains, *Ann. Appl. Prob.*, 1991, 1(1): 36.
52. Ingrassia, S., On the rate of convergence of the metropolis algorithm and Gibbs sampler by geometric bounds, *Ann. Appl. Prob.*, 1994, 4(2): 347.
53. Jerrum, M. R., Sinclair, A. J., Approximating the permanent, *SIAM J. Comput.*, 19889, 18: 1149.
54. 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.
55. Sullivan, W. G., The L^2 spectral gap of certain positive recurrent Markov chains and jump processes, *Z. Wahrs.*, 1994, 67: 387.
56. Thomas, L. E., Bound on the mass gap for finite volume stochastic Ising models at low temperature, *Comm. Math. Phys.*, 1989, 126: 1.
57. Stong, R., Eigenvalues of random walks on groups, *Ann. Prob.*, 1995, 23(4): 1961.
58. Sinclair, A. J., Jerrum, M. R., Approximate counting, uniform generation, and rapidly mixing Markov chains, *Inform. and Comput.*, 1989, 82: 93.
59. Coulhon, T., Grigor'yan, A., On-diagonal lower bounds for heat kernels and Markov chains, *Duke Univ. Math. J.*, 1997.
60. Chen, M. F., On ergodic region of Schlögl's model, in *Dirichlet Forms and Stoch. Proc.* (Eds. Ma, Z. M., Röckner, M., Yan, J. A., Walter de Gruyter, 1995, 87.

Acknowledgement Research supported in part by National Natural Science Foundation of China (Grant No. 19631060), Qiu Shi Science and Technology Foundation and the State Education Commission of China.

Chin. Sci. Bulletin, 1997, 42:18, 1497–1505.

Coupling, spectral gap and related topics (Ⅲ)

CHEN Mufa
Department of Mathematics, Beijing Normal University, Beijing 100875, China

Keywords: Diffusions, manifold, Markov chains, spectral gap, couplings.

This is the last one of a series of three papers. Here, we discuss six topics related to the spectral gap: the gradient estimate, the heat kernel and Harnack inequality, the logarithmic Sobolev inequality, the convergence in total variation, the algebraic convergence and the infinite-dimensional case. The perturbation of spectral gap and the logarithmic Sobolev constant under a linear transform is given (Theorem 5). A new proof for computing the logarithmic Sobolev constant in a basic case is also presented (Theorem 7).

5 Related topics.

5.1 Gradient estimate

This is the core of the Li-Yau's method[Ⅱ; 9, 17] (i.e., refs. [9] and [17] in the second paper of the series) for the estimation of λ_1. The coupling method does not need but can also be used to study the gradient estimate. There are different kinds of gradient estimate and the one we are talking is as follows: $|\nabla u(x)| \leqslant \text{const.}\|u\|_\infty$. Refer to [1] and [2] and references within. Here are two remarks. In the former paper, in the definition of the function $C(r)$, the constant $\sqrt{k(d-1)}$ can be replaced by

$$\sqrt{k(d-1)}\,\tanh\left(\frac{1}{2}\sqrt{\frac{k}{d-1}}\,s\right).$$

In the latter paper, one may obtain some new estimates by using the different distances as illustrated in Chen and Wang (1995) [I; 24].

5.2 Heat kernel and Harnack inequality

The well-known probabilistic proof of the Atiyah-Singer Index theorem depends on the short-time behavior of the heat kernel. From this, one sees the importance of the study on this topic. Even though these topics are well developed in geometry but it is still possible to make some addition. See Wang (1996)[1]. In the last paper, some probabilistic idea is adopted.

Combining these with what talked in the last two parts, one sees that a considerable progress has been made recently in the study of spectral theory in geometry and analysis. Compare with the books [Ⅱ; 11, 13, 15, 17].

5.3 Logarithmic Sobolev inequality

[1] Wang, F. Y., Sharp explicit lower bounds of heat kernels, 1996, preprint

Consider the operator $L = \Delta + \nabla V$ and set $\mathrm{d}\pi = e^V \mathrm{d}x$. Then the inequality means that

$$\int_M f^2 \log \frac{f^2}{\|f\|^2} \mathrm{d}\pi \leqslant \frac{2}{\alpha} \int_M |\nabla f|^2 \mathrm{d}\pi, \tag{5.1}$$

where $\|\cdot\|$ denotes the L^2-norm in $L^2(\pi)$. The largest constant α is called the *logarithmic Sobolev constant*. It is well known that $\lambda_1 \geqslant \alpha$, which explains the relation between λ_1 and α. Comparing (5.1) with the classical Sobolev inequality in $\mathbb{R}^d$:

$$\|f\|_p \leqslant \frac{1}{c} \|\nabla f\|_q, \qquad q < d, \quad p = \frac{dq}{d-q}, \tag{5.2}$$

we see that the former one contains the logarithmic factor more. This makes a serious difficulty. Note that the inequality is meaningful if one replace M with a general space and with a probability measure π on it (see the next part for more details). The main advantage of the inequality, in contrast to (5.2), is that it does not depend on the dimension of the space and hence it has become one of the main tools in the study of infinite-dimensional situation[3,4].

For compact manifolds with $V = 0$, a nice result due to Deuschel and Stroock (1990)[5] says that

$$\alpha \geqslant \max\left\{ \frac{\lambda_1}{d} + K,\ \frac{3\lambda_1 + Kd}{d+2} \right\}$$

which is sharp for the unit spheres. Combining this with Theorem 2, we get a general formula for the lower bound of α. Moreover, for non-compact manifolds or for general elliptic operators, we now have rather complete results for (5.1). See refs. [6], [7] and Wang (1996)[2] and references within. For Markov chains, the inequality is much difficult to handle. Even in the nearly trivial case that $M = \{0, 1\}$, it is still a non-trivial work to determine the optimal constant α (see Theorem 7 in the next part). One may refer to the recent papers [Ⅱ; 50, 45] for the study on finite Markov chains. However, nearly nothing is known for the following question.

Problem 13. When (5.1) holds for infinite Markov chains (with unbounded rates)?

To illustrate the role played by the geometric symmetry, we now discuss the perturbation of λ_1 and α under a linear transform.

Theorem 5. Consider the operators $L = \Delta + \nabla V$ and $\overline{L} = \Delta + \nabla \overline{V}$ for some $V \in C^2(\mathbb{R}^d)$ and $\overline{V}(x) = V(Mx)$ with $\det M \neq 0$. Denote by λ_1 and $\bar{\lambda}_1$ the corresponding first eigenvalues respectively. Then we have

$$\lambda_{\min}(MM^*)\lambda_1 \leqslant \bar{\lambda}_1 \leqslant \lambda_{\max}(MM^*)\lambda_1,$$

where $\lambda_{\max}(A)$ denotes the maximal eigenvalue of a symmetric matrix A. The same conclusion holds if λ_1 and $\bar{\lambda}_1$ are replaced by the logarithmic Sobolev constants α and $\bar{\alpha}$ respectively.

[2]Wang, F. Y., Logarithmic Sobolev inequalities on noncompact Riemannian manifolds, 1996, preprint

Proof. a) Set $\mathrm{d}\pi = Z^{-1}e^V\mathrm{d}x$ and $\mathrm{d}\bar{\pi} = \overline{Z}^{-1}e^{\overline{V}}\mathrm{d}x$, where Z and $\overline{Z}$ are normalizing constants. Define $\bar{g}(x) = f(Mx)$. Then

$$\bar{\pi}(\bar{g}) := \int \bar{g}\mathrm{d}\bar{\pi} = 0 \Longleftrightarrow \int f(Mx)e^{V(Mx)}\mathrm{d}x = 0 \Longleftrightarrow \int f\mathrm{d}\pi =: \pi(f) = 0.$$

$$\int \bar{g}^2 e^{\overline{V}}\mathrm{d}x = \int f(Mx)^2 e^{V(Mx)}\mathrm{d}x = \frac{1}{|\det M|}\int f^2 e^V \mathrm{d}x.$$

Thus, $\bar{g} \in L^2(\bar{\pi}) \Longleftrightarrow f \in L^2(\pi)$. Next, $(\nabla\bar{g})(x) = (M^*\nabla f)(Mx)$ and

$$|\nabla\bar{g}|^2(x) = (M^*\nabla f, M^*\nabla f)(Mx) = (MM^*\nabla f, \nabla f)(Mx).$$

Hence,

$$\begin{aligned}\frac{\int |\nabla\bar{g}|^2 e^{\overline{V}}\mathrm{d}x}{\int \bar{g}^2 e^{\overline{V}}\mathrm{d}x} &\leqslant \lambda_{\max}(MM^*)\frac{\int |\nabla f|^2(Mx)e^{V(Mx)}\mathrm{d}x}{\int f(Mx)^2 e^{V(Mx)}\mathrm{d}x} \\ &\leqslant \lambda_{\max}(MM^*)\frac{\int |\nabla f|^2 e^V \mathrm{d}x}{\int f^2 e^V \mathrm{d}x}.\end{aligned}$$

This proves that $\bar{\lambda}_1 \leqslant \lambda_{\max}(MM^*)\lambda_1$. The proof for the opposite inequality is similar.

b) Note that

$$\overline{Z} = \int e^{\overline{V}}\mathrm{d}x = \int e^{V(Mx)}\mathrm{d}x = \frac{1}{|\det M|}\int e^V \mathrm{d}x = \frac{Z}{|\det M|}. \tag{5.3}$$

We have

$$\begin{aligned}&\overline{Z}\int \bar{g}^2 \log \bar{g}^2 \mathrm{d}\bar{\pi} - \overline{Z}\int \bar{g}^2 \mathrm{d}\bar{\pi} \log \int \bar{g}^2 \mathrm{d}\bar{\pi} \\ &\quad = \int \bar{g}^2 \log \bar{g}^2 e^{\overline{V}}\mathrm{d}x - \int \bar{g}^2 e^{\overline{V}}\mathrm{d}x\bigg(\log \int \bar{g}^2 e^{\overline{V}}\mathrm{d}x - \log \overline{Z}\bigg) \\ &\quad = \int f(Mx)^2 \log f(Mx)^2 e^{V(Mx)}\mathrm{d}x \\ &\qquad - \int f(Mx)^2 e^{V(Mx)}\mathrm{d}x\bigg(\log \int f(Mx)^2 e^{V(Mx)}\mathrm{d}x - \log \overline{Z}\bigg) \\ &\quad = \frac{1}{|\det M|}\bigg[\int f^2 \log f^2 e^V \mathrm{d}x \\ &\qquad - \int f^2 e^V \mathrm{d}x\bigg(\log \int f^2 e^V \mathrm{d}x - \log|\det M| - \log \overline{Z}\bigg)\bigg] \\ &\quad = \frac{Z}{|\det M|}\bigg[\int f^2 \log f^2 \mathrm{d}\pi \\ &\qquad - \int f^2 \mathrm{d}\pi\bigg(\log \int f^2 \mathrm{d}\pi + \log Z - \log|\det M| - \log \overline{Z}\bigg)\bigg] \\ &\quad = \frac{Z}{|\det M|}\bigg[\int f^2 \log f^2 \mathrm{d}\pi - \int f^2 \mathrm{d}\pi \log \int f^2 \mathrm{d}\pi\bigg].\end{aligned} \tag{5.4}$$

On the other hand,

$$\int |\nabla \bar{g}|^2 e^{\overline{V}(x)} \mathrm{d}x = \int (MM^* \nabla f, \nabla f)(Mx) e^{V(Mx)} \mathrm{d}x$$
$$= \frac{Z}{|\det M|} \int |M \nabla f|^2 \mathrm{d}\pi, \tag{5.5}$$

Now, the second assertion of the theorem follows from (5.4) and (5.5). □

Part (1) of the corollary below answers a question proposed to the author by Yu. G. Kondratiev. Part (2) below improves [I; 24: Example 4.12]. One may extend [6; Corollary 1.6] in a similar way.

Corollary 6. (1) Take $M = U/m$ for some orthogonal matrix U and constant $m > 0$, then $\bar{\lambda}_1 = \lambda_1/m^2$ and $\bar{\alpha} = \alpha/m^2$, independent of d.
(2) Take $V(x) = -|x|^2/2$. Then $\bar{\lambda}_1 = \bar{\alpha} = \lambda_{\min}(MM^*)$.

Proof. Assertion (1) follows from Theorem 5 directly. We now prove assertion (2).

a) It follows from (1) that λ_1 and α are invariant under an orthogonal transform and hence we may assume that $\overline{V}(x) = -\sum_{i=1}^d m_i x_i^2/2$, where (m_i) are the eigenvalues of MM^*.

b) Now, since the components (x_i) are separated, we reduce the higher-dimensional case to dimension one (by additivity theorem [8; Theorem 2.6] and [4; Theorem 2.3]). That is, the λ_1 here equals the minimum of the λ_1's of the one-dimensional processes. The same conclusion holds for α.

c) Finally, since $\lambda_1 = \alpha = 1$, as another application of (1), we get $\bar{\lambda}_1 = \bar{\alpha} = \min_i m_i = \lambda_{\min}(MM^*)$. □

5.4 Convergence in total variation

Recall that the total variation distance of two probabilities μ and ν on a measurable space $(E, \mathscr{E})$ is defined by $\|\mu - \nu\|_{\mathrm{var}} = 2 \sup_{A \in \mathscr{E}} |\mu(A) - \nu(A)|$. Based on the *coupling inequality*[9]:

$$\|\mu P_t - \pi\|_{\mathrm{var}} \leqslant 2\mathbb{P}[T < \infty],$$

where π is the stationary measure of the process $P(t, x, dy)$, the most traditional topic in the study of coupling is the convergence in total variation[9],[10]. We are now interested in the exponential rate of this convergence. That is

$$\|\mu P_t - \pi\|_{\mathrm{var}} \leqslant C(\mu) e^{-\varepsilon t}, \qquad t \geqslant 0$$

for some constants $C(\mu) \geqslant 0$ and $\varepsilon > 0$ [11]. At the first look, this rate $\varepsilon_{\max}$ may be rather different from the spectral gap λ_1 since the latter describes another (i.e., L^2-) exponential convergence[8],[12]:

$$\|P_t f - \pi(f)\| \leqslant \|f - \pi(f)\| e^{-\lambda_1 t}, \qquad t \geqslant 0. \tag{5.6}$$

However, we have proved that in many cases, $\varepsilon_{\max} = \lambda_1$ and so our results provide automatically some general formula for the lower bound of $\varepsilon_{\max}$. See Wang (1996)[3] and Chen (1996)[4].

5.5 Algebraic convergence

A weaker convergence than (5.6) is the algebraic one:

$$\|P_t f - \pi(f)\| \leqslant V(f)/t^{\nu}, \qquad t > 0, \tag{5.7}$$

where $\nu > 0$ and V is a positive (may be infinity) functional on $L^2(\pi)$. There are several papers devoted to this topic, see for instance Liggett[15] and Deuschel[16]. From these papers, we learnt that the Lipschitz property of the semigroup with respect to some distance (not necessarily the Euclidean one) plays a critical role. On the other hand, as pointed in Chen (1994)[I; 10] the last property can be implied naturally by using the coupling method. Thus, one expects a further development on this topic.

Problem 14. Study the algebraic convergence for diffusions or for Markov chains.

5.6 Infinite-dimensional case

For infinite-dimensional situation, there are much more open problems. A large part of the study on mathematical physics concerns with the spectral theory. For instance, the main open problem in the study on loop space is to prove the existence or non-existence of the spectral gap[17]–[19]. My own interest in the field comes from the study on interacting particle systems. Here we discuss a standard model—the Ising model. For this, we need a little notations.

a) *State space.* $E = \{-1, +1\}^{\mathbf{Z}^d}$, endowed with the product topology. On which the set of probability measures is denote by $\mathscr{P}(E)$.

b) *Cylindrical functions.* Denoted by $Cy\ell\,(E)$ the set of functions depending on only finite number of coordinates $u \in \mathbf{Z}^d$.

c) *Speed functions.* $c(u, x) = \exp\big[-\beta \sum_{v:|v-u|=1} x_u x_v\big]$, $x = (x_u : u \in \mathbf{Z}^d) \in E$, where $|\cdot|$ is the usual Euclidean distance in $\mathbf{Z}^d$ and $\beta > 0$ is called the *inverse temperature*.

d) *Operator.* $\Omega f(x) = \sum_{u \in \mathbf{Z}^d} c(u, x)\big[f({}_u x) - f(x)\big]$ defined on $Cy\ell\,(E)$, where ${}_u x \in E$ is the flip of $x \in E$ at the site u: $({}_u x)_v = -x_u$ if $u = v$ and otherwise $= x_v$.

Let $\lambda_1(\beta)$ denote the first eigenvalue of Ω. We are interested in a recent program for describing the phase transitions:

> In the higher-dimensional case ($d \geqslant 2$), $\lambda_1(\beta)$ decreases from positive to zero as decreasing the temperature.

For the Ising model, the conclusion is proved by several authors. Refer to [20]–[23] and [Ⅱ; 56]. See also [5], [24]–[31], [Ⅱ; 25] and Bertini & Zegarlinski (1996)[5, 6] for related study.

---

[3]Wang, Y. Z., Convergence rate in total variation for diffusion processes, 1996, preprint

[4]Chen, M. F., Estimate of exponential convergence rate in total variation by spectral gap, 1996, preprint

[5]Bertini, L., Zegarlinski, B., Coercive inequality for Gibbs measures, 1996, preprint

[6]Bertini, L., Zegarlinski, B., Coercive inequality for Kawasaki dynamics: the product case, 1996, preprint

Problem 15. Prove the above result by using the coupling method.

A more traditional way to describe the phase transitions goes as follows. We say that $\pi \in \mathscr{P}(E)$ is *reversible* if

$$\int_E f\Omega g \mathrm{d}\pi = \int_E g\Omega f \mathrm{d}\pi, \qquad f, g \in Cy\ell\,(E). \tag{5.8}$$

Equivalently,

$$\int_E \pi(\mathrm{d}x)c(u,x)\big[f({}_u x) - f(x)\big] = 0, \qquad f \in Cy\ell\,(E), \quad u \in \mathbf{Z}^d \tag{5.9}$$

(cf. [10; Lemma 11.8 (1)]). The reversible measure π coincides with the *Gibbs state* in physics. The set of Gibbs states is denoted by $\mathscr{G}_\beta$, its cadinality is denoted by $|\mathscr{G}_\beta|$. Now, a famous result (cf. [32] or [10] for instance) says that we have $|\mathscr{G}_\beta| = 1$ when $d = 1$ and for $d \geqslant 2$, there exists $\beta_c^{(d)} \in (0, \infty)$ such that

$$\begin{aligned} |\mathscr{G}_\beta| &= 1, \qquad \text{if } \beta < \beta_c^{(d)} \\ |\mathscr{G}_\beta| &> 1, \qquad \text{if } \beta > \beta_c^{(d)}. \end{aligned}$$

In other words, the existence of phase transitions is equivalent to that Eq. (5.9) has multi-solutions π. There is also a variational principle: the Gibbs state minimizes the relative entropy (see ref. [32; Theorem 15.39]). Of course, one can replace the spin space $\{-1, +1\}$ by general manifold and replace Ω by differential operators. For which we still have (5.8) and the variational principle.

When $|\mathscr{G}_\beta| = 1$, we are in the ergodic region. One then expects an exponential ergodicity and hence $\lambda_1(\beta) > 0$. When $|\mathscr{G}_\beta| > 1$, the system is not ergodic and so $\lambda_1(\beta) = 0$. This explains the meaning of the program mentioned above.

6 Appendix: Logarithmic Sobolev constant.

The main purpose of the appendix is to compute the exact logarithmic Sobolev constant α in the simplest case that $E = \{0, 1\}$. Even though the proof here is rather elementary but it is worthy to be presented here since on the one hand the new proof is considerably simpler than the original one[Ⅱ; 50] and on the other hand this particular case is a key to deduce a non-trivial lower bound of α for any finite Markov chains[Ⅱ; 50].

First, we recall some general facts. Let π be a probability measure on a measurable state space $(E, \mathscr{E})$ and $(P_t)_{t\geqslant 0}$ be a semigroup of a Markov process with Dirichlet form $(D, \mathscr{D}(D))$. For $L = \Delta + \nabla V$ on M, the Dirichlet form is $D(f,f) = \int_M |\nabla f|^2 \mathrm{d}\pi\,(\mathrm{d}\pi = e^V \mathrm{d}x/Z)$ with $\mathscr{D}(D) \supset$ the set of all smooth functions with compact support. For countable E, we have a Q-matrix $Q = (q_{ij})$, the corresponding Dirichlet form is $D(f,f) = \frac{1}{2}\sum_{i,j\in E} \pi_i q_{ij}(f_i - f_j)^2$ with domain $\mathscr{D}(D) = \{f \in L^2(\pi) : D(f,f) < \infty\}$. Then, the logarithmic Sobolev inequality means that

$$\int f^2 \log \frac{f^2}{\|f\|^2} \mathrm{d}\pi \leqslant \frac{2}{\alpha} D(f,f), \qquad f \in \mathscr{D}(D) \tag{6.1}$$

Noticing that $D(|f|,|f|) \leqslant D(f,f)$, one may assume in (6.1) that $f \geqslant 0$. Then, we can rewrite $f^2/\|f\|^2$ as $\mathrm{d}\mu/\mathrm{d}\pi$ for a probability measure μ. Thus, (6.1) is equivalent to

$$\text{The relative entropy} \leqslant \frac{2}{\alpha}\,\text{The Donsker-Varadhan entropy}.$$

That is,

$$\int \mathrm{d}\mu \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \leqslant \frac{2}{\alpha}\, D\Big(\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}}, \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}}\Big). \tag{6.2}$$

When $\mu \not\ll \pi$, both sides of (6.2) are defined to be ∞. Refer to [10; (9.14)] for instance. The next result is due to Diaconis and Saloff-Coste (1996)[Ⅱ; 50].

Theorem 7. Let $\theta \in (0,1/2]$. Consider the Markov chain on $\{0,1\}$ with Q-matrix $\begin{pmatrix} -\theta, & \theta \\ 1-\theta, & \theta-1 \end{pmatrix}$. Then the logarithmic Sobolev constant is equal to $\alpha = \alpha(\theta) = \frac{2-4\theta}{\log[(1-\theta)/\theta]}$. When $\theta = \frac{1}{2}$, $\alpha = \lim\limits_{\theta\to 1/2} \alpha(\theta) = 1$.

Proof. a) Note that $\pi_0 = 1-\theta$ and $\pi_1 = \theta$. Take $\mu_0 = x$ and $\mu_1 = 1-x$, $x \in [0,1]$. Set $h(\theta) = (1-2\theta)^{-1}\log(1/\theta - 1)$. On the other hand, the Donsker-Varadhan entropy equals $\big(\sqrt{x\theta} - \sqrt{(1-x)(1-\theta)}\,\big)^2$ (cf. [10; Corollary 8.18], the result is due to Chen and Lu (1991)[33]), which can be deduced directly by using the Dirichlet form for countable state space E. Thus, we need only to show that

$$x\log\frac{x}{1-\theta} + (1-x)\log\frac{1-x}{\theta} \leqslant h(\theta)\big(\sqrt{x\theta} - \sqrt{(1-x)(1-\theta)}\,\big)^2, \qquad \theta \in (0,1/2],\ x \in [0,1]. \tag{6.3}$$

Before moving further, we mention that it is quite easy to guess the required answer $h(\theta)$ in (6.3). First, the equality in (6.3) holds at $x = \theta$ and $x = 1-\theta$. When $\theta < 1/2$, the latter one is clearly the degenerated case since both sides of (6.3) vanish. Secondly, when we plot the ratio

$$\frac{x\log\frac{x}{1-\theta} + (1-x)\log\frac{1-x}{\theta}}{\big(\sqrt{x\theta} - \sqrt{(1-x)(1-\theta)}\,\big)^2} \tag{6.4}$$

by using Mathematica, one sees that the ratio is less than $h(\theta)$ unless $x = \theta$. Finally, the Taylor expansion of (6.4) at $x = \theta$ equals $h(\theta) - k(\theta)(x-\theta)^2 + \mathrm{O}(x-\theta)^3$ for some $k(\theta) > 0$, $\theta \in (0,1/2]$.

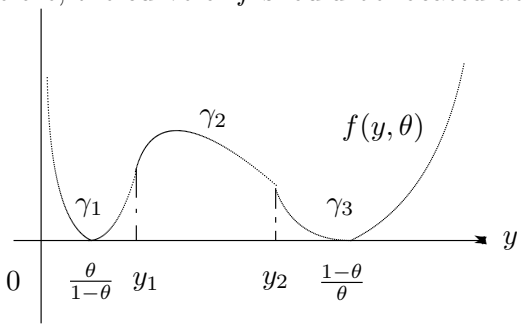
b) We now start to prove (6.3). Because

$$\frac{1-x}{x}\log\frac{1-x}{\theta} + \log\frac{x}{1-\theta} = \frac{1-x}{x}\log\frac{1-x}{x\theta} + \frac{1}{x}\log x + \log\frac{1}{1-\theta}.$$

By making a change of variables $y = (1-x)/x$ (and hence $x = 1/(1+y)$), we see that (6.3) holds iff

$$f(y,\theta) := h(\theta)\big(\sqrt{\theta} - \sqrt{y(1-\theta)}\,\big)^2 - y\log\frac{y}{\theta} + (1+y)\log(1+y) + \log(1-\theta) \geqslant 0. \tag{6.5}$$

c) The special case that $\theta = 1/2$ is easier and will be treated at the end of the proof. We now assume that $\theta \leqslant 1/2 - \varepsilon$ for some $\varepsilon \in (0, 1/2)$. To prove (6.5), our goal is to show that for each fixed θ, the continuous curve of $f(\cdot, \theta)$ can be successively divided into three parts γ_1, γ_2 and γ_3 having the properties: *i)* γ_1 and γ_3 are convex but γ_2 is concave, *ii)* the y-axis is the common tangent line to γ_1 and γ_3. Therefore, the curve of f should be located above the y-axis.



For this, it suffices to prove the following two assertions.

(1) $f(\cdot, \theta)$ and $\frac{\partial f}{\partial y}(\cdot, \theta)$ equal zero at $\frac{\theta}{1-\theta}$ and $\frac{1-\theta}{\theta}$.

(2) There exist $y_1 = y_1(\theta) < y_2 = y_2(\theta)$ in the open interval $\left(\frac{\theta}{1-\theta}, \frac{1-\theta}{\theta}\right)$ such that $\frac{\partial^2 f}{\partial y^2}(\cdot, \theta)$ is negative in (y_1, y_2) and positive out of $[y_1, y_2]$.

d) The proof of (1) is easy. Since (6.3) becomes equality at $x = \theta$ and $x = 1-\theta$, the assertion for f follows from the substitution $y = (1-x)/x$. The assertion for $\partial f/\partial y$ follows from (6.6) below.

$$\frac{\partial f}{\partial y} = \log \frac{\theta(1+y)}{y} - (1-\theta)h(\theta)\left(\sqrt{\frac{\theta}{y(1-\theta)}} - 1\right). \tag{6.6}$$

$$\frac{\partial^2 f}{\partial y^2} = \frac{1}{2y^{3/2}}\left(\sqrt{\theta(1-\theta)}\, h(\theta) - \frac{2\sqrt{y}}{1+y}\right). \tag{6.7}$$

Next, we show that the assertion (2) follows from

$$2 < h(\theta) < [\theta(1-\theta)]^{-1/2}, \qquad \theta < 1/2. \tag{6.8}$$

Actually, by (6.7), the second inequality of (6.8) implies that $\partial^2 f/\partial y^2$ has two roots y_1 and y_2 in $(0, \infty)$. Because $\sqrt{y}/(1+y)$ is unimodal in y which achieves the maximum $1/2$ at $y = 1$, it follows that $\partial^2 f/\partial y^2$ is negative in the (y_1, y_2) and positive out of $[y_1, y_2]$. On the other hand, since $\sqrt{y}/(1+y) = \sqrt{\theta(1-\theta)}$ at $y = \theta/(1-\theta)$ and $(1-\theta)/\theta$, by the first inequality of (6.8), one sees that $\partial^2 f/\partial y^2 > 0$ at $y = \theta/(1-\theta)$ and $(1-\theta)/\theta$. Hence $y_1, y_2 \in \left(\frac{\theta}{1-\theta}, \frac{1-\theta}{\theta}\right)$ and so the assertion (2) follows.

To prove (6.8), note that

$$m(\theta) := \log(1/\theta - 1) - 2(1-2\theta), \qquad m'(\theta) = 4 - [\theta(1-\theta)]^{-1}.$$

We have $m'(\theta) \leqslant 0$ and the equality holds iff $\theta = 1/2$. Thus,

$$m(\theta) \geqslant m(1/2 - \varepsilon) > m(1/2) = 0$$

for all $\theta \leqslant 1/2 - \varepsilon$. This proves the first half of (6.8). For the second one, set

$$n(\theta) = 1 - 2\theta - \sqrt{\theta(1-\theta)}\,\log(1/\theta - 1).$$

Then, by using the first half of (6.8), we get

$$\begin{aligned} n'(\theta) &= -2 + \frac{1}{\sqrt{\theta(1-\theta)}} - \frac{1-2\theta}{2\sqrt{\theta(1-\theta)}}\,\log\left(\frac{1}{\theta} - 1\right) \\ &\leqslant -2 + \frac{1}{\sqrt{\theta(1-\theta)}} - \frac{(1-2\theta)^2}{\sqrt{\theta(1-\theta)}} \\ &= -2 + 4\sqrt{\theta(1-\theta)} < 0, \qquad \theta < 1/2. \end{aligned}$$

Hence, $n(\theta) > n(1/2) = 0$ for all $\theta < 1/2$.

e) We now come to the special case that $\theta = 1/2$. Then, $h(\theta) = 2$ and $\frac{\theta}{1-\theta} = \frac{1-\theta}{\theta} = y_1 = y_2 = 1$. Because $\partial^2 f/\partial y^2 > 0$ unless $y = 1$, $\partial f/\partial y$ has only one zero-point $y = 1$ and so f attains its global minimum 0 at $y = 1$.

(*Received March* 29, 1997)

References

1. Wang, F. Y., Gradient estimates for generalized harmonic function on Riemannian manifolds, *Chin. Sci. Bull.*, 1994, 39(22): 1849.
2. Wang, F. Y., Gradient estimates on $\mathbb{R}^d$, *Canad. Math. Bull.*, 1994, *XX*(2): 1.
3. Bakry, D., L'hypercontractivité et son utilisation en théorie des semigroupes, *LNM*, **1581**, Springer, 1992.
4. Gross, L., Logarithmic Sobolev inequalities and contractivity of semigroups, *LNM*. **1563**, Springer, 1993.
5. Deuschel, J. D., Stroock, D. W., Hypercontractivity and spectral gap of symmetric diffusion with applications to the stochastic Ising models, *J. Funct. Anal.*, 1990, 92: 30.
6. Chen, M. F., Wang, F. Y., Estimates of logarithmic Sobolev constant — An improvement of Bakry–Emery criterion, *J. Funct. Anal.*, 1997, 144(2): 287.
7. Wang, F. Y., On estimates of logarithmic Sobolev constant (In Chinese), *J. Beijing Normal Univ.*, 1994, 30(4): 448.
8. Liggett, T. M., Exponential L_2 convergence of attractive reversible nearest particle systems, *Ann. Prob.*, 1989, 17: 403.
9. Lindvall, T., Lectures on the Coupling Method, Wiley, New York, 1992.
10. Chen, M. F., From Markov Chains to Non-Equilibrium Particle Systems, World Scientific, 1992.
11. Down, D., Meyn, S. P., Tweedie, R. L., Exponential and uniform ergodicity of Markov processes, *Ann. Prob.*, 1995, 23(4): 1671.
12. Chen, M. F., Exponential L^2-convergence and L^2-spectral gap for Markov processes, *Acta Math. Sin. New Ser.*, 1991, 7(1): 19.
13. Saloff-Coste, L., Precise estimates on the rate at which certain diffusions tend to equilibrium, *Math. Z.*, 1994, 217: 641.
14. Saloff-Coste, L., Convergence to equilibrium and logarithmic Sobolev constant on manifolds with Ricci curvature bounded below, *Coll. Math.*, 1994, LXVⅡ(1): 109.
15. Liggett, T. M., Exponential L_2 convergence of attractive reversible nearest particle systems: the critical case, *Ann. Prob.*, 1991, 19: 935.
16. Deuschel, J. D., Algebraic L^2 decay of attractive critical processes on the lattice, *Ann. Prob.*, 1991, 22: 261.

17. Gross, L., Analysis on loop groups, in *Stochastic analysis and applications in physics*, NATO ASI, Ser. C: Math. Phys. Sci., 449, Kluwer Acad. Publ., 1994.
18. Hsu, E. P., Logarithmic Sobolev inequality on path spaces, *Ann. of Math.*, 1996.
19. Wang, F. Y., Logarithmic Sobolev inequalities for diffusion processes with application to path space, *Chin. J. Appl. Prob. Stat.*, 1996, 12(3): 255.
20. Holley, R., Stroock, D. W., Uniform and L^2 convergence in one dimensional stochastic Ising models, *Comm. Math. Phys.*, 1989, 123: 85.
21. Minlos, R. A., Invariant subspaces of the stochastic Ising high temperature dynamics, *Markov Processes Relat. Fields*, 1996, 2: 263.
22. Minlos, R. A., Trisch, A., Complete spectral decomposition of the generator for one-dimensional Glauber dynamics, *Uspekhi Matem. Nauk* (In Russian), 1994, 49: 209.
23. Schonmann, R. H., Slow drop-driven relaxation of stochastic Ising models in the vicinity of the phase coexistence region, *Commun. Math. Phys.*, 1994, 161: 1.
24. Holley, R., Stroock, D. W., Logarithmic Sobolev inequalities and stochastic Ising models, *J. Stat. Phys.*, 1987, 46: 1159.
25. Lu, S. L., Yau, H. T., Spectral gap and logarithmic Sobolev inequality for Kawasaki and Glauber dynamics, *Comm. Math. Phys.*, 1993, 156: 399.
26. Martinelli, F., On the two dimensional dynamical Ising model in the phase coexistence region, *J. Stat. Phys.*, 1994, 76: 1179.
27. Sokal, A. D., Thomas, L. E., Absence of mass gap for a class of stochastic contour models, *J. Statis. Phys.*, 1988, 51(5/6): 907.
28. Stroock, D. W., Zegarlinski, B., The equivalence of the logarithmic Sobolev inequality and the Dobrushin-Shlosman mixing condition, *Comm. Math. Phys.*, 1992, 144(2): 303.
29. Wang, F. Y., Ergodicity for infinite-dimensional diffusion processes on manifolds, *Sci. Sin.* Ser. A, 1994, 37(2), 137.
30. Wang, F. Y., Uniqueness of Gibbs states and the L^2-convergence of infinite-dimensional reflecting diffusion processes, *Sci. Sin.* Ser. A, 1995, 32(8): 908.
31. Wang, F. Y., Estimates of logarithmic Sobolev constant for finite volume continuous spin systems, *J. Stat. Phys.*, 1995, 84(1/2): 277.
32. Georgii, H. O., Gibbs Measures and Phase Transitions, de Gruyter Studies in Math.9, 1988.
33. Chen, M. F. Lu, Y. G., On evaluating the rate function of large deviations for jump processes, *Acta Math. Sin.* New Ser. 1990, 6(3): 206.

Acknowledgement The author acknowledges Prof. F. Y. Wang for the fruitful cooperation. Partial results of the paper was obtained while the author visited Univ. of Bari, Italy in November, 1996. The author would like to acknowledge the warm hospitality and the financial support from Italian CNR, especially Professor Y. G. Lu for his valuable discussions. Finally, the research is supported in part by National Natural Science Foundation of China (Grant No. 19631060), Qiu Shi Science and Technology Foundation and the State Education Commission of China.

Acta Math. Sin. New Ser. 14:1, 9–16, 1998.

ESTIMATE OF EXPONENTIAL CONVERGENCE RATE IN TOTAL VARIATION BY SPECTRAL GAP

Mu-Fa Chen

(Beijing Normal University)

(October 22, 1996)

Abstract. This note is devoted to study the exponential convergence rate in the total variation for reversible Markov processes by comparing it with the spectral gap. It is proved that in a quite general setup, with a suitable restriction on the initial distributions, the rate is bounded from below by the spectral gap. Furthermore, in the compact case or for birth-death processes or half-line diffusions, the rate is shown to be equal to the spectral gap.

1. Introduction.

Let $P_t(x,\cdot)$ be the transition probability of a Markov process on a measurable state space $(E,\mathscr{E})$ with stationary distribution π. Denote by $\mathscr{P}$ the set of probability measures on $(E,\mathscr{E})$. Recall that for μ_1, $\mu_2 \in \mathscr{P}$, the variational norm of μ_1 and μ_2 is defined by $\|\mu_1-\mu_2\|_{\rm var} = 2\sup_{A\in\mathscr{E}}|\mu_1(A)-\mu_2(A)|$. The exponential convergence in the variational norm means that for every $\mu\in\mathscr{P}$,

$$\|\mu P_t-\pi\|_{\rm var} \leqslant C(\mu)e^{-\varepsilon t}, \qquad t\geqslant 0 \tag{1.1}$$

for some constants $\varepsilon>0$ and $C(\mu)\geqslant 0$. Sometimes, we will use a subset $\mathscr{P}_0\subset\mathscr{P}$ instead of the set of all initial distributions. Denote by $\sigma=\sigma(\mathscr{P}_0)$ the largest rate ε such that (1.1) holds for all $\mu\in\mathscr{P}_0$.

Throughout this note, we consider only reversible $P_t(x,\cdot)$: $\int_A \pi(dx)P_t(x,B) = \int_B \pi(dx)P_t(x,A)$ for all $A,B\in\mathscr{E}$ and $t\geqslant 0$. For which we have the L^2-exponential convergence:

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

2000 *Mathematics Subject Classification.* 60J25, 60J27, 60J60.

Key words and phrases. Total variation, spectral gap, Markov processes.

Research supported in part by NSFC, Qiu Shi Sci. & Tech. Found., DPFIHE, MCSEC and Univ. of Rome I, Italy.

where $\|\cdot\|_{p,\pi}$ denotes the L^p-norm in the space $L^p(\pi)$ (real) and $\pi(f) = \int f d\pi$. It is known that the largest exponential rate ε in (1.2) is given by the spectral gap:

$$\begin{aligned} \text{gap}\,(D) &= \inf\{-(f, Lf)_\pi : f \in \mathscr{D}(L), \pi(f) = 0 \text{ and } \|f\|_{2,\pi} = 1\} \\ &= \inf\{D(f,f) : f \in \mathscr{D}(D), \pi(f) = 0 \text{ and } \|f\|_{2,\pi} = 1\}, \end{aligned} \tag{1.3}$$

where L is the generator of P_t with domain $\mathscr{D}(L)$ in the L^2-sense and $(D(f,f), \mathscr{D}(D))$ is the corresponding Dirichlet form. In other words, the largest rate in (1.2) is just the first (non-trivial) eigenvalue λ_1 of the operator $-L$. Refer to [1] and [2], or [3; Chapter 9].

