Landscape modification meets spin systems: from torpid to rapid mixing, tunneling and annealing in the low-temperature regime

Michael Choi

Department of Statistics and Data Science and Yale-NUS College
National University of Singapore (NUS)

17th Workshop on Markov Processes and Related Topics

November 27th 2022

Introduction

- This talks centers around a technique that we call landscape modification.

Introduction

- This talks centers around a technique that we call landscape modification.
- I hope to convince you that this is a promising acceleration technique: this has successfully been applied to spin systems to yield rapidly mixing algorithms with a novel use of the global minimum value to adjust the landscape for acceleration, while the same algorithm on the original landscape mixes torpidly.

Introduction

- Let π be a discrete or continuous distribution.
  Goal: Sample from π or estimate $\pi(f)$, where

$$\pi(f) = \sum_x f(x)\pi(x), \quad \text{or} \quad \pi(f) = \int f(x)\pi(dx).$$

Introduction

- Let π be a discrete or continuous distribution.
  Goal: Sample from π or estimate $\pi(f)$, where
  $$\pi(f) = \sum_x f(x)\pi(x), \quad \text{or} \quad \pi(f) = \int f(x)\pi(dx).$$
- **Difficulty**: At times it is impossible to apply classical Monte Carlo methods, since π is often of the form
  $$\pi(x) = \frac{e^{-\beta H(x)}}{Z},$$
  where Z is a normalization constant that cannot be computed.

Introduction

- Let π be a discrete or continuous distribution.
  Goal: Sample from π or estimate $\pi(f)$, where
  $$\pi(f) = \sum_x f(x)\pi(x), \quad \text{or} \quad \pi(f) = \int f(x)\pi(dx).$$
- **Difficulty**: At times it is impossible to apply classical Monte Carlo methods, since π is often of the form
  $$\pi(x) = \frac{e^{-\beta H(x)}}{Z},$$
  where Z is a normalization constant that cannot be computed.
- **Idea of Markov chain Monte Carlo (MCMC)**:
  Construct a Markov chain that converges to π, which only depends on the ratio
  $$\frac{\pi(y)}{\pi(x)}.$$
  Thus there is no need to know Z.

The Metropolis-Hastings algorithm

- Two ingredients:
  (i). Target distribution: π
  (ii). Proposal chain with transition matrix $Q = (Q(x, y))_{x,y}$.

The Metropolis-Hastings algorithm

Algorithm 1: The Metropolis-Hastings algorithm

Input: Proposal chain Q, target distribution π

Given X_n, generate $Y_n \sim Q(X_n, \cdot)$

Take

$$X_{n+1} = \begin{cases} Y_n, & \text{with probability } \alpha(X_n, Y_n), \\ X_n, & \text{with probability } 1 - \alpha(X_n, Y_n), \end{cases}$$

where

$$\alpha(x,y) := \min \left\{ \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}, 1 \right\}$$

is known as the acceptance probability.

The Metropolis-Hastings algorithm

Definition

The Metropolis-Hastings algorithm, with proposal chain Q and target distribution π, is a Markov chain $X = (X_n)_{n \geq 1}$ with transition matrix

$$P(x,y) = \begin{cases} \alpha(x,y)Q(x,y), & \text{for } x \neq y, \\ 1 - \sum_{y;\ y \neq x} P(x,y), & \text{for } x = y. \end{cases}$$

The Metropolis-Hastings (MH) algorithm

Theorem

Given target distribution π and proposal chain Q, the Metropolis-Hastings chain is

- ***reversible**, that is, for all x, y,*

$$\pi(x)P(x,y) = \pi(y)P(y,x).$$

- *(Ergodic theorem of MH) If P is irreducible, then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(X_i) = \pi(f).$$

The Metropolis-Hastings algorithm

- Different choices of Q give rise to different MH algorithms

The Metropolis-Hastings algorithm

- Different choices of Q give rise to different MH algorithms
- **Symmetric MH**: We take a symmetric proposal chain with $Q(x, y) = Q(y, x)$, and so the acceptance probability is

$$\alpha(x, y) = \min\left\{\frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)}, 1\right\} = \min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\}.$$

The Metropolis-Hastings algorithm

- Different choices of Q give rise to different MH algorithms
- **Symmetric MH**: We take a symmetric proposal chain with $Q(x,y) = Q(y,x)$, and so the acceptance probability is

$$\alpha(x,y) = \min\left\{\frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}, 1\right\} = \min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\}.$$

- **Random walk MH**: We take a random walk proposal chain with $Q(x,y) = Q(y-x)$. E.g., $Q(x,\cdot)$ is the probability density function of $N(x,\sigma^2)$.

The Metropolis-Hastings algorithm

- Different choices of Q give rise to different MH algorithms
- **Symmetric MH**: We take a symmetric proposal chain with $Q(x,y) = Q(y,x)$, and so the acceptance probability is

$$\alpha(x,y) = \min\left\{\frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}, 1\right\} = \min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\}.$$

- **Random walk MH**: We take a random walk proposal chain with $Q(x,y) = Q(y-x)$. E.g., $Q(x,\cdot)$ is the probability density function of $N(x,\sigma^2)$.
- **Independence sampler**: Here we take $Q(x,y) = q(y)$, where $q(y)$ is a probability distribution. In words, $Q(x,y)$ does not depend on x.

Simulated annealing

- **Goal**: Find the global minimizer(s) of a target function U.

Simulated annealing

- **Goal**: Find the global minimizer(s) of a target function U.
- **Idea of simulated annealing**: Construct a **non-homogeneous** Metropolis-Hastings Markov chain that converges to π_∞, which is supported on the set of global minima of U.

Simulated annealing

- **Goal**: Find the global minimizer(s) of a target function U.
- **Idea of simulated annealing**: Construct a **non-homogeneous** Metropolis-Hastings Markov chain that converges to π_∞, which is supported on the set of global minima of U.
- Target distribution: Gibbs distribution $\pi_{T(t)}$ with temperature $T(t)$ that depends on time t

$$\pi_{T(t)}(x) = \frac{e^{-U(x)/T(t)}}{Z_{T(t)}},$$

$$Z_{T(t)} = \sum_x e^{-U(x)/T(t)}.$$

Proposal chain Q: symmetric

Simulated annealing

- The temperature cools down $T(t) \to 0$ as $t \to \infty$, and we expect the Markov chain get "frozen" at the set of global minima U_{min}:

$$\pi_\infty(x) := \lim_{t\to\infty} \pi_{T(t)}(x) = \begin{cases} \dfrac{1}{|U_{min}|}, & \text{for } x \in U_{min}, \\ 0, & \text{for } x \notin U_{min}. \end{cases}$$

$$U_{min} := \{x;\ U(x) \leq U(y) \text{ for all } y\}.$$

Simulated annealing

Algorithm 2: Simulated annealing

Input: Symmetric proposal chain Q, target distribution $\pi_{T(t)}$, temperature schedule $T(t)$

Given X_t, generate $Y_t \sim Q(X_t, \cdot)$

Take

$$X_{t+1} = \begin{cases} Y_t, & \text{with probability } \alpha_t(X_t, Y_t), \\ X_t, & \text{with probability } 1 - \alpha_t(X_t, Y_t), \end{cases}$$

where

$$\alpha_t(x, y) := \min\left\{\frac{\pi_{T(t)}(y)Q(y, x)}{\pi_{T(t)}(x)Q(x, y)}, 1\right\} = \min\left\{e^{\frac{U(x)-U(y)}{T(t)}}, 1\right\}$$

is the acceptance probability.

Optimal cooling schedule

- The temperature schedule $T(t)$ cannot be too slow: it may take too long for the Markov chain to converge

Optimal cooling schedule

- The temperature schedule $T(t)$ cannot be too slow: it may take too long for the Markov chain to converge
- $T(t)$ cannot converge to zero too fast: we can prove that with positive probability the Markov chain may get stuck at local minimum.

Optimal cooling schedule

- The temperature schedule $T(t)$ cannot be too slow: it may take too long for the Markov chain to converge
- $T(t)$ cannot converge to zero too fast: we can prove that with positive probability the Markov chain may get stuck at local minimum.

Theorem (Hajek '88, Holley and Stroock '88)

The Markov chain generated by simulated annealing converges to π_∞ if and only if for any $\epsilon > 0$,

$$T(t) = \frac{c + \epsilon}{\ln(t+1)},$$

where c is known as the optimal hill-climbing constant that depends on the target function U and proposal chain Q.

Simulated annealing (SA)

- Let $U : \mathbb{R}^d \to \mathbb{R}$ be a differentiable target function to minimize.

Simulated annealing (SA)

- Let $U : \mathbb{R}^d \to \mathbb{R}$ be a differentiable target function to minimize.
- Overdamped Langevin diffusion $(\mathcal{Z}_t)_{t \geq 0}$:

Definition (Overdamped Langevin)

The SDE of overdamped Langevin is given by

$$d\mathcal{Z}_t = -\nabla U(\mathcal{Z}_t)\, dt + \sqrt{2\epsilon_t} dB_t, \tag{1}$$

where $(B_t)_{t \geq 0}$ is the standard d-dimensional Brownian motion and $(\epsilon_t)_{t \geq 0}$ is the temperature or cooling schedule.

Simulated annealing (SA)

- Let $U : \mathbb{R}^d \to \mathbb{R}$ be a differentiable target function to minimize.
- Overdamped Langevin diffusion $(\mathcal{Z}_t)_{t \geq 0}$:

Definition (Overdamped Langevin)

The SDE of overdamped Langevin is given by

$$d\mathcal{Z}_t = -\nabla U(\mathcal{Z}_t)\, dt + \sqrt{2\epsilon_t} dB_t, \tag{1}$$

where $(B_t)_{t \geq 0}$ is the standard d-dimensional Brownian motion and $(\epsilon_t)_{t \geq 0}$ is the temperature or cooling schedule.

- The instantaneous stationary distribution at time t is the Gibbs distribution

$$\mu^0_{\epsilon_t}(x) \propto e^{-\frac{1}{\epsilon_t} U(x)}.$$

Simulated annealing (SA)

- Let $U : \mathbb{R}^d \to \mathbb{R}$ be a differentiable target function to minimize.
- Overdamped Langevin diffusion $(\mathcal{Z}_t)_{t \geq 0}$:

Definition (Overdamped Langevin)

The SDE of overdamped Langevin is given by

$$d\mathcal{Z}_t = -\nabla U(\mathcal{Z}_t)\, dt + \sqrt{2\epsilon_t} dB_t, \tag{1}$$

where $(B_t)_{t \geq 0}$ is the standard d-dimensional Brownian motion and $(\epsilon_t)_{t \geq 0}$ is the temperature or cooling schedule.

- The instantaneous stationary distribution at time t is the Gibbs distribution

$$\mu^0_{\epsilon_t}(x) \propto e^{-\frac{1}{\epsilon_t} U(x)}.$$

- The overdamped Langevin diffusion is widely used in sampling, e.g. ULA or MALA (Roberts and Tweedie '96)

Simulated annealing (SA)

- Convergence of SA depends on a constant E_* that is called the **critical height** or the hill-climbing constant.