The main purpose of the note is to estimate the exponential rate σ in variational norm in terms of λ_1. Since these two types of convergence are rather different and so one may wonder at the first look if their rates are comparable. Nevertheless, they do have close relation as one will see very soon. To state the main results of the note, we should explain some conditions. The main hypothesis we need is the following:

(H_1). For each $t > 0$, the transition probability $P_t(x, \cdot)$ has a density $p_t(x,y)$ with respect to a reference measure λ on $(E, \mathscr{E})$. Moreover, $p_t(x,y)$ is joint measurable in (x,y).

Then, it is easy to show that the reversible measure π also has a density $\pi(x)$ with respect to λ (Lemma 2.1). For simplicity, we assume that

(H_2). $\pi(x) > 0$ everywhere.

We can now state our first result as follows.

Theorem 1.1. Let (H_1) and (H_2) hold. Take

$$\mathscr{P}_0 = \{\mu(dx) = \mu(x)\lambda(dx) : \mu/\pi \in L^2(\pi)\},$$

where μ/π denotes the function $\mu(x)/\pi(x)$, $x \in E$. Then, we have $\sigma(\mathscr{P}_0) \geqslant \lambda_1$.

To go to the opposite direction, we need another condition. Suppose that for the semigroup P_t induced by $P_t(x, \cdot)$, we have an extended generator L^{e}. That is,

$$\frac{d}{dt} P_t f(x) = P_t L^{\text{e}} f(x)$$

for all $x \in E$ and suitable measurable function f (not necessarily bounded). The set of such functions f consists of the domain $\mathscr{D}(L^{\text{e}})$. Next, denote by g the eigenfunction of λ_1, that is, $L^{\text{e}} g(x) = -\lambda_1 g(x)$ for all $x \in E$ (some boundary condition may be included).

(H_3). There exists an eigenfunction $g \in \mathscr{D}(L^{\text{e}})$, which is bounded from below (or above) and $\pi(g) = 0$.

Theorem 1.2. Under (H_1)–(H_3), there exists a probability measure μ_0 such that $\|\mu_0 P_t - \pi\|_{\text{var}} = \|\mu_0 - \pi\|_{\text{var}} e^{-\lambda_1 t}$ for all $t \geqslant 0$.

We now consider two typical classes of Markov processes.

Corollary 1.3. Let E be discrete, take λ to be the counting measure and assume that the reversible Markov chain is irreducible.

(1) When E is finite, we have $\sigma(\mathscr{P}) = \lambda_1$.
(2) For birth-death processes, the assumptions of Theorem 1.1 and Theorem 1.2 hold. If moreover, the eigenfunction of λ_1 belongs to $L^2(\pi)$, then $\sigma(\mathscr{P}_0) = \lambda_1$ for $\mathscr{P}_0$ given in Theorem 1.1.[1]

Next, consider diffusion processes on a domain $E \subset \mathbb{R}^d$ with operator

$$L = \sum_{i,j=1}^{d} a_{ij}(x)\partial_i\partial_j + \sum_i b_i(x)\partial_i,$$

where

$$b_i = \sum_{j=1}^{d}(a_{ij}\partial_j V + \partial_j a_{ij}).$$

The form of b_i comes from the self-adjoint property of the operator L. For simplicity, assume that a_{ij} and V are all smooth functions and (a_{ij}) is positive definite. In the case that $\partial E \neq \emptyset$, the reflecting boundary is imposed.

Corollary 1.4. Consider the diffusion processes as above and take λ to be the Lebesgue measure.

(1) When E is compact, we have $\sigma(\mathscr{P}_0) = \lambda_1$ for $\mathscr{P}_0 = \{\mu(dx) = \mu(x)dx : \mu(x)$ is continuous and $\mu/\pi \in L^2(\pi)\}$.
(2) When E is a half-line, the assumptions of Theorem 1.1 and Theorem 1.2 hold. If moreover, the eigenfunction of λ_1 belongs to $L^2(\pi)$, then $\sigma(\mathscr{P}_0) = \lambda_1$ for $\mathscr{P}_0$ given in part (1) of the corollary.

We are now ready to make some remarks on the above results. Certainly, part (1) of Corollary 1.4 is also meaningful for the diffusions on compact manifolds. This needs a small modification and hence is omitted here. Clearly, Theorem 1.1 is most useful in practice since one needs mainly an upper bound for the convergence. However, one may argue about the use of the theorem since it transfers one difficult problem to the another. It is well known that the estimate of the lower bounds of λ_1 is much harder to handle than the upper ones for which the variational formula

[1] Addition to the original proof. The second assertion can be improved as follows: Moreover, $\sigma(\mathscr{P}_1) = \sigma(\mathscr{P}_0) = \lambda_1 \geqslant \sigma(\mathscr{P})$, where $\mathscr{P}_1 = \{\delta_i : i \in E\}$ and $\mathscr{P}_0$ is the same as in Theorem 1.1.

Proof The conclusion "$\lambda_1 \geqslant \sigma(\mathscr{P})$" follows from Theorem 1.2. Since $\mathscr{P}_1 \subset \mathscr{P}_0$, by Theorem 1.1, we have

$$\lambda_1 \leqslant \sigma(\mathscr{P}_0) \leqslant \sigma(\mathscr{P}_1).$$

Next, write $C_i = C(\delta_i)$. By definition,

$$C_i \exp[-\sigma(\mathscr{P}_1)t] \geqslant \|\delta_i P_t - \pi\|_{\mathrm{Var}} = \sum_j |p_{ij}(t) - \pi_j| \geqslant |p_{ij}(t) - \pi_j|.$$

Combining this with [3; Proposition 9.20], we obtain $\sigma(\mathscr{P}_1) \leqslant \hat{\alpha} = \lambda_1$. $\square$

(1.3) is available. Fortunately, we now also have some general formulas for the lower bounds of λ_1 and this is indeed the starting point of the present paper. Since the new formulas are very essential for our purpose (for instance, based on Theorem 1.1, these formulas give us automatically some general formulas for the lower bounds of $\sigma(\mathscr{P}_0)$ considered in Theorem 1.1), the reader are urged to refer to [4]–[6] for the details. Refer also to [1], [7], [8]–[12] for different estimates of λ_1 and references within.

The key for proving the above theorems is the following simple observation

$$\|\mu P_t - \pi\|_{\rm var} = \|P_t(\mu/\pi - 1)\|_{1,\pi}, \tag{1.4}$$

which will be proved in the next section. This comes as far as I know from [10; Proposition 1] for Markov chains with finite state space and it enables the author to pass through from finite state space to infinite one. Actually, noticing that $\mu/\pi - 1$ is an $L^1(\pi)$-function with mean zero, by Cauchy-Schwarz inequality and (1.4), we get

$$\|\mu P_t - \pi\|_{\rm var} \leqslant \|P_t(\mu/\pi - 1)\|_{2,\pi} \leqslant \|\mu/\pi - 1\|_{2,\pi} e^{-\lambda_1 t}, \tag{1.5}$$

which gives us the conclusion of Theorem 1.1. From (1.4), Theorem 1.2 also follows easily. Again, because μ/π is an $L^1(\pi)$-function with mean zero, one sees that the rate σ is close related to (indeed it is bigger or equal to) the exponential L^1-convergence rate defined in a similar way as in (1.2). Certainly, the exponential L^1-convergence rate is more difficult to handle than the one in the L^2-sense and it is not the aim of the note. But if one replaces σ with the exponential L^1-convergence rate, all the above results remain the same and moreover the condition "g is bounded from below" given in (H_3) can be removed. We have thus explained the main reason why the rate σ can be related to λ_1.

The meaning of Theorem 1.2 is mainly theoretical, it points out the main situation for which $\sigma = \lambda_1$. To check condition (H_3) is far away to be easy, even for birth-death processes or for half-line diffusions, as one will see in the next section. It is interesting that for birth-death processes, the rates σ, λ_1 and the rate for the exponential ergodicity can all be the same. Refer to [2] or [3; Theorem 9.1].

For the remainder of this section, we discuss the constant

$$C(\mu) = \|\mu/\pi - 1\|_{2,\pi} = \big(\int \mu(dx)[\mu/\pi](x) - 1\big)^{1/2}, \qquad \mu/\pi \in L^2(\pi)$$

appeared in (1.5). The use of Cauchy-Schwarz inequality is natural to obtain the rate λ_1 but it does enlarge the constant from $\|\mu/\pi - 1\|_{1,\pi}$ to $\|\mu/\pi - 1\|_{2,\pi}$ and so the constant $\|\mu/\pi - 1\|_{2,\pi}$ given in (1.5) can not be sharp in general. For Markov chains, this constant is good enough since it contains all measures $\mu = \delta_i$ with single mass at $i \in E$. Moreover, for finite state space,

$$\|\mu/\pi - 1\|_{2,\pi} \leqslant \|\mu/\pi - 1\|_{1,\pi}/\sqrt{\pi_*}, \tag{1.6}$$

where $\pi_* = \min_i \pi_i$. When restricting to the class $\{\mu : \mu = \delta_i, i \in E\}$, we indeed have $\|\mu/\pi - 1\|_{2,\pi} \leqslant \sqrt{1/\pi_* - 1} < 1/\sqrt{\pi_*}$.

For diffusions, it seems that the assumption of Theorem 1.1 rules out the initial distribution δ_x with single mass at x. But this is indeed not a restriction since we can often use $\mu_s(dy) = p_s(x,y)\lambda(dy)$ instead of δ_x. Then, we have

$$\|\delta_x P_t - \pi\|_{\mathrm{var}} \leqslant \big(\|p_s(x,\cdot)/\pi - 1\|_{2,\pi} e^{\lambda_1 s}\big) e^{-\lambda_1 t}.$$

The coefficient on the right-hand side is usually bounded in the compact case[2]. Refer to [11] and [13] for some different treatments. Thus, the main restriction of Theorem 1.1 is the L^2-integrability of the function $[\mu/\pi](x)$. Since the ergodicity is not a problem for our purpose, we are now interested in the convergence rate and so it is meaningful to obtain the exact rate even with a restriction on the initial distributions.

We now look at a possibly different way to describe the exponential convergence rate in the total variation. Note that for every finite signed measure ν, $\|\nu\|_{\mathrm{var}} = \sup_{f\in\mathscr{E}:|f|\leqslant 1} |\nu(f)|$. We have for every bounded signed measure ν that

$$\|\nu P_{t+s}\|_{\mathrm{var}} = \sup_{|f|\leqslant 1} |\nu P_t P_s f| \leqslant \sup_{|g|\leqslant 1} |\nu P_t g| = \|\nu P_t\|_{\mathrm{var}}.$$

Hence, $t \to \|\nu P_t\|_{\mathrm{var}}$ is non-increasing. Next, define

$$\underline{\sigma}(t) = -\sup_{\mu\neq\pi} \log\big[\|\mu P_t - \pi\|_{\mathrm{var}}/\|\mu - \pi\|_{\mathrm{var}}\big].$$

Then,

$$\|\mu P_{t+s} - \pi\|_{\mathrm{var}} = \|(\mu P_t)P_s - \pi\|_{\mathrm{var}} \leqslant \|\mu P_t - \pi\|_{\mathrm{var}} e^{-\underline{\sigma}(s)} \leqslant \|\mu - \pi\|_{\mathrm{var}} e^{-\underline{\sigma}(t)-\underline{\sigma}(s)}.$$

It follows that $t \to \underline{\sigma}(t)$ is super-additive and moreover $\underline{\sigma}(t) \downarrow \underline{\sigma}(0) = 0$ as $t \downarrow 0$. Therefore,

$$\underline{\sigma} := \lim_{t\downarrow 0} \frac{\underline{\sigma}(t)}{t} = \inf_{t>0} \frac{\underline{\sigma}(t)}{t}$$

is well defined. This suggests us to use $\sigma = \underline{\sigma}$ in (1.1) with constant $C(\mu) = \|\mu - \pi\|_{\mathrm{var}}$. In other words, $\|\mu - \pi\|_{\mathrm{var}}$ may be the correct constant in (1.1) rather than $\|\mu/\pi - 1\|_{2,\pi}$. To check this conjecture, let us start from $E = \{0,1\}$. Then for this Markov chain, we have

$$Q = \begin{pmatrix} -b & b \\ a & -a \end{pmatrix},\ a,b>0, \quad P_t = e^{tQ} = \frac{1}{a+b}\begin{pmatrix} a+be^{-\lambda_1 t}, & b\big(1-e^{-\lambda_1 t}\big) \\ a\big(1-e^{-\lambda_1 t}\big), & b+ae^{-\lambda_1 t} \end{pmatrix},$$

where $\lambda_1 = a + b$. It is now easy to check that we indeed have an equality: $\|\mu P_t - \pi\|_{\mathrm{var}} = \|\mu - \pi\|_{\mathrm{var}} e^{-\lambda_1 t}$ for all $t \geqslant 0$. Everything is so nice for this trivial case (and it also indicates the difference between the present rate and the logarithmic Sobolev constant). But if we go one more step ahead, that is considering $E = \{0,1,2\}$, then the conjecture is wrong.

[2]In the reversible case, we have $\int p_s(x,y)^2\pi(dy) = \int p_s(x,y)p_s(y,x)\pi(dy) = p_{2s}(x,x) < \infty$.

Example 1.5. Take $E = \{0, 1, 2\}$ and

$$Q = \begin{pmatrix} -b_0, & b_0, & 0 \\ a_1, & -(a_1+b_1), & b_1 \\ 0, & a_2, & -a_2 \end{pmatrix}, \qquad b_0, b_1, a_1, a_2 > 0.$$

Then, by some elementary computations, we obtain

$$P_t = e^{tQ} = \Pi + \big[e^{-\lambda_1 t}C_1 - e^{-\lambda_2 t}C_2\big]/(\lambda_2 - \lambda_1),$$

where Π is the matrix having the same rows as the distribution π and

$$\begin{aligned} \lambda_1 &= 2^{-1}\big[a_1 + a_2 + b_0 + b_1 - \sqrt{(a_1 - a_2 + b_0 - b_1)^2 + 4a_1b_1}\,\big] \\ \lambda_2 &= 2^{-1}\big[a_1 + a_2 + b_0 + b_1 + \sqrt{(a_1 - a_2 + b_0 - b_1)^2 + 4a_1b_1}\,\big] \\ C_1 &= Q + \lambda_2(I - \Pi), \qquad C_2 = Q + \lambda_1(I - \Pi). \end{aligned}$$

Certainly, $\lambda_0 = 0$, λ_1 and λ_2 are the eigenvalues of $-Q$. (To check the formula of P_t, one may also need the facts that $\Pi Q = Q\Pi = 0$ and $Q^2 + (\lambda_1 + \lambda_2)Q + \lambda_1\lambda_2(I - \Pi) = 0$). Now, noticing that $\mu\Pi = \pi$ and $\pi Q = 0$, it is easy to show that

$$\mu P_t - \pi = e^{-\lambda_1 t}(\mu - \pi)\left[I + \frac{1 - e^{(\lambda_1 - \lambda_2)t}}{\lambda_2 - \lambda_1}(Q + \lambda_1 I)\right]. \tag{1.7}$$

To get a more concrete impression, take $b_0 = b_1 = 1$, $a_1 = a_2 = n^2$ and $\mu_0 = \mu_1 = 1/10$. Then, it follows from (1.7) that

$$\frac{\|\mu P_t - \pi\|_{\rm var}}{\|\mu - \pi\|_{\rm var} e^{-\lambda_1 t}} \approx \frac{\|(\mu - \pi)(Q + \lambda_2 I)\|_{\rm var}}{(\lambda_2 - \lambda_1)\|\mu - \pi\|_{\rm var}} \approx \frac{4}{9}n$$

for large enough t and n. From this, one sees that the constant $C(\mu)$ can not be $\|\mu - \pi\|_{\rm var}$, except one uses a smaller σ instead of λ_1 but then the resulting rate is not closely relative to the eigenvalues of the generator Q. Since the rate is more essential than the constant, it is natural to take $\sigma = \lambda_1$. However, we do not know at the moment how to choose a better but still simple constant $C(\mu)$ instead of $\|\mu/\pi - 1\|_{2,\pi}$. This example also indicates some critical distinction of the L^1-convergence and the L^2-convergence.

2. Proofs.

We begin this section with a simple observation.

Lemma 2.1. Under (H_1), we have $\pi \ll \lambda$.

Proof. Since π is a reversible measure and hence stationary and moreover for fixed t, $(x, y) \to p_t(x, y)$ is measurable, we have by Fubini theorem that

$$1 \geqslant \pi(A) = \int \pi(dx) \int_A \lambda(dy) p_t(x, y) = \int_A \lambda(dy)\left[\int \pi(dx) p_t(x, y)\right], \quad t > 0. \quad \square$$

Proof of Theorem 1.1. a) Recall that for each pair $\mu_k(dx) = \mu_k(x)\lambda(dx)$, $k = 1, 2$ we have

$$\|\mu_1 - \mu_2\|_{\text{var}} = 2 \sup_{A\in\mathscr{E}} |\mu_1(A) - \mu_2(A)| = \int |\mu_1(x) - \mu_2(x)|\lambda(dx).$$

b) We now show that $\pi(x)p_t(x, y) = \pi(y)p_t(y, x)$, $\lambda \times \lambda$-a.e. (x, y). For every A, $B \in \mathscr{E}$, by using the reversibility of P_t, we have

$$\int_A \lambda(dx) \int_B \lambda(dy)\pi(x)p_t(x, y) = \int_A \pi(dx)P_t(x, B) = \int_B \pi(dx)P_t(x, A).$$

Similarly,

$$\int_A \lambda(dx) \int_B \lambda(dy)\pi(y)p_t(y, x) = \int_B \pi(dy)P_t(y, A) = \int_B \pi(dx)P_t(x, A).$$

Thus, the equality

$$\iint_C \lambda(dx)\lambda(dy)\pi(x)p_t(x, y) = \iint_C \lambda(dx)\lambda(dy)\pi(y)p_t(y, x)$$

holds for all product measurable set $C = A \times B$ and hence for all $C \in \mathscr{E}$ by the monotone class theorem.

c) We now prove that $(\mu/\pi - 1)(y)\pi(y)p_t(y, x)$ is integrable with respect to $\lambda \times \lambda$. Actually, by Fubini theorem and a), we have

$$\begin{aligned}
&\int \lambda(dx) \int \lambda(dy)|\mu/\pi - 1|(y)\pi(y)p_t(y, x)\\
&\quad = \int \lambda(dy)|\mu/\pi - 1|(y)\pi(y) \int \lambda(dx)p_t(y, x)\\
&\quad = \int \lambda(dy)|\mu/\pi - 1|(y)\pi(y) = \|\mu - \pi\|_{\text{var}} \leqslant 2.
\end{aligned}$$

d) We now prove (1.4).

$$\begin{aligned}
\|\mu P_t - \pi\|_{\text{var}} &= \|\mu P_t - \pi P_t\|_{\text{var}}\\
&= \int \lambda(dx)\left| \int \mu(dy)p_t(y, x) - \int \pi(dy)p_t(y, x)\right| \quad \text{(by } a\text{))}\\
&= \int \lambda(dx)\left| \int \lambda(dy)[\mu/\pi - 1](y)\pi(y)p_t(y, x)\right|\\
&= \int \lambda(dx)\pi(x)\left| \int \lambda(dy)p_t(x, y)[\mu/\pi - 1](y)\right| \quad \text{(by } c\text{) and } b\text{))}\\
&= \int \pi(dx)\big|P_t(\mu/\pi - 1)\big|(x) = \|P_t(\mu/\pi - 1)\|_{1,\pi}.
\end{aligned}$$

We have thus proved (1.4). It should be pointed out that up to now, we do not need the condition that $\|\mu/\pi - 1\|_{2,\pi} < \infty$. It is needed only in the last step to deduce the conclusion of Theorem 1.1 (i.e. (1.5)) from (1.4). □

Proof of Theorem 1.2. Let g be an eigenfunction of λ_1 and satisfy (H_3). Replacing g with $-g$ if necessary, we may assume that g is bounded from below. Take $\mu_0(x) = \pi(x)\big(g(x)/\big[-\inf_{z\in E} g(z)\big] + 1\big)$. Noticing that $\pi(g) = 0$, it is easy to check that $\mu_0(dx) := \lambda(dx)\mu_0(x) \in \mathscr{P}$. On the other hand, since

$$\frac{d}{dt}P_t g(x) = P_t L^e g(x) = -\lambda_1 P_t g(x), \qquad x \in E,$$

it follows that

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

Combining these facts with (1.4), we obtain the assertion of Theorem 1.2. □

The remainder of this section is devoted to the proofs of Corollary 1.3 and Corollary 1.4. The compact case is easy since the function $[\mu/\pi](x)$ is bounded by the assumptions. Thus, what we need is to prove the non-compact case. The main difficulty to check (H_3) is that it uses not only the eigenvalue λ_1 but also the eigenfunction g, both of them are unknown explicitly. Fortunately, an essential part of the proofs have completed in [4] and [5]. For instance, it was proved there that the eigenfunction g with $g(0) = -1$ should be strictly increasing and belongs to $L^1(\pi)$. So the main job in the present proofs is to show that $\pi(g) = 0$. The proofs for the corollaries will be completed by the following lemmas respectively.

Lemma 2.2. Consider the birth-death process with birth rate b_i $(i \geqslant 0)$ and death rate a_i $(i \geqslant 1)$. Let g be the eigenfunction of $\lambda_1 > 0$ with $g_0 = -1$. Then, g is strictly increasing and $\pi(g) = 0$.

Proof. a) By [4; Lemma 4.2], we know that g is strictly increasing, $g \in L^1(\pi)$ and $c := \lim_{n\to\infty} b_n\mu_n\ (g_{n+1} - g_n) = -\pi(g) \geqslant 0$, where μ_n is the following sequence induced by the rates (a_i, b_i):

$$\mu_0 = 1, \qquad \mu_n = \frac{b_0 b_1 \cdots b_{n-1}}{a_1 a_2 \cdots a_n}, \qquad n \geqslant 1, \qquad \mu := \textstyle\sum_{n\geqslant 0}\mu_n.$$

One should not be confused by this sequence with the probability μ used before.

b) Set $u_i = g_{i+1} - g_i$ and $v_i = u_{i+1}/u_i$, $i \geqslant 0$. Then, it is easy to check (as in [4]) that $R_i := a_{i+1} + b_i - a_i/v_{i-1} - b_{i+1}v_i \equiv \lambda_1 > 0$ for all $i \geqslant 0$. Thus, part (1) of [4; Theorem 1.1] says that the sequence (v_i) achieves at the sharp estimate of λ_1.

c) Suppose that $c > 0$. Following the proof of [4; Lemma 2.1], define $w_i = a_i u_{i-1} - b_i u_i + c/(\mu - \mu_0)$, $i \geqslant 1$. It was proved there that w_i is strictly increasing, $w \in L^1(\pi)$ and $\sum_{i\geqslant 1}\mu_i w_i > 0$. By induction, it follows that $\sum_{j\geqslant i}\mu_j w_j > 0$ for all $i \geqslant 1$.

d) Since $w \in L^1(\pi)$, $\sum_{j\geqslant i}\mu_j w_j \to 0$ as $i \to \infty$ and moreover

$$c = \lim_{n\to\infty} b_n\mu_n u_n > 0,$$

there exists i_0 such that

$$0 < (b_i\mu_i u_i)^{-1} \sum_{j\geqslant i+1}\mu_j w_j < 1/2$$

for all $i \geqslant i_0$. Next, since

$$\sum_{j\geqslant i+1}\mu_j w_j = b_i\mu_i u_i - \frac{c}{\mu-\mu_0}\sum_{1\leqslant j\leqslant i}\mu_j, \qquad i \geqslant 0 \quad \text{(cf. [4; (2.3)])},$$

we have

$$\begin{aligned} 0 &< \min_{1\leqslant i\leqslant i_0-1} \frac{1}{b_i\mu_i u_i}\sum_{j\geqslant i+1}\mu_j w_j \\ &= 1 - \max_{1\leqslant i\leqslant i_0-1} \frac{c}{b_i\mu_i u_i(\mu-\mu_0)}\sum_{1\leqslant j\leqslant i}\mu_j \\ &=: \varepsilon < 1. \end{aligned}$$

Thus, for each $i \geqslant 1$, we have

$$I_i(w) := \frac{b_i\mu_i(w_{i+1}-w_i)}{\sum_{j\geqslant i+1}\mu_j w_j} = \frac{b_i\mu_i u_i}{\sum_{j\geqslant i+1}\mu_j w_j}R_i \geqslant (\varepsilon^{-1}\wedge 2)\lambda_1 > \lambda_1 \quad \text{(by } c\text{) and } b)).$$

e) When $i = 0$, we have

$$\begin{aligned} I_0(w) &:= b_0\left[1 + \frac{w_1}{\sum_{j\geqslant 1}\mu_j w_j}\right] \\ &= b_0\left[1 + \frac{w_1}{b_0\mu_0 u_0}\right] \\ &= R_0 + \frac{c}{u_0(\mu-\mu_0)} \\ &= \lambda_1 + \frac{c}{u_0(\mu-\mu_0)} \\ &> \lambda_1. \end{aligned}$$

f) Combining e) with d), we get $\inf_{i\geqslant 0} I_i(w) > \lambda_1$, which is a contradiction to [4; Theorem 1.1]. □

Lemma 2.3. Let g be the eigenfunction of $\lambda_1 > 0$ of the elliptic operator $L = a(x)d^2/dx^2 + b(x)d/dx$ on the half line $[x_0, \infty)$ with smooth coefficients and reflecting boundary on $x_0 \in \mathbb{R}$. That is $Lg = -\lambda_1 g$ and $g'(x_0) = 0$. Then, we have $g' \neq 0$ on (x_0, ∞) and $\pi(g) = 0$ [3].

Proof. a) The first assertion was proved in [5; Lemma 6.4]. Without loss of generality, assume that $g' > 0$ on (x_0, ∞).

b) Set $f = g'$ and $f_1 = -af' - bf$. Then, it follows from [5; Lemma 6.2] that

$$c := -\pi(g) = \lim_{x\to\infty} f(x)e^{C(x)} \geqslant 0,$$

where $C(x) = \int_{x_0}^x b/a$. Clearly, $f_1'/f = -(Lg)'/g' = \lambda_1$ on (x_0, ∞). Hence, the lower bound given by [5; (2.2)] is exact.

[3] See also Appendix to the paper [8] in this book.

c) Because $-\big(fe^C\big)' = f_1e^C/a$, we have $\int_x^\infty f_1e^C/a = f(x)e^{C(x)} - c$ and so

$$\frac{e^{-C(x)}}{f_1'(x)}\int_x^\infty \frac{f_1e^C}{a} = \frac{f(x)}{f_1'(x)} - \frac{cf(x)}{f_1'(x)}\cdot\frac{e^{-C(x)}}{f(x)} = \frac{1}{\lambda_1} - \frac{c}{\lambda_1 f(x)e^{C(x)}}, \qquad x > x_0.$$

d) Suppose that $c > 0$ and set $f_2 = c/Z + f_1$, where $Z = \int_{x_0}^\infty e^C/a$. It is rather simple to check, as was did in [5; Remark 2.2 (1)], that $f_2' > 0$ on (x_0, ∞) and $\pi(f_2) \geqslant 0$.[4] Now

$$\begin{aligned} I(f_2)(x) :&= \frac{e^{-C(x)}}{f_2'(x)}\int_x^\infty \frac{f_2e^C}{a} \\ &= \frac{e^{-C(x)}}{f_1'(x)}\int_x^\infty \frac{f_1e^C}{a} + \frac{ce^{-C(x)}}{Zf_1'(x)}\int_x^\infty \frac{e^C}{a} \\ &= \frac{1}{\lambda_1} - \frac{c}{\lambda_1 f(x)e^{C(x)}} + \frac{c}{\lambda_1 Zf(x)e^{C(x)}}\int_x^\infty \frac{e^C}{a} \qquad \text{(by } c\text{)} \\ &= \frac{1}{\lambda_1} - \frac{c}{\lambda_1 f(x)e^{C(x)}}\cdot\frac{1}{Z}\int_{x_0}^x \frac{e^C}{a}, \qquad x > x_0. \end{aligned}$$

Thus,

$$0 < I(f_2)^{-1}(x) = \lambda_1\left[1 - \frac{c}{f(x)e^{C(x)}}\cdot\frac{1}{Z}\int_{x_0}^x \frac{e^C}{a}\right]^{-1}, \qquad x > x_0. \tag{2.1}$$

Note that $\lim_{x\to\infty} f(x)e^{C(x)} = c$ and $Z = \int_{x_0}^\infty e^C/a$. From (2.1), it should be easy (and similar to the proofs d) and e) of the previous lemma) to conclude that $\inf_{x>x_0} I(f_2)^{-1}(x) > \lambda_1$, which is a contradiction to [5; Theorem 2.1].[5] $\square$

Acknowledgement. The paper was done while the author visited Dept. of Math., Univ. of Roma "La Sapienza" in October, 1996. The author would like to acknowledge the warm hospitality and the financial support of Dept., especially Professor E. Scacciatelli for his valuable discussions.

[4]Addition to the original proof. Actually, $f_2' = f_1' = \lambda_1 f = \lambda_1 g' > 0$,

$$\pi(f_2) = \frac{c}{Z} + \frac{1}{Z}\int_{x_0}^\infty \frac{f_1e^C}{a} = \frac{c}{Z} + \frac{1}{Z}\big[f(x_0)e^{C(x_0)} - c\big] = 0.$$

[5]Addition to the original proof. Set

$$A(x) = 1 - \frac{c}{f(x)e^{C(x)}}\cdot\frac{1}{Z}\int_{x_0}^x \frac{e^C}{a}.$$

Since $\lim\limits_{x\to\infty} A(x) = 0$, we have $A(x)^{-1} > 2$ for all large enough x. It suffices to consider the local region of x and then to show that $A(x_0) < 1$. For this, we have

$$\lim_{x\to x_0}\frac{c}{f(x)e^{C(x)}}\cdot\frac{1}{Z}\int_{x_0}^x \frac{e^C}{a} = \frac{c}{Z}\lim_{x\to x_0}\frac{e^{C(x)}}{(fe^C)'(x)a(x)} = \frac{c}{Z}\lim_{x\to x_0}\frac{1}{-f_1(x)} = \frac{c}{Z}\cdot\frac{1}{-\lambda_1 g(x_0)} > 0.$$

References

[1] Liggett, T. M. (1989), *Exponential L_2 convergence of attractive reversible nearest particle systems*, Ann. Probab. 17, 403-432.

[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*, Singapore, World Scientific.

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

[5] Chen, M. F. and Wang, F. Y. (1997), *Estimation of spectral gap for elliptic operators*, Trans. Amer. Math. Soc. 349, 1209–1237.

[6] Chen, M. F. and Wang, F. Y. (1997), *General formula for lower bound of the first eigenvalue on Riemannian manifolds*, Sci. Sin. 27:1, 34–42 (Chinese Edition); 40:4, 384–394 (English Edition).

[7] Chen, M. F. and Wang, F. Y. (1995), *Estimation of the first eigenvalue of second order elliptic operators*, J. Funct. Anal. 131:2, 345–363.

[8] Diaconis, P. and Stroock, D. W. (1991), *Geometric bounds for eigenvalues of Markov chains*, Ann. Appl. Prob. 1:1, 36–61.

[9] Diaconis, P. and Saloff-coste, L. (1996), *Nash inequality for finite Markov chains*, J. Theor. Prob. 9:2, 459–510.

[10] Rosenthal, J. S. (1996), *Markov chain convergence: From finite to infinite*, Stoch. Proc. Appl. 62:1, 55–72.

[11] Saloff-Coste, L. (1994), *Convergence to equilibrium and Logarithmic Sobolev constant on manifolds with Ricci curvature bounded below*, Coll. Math., 109–121.

[12] Sinclair, A. J. and Jerrum, M. R. (1989), *Approximate counting, uniform generation, and rapidly mixing Markov chains*, Inform. and Comput.82, 93–133.

[13] Wang, Y. Z. (1996), *Convergence rate in total variation for diffusion processes*, preprint.

Received March 10, 1997. Accepted June 28, 1997.

Appendix: Order-three Semigroup (unpublished note).

In this note, we study the semigroup $\{P_t\}_{t\geqslant 0}$ generated by the 3×3 Q-matrix

$$Q=\begin{pmatrix} -q_{01}-q_{02} & q_{01} & q_{02} \\ q_{10} & -q_{10}-q_{12} & q_{12} \\ q_{20} & q_{21} & -q_{20}-q_{21} \end{pmatrix}.$$

Its eigenvalues are as follows.

$$\lambda_0=0, \qquad \lambda_1=\frac{1}{2}\big(-\Theta+\sqrt{\Delta}\,\big), \qquad \lambda_2=\frac{1}{2}\big(-\Theta-\sqrt{\Delta}\,\big),$$

where

$$\begin{aligned}
\Theta &= \sum_{i\neq j} q_{ij} = q_{01}+q_{02}+q_{10}+q_{12}+q_{20}+q_{21},\\
\Delta &= \Theta^2-4Z\\
&= (q_{10}-q_{21}+q_{01}-q_{12}+q_{02}+q_{20})^2\\
&\quad -4\big(q_{10}(q_{02}-q_{12}+q_{20})+q_{20}(q_{01}-q_{21})\big),\\
Z &= q_{12}q_{20}+q_{10}(q_{20}+q_{21})+q_{02}(q_{10}+q_{12}+q_{21})+q_{01}(q_{12}+q_{20}+q_{21}).
\end{aligned}$$

Since $Z\geqslant 0$, we have $\Theta^2\geqslant\Delta$. Clearly, λ_1 and λ_2 are real when $\Delta\geqslant 0$. Otherwise, λ_1 and λ_2 are complex imaginary numbers.

We assume that $Q\neq 0$. Otherwise, the problem is trivial. Then $\Theta>0$. We also assume that $Z>0$, which holds whenever Q is irreducible. Otherwise $\lambda_1=0$ and Q is degenerated since there exists at least one zero row. Under these assumptions, we have $\lambda_1,\ \lambda_2\neq 0$. Moreover, the Q-matrix has uniquely an invariant probability measure as follows.

$$\begin{aligned}
\pi_0 &= [q_{12}q_{20}+q_{10}(q_{20}+q_{21})]/Z,\\
\pi_1 &= [q_{02}q_{21}+q_{01}(q_{20}+q_{21})]/Z,\\
\pi_2 &= [q_{01}q_{12}+q_{02}(q_{10}+q_{12})]/Z.
\end{aligned}$$

Proposition A.1. Let $Q\neq 0$ and $Z>0$. Then the semigroup $\{P_t=e^{tQ}\}_{t\geqslant 0}$ takes the following form.

(1) If $\Delta\neq 0$ (in particular, if $\Delta>0$), then

$$\begin{aligned}
P_t &= \Pi+\frac{1}{\lambda_1-\lambda_2}\Big[e^{\lambda_1 t}\big(Q-\lambda_2(I-\Pi)\big)-e^{\lambda_2 t}\big(Q-\lambda_1(I-\Pi)\big)\Big]\\
&= \Pi+\frac{e^{\lambda_1 t}-e^{\lambda_2 t}}{\lambda_1-\lambda_2}Q+\frac{\lambda_1 e^{\lambda_2 t}-\lambda_2 e^{\lambda_1 t}}{\lambda_1-\lambda_2}(I-\Pi),
\end{aligned}$$

where $\Pi=(\pi_{ij})$: $\pi_{ij}=\pi_j$ for all i and j.

(2) If $\Delta=0$, then $\lambda_1=\lambda_2<0$ and

$$\begin{aligned}
P_t &= \Pi+\big(I-\Pi+t(Q-\lambda_1(I-\Pi))\big)e^{\lambda_1 t}\\
&= \Pi+\big(tQ+(1-\lambda_1 t)(I-\Pi)\big)e^{\lambda_1 t}.
\end{aligned}$$

(3) If $\Delta < 0$, then

$$P_t = \Pi + \left[\cos(\beta t)(I-\Pi) - \frac{1}{\beta}\sin(\beta t)\Big(\alpha(I-\Pi) - Q\Big)\right]e^{\alpha t}$$
$$= \Pi + \left[\frac{1}{\beta}\sin(\beta t)Q + \left(\cos(\beta t) - \frac{\alpha}{\beta}\sin(\beta t)\right)(I-\Pi)\right]e^{\alpha t},$$

where

$$\alpha = \mathrm{Re}(\lambda_1) = -\frac{1}{2}\Theta, \qquad \beta = \mathrm{Im}(\lambda_1) = \frac{\sqrt{-\Delta}}{2}.$$

We remark that the first case is more essential since of which, the third one is a reorganization by eliminating the imaginary parts and the second one is a limit as $\lambda_2 \uparrow \lambda_1$.

Proof of Proposition A.1. Since $\det(\lambda I - Q) = \lambda(\lambda - \lambda_1)(\lambda - \lambda_2)$, by the Cayley-Hamilton theorem, we have

$$Q(Q - \lambda_1 I)(Q - \lambda_2 I) = 0.$$

This means that the element $a_{ij}^{(n)}$ of Q^n should satisfy the following recurrence equation.

$$a_{ij}^{(n+2)} - (\lambda_1 + \lambda_2)a_{ij}^{(n+1)} + \lambda_1\lambda_2 a_{ij}^{(n)} = 0, \qquad n \geqslant 1.$$

When $\Delta \neq 0$, then $\lambda_1 \neq \lambda_2$ and the general solution of $a_{ij}^{(n)}$ is given by

$$a_{ij}^{(n)} = c_{ij}^{(1)}\lambda_1^n + c_{ij}^{(2)}\lambda_2^n, \qquad n \geqslant 1$$

for some constants $C_1 = \big(c_{ij}^{(1)}\big)$ and $C_2 = \big(c_{ij}^{(2)}\big)$. That is

$$Q^n = C_1\lambda_1^n + C_2\lambda_2^n, \qquad n \geqslant 1.$$

If we adopt the convention $0^0 = 1$, then we can write

$$Q^n = \lambda_0^n C_0 + \lambda_1^n C_1 + \lambda_2^n C_2, \qquad n \geqslant 0 \tag{A.1}$$

with $C_0 = I - C_1 - C_2$. Now, we have

$$P_t = e^{tQ} = \sum_{n=0}^{\infty}\frac{t^n Q^n}{n!} = C_0 e^{\lambda_0 t} + C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t} = C_0 + C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}.$$

That is

$$P_t = C_0 + C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}. \tag{A.2}$$

Case 1: $\Delta \neq 0$.