Simulated annealing (SA)

- Convergence of SA depends on a constant E_* that is called the **critical height** or the hill-climbing constant.
- $$E_* := \sup_{x,y \in \mathbb{R}^d} \inf_{\gamma \in \Gamma_{x,y}} \left\{ \sup_t \{U(\gamma(t))\} - U(x) - U(y) + \inf U \right\},$$

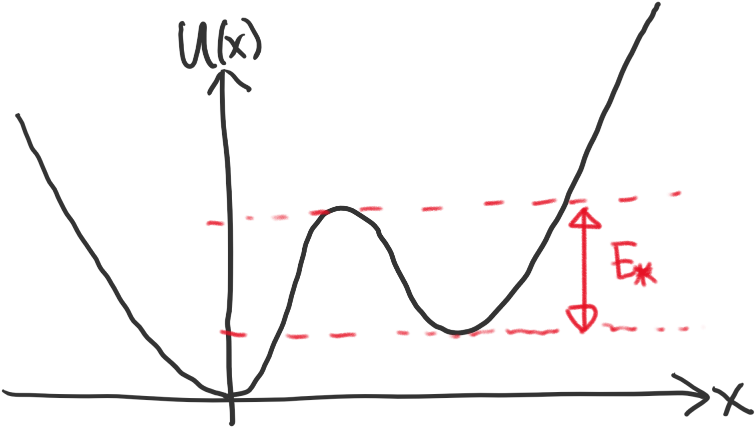
  where for two points $x, y \in \mathbb{R}^d$, we write $\Gamma_{x,y}$ to be the set of C^1 parametric curves that start at x and end at y.

Simulated annealing (SA)

- Convergence of SA depends on a constant E_* that is called the **critical height** or the hill-climbing constant.
- $$E_* := \sup_{x,y \in \mathbb{R}^d} \inf_{\gamma \in \Gamma_{x,y}} \left\{ \sup_t \{U(\gamma(t))\} - U(x) - U(y) + \inf U \right\},$$

  where for two points $x, y \in \mathbb{R}^d$, we write $\Gamma_{x,y}$ to be the set of C^1 parametric curves that start at x and end at y.
- Intuitively speaking, E_* is the largest hill one need to climb starting from a local minimum to a fixed global minimum.

What is E_*?



Convergence of SA

Theorem (Convergence of SA (Chiang et al. '87, Holley et al. '89, Jacquot '92, Miclo '92 ...))

Under the logarithmic cooling schedule of the form

$$\epsilon_t = \frac{E}{\ln t}, \quad \textit{large enough}\, t, \tag{2}$$

where $E > E_*$*, for any* $\delta > 0$ *we have*

$$\lim_{t \to \infty} \mathbb{P}\left(U(\mathcal{Z}_t) > \inf U + \delta\right) = 0.$$

Improved simulated annealing (ISA)

- Many techniques have been developed in the literature to accelerate the convergence of Langevin diffusion, e.g. preconditioning (Li et al. '16), use of Lévy noise (Simsekli '17), generalized Langevin dynamics (Chak et al. '20), anti-symmetric perturbation of drift (Hwang et al. '93, Duncan et al. '17)...

Improved simulated annealing (ISA)

- Many techniques have been developed in the literature to accelerate the convergence of Langevin diffusion, e.g. preconditioning (Li et al. '16), use of Lévy noise (Simsekli '17), generalized Langevin dynamics (Chak et al. '20), anti-symmetric perturbation of drift (Hwang et al. '93, Duncan et al. '17)...
- In our talk today we will focus on a variant of overdamped Langevin diffusion with **state-dependent** diffusion coefficient, introduced by Fang et al. (SPA '97)

Improved simulated annealing (ISA)

- Improved overdamped Langevin diffusion $(Z_t)_{t \geq 0}$:

Definition (Improved overdamped Langevin)

The SDE of improved overdamped Langevin is given by

$$dZ_t = -\nabla U(Z_t)\, dt + \sqrt{2\left(f((U(Z_t) - c)_+) + \epsilon_t\right)}\, dB_t. \quad (3)$$

Improved simulated annealing (ISA)

- Improved overdamped Langevin diffusion $(Z_t)_{t \geq 0}$:

Definition (Improved overdamped Langevin)

The SDE of improved overdamped Langevin is given by

$$dZ_t = -\nabla U(Z_t)\, dt + \sqrt{2\left(f((U(Z_t) - c)_+) + \epsilon_t\right)}\, dB_t. \tag{3}$$

- Two parameters are introduced:
  - c: It is chosen such that $c > \inf U$
  - $f : \mathbb{R} \to \mathbb{R}^+$ twice-differentiable, non-negative, bounded and non-decreasing with $f(0) = f'(0) = f''(0) = 0$.

Improved simulated annealing (ISA)

- Improved overdamped Langevin diffusion $(Z_t)_{t\geq 0}$:

Definition (Improved overdamped Langevin)

The SDE of improved overdamped Langevin is given by

$$dZ_t = -\nabla U(Z_t)\,dt + \sqrt{2\left(f((U(Z_t)-c)_+)+\epsilon_t\right)}\,dB_t. \tag{3}$$

- Two parameters are introduced:
  - c: It is chosen such that $c > \inf U$
  - $f : \mathbb{R} \to \mathbb{R}^+$ twice-differentiable, non-negative, bounded and non-decreasing with $f(0) = f'(0) = f''(0) = 0$.
- The instantaneous stationary distribution at time t is

$$\mu^f_{\epsilon_t}(x) \propto \frac{1}{f((U(x)-c)_+)+\epsilon_t} \exp\left(-\int_{\inf U}^{U(x)} \frac{1}{f((u-c)_+)+\epsilon_t}\,du\right)$$

Improved simulated annealing (ISA)

- Improved overdamped Langevin diffusion $(Z_t)_{t\geq 0}$:

Definition (Improved overdamped Langevin)