We have $\lambda_1 \neq \lambda_2$. In the case that $\Delta > 0$, we indeed have $\lambda_2 < \lambda_1 < 0$. Otherwise, we have $\mathrm{Re}\,(\lambda_1) = \mathrm{Re}\,(\lambda_2) = -\Theta/2 < 0$. Letting $t \to \infty$ in (A.2), we obtain $\Pi = C_0$. Thus,

$$P_t = \Pi + C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}.$$

Now, the initial conditions for (A.1) give us

$$\begin{cases} I = \Pi + C_1 + C_2 \\ Q = \lambda_1 C_1 + \lambda_2 C_2. \end{cases}$$

From which, we obtain

$$C_1 = \frac{1}{\lambda_1 - \lambda_2}\big(Q - \lambda_2(I - \Pi)\big),$$
$$C_2 = -\frac{1}{\lambda_1 - \lambda_2}\big(Q - \lambda_1(I - \Pi)\big).$$

Combining this with (A.2) and noting that $C_0 = \Pi$, we obtain part (1) of the proposition.

Case 2: $\Delta = 0$.

We have $\lambda_1 = \lambda_2 < 0$. Hence

$$Q^n = C_0\lambda_0^n + (C_1 + nC_2)\lambda_1^n, \qquad n \geqslant 0, \tag{A.3}$$
$$P_t = \Pi + (C_1 + \lambda_1 t C_2)e^{\lambda_1 t}. \tag{A.4}$$

By (A.3), we have

$$\begin{cases} I = \Pi + C_1 \\ Q = \lambda_1(C_1 + C_2). \end{cases}$$

From which, we get

$$C_1 = I - \Pi,$$
$$C_2 = Q/\lambda_1 - (I - \Pi).$$

The second assertion now follows from (A.4).

Case 3: $\Delta < 0$.

We have $\lambda_1 = \alpha + i\beta$ and $\lambda_2 = \alpha - i\beta$ with $\alpha < 0$ and $\beta > 0$.

The assertion follows from part (1) by eliminating the imaginary parts. Actually, we have $\lambda_1 - \lambda_2 = 2i\beta$. Since the last two terms in the first formula of P_t in part (1) are conjugate, it suffices to compute

$$\mathrm{Re}\bigg(\frac{Q - \lambda_2(I - \Pi)}{\lambda_1 - \lambda_2}e^{\lambda_1 t}\bigg) = \mathrm{Re}\bigg(\frac{Q - (\alpha - i\beta)(I - \Pi)}{2i\beta}e^{i\beta t}\bigg)e^{\alpha t}.$$

Because

$$\frac{e^{i\beta t}}{2i\beta} = \frac{\cos(\beta t) + i\sin(\beta t)}{2i\beta} = \frac{\sin(\beta t) - i\cos(\beta t)}{2\beta},$$

we have

$$\begin{aligned}
&2\mathrm{Re}\Big(\frac{Q-(\alpha-i\beta)(I-\Pi)}{2i\beta}e^{i\beta t}\Big)\\
&\quad=\frac{1}{\beta}\mathrm{Re}\Big(\big(\sin(\beta t)-i\cos(\beta t)\big)\big(Q-\alpha(I-\Pi)+i\beta(I-\Pi)\big)\Big)\\
&\quad=\frac{1}{\beta}\Big(\sin(\beta t)\big(Q-\alpha(I-\Pi)\big)+\beta\cos(\beta t)(I-\Pi)\Big).
\end{aligned}$$

Applying the first formula in part (1) of the proposition, we obtain the required assertion. □

Example A.2. Take

$$Q=\begin{pmatrix}-1/2 & 1/2 & 0\\ 0 & -1 & 1\\ 1 & 0 & -1\end{pmatrix}$$

Then we have $\lambda_1=-5/4+\sqrt{7}\,i/4$, $\lambda_2=-5/4-\sqrt{7}\,i/4$, $\pi_0=1/2$, $\pi_1=\pi_2=1/4$. Hence

$$\Pi=\begin{pmatrix}1/2 & 1/4 & 1/4\\ 1/2 & 1/4 & 1/4\\ 1/2 & 1/4 & 1/4\end{pmatrix}\qquad I-\Pi=\begin{pmatrix}1/2 & -1/4 & -1/4\\ -1/2 & 3/4 & -1/4\\ -1/2 & -1/4 & 3/4\end{pmatrix}$$

and so by part (3) of Proposition A.1,

$$\begin{aligned}
P_t=&\begin{pmatrix}1/2 & 1/4 & 1/4\\ 1/2 & 1/4 & 1/4\\ 1/2 & 1/4 & 1/4\end{pmatrix}+\Bigg[\frac{1}{\sqrt{7}}\sin\big(\sqrt{7}\,t\big)\begin{pmatrix}-1/2 & 1/2 & 0\\ 0 & -1 & 1\\ 1 & 0 & -1\end{pmatrix}\\
&+\Big(\cos\big(\sqrt{7}\,t\big)+\frac{5}{\sqrt{7}}\sin\big(\sqrt{7}\,t\big)\Big)\begin{pmatrix}1/2 & -1/4 & -1/4\\ -1/2 & 3/4 & -1/4\\ -1/2 & -1/4 & 3/4\end{pmatrix}\Bigg]e^{-5t/4}.
\end{aligned}$$

Clearly, the exponentially ergodic convergence rate is $5/4$ but the L^2-spectral gap is 1. Note that the equation

$$Qf=-\frac{5}{4}f$$

has no nontrivial solution f since $-5/4$ is not an eigenvalue of Q. Besides, in the exponentially ergodic criterion, one needs $\lambda<q_i$ for all i. In the present case, $\min_i q_i=1/2$ is smaller than $5/4$. Actually, if we regard $\{0\}$ as an forbidden set, and solve the equation

$$\begin{cases}\sum_j q_{1j}y_j=-\lambda y_1-1\\ \sum_j q_{2j}y_j=-\lambda y_2-1,\end{cases}$$

that is,

$$\begin{cases}-y_1+y_2=-\lambda y_1-1\\ y_0-y_2=-\lambda y_2-1,\end{cases}$$

we obtain

$$y_1 = \frac{2-\lambda+y_0}{(1-\lambda)^2}, \qquad y_2 = \frac{1+y_0}{1-\lambda}.$$

Since $y_0 \geqslant 0$, in order for $y_2 \geqslant 0$, it is clearly necessary that $\lambda < 1 = q_1 = q_2$. If we replace $\{0\}$ by $\{2\}$, then the condition "$\lambda < 1/2 = \min_i q_i$" becomes necessary. This shows that the parameter λ here and the spectral gap are all smaller than the rate $5/4$.

Department of Mathematics, Beijing Normal University, Beijing 100875, The People's Republic of China.

Ann. Prob. 28:1, (2000), 235–257

CHEEGER'S INEQUALITIES FOR GENERAL SYMMETRIC FORMS AND EXISTENCE CRITERIA FOR SPECTRAL GAP

Mu-Fa Chen and Feng-Yu Wang

(Beijing Normal University)
Received May 1998;
Revised July 1999

Abstract. In this paper, some new forms of Cheeger's inequalities are established for general (maybe unbounded) symmetric forms (Theorems 1.1 and 1.2), the resulting estimates improve and extend the ones obtained by Lawler and Sokal for bounded jump processes. Furthermore, some existence criteria for spectral gap of general symmetric forms or general reversible Markov processes are presented (Theorems 1.4 and 3.1), based on Cheeger's inequalities and a relationship between the spectral gap and the first Dirichlet and Neumann eigenvalues on local region.

1. Introduction.

Cheeger's inequalities [2] are well known and widely used in geometric analysis, they provide a practical way to estimate the first eigenvalue of Laplacian in terms of volumes. These inequalities were established for bounded jump processes by Lawler and Sokal [8] (in which a detailed comment on the earlier study and references are included). The first aim of this paper is to establish the inequalities for general (maybe unbounded) symmetric forms.

Let $(E, \mathscr{E}, \pi)$ be a probability space satisfying $\{(x,x) : x \in E\} \in \mathscr{E} \times \mathscr{E}$. Consider the symmetric form D with domain $\mathscr{D}(D)$,

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

2000 *Mathematics Subject Classification.* 60J25, 60J75, 47A75.

Key words and phrases. Cheeger's inequality, spectral gap, Neumann and Dirichlet eigenvalue, jump process.

Supported in part by NSFC (No. 19631060), Qiu Shi Sci. and Tech. Found., DPFIHE, MCSEC and MCMCAS.

where J and K are nonnegative and J is symmetric: $J(\mathrm{d}x,\mathrm{d}y) = J(\mathrm{d}y,\mathrm{d}x)$. Without loss of generality, we assume that $J(\{(x,x)\colon x\in E\}) = 0$.

We are interested in the following two quantities:

$$\lambda_0 = \inf\{D(f,f)\colon\ \pi(f^2) = 1\}, \tag{1.1}$$

$$\lambda_1 = \inf\{D(f,f)\colon\ \pi(f) = 0,\ \pi(f^2) = 1\}. \tag{1.2}$$

We remark that in these definitions, the usual condition "$f\in\mathscr{D}(D)$" is not needed since $D(f,f) = \infty$ for all $f\in L^2(\pi)\setminus\mathscr{D}(D)$. We do not even assume in some cases the density of $\mathscr{D}(D)$ in $L^2(\pi)$. In what follows, whenever λ_1 is considered, the killing measure $K(\mathrm{d}x)$ is set to zero. In this case, we have $\lambda_0 = 0$ and λ_1 is known as the spectral gap of the symmetric form $(D,\mathscr{D}(D))$.

Define Cheeger's constants as follows:

$$h = \inf_{\pi(A)>0}\frac{J(A\times A^c)+K(A)}{\pi(A)}, \tag{1.3}$$

$$k = \inf_{\pi(A)\in(0,1)}\frac{J(A\times A^c)}{\pi(A)\pi(A^c)}, \tag{1.4}$$

$$k' = \inf_{\pi(A)\in(0,1/2]}\frac{J(A\times A^c)}{\pi(A)} = \inf_{\pi(A)\in(0,1)}\frac{J(A\times A^c)}{\pi(A)\wedge\pi(A^c)}, \tag{1.5}$$

where $a\wedge b = \min\{a,b\}$. Clearly,

$$k/2 \leqslant k' \leqslant k$$

and it is easy to see that k' can be varied over whole $(k/2,k)$. For instance, take $E=\{0,1\}$, $K=0$, $J(\{i\}\times\{j\}) = 1$ for $i\neq j$ and $\pi(0) = p\leqslant 1/2$, $\pi(1) = 1-p$. Then $k'/k = 1-p$.

Recall that for a given reversible jump process, we have a q-pair $(q(x), q(x,\mathrm{d}y))$: $q(x,E)\leqslant q(x)\leqslant\infty$ for all $x\in E$. Throughout the paper, we assume that $q(x)<\infty$ for all $x\in E$. The reversibility simply means that the measure $\pi(\mathrm{d}x)q(x,\mathrm{d}y)$ is symmetric, which gives us automatically a measure J. Then the killing measure is given by $K(\mathrm{d}x) = \pi(\mathrm{d}x)d(x)$, where $d(x) = q(x) - q(x,E)$ is called the non-conservative quantity in the context of jump processes. A jump process is called bounded if $\sup_{x\in E} q(x)<\infty$. In this case [or more generally, if $\|J(\cdot,E)+K\|_{\mathrm{op}}<\infty$, where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm from $L^1_+(\pi) := \{f\in L^1(\pi)\colon\ f\geqslant 0\}$ to $\mathbb{R}_+$], for the corresponding form, we have $\mathscr{D}(D) = L^2(\pi)$. For more details, refer to [3].

Theorem [Lawler & Sokal(1988)]. Take $J(\mathrm{d}x,\mathrm{d}y) = \pi(\mathrm{d}x)q(x,\mathrm{d}y)$ and suppose that $\|J(\cdot,E)+K/2\|_{\mathrm{op}}\leqslant M<\infty$. Then we have

$$h\geqslant\lambda_0\geqslant\frac{h^2}{2M}. \tag{1.6}$$

Next, if additionally $K=0$, then

$$k\geqslant\lambda_1\geqslant\max\left\{\frac{\kappa k^2}{8M},\frac{k'^2}{2M}\right\}, \tag{1.7}$$

where

$$\kappa = \inf_{X,Y} \sup_{c\in\mathbb{R}} \frac{(\mathbb{E}|(X+c)^2 - (Y+c)^2|)^2}{1+c^2} \geqslant 1,$$

the infimum is taken over all i.i.d. random variables X and Y with $\mathbb{E}X = 0$ and $\mathbb{E}X^2 = 1$.

In what follows, we consider directly the general symmetric measure J whenever possible. In other words, we do not require the existence of a kernel of a modification of $J(\mathrm{d}x,\cdot)/\pi(\mathrm{d}x)$, for which some extra conditions on $(E,\mathscr{E})$ are needed.

We now turn to discuss our general setup. Note that the lower bounds given in (1.6) and (1.7) decrease to zero as $M \uparrow \infty$. So the results would lose their meaning if we go directly from the bounded case to unbounded forms. More seriously, when we adopt a general approximation procedure to reduce the unbounded case to the bounded one (cf. [3], Theorem 9.12), the lower bounds given above usually vanish as we go to the limit. To overcome the difficulty, one needs some trick. Here we propose a comparison technique, that is, comparing the original form with some other forms introduced below.

Take and fix a nonnegative, symmetric function $r \in \mathscr{E}\times\mathscr{E}$ and a nonnegative function $s \in \mathscr{E}$ such that

$$\|J^{(1)}(\cdot,E) + K^{(1)}\|_{\mathrm{op}} \leqslant 1, \qquad L^1_+(\pi) \to \mathbb{R}_+, \tag{1.8}$$

where

$$J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y) = I_{\{r(x,y)^\alpha>0\}}\frac{J(\mathrm{d}x,\mathrm{d}y)}{r(x,y)^\alpha}, \quad K^{(\alpha)}(\mathrm{d}x) = I_{\{s(x)^\alpha>0\}}\frac{K(\mathrm{d}x)}{s(x)^\alpha}, \qquad \alpha\geqslant 0.$$

Throughout the paper, we adopt the convention that $r^0 = 1$ and $s^0 = 1$ for $r, s \geqslant 0$. For jump processes, one may simply choose

$$r(x,y) = q(x)\vee q(y) = \max\{q(x), q(y)\} \quad \text{and} \quad s(x) = q(x).$$

We remark that when $\alpha < 1$, the operator $J^{(\alpha)}(\cdot,E) + K^{(\alpha)}$ from $L^1_+(\pi)$ to $\mathbb{R}_+$ may no longer be bounded. Correspondingly, we have symmetric forms $D^{(\alpha)}$ defined by $(J^{(\alpha)}, K^{(\alpha)})$. Therefore, with respect to the form $D^{(\alpha)}$, according to (1.1)—(1.5), we can define $\lambda_0^{(\alpha)}$, $\lambda_1^{(\alpha)}$ and Cheeger's constants $h^{(\alpha)}$, $k^{(\alpha)}$ and $k^{(\alpha)'}$ ($\alpha \geqslant 0$). However, in what follows, we need only three cases, $\alpha = 0$, $1/2$ and 1. When $\alpha = 0$, we return to the original form and so the superscript "(α)" is omitted from our notations.

The next two results are our new forms of Cheeger's inequalities.

Theorem 1.1. Suppose that (1.8) holds. We have

$$\lambda_0 \geqslant \frac{{h^{(1/2)}}^2}{2-\lambda_0^{(1)}} \geqslant \frac{{h^{(1/2)}}^2}{1+\sqrt{1-{h^{(1)}}^2}}. \tag{1.9}$$

Theorem 1.2. Let $K = 0$ and (1.8) hold. Then, we have

$$\lambda_1 \geqslant \left(\frac{k^{(1/2)}}{\sqrt{2} + \sqrt{2 - \lambda_1^{(1)}}} \right)^2, \tag{1.10}$$

$$\lambda_1 \geqslant \frac{{k^{(1/2)}}'^2}{1 + \sqrt{1 - {k^{(1)}}'^2}}. \tag{1.11}$$

When $\|J(\cdot, E)+K\|_{\rm op} \leqslant M < \infty$, the simplest choice of r and s are $r(x, y) \equiv M$ and $s(x) \equiv M$. Then, (1.8) holds and moreover $h^{(1/2)} = h/\sqrt{M}$, ${k^{(1/2)}}' = k'/\sqrt{M}$, $h^{(1)} = h/M$ and ${k^{(1)}}' = k'/M$. Hence, by (1.9) and (1.11), we get

$$\lambda_0 \geqslant M\big(1 - \sqrt{1 - h^2/M^2}\,\big) = \frac{h^2}{M\big(1 + \sqrt{1 - h^2/M^2}\,\big)} \in \left[\frac{h^2}{2M}, \frac{h^2}{M}\right].$$

and

$$\lambda_1 \geqslant M\Big(1 - \sqrt{1 - {k'}^2/M^2}\,\Big) = \frac{{k'}^2}{M\big(1 + \sqrt{1 - {k'}^2/M^2}\,\big)} \in \left[\frac{{k'}^2}{2M}, \frac{{k'}^2}{M}\right]. \tag{1.12}$$

Therefore, for the lower bounds, (1.9) improves the second part of (1.6) and (1.11) improves the second part of (1.7). More essentially, the lower bound (1.11) is often good enough so that the approximation procedure ([3], Theorem 9.12) mentioned above becomes practical. However, we will not go in this direction. In the context of Markov chains on finite graphs, (1.12) was obtained before by Chung [6]. Applying (1.12) to $J^{(1)}$, we get

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

From this and (1.10), we obtain

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

which is indeed controlled by (1.11) since $k^{(\alpha)} \leqslant 2{k^{(\alpha)}}'$. This means that (1.11) is usually more practical than (1.10) except a good lower bound of $\lambda_1^{(1)}$ is known in advance. However, (1.10) and (1.11) are not comparable even in the case of $E = \{0, 1\}$. See also the discussion in the second paragraph below Lemma 2.2.

In view of Theorem 1.2, we have $\lambda_1 > 0$ whenever $k^{(1/2)} > 0$. We now study some more explicit conditions for the Cheeger's constants appearing in Theorem 1.2 to be positive. To state the result, we should use the operators corresponding to the forms. For a jump process, the operator corresponding to $(D^{(\alpha)}, \mathscr{D}(D^{(\alpha)}))$ can be expressed by the following simple form

$$\Omega^{(\alpha)} f(x) = \int I_{[r(x,y)^\alpha > 0]} \frac{q(x, {\rm d}y)}{r(x, y)^\alpha} [f(y) - f(x)] - I_{[s(x)^\alpha > 0]} \frac{d(x)}{s(x)^\alpha} f(x).$$

Next, we need some local quantities of λ_0 and λ_1. First, for $B \in \mathscr{E}$ with $\pi(B) \in (0,1)$, let $\lambda_1^{(\alpha)}(B)$ and $k^{(\alpha)}(B)$ be defined by (1.2) and (1.4) with E, π and D replaced, respectively, by B, $\pi^B := \pi(\cdot \cap B)/\pi(B)$ and

$$D_B^{(\alpha)}(f,f) = \frac{1}{2}\int_{B\times B} J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y)(f(y)-f(x))^2. \tag{1.13}$$

Second, define

$$\lambda_0^{(\alpha)}(B) = \inf\big\{D^{(\alpha)}(f,f)\colon\ \pi(f^2)=1,\ f|_{B^c}=0\big\}.$$

As usual, we call $\lambda_0^{(\alpha)}(B)$ and $\lambda_1^{(\alpha)}(B)$, respectively, the (generalized) first Dirichlet and Neumann eigenvalue on B. It is a simple matter to check that as in (1.7), $k^{(\alpha)}(B) \geqslant \lambda_1^{(\alpha)}(B)$.

For $A \in \mathscr{E}$, put $M_A^{(\alpha)} = (\operatorname{ess\,sup}_\pi)_A J^{(\alpha)}(\mathrm{d}x, A^c)/\pi(\mathrm{d}x)$, where $\operatorname{ess\,sup}_\pi$ denotes the essential supremum with respect to π.

Theorem 1.3. Let $K = 0$. Given $\alpha \geqslant 0$ and $B \in \mathscr{E}$ with $\pi(B) > 1/2$, suppose that there exist a function φ with

$$\delta_1(\varphi) := \operatorname{ess\,sup}_{J^{(\alpha)}}|\varphi(x)-\varphi(y)| < \infty$$

and a symmetric operator $\big(\Omega^{(\alpha)}, \mathscr{D}\big(\Omega^{(\alpha)}\big)\big)$ corresponding to the form $\big(D^{(\alpha)}, \mathscr{D}\big(D^{(\alpha)}\big)\big)$ such that $\mathscr{D}\big(\Omega^{(\alpha)}\big) \supset \{I_A\colon\ A \in \mathscr{E},\ A \subset B\}$ and $\gamma_{B^c} := -\sup_{B^c}\Omega^{(\alpha)}\varphi > 0$. Then, we have

$$k^{(\alpha)} \geqslant {k^{(\alpha)}}' \geqslant \frac{k^{(\alpha)}(B)\,\gamma_{B^c}\,[2\pi(B)-1]}{k^{(\alpha)}(B)\,\delta_1(\varphi)\,[2\pi(B)-1] + 2\pi(B)^2\big[\delta_1(\varphi)M_B^{(\alpha)} + \gamma_{B^c}\big]}.$$

Usually, for locally compact E, we have $k^{(\alpha)}(B) > 0$ and $M_B^{(\alpha)} < \infty$ for all compact B. Then the result means that ${k^{(\alpha)}}' > 0$ provided $\delta_1(\varphi) < \infty$ and $\gamma_{B^c} > 0$ for large enough B.

Up to now, we have discussed the lower bound of λ_1 by using the Cheeger's constants. However, Theorem 1.3 is indeed a modification of the second approach we are going to study, that is, estimating λ_1 in terms of local λ_0 and λ_1 on subsets of E. The last method has been used recently in the context of diffusions by Wang [10] and is extended here to general reversible processes. The details of the next two results for the general situation are delayed to Section 3. Here, we restrict ourselves to the symmetric forms introduced above.

This is the place to state our first criterion for $\lambda_1 > 0$.

Theorem 1.4. Let $K = 0$. Then for any $A \subset B$ with $0 < \pi(A), \pi(B) < 1$, we have

$$\frac{\lambda_0(A^c)}{\pi(A)} \geqslant \lambda_1 \geqslant \frac{\lambda_1(B)[\lambda_0(A^c)\pi(B) - 2M_A\pi(B^c)]}{2\lambda_1(B) + \pi(B)^2[\lambda_0(A^c) + 2M_A]}. \tag{1.14}$$

As we mentioned before, usually, $\lambda_1(B) > 0$ for all compact B. Hence the result means 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 (1.14) becomes positive.

Finally, we present an upper bound of λ_1 which provides us a necessary condition for $\lambda_1 > 0$ and can qualitatively be sharp as illustrated by Example 4.5. For some related works, refer to [1] and references therein.

Theorem 1.5. Let $K = 0$, $r > 0$, J-a.e. and (1.8) hold. If there exists $\varphi \geqslant 0$ such that

$$0 < \delta_2(\varphi) := \operatorname{ess\,sup}_J |\varphi(x) - \varphi(y)|^2 r(x,y) < \infty,$$

then

$$\lambda_1 \leqslant \frac{\delta_2(\varphi)}{4} \inf\Big\{\varepsilon^2 \colon \varepsilon \geqslant 0,\, \pi\big(e^{\varepsilon\varphi}\big) = \infty\Big\}.$$

Consequently, $\lambda_1 = 0$ if there exists $\varphi \geqslant 0$ with $0 < \delta_2(\varphi) < \infty$ such that $\pi\big(e^{\varepsilon\varphi}\big) = \infty$ for all $\varepsilon > 0$. In particular, when $J(\mathrm{d}x, \mathrm{d}y) = \pi(\mathrm{d}x)q(x, \mathrm{d}y)$, $\delta_2(\varphi)$ can be replaced by

$$\delta_2'(\varphi) := \operatorname{ess\,sup}_\pi \int |\varphi(x) - \varphi(y)|^2 q(x, \mathrm{d}y) < \infty,$$

without using the function r and (1.8).

To have a test for the new forms of Cheeger's constants, we introduce the following result.

Corollary 1.6. Let $J(\mathrm{d}x, \mathrm{d}y) = j(x,y)\pi(\mathrm{d}x)\pi(\mathrm{d}y)$ for some symmetric function $j(x,y)$ having the properties: $j(x,x) = 0$ and $j(x) := \int j(x,y)\pi(\mathrm{d}y) < \infty$ for all $x \in E$. Take $r(x,y) = j(x) \vee j(y)$. Then

$$k^{(\alpha)'} \geqslant \frac{1}{2} \inf_{x \neq y} \frac{j(x,y)}{[j(x) \vee j(y)]^\alpha}. \tag{1.15}$$

Proof. Denote by $C^{(\alpha)}$ the right-hand side of (1.15). Note that

$$\begin{aligned} \frac{J^{(\alpha)}(A \times A^c)}{\pi(A)} &= \frac{1}{\pi(A)} \int_{A \times A^c} \pi(\mathrm{d}x)\pi(\mathrm{d}y) \frac{j(x,y)}{[j(x) \vee j(y)]^\alpha} \\ &\geqslant \inf_{x \neq y} \frac{j(x,y)}{[j(x) \vee j(y)]^\alpha} \pi(A^c) \\ &= 2C^{(\alpha)}\pi(A^c). \end{aligned}$$

Hence

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

as required. □

The corollary shows that our results are meaningful in a very general setup. Here are two more explicit examples.

(1) Let $j(x,y) = 1$ for $x \neq y$ and $j(x,x) = 0$. Then, by (1.15), we have $k^{(\alpha)'} \geqslant 1/2$. Hence $\lambda_1 \geqslant 1/2(2 + \sqrt{3})$ by (1.11). The precise value of λ_1 is equal to 1.

(2) Let $E = \mathbb{Z}$ and $j(x,y) = |x^2 - y^2|$. Suppose that $c := \pi(x^2) < \infty$. Then $j(x) \leqslant x^2 + c$ for all x and

$$k^{(1/2)'} \geqslant \frac{1}{2} \inf_{x \neq y} \frac{|x| + |y|}{\sqrt{x^2 + y^2 + c}} \geqslant \frac{1}{2\sqrt{c+1}}.$$

Hence

$$\lambda_1 \geqslant \frac{1}{2} k^{(1/2)'2} \geqslant \frac{1}{8(c+1)}.$$

Certainly, the estimate (1.15) is very rough. However, Theorems 1.1 and 1.2 can actually be sharp as illustrated by Examples 4.6 and 4.7 in Section 4.

We mention that the study on the leading eigenvalue of a bounded integral operator is indeed included in our general setup. Consider the operator P on $L^2(\pi)$: $Pf(x) = \int p(x, \mathrm{d}y) f(y)$, generated by an arbitrary nonnegative kernel $p(x, \mathrm{d}y)$ with $M := \sup_x p(x, E) < \infty$. Let $\pi(\mathrm{d}x)p(x, \mathrm{d}y)$ be symmetric for a moment. Clearly, the spectrum of P on $L^2(\pi)$ is determined by that of $M - P$. Note that

$$\big\langle f,\, (M-P)f\big\rangle_\pi = \frac{1}{2}\int \pi(\mathrm{d}x)p(x,\mathrm{d}y)\big[f(x)-f(y)\big]^2 + \int \pi(\mathrm{d}x)\big[M-p(x,E)\big]f(x)^2.$$

Thus, the largest (non-trivial) eigenvalue of the integral operator P can be deduced from λ_0 or λ_1 treated in the paper. Finally, by using a symmetrizing procedure, all the results presented here can be extended to the nonsymmetric forms. Refer to [3], Chapter 9, or [8], for instance.

The remainder of the paper is organized as follows. Section 2 is devoted to the proofs of Theorems 1.1—1.3. At the end of the section, a different approach for handling unbounded symmetric forms is presented. A general existence criterion for spectral gap is presented in Section 3, which also contains the proofs of Theorems 1.4 and 1.5. All the results concerning the spectral gap are illustrated by Markov chains in the last section.

2. Proofs of Theorems 1.1—1.3.

We begin this section with the functional representation of Cheeger's constants. The proof is essentially the same as in [8] and [9], Section 3.3, for the bounded situation and hence omitted.

Lemma 2.1. For every $\alpha \geqslant 0$, we have

$$h^{(\alpha)} = \inf\left\{\frac{1}{2}\int J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y)|f(x)-f(y)| + K^{(\alpha)}(f)\colon\ f\geqslant 0,\ \pi(f)=1\right\},$$

$$k^{(\alpha)} = \inf\left\{\int J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y)|f(x)-f(y)|\colon f\in L^1_+(\pi),\ \int\pi(\mathrm{d}x)\pi(\mathrm{d}y)|f(x)-f(y)|=1\right\}$$

$$= \inf\left\{\int J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y)|f(x)-f(y)|\colon\ f\in L^1_+(\pi),\ \pi(|f-\pi(f)|)=1\right\},$$

$$k^{(\alpha)'} = \inf\left\{\frac{1}{2}\int J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y)|f(x)-f(y)|\colon\ f\in L^1_+(\pi),\ \min_{c\in\mathbb{R}}\pi(|f-c|)=1\right\}.$$

Proof of Theorem 1.1. The idea of the proof is based on [8].

Let $E^* = E\cup\{\infty\}$. For any $f\in\mathscr{E}$, define f^* on E^* by setting $f^* = fI_E$. Next, define $J^{*(\alpha)}$ on $E^*\times E^*$ by

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

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

$$\int J^{(\alpha)}(\mathrm{d}x,E)f(x)^2+K^{(\alpha)}(f^2)=\int_{E^*}J^{*(\alpha)}(\mathrm{d}x,E^*)f^*(x)^2, \tag{2.1}$$

$$D^{(\alpha)}(f,f)=\frac{1}{2}\int_{E^*\times E^*}J^{*(\alpha)}(\mathrm{d}x,\mathrm{d}y)(f^*(y)-f^*(x))^2, \tag{2.2}$$

$$\begin{aligned}&\frac{1}{2}\int J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y)|f(y)-f(x)|+\int K^{(\alpha)}(\mathrm{d}x)|f(x)|\\&\quad=\frac{1}{2}\int_{E^*\times E^*}J^{*(\alpha)}(\mathrm{d}x,\mathrm{d}y)|f^*(y)-f^*(x)|.\end{aligned} \tag{2.3}$$

Therefore, for f with $\pi(f^2)=1$, by (2.1)–(2.3), (1.8), Lemma 2.1 and the Cauchy-Schwarz inequality,

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

This implies that $D^{(1)}(f,f)\geqslant 1-\sqrt{1-h^{(1)^2}}$ and so

$$\lambda_0^{(1)}\geqslant 1-\sqrt{1-h^{(1)^2}}. \tag{2.4}$$

Next, by (1.8), Lemma 2.1 and another use of the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}h^{(1/2)^2}&\leqslant\left\{\frac{1}{2}\int J^{*(1/2)}(\mathrm{d}x,\mathrm{d}y)\big|f^*(y)^2-f^*(x)^2\big|\right\}^2\\&\leqslant\frac{1}{2}D(f,f)\int J^{*(1)}(\mathrm{d}x,\mathrm{d}y)\big[f^*(y)+f^*(x)\big]^2\\&\leqslant D(f,f)\big[2-D^{(1)}(f,f)\big]\\&\leqslant D(f,f)\big[2-\lambda_0^{(1)}\big].\end{aligned} \tag{2.5}$$

From this and (2.4), the required assertion follows. □

Proof of Theorem 1.2. (a) First, we prove (1.10). Let $f \in \mathscr{D}(D)$ with $\pi(f) = 0$ and $\pi(f^2) = 1$. Set $g = f + c$, $c \in \mathbb{R}$. Similarly to (2.5), we have

$$\begin{aligned}\left\{\int J^{(1/2)}(\mathrm{d}x, \mathrm{d}y)\big|g(y)^2 - g(x)^2\big|\right\}^2 &\leqslant 4D(f,f)\big[2(1+c^2) - D^{(1)}(f,f)\big] \\ &\leqslant 4D(f,f)[2(1+c^2) - \beta]\end{aligned}$$

for all $\beta\colon\ 0 \leqslant \beta < \lambda_1^{(1)} \leqslant 2$. Hence by Lemma 2.1, we have

$$\begin{aligned}D(f,f) &\geqslant \frac{1}{4[2(1+c^2)-\beta]}\left\{\int J^{(1/2)}(\mathrm{d}x, \mathrm{d}y)\big|g(y)^2 - g(x)^2\big|\right\}^2 \\ &\geqslant \frac{\kappa_\beta}{4} {k^{(1/2)}}^2 \end{aligned} \tag{2.6}$$

where κ_β is the same as κ defined below (1.7) but replacing the denominator $1+c^2$ with $2(1+c^2)-\beta$. To estimate κ_β, we adopt an optimizing procedure which will be used several times subsequently. Set $\gamma = \mathbb{E}|X| \in (0,1]$. It is known that

$$\lim_{c\to\pm\infty} \frac{\big(\mathbb{E}\big|(X+c)^2 - (Y+c)^2\big|\big)^2}{2(1+c^2)-\beta} = 2(\mathbb{E}|X-Y|)^2 \geqslant 2(\mathbb{E}|X|)^2 = 2\gamma^2$$

and when $c = 0$, $\mathbb{E}\big|X^2 - Y^2\big| \geqslant 2(1 - \mathbb{E}|X|) = 2(1-\gamma)$ (cf. [8] or [3], Section 9.2). Thus,

$$\kappa_\beta \geqslant \inf_{\gamma\in(0,1]} \max\left\{2\gamma^2,\ \frac{4(1-\gamma)^2}{2-\beta}\right\}. \tag{2.7}$$

We now need an elementary fact.

Lemma 2.2. Let f and g be continuous functions on $[0,1]$ and satisfy $f(0) < g(0)$ and $f(1) > g(1)$. Suppose that f is increasing and g is decreasing. Then

$$\inf_{\gamma\in[0,1]} \max\{f(\gamma), g(\gamma)\} = f(\gamma_0),$$

where γ_0 is the unique solution to the equation $f = g$ on $[0,1]$.

Applying Lemma 2.2 to (2.7), we get

$$\kappa_\beta \geqslant \frac{4}{\big(\sqrt{2} + \sqrt{2-\beta}\big)^2}.$$

Combining this with (2.6) and then letting $\beta \uparrow \lambda_1^{(1)}$, we obtain (1.10).

It is worthy to mention that the estimate just proved can be sharp. To see this, simply consider $E = \{0,1\}$, $J(\{i\},\{j\}) = 1\,(i \neq j)$ and $\pi_0 = \pi_1 = 1/2$. Then $k^{(1/2)} = \lambda_1^{(1)} = \lambda_1 = 2$. Moreover, the same example shows that in contrast to (1.9), the analog of (1.9) "$\lambda_1 \geqslant {k^{(1/2)}}^2/\big[4\big(2-\lambda_1^{(1)}\big)\big]$" or "$\lambda_1 \geqslant {k^{(1/2)}}'^2/\big[2-\lambda_1^{(1)}\big)\big]$" does not hold.

(b) For any $B \subset E$ with $\pi(B) > 0$, define a local form as follows.

$$\widetilde{D}_B^{(\alpha)}(f,f) = \frac{1}{2}\int_{B\times B} J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y)[f(y)-f(x)]^2 + \int_B J^{(\alpha)}(\mathrm{d}x, B^c)f(x)^2.$$

Obviously, $\widetilde{D}_B^{(\alpha)}(f,f) = \widetilde{D}_B^{(\alpha)}(fI_B, fI_B)$. Moreover, it is easy to see that

$$\lambda_0(B) = \inf\big\{\widetilde{D}_B(f,f)\colon\ \pi(f^2 I_B) = 1\big\}.$$

Let

$$\begin{aligned} h_B^{(\alpha)} &= \inf_{A\subset B,\,\pi(A)>0} \frac{J^{(\alpha)}(A\times(B\setminus A)) + J^{(\alpha)}(A\times B^c)}{\pi(A)} \\ &= \inf_{A\subset B,\,\pi(A)>0} \frac{J^{(\alpha)}(A\times A^c)}{\pi(A)}. \end{aligned} \tag{2.8}$$

Applying Theorem 1.1 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

$$\lambda_0(B) \geqslant {h_B^{(1/2)}}^2 \Big/ \Big[1 + \sqrt{1 - {h_B^{(1)}}^2}\,\Big].$$

We now come to another key point of the proof. In [8], the proof is based on the estimate $\lambda_1 \geqslant \inf_B\{\lambda_0(B)\vee\lambda_0(B^c)\}$. However, we are unable to prove this in the present setup. Instead, we prove the following weaker result which is enough for our purpose.

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

For each $\varepsilon > 0$, choose f_ε with $\pi(f_\varepsilon) = 0$ and $\pi(f_\varepsilon^2) = 1$ such that $\lambda_1 + \varepsilon \geqslant D(f_\varepsilon, f_\varepsilon)$. 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

$$\begin{aligned} \lambda_1 + \varepsilon &\geqslant D(f_\varepsilon - c_\varepsilon,\, f_\varepsilon - c_\varepsilon) \\ &= \frac{1}{2}\int J(\mathrm{d}x,\mathrm{d}y)\big[\big|f_\varepsilon^+(y) - f_\varepsilon^+(x)\big| + \big|f_\varepsilon^-(y) - f_\varepsilon^-(x)\big|\big]^2 \\ &\geqslant \frac{1}{2}\int J(\mathrm{d}x,\mathrm{d}y)\big(f_\varepsilon^+(y) - f_\varepsilon^+(x)\big)^2 + \frac{1}{2}\int J(\mathrm{d}x,\mathrm{d}y)\big(f_\varepsilon^-(y) - f_\varepsilon^-(x)\big)^2 \\ &\geqslant \lambda_0\big(B_\varepsilon^+\big)\pi\big(\big(f_\varepsilon^+\big)^2\big) + \lambda_0\big(B_\varepsilon^-\big)\pi\big(\big(f_\varepsilon^-\big)^2\big) \\ &\geqslant \inf_{\pi(B)\leqslant 1/2} \lambda_0(B)\pi\big(\big(f_\varepsilon^+\big)^2 + \big(f_\varepsilon^-\big)^2\big) \\ &= (1 + c_\varepsilon^2) \inf_{\pi(B)\leqslant 1/2} \lambda_0(B) \\ &\geqslant \inf_{\pi(B)\leqslant 1/2} \lambda_0(B). \end{aligned}$$

Because ε is arbitrary, we obtain the required conclusion.

Finally, combining the above two assertions, we obtain

$$\begin{aligned}\lambda_1 &\geqslant \inf_{\pi(B)\leqslant 1/2} \frac{{h_B^{(1/2)}}^2}{1+\sqrt{1-{h_B^{(1)}}^2}} \\ &\geqslant \inf_{\pi(B)\leqslant 1/2} \frac{\inf_{\pi(B)\leqslant 1/2} {h_B^{(1/2)}}^2}{1+\sqrt{1-{h_B^{(1)}}^2}} \\ &\geqslant \frac{\inf_{\pi(B)\leqslant 1/2} {h_B^{(1/2)}}^2}{1+\sqrt{1-\inf_{\pi(B)\leqslant 1/2} {h_B^{(1)}}^2}} \\ &= \frac{{k^{(1/2)}}'^2}{1+\sqrt{1-{k^{(1)}}'^2}} \quad \square\end{aligned}$$

Proof of Theorem 1.3. The proof is split into two lemmas given below. Noticing that α is fixed, we may and will omit the superscript "(α)" everywhere in the next two lemmas and their proofs for simplicity. $\square$

Lemma 2.3. Let $B \in \mathscr{E}$ with $2\pi(B) > 1$. Then

$$k' \geqslant \frac{h_{B^c} k(B)(2\pi(B)-1)}{k(B)(2\pi(B)-1) + 2\pi(B)^2(M_B + h_{B^c})},$$

where h_B is defined by (2.8).

Proof. We need only to consider the case that $h_{B^c}k(B) > 0$. For any $A \in \mathscr{E}$ with $\pi(A) \in (0, 1/2]$, let $\gamma = \pi(AB)/\pi(A)$. Then

$$\begin{aligned}\frac{J(A\times A^c)}{\pi(A)} &= \frac{1}{2\pi(A)} \int J(\mathrm{d}x, \mathrm{d}y)\big[I_A(y) - I_A(x)\big]^2 \\ &\geqslant \frac{1}{2\pi(A)} \int_{B\times B} J(\mathrm{d}x, \mathrm{d}y)\big[I_A(y) - I_A(x)\big]^2 \\ &\geqslant \frac{k(B)\pi^B(A)\pi^B(A^c)}{\pi(A)} \\ &\geqslant \frac{\pi(B)-1/2}{\pi(B)^2} k(B)\gamma. \end{aligned} \tag{2.9}$$

Here, in the last step, we have used $\pi(AB) \leqslant \pi(A) \leqslant 1/2$. On the other hand, we have

$$\begin{aligned} h_{B^c}\pi(AB^c) &\leqslant \frac{1}{2}\int J(\mathrm{d}x, \mathrm{d}y)\big[I_{AB^c}(x) - I_{AB^c}(y)\big]^2 \\ &= \frac{1}{2}\int J(\mathrm{d}x, \mathrm{d}y)\big|I_{A^c\cup B}(x) - I_{A^c\cup B}(y)\big|. \end{aligned}$$

Noticing that J is symmetric and

$$\big|I_{A^c\cup B}(x)-I_{A^c\cup B}(y)\big|\leqslant\big|I_{A^c}(x)-I_{A^c}(y)\big|+I_{B\times B^c+B^c\times B}\big|I_{AB}(x)-I_{AB}(y)\big|,$$

we obtain

$$h_{B^c}(1-\gamma)=\frac{h_{B^c}\pi(AB^c)}{\pi(A)}\leqslant\frac{J(A\times A^c)}{\pi(A)}+M_B\gamma.$$

Combining this with (2.9) and applying Lemma 2.2, we get

$$\begin{aligned}\frac{J(A\times A^c)}{\pi(A)}&\geqslant\inf_{\gamma\in[0,1]}\max\Big\{(\pi(B)-1/2)\pi(B)^{-2}k(B)\gamma,\ h_{B^c}-\big(M_B+h_{B^c}\big)\gamma\Big\}\\&=\frac{h_{B^c}k(B)(2\pi(B)-1)}{k(B)(2\pi(B)-1)+2\pi(B)^2(M_B+h_{B^c})}.\quad\square\end{aligned}$$

Lemma 2.4. Let φ satisfy $\delta_1(\varphi)<\infty$. If $\gamma_B=-\sup_B\Omega\varphi>0$, then

$$h_B\geqslant\gamma_B/\delta_1(\varphi)>0.$$

Proof. For any $A\subset B$, we have

$$\begin{aligned}\gamma_B\pi(A)&\leqslant\int_A[-\Omega\varphi]\mathrm{d}\pi\\&=\frac12\int J(\mathrm{d}x,\mathrm{d}y)(I_A(x)-I_A(y))(\varphi(x)-\varphi(y))\\&\leqslant\frac{\delta_1(\varphi)}{2}\int J(\mathrm{d}x,\mathrm{d}y)|I_A(x)-I_A(y)|\\&=\delta_1(\varphi)J(A\times A^c).\end{aligned}$$

Hence, $h_B\geqslant\gamma_B/\delta_1(\varphi)$. $\square$

To conclude this section, we discuss a different way to deal with general symmetric forms. In contrast to the previous approach, we now keep (J,K) to be the same but change the L^2-space. To do so, let p be a measurable function and satisfy $\alpha_p:=\operatorname{ess\,inf}_\pi p>0$, $\beta_p:=\pi(p)<\infty$ and $\|J(\cdot,E)+K\|_{\mathrm{op}}\leqslant\beta_p\,(L^1_+(\pi_p)\to\mathbb{R}_+)$, where $\pi_p=p\pi/\beta_p$. For jump processes, one may take $p(x)=q(x)\vee r$ for some $r\geqslant0$. From this, one sees the main restriction of the present approach: $\int\pi(\mathrm{d}x)q(x)<\infty$, since we require that $\pi(p)<\infty$. Except this point, the approach is not comparable with the previous one (see Examples 4.5 and 4.7 given below).

Next, define h_p, k_p and k_p' by (1.3)–(1.5), respectively, with π replaced by π_p and then divided by β_p. For instance,

$$k_p'=\inf_{\pi_p(A)\leqslant1/2}J(A\times A^c)/\pi(pI_A).$$

Theorem 2.5. Let p, α_p, β_p and π_p be given above. Define $\lambda_{p,i}\,(i=0,1)$ by (1.1) and (1.2) with π replaced by π_p. Then, we have

$$\lambda_i \geqslant \frac{\alpha_p}{\beta_p}\,\lambda_{p,i}, \qquad i=0,\ 1. \tag{2.10}$$

In particular,

$$\lambda_0 \geqslant \alpha_p\Big(1-\sqrt{1-h_p^2}\Big) \tag{2.11}$$

and when $K=0$,

$$\lambda_1 \geqslant \max\left\{\frac{\kappa}{8}\alpha_p k_p^2,\ \alpha_p\Big(1-\sqrt{1-k_p'^2}\Big)\right\}. \tag{2.12}$$

Proof. (a) We prove that $L^\infty(\pi)$ is dense in $\mathscr{D}(D)$ in the D-norm: $\|f\|_D^2 = D(f,f)+\pi(f^2)$. The proof is similar to [3], Lemma 9.7. First, we show that $L^\infty(\pi)\subset\mathscr{D}(D)$. Because $1\in L^1(\pi_p)$ and $\|J(\cdot,E)+K\|_{\mathrm{op}}\leqslant\beta_p$, we have $J(E,E)+K(E)\leqslant\beta_p<\infty$. Thus,

$$\begin{aligned} D(f,f) &\leqslant \int J(\mathrm{d}x,\mathrm{d}y)\big[f(y)^2+f(x)^2\big]+\int K(\mathrm{d}x)f(x)^2\\ &\leqslant 2\|f\|_\infty^2\big(J(E,E)+K(E)\big)\\ &<\infty, \end{aligned}$$

and hence $f\in\mathscr{D}(D)$. Next, let $f\in\mathscr{D}(D)$ and set $f_n=(-n)\vee(f\wedge n)$. Then $f_n\in\mathscr{D}(D)$,

$$|f_n(y)-f_n(x)|\leqslant|f(y)-f(x)| \qquad\text{and}\qquad |f_n(x)|\leqslant|f(x)| \tag{2.13}$$

for all x, y and n. Clearly, $\pi\big((f_n-f)^2\big)\to 0$. Moreover, since

$$D(f_n-f,f_n-f)\leqslant 4D(f,f)<\infty$$

by (2.13), we have $D(f_n-f,f_n-f)\to 0$ by (2.13) and the dominated convergence theorem. Therefore, $\|f_n-f\|_D\to 0$.

(b) Here, we prove (2.10) for $i=1$ only since the proof for $i=0$ is similar and even simpler. Then, (2.11) and (2.12) follows from (1.7) and the comment right after Theorem 1.2 with $M=\beta_p$.

Because $L^\infty(\pi)\subset L^2(\pi_p)$ and $L^2(\pi_p)$ is just the domain of the form $D(f,f)$ on $L^2(\pi_p)$, by definition of λ_1 and $\lambda_{p,1}$, it suffices to show that

$$\pi_p(f^2)-\pi_p(f)^2\geqslant\big[\pi(f^2)-\pi(f)^2\big]\alpha_p/\beta_p$$

for every $f\in L^\infty(\pi)$. The proof goes as follows.

$$\begin{aligned} \pi_p\big(f^2\big)-\pi_p(f)^2 &= \inf_{c\in\mathbb{R}}\int\big(f(x)-c\big)^2\pi_p(\mathrm{d}x)\\ &= \beta_p^{-1}\inf_{c\in\mathbb{R}}\int\big(f(x)-c\big)^2p(x)\pi(\mathrm{d}x)\\ &\geqslant \frac{\alpha_p}{\beta_p}\inf_{c\in\mathbb{R}}\int\big(f(x)-c\big)^2\pi(\mathrm{d}x)\\ &= \frac{\alpha_p}{\beta_p}\big[\pi\big(f^2\big)-\pi(f)^2\big]. \quad\square \end{aligned}$$

3. A criterion for the existence of spectral gap. Proofs of Theorems 1.4 and 1.5.

To state our main criterion, we need some preparation.

Let E be a locally compact separable metric space with Borel field $\mathscr{E}$ and π be a probability measure with $\mathrm{supp}(\pi) = E$. Denote by $C_b(E)$ [resp. $C_0(E)$] the set of all bounded continuous functions (resp. with compact support) on E.

Next, let $(D, \mathscr{D}(D))$ be a regular conservative Dirichlet form on $L^2(\pi)$. By Beurling-Deny's formula, the form can be expressed as follows:

$$D(f,f)=D^{(c)}(f,f)+\frac{1}{2}\int J(\mathrm{d}x,\mathrm{d}y)(f(x)-f(y))^2, \quad f\in\mathscr{D}(D)\cap C_0(E) \tag{3.1}$$

where $\mathscr{D}(D^{(c)}) = \mathscr{D}(D) \cap C_0(E)$ and satisfies a strong local property; J is a symmetric Radon measure on the product space $E \times E$ off diagonal. Moreover, there exists a finite, nonnegative Radon measure $\mu^c_{\langle f\rangle}$ such that

$$D^{(c)}(f,f) = \frac{1}{2}\int_E \mathrm{d}\mu^c_{\langle f\rangle}, \qquad f\in\mathscr{D}(D)\cap C_b(E).$$

Theorem 3.1. Let $\mathscr{C} \subset \mathscr{D}(D)\cap C_0(E)$ be dense in $\mathscr{D}(D)$ in the D-norm: $\|f\|_D^2 = D(f,f) + \pi(f^2)$. Set $\mathscr{C}_L = \{f + c\colon\ f \in \mathscr{C},\ c \in \mathbb{R}\}$. Given A, $B \in \mathscr{E}$, $A \subset B$ with $0 < \pi(A)$, $\pi(B) < 1$. Suppose that the following conditions hold.

(i) There exists a conservative Dirichlet form $(D_B, \mathscr{D}(D_B))$ on the square-integrable functions on B with respect to π^B such that $\mathscr{C}_L|_B \subset \mathscr{D}(D_B)$ and

$$D(f,f) \geqslant D_B(fI_B, fI_B), \qquad f\in\mathscr{C}_L.$$

(ii) There exists a function $h \in \mathscr{C}_L$: $0 \leqslant h \leqslant 1$, $h|_A = 0$ and $h|_{B^c} = 1$ such that

$$c(h) := \sup_{f\in\mathscr{C}_L}\frac{1}{\pi\big(f^2I_B\big)}\left[\frac{1}{2}\int f^2\mathrm{d}\mu^c_{\langle h\rangle}+\int_{B\times A^c}J(\mathrm{d}x,\mathrm{d}y)\big[f(1-h)(y)-f(1-h)(x)\big]^2\right] < \infty.$$

Then, we have

$$\frac{\lambda_0(A^c)}{\pi(A)} \geqslant \lambda_1 \geqslant \frac{\lambda_1(B)[\lambda_0(A^c)\pi(B) - 2c(h)\pi(B^c)]}{2\lambda_1(B) + \pi(B)^2[\lambda_0(A^c) + 2c(h)]}.$$

Proof. The upper bound is easy. Simply take $f \in \mathscr{D}(D)$ with $f|_A = 0$ and $\pi(f^2) = 1$. Then

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

Hence $\lambda_1 \leqslant D(f,f)/\pi(A)$ which gives us $\lambda_1 \leqslant \lambda_0\big(A^c\big)/\pi(A)$.

For the lower bound, let $f \in \mathscr{C}_L$ with $\pi(f) = 0$ and $\pi\big(f^2\big) = 1$. Set $\gamma = \pi\big(f^2I_B\big)$.

(a) By condition (i), we have

$$\begin{aligned} D(f,f) &\geqslant D_B(fI_B, fI_B) \\ &\geqslant \lambda_1(B)\pi(B)^{-1}\big[\pi\big(f^2 I_B\big) - \pi(B)^{-1}\pi\big(fI_B\big)^2\big] \\ &= \lambda_1(B)\pi(B)^{-1}\big[\pi\big(f^2 I_B\big) - \pi(B)^{-1}\pi\big(fI_{B^c}\big)^2\big] \\ &\geqslant \lambda_1(B)\pi(B)^{-1}\big[\gamma - \pi(B)^{-1}\pi\big(f^2 I_{B^c}\big)\pi\big(B^c\big)\big] \\ &= \lambda_1(B)\pi(B)^{-2}\big[\gamma - \pi\big(B^c\big)\big]. \end{aligned} \tag{3.2}$$

(b) Let ρ be the metric in E. By the construction of $\mu^c_{\langle f\rangle}$ ([7], Section 3.2), there exist a sequence of relatively compact open sets G_ℓ increasing to E, a sequence of symmetric, nonnegative Radon measures σ_{β_n} and a sequence δ_ℓ such that

$$\int_E g\mathrm{d}\mu^c_{\langle f\rangle} = \lim_{\ell\to\infty}\lim_{\beta_n\to\infty}\beta_n \int_{G_\ell\times G_\ell,\, \rho(x,y)<\delta_\ell} [f(x)-f(y)]^2 g(x)\sigma_{\beta_n}(\mathrm{d}x,\mathrm{d}y)$$
$$f, g \in \mathscr{D}(D)\cap C_0(E).$$

From this and

$$[(fh)(x)-(fh)(y)]^2 \leqslant 2h(y)^2[f(x)-f(y)]^2 + 2f(x)^2[h(x)-h(y)]^2,$$

it follows that

$$\int \mathrm{d}\mu^c_{\langle fh\rangle} \leqslant 2\int h^2\mathrm{d}\mu^c_{\langle f\rangle} + 2\int f^2\mathrm{d}\mu^c_{\langle h\rangle},$$

first for f, $h\in\mathscr{D}(D)\cap C_0(E)$ and then for f, $h\in\mathscr{D}(D)\cap C_b(E)$ (cf. [7], Section 3.2). Hence

$$D^{(c)}(fh,fh) = \frac{1}{2}\int \mathrm{d}\mu^c_{\langle fh\rangle} \leqslant 2D^{(c)}(f,f) + \int f^2\mathrm{d}\mu^c_{\langle h\rangle}. \tag{3.3}$$

On the other hand, since

$$|(fh)(x)-(fh)(y)| \leqslant |f(x)-f(y)| + I_{B\times A^c\cup A^c\times B}(x,y)|f(1-h)(x)-f(1-h)(y)|,$$

we have

$$\begin{aligned} &\int J(\mathrm{d}x,\mathrm{d}y)[(fh)(x)-(fh)(y)]^2 \\ &\quad\leqslant 2\int J(\mathrm{d}x,\mathrm{d}y)[f(x)-f(y)]^2 \\ &\qquad + 4\int_{B\times A^c} J(\mathrm{d}x,\mathrm{d}y)[f(1-h)(x)-f(1-h)(y)]^2. \end{aligned} \tag{3.4}$$

Thus, combining (3.1), (3.3), (3.4) with condition (ii), we get

$$\begin{aligned} D(fh,fh) &\leqslant 2D(f,f) + \int f^2\mathrm{d}\mu^c_{\langle h\rangle} + 2\int_{B\times A^c} J(\mathrm{d}x,\mathrm{d}y)\big[f(1-h)(x)-f(1-h)(y)\big]^2 \\ &\leqslant 2D(f,f) + 2c(h)\pi\big(f^2 I_B\big) \\ &= 2D(f,f) + 2\gamma c(h). \end{aligned}$$

That is,

$$\begin{aligned} D(f,f) &\geqslant \frac{1}{2}D(fh,fh)-\gamma c(h)\\ &\geqslant \frac{1}{2}\lambda_0\big(A^c\big)\pi\big(f^2h^2\big)-\gamma c(h)\\ &\geqslant \frac{1}{2}\lambda_0\big(A^c\big)\pi\big(f^2I_{B^c}\big)-\gamma c(h)\\ &= \frac{1}{2}\lambda_0\big(A^c\big)(1-\gamma)-\gamma c(h). \end{aligned} \tag{3.5}$$

Combining (3.2) with (3.5), we obtain

$$\begin{aligned} D(f,f) &\geqslant \inf_{\gamma\in[0,1]}\max\left\{\frac{\lambda_1(B)}{\pi(B)^2}\big(\gamma-\pi\big(B^c\big)\big),\ \frac{1}{2}\lambda_0\big(A^c\big)(1-\gamma)-\gamma c(h)\right\}\\ &= \lambda_1(B)\pi(B)^{-2}\big(\gamma_0-\pi\big(B^c\big)\big). \end{aligned} \tag{3.6}$$

The assertion of the theorem now follows from (3.6) and Lemma 2.2. □

Theorem 3.1 is effective for diffusions as was shown in [10] with a more direct proof (in this case the Dirichlet form is explicit). We now apply the theorem to jump processes.

Proof of Theorem 1.4. First, the topological assumptions of Theorem 3.1 are unnecessary in the present context. To see that condition (i) is fulfilled, simply take D_B to be the one defined by (1.13). For condition (ii), take $h = I_{A^c}$. Then

$$\begin{aligned} \int_{B\times A^c} J(\mathrm{d}x,\mathrm{d}y)[(fI_A)(x)-(fI_A)(y)]^2 &= \int_{A\times A^c} J(\mathrm{d}x,\mathrm{d}y)f(x)^2\\ &\leqslant M_A\pi\big(f^2I_A\big)\\ &\leqslant M_A\pi\big(f^2I_B\big). \end{aligned}$$

This means that condition (ii) holds with $c(h) = M_A$. We have thus proved Theorem 1.4. □

The application of Theorem 3.1 (or Theorem 1.4) requires some estimates of $\lambda_0(A^c)$ and $\lambda_1(B)$, which may be obtained from Theorems 1.1 and 1.2. These estimates are usually in the qualitative sense good enough for $\lambda_1(B)$, for which there are also quite a lot of publications, including the authors' study, in the past years. However, for $\lambda_0(A^c)$, the bound presented above may not be sharp enough, especially in the unbounded situation. For this reason, we now introduce a different result.

Theorem 3.2. Let E be a metric space with Borel field $\mathscr{E}$ and let (x_t) be a reversible right-continuous Markov process valued in E with weak generator Ω. Suppose that the corresponding Dirichlet form is regular. Next, fix a closed set B. Suppose additionally that the following conditions hold:

(i) There exists a function φ satisfying $\varphi|_B = 0$, $\varphi|_{B^c} > 0$ and $\sup_{B^c}\Omega\varphi/\varphi =: -\delta < 0$.
(ii) There exists a sequence of open sets (E_n): $E_0 \supset B$, $E_n \uparrow E$ such that φ is bounded below on each $E_n \setminus B$ by a positive constant.
(iii) The first Dirichlet eigenfunction of Ω on each $E_n \setminus B$ is bounded above.

Then we have $\lambda_0(B^c) \geqslant \delta$. In particular, for jump processes, the condition "$\varphi|_B = 0$" given in (i) can be removed.

Clearly, conditions (ii) and (iii) with compact B are fulfilled for diffusions or Markov chains. Thus, the key condition here is the first one.

Proof of Theorem 3.2. The last assertion follows by replacing φ with φI_{B^c}. Indeed,

$$\begin{aligned}\Omega(\varphi I_{B^c})(x) &= \int q(x,\mathrm{d}y)\big[(\varphi I_{B^c})(y) - (\varphi I_{B^c})(x)\big]\\ &\leqslant \int q(x,\mathrm{d}y)[\varphi(y) - (\varphi I_{B^c})(x)]\\ &= \Omega\varphi(x)\\ &\leqslant -\delta(\varphi I_{B^c})(x) \quad \text{on} \quad B^c.\end{aligned}$$

We are now going to prove the main assertion of the theorem. Set $\tau_B = \inf\{t \geqslant 0\colon\ x_t \in B\}$. Then, by condition (i) plus a truncating argument if necessary, we get

$$\mathbb{E}^x\varphi(x_{t\wedge\tau_B}) \leqslant \varphi(x)e^{-\delta t}, \qquad t \geqslant 0,\ x \notin B.$$

Next, let $u_n\ (\geqslant 0)$ be the first Dirichlet eigenfunction of Ω on $E_n \setminus B$. Set $\tau = \inf\{t \geqslant 0\colon\ x_t \notin E_n \setminus B\}$. Then, by conditions (ii) and (iii), there exists $c_1 > 0$ such that $u_n(x_{t\wedge\tau}) \leqslant c_1\varphi(x_{t\wedge\tau_B})$ and so

$$u_n(x)e^{-\lambda_0(E_n\setminus B)t} = \mathbb{E}^x u_n(x_{t\wedge\tau}) \leqslant c_1\mathbb{E}^x\varphi(x_{t\wedge\tau_B}) \leqslant c_1\varphi(x)e^{-\delta t}, \qquad x \in E_n \setminus B.$$

This implies that $\lambda_0(E_n \setminus B) \geqslant \delta$. Finally, because the Dirichlet form is regular, it is easy to show that $\lambda_0(B^c) = \lim_{n\to\infty}\lambda_0(E_n \setminus B)$ and so the required assertion follows. □

For the remainder of this section, we turn to study the upper bound of λ_1.

Let $(D, \mathscr{D}(D))$ be a general conservative Dirichlet form and let $P(t,x,\mathrm{d}y)$ be the corresponding transition probability. Fix $\varphi \geqslant 0$. Suppose that $\varphi \wedge n \in \mathscr{D}(D)$ for every $n \geqslant 1$. Set $f_n = \exp[\varepsilon(\varphi \wedge n)/2]$. Since the function $e^{\alpha x}$ is locally Lipschitz continuous and $\varphi \wedge n$ is bounded, by the elementary spectral representation theory, we have

$$\begin{aligned}D(f_n, f_n) &= \lim_{t\to 0}\frac{1}{2t}\int \pi(\mathrm{d}x)P(t,x,\mathrm{d}y)[f_n(x) - f_n(y)]^2\\ &\leqslant \frac{\varepsilon^2}{4}C(\varphi,n)\lim_{t\to 0}\frac{1}{2t}\int \pi(\mathrm{d}x)P(t,x,\mathrm{d}y)[(\varphi\wedge n)(x) - (\varphi\wedge n)(y)]^2\\ &\leqslant \frac{\varepsilon^2}{4}C(\varphi,n)D(\varphi\wedge n, \varphi\wedge n)\\ &< \infty,\end{aligned}$$

where $C(\varphi,n)$ is the Lipschitz norm of $e^{\varepsilon x/2}$ on the range of $\varphi \wedge n$. This leads us to introduce the following constant:

$$\delta(\varepsilon,\varphi) = \varepsilon^{-2}\sup_{n\geqslant 1} D(f_n,f_n)/\pi(f_n^2).$$

Theorem 3.3. Let $(D, \mathscr{D}(D))$, φ, f_n and $\delta(\varepsilon,\varphi)$ be as above. Then, we have

$$\lambda_1 \leqslant \inf\{\varepsilon^2\delta(\varepsilon,\varphi)\colon\ \pi(e^{\varepsilon\varphi}) = \infty\}.$$

Proof. We need to show that if $\pi(e^{\varepsilon\varphi}) = \infty$, then $\lambda_1 \leqslant \varepsilon^2\delta(\varepsilon,\varphi)$. For $n \geqslant 1$, we have

$$\lambda_1 \leqslant \frac{D(f_n,f_n)}{\pi(f_n^2) - \pi(f_n)^2}. \tag{3.7}$$

For every $m \geqslant 1$, choose $r_m > 0$ such that $\pi(\varphi \geqslant r_m) \leqslant 1/m$. Then

$$\pi(I_{[\varphi\geqslant r_m]}f_n^2)^{1/2} \geqslant \sqrt{m}\,\pi(I_{[\varphi\geqslant r_m]}f_n) \geqslant \sqrt{m}\,\pi(f_n) - \sqrt{m}\,\exp(\varepsilon r_m/2).$$

Hence

$$\pi(f_n)^2 \leqslant \Big[\sqrt{\pi(f_n^2)}/\sqrt{m} + \exp(\varepsilon r_m/2)\Big]^2. \tag{3.8}$$

On the other hand, by assumption, we have

$$D(f_n,f_n) \leqslant \varepsilon^2\delta(\varepsilon,\varphi)\pi(f_n^2). \tag{3.9}$$

Noticing that $\pi(f_n^2) \uparrow \infty$, combining (3.9) with (3.7) and (3.8) and then letting $n \uparrow \infty$, we obtain

$$\lambda_1 \leqslant \varepsilon^2\delta(\varepsilon,\varphi)/\big[1 - m^{-1}\big].$$

The proof is completed by setting $m \uparrow \infty$. □

Proof of Theorem 1.5. It suffices to prove the first assertion because the remainder of the proof is similar. Let f_n be given as in Theorem 3.3. Note that by the 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,

$$\begin{aligned} D(f_n,f_n) &= \frac{1}{2}\int J(\mathrm{d}x,\mathrm{d}y)[f_n(x) - f_n(y)]^2 \\ &\leqslant \frac{\varepsilon^2}{8}\int J^{(1)}(\mathrm{d}x,\mathrm{d}y)[\varphi(x)-\varphi(y)]^2 r(x,y)\big[f_n(x)\vee f_n(y)\big]^2 \\ &\leqslant \frac{\varepsilon^2}{4}\delta_2(\varphi)\pi(f_n^2). \end{aligned}$$

The conclusion now follows from Theorem 3.3 with $\delta(\varepsilon,\varphi) = \frac{1}{4}\delta_2(\varphi)$. □

4. Spectral gap for Markov chains.

Usually, the power of a result for general jump processes should be justified by Markov chains.

Let E be countable and (q_{ij}) be a regular and irreducible Q-matrix, reversible with respect to $\pi = (\pi_i)$. As usual, let $q_i = \sum_{j\neq i} q_{ij}$. Then $K = 0$ and

$$\Omega f(i) = \sum_{j\neq i} q_{ij}[f_j - f_i].$$

The density of the symmetric measure with respect to the counting measure becomes $J(i,j) = \pi_i q_{ij}\,(i \neq j)$. For simplicity, we consider only two typical situations: $E = \mathbb{Z}_+$ or $E = \mathbb{Z}^d$ and take $r(i,j) = 1/(q_i \vee q_j)$. Denote by $|i|$ the L^1-norm, that is, $|i| = \sum_{k=1}^d |i_k|$ for $i = (i_1, \cdots, i_d) \in \mathbb{Z}^d$.

A combination of Theorem 1.2 and the next result provides us with a simple condition for the existence of spectral gap for birth-death processes and the result seems to be new to our knowledge, even for such a simple situation (cf. [4]).

Theorem 4.1. Consider the birth-death process on $\mathbb{Z}_+$ with birth rates (b_i) and death rates (a_i):

(i) Take $r_{ij} = (a_i+b_i)\vee(a_j+b_j)\,(i \neq j)$. Then $k^{(\alpha)'} > 0$ (equivalently, $k^{(\alpha)} > 0$) iff there exists a constant $c > 0$ such that

$$\frac{\pi_i a_i}{[(a_i + b_i) \vee (a_{i-1} + b_{i-1})]^\alpha} \geqslant c \sum_{j\geqslant i} \pi_j, \qquad i \geqslant 1. \tag{4.1}$$

Then, we indeed have $k^{(\alpha)'} \geqslant c$. Furthermore,

$$k^{(\alpha)} \geqslant \inf_{i\geqslant 1} \frac{\pi_i a_i}{[(a_i + b_i) \vee (a_{i-1} + b_{i-1})]^\alpha (1 - \pi_i) \sum_{j\geqslant i} \pi_j}.$$

(ii) Let $\sum_i \pi_i(a_i + b_i) < \infty$. Take $p_i = a_i + b_i$. Then we have $k'_p > 0$ (equivalently, $k_p > 0$) iff

$$\inf_{i\geqslant 1} \frac{\pi_i a_i}{\sum_{j\geqslant i} \pi_j p_j} > 0$$

and moreover,

$$k'_p \geqslant \inf_{i\geqslant 1} \frac{\pi_i a_i}{\sum_{j\geqslant i} \pi_j p_j}, \qquad k_p \geqslant \inf_{i\geqslant 1} \frac{\pi_i a_i}{(1 - \pi_i p_i/\beta_p) \sum_{j\geqslant i} \pi_j p_j}.$$

Roughly speaking, (4.1) holds if π_j has exponential decay. For polynomial decay, (4.1) can still be true when $\alpha = 1/2$. See Example 4.5.

Proof. Here we prove part (i) only since the proof of part (ii) is similar.

(a) Let $k^{(\alpha)} > 0$. Take $A = I_i = \{i, i+1, \cdots\}$ for a fixed $i > 0$ and

$$J^{(\alpha)}(i,j) = \frac{\pi_i q_{ij}}{[q_i \vee q_j]^\alpha} = \begin{cases} \dfrac{\pi_i a_i}{[(a_i + b_i) \vee (a_{i-1} + b_{i-1})]^\alpha} =: \pi_i \tilde{a}_i, & \text{if } j = i-1 \\ \dfrac{\pi_i b_i}{[(a_i + b_i) \vee (a_{i+1} + b_{i+1})]^\alpha} =: \pi_i \tilde{b}_i, & \text{if } j = i+1. \end{cases}$$

Then

$$k^{(\alpha)'} \leqslant k^{(\alpha)} \leqslant \frac{J^{(\alpha)}(A\times A^c)}{\pi(A)\pi(A^c)} = \frac{\pi_i\tilde{a}_i}{\big(\sum_{j\geqslant i}\pi_j\big)\big(\sum_{j<i}\pi_j\big)} \leqslant \frac{\pi_i\tilde{a}_i}{\pi_0\sum_{j\geqslant i}\pi_j}.$$

This proves the necessity of the condition.

(b) Next, assume that the condition holds. Then for each A with $\pi(A)\in(0,1)$, since the symmetry of A and A^c, we may assume that $0\notin A$. Set $i_0=\min A\geqslant 1$. Then, $A\subset I_{i_0}$, $A^c\subset E\setminus\{i_0\}$ and so

$$\frac{J^{(\alpha)}(A\times A^c)}{\pi(A)\wedge\pi(A^c)} \geqslant \frac{\pi_{i_0}\tilde{a}_{i_0}}{\sum_{j\geqslant i_0}\pi_j} \geqslant c, \qquad \frac{J^{(\alpha)}(A\times A^c)}{\pi(A)\pi(A^c)} \geqslant \frac{\pi_{i_0}\tilde{a}_{i_0}}{(1-\pi_{i_0})\sum_{j\geqslant i_0}\pi_j}.$$

Because A is arbitrary, we obtain the required assertions. □

Theorem 4.2. Let $E=\mathbb{Z}_+$. Suppose that (q_{ij}) has finite range R, that is, $q_{ij}=0$ whenever $|i-j|>R$. Then, we have $\lambda_1>0$ provided

$$\overline{\lim_{i\to\infty}}\sum_j \frac{q_{ij}}{\sqrt{q_i\vee q_j}}(j-i)<0.$$

Proof. Simply take $\varphi_i=i+1$ and $B=\{0,1,\cdots,n\}$ for large n in Theorem 1.3 and then apply Theorem 1.2. □

Similarly, we have the following result.

Theorem 4.3. Let $E=\mathbb{Z}^d$. Suppose that (q_{ij}) has finite range R. Then, we have $\lambda_1>0$ provided

$$\overline{\lim_{|i|\to\infty}}\sum_j \frac{q_{ij}}{\sqrt{q_i\vee q_j}}[|j|-|i|]<0.$$

Proof. Take $\varphi_i=|i|+1$ in Theorem 1.3 and then apply Theorem 1.2. □

Theorem 4.4. Let $E=\mathbb{Z}^d$. If there exists a positive function φ such that

$$\overline{\lim_{|i|\to\infty}}\,\Omega\varphi/\varphi<0,$$

then $\lambda_1>0$.

Proof. Apply Theorem 1.2, Theorem 3.2 and then Theorem 1.4 to the finite sets $\{i\colon |i|\leqslant n\}$. □

The following example, taken from [4], is especially rare and interesting since it exhibits the critical phenomena for the existence of spectral gap. It is now used to justify the power of our results and we should see soon what will happen. Similar examples for diffusion were given in [5] and [10].

Example 4.5. Let $E = \mathbb{Z}_+$ and $a_i = b_i = i^\gamma$ $(i \geqslant 1)$ for some $\gamma > 0$, $a_0 = 0$ and $b_0 = 1$. Then $\lambda_1 > 0$ iff $\gamma \geqslant 2$.

Proof. (a) By part (i) of Theorem 4.1, we have $k^{(1/2)} > 0$ iff $\gamma \geqslant 2$. Thus, by Theorem 1.2, we have $\lambda_1 > 0$ for all $\gamma \geqslant 2$.

(b) Applying Theorem 1.5 to $\varphi_i = 1 + i^{1-\gamma/2}$, it follows that $\lambda_1 = 0$ for all $\gamma \in (1, 2)$.

(c) The conditions of Theorem 4.2 hold whenever $\gamma \geqslant 2$. Hence $\lambda_1 > 0$ for all $\gamma \geqslant 2$.

(d) Next, taking $\varphi_i = \sqrt{i}$ $(i \geqslant 1)$, we see that $\Omega\varphi(i)/\varphi_i = -\frac{1}{4}i^{\gamma-2} + O\big(i^{\gamma-3}\big)$. Then

$$\lim_{i\to\infty} \frac{1}{\varphi_i}\Omega\varphi(i) = \begin{cases} -\infty, & \text{if } \gamma > 2 \\ -\frac{1}{4}, & \text{if } \gamma = 2. \end{cases}$$

By Theorem 4.4, we have $\lambda_1 > 0$ for all $\gamma \geqslant 2$.

On the other hand, take $f_n(i) = i^{(\gamma-1)/2} \wedge n^{(\gamma-1)/2}$ and $A = \{0\}$. Then

$$\begin{aligned} \lambda_0(A^c) &\leqslant \varliminf_{n\to\infty} \frac{\sum_{i,j\geqslant 0} \pi_i q_{ij}[f_n(j) - f_n(i)]^2}{2\sum_{i\geqslant 0} \pi_i f_n(i)^2} \\ &= \varliminf_{n\to\infty} \frac{\sum_{i\geqslant 0} \pi_i q_{i,i+1}[f_n(i+1) - f_n(i)]^2}{2\sum_{i\geqslant 0} \pi_i f_n(i)^2} \\ &\leqslant \varliminf_{n\to\infty} \frac{1 + (\gamma-1)^2 \sum_{i=1}^n i^{\gamma-3}}{\sum_{i=1}^n i^{-1}} = 0, \qquad 1 < \gamma < 2. \end{aligned}$$

By Theorem 1.4, we get $\lambda_1 \leqslant \lambda_0(A^c)/\pi(A) = 0$. The case that $\gamma \leqslant 1$ can be ignored since then the chain is not positive recurrent. □

Thus, we have seen that all the results presented in this paper, except Theorem 2.5 which does not work for this example, are qualitatively sharp for this example since every one covers the required region and there is no gap left. Finally, taking $\alpha = 0$ in part (i) of Theorem 4.1, we obtain $k \geqslant \big(\sum_{i=1}^\infty i^{-\gamma}\big)^{-1} > 0$ for all $\gamma > 1$. In other words, we have $k > 0$ but $\lambda_1 = 0$ for all $\gamma \in (1, 2)$. Therefore, the condition "$k > 0$" is not enough but "$k^{(1/2)} > 0$" is sufficient for $\lambda_1 > 0$.

The next two examples show that the two approaches used in the paper for Cheeger's inequalities may all attain sharp estimates but they are not comparable (remember that Theorem 2.5 is not suitable for Example 4.5). We mention that as far as we know, no optimal estimate provided by Cheeger's technique has appeared before.

Example 4.6. Let $E = \mathbb{Z}_+$ and take $a_i \equiv a$ and $b_i \equiv b$ with $a > b > 0$. Then, both Theorem 1.2 and Theorem 2.5 are sharp.

Proof. This is a standard example which is often used to justify the power of a method. It is well known that $\lambda_1 = \big(\sqrt{a} - \sqrt{b}\big)^2$ (cf. [3], Example 9.22 and [4]).

(a) By part (i) of Theorem 4.1, we have

$$k^{(\alpha)\prime} \geqslant \inf_{i\geqslant 1} \frac{\pi_i a_i}{(a+b)^\alpha \sum_{j\geqslant i} \pi_j} = \frac{a-b}{(a+b)^\alpha}.$$

Then, by Theorem 1.2, we get $\lambda_1 \geqslant \left(\sqrt{a} - \sqrt{b}\right)^2$.