The SDE of improved overdamped Langevin is given by

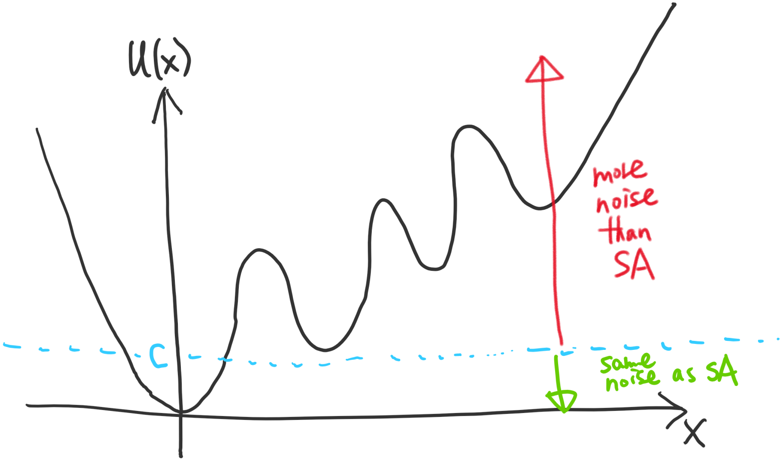
$$dZ_t = -\nabla U(Z_t)\,dt + \sqrt{2\,(f((U(Z_t)-c)_+) + \epsilon_t)}\,dB_t. \qquad (3)$$

- Two parameters are introduced:
  - c: It is chosen such that $c > \inf U$
  - $f : \mathbb{R} \to \mathbb{R}^+$ twice-differentiable, non-negative, bounded and non-decreasing with $f(0) = f'(0) = f''(0) = 0$.
- The instantaneous stationary distribution at time t is

$$\mu^f_{\epsilon_t}(x) \propto \frac{1}{f((U(x)-c)_+) + \epsilon_t} \exp\left(-\int_{\inf U}^{U(x)} \frac{1}{f((u-c)_+) + \epsilon_t}\,du\right)$$

- If $f = 0$, then $\sqrt{2\,(f((U(Z_t)-c)_+) + \epsilon_t)} = \sqrt{2\epsilon_t}$, which reduces to the classical overdamped Langevin.

Idea of ISA



Convergence of ISA

- The idea of using state-dependent noise makes sense intuitively. However, is there convergence guarantee that this improved Langevin dynamics ISA converges faster?

Convergence of ISA

- The idea of using state-dependent noise makes sense intuitively. However, is there convergence guarantee that this improved Langevin dynamics ISA converges faster?
- Yes.

Convergence of ISA

Theorem (Convergence of ISA (Fang et al. '97))

Under the logarithmic cooling schedule of the form

$$\epsilon_t = \frac{E}{\ln t}, \quad \textit{large enough } t,$$

where $E > c_*$*, for any* $\delta > 0$ *we have*

$$\lim_{t\to\infty} \mathbb{P}\left(U(Z_t) > \inf U + \delta\right) = 0.$$

Convergence of ISA

Theorem (Convergence of ISA (Fang et al. '97))

Under the logarithmic cooling schedule of the form

$$\epsilon_t = \frac{E}{\ln t}, \quad \textit{large enough } t,$$

where $E > c_*$, *for any* $\delta > 0$ *we have*

$$\lim_{t \to \infty} \mathbb{P}\left(U(Z_t) > \inf U + \delta\right) = 0.$$

- Key ingredient in the proof: both the relaxation time (i.e. inverse of the spectral gap) and the log-Sobolev constant are of the order $\mathcal{O}\left(\exp\left\{\frac{c_*}{\epsilon_t}\right\}\right)$.

c_*: the clipped critical height

- Recall the critical height E_* in SA:

$$E_* = \sup_{x,y \in \mathbb{R}^d} \inf_{\gamma \in \Gamma_{x,y}} \left\{ \sup_t \{U(\gamma(t))\} - U(x) - U(y) + \inf U \right\}$$

c_*: the clipped critical height

- Recall the critical height E_* in SA:

$$E_* = \sup_{x,y \in \mathbb{R}^d} \inf_{\gamma \in \Gamma_{x,y}} \left\{ \sup_t \{U(\gamma(t))\} - U(x) - U(y) + \inf U \right\}$$

- The **clipped critical height** c_* is defined to be

$$c_* := \sup_{x,y \in \mathbb{R}^d} \inf_{\gamma \in \Gamma_{x,y}} \left\{ \sup_t \{U(\gamma(t)) \wedge c\} - U(x) \wedge c - U(y) \wedge c + \inf U \right\}.$$

c_*: the clipped critical height

- Recall the critical height E_* in SA:

$$E_* = \sup_{x,y \in \mathbb{R}^d} \inf_{\gamma \in \Gamma_{x,y}} \left\{ \sup_t \{U(\gamma(t))\} - U(x) - U(y) + \inf U \right\}$$

- The **clipped critical height** c_* is defined to be

$$c_* := \sup_{x,y \in \mathbb{R}^d} \inf_{\gamma \in \Gamma_{x,y}} \left\{ \sup_t \{U(\gamma(t)) \wedge c\} - U(x) \wedge c - U(y) \wedge c + \inf U \right\}.$$

- One way to understand c_*: pretend that we are minimizing $U \wedge c$ instead

c_*: the clipped critical height

- Recall the critical height E_* in SA:

$$E_* = \sup_{x,y \in \mathbb{R}^d} \inf_{\gamma \in \Gamma_{x,y}} \left\{ \sup_t \{U(\gamma(t))\} - U(x) - U(y) + \inf U \right\}$$

- The **clipped critical height** c_* is defined to be

$$c_* := \sup_{x,y \in \mathbb{R}^d} \inf_{\gamma \in \Gamma_{x,y}} \left\{ \sup_t \{U(\gamma(t)) \wedge c\} - U(x) \wedge c - U(y) \wedge c + \inf U \right\}.$$