(b) Take $p_i \equiv a + b$. Then by part (ii) of Theorem 4.1,

$$k'_p \geqslant \inf_{i\geqslant 1} \frac{\pi_i a_i}{\sum_{j\geqslant i} \pi_j p_j} = \frac{a-b}{a+b}.$$

The same estimate as in (a) now follows from Theorem 2.5. □

Example 4.7. Let $E = \mathbb{Z}_+$ and take $q_{0k} = \beta_k > 0$ (be careful to distinguish the sequence (β_k) and the constant β_p), $q_{k0} = 1/2$ for $k \geqslant 1$ and $q_{ij} = 0$ for all other $i \neq j$. Assume that $q_0 = \sum_{k\geqslant 1} \beta_k < \infty$. Then, Theorem 2.5 is sharp for all q_0 but Theorem 1.2 is sharp only for $q_0 \leqslant 1/2$.

Proof. From $\pi_0 q_{0k} = \pi_k q_{k0}$, it follows that $\pi_k = 2\pi_0\beta_k$, $k \geqslant 1$ and $\pi_0 = (1 + 2q_0)^{-1}$. An interesting point of the example is that the decay of $\sum_{j\geqslant i} \pi_j$ as $i \to \infty$ can be arbitrarily slow, not necessarily exponential. The last condition is necessary for $\lambda_1 > 0$ for the birth-death processes with rates bounded below (by a positive constant) and above (cf. [3], Corollary 9.19 (4)).

(a) Take $p_i = q_i \vee (1/2)$, then $\alpha_p = 1/2$. Without loss of generality, assume that $0 \notin A$. Then

$$\begin{aligned}
\frac{1}{\beta_p} \frac{J(A\times A^c)}{\pi_p(A)\wedge\pi_p(A^c)} &= \frac{\sum_{i\in A} \pi_i q_{i0}}{\left(\sum_{i\in A} 2\pi_0\beta_i p_i\right) \wedge \left(\pi_0 p_0 + \sum_{i\notin A,\, i\neq 0} 2\pi_0\beta_i p_i\right)} \\
&= \frac{\sum_{i\in A} \beta_i}{\left(\sum_{i\in A} 2\beta_i p_i\right) \wedge \left(p_0 + \sum_{i\notin A,\, i\neq 0} 2\beta_i p_i\right)} \\
&= \frac{\sum_{i\in A} \beta_i}{\left(\sum_{i\in A} \beta_i\right) \wedge \left(p_0 + \sum_{i\notin A,\, i\neq 0} \beta_i\right)} \\
&\geqslant 1.
\end{aligned}$$

This gives us $k'_p \geqslant 1$ and hence by Theorem 2.5,

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

Actually, every equality in the last line must hold.

(b) Again, assume that $0 \notin A$. Then

$$\begin{aligned}
\frac{J^{(\alpha)}(A\times A^c)}{\pi(A)\wedge\pi(A^c)} &= \frac{\sum_{i\in A} \pi_i q_{i0}(q_i \vee q_0)^{-\alpha}}{\left(\sum_{i\in A} 2\pi_0\beta_i\right) \wedge \left(\pi_0 + \sum_{i\notin A,\, i\neq 0} 2\pi_0\beta_i\right)} \\
&= \frac{1}{2} \frac{\sum_{i\in A} 2\beta_i}{\left(\frac{1}{2} \vee q_0\right)^{\alpha} \left[\left(\sum_{i\in A} 2\beta_i\right) \wedge \left(1 + \sum_{i\notin A,\, i\neq 0} 2\beta_i\right)\right]} \\
&= \frac{1}{2} \frac{1}{\left(\frac{1}{2} \vee q_0\right)^{\alpha}} \frac{\sum_{i\in A} \beta_i}{\left(\sum_{i\in A} \beta_i\right) \wedge \left(1/2 + \sum_{i\notin A,\, i\neq 0} \beta_i\right)} \\
&= \frac{1}{2} \frac{1}{\left(\frac{1}{2} \vee q_0\right)^{\alpha}} \frac{1}{1 \wedge \left[\left(1/2 + \sum_{i\notin A,\, i\neq 0} \beta_i\right) \Big/ \sum_{i\in A} \beta_i\right]}.
\end{aligned}$$

Because $\left(1/2 + \sum_{i\notin A,\, i\neq 0} \beta_i\right) \Big/ \sum_{i\in A} \beta_i$ decreases when A increases, by setting $A = \{i\}$ for a large enough $i \neq 0$, it follows that

$$k^{(\alpha)'} = \inf_{A:\, 0\notin A} \frac{J^{(\alpha)}(A \times A^c)}{\pi(A) \wedge \pi(A^c)} = \frac{1}{2}\left(\frac{1}{2} \vee q_0\right)^{-\alpha}.$$

By Theorem 1.2, we get

$$\lambda_1 \geqslant \frac{1}{2}\left\{1 \vee (2q_0) + \sqrt{\left(1 \vee (2q_0)\right)^2 - 1}\,\right\}^{-1}.$$

Thus, the lower bound is equal to $1/2 = \lambda_1$ iff $q_0 \leqslant 1/2$. □

The following counterexample shows the limitation of Cheeger's inequalities. Of course, the example can be easily handled with the help of some comparison technique. However, this suggests to us that sometimes it is necessary to examine a model carefully before applying the inequalities.

Example 4.8. Consider the birth-death process with $a_{2i-1} = (2i-1)^2$, $a_{2i} = (2i)^4$ and $b_i = a_i$ for all $i \geqslant 1$. Then, we have $k^{(1/2)'} = 0$ and so Theorem 1.2 is not applicable.

Proof. First, applying Theorem 4.4 to $\varphi_i = \sqrt{i}$ or comparing the chain with the one with rates $a_i = b_i = (2i)^2$, one sees that $\lambda_1 > 0$. Next, because $\mu_i = 1/a_i$ (and hence $\pi_i = \mu_i/Z$, where Z is the normalizing constant), we have $\sum_{j\geqslant i} \mu_j = O(i^{-1})$. However, $\sqrt{a_i \vee a_{i-1}} = O(i^2)$. Hence $\sup_{i\geqslant 1} \sqrt{a_i \vee a_{i-1}} \sum_{j\geqslant i} \mu_j = \infty$. This gives us $k^{(1/2)'} = 0$ by part (i) of Theorem 4.1.

Note that the choice $r_{ij} = q_i \vee q_j$ $(i \neq j)$ is usually not optimal in the sense for which (1.8) often becomes inequality rather than equality. However, the improvement provided by an optimal r_{ij} is still not enough to cover this example and so the problem is really due to the limitation of the technique. □

Acknoledgement. We are grateful to a referee for very careful comments on the first version of the paper, which was distributed as MSRI Preprint No. 1998-024. The results were also announced in Chin. Sci. Bull. 1998, 43:14, 1475–1477.

References

1. Bobkov, S. and Ledoux, M. (1997), *Poincaré inequalities and Talagrand's concentration phenomenon for the exponential distribution*, Prob. Th. Rel. Fields 107, 383–400.
2. 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, Ed. R. C. Gunning, Princeton U. Press, Princeton.
3. Chen, M. F. (1992), *From Markov Chains to Non-Equilibrium Particle Systems*, Singapore, World Scientific.
4. Chen, M. F. (1996), *Estimation of spectral gap for Markov chains*, Acta Math. Sin. New Ser. 12:4, 337–360.
5. Chen, M. F. and Wang, F. Y. (1997), *Estimation of spectral gap for elliptic operators*, Trans. Amer. Math. Soc. 349, 1209–1237.
6. Chung, F. R. K. (1997), *Spectral Graph Theory*, CBMS, **92**, AMS, Providence, Rhode Island.

7. Fukushima, M., Oshima, Y. and Takeda, M. (1994), *Dirichlet Forms and Symmetric Markov Processes*, Walter de Gruyter & Co.
8. 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.
9. Saloff-Coste, L. (1997), *Lectures on finite Markov chains*, LNM **1665**, 301–413, Springer-Verlag.
10. Wang, F. Y. (1999), *Existence of spectral gap for elliptic operators.*, Arkiv For Math. 37:3, 395–407.

5. Appendix. Proof of Lemma 2.1, for referee's reference but not for publication.

Because $\alpha \geqslant 0$ is fixed, we can omit the superscript "(α)" everywhere in the proof. Denote by $\tilde{h}$, $\tilde{k}$ and $\tilde{k}'$ the right-hand sides of the above quantities. By taking $f = I_A$, we obtain $h \geqslant \tilde{h}$, $k \geqslant \tilde{k}$ and $k' \geqslant \tilde{k}'$. We now prove the reverse inequalities.

(a) For any $f \geqslant 0$ with $\pi(f) = 1$, let $A_\gamma = \{f > \gamma\}$, $\gamma \geqslant 0$. By the symmetry of J, we have

$$\begin{aligned}
&\frac{1}{2}\int J(\mathrm{d}x, \mathrm{d}y)|f(y) - f(x)| + K(f) \\
&= \int_{\{f(x)>f(y)\}} J(\mathrm{d}x, \mathrm{d}y)[f(x) - f(y)] + K(f) \\
&= \int_0^\infty \mathrm{d}\gamma\Big\{J\big(\{f(x) > \gamma \geqslant f(y)\}\big) + K\big(\{f > \gamma\}\big)\Big\} \\
&= \int_0^\infty \big[J\big(A_\gamma \times A_\gamma^c\big) + K(A_\gamma)\big]\mathrm{d}\gamma \\
&\geqslant h\int_0^\infty \pi\big(A_\gamma\big)\mathrm{d}\gamma \\
&= h\pi(f) \\
&= h.
\end{aligned}$$

Hence $\tilde{h} \geqslant h$.

(b) For any $f \in L^1_+(\pi)$ with $\int \pi(\mathrm{d}x)\pi(\mathrm{d}y)|f(x) - f(y)| = 1$, by a), we have

$$\begin{aligned}
\int J(\mathrm{d}x, \mathrm{d}y)|f(x) - f(y)| &= 2\int \mathrm{d}\gamma J(A_\gamma \times A_\gamma^c) \\
&\geqslant 2k\int_0^\infty \mathrm{d}\gamma\ (\pi \times \pi)(A_\gamma \times A_\gamma^c) \\
&= k\int_0^\infty \pi(\mathrm{d}x)\pi(\mathrm{d}y)|f(x) - f(y)| \\
&= k.
\end{aligned}$$

This proves the first equality of $k^{(\alpha)}$.

Next, we show that

$$\int |f - \pi(f)|\mathrm{d}\pi = \sup_{g:\,\pi(g)=0,\,\inf_{c\in\mathbb{R}}\|g-c\|_\infty\leqslant 1} \int fg\mathrm{d}\pi, \tag{5.1}$$

where $\|\cdot\|_p$ denotes the L^p-norm. First, let $\pi(g) = 0$ with $\inf_{c\in\mathbb{R}} \|g - c\|_\infty \leqslant 1$. Then, because $\pi(g) = 0$ and $\pi(f - \pi(f)) = 0$, we have

$$\int f g \mathrm{d}\pi = \int (f - \pi(f)) g \mathrm{d}\pi = \int (f - \pi(f))(g - c) \mathrm{d}\pi$$

for all $c \in \mathbb{R}$. Hence, by Hölder inequality, we have

$$\left| \int f g \mathrm{d}\pi \right| \leqslant \|f - \pi(f)\|_1 \|g - c\|_\infty$$

for all c. This gives us

$$\left| \int f g \mathrm{d}\pi \right| \leqslant \|f - \pi(f)\|_1 \inf_c \|g - c\|_\infty \leqslant \|f - \pi(f)\|_1.$$

On the other hand, for a given $f \in L^1(\pi)$, set $A_f^+ = \{f \geqslant \pi(f)\}$ and $A_f^- = \{f < \pi(f)\}$. Take $g_0 = I_{A_f^+} - I_{A_f^-} - \pi(A_f^+) + \pi(A_f^-)$. Then, $g_0 \in L^\infty(\pi)$ and $\pi(g_0) = 0$. Finally, take $c_0 = 1 - 2\pi(A_f^+)$. Then, it is easy to check that $\inf_c \|g_0 - c\|_\infty = \|g_0 - c_0\|_\infty = 1$. Therefore, we have

$$\int f g_0 \mathrm{d}\pi = \int |f - \pi(f)| \mathrm{d}\pi$$

as required.

We now prove the second equality of $k^{(\alpha)}$. Let $f \geqslant 0$ and set $A_\gamma = \{f \geqslant \gamma\}$. Again, by using (a) and (5.1), we have

$$\begin{aligned}
\int J(\mathrm{d}x, \mathrm{d}y)|f(y) - f(x)| &\geqslant 2k \int_0^\infty \mathrm{d}\gamma \pi(A_\gamma)\pi(A_\gamma^c) \\
&= k \int_0^\infty \mathrm{d}\gamma \int \left| I_{A_\gamma} - \pi(A_\gamma) \right| \mathrm{d}\pi \\
&= k \int_0^\infty \mathrm{d}\gamma \sup_{g:\, \pi(g)=0,\, \inf_{c\in\mathbb{R}} \|g-c\|_\infty \leqslant 1} \int I_{A_\gamma} g \mathrm{d}\pi \\
&\geqslant k \sup_{g:\, \pi(g)=0,\, \inf_{c\in\mathbb{R}} \|g-c\|_\infty \leqslant 1} \int_0^\infty \mathrm{d}\gamma \int I_{A_\gamma} g \mathrm{d}\pi \\
&= k \sup_{g:\, \pi(g)=0,\, \inf_{c\in\mathbb{R}} \|g-c\|_\infty \leqslant 1} \int f g \mathrm{d}\pi \\
&= k \int |f - \pi(f)| \mathrm{d}\pi.
\end{aligned}$$

Therefore, we obtain $\tilde{k} \geqslant k$.

(c) Choose $c_0 \in \mathbb{R}$ such that $\pi(f < c_0)$, $\pi(f > c_0) \leqslant 1/2$. Let $f_\pm = (f - c_0)^\pm$. Then we have $f_+ + f_- = |f - c_0|$ and $\pi(|f - c_0|) = \min_c \pi(|f - c|)$. For any $\gamma \geqslant 0$,

define $A_\gamma^\pm = \{f_\pm > \gamma\}$. We have

$$\begin{aligned}\frac{1}{2}\int J(\mathrm{d}x,\mathrm{d}y)|f(y)-f(x)| &= \frac{1}{2}\int J(\mathrm{d}x,\mathrm{d}y)\big[|f_+(y)-f_+(x)|+|f_-(y)-f_-(x)|\big]\\ &= \int_0^\infty \big[J\big(A_\gamma^+\times A_\gamma^{+c}\big)+J\big(A_\gamma^-\times A_\gamma^{-c}\big)\big]\mathrm{d}\gamma\\ &\geqslant k'\int_0^\infty \big[\pi\big(A_\gamma^+\big)+\pi\big(A_\gamma^-\big)\big]\mathrm{d}\gamma = k'\pi(f_++f_-)\\ &= k'\pi(|f-c_0|)\\ &= k'\min_c \pi(|f-c|).\end{aligned}$$

This implies that $\tilde{k}' \geqslant k'$. $\square$

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

Sci. Sin. (A) 1999, 42:8, 805–815

ANALYTIC PROOF OF DUAL VARIATIONAL FORMULA FOR THE FIRST EIGENVALUE IN DIMENSION ONE*

Mu-Fa Chen

(Dept. of Math., Beijing Normal University, 100875)
Received April 13, 1998

Abstract. The first non-zero eigenvalue is the leading term in the spectrum of a self-adjoint operator. It plays a critical role in various applications and is treated in a large number of textbooks. There is a well known variational formula for it (called the Min-Max Principle) which is especially effective for an upper bound of the eigenvalue. However, for the lower bound of the spectral gap, some dual variational formulas have been obtained only very recently. The original proofs are probabilistic. Some analytic proofs in one-dimensional case and certain extension is made in the paper.

Keywords The first eigenvalue variational formula Neumann and Dirichlet eigenvalue elliptic operator infinite matrix

§1. Introduction. Neumann Eigenvalue

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

$$\pi(\mathrm{d}x) = \frac{1}{Za(x)} \exp[C(x)]\mathrm{d}x.$$

*Project supported in part by National Natural Science Foundation of China (No. 19631060), Qiu Shi Science & Technology Foundation, DPFIHE, MCSEC and MCMCAS.

On $L^2(\pi)$, the operator L has the 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.

$$\lambda_1 = \inf\{D(f,f) : f \in C^1[0,D],\ \pi(f) = 0 \text{ and } \pi(f^2) = 1\}, \tag{1.1}$$

where $D(f,f) = \int_0^D a(x)f'(x)^2\pi(\mathrm{d}x)$ and $\pi(f) = \int f\mathrm{d}\pi$. Actually, this formula is valid in completely general situation (refer to [1; 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.

Before moving further, let us make a remark about the definition of λ_1. In the compact case (i.e., $D < \infty$), the spectrum of L is discrete and hence $\lambda_1 > 0$. This may no longer be true in the non-compact case and moreover, an eigenfunction g (i.e. $Lg = -\lambda_1 g$) with respect to λ_1 given by (1.1) may not exist. Therefore, λ_1 may not be an eigenvalue in the ordinary sense. Compared with the solution to the equation, much weaker regularity condition on the coefficients a and b is needed in (1.1). Because of these reasons, hereinafter, we adopt (1.1) or equivalently,

$$\lambda_1 = \inf\{-(f,Lf) : f \in \mathscr{D}(L),\ \pi(f) = 0,\ \pi(f^2) = 1\},$$

as the definition of λ_1(cf. [1; Chapter 9]), here L is regarded as the L^2-operator with domain $\mathscr{D}(L)$.

It is well known that estimating the lower bound of λ_1 is a much harder problem. Even in the present simple situation, only very recently the following dual variational formula has been presented[2].

Theorem 1.1. Let $\mathscr{F} = \{f : f' > 0 \text{ on } (0,D) \text{ and } \pi(f) \geqslant 0\}$. Then, we have

$$\lambda_1 \geqslant \sup_{f\in\mathscr{F}} \inf_{x\in(0,D)} \left\{ \frac{e^{-C(x)}}{f'(x)} \int_x^D \frac{f(u)c^{C(u)}}{a(u)}\mathrm{d}u \right\}^{-1}. \tag{1.2}$$

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 word "dual" comes from the fact that the "inf" in (1.1) is replaced with "sup" in (1.2). It is now quite easy to get a meaningful lower bound of λ_1 by applying (1.2) to a suitable test function $f \in \mathscr{F}$. Note that there is no common point between (1.1) and (1.2). 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. The result was proved in [2] by using a probabilistic approach (i.e., the coupling method). The main purpose of this note is to introduce an analytic proof of (1.2) based on (1.1), as well as all the one-dimensional results presented in [2]–[4] with some extension. It is worthy to mention that for the higher dimensional situation, the coupling method enables us to reduce the general problem to compute the distance, which then turns to the problem on the half-line only. In other words, the one-dimensional result is the key step from the

result of ref. [5] to the general formulas for the lower bound of spectral gap in refs. [2]–[4]. Finally, one may combine Theorem 1.1 with Bakry and Qian[1] to deduce an analytic proof for the general formula for Laplacian on compact manifolds.

The remainder of the note is organized as follows. The proof of Theorem 1.1 is given right below to illustrate more or less the main technique adopted in the paper. The approach enables us to avoid a localizing procedure used in the original proofs. In the next section, we study the full-line or the mixed eigenvalue problem on an interval. The last section is devoted to the discrete space, i.e., we deal with some infinite matrices instead of the differential operators studied here. In particular, we will show by an example the limitation of the present technique.

Proof of Theorem 1.1. 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

$$\begin{aligned}
1 &= \frac{1}{2} \int_0^D \pi(\mathrm{d}x)\pi(\mathrm{d}y)[g(y) - g(x)]^2 \\
&= \int_{\{x \leqslant y\}} \pi(\mathrm{d}x)\pi(\mathrm{d}y) \bigg(\int_x^y g'(u)\sqrt{f'(u)}/\sqrt{f'(u)}\, \mathrm{d}u \bigg)^2 \\
&\leqslant \int_{\{x \leqslant y\}} \pi(\mathrm{d}x)\pi(\mathrm{d}y) \int_x^y g'(u)^2 f'(u)^{-1} \mathrm{d}u \int_x^y f'(\xi)\mathrm{d}\xi \\
&\qquad \text{(by Cauchy-Schwarz inequality)} \\
&= \int_{\{x \leqslant y\}} \pi(\mathrm{d}x)\pi(\mathrm{d}y) \int_x^y a(u)g'(u)^2 e^{C(u)} \frac{e^{-C(u)}}{a(u)f'(u)} \mathrm{d}u \big[f(y) - f(x)\big] \\
&= \int_0^D a(u)g'(u)^2 \pi(\mathrm{d}u) \frac{Ze^{-C(u)}}{f'(u)} \int_0^u \pi(\mathrm{d}x) \int_u^D \pi(\mathrm{d}y) \big[f(y) - f(x)\big].
\end{aligned} \tag{1.3}$$

But

$$\begin{aligned}
&\int_0^u \pi(\mathrm{d}x) \int_u^D \pi(\mathrm{d}y) \big[f(y) - f(x)\big] \\
&\quad = \int_0^u \pi(\mathrm{d}x) \int_u^D f(y)\pi(\mathrm{d}y) - \int_0^u f(x)\pi(\mathrm{d}x) \int_u^D \pi(\mathrm{d}y) \\
&\quad = \int_u^D f(y)\pi(\mathrm{d}y) - \int_u^D \pi(\mathrm{d}x) \int_u^D f(y)\pi(\mathrm{d}y) - \int_0^u f(x)\pi(\mathrm{d}x) \int_u^D \pi(\mathrm{d}y) \\
&\quad = \int_u^D f(y)\pi(\mathrm{d}y) - \bigg[\int_u^D \pi(\mathrm{d}x) \bigg] \int_0^D f(y)\pi(\mathrm{d}y) \leqslant
\end{aligned}$$

[1]Bakry, D. Qian, Z. M., Comparison theorem for spectral gap via dimension, diameter and Ricci curvature, preprint 1998

$$\leqslant \int_u^D f(y)\pi(\mathrm{d}y) \quad (\text{since } \pi(f) \geqslant 0)$$
$$= \frac{1}{Z}\int_u^D \frac{f(y)e^{C(y)}}{a(y)}\,\mathrm{d}y.$$

Combining this with (1.3), 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 (1.2) follows by making the infimum over g and then the supremum over $f \in \mathscr{F}$.

b) The proof of the last assertion of Theorem 1.1 is more technical but it was largely done in [2]. By the assumption, there exists an $f \in C^2[0,D]$ such that $af''+bf' = -\lambda_1 f$ on $[0,D]$ with $f'(0) = f'(D) = 0$. Then, by [2; Proposition 6.4] and the discussion above [2; Lemma 6.2], it follows that $\{I(f)(x)\}^{-1} \geqslant \lambda_1$ for all $x \in (0,D)$. Thus, the equality in (1.2) must hold. Indeed, by [6; Lemma 2.3], we have moreover $\pi(f) = 0$. □

§2. The closed and Mixed Eigenvalues

In this section, we first study the closed eigenvalue problem for the operator $L = a(x)\,\mathrm{d}^2/\mathrm{d}x^2 + b(x)\,\mathrm{d}/\mathrm{d}x$ on the full-line and then the mixed eigenvalue on an interval.

In the present situation, we use the same function $C(x)$ and the probability measure $\pi(\mathrm{d}x)$ introduced in the last section but now

$$Z := \int_{\mathbb{R}} e^{C(x)}/a(x)\mathrm{d}x < \infty.$$

The definition of λ_1 is the same as in (1.1) but with redefined

$$D(f,f) = \int_{\mathbb{R}} a(x)f'(x)^2\pi(\mathrm{d}x).$$

Next, set

$$\mathscr{F} = \{f \in C^1(\mathbb{R}) : f' > 0 \text{ and } \pi(f) = 0\}.$$

For each $f \in \mathscr{F}$, denote by $x_0 = x_0(f)$ the unique zero-point of f and put

$$I^{\pm}(f)(x) = \frac{e^{-C(x)}}{f'(x)}\int_x^{\pm\infty} \frac{f(u)e^{C(u)}\mathrm{d}u}{a(u)}, \qquad \pm(x-x_0) > 0,$$
$$\delta^{\pm}(f) = \sup_{\pm(x-x_0)>0} I^{\pm}(f)(x).$$

Theorem 2.1. We have $\lambda_1 \geqslant \sup_{f\in\mathscr{F}} [\delta^+(f) \vee \delta^-(f)]^{-1}$.

Under mild assumption, the above equality can also hold. Refer to [2; §7].

Proof of Theorem 2.1. Given $f \in \mathscr{F}$, to simplify the notation, assume that $x_0 = 0$. Similar to (1.3), we have

$$\begin{aligned} 1 &\leqslant \int_{\{x\leqslant y\}} \pi(\mathrm{d}x)\pi(\mathrm{d}y) \int_x^y \frac{g'(u)^2}{f'(u)}\mathrm{d}u \big[f(y)-f(x)\big] \\ &= \int_0^\infty a(u)g'(u)^2\pi(\mathrm{d}u)\frac{Ze^{-C(u)}}{f'(u)} \int_{-\infty}^u \pi(\mathrm{d}x)\int_u^\infty \pi(\mathrm{d}y)\big[f(y)-f(x)\big] \\ &\quad + \int_{-\infty}^0 a(u)g'(u)^2\pi(\mathrm{d}u)\frac{Ze^{-C(u)}}{f'(u)} \int_{-\infty}^u \pi(\mathrm{d}x)\int_u^\infty \pi(\mathrm{d}y)\big[f(y)-f(x)\big]. \end{aligned} \tag{2.1}$$

Next, for each $u \geqslant 0$, we have

$$\begin{aligned} &\int_{-\infty}^u \pi(\mathrm{d}x)\int_u^\infty \pi(\mathrm{d}y)\big[f(y)-f(x)\big] \\ &\quad = \int_{-\infty}^u \pi(\mathrm{d}x)\int_u^\infty f(y)\pi(\mathrm{d}y) - \int_{-\infty}^u f(x)\pi(\mathrm{d}x)\int_u^\infty \pi(\mathrm{d}y) \\ &\quad = \int_u^\infty f(y)\pi(\mathrm{d}y) - \bigg[\int_u^\infty \pi(\mathrm{d}x)\bigg]\bigg[\int_u^\infty f(y)\pi(\mathrm{d}y) + \int_{-\infty}^u f(y)\pi(\mathrm{d}y)\bigg] \\ &\quad = \int_u^\infty f(y)\pi(\mathrm{d}y) - \bigg[\int_u^\infty \pi(\mathrm{d}x)\bigg]\int_{-\infty}^{+\infty} f(y)\pi(\mathrm{d}y) \\ &\quad = \int_u^\infty f(y)\pi(\mathrm{d}y) \quad (\text{since } \pi(f)=0). \end{aligned} \tag{2.2}$$

On the other hand, for each $u \leqslant 0$, we have

$$\begin{aligned} &\int_{-\infty}^u \pi(\mathrm{d}x)\int_u^\infty \pi(\mathrm{d}y)\big[f(y)-f(x)\big] \\ &\quad = \int_{-\infty}^u \pi(\mathrm{d}x)\int_u^\infty f(y)\pi(\mathrm{d}y) - \int_{-\infty}^u f(x)\pi(\mathrm{d}x)\int_u^\infty \pi(\mathrm{d}y) \\ &\quad = \int_u^{-\infty} f(x)\pi(\mathrm{d}x) + \int_{-\infty}^u f(x)\pi(\mathrm{d}x)\int_{-\infty}^u \pi(\mathrm{d}y) \\ &\qquad + \int_{-\infty}^u \pi(\mathrm{d}x)\int_u^\infty f(y)\pi(\mathrm{d}y) \\ &\quad = \int_u^{-\infty} f(x)\pi(\mathrm{d}x) + \bigg[\int_{-\infty}^u \pi(\mathrm{d}x)\bigg]\int_{-\infty}^{+\infty} f(y)\pi(\mathrm{d}y) \\ &\quad = \int_u^{-\infty} f(x)\pi(\mathrm{d}x). \end{aligned} \tag{2.3}$$

Combining (2.1)-(2.3), we obtain

$$\begin{aligned}1 &\leqslant \int_0^\infty a(x)g'(x)^2\pi(\mathrm{d}x)\frac{e^{-C(x)}}{f'(x)}\int_x^\infty \frac{f(u)e^{C(u)}}{a(u)}\mathrm{d}u \\ &\quad + \int_{-\infty}^0 a(x)g'(x)^2\pi(\mathrm{d}x)\frac{e^{-C(x)}}{f'(x)}\int_x^{-\infty} \frac{f(u)e^{C(u)}}{a(u)}\mathrm{d}u \\ &\leqslant \delta^+(f)\int_0^\infty a(x)g'(x)^2\pi(\mathrm{d}x) + \delta^-(f)\int_{-\infty}^0 a(x)g'(x)^2\pi(\mathrm{d}x) \\ &\leqslant \big[\delta^+(f)\vee\delta^-(f)\big]\int_{-\infty}^{+\infty} a(x)g'(x)^2\pi(\mathrm{d}x).\end{aligned}$$

Now, the assertion of the theorem follows immediately. □

We now turn to study the mixed or Dirichlet eigenvalue problem. That is the same operator on $[0,D]\,(D\leqslant\infty)$ with Dirichlet boundary at 0 and Neumann boundary at D when $D<\infty$. The problem not only has its own interest but also plays a key role in the higher-dimensional situation since the coupling method reduces the general case to the present one [3]–[5].

Recall that

$$\lambda_1 = \inf\{D(f,f) : f(0)=0,\ f'(D)=0 \text{ and } \pi(f^2)=1\},$$

where $D(f,f)$ is the same as in Section 1.

Theorem 2.2. Set $\mathscr{F}=\{f\in C^1[0,D] : f(0)=0 \text{ and } f'>0 \text{ on } (0,D)\}$. Then we have

$$\lambda_1 \geqslant \sup_{f\in\mathscr{F}}\inf_{x\in(0,D)} I(f)(x)^{-1},$$

where $I(f)$ is the same as in Section 1. 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 $D<\infty$ [See also Appendix to the paper [8] in this book].

Proof. a) The proof for the first assertion is quite similar to the proof a) of Theorem 1.1. Given $f\in\mathscr{F}$, for every g with $g(0)=0$ and $\pi(g^2)=1$, we have

$$\begin{aligned}1 &= \int_0^D g(x)^2\pi(\mathrm{d}x) \\ &= \int_0^D \big(g(x)-g(0)\big)^2\pi(\mathrm{d}x) \\ &= \int_0^D \left(\int_0^x \frac{g'(u)\sqrt{f'(u)}}{\sqrt{f'(u)}}\mathrm{d}u\right)^2\pi(\mathrm{d}x) \\ &\leqslant \int_0^D \pi(\mathrm{d}x)\int_0^x \frac{g'(u)^2}{f'(u)}\mathrm{d}u\int_0^x f'(\xi)\mathrm{d}\xi \\ &= \int_0^D \pi(\mathrm{d}u)a(u)g'(u)^2\frac{Ze^{-C(u)}}{f'(u)}\int_u^D \pi(\mathrm{d}x)f(x) \\ &\leqslant D(g,g)\sup_{x\in(0,D)}\frac{e^{-C(x)}}{f'(x)}\int_x^D\frac{f(u)e^{C(u)}}{a(u)}\mathrm{d}u.\end{aligned}$$

b) To prove the last assertion, we again follow [2]. By assumption, there exists an f such that $-af'' - bf' = \lambda_1 f$ with $f(0) = 0$ and $f'(D) = 0$ when $D < \infty$. We first show that $f' > 0$ on $(0, D)$. Note that [2; Lemma 6.2 and Lemma 6.3] are valid. Moreover, the proof a) of [2; Proposition 6.3] shows that f is not a constant on $[0, p)$ and $f(p) \neq 0$, provided $f'(p) = 0$ for some $p \in (0, D)$. Take $g = fI_{[0,p]} + f(p)I_{(p,D]}$. Then g is not a constant, $g(0) = 0$ and $g'(D) = 0$. Now we have

$$\pi(ag'^2) = \int_0^p af'^2 \mathrm{d}\pi = -\int_0^p (fLf)\mathrm{d}\pi = \lambda_1 \int_0^p f^2 \mathrm{d}\pi,$$
$$\pi(g^2) = \int_0^p f^2 \mathrm{d}\pi + f(p)^2 \pi(p, D).$$

Hence

$$\lambda_1 \leqslant \frac{\pi(ag'^2)}{\pi(g^2)} \leqslant \frac{\lambda_1 \int_0^p f'^2 \mathrm{d}\pi}{\int_0^p f^2 \mathrm{d}\pi + f(p)^2 \pi(p, D)} < \lambda_1.$$

We have thus proved $f' > 0$ on $[0, D)$. Next, since

$$(f'e^C)' = (af'' + bf')e^C/a,$$

by the boundary condition at D, we get

$$-\int_x^D \frac{\lambda_1 f}{a} e^C = (f'e^C)|_x^D = (f'e^C)(D) - (f'e^C)(x) = -(f'e^C)(x).$$

That is $I(f) \equiv \lambda_1^{-1}$ on $(0, D)$. $\square$

§3. Discrete Case

Consider a class of matrices $Q = (q_{ij})$ 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 $\pi_i q_{ij} = \pi_j q_{ji}$ for a probability measure $(\pi_i > 0 : i \in E)$ and all i, j. Then the corresponding operator

$$\Omega f(i) := \sum_j q_{ij}(f_j - f_i), \qquad i \in E$$

becomes symmetric on $L^2(\pi)$, for which we have

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

with domain $\mathscr{D}(D) = \{f \in L^2(\pi) : D(f, f) < \infty\}$. Next, by [1; Theorem 9.9 and Theorem 6.61], we have

$$\lambda_1 = \inf\{D(f, f) : \pi(f) = 0 \text{ and } \pi(f^2) = 1\}.$$

We now define a graph structure associated with the matrix $Q = (q_{ij})$. We call $\langle ij\rangle$ an edge if $q_{ij} > 0$ $(i \neq j)$. The adjacent edges $\langle ii_1\rangle, \langle i_1i_2\rangle, \cdots, \langle i_nj\rangle$ (i, j and i_k's are different) consists a path from i to j. Assume that for each pair $i \neq j$, there exists a path from i to j. Choose and fix such a path γ_{ij}. Next, define a positive weight function $\{w(e)\}$ on the edges $e = \langle ij\rangle$ and set $|\gamma_{ij}|_w = \sum_{e\in\gamma_{ij}} w(e)$. Put $a(e) = \pi_i q_{ij}$ if $e = \langle ij\rangle$ and set

$$I(w)(e) = \frac{1}{a(e)w(e)} \sum_{\{i,j\}:\, \gamma_{ij}\ni e} |\gamma_{ij}|_w \pi_i \pi_j,$$

where $\{i, j\}$ denotes the disordered pair of i and j.

Theorem 3.1. We have $\lambda_1 \geqslant \sup_{w\in\mathscr{W}} \inf_e I(w)(e)^{-1}$.

Proof. For simplicity, we write $f(e) = f_j - f_i$ if $e = \langle ij\rangle$. By Cauchy-Schwarz inequality, we get

$$(f_i - f_j)^2 = \Big(\sum_{e\in\gamma_{ij}} f(e)\Big)^2 \leqslant \Big(\sum_{e\in\gamma_{ij}} \frac{f(e)^2}{w(e)}\Big)|\gamma_{ij}|_w.$$

Thus, for each f with $\pi(f) = 0$ and $\pi(f^2) = 1$, we have

$$\begin{aligned}
1 &= \frac{1}{2}\sum_{i,j} \pi_i\pi_j (f_i - f_j)^2 \\
&= \sum_{\{i,j\}} \pi_i\pi_j \Big(\sum_{e\in\gamma_{ij}} f(e)\Big)^2 \\
&\leqslant \sum_{\{i,j\}} \pi_i\pi_j \Big(\sum_{e\in\gamma_{ij}} \frac{f(e)^2}{w(e)}\Big)|\gamma_{ij}|_w \\
&= \sum_e a(e) f(e)^2 \frac{1}{a(e)w(e)} \sum_{\{i,j\}:\, \gamma_{ij}\ni e} |\gamma_{ij}|_w \pi_i\pi_j \\
&\leqslant D(f, f) \sup_e I(w)(e). \quad \square
\end{aligned}$$

From the proof, one sees that the use of the graphic structure is quite natural since only those pair $\{i, j\}$ with $q_{ij} > 0$ appear in the Dirichlet form $D(f, f)$. However, we now show by an example that the graphic structure is not completely necessary.

Example. Take $E = \mathbb{Z}_+$, $q_{0i} = \beta_i > 0$, $q_0 = \sum_{k=1}^{\infty} \beta_k < \infty$, $q_{i0} = 1/2$ $(i \geqslant 1)$ and $q_{ij} = 0$ for other $i \neq j$. Then $\pi_0 = (1 + 2q_0)^{-1}$, $\pi_k = 2\pi_0\beta_k$ $(k \geqslant 1)$. It is easy to check that $\lambda_1 = 1/2$.

For each $i \neq 0$, there is only one path (without loop) γ_{0i} consisting of the single edge $\langle 0i\rangle$ and for each pair $i, j \neq 0$, there is only one path γ_{ij} consisting of the

edges $\langle i0\rangle$ and $\langle 0j\rangle$. Denoting by w_ℓ a weight on the edge $\langle 0\ell\rangle$ ($\ell \geqslant 1$), we have

$$\begin{aligned} I_\ell(w) &= \frac{1}{a(\langle 0\ell\rangle)w_\ell} \sum_{\{i,j\}:\,\gamma_{ij}\ni\langle 0\ell\rangle} \pi_i\pi_j|\gamma_{ij}|_w \\ &= \frac{1}{\pi_0\beta_\ell w_\ell}\Big[\sum_{j\neq 0,\ell}\pi_\ell\pi_j(w_\ell+w_j)+\pi_0\pi_\ell w_\ell\Big] \\ &= \frac{1}{\pi_0\beta_\ell}\Big[\sum_{j\neq 0,\ell}4\pi_0^2\beta_\ell\beta_j(1+w_j/w_\ell)+2\pi_0^2\beta_\ell\Big] \\ &= 2\pi_0\Big[1+\sum_{j\neq 0,\ell}2\beta_j(1+w_j/w_\ell)\Big] \\ &= 2\pi_0\Big[1+2\sum_{j\geqslant 1}\beta_j+2\sum_{j\geqslant 1}\beta_j w_j/w_\ell-4\beta_\ell\Big] \\ &= 2\pi_0\Big[1+2q_0+2\sum_{j\geqslant 1}\beta_j w_j/w_\ell-4\beta_\ell\Big]. \end{aligned}$$

By Choosing $w_i\equiv 1$, we get a non-trivial lower bound:

$$\lambda_1 \geqslant \inf_{\ell\geqslant 1} I_\ell(w)^{-1} = \big[2\pi_0(1+4q_0)\big]^{-1} = \frac{1+2q_0}{2+8q_0} < \frac{1}{2}.$$

Is it possible to get the sharp bound by choosing a better (w_i)? To see this, consider the set

$$\mathscr{W}_1 = \Big\{w\in\mathscr{W}:\ C(w) := \sup_{\ell\geqslant 1}\Big[\sum_{j\geqslant 1}\beta_j w_j/w_\ell - 2\beta_\ell\Big] < \infty\Big\}.$$

Then, for each $w\in\mathscr{W}_1$, we have $\sum_{j\geqslant 1}\beta_j w_j \leqslant \big(C(w)+2\beta_\ell\big)w_\ell < \infty$. This implies that $\inf_{\ell\geqslant 1} w_\ell > 0$. Because of the homogeneous, we may assume, for each $w\in\mathscr{W}_1$, that $\inf_{\ell\geqslant 1} w_\ell = 1$. Then $C(w) \geqslant \sup_{\ell\geqslant 1}\big\{q_0/w_\ell - 2\sup_{j\geqslant 1}\beta_j\big\} = q_0 - 2\sup_{\ell\geqslant 1}\beta_\ell$, $w\in\mathscr{W}_1$. Hence

$$\begin{aligned} \sup_{w\in\mathscr{W}}\inf_{\ell\geqslant 1} I_\ell(w)^{-1} &= \sup_{w\in\mathscr{W}_1}\Big\{\sup_{\ell\geqslant 1} I_\ell(w)\Big\}^{-1} \\ &= \frac{1}{2\pi_0}\sup_{w\in\mathscr{W}_1}\{1+2q_0+2C(w)\}^{-1} \\ &= \frac{1}{2\pi_0}\Big\{1+2q_0+2\inf_{w\in\mathscr{W}_1} C(w)\Big\}^{-1} \\ &\leqslant \frac{1}{2\pi_0}\Big\{1+2q_0+2\Big[q_0-2\sup_{\ell\geqslant 1}\beta_\ell\Big]\Big\}^{-1} \\ &< \frac{1}{2\pi_0(1+2q_0)} \\ &= 1/2 \\ &= \lambda_1, \end{aligned}$$

whenever $q_0 > 2\sup_{\ell\geqslant 1}\beta_\ell$ (a simple example is $\beta_n = n^{-1-\varepsilon}$ for small enough $\varepsilon > 0$). This means that the estimate provided by Theorem 3.1 may not be sharp due to the specific graphic structure. However, it was proved that the coupling method does achieve the sharp estimate[3] (Here, we mention that the condition "$g_2 \geqslant 0 \vee \cdots$" in [3; Example 3.5] can be removed). The sharp estimate can be also achieved by using the Cheeger's inequality[7]. □

Of course, in practice, the key of Theorem 3.1 is the choice of $\{w(e)\}$ especially for infinite E. For finite E, the theorem was appeared in [8] with the simple choice $w(e) = a(e)^{-1}$. Ref.[9] used $w(e)^{-1}$ instead of $w(e)$ used in this paper. Our representation has a meaning that the value of the weight function at an edge is given by the difference of the eigenfunction at the two endpoints of the edge. This corresponds the derivative of the eigenfunction appeared in Theorem 1.1. The idea is illustrated by the following variational formula due to [3].

Theorem 3.2. Let $E = \{0, 1, 2, \cdots, N\}, N \leqslant \infty$, $q_{i,i+1} = b_i > 0$ $(0 \leqslant i \leqslant N-1)$, $q_{i,i-1} = a_i > 0$ $(1 \leqslant i \leqslant N)$ 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

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

Then, we have

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

Proof. a) Recall that the distribution (π_i) is determined by $\pi_i = \mu_i/\sum_{j\geqslant 0}\mu_j$, $0 \leqslant i \leqslant N$. Denote by e_i the edge $\langle i, i+1\rangle$. Clearly, for each pair $i < j$, there is only one path without loop consisting of $e_i, e_{i+1}, \cdots, e_{j-1}$. Take $w(e_i) = w_{i+1} - w_i$. Then

$$|\gamma_{k\ell}|_w = (w_{k+1} - w_k) + \cdots + (w_\ell - w_{\ell-1}) = w_\ell - w_k.$$

Thus,

$$\begin{aligned}\sum_{\{k,\ell\}:\,\gamma_{k\ell}\ni e_i}|\gamma_{k\ell}|_w\pi_k\pi_\ell &= \sum_{k=0}^{i}\sum_{\ell=i+1}^{N}\pi_k\pi_\ell(w_\ell - w_k)\\ &= \sum_{k=0}^{i}\pi_k\sum_{\ell=i+1}^{N}\pi_\ell w_\ell - \sum_{k=0}^{i}\pi_k w_k\sum_{\ell=i+1}^{N}\pi_\ell\\ &= \sum_{\ell=i+1}^{N}\pi_\ell w_\ell - \left(\sum_{k=i+1}^{N}\pi_k\right)\sum_{\ell=i+1}^{N}\pi_\ell w_\ell - \sum_{k=0}^{i}\pi_k w_k\sum_{\ell=i+1}^{N}\pi_\ell\end{aligned}$$

$$= \sum_{\ell=i+1}^{N} \pi_\ell w_\ell - \left(\sum_{k=i+1}^{N} \pi_k\right)\left(\sum_{\ell=0}^{N} \pi_\ell w_\ell\right)$$
$$\leqslant \sum_{\ell=i+1}^{N} \pi_\ell w_\ell, \qquad 0 \leqslant i \leqslant N-1.$$

By Theorem 3.1, we have proved the assertion replacing "=" with "$\geqslant$".

b) To prove that the equality holds, we first show that the present formula coincides with part (2) of [3; Theorem 1.1]. In the last result, w_0 was left to be free and the condition "$\sum_{i=0}^{N} \mu_i w_i \geqslant 0$" here was replaced by "$\sum_{i=1}^{N} \mu_i w_i > 0$". Because of the strictly increasing property, the latter can be implied by the former one. Actually, the conclusion is trivial when $w_0 < 0$ since

$$\sum_{i=1}^{N} \mu_i w_i \geqslant -\mu_0 w_0 = -w_0.$$

On the other hand, if $w_0 \geqslant 0$, then we must have $w_1 > 0$ and hence

$$\sum_{i=1}^{N} \mu_i w_i \geqslant w_1 \sum_{i=1}^{N} \mu_i > 0.$$

Besides,

$$b_0(1+w_1)/\textstyle\sum_{j=1}^{N} \mu_j w_j \geqslant \mu_0 b_0(-w_0+w_1)/\sum_{j=1}^{N} \mu_j w_j$$

whenever $w_0 \geqslant 0$, otherwise we replace (w_i) by $(\widetilde{w}_i = w_i/|w_0|)$. It follows that the initial condition $I_0(w)$ given in [3; (1.3)] is also included in the present $I_0(\widetilde{w})$ for suitable $(\widetilde{w}_i)$ which may be different from the original (w_i). We have thus proved the required assertion.

Finally, we return to prove the equality mentioned above. The key fact is that the eigenfunction g of λ_1 must be strictly increasing [3; Lemma 4.2] and so we may take $w = g$. Moreover, we indeed have $\pi(g) = 0$ (cf. [6; Lemma 2.2]), this gives us $\pi(w) \geqslant 0$ as we required. Then, some computation shows that $I_i(w) \equiv \lambda_1^{-1}$ for this specific $w = g$.

We now prove the strictly increasing property of the eigenfunction g (since the proof d) of [3; Lemma 4.2] contains an error). For convenience, we set $a_0 = 0$ and $b_N = 0$ when $N < \infty$. Let $\lambda_1 > 0$ and g be a solution to the equation $\Omega g = -\lambda_1 g$ with $g_0 < 0$. By [3; Lemma 4.1], we have

$$\pi_n b_n(g_{n+1} - g_n) = -\lambda_1 \textstyle\sum_{i=0}^{n} \pi_i g_i, \qquad 0 \leqslant n \leqslant N. \tag{3.1}$$

Hence $g_1 > g_0$. Suppose that there exists an n with $1 \leqslant n \leqslant N-1$ such that

$$g_0 < g_1 < \cdots < g_{n-1} < g_n \geqslant g_{n+1} \tag{3.2}$$

We are going to prove this is impossible.

By (3.1), we have $g_k < (\text{resp.}=)\, g_{k+1} \Longleftrightarrow \sum_{i=0}^{k} \pi_i g_i < (\text{resp.}=)\, 0$ for $0 \leqslant k \leqslant N-1$. Again, as in [3], set $\widetilde{g}_n = -\sum_{i=0}^{n-1} \pi_i g_i/\pi_n$. Then it was proved in [3] that

$$\sum_{i \leqslant n-1} \pi_i g_i + \pi_n \widetilde{g}_n = 0, \tag{3.4}$$

$$g_n \geqslant \widetilde{g}_n = a_n(g_n - g_{n-1})/\lambda_1 > 0. \tag{3.5}$$

Take $\overline{g}_i = g_i I_{[i<n]} + g_n I_{[i \geqslant n]}$. Then, we have

$$\sum_i \pi_i \overline{g}_i^2 = \sum_{i \leqslant n-1} \pi_i g_i^2 + g_n^2 \sum_{i=n}^{N} \pi_i,$$

$$\sum_i \pi_i \overline{g}_i = \sum_{i \leqslant n-1} \pi_i g_i + g_n \sum_{i=n}^{N} \pi_i = g_n \sum_{i=n}^{N} \pi_i - \pi_n \widetilde{g}_n \qquad (\text{by (3.4)}).$$

Hence

$$\sum_i \pi_i \overline{g}_i^2 - \left(\sum_i \pi_i \overline{g}_i \right)^2 = \sum_{i \leqslant n-1} \pi_i g_i^2 + g_n^2 \sum_{i=n}^{N} \pi_i - \left(g_n \sum_{i=n}^{N} \pi_i - \pi_n \widetilde{g}_n \right)^2. \tag{3.6}$$

Next,

$$\begin{aligned} -\sum_i \pi_i (\overline{g} \Omega \overline{g})(i) &= \lambda_1 \sum_{i \leqslant n-1} \pi_i g_i^2 + \pi_n a_n g_n (g_n - g_{n-1}) \\ &= \lambda_1 \sum_{i \leqslant n-1} \pi_i g_i^2 + \lambda_1 \pi_n g_n \widetilde{g}_n \qquad (\text{by (3.5)}). \end{aligned} \tag{3.7}$$

We now prove that

$$\pi_n g_n \widetilde{g}_n < g_n^2 \sum_{i=n}^{N} \pi_i - \left(g_n \sum_{i=n}^{N} \pi_i - \pi_n \widetilde{g}_n \right)^2. \tag{3.8}$$

Since $g_n > 0$ by (3.5), (3.8) is equivalent to

$$\pi_n \frac{\widetilde{g}_n}{g_n} < \sum_{i=n}^{N} \pi_i - \left(\sum_{i=n}^{N} \pi_i - \pi_n \frac{\widetilde{g}_n}{g_n} \right)^2,$$

or

$$\left(\sum_{i=n}^{N} \pi_i - \pi_n \frac{\widetilde{g}_n}{g_n} \right)^2 < \sum_{i=n}^{N} \pi_i - \pi_n \frac{\widetilde{g}_n}{g_n}.$$

The last inequality holds because $0 < \widetilde{g}_n \leqslant g_n$, $0 < \sum_{i=n}^N \pi_i - \pi_n \widetilde{g}_n / g_n = \sum_{i \geqslant n+1} \pi_i + \pi_n(1 - \tilde{g}_n/g_n) < 1$. We have thus proved (3.8). Combining (3.6)–(3.8), we get

$$\begin{aligned}
\lambda_1 &\leqslant \frac{-\sum_i \pi_i (\overline{g}\Omega\overline{g})(i)}{\sum_i \pi_i \overline{g}_i^2 - (\sum_i \pi_i \overline{g}_i)^2} \\
&= \frac{\lambda_1 \sum_{i \leqslant n-1} \pi_i g_i^2 + \lambda_1 \pi_n g_n \widetilde{g}_n}{\sum_{i \leqslant n-1} \pi_i g_i^2 + g_n^2 \sum_{i=n}^N \pi_i - \left(g_n \sum_{i=n}^N \pi_i - \pi_n \widetilde{g}_n\right)^2} \\
&< \lambda_1,
\end{aligned}$$

which is a contradiction. □

Having Theorem 3.2 at hand, it should not be difficult to study the birth-death processes on the whole $\mathbb{Z}$, as a parallel to the diffusion on the whole line. However, we will not go to this direction.

We now turn to study the Dirichlet eigenvalue for general Markov chains. Fix a point, say $0 \in E$. Then the Dirichlet eigenvalue is defined by

$$\lambda_1 = \inf\{D(f,f) : f(0) = 0 \text{ and } \pi(f^2) = 1\}.$$

For each $i \in E$, choose a path γ_i from 0 to i (without loop). Again, choose a positive weight function $\{w(e)\}$ on the edges and define $|\gamma_i|_w = \sum_{e \in \gamma_i} w(e)$,

$$I(w)(e) = \frac{1}{a(e)w(e)} \sum_{i \neq 0:\, \gamma_i \ni e} |\gamma_i|_w \pi_i.$$

Theorem 3.3. We have $\lambda_1 \geqslant \sup_w \inf_e I(w)(e)^{-1}$.

Proof.

$$\begin{aligned}
1 &= \sum_{i \neq 0} \pi_i f_i^2 \\
&= \sum_{i \neq 0} \pi_i (f_i - f_0)^2 \\
&= \sum_{i \neq 0} \pi_i \left(\sum_{e \in \gamma_i} f(e) \right)^2 \\
&\leqslant \sum_{i \neq 0} \pi_i \sum_{e \in \gamma_i} \frac{f(e)^2}{w(e)} |\gamma_i|_w \\
&= \sum_e a(e) f(e)^2 I(w)(e) \\
&\leqslant D(f,f) \sup_e I(w)(e). \quad \square
\end{aligned}$$

Theorem 3.4. Let (a_i, b_i), (μ_i), $(I_i(w))$ be the same as in Theorem 3.2 but replace $\mathscr{W}$ by the set of all strictly increasing sequence (w_i) with $w_0 = 0$. Then we have

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

When $b_0 = 0$, the conclusion remains true if one redefines

$$I_i(w) = \frac{1}{a_{i+1}\tilde{\mu}_{i+1}(w_{i+1} - w_i)} \sum_{j=i+1}^{N} \tilde{\mu}_j w_j,$$

where

$$\tilde{\mu}_1 = 1, \quad \tilde{\mu}_n = \frac{b_1 \cdots b_{n-1}}{a_2 \cdots a_n}, \quad 2 \leqslant n \leqslant N,$$

$$\widetilde{D}(f) = \sum_{1\leqslant i\leqslant N-1} \tilde{\pi}_i b_i (f_{i+1} - f_i)^2 + \tilde{\pi}_1 a_1 f_1^2,$$

$$\lambda_1 = \inf\{\widetilde{D}(f) : f_0 = 0, \tilde{\pi}(f^2) = 1\}.$$

Proof. a) Again, let e_i be the edge $\langle i, i+1\rangle$. For each $i \geqslant 1$, there is a path consisting of $e_0, e_1, \cdots, e_{i-1}$. Take $w(e_i) = w_{i+1} - w_i$. Then

$$\sum_{k:\,\gamma_k \ni e_i} |\gamma_k|_w \pi_k = \sum_{k=i+1}^{N} (w_k - w_0)\pi_k = \sum_{k=i+1}^{N} \pi_k w_k.$$

Now, the inequality "$\lambda_1 \geqslant \cdots$" follows from Theorem 3.3.

b) The remainder of the proof is similar to the second part of the proof of Theorem 3.2. However, we still present the details here for completeness. Let $\lambda_1 > 0$ and $g \not\equiv 0$ with $g_0 = 0$ be a solution to the equation $\Omega g(i) = -\lambda_1 g_i$, $1 \leqslant i \leqslant N$. Here, we adopt the convention that $a_0 = 0$ and $b_N = 0$. The key to prove the equality is to show the strictly monotonicity of (g_i). Once this is done, without less of generality, assume that $g_i \uparrow$, then we have

$$I_i(g) = \frac{1}{a_{i+1}\mu_{i+1}(g_{i+1} - g_i)} \sum_{j=i+1}^{N} \mu_j g_j \equiv \frac{1}{\lambda_1} \tag{3.9}$$

for all $0 \leqslant i \leqslant N-1$ and hence the required assertion follows.

c) To see that (3.9) holds, first, we show that

$$-\lambda_1 \sum_{1}^{n} \pi_i g_i = \pi_{n+1} a_{n+1}(g_{n+1} - g_n) - \pi_1 a_1 g_1, \qquad 1 \leqslant n \leqslant N. \tag{3.10}$$

Here, we use the convention $a_{N+1} = 0$ provided $N < \infty$. The proof is easy:

$$\begin{aligned}
-\lambda_1 \sum_{1}^{n} \pi_i g_i &= \sum_{1}^{n} \pi_i \Omega g(i) \\
&= \sum_{1}^{n} \big[\pi_i a_i (g_{i-1} - g_i) + \pi_i b_i (g_{i+1} - g_i)\big] \\
&= \sum_{1}^{n} \big[- \pi_i a_i (g_i - g_{i-1}) + \pi_{i+1} a_{i+1} (g_{i+1} - g_i)\big] \\
&= \pi_{n+1} a_{n+1}(g_{n+1} - g_n) - \pi_1 a_1 g_1.
\end{aligned}$$

[Added to the original proof: Let $u_i = g_{i+1} - g_i$, $0 \leqslant i \leqslant N-1$. Even though it is not necessary but for specificity, we set $u_N = 1$ when $N < \infty$. By eigen-equation, we have

$$b_i u_i - a_i u_{i-1} = -\lambda_1 g_i, \qquad 1 \leqslant i \leqslant N.$$

Then,

$$R_i(u) := (a_{i+1}u_i - b_{i+1}u_{i+1} - a_i u_{i-1} + b_i u_i)/u_i = \lambda_1 > 0, \qquad 1 \leqslant i \leqslant N-1.$$

By (3.10) and the assumption $g_i \uparrow\uparrow$, $g_0 = 0$, it follows that

$$0 \leqslant \mu_{n+1}a_{n+1}u_n = \mu_1 a_1 g_1 - \lambda_1 \sum_{i=1}^{n} \mu_i g_i \leqslant \mu_1 a_1 g_1, \qquad 1 \leqslant i \leqslant N-1.$$

Thus, $\mu_{n+1}a_{n+1}u_n$ is decreasing in n and

$$0 \leqslant c := \lim_{n \to N} \mu_{n+1}a_{n+1}u_n \leqslant \mu_1 a_1 g_1.$$

Note that $c = 0$ when $N < \infty$. Next, let

$$w_i = a_i u_{i-1} - b_i u_i + c/(\mu - \mu_0) = \lambda_1 g_i + c/(\mu - \mu_0) > 0, \qquad 1 \leqslant i \leqslant N.$$

Then $(w_{i+1} - w_i)/u_i = R_i(u) = \lambda_1 > 0$ for all $1 \leqslant i \leqslant N-1$. This implies that $w_i \uparrow\uparrow$. Therefore,

$$\begin{aligned}
\sum_{j=i+1}^{N} \mu_j w_j &= \sum_{j=i+1}^{N} (\mu_j a_j u_{j-1} - \mu_j b_j u_j) + \frac{c}{\mu - \mu_0} \sum_{j=i+1}^{N} \mu_j \\
&= \sum_{j=i+1}^{N} (\mu_j a_j u_{j-1} - \mu_{j+1} a_{j+1} u_j) + \frac{c}{\mu - \mu_0} \sum_{j=i+1}^{N} \mu_j \\
&= \mu_{i+1}a_{i+1}u_i - c + \frac{c}{\mu - \mu_0} \sum_{j=i+1}^{N} \mu_j \\
&= \mu_{i+1}a_{i+1}u_i - \frac{c}{\mu - \mu_0} \sum_{1 \leqslant j \leqslant i} \mu_j, \qquad 0 \leqslant i \leqslant N-1.
\end{aligned}$$

Define additionally $w_0 = 0$. Since $w_1 > 0$, it is clear that $w \in \mathscr{W}$. We have

$$\begin{aligned}
I_i(w)^{-1} &= \mu_{i+1}a_{i+1}(w_{i+1} - w_i) \Big/ \sum_{j=i+1}^{N} \mu_j w_j \\
&= \mu_{i+1}a_{i+1}R_i(u)u_i \Big/ \left[\mu_{i+1}a_{i+1}u_i - \frac{c}{\mu - \mu_0} \sum_{1 \leqslant j \leqslant i} \mu_j\right] \\
&= \lambda_1 \left[1 - \frac{c}{(\mu - \mu_0)\mu_{i+1}a_{i+1}u_i} \sum_{1 \leqslant j \leqslant i} \mu_j\right]^{-1} \\
&\geqslant \lambda_1, \qquad 1 \leqslant i \leqslant N-1.
\end{aligned} \tag{3.11}$$

$$\begin{aligned} I_0(w)^{-1} &= \mu_1 a_1 w_1 \Big/ \sum_{j=1}^{N} \mu_k w_j \\ &= \mu_1 a_1 \left(\lambda_1 u_0 + \frac{c}{\mu - \mu_0}\right) \Big/ a_1 \mu_1 u_0 \\ &= \lambda_1 + \frac{c}{(\mu - \mu_0) u_0} \\ &\geqslant \lambda_1. \end{aligned} \tag{3.12}$$

Collecting these two estimates together, we get

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

Combining this with proof a), we know that

$$\inf_{0 \leqslant i \leqslant N-1} I_i(w)^{-1} = \lambda_1. \tag{3.13}$$

When $N < \infty$, we have $c = 0$ and so $w_i = \lambda_1 g_i$. Hence (3.9) holds by using (3.11) and (3.12) with this w and $c = 0$. We now show that when $N = \infty$, we still have $c = 0$ and so (3.9) also holds. Otherwise, since $\mu_{i+1} a_{i+1} u_i$ is decreasing in i, we have $\inf_{1 \leqslant i \leqslant N-1} I_i(w)^{-1} = I_1(w)^{-1}$. From this, by using (3.11) and (3.12), we must have a contradiction with (3.13) provided $c > 0$.

We have thus completed the proof of (3.9) under the assumption that $g_i \uparrow\uparrow$.

This note is published in the author's book: Eigenvalues, Inequalities, and Ergodic Theory, Springer 2005, §3.8.]

d) We now prove the strictly monotonicity of the eigenfunction (g_i) of λ_1. By (3.10), we have $g_1 \neq 0$. Otherwise, by induction, we would have $g_i \equiv 0$ for all $i \geqslant 1$. Thus, we may assume that $g_1 > 0$. Suppose that there is an n with $1 \leqslant n \leqslant N-1$ such that

$$0 = g_0 < g_1 < \cdots < g_{n-1} < g_n \geqslant g_{n+1}.$$

Define $\overline{g}_i = g_i I_{[i<n]} + g_n I_{[i \geqslant n]}$. Then, we have

$$\sum_i \pi_i \overline{g}_i^2 = \sum_{i \leqslant n-1} \pi_i g_i^2 + g_n^2 \sum_{i=n}^{N} \pi_i,$$

$$-\sum_i \pi_i (\overline{g} \Omega \overline{g})(i) = \lambda_1 \sum_{i \leqslant n-1} \pi_i g_i^2 + \pi_n a_n g_n (g_n - g_{n-1}).$$

Note that

$$\lambda_1 g_n = -\Omega g(n) = b_n (g_n - g_{n+1}) + a_n (g_n - g_{n-1}) \geqslant a_n (g_n - g_{n-1}).$$

We have

$$\pi_n a_n g_n (g_n - g_{n-1}) \leqslant \lambda_1 \pi_n g_n^2 < \lambda_1 g_n^2 \sum_{i=n}^{N} \pi_i.$$

Therefore,

$$\lambda_1 \leqslant \frac{-\sum_i \pi_i(\overline{g}\Omega\overline{g})(i)}{\sum_i \pi_i \overline{g}_i^2} = \frac{\lambda_1 \sum_{i\leqslant n-1} \pi_i g_i^2 + \pi_n a_n g_n(g_n - g_{n-1})}{\sum_{i\leqslant n-1} \pi_i g_i^2 + g_n^2 \sum_{i=n}^N \pi_i} < \lambda_1,$$

which is a contradiction.

e) As for the last assertion of the theorem, simply note that in the above proofs a)–d), we make no use of π_0 (recall that $g_0 = 0$) and b_0. Moreover, the original $I_i(w)$ is homogeneous in (μ_i). [Added in proof: Actually, when $b_0 > 0$,

$$\begin{aligned}
\lambda_1 &= \inf_{f_0=0, f\neq 0} \frac{\sum_{i\geqslant 0} \pi_i b_i (f_{i+1} - f_i)^2}{\sum_{i\geqslant 0} \pi_i f_i^2} \\
&= \inf_{f\neq 0} \frac{\sum_{i\geqslant 1} \pi_i b_i (f_{i+1} - f_i)^2 + \pi_1 a_1 f_1^2}{\sum_{i\geqslant 1} \pi_i f_i^2} \\
&= \inf_{f\neq 0} \frac{\sum_{i\geqslant 1} \mu_i b_i (f_{i+1} - f_i)^2 + \mu_1 a_1 f_1^2}{\sum_{i\geqslant 1} \mu_i f_i^2} \\
&= \inf_{f\neq 0} \frac{\sum_{i\geqslant 1} \tilde{\mu}_i b_i (f_{i+1} - f_i)^2 + \tilde{\mu}_1 a_1 f_1^2}{\sum_{i\geqslant 1} \tilde{\mu}_i f_i^2}.
\end{aligned}$$

Thus, we are studying the process with Dirichlet form

$$\widetilde{D}(f) = \sum_{i\geqslant 1} \tilde{\pi}_i b_i (f_{i+1} - f_i)^2 + \tilde{\pi}_1 a_1 f_1^2 \qquad \left(\tilde{\pi}_i := \tilde{\mu}_i \Big/ \sum_{j\geqslant 1} \tilde{\mu}_j\right)$$

on the state space $\{1, 2, \cdots\}$ and with killing rate a_1. No role is played by b_0.] □

References

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

[2] Chen, M. F. Wang, F. Y., *Estimation of spectral gap for elliptic operators*, Trans. Amer. Math. Soc. 1997, 349, 1209.

[3] Chen, M. F., *Estimation of spectral gap for Markov chains*, Acta Math. Sin. New Ser. 1996, 12:4, 337.

[4] Chen, M. F. Wang, F. Y., *General formula for lower bound of the first eigenvalue on Riemannian manifolds*, 1997, 40:4, 384.

[5] Chen, M. F. Wang, F. Y., *Application of coupling method to the first eigenvalue on manifold*, Sci. Sin.(A) 1994, 37:1, 1.

[6] Chen, M. F., *Estimate of exponential convergence rate in total variation by spectral gap*, Acta Math. Sin. New Ser., 1998, 14:1, 9.

[7] 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.

[8] Diaconis, P. Stroock, D. W., *Geometric bounds for eigenvalues of Markov chains*, Ann. Appl. Prob. 1991, 1:1, 36.

[9] Saloff-Coste, L., *Lectures on finite Markov chains*, in "Lectures on Probability Theory and Statistics", edited by Bernard, P., LNM **1665**, Berlin, Springer-Verlag, 1997, 301.

Acta Math. Sin. Eng. Ser. 15:3 (1999), 353–370

NASH INEQUALITIES FOR GENERAL SYMMETRIC FORMS

Mu-Fa Chen

(Beijing Normal University)
Received March 4, 1999; accepted May 7, 1999.

Abstract. This paper deals with the Nash inequalities and the related ones for general symmetric forms which can be very much unbounded. Some sufficient conditions in terms of the isoperimetric inequalities and some necessary conditions for the inequalities are presented. The resulting conditions can be sharp qualitatively as illustrated by some examples. It turns out that for a probability measure, the Nash inequalities are much stronger than the Poincaré and the logarithmic Sobolev inequalities in the present context.

1. Introduction

Let $(E, \mathscr{E}, \pi)$ be a σ-finite measure space and denote by $L^p(\pi)$ the usual L^p-space of real measurable functions with norm $\|\cdot\|_p$ $(p \in [1, \infty])$. Given a symmetric form $D(f, g)$ with domain $\mathscr{D}(D)$ on $L^2(\pi)$, we are interested in the inequality

$$\|f\|_2^{2+4/\nu} \leqslant \eta_1^{-1}\big[D(f,f) + \delta\|f\|_2^2\big]\,\|f\|_p^{4/\nu}, \qquad f \in L^2(\pi) \tag{1.1}$$

for some constants $\delta \in [0, \infty)$, $p \in [1, 2]$ and ν, $\eta_1 = \eta_1(\delta, p, \nu) \in (0, \infty)$. When π is a probability measure and $D(1, 1) = 0$, the inequality (1.1) with $\delta = 0$ is meaningless for constant f. In and only in this case, we consider an alternative inequality as follows.

$$\mathrm{Var}_\pi(f)^{1+2/\nu} \leqslant \eta_2^{-1} D(f,f)\,\|f\|_p^{4/\nu}, \qquad f \in L^2(\pi) \tag{1.2}$$

for some constants $p \in [1, 2]$, ν, $\eta_2 = \eta_2(p, \nu) \in (0, \infty)$.

The situation where $\nu = \infty$ in (1.1) and (1.2) is excluded since it can be reduced to the case of $p = 2$. When $\pi(E) = 1$, the inequalities usually become stronger for smaller p since $L^1(\pi) \supset L^p(\pi)$. In particular, in the strongest case $p = 1$, they

2000 *Mathematics Subject Classification.* 60J25, 60J75, 47A50.

Key words and phrases. Nash inequality, symmetric form, isoperimetric inequality, jump process.

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

are called the Nash inequalities[6]. In the weakest case $p = 2$, (1.2) is equivalent to the Poincaré inequality:

$$\mathrm{Var}_\pi(f) \leqslant \lambda_1^{-1} D(f,f), \qquad f \in L^2(\pi) \tag{1.3}$$

for some $\lambda_1 > 0$. To see this, replacing f with $f - \pi(f)$ in (1.2), where $\pi(f) = \int f \mathrm{d}\pi$, we get (1.3). Noticing that

$$\mathrm{Var}_\pi(f) = \inf_c \|f - c\|_2^2 \leqslant \|f\|_2^2,$$

(1.3) implies (1.2). Finally, when $p \in [1,2)$, as we will see very soon, (1.2) implies the logarithmic Sobolev inequality

$$\int f^2 \log[f^2/\|f\|_2^2]\,\mathrm{d}\pi \leqslant \sigma^{-1} D(f,f), \qquad f \in L^2(\pi) \tag{1.4}$$

for some $\sigma > 0$. Here and in what follows, the constants η_1, η_2, λ_1 and σ denote the largest one for which the corresponding inequality holds.

We now explain the probabilistic meaning of (1.1) and (1.2). Suppose that $(D, \mathscr{D}(D))$ is deduced from a symmetric Markov semigroup $(P_t)_{t\geqslant 0}$ on $L^2(\pi)$. Then, under some mild assumptions, following the proof of Carlen, Kusuoka and Stroock [2; Theorem 2.1], it can be checked that (1.1) and (1.2) are equivalent respectively to

$$\|P_t\|_{p\to q} \leqslant \left(\frac{\nu}{2\eta_1 t}\right)^{\nu/2} e^{\delta t}, \qquad t > 0 \tag{1.5}$$

and

$$\|P_t - \pi\|_{p\to q} \leqslant \left(\frac{\nu}{2\eta_2 t}\right)^{\nu/2}, \qquad t > 0, \tag{1.6}$$

where $\|\cdot\|_{p\to q}$ denote the operator norm from $L^p(\pi)$ to $L^q(\pi)$, $p^{-1} + q^{-1} = 1$. We remark that one may get different constant η_1 when we go back from (1.5) to (1.1) and similarly from (1.6) to (1.2). Thus, the inequalities (1.1) and (1.2) describe the uniformly algebraic decay of the semigroup $(P_t)_{t\geqslant 0}$.

On the other hand, when $\pi(E) = 1$, it is well known that (1.3) is equivalent to

$$\|P_t f - \pi(f)\|_2 \leqslant \|f - \pi(f)\|_2\, e^{-\lambda_1 t}, \qquad t \geqslant 0, \quad f \in L^2(\pi) \tag{1.7}$$

(cf. [3; Chapter 9]). Besides, by the well known Gross theorem, (1.4) is equivalent to

$$\|P_t\|_{p\to q} \leqslant 1, \quad \text{for all } 1 < p < q < \infty \text{ with } e^{4\sigma t} \geqslant (q-1)/(p-1) \tag{1.8}$$

Note that the proof of (1.6) comes from

$$\|P_t - \pi\|_{1\to\infty} \leqslant \|P_{t/2} - \pi\|_{1\to 2}^2 \leqslant (\nu/(\eta_2 t))^\nu < \infty.$$

Hence, we have

$$\|P_t\|_{1\to 2} \leqslant \|P_t - \pi\|_{1\to 2} + \|\pi\|_{1\to 2} < \infty \quad \text{for all } t > 0.$$

Thus, once (1.2) holds with $p \in [1,2)$, we have not only $\|P_t\|_{p\to 2} < \infty$ but also $\lambda_1 > 0$. Hence by (cf. [1; Theorem 3.6 and Proposition 3.9]), (1.4) holds. Similarly, when π is a probability measure and $D(1,1) = 0$, (1.1) (with $\delta \neq 0$) plus the existence of spectral gap also gives us (1.4). However, the inverse statement is not true in general, i.e., (1.4) or (1.8) is still not strong enough to imply (1.2) for any $p \in [1,2)$ as will be shown in the next section. The reason is that the hypercontractivity (1.8) does not guarantee an algebraic decay of the semigroup, especially there is not enough information for sufficient small t.

The above discussion exhibits a very interesting phenomena. When p increases from 1 to 2, the inequality (1.2) is believed to be weaker and weaker, but each one with $p < 2$ is stronger than (1.4) and at the end point $p = 2$, it becomes weaker than (1.4) suddenly. The intuitive reason for this phenomena is that the function $\log x$ is slower increasing than any x^γ ($\gamma > 0$) as $x \to \infty$.

However, it is much more interesting that when p varies over $[1,2)$, the inequalities given by (1.1) are qualitatively equivalent, and so the ones given by (1.5) for all $p \in [1,2)$, in the sense that the positive constants ν and η_1 are allowed to be different. The proof is rather easy. By Hölder inequality, we have

$$\|f\|_p \leqslant \left[\int f^{(2-p)\cdot\frac{1}{2-p}}\,\mathrm{d}\pi\right]^{2/p-1}\left[\int f^{(2p-2)\cdot\frac{1}{p-1}}\,\mathrm{d}\pi\right]^{1-1/p} = \|f\|_1^{2/p-1}\|f\|_2^{2-2/p}.$$

Thus, if (1.1) holds for some $p \in (1,2)$, then

$$\begin{aligned}\|f\|_2^{2+4/\nu} &\leqslant \eta_1^{-1}\big[D(f,f)+\delta\|f\|_2^2\big]\,\|f\|_p^{4/\nu}\\ &\leqslant \eta_1^{-1}\big[D(f,f)+\delta\|f\|_2^2\big]\,\|f\|_1^{4(2/p-1)/\nu}\|f\|_2^{4(2-2/p)/\nu}.\end{aligned}$$

Dividing both sides by $\|f\|_2^{4(2-2/p)/\nu} < \infty$, we obtain

$$\|f\|_2^{2+4(2/p-1)/\nu} \leqslant \eta_1^{-1}\big[D(f,f)+\delta\|f\|_2^2\big]\,\|f\|_1^{4(2/p-1)/\nu}$$

which is nothing but (1.1) with $p = 1$, and with ν being replaced by $\nu/[2/p-1]$. The same conclusion holds for (1.2) and (1.6). Thus, in what follows, when talking about (1.1) and (1.2), we will always fix $p = 1$.

The symmetric form $(D, \mathscr{D}(D))$ considered in the paper is as follows:

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

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

The typical form in our mind comes form the symmetrizable jump process for which we have a q-pair $(q(x), q(x,\mathrm{d}y))$: $q(x,E) \leqslant q(x) \leqslant \infty$ for all $x \in E$. Throughout the paper, we assume that $q(x) < \infty$ for all $x \in E$. The symmetrizable property simply means that the measure $\pi(\mathrm{d}x)q(x,\mathrm{d}y)$ is symmetric, which gives us automatically a measure J. Then, the killing measure is

given by $K(\mathrm{d}x) = \pi(\mathrm{d}x)[q(x) - q(x, E)]$. For more details, refer to [3]. Next, if $[J(\mathrm{d}x, E) + K(\mathrm{d}x)]/\pi(\mathrm{d}x)$ is bounded (π-a.e.), then for the corresponding form, we have $\mathscr{D}(D) = L^2(\pi)$.

Clearly, the inequalities (1.1) and (1.2) are not easy to check directly, the goal of the paper is to find some more explicit conditions. For sufficient conditions, we use the discrete analog of the isoperimetric inequalities, introduced by Varopoulos[8] (See also Saloff-Coste [7]) for Markov chains with probability kernel (i.e., the operators are bounded above by 1). However, here we handle with the general symmetric forms which can be very much unbounded and have not been studied in the literature as far as we know. To overcome this difficulty, we need some ideas developed in our previous study on the Cheeger's inequalities [5] in which the spectral gap λ_1 was studied in detail. The first idea adopted here is a boundizing procedure. Take and fix a non-negative, symmetric function $r \in \mathscr{E} \times \mathscr{E}$ and a non-negative function $s \in \mathscr{E}$ such that

$$[J^{(1)}(\mathrm{d}x, E) + K^{(1)}(\mathrm{d}x)]/\pi(\mathrm{d}x) \leqslant 1, \qquad \pi\text{-a.e.}, \tag{1.9}$$

where

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

Throughout the paper, we adopt the convention that $r^0 = 1$ and $s^0 = 1$ for $r, s \geqslant 0$. For jump processes, when π is a probability measure, one may simply choose $r(x, y) = q(x) \vee q(y) = \max\{q(x), q(y)\}$ and $s(x) = q(x)$. Correspondingly, we have symmetric forms $\big(D^{(\alpha)}, \mathscr{D}\big(D^{(\alpha)}\big)\big)$ defined by $(J^{(\alpha)}, K^{(\alpha)})$. However, in what follows, we need only three cases $\alpha = 0$, $1/2$ and 1. When $\alpha = 0$, we return to the original form and so the superscript "(α)" is omitted from our notations. We remark that when $\alpha < 1$, $[J^{(\alpha)}(\mathrm{d}x, E) + K^{(\alpha)}(\mathrm{d}x)]/\pi(\mathrm{d}x)$ may no longer be bounded (π-a.e.).