- One way to understand c_*: pretend that we are minimizing $U \wedge c$ instead
- We can show that the following two statements hold:
  - $c_* \leq E_*$
  - $c_* \leq c - \inf U$

Change the target function from U to ϵH_ϵ

- Recall $\mu^f_{\epsilon_t}$:

$$\mu^f_{\epsilon_t}(x) \propto \frac{1}{f((U(x)-c)_+)+\epsilon_t} \exp\left(-\int_{\inf U}^{U(x)} \frac{1}{f((u-c)_+)+\epsilon_t}\, du\right)$$

Change the target function from U to ϵH_ϵ

- Recall $\mu^f_{\epsilon_t}$:

$$\mu^f_{\epsilon_t}(x) \propto \frac{1}{f((U(x)-c)_+)+\epsilon_t} \exp\left(-\int_{\inf U}^{U(x)} \frac{1}{f((u-c)_+)+\epsilon_t}\, du\right)$$

- Let's define H_{ϵ_t}:

$$H_\epsilon(x) := \int_{\inf U}^{U(x)} \frac{1}{f((u-c)_+)+\epsilon}\, du + \ln\left(f((U(x)-c)_+)+\epsilon\right).$$

so that

$$\mu^f_{\epsilon_t}(x) \propto e^{-H_{\epsilon_t}(x)}.$$

Change the target function from U to ϵH_ϵ

- Recall $\mu^f_{\epsilon_t}$:

$$\mu^f_{\epsilon_t}(x) \propto \frac{1}{f((U(x)-c)_+)+\epsilon_t} \exp\left(-\int_{\inf U}^{U(x)} \frac{1}{f((u-c)_+)+\epsilon_t}\, du\right)$$

- Let's define H_{ϵ_t}:

$$H_\epsilon(x) := \int_{\inf U}^{U(x)} \frac{1}{f((u-c)_+)+\epsilon}\, du + \ln\left(f((U(x)-c)_+)+\epsilon\right).$$

so that

$$\mu^f_{\epsilon_t}(x) \propto e^{-H_{\epsilon_t}(x)}.$$

- In SA,

$$\mu^0_{\epsilon_t}(x) \propto e^{-(1/\epsilon_t)U(x)}.$$

We can understand as if the optimization landscape is modified from $(1/\epsilon_t)U(x)$ to $H_{\epsilon_t}(x)$.

Change the target function from U to ϵH_ϵ

- Recall $\mu^f_{\epsilon_t}$:

$$\mu^f_{\epsilon_t}(x) \propto \frac{1}{f((U(x)-c)_+)+\epsilon_t} \exp\left(-\int_{\inf U}^{U(x)} \frac{1}{f((u-c)_+)+\epsilon_t}\, du\right)$$

- Let's define H_{ϵ_t}:

$$H_\epsilon(x) := \int_{\inf U}^{U(x)} \frac{1}{f((u-c)_+)+\epsilon}\, du + \ln\left(f((U(x)-c)_+)+\epsilon\right).$$

so that

$$\mu^f_{\epsilon_t}(x) \propto e^{-H_{\epsilon_t}(x)}.$$

- In SA,

$$\mu^0_{\epsilon_t}(x) \propto e^{-(1/\epsilon_t)U(x)}.$$

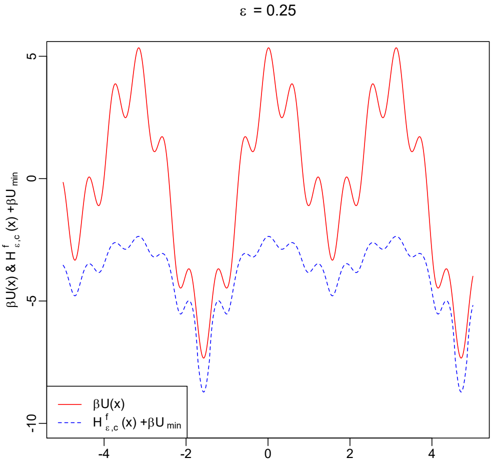
We can understand as if the optimization landscape is modified from $(1/\epsilon_t)U(x)$ to $H_{\epsilon_t}(x)$.
- The idea of state-dependent noise is embedded in the modified optimization landscape.

Idea of IKSA: landscape modification

- Consider the function

$$U_0(x) = \cos(2x) + \frac{1}{2}\sin(x) + \frac{1}{3}\sin(10x).$$

We take $\epsilon = 0.25$, $c = -1.5$ and $f = \arctan$.



Landscape modification in the wild

Image source: `https://kdlandscapingandsnowplowingbuffalo.com/renovation-landscape-modification/`

Improved kinetic simulated annealing (IKSA)

Definition (Improved kinetic Langevin)

The SDE of improved kinetic Langevin is given by

$$\begin{aligned} dX_t &= Y_t\,dt, \\ dY_t &= -\frac{1}{\epsilon_t} Y_t\,dt - \epsilon_t \nabla H_{\epsilon_t}(X_t)\,dt + \sqrt{2}\,dB_t. \end{aligned}$$

- The instantaneous stationary distribution at time t is the product distribution of $\mu^f_{\epsilon_t}$ and a Gaussian distribution with mean 0 and variance ϵ_t:

$$\pi^f_{\epsilon_t}(x,y) \propto \mu^f_{\epsilon_t}(x) e^{-\frac{\|y\|^2}{2\epsilon_t}} \propto e^{-H_{\epsilon_t}(x)} e^{-\frac{\|y\|^2}{2\epsilon_t}}.$$

Improved kinetic simulated annealing (IKSA)

Definition (Improved kinetic Langevin)

The SDE of improved kinetic Langevin is given by

$$
\begin{aligned}
dX_t &= Y_t \, dt, \\
dY_t &= -\frac{1}{\epsilon_t} Y_t \, dt - \epsilon_t \nabla H_{\epsilon_t}(X_t) \, dt + \sqrt{2} \, dB_t.
\end{aligned}
$$

- The instantaneous stationary distribution at time t is the product distribution of $\mu^f_{\epsilon_t}$ and a Gaussian distribution with mean 0 and variance ϵ_t:

$$\pi^f_{\epsilon_t}(x, y) \propto \mu^f_{\epsilon_t}(x) e^{-\frac{\|y\|^2}{2\epsilon_t}} \propto e^{-H_{\epsilon_t}(x)} e^{-\frac{\|y\|^2}{2\epsilon_t}}.$$

- If $f = 0$, then $\nabla U(X_t) = \epsilon_t \nabla H_{\epsilon_t}(X_t)$, which reduces to the classical kinetic Langevin.

Spin systems

- We would like to apply landscape modification to the Metropolis-Hastings algorithm in the context of spin systems.

Spin systems

- We would like to apply landscape modification to the Metropolis-Hastings algorithm in the context of spin systems.
- In many examples of spin systems of interest, the global minimum value $\min U$ is known explicitly. This piece of information can be utilized in the tuning of c in landscape modification, leading to accelerated samplers or optimizers.

MH chain on the original landscape

- **State space**: $\Sigma_N := \{-1, +1\}^N$, $N \in \mathbb{N}$.

MH chain on the original landscape

- **State space**: $\Sigma_N := \{-1, +1\}^N$, $N \in \mathbb{N}$.
- **Goal**: sample from $\pi^0_\beta \propto \exp\{-\beta U\}$ in the low-temperature regime (i.e. the inverse temperature β is large), where U is the target Hamiltonian function specified by the spin system of interest.

MH chain on the original landscape

- **State space**: $\Sigma_N := \{-1, +1\}^N$, $N \in \mathbb{N}$.
- **Goal**: sample from $\pi^0_\beta \propto \exp\{-\beta U\}$ in the low-temperature regime (i.e. the inverse temperature β is large), where U is the target Hamiltonian function specified by the spin system of interest.
- **Algorithm**: MH algorithm with target distribution π^0_β and base chain being the simple random walk proposal on Σ_N with transition matrix $P^{SRW} = (P^{SRW}(\eta, \sigma))_{\eta,\sigma \in \Sigma_N}$ given by

$$P^{SRW}(\eta, \sigma) := \frac{1}{N} \mathbf{1}_{\{\text{there exists } i \text{ such that } \eta(i) = -\sigma(i) \text{ and } \eta(j) = \sigma(j) \text{ for all } j \neq i\}}.$$

MH chain on the original landscape

- **State space**: $\Sigma_N := \{-1, +1\}^N$, $N \in \mathbb{N}$.
- **Goal**: sample from $\pi^0_\beta \propto \exp\{-\beta U\}$ in the low-temperature regime (i.e. the inverse temperature β is large), where U is the target Hamiltonian function specified by the spin system of interest.
- **Algorithm**: MH algorithm with target distribution π^0_β and base chain being the simple random walk proposal on Σ_N with transition matrix $P^{SRW} = (P^{SRW}(\eta, \sigma))_{\eta, \sigma \in \Sigma_N}$ given by

$$P^{SRW}(\eta, \sigma) := \frac{1}{N} \mathbf{1}_{\{\text{there exists } i \text{ such that } \eta(i) = -\sigma(i) \text{ and } \eta(j) = \sigma(j) \text{ for all } j \neq i\}}.$$

- This is the baseline algorithm that we will be comparing with.

MH chain on the modified landscape

- Consider the following modified Hamiltonian:

$$\mathcal{U}^f_{\alpha,c,1/\beta}(\sigma) = \int_{\min U}^{U(\sigma)} \frac{1}{\alpha f((u-c)_+) + 1/\beta}\, du.$$

MH chain on the modified landscape

- Consider the following modified Hamiltonian:

$$\mathcal{U}^f_{\alpha,c,1/\beta}(\sigma) = \int_{\min U}^{U(\sigma)} \frac{1}{\alpha f((u-c)_+) + 1/\beta}\, du.$$

- For this talk we are interested in taking $f(x) = x^2$ and $\alpha = \beta$, which gives

$$\mathcal{U}^f_{\beta,c,1/\beta}(\sigma) = \beta(U(\sigma) \wedge c - \min U) + \arctan(\beta(U(\sigma) - c)_+).$$

MH chain on the modified landscape

- Consider the following modified Hamiltonian:

$$\mathcal{U}^f_{\alpha,c,1/\beta}(\sigma) = \int_{\min U}^{U(\sigma)} \frac{1}{\alpha f((u-c)_+) + 1/\beta}\, du.$$

- For this talk we are interested in taking $f(x) = x^2$ and $\alpha = \beta$, which gives

$$\mathcal{U}^f_{\beta,c,1/\beta}(\sigma) = \beta(U(\sigma) \wedge c - \min U) + \arctan(\beta(U(\sigma) - c)_+).$$

- The modified landscape exhibits a balance between exploration and exploitation: the landscape is flattened above c to encourage exploration, while the original landscape is utilized below c to encourage exploitation.

MH chain on the modified landscape

- **Algorithm**: MH algorithm with target distribution

$$\pi^f_{\beta,c}(\sigma) \propto e^{-\mathcal{U}^f_{\beta,c,1/\beta}(\sigma)}.$$

and base chain being the simple random walk proposal on Σ_N with transition matrix P^{SRW}.