Next, for each $B \in \mathscr{E}$, define

$$\lambda_0^{(\alpha)}(B) = \inf\big\{D^{(\alpha)}(f, f) :\ f|_{B^c} = 0 \text{ and } \pi(f^2) = 1\big\}$$

and set $\lambda_0^{(\alpha)} = \lambda_0^{(\alpha)}(E)$. For the use of the results below, we mention that by [5; (2.4)], we have

$$\lambda_0^{(1)} \geqslant 1 - \sqrt{1 - {h^{(1)}}^2} = {h^{(1)}}^2 \Big/ \Big[1 + \sqrt{1 - {h^{(1)}}^2}\Big],$$

and the proof (b) of [5; Theorem 1.2] gives us

$$\inf_{\pi(B) \leqslant 1/2} \lambda_0^{(1)}(B) \geqslant {k^{(1)}}'^2 \Big/ \Big[1 + \sqrt{1 - {k^{(1)}}'^2}\Big],$$

where $h^{(1)}$ and ${k^{(1)}}'$ are Cheeger's constants:

$$h^{(\alpha)} = \inf_{\pi(A) > 0} \frac{J^{(\alpha)}(A \times A^c) + K^{(\alpha)}(A)}{\pi(A)},$$

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

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

Theorem 1.1. Given constants $\delta \in [0,\infty)$ and $\nu \in [1,\infty)$. Define

$$S_{\nu,\delta} = \inf_{\pi(A)\in(0,\infty)} \frac{J^{(1/2)}(A\times A^c) + K^{(1/2)}(A) + \delta\pi(A)}{\pi(A)^{(\nu-1)/\nu}}. \tag{1.10}$$

Then

$$\|f\|_2^{2+4/\nu} \leqslant \big(2-\lambda_0^{(1)}\big) S_{\nu,\delta}^{-2} D(f,f)\,\|f\|_1^{4/\nu}, \qquad \text{if } \delta = 0 \tag{1.11}$$

$$\|f\|_2^{2+4/\nu} \leqslant 2\big[\big(2-\lambda_0^{(1)}\big) S_{\nu,\delta}^{-2} D(f,f) + \delta^2\|f\|_2^2\big]\,\|f\|_1^{4/\nu}, \quad \text{if } \delta \neq 0,\ f\in L^2(\pi). \tag{1.12}$$

Theorem 1.2. Let π be a probability measure and $K(\mathrm{d}x) = 0$. Define the isoperimetric constant I_ν as follows:

$$I_\nu = \inf_{0<\pi(A)\leqslant 1/2} \frac{J^{(1/2)}(A\times A^c)}{\pi(A)^{(\nu-1)/\nu}} = \inf_{0<\pi(A)<1} \frac{J^{(1/2)}(A\times A^c)}{\big[\pi(A)\wedge\pi(A^c)\big]^{(\nu-1)/\nu}}.$$

Then

$$\mathsf{Var}_\pi(f)^{1+2/\nu} \leqslant \min\Big\{2,\, 2^{2/\nu}\Big(2 - \inf_{\pi(B)\leqslant 1/2} \lambda_0^{(1)}(B)\Big)\Big\} I_\nu^{-2} D(f,f)\,\|f\|_1^{4/\nu}, \quad f\in L^2(\pi). \tag{1.13}$$

When $\nu = \infty$, $S_{\nu,\delta} = (1-\delta)h^{(1/2)}$ $(\delta < 1)$ and $I_\nu = k^{(1/2)'}$ (recall the notations $h^{(\alpha)}$ and $k^{(\alpha)'}$ given in two lines above Theorem 1.1). This is just the case studied in [5]. Next, define

$$S_\nu(r) = \inf_{\pi(A)\in(0,r]} \frac{J^{(1/2)}(A\times A^c) + K^{(1/2)}(A)}{\pi(A)^{(\nu-1)/\nu}}, \qquad r\in(0,\infty).$$

Then, a sufficient condition for $\lim_{\delta\to\infty} S_{\nu,\delta} > 0$ is that $S_\nu(0) := \lim_{r\to 0} S_\nu(r) > 0$. This is easy to check:

$$S_{\nu,\delta} = \inf_{\pi(A)\in(0,r]}[\cdots] \bigwedge \inf_{\pi(A)\in(r,\infty)}[\cdots] \geqslant S_\nu(r)\wedge[\delta r^{1/\nu}] > 0.$$

Conversely, if there is a sequence $\{A_n\}\subset\mathscr{E}$ such that $\pi(A_n)\to 0$, then the inverse implication also holds since

$$0 < S_{\nu,\delta} \leqslant \varliminf_{r\to 0}[S_\nu(r) + \delta r^{1/\nu}] = S_\nu(0).$$

The above two theorems are an improvement even in the case of finite Markov chains, on the results given in [7].

To illustrate the application of Theorem 1.2, consider the regular birth-death process on $\mathbb{Z}_+$ with birth rates (b_i) and death rates (a_i). Then $K(\mathrm{d}x) = 0$, $J_{ij} = \pi_i b_i$ if $j = i+1$, $J_{ij} = \pi_i a_i$ if $j = i-1$ and $J_{ij} = 0$ otherwise.

Corollary 1.3. For birth-death process with $\pi(E) = 1$, take

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

Then

(1) $I_\nu > 0$ for some $\nu \geqslant 1$ iff there exists a constant $c > 0$ such that

$$\frac{\pi_i a_i}{\sqrt{r_{i,i-1}}} \geqslant c \Bigg[\sum_{j \geqslant i} \pi_j \Bigg]^{(\nu-1)/\nu}, \qquad i \geqslant 1. \tag{1.14}$$

If so, we indeed have $I_\nu \geqslant c$.

(2) $S_{\nu,\delta} > 0$ (with $\delta > 0$) for some $\nu > 1$ iff (1.14) holds.

The inequalities (1.11)–(1.13) provides us some lower bounds of $\eta_k = \eta_k(\delta, \nu)$, $k = 1, 2$. For instance, from (1.11), it follows that $\eta_1(0, \nu) \geqslant \big(2 - \lambda_0^{(1)}\big)^{-1} S_{\nu,0}^2$. As usual, some rough upper bounds are easier to obtain. To see this, define $\overline{S}_{\nu,\delta}$ and $\overline{I}_\nu$ in the same way as $S_{\nu,\delta}$ and I_ν, but except for replacing $J^{(1/2)}$ and $K^{(1/2)}$ with J and K respectively. By setting $f = I_A$ with $\pi(A) \in (0, \infty)$ in (1.1) and $f = I_A - \pi(A)$ with $\pi(A) \in (0, 1)$ in (1.2), one deduces that $\eta_1 \leqslant \overline{S}_{\nu/2,\delta}$ and $\eta_2 \leqslant 4^{1+1/\nu} \overline{I}_{\nu/2}$.

We now introduce some more precise necessary conditions for (1.1) and (1.2). To state the result, we need the following condition for a test function φ (which is often chosen to be an elementary function):

$$\varphi \geqslant 0, \quad \pi(e^\varphi) = \infty, \quad \lim_{n \to \infty} \pi(\varphi > n) = 0, \quad \pi(\varphi < c) < \infty \, (\forall c > 0). \tag{1.15}$$

The next result is a modification of [5; Theorem 1.5].

Theorem 1.4. Let $\|K\|_{2 \to 2} < \infty$. Then the inequalities (1.1) and (1.2) do not hold (i.e., η_1 and $\eta_2 = 0$) if one of the following conditions holds.

(1) (1.9) holds and $r > 0$. There exists φ satisfying (1.15) and

$$\operatorname{ess\,sup}_J |\varphi(x) - \varphi(y)|^2 r(x, y) < \infty.$$

(2) $J(\mathrm{d}x, \mathrm{d}y) = \pi(\mathrm{d}x) q(x, \mathrm{d}y)$. There exists φ satisfying (1.15) and

$$\operatorname{ess\,sup}_\pi \int |\varphi(x) - \varphi(y)|^2 q(x, \mathrm{d}y) < \infty.$$

(3) π is a probability measure, the support of π contains infinite disjoint sets and $J(\mathrm{d}x, E)/\pi(\mathrm{d}x)$ is π-a.e. bounded.

Part (3) of the theorem tells us that in order to study (1.2) for infinite E, it is necessary to consider the unbounded operators. Roughly speaking, this theorem requires

$$D(f_n, f_n) \sim \|f_n\|_2^2$$

and allows $\|f_n\|_1 \to \infty$. The next result allows

$$D(f_n, f_n) \sim \|f_n\|_2^{2+4/\nu}$$

but requires $\|f_n\|_1$ to be bounded.

Theorem 1.5. Given φ and ψ with

$$\varphi,\, \psi > 0, \quad \|\varphi\|_1 < \infty, \quad \|\varphi\|_2 = \infty \quad \text{and}$$
$$C_1 := \operatorname{ess\,sup}_J I_{\{\varphi(y)>\varphi(x)\}}[\varphi(y)-\varphi(x)]/\psi(y) < \infty. \tag{1.16}$$

Let $f_n = \varphi \wedge N_n$, where $N_n \to \infty$ as $n \to \infty$. If

$$C_1 \int_{\{f_n(y)<f_n(x)\}} J(\mathrm{d}x, \mathrm{d}y)\psi(x)^2 + \int K(\mathrm{d}x) f_n(x)^2 \leqslant C_2(n)\|f_n\|_2^{2+4/v}, \tag{1.17}$$

then $\eta_1,\, \eta_2 \leqslant \|\varphi\|_1^{4/\nu} \varliminf_{n\to\infty} C_2(n)$.

The simplest choice of ψ used in (1.16) is nothing but φ and then $C_1 \leqslant 1$. It is usually taken to be the derivative of φ. As a consequence of Theorem 1.4 and Theorem 1.5, we have the following result.

Corollary 1.6. For birth-death process with $\pi(E) = 1$, we have η_1 (with $\delta > 0$), $\eta_2 = 0$ if one of the following conditions holds.

(1) There exists ψ such that $\psi_i \geqslant \pi_i\, (i \gg 1)$,

$$\sum_i \psi_i = \infty \quad \text{and} \quad \sup_{i\geqslant 1}\left[a_i\left(\log\frac{\psi_i a_i}{\psi_{i-1}b_{i-1}}\right)^2 + b_i\left(\log\frac{\psi_{i+1}b_i}{\psi_i a_{i+1}}\right)^2\right] < \infty. \tag{1.18}$$

(2) $\overline{\lim}_{i\to\infty} b_i/a_{i+1} =: \rho^{-1} \in [0,1)$, $a_i \uparrow$ as $i \uparrow$. There exists ψ such that $\psi_i \geqslant \pi_i\, (i \gg 1)$, $\sum_i \psi_i = \infty$, $\lim_{i\to\infty} \psi_{i+1}/\psi_i =: \gamma \in [1, \rho)$ and

$$\varliminf_{n\to\infty} a_n \Big/ \left(\sum_{i\leqslant n} \psi_i\right)^{2/\nu} = 0.$$

The proofs of the above results are delayed to Section 3. In the next section, we introduce some more concrete corollaries and illustrate the power of the results by some examples. In particular, we show that all of them can be sharp qualitatively. However, we should mention that the method of isoperimetric inequalities does have certain limitation as illustrated by [5; Example 4.8].

2. Corollaries and Examples

In this section, we first introduce some criteria for (1.1), (1.2) and (1.4) for some more specific but typical birth-death processes. Their proofs are delayed again to the next section. However, one may first ignore the corollaries and jump to look at the examples given in the second part of the section.

Note that in the qualitative study, we need only the asymptotic behavior of the quantity and so one may ignore a finite number of terms or a positive factor. In particular, we write $A(i) \sim B(i)$ if either $\lim_{i\to\infty} B(i) \in [0,\infty)$ and $\lim_{i\to\infty} A(i) = c\lim_{i\to\infty} B(i)$ or $\lim_{i\to\infty} B(i) = \infty$ but still $\lim_{i\to\infty} A(i)/B(i) = c$ for some constant $c \in (0,\infty)$, and write $A(i) \gtrsim B(i)$ if $A(i) \geqslant cB(i)$ for all large enough i and a constant $c \in (0,\infty)$.

Unless otherwise stated, the measure π considered in this section is a probability. The first two corollaries deal with the case of $a_i = b_i$ which is related to the polynomial decay of (π_i). Corollaries 2.2 and 2.4 are devoted to the logarithmic Sobolev inequality, they may be regarded as an addition to [9].

Corollary 2.1. Let $a(x) \in C^1([1,\infty))$ be strictly increasing and satisfy

$$\int_1^\infty \frac{\mathrm{d}x}{a(x)} < \infty.$$

Take $b_i = a_i = a(i)\,(i \geqslant 1)$.

(1) If $a(x) \gtrsim x^\gamma$ for some $\gamma > 2$, then $I_\nu,\ S_{\nu,\delta}\,(\delta > 0) > 0$ for $\nu \geqslant 2(\gamma-1)/(\gamma-2)$, and hence (1.1) with $\delta > 0$ and (1.2) hold.

(2) Suppose additionally that

$$\sup_{x\geqslant 1, \varepsilon\in(0,1)} a(x)/a(x-\varepsilon) < \infty \quad \text{and} \quad \overline{\lim_{x\to\infty}}\, x[\log a(x)]' < \infty.$$

If $a(x) \lesssim x^2$, then (1.1) and (1.2) do not hold for all $\delta \geqslant 0$ and $\nu > 0$.

(3) Suppose additionally that the limit $\xi := \lim_{x\to\infty} x(\log a(x))'$ exists. Fix $\delta,\ \nu > 0$. Then (1.1) and (1.2) do not hold if either $\xi \leqslant 1$ or $\xi \in (1,\infty]$ but still

$$\sup_{x\geqslant 1,\ \varepsilon\in(0,1)} [\log a(x-\varepsilon)]'/[\log a(x)]' < \infty, \qquad \sum_{i\geqslant 1} a_i^{-1/2} i^{s/2} < \infty$$

and

$$\lim_{x\to\infty} \big[\sqrt{a(x)}\,\big]'/x^{(s+1)/\nu} = 0$$

for some $s = s(\nu) > -1$.

Corollary 2.2. Suppose that $a(x)$, a_i and b_i be the same as in Corollary 2.1 and that the limit $\xi := \lim_{x\to\infty} x(\log a(x))'$ exists.

(1) Let $a(x) \in C^2([1,\infty))$. Then (1.4) does not hold if either $\xi \leqslant 1$ or $\xi > 1$ but still

$$\sup_{x\geqslant 1,\ \varepsilon\in(0,1)} [\log a(x-\varepsilon)]'/[\log a(x)]' < \infty \ \text{ and } \ \lim_{x\to\infty} a''(x)/\log a(x) = 0.$$

(2) Let $a(x) \in C^3([1,\infty))$. Then (1.4) holds provided $\xi > 1$,

$$\lim_{x\to\infty} \big(\sqrt{a(x)}\,\big)' = \infty, \qquad \lim_{x\to\infty} a(x)\big[\big(\sqrt{a(x)}\,\big)'^2\big]' = \infty$$

and

$$\overline{\lim_{x\to\infty}}\, a(x)\big\{1/\big[\big(\sqrt{a(x)}\,\big)'^2\big]'\big\}' < \infty.$$

The next two corollaries deal with the case of (π_i) being exponential decay.

Corollary 2.3. Let $\overline{\lim}_{i\to\infty} b_i/a_{i+1} =: \rho^{-1} \in [0,1)$. Then

(1) $I_\nu,\ S_{\nu,\delta}(\delta > 0) > 0\,(\nu \geqslant 2)$ if $b_i \leqslant a_i\,(i \gg 1)$ and $a_i \gtrsim \pi_i^{-2/\nu}$.

(2) (1.1) and (1.2) do not hold if there exists $\gamma \in [1,\rho)$ such that

$$\underline{\lim_{n\to\infty}}\, a_n/\gamma^{2n/\nu} = 0.$$

Corollary 2.4. Let $\overline{\lim}_{i\to\infty} b_i/a_{i+1} =: \rho^{-1} \in [0,1)$. Then

(1) (1.4) does not hold if $\lim_{n\to\infty} a_n/\log \pi_n^{-1} = 0$.
(2) (1.4) holds if $b_i \leqslant a_i$ $(i \gg 1)$, $a_i \uparrow$ and $\underline{\lim}_{n\to\infty} a_n/\log \pi_n^{-1} > 0$.

It is now appropriate to present some examples. The first example below is standard, for which $\pi(E) = \infty$. The example also shows that the inequality (1.1) is not so restrictive in the case of $\pi(E) = \infty$. Refer to [2] or [8] for more information.

Example 2.5. For the simple random walk $P = (p_{ij})$ on $\mathbb{Z}^d$, $J_{ij} = \pi_i p_{ij}$ $(i \neq j)$, the Sobolev constant $S_{\nu,\delta} \geqslant S_{\nu,0} > 0$ for all $\nu \geqslant 1$.

Proof. Simply use that fact that $J_{ij} =$ positive constant for all $i \neq j$ with $|i-j| = 1$ and apply Theorem 1.1. □

The next example shows that the inequalities (1.1), (1.2) and (1.4) are stronger than the Poincaré inequality (1.3) only at the critical point.

Example 2.6. Take $a_i = b_i = i^\gamma$ $(i \geqslant 1, \gamma > 1)$. Then (1.3) holds iff $\gamma \geqslant 2$.

(1) (1.1) with $\delta > 0$ and (1.2) hold for some $\nu > 0$ iff $\gamma > 2$. Then, we must have $\nu \geqslant 2(\gamma-1)/(\gamma-2)$.
(2) (1.4) holds iff $\gamma > 2$.

Proof. Refer to [4] or [5] for a proof about (1.3). The first assertion and the sufficiency of the second one in part (1) follow from the first two parts of Corollary 2.1. Moreover, Part (3) of the corollary removes the region $\nu < 2(\gamma-1)/(\gamma-2)$ and then proves the necessity of the second assertion in (1). Part (2) follows from Corollary 2.2. □

Before moving further, we mention that the inequality (1.3) holds for all the examples given below.

Example 2.7. Take $a_i = b_i = i^2 \log^\gamma(i+1)$ $(i \geqslant 1, \gamma \in \mathbb{R})$. Then

(1) (1.1) and (1.2) do not hold for all γ.
(2) (1.4) holds iff $\gamma \geqslant 1$.

Proof. Part (1) follows from Part (3) of Corollary 2.1. Part (2) is due to [9] and follows from Corollary 2.2. □

Example 2.8. Take $\pi_i \sim \rho^{-i}$ for some $\rho > 1$, $a_i = \rho^{\beta i}$ $(\beta \in (0,1))$. Then

(1) (1.1) with $\delta > 0$ and (1.2) hold if $\beta \geqslant 2/\nu$ with $\nu > 2$. Conversely, (1.1) and (1.2) do not hold if $\beta < 2/\nu$.
(2) (1.4) holds for all $\beta \in (0,1)$.

Proof. Part (1) follows from Corollary 2.3 and Part (2) follows from Corollary 2.4. □

The next result is very surprising. It indicates a big jump from (1.2) to (1.3) and shows that the Nash inequalities are much stronger than the Poincaré and the logarithmic Sobolev ones.

Example 2.9. Take $b_i = bi^\beta$ $(i \geqslant 1)$, $b_0 = 1$, $a_i = i^\gamma$, $\gamma \geqslant \beta \geqslant 0$. Assume also that $b < 1$ when $\beta = \gamma$. Then

(1) (1.1) with $\delta > 0$ (resp., (1.2)) does not hold.
(2) (1.4) holds iff either $\beta = \gamma \geqslant 1$ or $\beta < \gamma > 1$.

Proof. From [4], it follows that $\lambda_1 > 0$. Next, Part (1) follows from Part (2) of Corollary 2.3 and Part (2) follows from Corollary 2.4. □

Of course, the stronger convergence is less common. However, it has its own use. For instance, in the study of Markov Chain Monte Carlo, one looks for rapidly convergent symmetric form for a given distribution. For this, the stronger convergence may be more helpful. Roughly speaking, as we have seen from the above corollaries and examples, in order for I_ν (or $S_{\nu,\delta}$ $(\delta > 0)) > 0$, when (π_i) is polynomial (resp., exponential) decay, (a_i) should be polynomial (resp., exponential) growth.

Note that Theorem 1.2 is deduced from Theorem 1.1, we have seen that all the results in the paper can be sharp qualitatively. We now want to know how about the constants given in (1.11) and (1.13).

Example 2.10. Consider the Markov chain with state $\{0, 1\}$. Let $q_{01} = q_{10} = 1$. Then the coefficient in (1.11) is exact and the one in (1.13) has only an extra factor $2^{1+4/\nu}$ for every $\nu \geqslant 1$.

Proof. Note that $\pi_0 = \pi_1 = 1/2$, $J_{01} = J_{01}^{(1)} = \pi_0 q_{01} = 1/2$.

(a) Take $B = \{1\}$. Let $f_0 = 0$ and $f_1 = 2$. Then $\|f\|_1 = 1$, $\|f\|_2^2 = 2$ and

$$D(f, f) = \pi_0 q_{01}(f_1 - f_0)^2 = 2.$$

Thus,

$$\lambda_0(B) = D(f, f)/\|f\|_2^2 = 1$$

and $D(f, f)/\|f\|_2^{2+4/\nu} = 2^{-2/\nu}$. On the other hand,

$$S_\nu(B) := J_{01}/\pi_1^{(\nu-1)/\nu} = 2^{-1/\nu}.$$

Therefore,

$$(2 - \lambda_0(B))S_\nu(B)^{-2} = 2^{2/\nu}.$$

This means that constant given in (1.11) is exact.

(b) Note that

$$I_\nu = J_{01}/[\pi_0 \wedge \pi_1]^{(\nu-1)/\nu} = 2^{-1/\nu} = S_\nu(B).$$

Next, by symmetry, $\lambda_0(\{0\}) = 1$ and so

$$2^{2/\nu}\Big(2 - \inf_{\pi(B)\leqslant 1/2} \lambda_0(B)\Big) I_\nu^{-2} = 2^{4/\nu}.$$

However, $\eta_2 = 2/1^{1+2/\nu} = 2$ for all $\nu > 0$, the infimum is achieved by $f_0 = -f_1 = 1$. This means that there is an extra factor $2^{1+4/\nu}$ in (1.13). The factor 2 comes from the use of the inequality in (3.8) and the factor $2^{4/\nu}$ comes from the second inequality of (3.7). □

3. Proofs

To prove Theorem 1.1, we need some preparation. Let $f \in L^1_+(\pi)$. Set $F_t = \{f \geqslant t\}$ and $f_t = I_{F_t}$. Then we have

$$f(x) = \int_0^{\|f\|_u} f_t(x)\mathrm{d}t \quad \text{and} \quad \pi(f) = \int_0^{\|f\|_u} \pi(F_t)\mathrm{d}t,$$

where $\|f\|_u = \sup|f| \leqslant \infty$.

Lemma 3.1 (Co-Area Formula).

$$\int J^{(\alpha)}(\mathrm{d}x, \mathrm{d}y)|f(y) - f(x)| = 2\int_0^{\|f\|_u} J^{(\alpha)}(F_t \times F_t^c)\mathrm{d}t.$$

Proof. The proof is standard. Refer to [7; Chapter 3] for instance. □

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 $f^* = fI_E$. Next, define $J^{*(\alpha)}$ on $E^* \times E^*$ by

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

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

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

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

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

We remark that in proving Theorem 1.1 and Theorem 1.2, it suffices to consider a bounded $f \in \mathscr{D}(D) \cap L^1(\pi)$ only. Actually, for $f \in \mathscr{D}(D) \cap L^1(\pi)$, define $f_n = (-n) \vee f \wedge n$. Since $|f_n(y) - f_n(x)| \leqslant |f(y) - f(x)|$ and $|f_n| \leqslant |f|$, we have

$$D(f_n - f, f_n - f) \leqslant 4D(f,f), \qquad D(f_n - f, f_n - f) \to 0$$

and $\|f_n - f\|_p \to 0$ as $n \to \infty$ for all $p \in [1,2]$. Hence, f_n is bounded and belongs to $\mathscr{D}(D) \cap L^1(\pi)$.

Proof of Theorem 1.1. To prove (1.11), it suffices to consider a bounded $g \in \mathscr{D}(D) \cap L^1_+(\pi)$ since $D(|g|,|g|) \leqslant D(g,g)$. Set $C = S^{-1}_{\nu,\delta}$ and $q = \nu/(\nu-1) \in (1,\infty]$ (do not confuse with the conjugate exponent of p used elsewhere),

$$G_t = \{x \in E : g(x) \geqslant t\}, \qquad G^*_t = \{x \in E^* : g^*(x) \geqslant t\}$$

and $g_t = I_{G_t}$. Then we have $G^*_t = G_t$ and $\pi(G_t) \leqslant t^{-1}\pi(g) < \infty$ whenever $t > 0$. Recall that in the present situation, $\|g\|_\infty = \operatorname{ess\,sup}_\pi |g| = \|g\|_u$. Fix $q < \infty$ for a moment. Then, we have

$$\begin{aligned}
\|g\|_q &\leqslant \int_0^{\|g\|_u} \|g_t\|_q \mathrm{d}t \quad \text{(by Hölder-Minkowski inequality)} \\
&= \int_0^{\|g\|_u} \pi(G_t)^{1/q} \mathrm{d}t \\
&\leqslant C \int_0^{\|g\|_u} \big[J^{(1/2)}\big(G_t \times E \setminus G_t\big) + K^{(1/2)}(G_t) + \delta\pi(G_t)\big]\mathrm{d}t \\
&\qquad \text{(by assumption)} \\
&= C \int_0^{\|g\|_u} \big[J^{*(1/2)}\big(G^*_t \times E^* \setminus G^*_t\big) + \delta\pi(G_t)\big]\mathrm{d}t \\
&= C\left[\frac{1}{2}\int_{E^*\times E^*} J^{*(1/2)}(\mathrm{d}x,\mathrm{d}y)|g^*(y) - g^*(x)| + \delta\|g\|_1\right]. \qquad (3.1)
\end{aligned}$$

Here in the last step, we have used the co-area formula for the symmetric form J^*. It is easy to check that the proof remains true even if $q = \infty$. By taking $g = I_A$ with $\pi(A) < \infty$, (3.1) is reduced to (1.10) and hence we have proved that (1.10) and (3.1) are actually equivalent.

Next, by Cauchy-Schwarz inequality,

$$\begin{aligned}
&\int_{E^*\times E^*} J^{*(1/2)}(\mathrm{d}x,\mathrm{d}y)|g^*(y)^2 - g^*(x)^2| \\
&\qquad = \int_{E^*\times E^*} J^{*(1/2)}(\mathrm{d}x,\mathrm{d}y)|g^*(y) - g^*(x)|\,|g^*(y) + g^*(x)| \\
&\qquad \leqslant \sqrt{2D(g,g)}\left[\int_{E^*\times E^*} J^{*(1)}(\mathrm{d}x,\mathrm{d}y)[g^*(y) + g^*(x)]^2\right]^{1/2} \\
&\qquad = \sqrt{2D(g,g)}\Bigg[\int_{E^*\times E^*} J^{*(1)}(\mathrm{d}x,\mathrm{d}y)[2g^*(y)^2 + 2g^*(x)^2] \\
&\qquad\qquad - \int_{E^*\times E^*} J^{*(1)}(\mathrm{d}x,\mathrm{d}y)[g^*(y) - g^*(x)]^2\Bigg]^{1/2} \\
&\qquad \leqslant 2\sqrt{D(g,g)}\,\left[2\|g\|_2^2 - D^{(1)}(g,g)\right]^{1/2} \quad \text{(by (1.9))} \\
&\qquad \leqslant 2\sqrt{\big(2 - \lambda^{(1)}_0\big)D(g,g)}\,\|g\|_2. \qquad (3.2)
\end{aligned}$$

Applying (3.1) to g^2 and then using (3.2) we get

$$\begin{aligned}\|g\|_{2q}^2 &\leqslant C\bigg[\frac{1}{2}\int_{E^*\times E^*} J^{*(1/2)}(\mathrm{d}x,\mathrm{d}y)|g^*(y)^2-g^*(x)^2|+\delta\|g\|_2^2\bigg]\\ &\leqslant C\Big[\sqrt{\big(2-\lambda_0^{(1)}\big)D(g,g)}\ \|g\|_2+\delta\|g\|_2^2\Big]. \end{aligned}\tag{3.3}$$

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

$$\|g\|_2 \leqslant \|g\|_1^{1/(\nu+1)}\|g\|_{2q}^{\nu/(\nu+1)}.\tag{3.4}$$

Combining (3.4) with (3.3), we get

$$\|g\|_2 \leqslant \bigg\{C\Big[\sqrt{\big(2-\lambda_0^{(1)}\big)D(g,g)}\ \|g\|_2+\delta\|g\|_2^2\Big]\bigg\}^{\nu/2(\nu+1)}\|g\|_1^{1/(\nu+1)}.\tag{3.5}$$

From this, (1.11) and (1.12) follow immediately. □

We now turn to prove Theorem 1.2.

Proof of Theorem 1.2. (a) By assumption, $K(\mathrm{d}x) = 0$. Thus, for every f with $f|_{B^c} = 0$, we have

$$\begin{aligned}D^{(\alpha)}(f,f) &= \frac{1}{2}\int_{B\times B} J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y)[f(y)-f(x)]^2+\int_B J^{(\alpha)}(\mathrm{d}x,B^c)f(x)^2\\ &=: D_B^{(\alpha)}(f,f).\end{aligned}$$

Then,

$$\lambda_0^{(\alpha)}(B) = \big\{D_B^{(\alpha)}(fI_B,fI_B) : \pi\big(f^2I_B\big)=1\big\}.$$

Define

$$\begin{aligned}S_\nu(B) &= \inf_{A\subset B,\pi(A)>0}\frac{J^{(1/2)}(A\times(B\setminus A))+J^{(1/2)}(A\times B^c)}{\pi(A)^{(\nu-1)/\nu}}\\ &= \inf_{A\subset B,\pi(A)>0}\frac{J^{(1/2)}(A\times A^c)}{\pi(A)^{(\nu-1)/\nu}}.\end{aligned}$$

Then, applying Theorem 1.1 to the form D_B with $\delta = 0$ and using (1.11) with $S_{\nu,\delta} = S_\nu(B)$, we obtain

$$\|fI_B\|_2^{2+4/\nu} \leqslant \big(2-\lambda_0^{(1)}(B)\big)S_\nu(B)^{-2}D_B(fI_B,fI_B)\,\|fI_B\|_1^{4/\nu}.\tag{3.6}$$

(b) Fix a bounded $g\in\mathscr{D}(D)\subset L^2(\pi)$ with median c. Define $g_\pm = (g-c)^\pm$ and $B_\pm = \{g_\pm>0\}$. Then $\pi(B_\pm)\leqslant 1/2$. By (3.6), we get

$$\begin{aligned}\|g_\pm\|_2^{2+4/\nu} &\leqslant \big(2-\lambda_0^{(1)}(B_\pm)\big)S_\nu(B_\pm)^{-2}D_{B_\pm}(g_\pm,g_\pm)\|g_\pm\|_1^{4/\nu}\\ &\leqslant \big(2-\lambda_0^{(1)}(B_\pm)\big)S_\nu(B_\pm)^{-2}D_{B_\pm}(g_\pm,g_\pm)\|g-c\|_1^{4/\nu}.\end{aligned}\tag{3.7}$$

On the other hand,

$$\begin{aligned} D(g,g) &= D(g-c,g-c) \\ &= \frac{1}{2}\int J(\mathrm{d}x,\mathrm{d}y)\big[|g_+(y)-g_+(x)|+|g_-(y)-g_-(x)|\big]^2 \\ &\geqslant \frac{1}{2}\int J(\mathrm{d}x,\mathrm{d}y)\big[g_+(y)-g_+(x)\big]^2+\frac{1}{2}\int J(\mathrm{d}x,\mathrm{d}y)\big[g_-(y)-g_-(x)\big]^2 \\ &= D(g_+,g_+)+D(g_-,g_-) \\ &= D_{B_+}(g_+,g_+)+D_{B_-}(g_-,g_-); \end{aligned} \tag{3.8}$$

$$\begin{aligned} S_\nu(B_+)\wedge S_\nu(B_-) &\geqslant \inf_{\pi(B)\leqslant 1/2} S_\nu(B) \\ &= \inf_{\pi(B)\leqslant 1/2}\ \inf_{A\subset B,\pi(A)>0} \frac{J^{(1/2)}(A\times A^c)}{\pi(A)^{(\nu-1)/\nu}} \\ &= \inf_{0<\pi(A)\leqslant 1/2} \frac{J^{(1/2)}(A\times A^c)}{\pi(A)^{(\nu-1)/\nu}} = I_\nu; \end{aligned} \tag{3.9}$$

$$\|g-c\|_2^{2+4/\nu} = \big(\|g_+\|_2^2+\|g_-\|_2^2\big)^{1+2/\nu} \leqslant 2^{2/\nu}\big(\|g_+\|_2^{2+4/\nu}+\|g_-\|_2^{2+4/\nu}\big). \tag{3.10}$$

Combining (3.7)–(3.10) together, we get

$$\begin{aligned} &2^{-2/\nu}\|g-c\|_2^{2+4/\nu} \\ &\leqslant \big[\big(2-\lambda_0^{(1)}(B_+)\big)S_\nu(B_+)^{-2}D_{B_+}(g_+,g_+) \\ &\qquad +\big(2-\lambda_0^{(1)}(B_-)\big)S_\nu(B_-)^{-2}D_{B_-}(g_-,g_-)\big]\,\|g-c\|_1^{4/\nu} \\ &\leqslant \Big(2-\inf_{\pi(B)\leqslant 1/2}\lambda_0^{(1)}(B)\Big)I_\nu^{-2}\Big[D_{B_+}(g_+,g_+)+D_{B_-}(g_-,g_-)\Big]\,\|g-c\|_1^{4/\nu} \\ &\leqslant \Big(2-\inf_{\pi(B)\leqslant 1/2}\lambda_0^{(1)}(B)\Big)I_\nu^{-2}D(g,g)\,\|g-c\|_1^{4/\nu}. \end{aligned}$$

We obtain

$$\|g-c\|_2^{2+4/\nu} \leqslant 2^{2/\nu}\Big(2-\inf_{\pi(B)\leqslant 1/2}\lambda_0^{(1)}(B)\Big)I_\nu^{-2}D(g,g)\,\|g-c\|_1^{4/\nu}.$$

(c) Finally, since $\mathrm{Var}_\pi(g)=\inf_\alpha\|g-\alpha\|_2^2$ and c is a median of g, we obtain

$$\mathrm{Var}_\pi(g)^{1+2/\nu} \leqslant 2^{2/\nu}\Big(2-\inf_{\pi(B)\leqslant 1/2}\lambda_0^{(1)}(B)\Big)I_\nu^{-2}D(g,g)\,\|g\|_1^{4/\nu} \tag{3.11}$$

as required. □

We now present an alternative proof of Theorem 1.2 which will gives us the same inequality (3.11) but with a different constant. To do so, we need the following result.

Lemma 3.2. The following variational formula holds.

$$I_\nu = \inf\left\{\frac{\frac{1}{2}\int J^{(1/2)}(\mathrm{d}x,\mathrm{d}y)|f(y)-f(x)|}{\inf_{c:c \text{ is a median of } f}\|f-c\|_{\nu/(\nu-1)}} : f\in L^1(\pi) \text{ is non-constant}\right\}.$$

Proof. Similar to the proof given in [7; Chapter 3]. □

Alternative Proof of Theorem 1.2. 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 the definition of f and Lemma 3.2, we obtain

$$\|g-c\|_{2q}^2=\|f\|_q\leqslant\frac{1}{2}I_\nu^{-1}\int J^{(1/2)}(\mathrm{d}x,\mathrm{d}y)|f(y)-f(x)|. \tag{3.12}$$

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 equality and following the proof of (3.2), we get

$$\begin{aligned}&\int J^{(1/2)}(\mathrm{d}x,\mathrm{d}y)|f(y)-f(x)|\\ &\quad\leqslant\sqrt{2D(g,g)}\left[\int J^{(1)}(\mathrm{d}x,\mathrm{d}y)[|g(y)-c|+|g(x)-c|]^2\right]^{1/2}\\ &\quad\leqslant 2\sqrt{2D(g,g)}\,\|g-c\|_2.\end{aligned} \tag{3.13}$$

Combining (3.12) with (3.13) together, we get

$$\|g-c\|_{2q}^2\leqslant 2I_\nu^{-1}\sqrt{2D(g,g)}\,\|g-c\|_2.$$

Now, by using Hölder inequality (3.4), it follows that

$$\|g-c\|_2\leqslant\left[I_\nu^{-1}\sqrt{2D(g,g)}\,\|g-c\|_2\right]^{\nu/2(\nu+1)}\|g-c\|_1^{1/(\nu+1)}.$$

Thus,

$$\|g-c\|_2^{2(1+2/\nu)}\leqslant 2I_\nu^{-2}D(g,g)\,\|g-c\|_1^{4/\nu}$$

and hence

$$\mathrm{Var}_\pi(g)^{1+2/\nu}\leqslant 2I_\nu^{-2}D(g,g)\,\|g\|_1^{4/\nu}.\quad\square$$

The above two proofs show that one may replace $\|f\|_1$ by $\|f-c\|_1$ (where c is the median of f) on the right-hand side of (1.13). However, the resulting

inequality is only formally stronger than but actually equivalent to the original one.

Proof of Corollary 1.3. (a) The proof of Part (1) is very much the same as the proof of [5; Part (1) of Theorem 4.1].

(b) By the remark after Theorem 1.2, we need only to consider $S_\nu(0)$. Take $r < \pi_0$. Then $0 \notin A$ whenever $\pi(A) \leqslant r$. Next, set $i = \inf A \geqslant 1$. Then

$$\frac{J^{(1/2)}(A \times A^c)}{\pi(A)^{1-1/\nu}} \geqslant \frac{\pi_i a_i}{\sqrt{r_{i,i-1}}} \cdot \frac{1}{\pi(A)^{1-1/\nu}} \geqslant \frac{\pi_i a_i}{\sqrt{r_{i,i-1}}\big(\sum_{j \geqslant i} \pi_j\big)^{1-1/\nu}}.$$

This proves the sufficiency of (1.14).