MH chain on the modified landscape

- **Algorithm**: MH algorithm with target distribution

$$\pi^f_{\beta,c}(\sigma) \propto e^{-\mathcal{U}^f_{\beta,c,1/\beta}(\sigma)}.$$

  and base chain being the simple random walk proposal on Σ_N with transition matrix P^{SRW}.
- **Intuition**: in the low-temperature regime, the bias between the original target π^0_β and $\pi^f_{\beta,c}$ is small. The MH chain on the modified landscape mixes "fast", while the MH chain on the original landscape mixes "slowly" due to the landscape.

Mixing time parameters

To quantify the time it takes for the chain to mix, we introduce the following parameters:

- (Total variation mixing time to π_β^0 by $X^f_{\alpha,c,1/\beta}$ (resp. X_β^0) on the modified (resp. original) landscape)

$$t^f_{mix}(\varepsilon) := \inf\left\{t \geq 0;\ \sup_{\sigma \in \Sigma_N} \left\|(P^f_{\alpha,c,1/\beta})^t(\sigma,\cdot) - \pi^0_\beta\right\|_{TV} \leq \varepsilon\right\}.$$

$$t^0_{mix}(\varepsilon) := \inf\left\{t \geq 0;\ \sup_{\sigma \in \Sigma_N} \left\|(P^0_\beta)^t(\sigma,\cdot) - \pi^0_\beta\right\|_{TV} \leq \varepsilon\right\}.$$

Mixing time parameters

To quantify the time it takes for the chain to mix, we introduce the following parameters:

- (Total variation mixing time to π^0_β by $X^f_{\alpha,c,1/\beta}$ (resp. X^0_β) on the modified (resp. original) landscape)

$$t^f_{mix}(\varepsilon) := \inf\left\{t \geq 0;\ \sup_{\sigma \in \Sigma_N} \left\|(P^f_{\alpha,c,1/\beta})^t(\sigma,\cdot) - \pi^0_\beta\right\|_{TV} \leq \varepsilon\right\}.$$

$$t^0_{mix}(\varepsilon) := \inf\left\{t \geq 0;\ \sup_{\sigma \in \Sigma_N} \left\|(P^0_\beta)^t(\sigma,\cdot) - \pi^0_\beta\right\|_{TV} \leq \varepsilon\right\}.$$

- (First time reaching $\min U$ with high probability)

$$\mathcal{T}^f(\varepsilon) := \inf\left\{t \geq 0;\ \inf_{\sigma \in \Sigma_N} \mathbb{P}_\sigma(U(X^f_{\alpha,c,1/\beta}(t)) = \min U) \geq 1 - \varepsilon\right\},$$

$$\mathcal{T}^0(\varepsilon) := \inf\left\{t \geq 0;\ \inf_{\sigma \in \Sigma_N} \mathbb{P}_\sigma(U(X^0_\beta(t)) = \min U) \geq 1 - \varepsilon\right\}.$$

Main results: from torpid to rapid mixing

Theorem

For low enough temperature, the following holds:

- *(Torpid total variation mixing time with exponential dependence on N using X_β^0)*

$$t^0_{mix}(\varepsilon) = \Omega\left(\frac{4^N}{\varepsilon}\ln\left(\frac{1}{2\varepsilon}\right)\right).$$

- *(Rapid total variation mixing time with polynomial dependence on N and β using $X^f_{\beta,c,1/\beta}$)*

$$t^f_{mix}(\varepsilon) = \mathcal{O}\left(N^3\left(\ln\left(\frac{2}{\varepsilon}\right) + \beta(c - \min U) + \frac{\pi}{2} + N\ln 2\right)\right).$$

Main results: from torpid to rapid mixing

Theorem

For low enough temperature, the following holds:

- *(X_β^0 takes at least exponential in N time to reach $\min U$)*

$$\mathcal{T}^0(\varepsilon) = \Omega\left(\frac{2^N}{\varepsilon}\right).$$

- *($X^f_{\beta,c,1/\beta}$ reaches $\min U$ in polynomial in N time with high probability)*

$$\mathcal{T}^f(\varepsilon) = \mathcal{O}\left(N^3\left(\ln\left(\frac{2}{\varepsilon}\right) + \beta(c - \min U) + \frac{\pi}{2} + N\ln 2\right)\right).$$

Application: Ising model on the complete graph

- Let $G_N = (V_N, E_N)$ be a graph with $V_N = [\![N]\!]$. For $\sigma \in \Sigma_N$, we consider the Ising model on the graph G_N where the Hamiltonian function is given by

$$U(\sigma) = -\frac{J}{2} \sum_{(v,w)\in E_N} \sigma_v \sigma_w - \frac{h}{2} \sum_{v \in [\![N]\!]} \sigma_v,$$

  where $J > 0$ is the pairwise interaction constant and $h > 0$ is the external magnetic field. In particular, in this subsection we focus on the complete graph $G_N = K_N$.

Application: Ising model on the complete graph

- Let $G_N = (V_N, E_N)$ be a graph with $V_N = [\![N]\!]$. For $\sigma \in \Sigma_N$, we consider the Ising model on the graph G_N where the Hamiltonian function is given by

$$U(\sigma) = -\frac{J}{2} \sum_{(v,w) \in E_N} \sigma_v \sigma_w - \frac{h}{2} \sum_{v \in [\![N]\!]} \sigma_v,$$

  where $J > 0$ is the pairwise interaction constant and $h > 0$ is the external magnetic field. In particular, in this subsection we focus on the complete graph $G_N = K_N$.
- For this model, $\min U = U(+\mathbf{1})$.

Application: Ising model on the complete graph

Corollary

Suppose we set $c = U(+\mathbf{1}) + \delta$, where δ is chosen small enough

- $$t^0_{mix}(e^{-N}) = \Omega\left(e^{DN^3/\delta} N\right),$$

  while

  $$t^f_{mix}(e^{-N}) = \mathcal{O}\left(N^3\left(\ln\left(2e^N\right) + \beta\delta + \frac{\pi}{2} + N\ln 2\right)\right),$$

  where $D = D(J,h) > 0$ is a universal constant that depends on J, h.