To prove the necessity, simply take $A = \{i, i+1, \cdots\}\,(i \geqslant 1)$ with $\pi(A) \leqslant r$. Then

$$0 < S_\nu(r) \leqslant \frac{\pi_i a_i}{\sqrt{r_{i,i-1}}\,\big(\sum_{j \geqslant i} \pi_j\big)^{1-1/\nu}}.$$

This implies (1.14). □

Proof of Theorem 1.4. (a) We show that under the first condition of Part (3), there also exists φ satisfying (1.15). Let $\{A_n\}_{n \geqslant 1}$ satisfy $A_n \cap A_m = \emptyset\,(n \neq m)$ and $\pi(A_n) > 0$ for all n. Set $\psi(x) = 1 + \pi(A_n)^{-1}$ if $x \in A_n$ and $= 1$ if $x \notin \cup_n A_n$. Then $\pi(\psi) = \infty$. Because π is a probability measure, the assertion now follows by setting $\varphi = \log \psi$.

Next, set $f_n = \exp[\varphi \wedge n]$ and $\delta_1(\varphi) = \operatorname{ess\,sup}_J |\varphi(x) - \varphi(y)|^2 r(x,y)$.

(b) We claim that $D(f_n, f_n) \leqslant C_1 \|f_n\|_2^2$ for some constant $C_1 = C_1(\varphi)$ provided one of the conditions of the theorem holds. First, assume (1). By using the Mean Value Theorem, we have

$$|e^A - e^B| \leqslant |A - B| e^{A \vee B} = |A - B|(e^A \vee e^B)$$

for all A, $B \geqslant 0$. Hence, by (1.9) and the assumption, we get

$$\begin{aligned}
&\frac{1}{2}\int J(\mathrm{d}x, \mathrm{d}y)[f_n(x) - f_n(y)]^2 \\
&\quad \leqslant \frac{1}{2}\int J^{(1)}(\mathrm{d}x, \mathrm{d}y)[\varphi(x) - \varphi(y)]^2 r(x,y)\big[f_n(x) \vee f_n(y)\big]^2 \\
&\quad \leqslant \int J^{(1)}(\mathrm{d}x, \mathrm{d}y)[\varphi(x) - \varphi(y)]^2 r(x,y) f_n(x)^2 \\
&\quad \leqslant \delta_1(\varphi)\|f_n\|_2^2.
\end{aligned}$$

The required assertion follows since $K(\mathrm{d}x)$ is bounded on $L^2(\pi)$. The proof is similar and even simple for the other two cases.

(c) For every $m \geqslant 1$, by (1.15), one may choose $r_m > 0$ such that

$$\pi(\varphi \geqslant r_m) \leqslant 1/m.$$

By Chebyshev's and Cauchy-Schwarz inequalities, we obtain

$$\|f_n\|_1 \leqslant \|f_n\|_2 m^{-1/2} + e^{r_m}\pi[\varphi < r_m].$$

(d) By assumption, $\|f_n\|_2 \uparrow \infty$ as $n \to \infty$. Hence, we have

$$\begin{aligned}
\eta_2 &\leqslant \frac{D(f_n, f_n)\|f_n\|_1^{4/\nu}}{\mathrm{Var}_\pi(f_n)^{1+2/\nu}} \\
&\leqslant \frac{C_1\|f_n\|_2^2\left\{\|f_n\|_1 m^{-1/2} + e^{r_m}\pi[\varphi < r_m]\right\}^{4/\nu}}{\left[\|f_n\|_2^2 - \left\{\|f_n\|_1 m^{-1/2} + e^{r_m}\pi[\varphi < r_m]\right\}^2\right]^{1+2/\nu}} \\
&= \frac{C_1\left\{m^{-1/2} + \|f_n\|_2^{-1} e^{r_m}\pi[\varphi < r_m]\right\}^{4/\nu}}{\left[1 - \left\{m^{-1/2} + \|f_n\|_2^{-1} e^{r_m}\pi[\varphi < r_m]\right\}^2\right]^{1+2/\nu}} \\
&\to \frac{C_1 m^{-2\nu}}{\left[1 - m^{-1}\right]^{1+2/\nu}} \qquad \text{as} \quad n \to \infty \\
&\to 0, \qquad \text{as} \quad m \to \infty.
\end{aligned}$$

This proves that $\eta_2 = 0$ and hence (1.2) does not hold. The proof for $\eta_1 = 0$ needs only a little modification in the last step. □

Proof of Theorem 1.5. Note that on the set $\{f_n(y) > f_n(x)\}$, we have

$$\begin{aligned}
0 &< f_n(y) - f_n(x) \\
&= \begin{cases} \varphi(y) - \varphi(x), & \text{if } \varphi(y),\ \varphi(x) < N_n \\ N_n - \varphi(x) \leqslant \varphi(y) - \varphi(x), & \text{if} \quad \varphi(y) \geqslant N_n \quad \text{and} \quad \varphi(x) < N_n. \end{cases}
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
&\frac{1}{2}\int J(\mathrm{d}x, \mathrm{d}y)[f_n(y) - f_n(x)]^2 \\
&= \int_{\{f_n(y) > f_n(x)\}} J(\mathrm{d}x, \mathrm{d}y)[f_n(y) - f_n(x)]^2 \\
&\leqslant \int_{\{f_n(y) > f_n(x)\}} J(\mathrm{d}x, \mathrm{d}y) I_{\{\varphi(y) > \varphi(x)\}}[\varphi(y) - \varphi(x)]^2 \\
&\lesssim \int_{\{f_n(y) > f_n(x)\}} J(\mathrm{d}x, \mathrm{d}y)\psi(y)^2 \\
&= \int_{\{f_n(y) < f_n(x)\}} J(\mathrm{d}x, \mathrm{d}y)\psi(x)^2.
\end{aligned}$$

Next, without loss of generality, assume that $\|\varphi\|_1 = 1$. Then,

$$\|f_n\|_1 \leqslant \|\varphi\|_1 = 1.$$

By (1.17), we get

$$\eta_2 \leqslant \frac{D(f_n,f_n)\|f_n\|_1^{4/\nu}}{\mathrm{Var}_\pi(f_n)^{1+2/\nu}} \lesssim \frac{C_2(n)\|f_n\|_2^{2+4/\nu}\|\varphi\|_1^{4/\nu}}{\left(\|f_n\|_2^2-\|\varphi\|_1^2\right)^{1+2/\nu}} = \frac{C_2(n)}{\left(1-\|f_n\|_2^{-2}\right)^{1+2/\nu}},$$

$$\eta_1 \leqslant \frac{[D(f_n,f_n)+\delta\|f_n\|_2^2]\,\|f_n\|_1^{4/\nu}}{\|f_n\|_2^{2+4/\nu}} \leqslant C_2(n)+\delta\|f_n\|_2^{-4/\nu}.$$

Since $\|f_n\|_2 \to \|\varphi\|_2 = \infty$, the conclusion follows by setting $n \to \infty$. □

Proof of Corollary 1.6. (a) Take $\varphi_i = \log[\psi_i/\pi_i]$. Noting that

$$\pi_i \sim \frac{b_0 b_1 \cdots b_{i-1}}{a_1 a_2 \cdots a_i},$$

Part (1) of the corollary follows from Part (2) of Theorem 1.4.

(b) To prove Part (2), set $\varphi_i = (\psi_i/\pi_i)^{1/2}$. Then

$$\|\varphi\|_2^2 = \sum_i \psi_i = \infty.$$

Moreover

$$\left(\frac{\pi_{i+1}}{\pi_i}\right)\left(\frac{\psi_{i+1}}{\psi_i}\right) = \left(\frac{b_i}{a_{i+1}}\right)\left(\frac{\psi_{i+1}}{\psi_i}\right) \leqslant \gamma/\rho < 1, \qquad i \gg 1.$$

$$\frac{\varphi_{i+1}}{\varphi_i} = \left(\frac{\psi_{i+1}\pi_i}{\psi_i\pi_{i+1}}\right)^{1/2} = \left(\frac{a_{i+1}\psi_{i+1}}{b_i\psi_i}\right)^{1/2} \geqslant (\rho\gamma)^{1/2} > 1, \quad i \gg 1. \tag{3.14}$$

Then $\|\varphi\|_1 = \sum_i \pi_i\psi_i < \infty$ and so (1.16) with $\psi = \varphi$ is satisfied.

Next, by (3.14), there is an N such that $\varphi_{i+1} > \varphi_i$ for all $i \geqslant N$. Thus, we have $N_n := \varphi_n \to \infty$, as $n \to \infty$ and furthermore

$$f_n(i) = \varphi_i \wedge \varphi_n = \varphi_{i\wedge n}$$

for large enough n. On the other hand, by assumption,

$$\frac{1}{\pi_i}\sum_{j\geqslant i}\pi_j = \sum_{j\geqslant i}\frac{b_i b_{i+1}\cdots b_{j-1}}{a_{i+1}a_{i+2}\cdots a_j} \lesssim \sum_{j\geqslant i}\rho^{i-j} = \frac{\rho}{\rho-1}.$$

Hence

$$\pi_i \leqslant \sum_{j\geqslant i}\pi_j \lesssim \frac{\rho\pi_i}{\rho-1} \quad (=\pi_i \text{ if } \rho=\infty). \tag{3.15}$$

Therefore,

$$\|f_n\|_2^2 = \sum_{i\leqslant n}\psi_i + \frac{\psi_n}{\pi_n}\sum_{j>n}\pi_j \lesssim \sum_{i\leqslant n}\psi_i.$$

Thus, for a large enough n, we have

$$\int_{\{f_n(y)<f_n(x)\}} J(\mathrm{d}x,\mathrm{d}y)\varphi(x)^2 = \sum_{i\leqslant N}[\pi_i b_i + \pi_i a_i]\varphi_i^2 + \sum_{N<i\leqslant n}\pi_i a_i\varphi_i^2 \lesssim \sum_{1\leqslant i\leqslant n} a_i\psi_i.$$

Then

$$C_2(n) = \frac{\sum_{1\leqslant i\leqslant n} a_i\psi_i}{\big(\sum_{i\leqslant n}\psi_i\big)^{1+2/\nu}} \leqslant \frac{a_n}{\big(\sum_{i\leqslant n}\psi_i\big)^{2/\nu}} \to 0 \quad \text{as} \quad n\to\infty.$$

Therefore, the required assertion follows from Theorem 1.5 with $\psi=\varphi$. □

Proof of Corollary 2.1. (a) For simplicity, write

$$A(x) = \int_x^\infty \frac{\mathrm{d}y}{a(y)}.$$

To show that

$$\inf_{x\geqslant 1}\frac{1}{a(x)^{q/2}}\Big/A(x) > 0, \qquad q := \nu/(\nu-1),$$

since $1/a(x)\to 0$ and $A(x)\to 0$ as $x\to\infty$, by the Mean Value Theorem, it suffices to prove that

$$\inf_{x\geqslant 1}\frac{a'(x)}{a(x)^{q/2+1}}\Big/\frac{1}{a(x)} > 0.$$

That is

$$\inf_{x\geqslant 1}\frac{a'(x)}{a(x)^{q/2}} > 0.$$

Now, Part (1) follows by solving this inequality and using Corollary 1.3.

(b) Take $\varphi(x)=a(x)/x$. Then by the assumption, we have

$$\begin{aligned}
&\sup_{x\geqslant 1} a(x)\big[\log\varphi(x\pm 1)-\log\varphi(x)\big]^2\\
&\quad= \sup_{x\geqslant 1} a(x)\big[(\log\varphi)'(x+\varepsilon)\big]^2 \quad (|\varepsilon|<1)\\
&\quad\lesssim \sup_{x\geqslant 1} a(x)\big[(\log\varphi)'(x)\big]^2\\
&\quad= \sup_{x\geqslant 1} a(x)\big[-a(x)/x + a'(x)\big]^2/a(x)^2\\
&\quad\lesssim \sup_{x\geqslant 1} a(x)/x^2.
\end{aligned}$$

Then the second conclusion follows from Part (1) of Corollary 1.6.

(c) It is known that $\lambda_1=0$ and hence $\sigma=0$ whenever

$$\lim_{x\to\infty} a(x)/x^{1+\varepsilon} = 0 \quad (0<\varepsilon<1)^{[4]}.$$

Thus, we may assume in what follows that

$$\xi = \lim_{x\to\infty} x(\log a(x))' > 1. \tag{3.16}$$

Otherwise, we would have $\lim_{x\to\infty} x(\log a(x))' \leqslant 1$, and hence $a(x) \lesssim x^{1+\varepsilon}$ for all small $\varepsilon > 0$. The condition (3.16) implies that

$$A(x) = \int_x^\infty \frac{\mathrm{d}y}{a(y)} \lesssim \frac{x}{a(x)}, \tag{3.17}$$

and

$$a(x) \gtrsim x^{1+\varepsilon} \qquad \text{for some } \ \varepsilon > 0. \tag{3.18}$$

Combining (3.16) with (3.18) together, we get

$$\lim_{x\to\infty} a'(x) \geqslant \lim_{x\to\infty} a(x)/x = \infty. \tag{3.19}$$

Next, take $\varphi(x) = \sqrt{x^s a(x)}$. Then

$$\|\varphi\|_2^2 = \sum_i \pi_i \varphi_i^2 \sim \sum_i i^s = \infty$$

since $s > -1$. On the other hand, we have

$$\|\varphi\|_1 = \sum_i \pi_i \varphi_i \sim \sum_{i\geqslant 1} a_i^{-1/2} i^{s/2} < \infty.$$

Finally, since $\varphi' > 0$ by (3.16), one may take $N_n = \varphi(n)$ and so $f_n(x) = \varphi(x \wedge n)$. Then

$$\sum_i \pi_i f_n(i)^2 \geqslant \sum_{i\leqslant n} \pi_i \varphi_i^2 \sim n^{s+1}.$$

Next, since for $\varepsilon \in (0,1)$,

$$\begin{aligned}\frac{\varphi'(x-\varepsilon)}{\varphi'(x)} &= \left(\frac{x-\varepsilon}{x}\right)^{s/2-1} \sqrt{\frac{a(x-\varepsilon)}{a(x)}}\ \frac{s + (x-\varepsilon)\big(\log a(x-\varepsilon)\big)'}{s + x\big(\log a(x)\big)'} \\ &\lesssim \frac{\big(\log a(x-\varepsilon)\big)'}{\big(\log a(x)\big)'},\end{aligned}$$

by assumption, we have $\sup_{x\geqslant 1,\, \varepsilon\in(0,1)} \varphi'(x-\varepsilon)/\varphi'(x) < \infty$. Hence

$$\begin{aligned} D(f_n, f_n) &= \sum_{i\leqslant n} \pi_i a_i [f_n(i-1) - f_n(i)]^2 \\ &\lesssim \sum_{i\leqslant n} \varphi'(i-\varepsilon_i)^2 \\ &\lesssim \int_1^n \varphi'(x)^2 \mathrm{d}x \\ &= \frac{1}{2}\int_1^n x^{s-2} a(x) \big[s + x(\log a(x))'\big]^2 \mathrm{d}x \\ &\lesssim \int_1^n x^{s-2} a(x) \big[x(\log a(x))'\big]^2 \mathrm{d}x \quad \text{(by (3.16))} \\ &= \int_1^n x^s a'(x)^2/a(x) \mathrm{d}x. \end{aligned}$$

Therefore

$$\begin{aligned}\frac{D(f_n,f_n)}{\|f_n\|_2^{2+4/\nu}} &\lesssim \int_1^x \frac{y^s a'(y)^2}{a(y)}\mathrm{d}y \Big/ x^{s+1+2(s+1)/\nu} \\ &\sim \big[\big(\sqrt{a(x)}\,\big)'/x^{(s+1)/\nu}\big]^2 \\ &\to 0, \qquad \text{as } x=n\to\infty.\end{aligned}$$

The third assertion now follows from Theorem 1.5. □

Proof of Corollary 2.2. Assume that $\xi > 1$. Otherwise, it was treated in the proof (c) of Corollary 2.1.

(a) Take $\varphi(x) = \sqrt{xa(x)}$. Then $\|\varphi\|_2^2 = \infty$. Next, set $f_n(x) = \varphi(x \wedge n)$. Then

$$\|f_n\|_2^2 = \sum_{i\leqslant n}\pi_i\varphi_i^2 + \varphi_n^2\sum_{i>n}\pi_i \sim \int_1^n x\mathrm{d}x + a_n nA(n) \sim n^2 \quad \text{(by (3.17))}.$$

$$\begin{aligned}\sum_i \pi_i f_n(i)^2\log f_n(i)^2 &= \sum_{i\leqslant n}\pi_i\varphi_i^2\log\varphi_i^2 + \varphi_n^2\big(\log\varphi_n^2\big)\sum_{i>n}\pi_i \\ &\sim \int_1^n x\log[xa(x)]\mathrm{d}x + \big(a_n n\log[na_n]\big)A(n) \\ &\sim \int_1^n x\log a(x)\mathrm{d}x.\end{aligned}$$

Here in the last step, we have used the fact that $a(x) \geqslant x\,(x \gg 1)$ and the inequality:

$$xa(x)A(x)\log a(x) \leqslant c_1 + c_2\int_1^\infty x\log a(x)\mathrm{d}x.$$

To see this, it suffices to show that

$$\big[a(x)\log a(x) + xa'(x)\log a(x) + xa'(x)\big]A(x) - x\log a(x) \leqslant c_2 x\log a(x).$$

Dividing both sides by $xa'(x)\log a(x)$, because of $a(x), a'(x) \to \infty$, it is enough to show that $\big[1 + a(x)/(xa'(x))\big]A(x) \leqslant c_2$. Now, the required assertion follows from (3.16).

Next,

$$\sum_i \pi_i f_n(i)^2\log\|f_n\|_2^2 \sim \Big\{\int_1^n x\mathrm{d}x + a_n nA(n)\Big\}\log\|f_n\|_2^2 \sim n^2\log n,$$

$$\begin{aligned}\sum_i \pi_i a_i \big[f_n(i-1)-f_n(i)\big]^2 &\lesssim \sum_{i\leqslant n}\varphi'(i-\varepsilon_i)^2\\ &\lesssim \int_1^n \varphi'(x)^2\mathrm{d}x\\ &= \frac{1}{2}\int_1^n \frac{[a(x)+xa'(x)]^2}{xa(x)}\mathrm{d}x\\ &= \int_1^n \frac{a(x)}{x}\big[1+xa'(x)/a(x)\big]^2\mathrm{d}x\\ &\sim \int_1^n \frac{a(x)}{x}\big[xa'(x)/a(x)\big]^2\mathrm{d}x \quad \text{(by (3.16))}\\ &= \int_1^n \frac{xa'(x)^2}{a(x)}\mathrm{d}x.\end{aligned}$$

Therefore,

$$\begin{aligned}D(f_n,f_n)\Big/\int f_n^2\log\big[f_n^2/\|f_n\|_2^2\big]\mathrm{d}\pi &\lesssim \int_1^x \frac{ya'(y)^2}{a(y)}\mathrm{d}y\Big/\int_1^x y\log a(y)\mathrm{d}y\\ &\lesssim \frac{a'(x)^2}{a(x)\log a(x)}\\ &\lesssim \frac{a''(x)}{\log a(x)+1} \quad \text{(by (3.19))}\\ &\sim \frac{a''(x)}{\log a(x)}.\end{aligned}$$

Here, in the second step, we have used the fact that

$$\int_1^x ya'(y)^2/a(y)\mathrm{d}y \gtrsim \int_1^x a'(y)\mathrm{d}y \sim a(x)\to\infty$$

as $x\to\infty$. We have thus proved Part (1) of the corollary.

(b) By [9; Theorem 1.2], (1.4) holds provided

$$\inf_{i\geqslant 1}\frac{\pi_i a_i}{\sqrt{r_{i,i-1}}}\Big/\Big[\sum_{j\geqslant i}\pi_j\Big]\Big\{\log\Big[\Big(\sum_{j\geqslant i}\pi_j\Big)^{-1}+e\Big]\Big\}^{1/2}>0, \tag{3.20}$$

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

$$A(x)=\int_x^\infty \frac{\mathrm{d}y}{a(y)}.$$

Since $a(x)\uparrow\infty$ and

$$A(x)\log A(x)=[\log A(x)]/A(x)^{-1}\sim A(x)\sim 0,$$

by assumption, we have

$$\begin{aligned}\frac{1}{\sqrt{a(x)}}\Big/A(x)\big[-\log A(x)\big]^{1/2} &\sim -\frac{a'(x)}{a(x)^{3/2}}\Big/\bigg[A'(x)\bigg(\sqrt{-\log A(x)}-\frac{1}{2\sqrt{-\log A(x)}}\bigg)\bigg]\\ &\sim -\frac{a'(x)}{a(x)^{3/2}}\Big/A'(x)\sqrt{-\log A(x)}\\ &= \big(\sqrt{a(x)}\,\big)'/\sqrt{-\log A(x)}\\ &= \Big[-\log A(x)/\big(\sqrt{a(x)}\,\big)'^2\Big]^{-1/2}\\ &= \Big[-A'(x)A(x)^{-1}/\big[\big(\sqrt{a(x)}\,\big)'^2\big]'\Big]^{-1/2}\\ &= \bigg[\frac{1}{a(x)\big[\big(\sqrt{a(x)}\,\big)'^2\big]'}\Big/A(x)\bigg]^{-1/2}\\ &\sim \bigg[\bigg(\frac{1}{a(x)\big[\big(\sqrt{a(x)}\,\big)'^2\big]'}\bigg)'\Big/A'(x)\bigg]^{-1/2}\\ &= \bigg[-a(x)\bigg(\frac{1}{a(x)\big[\big(\sqrt{a(x)}\,\big)'^2\big]'}\bigg)'\bigg]^{-1/2}.\end{aligned}$$

By assumption, this implies (3.20). □

Proof of Corollary 2.3. (a) By Corollary 1.3, it suffices to show that

$$\pi_i\sqrt{a_i}\gtrsim\bigg(\sum_{j\geqslant i}\pi_j\bigg)^{1/q},\qquad q:=(\nu-1)/\nu \tag{3.21}$$

for some $q\in(1,\infty)$. We prove that $q\neq\infty$ under the assumption. Because of $b_i=\pi_{i+1}a_{i+1}/\pi_i$ and $b_i\leqslant a_i$, we have

$$\pi_i a_i\leqslant\pi_{i-1}a_{i-1}\leqslant\cdots\leqslant\pi_1 a_1$$

and so

$$a_i\lesssim\pi^{-i}. \tag{3.22}$$

Combining this with (3.21) gives us the required assertion.

Next, by (3.15), it is enough for (3.21) that $\sqrt{a_i}\gtrsim\pi_i^{1/q-1}$. That is, $a_i\gtrsim\pi_i^{2/q-2}$. Combining this with (3.22) proves $I_\nu>0$ as well as $S_{\nu,\delta}>0$ and then Part (1) of the corollary.

(b) Part (2) is a simple application of Part (2) of Corollary 1.6 with $\psi_i=\gamma^i$. □

Proof of Corollary 2.4. (a) Take $\psi_i=i^\gamma$ ($\gamma>-1$) and set $\varphi_i=\sqrt{\psi_i/\pi_i}$. Then $\|\varphi\|_2=\infty$. Next, set $f_n(i)=\varphi_{i\wedge n}$. Then $\|f_n\|_2^2\sim n^{1+\gamma}$. Moreover,

$$\sum_i\pi_i a_i\big[f_n(i-1)-f_n(i)\big]^2\sim\sum_{i\leqslant n}a_i\psi_i\bigg[\sqrt{\bigg(\frac{i-1}{i}\bigg)^\gamma\frac{b_{i-1}}{a_i}}-1\bigg]^2\sim\sum_{i\leqslant n}a_i\psi_i.$$

On the other hand, by (3.15) and assumption, we have

$$\begin{aligned}\sum_i \pi_i f_n(i)^2 \log f_n(i)^2 &= \sum_{i\leqslant n} \psi_i \log \frac{\psi_i}{\pi_i} + \frac{\psi_n}{\pi_n}\left[\log \frac{\psi_n}{\pi_n}\right] \sum_{k>n} \pi_k \\ &\sim \sum_{i\leqslant n} \psi_i \log \frac{\psi_i}{\pi_i} \\ &\sim \sum_{i\leqslant n} \psi_i \log \pi_i^{-1}.\end{aligned}$$

Besides,

$$\sum_i \pi_i f_n(i)^2 \log \|f_n\|_2^2 \sim \sum_{i\leqslant n} \psi_i \log n.$$

Thus,

$$D(f_n, f_n)/\int f_n^2 \log\left[f_n^2/\|f_n\|_2^2\right] \mathrm{d}\pi \lesssim \sum_{i\leqslant n} a_i\psi_i / \sum_{i\leqslant n} \psi_i \log \pi_i^{-1}.$$

The conclusion now follows by using Stolz Theorem.

(b) By (3.15) and (3.20), we have

$$\theta_i := \pi_i\sqrt{a_i}\Big/\left[\sum_{j\geqslant i}\pi_j\right]\left\{\log\left[\left(\sum_{j\geqslant i}\pi_j\right)^{-1} + e\right]\right\}^{1/2} \sim \left(a_i/\log \pi_i^{-1}\right)^{1/2}.$$

Thus, $\underline{\lim}_{n\to\infty} \theta_n > 0$ iff $\underline{\lim}_{n\to\infty} a_n/\log \pi_n^{-1} > 0$. $\square$

References

1. Bakry D., *L'hypercontractivité et son utilisation en théorie des semigroupes*, LNM, Springer, 1992, **1581**.
2. Carlen E. A. Kusuoka S. Stroock D. W., *Upper bounds for symmetric Markov transition functions*, Ann. Inst. Henri Poincaré 1987, no. 2, 245–287.
3. Chen M. F., *From Markov Chains to Non-Equilibrium Particle Systems*, World Scientific, 1992.
4. Chen M. F., *Estimation of spectral gap for Markov chains*, Acta Math Sin New Ser 1996, 12:4, 337–360.
5. Chen M. F. Wang F. Y., *Cheeger's inequalities for general symmetric forms and existence criteria for spectral gap*, Preprint. Abstract. Chin Sci Bulletin 1998, 43:14, 1475–1477 (Chinese Edition); 1998, 43:18, 1516–1519 (English Edition).
6. Nash J., *Continuity of solutions of parabolic and elliptic equations*, Amer J Math 1958, 80, 931–954.
7. Saloff-Coste L., *Lectures on finite Markov chains*, LNM Springer-Verlag, 1997, **1665**, 301–413,.
8. Varopoulos N. Th., *Isoperimetric inequalities and Markov chains*, J Funct Anal 1985, 63, 215–239.
9. Wang F. Y., *Sobolev type inequalities for general symmetric forms*, to appear in Proc Amer Math Soc 1999.

4. Appendix. For referee's reference but not for publication

Proof of the equivalence of (1.2) *and* (1.6). In order to use the spectral theory of the symmetric semigroups, some standard conditions are needed in the proof but we omit the details here. One may refer to [2].

Assume that $(D, \mathscr{D}(D))$ determines a symmetric semigroup $(P_t)_{t\geqslant 0}$ with generator $(\Omega, \mathscr{D}(\Omega))$ on $L^2(\pi)$.

(a) (1.2) $\Longrightarrow$ (1.6). Let $f \in \mathscr{D}(\Omega) \subset L^2(\pi)$ with $\|f\|_p = 1$. Set $f_t = P_t f$ and

$$u_t = \|f_t - \pi(f)\|_2^2 = \mathrm{Var}_\pi(f_t) \quad (\text{since } \pi(f_t) = \pi(f)).$$

Then, noticing that $\|f_t\|_p \leqslant \|f\|_p = 1$ and $D(1,1) = \Omega 1 = 0$, by (1.2), we obtain

$$-u_t' = 2D(f_t, f_t) \geqslant 2\eta_2 \mathrm{Var}_\pi(f_t)^{1+2/\nu} = 2\eta_2 u_t^{1+2/\nu}.$$

Next, set

$$v_t = \frac{\nu}{4\eta_2} u_t^{-2/\nu}.$$

Then, $v_0 \geqslant 0$,

$$v_t' = -\frac{1}{2\eta_2} u_t^{-2/\nu-1} u_t' \geqslant 1.$$

Hence, $v_t \geqslant t$ and so

$$u_t \leqslant \Big(\frac{\nu}{4\eta_2 t}\Big)^{\nu/2}.$$

In other words, we have

$$\|P_t\|_{p\to 2} \leqslant \Big(\frac{\nu}{4\eta_2 t}\Big)^{\nu/4}.$$

Therefore,

$$\|P_t\|_{p\to q} \leqslant \|P_{t/2}\|_{p\to 2}\, \|P_{t/2}\|_{2\to q} = \|P_{t/2}\|_{p\to 2}^2 \leqslant \left(\frac{\nu}{2\eta_2 t}\right)^{\nu/2}, \qquad \frac{1}{p} + \frac{1}{q} = 1.$$

(b) (1.6) $\Longrightarrow$ (1.2). Let $f \in \mathscr{D}(\Omega)$ with $\|f\|_p = 1$. Set $f_t = P_t f - \pi(f)$. By (1.6), we have

$$\|f_t\|_q \leqslant \left(\frac{\nu}{2\eta_2 t}\right)^{\nu/2} \|f\|_p = \left(\frac{\nu}{2\eta_2 t}\right)^{\nu/2}.$$

Because

$$|(f, f_t)| = \left| \int \pi(\mathrm{d}x) f(x) f_t(x) \right| \leqslant \|f_t\|_q \|f\|_p, \quad f_t = f - \pi(f) - \int_0^t \Omega P_s f \mathrm{d}s,$$

we have

$$\begin{aligned} \left(\frac{\nu}{2\eta_2 t}\right)^{\nu/2} &\geqslant (f, f_t) = \|f\|_2^2 - (f, \pi(f)) - \int_0^t (f, \Omega P_s f) \mathrm{d}s \\ &\geqslant \mathrm{Var}_\pi(f) - tD(f,f). \end{aligned} \tag{4.1}$$

Put $B = \left(\frac{\nu}{2\eta_2}\right)^{\nu/2}$, $h_t = Bt^{-\nu/2} - \mathrm{Var}_\pi(f) + tD(f,f)$. Then

$$h_t' = -\frac{\nu B}{2t^{\nu/2+1}} + D(f,f).$$

We get the minimum point of h_t:

$$t_0^{\nu/2+1} = \frac{\nu B}{2D(f,f)}. \tag{4.2}$$

Combining (4.1) with (4.2), we get

$$\frac{B}{t_0^{\nu/2+1}} \geqslant \frac{1}{t_0}\mathrm{Var}_\pi(f) - D(f,f).$$

That is,

$$\mathrm{Var}_\pi(f) \leqslant t_0\left[D(f,f) + \frac{B}{t_0^{\nu/2+1}}\right] = \left(1+\frac{2}{\nu}\right)\left[\frac{\nu B}{2}\right]^{\frac{1}{\nu/2+1}} D(f,f)^{\frac{\nu}{\nu+2}}.$$

Or,

$$\mathrm{Var}_\pi(f)^{1+2/\nu} \leqslant \left[1+\frac{2}{\nu}\right]^{1+2/\nu}\left[\frac{\nu B}{2}\right]^{2/\nu} D(f,f) = \frac{1}{\eta_2}\left[1+\frac{\nu}{2}\right]^{1+2/\nu} D(f,f)\,\|f\|_p^{4/\nu}.$$

□

Proof of Lemma 3.1.

$$\begin{aligned}\int J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y)|f(y)-f(x)| &= 2\int_{[f(y)>f(x)]} J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y)[f(y)-f(x)]\\ &= 2\int_{[f(y)>f(x)]} J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y)\int_{f(x)}^{f(y)}\mathrm{d}t\\ &= 2\int_0^{\|f\|_u}\mathrm{d}t\int_{[f(y)\geqslant t>f(x)]} J^{(\alpha)}(\mathrm{d}x,\mathrm{d}y)\\ &= 2\int_0^{\|f\|_u} J^{(\alpha)}(F_t\times F_t^c)\mathrm{d}t. \quad \square\end{aligned}$$

Proof of Lemma 3.2. Denote by J_ν the right-hand side of the formula given in the lemma. 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

$$\begin{aligned}
&\frac{1}{2}\int J(\mathrm{d}x,\mathrm{d}y)|f(y)-f(x)| \\
&= \frac{1}{2}\int J(\mathrm{d}x,\mathrm{d}y)\big[|f_+(y)-f_+(x)| + |f_-(y)-f_-(x)|\big] \\
&= \int_0^{\|f\|_u} \big[J\big(F_t^+ \times (F_t^+)^c\big) + J\big(F_t^- \times (F_t^-)^c\big)\big]\mathrm{d}t \quad \text{(by co-area formula)} \\
&\geqslant I_\nu \int_0^{\|f\|_u} \big[\pi(F_t^+)^{1/q} + \pi(F_t^-)^{1/q}\big]\mathrm{d}t.
\end{aligned}$$

Note that by Theorem 4.1 below,

$$\pi(F_t^\pm)^{1/q} = \|I_{F_t^\pm}\|_q = \sup_{\|g\|_r \leqslant 1} \langle I_{F_t^\pm}, g\rangle, \qquad \frac{1}{r}+\frac{1}{q}=1.$$

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

$$\frac{1}{2}\int J(\mathrm{d}x,\mathrm{d}y)|f(y)-f(x)| \geqslant I_\nu \|f-c\|_q. \quad \square$$

Proof of Part (1) *of Corollary* 1.3. (a) Let $I_\nu > 0$. Take $A = I_i = \{i, i+1, \cdots\}$ for a fixed $i > 0$ and

$$J^{(\alpha)}(i,j) = \frac{\pi_i q_{ij}}{[q_i \vee q_j]^\alpha} = \begin{cases} \dfrac{\pi_i a_i}{[(a_i+b_i)\vee(a_{i-1}+b_{i-1})]^\alpha} =: \pi_i \tilde{a}_i, & \text{if } j = i-1 \\ \dfrac{\pi_i b_i}{[(a_i+b_i)\vee(a_{i+1}+b_{i+1})]^\alpha} =: \pi_i \tilde{b}_i, & \text{if } j = i+1. \end{cases}$$

Then

$$2I_\nu \leqslant \frac{J^{(\alpha)}(A\times A^c)}{[\pi(A)\wedge\pi(A^c)]^{1/q}} = \frac{\pi_i \tilde{a}_i}{\big[\big(\sum_{j\geqslant i}\pi_j\big)\wedge\big(\sum_{j<i}\pi_j\big)\big]^{1/q}} \leqslant \frac{\pi_i \tilde{a}_i}{\big[\pi_0 \sum_{j\geqslant i}\pi_j\big]^{1/q}},$$

where $q := (\nu-1)/\nu$. This proves the necessity of the condition.

(b) Next, assume that the condition holds. Then for each A with $\pi(A) \in (0,1)$, since the symmetry of $J^{(\alpha)}$, we may assume that $0 \notin A$. Set $i_0 = \min A \geqslant 1$. Then, $A \subset I_{i_0}$, $A^c \subset E \setminus \{i_0\}$ and so

$$\frac{J^{(\alpha)}(A\times A^c)}{\big[\pi(A)\wedge\pi(A^c)\big]^{1/q}} \geqslant \frac{\pi_{i_0}\tilde{a}_{i_0}}{\big[\sum_{j\geqslant i_0}\pi_j\big]^{1/q}} \geqslant c.$$

Because A is arbitrary, we obtain the required assertions. $\square$

Theorem 4.1. Let $p \geqslant 1$. Then $\|f\|_p \leqslant F$ iff $\|fg\|_1 \leqslant FG$ holds for all g satisfying $\|g\|_q \leqslant G$.

Theorem 4.2 (Hölder-Minkowski inequality). Let μ and ν be σ-finite non-negative measures on $(E_1, \mathscr{E}_1)$ and $(E_2, \mathscr{E}_2)$ respectively, $p \in [1, \infty)$ and $f \geqslant 0$. Then

$$\left\{ \int_{E_1} \mu(\mathrm{d}x) \left[\int_{E_2} f(x,y)\nu(\mathrm{d}y) \right]^p \right\}^{1/p} \leqslant \int_{E_2} \nu(\mathrm{d}y) \left[\int_{E_1} f(x,y)^p \mu(\mathrm{d}x) \right]^{1/p}.$$

When ν is a finite counting measure, this is the usual Minkowski inequality. The theorem says that the $L^p(\mu)$-norm of the integral of a bivariate function f w.r.t. ν is controlled by the integral w.r.t. ν of the $L^p(\mu)$-norm of f:

$$\big\| \|f\|_{\nu,1} \big\|_{\mu,p} \leqslant \big\| \|f\|_{\mu,p} \big\|_{\nu,1}, \qquad p \in [1, \infty).$$

Proof of Theorem 4.1. The necessity comes from the Hölder inequality. To prove the sufficiency, assume that $\|f\|_p > F$. Set $f_n = f \wedge n$. Then for large enough n, we have $\|f_n\|_p > F$. Take $g = f_n^{p-1} G / \|f_n\|_p^{p/q}$. Then $\|g\|_q = G$ and moreover,

$$\|fg\|_1 \geqslant \|f_n g\|_1 = G\|f_n^p\|_1 / \|f_n\|_p^{p/q} = G\|f_n\|_p > FG,$$

which is a contradiction. □

Proof of Theorem 4.2. Set $J(x) = \int_{E_2} f(x,y)\nu(\mathrm{d}y)$. Then by Theorem 4.1, we know that

$$\int_{E_1} J^p(x)\mu(\mathrm{d}x) \leqslant M^p \tag{4.3}$$

iff $\int_{E_1} Jg\mathrm{d}\mu \leqslant M$ holds for all g satisfying $\int g^q \mathrm{d}\mu \leqslant 1$. However,

$$\begin{aligned}
\int_{E_1} Jg\mathrm{d}\mu &= \int_{E_1} g(x)\mu(\mathrm{d}x) \int_{E_2} f(x,y)\nu(\mathrm{d}y) \\
&= \int_{E_2} \nu(\mathrm{d}y) \int_{E_1} \mu(\mathrm{d}x) g(x) f(x,y) \quad \text{(by Fubini Theorem)} \\
&\leqslant \int_{E_2} \nu(\mathrm{d}y) \left[\int_{E_1} \mu(\mathrm{d}x) f(x,y)^p \right]^{1/p} \left[\int_{E_1} \mu(\mathrm{d}x) g(x)^q \right]^{1/q} \\
&\qquad \text{(by Hölder inequality)} \\
&\leqslant \int_{E_2} \nu(\mathrm{d}y) \left[\int_{E_1} \mu(\mathrm{d}x) f(x,y)^p \right]^{1/p}.
\end{aligned}$$

This shows that we can take the right-hand side of the last inequality as the required upper bound M in (4.3). □

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