- $$\mathcal{T}^f(e^{-N}) = \mathcal{O}\left(N^3\left(\ln\left(2e^N\right) + \beta\delta + \frac{\pi}{2} + N\ln 2\right)\right).$$

Application: Derrida's random energy model (REM)

- Let $(X_\sigma)_{\sigma \in \Sigma_N}$ be a family of i.i.d. standard normal random variables. At a spin configuration $\sigma \in \Sigma_N$, the value of the random Hamiltonian function at σ is

$$U(\sigma) = -\sqrt{N} X_\sigma.$$

Application: Derrida's random energy model (REM)

- Let $(X_\sigma)_{\sigma \in \Sigma_N}$ be a family of i.i.d. standard normal random variables. At a spin configuration $\sigma \in \Sigma_N$, the value of the random Hamiltonian function at σ is

$$U(\sigma) = -\sqrt{N} X_\sigma.$$

- It is known that the maximum of X_σ over $\sigma \in \Sigma_N$, when normalized by $\sqrt{N}$, converges in probability to $\sqrt{2 \ln 2}$, that is, for any $\epsilon > 0$ we have

$$\lim_{N \to \infty} \mathbb{P}\left(\left| \frac{1}{\sqrt{N}} \max_{\sigma \in \Sigma_N} X_\sigma - \sqrt{2 \ln 2} \right| > \epsilon \right) = 0.$$

Application: Derrida's random energy model (REM)

Corollary

Suppose we set $c = -N\sqrt{2\ln 2} + \frac{N^{1/4}}{4}$, $\delta = -N\sqrt{2\ln 2} + \frac{N^{1/4}}{4} - \min U$. *Note that w.h.p.* $\delta = \Omega(N^{1/4} - \ln N)$. *For large enough* N *and low enough temperature, w.h.p. the following holds:*

-
$$t^0_{mix}(e^{-N}) = \Omega\left(e^{\beta\left(N\sqrt{2\ln 2} - C_1\sqrt{N\ln N}\right) - (\ln 4)N} N\right),$$

  while

$$t^f_{mix}(e^{-N}) = \mathcal{O}\left(N^3\left(\ln\left(2e^N\right) + \beta\delta + \frac{\pi}{2} + N\ln 2\right)\right).$$

-
$$\mathcal{T}^f(e^{-N}) = \mathcal{O}\left(N^3\left(\ln\left(2e^N\right) + \beta\delta + \frac{\pi}{2} + N\ln 2\right)\right).$$

Concluding remarks

- This talks centers around a technique that we call landscape modification.

Concluding remarks

- This talks centers around a technique that we call landscape modification.
- This has successfully been applied to spin systems to yield rapidly mixing algorithms with a novel use of the global minimum value to adjust the landscape for acceleration, while the same algorithm on the original landscape mixes torpidly.

Concluding remarks

- This talks centers around a technique that we call landscape modification.
- This has successfully been applied to spin systems to yield rapidly mixing algorithms with a novel use of the global minimum value to adjust the landscape for acceleration, while the same algorithm on the original landscape mixes torpidly.
- The transformation is not only limited to this setup. In fact it is broadly applicable to any gradient-based or difference-based optimization or sampling algorithm.

Concluding remarks

- This talks centers around a technique that we call landscape modification.
- This has successfully been applied to spin systems to yield rapidly mixing algorithms with a novel use of the global minimum value to adjust the landscape for acceleration, while the same algorithm on the original landscape mixes torpidly.
- The transformation is not only limited to this setup. In fact it is broadly applicable to any gradient-based or difference-based optimization or sampling algorithm.
- There are also quite a few techniques that share the spirit of landscape modification that we are aware:
    - Olivier Catoni's energy transformation algorithm, which can be further traced back to the work of Robert Azencott
    - Preconditioning
    - Importance sampling
    - Quantum annealing

Catoni's energy transformation algorithm

Probab. Theory Relat. Fields 110, 69–89 (1998)

Probability Theory and Related Fields

The energy transformation method for the Metropolis algorithm compared with Simulated Annealing

Olivier Catoni

DIAM – Intelligence Artificielle et Mathématiques, Laboratoire de Mathématiques de l'Ecole Normale Supérieure, UA 762 du CNRS, 45, rue d'Ulm, F-75 005 Paris, France

Quantum annealing, MCMC and D-Wave

Image source: Wang et al. Statistical Science '16

Statistical Science
2016, Vol. 31, No. 3, 362–398
DOI: 10.1214/16-STS560

Quantum Annealing with Markov Chain Monte Carlo Simulations and D-Wave Quantum Computers

Yazhen Wang, Shang Wu and Jian Zou

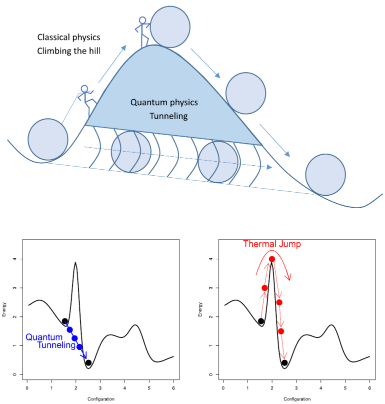


FIG. 1. *A cartoon illustration of quantum tunneling vs. thermal climbing on the top panel with annealing elucidations of quantum tunneling on the left bottom panel and thermal climbing on the right bottom panel.*

Ongoing work

- Landscape modification applied to Sequential Monte Carlo (SMC) (with Kengo Kamatani at ISM Tokyo)
- Finding maximum independent set in graphs
- Other NP-hard problems?

Thank you! Question(s)?