Accelerated simulated annealing with fast cooling

• Michael Choi (cuhk-sz)

Outline

1. Prelimaries
   1.1 Metropolis-Hastings (MH) algorithm M_1
   1.2 Basic properties
   1.3 Accelerated MH algorithm. M_2
   1.4 ~~Geometric~~ Comparison of M_1 and M_2
   1.5 Simulated annealing and its variant.
2. Main results
   2.1 Proofs

Throughout the talk, we only consider finite state space Markov chain.

.1 MH algorithm M_1.

• Given a target distribution π that we want to sample from, we would like to construct a Markov chain that converges to π from a known Markov chain with transition matrix Q,

(Example : π is the posterior distribution in Bayesian model)

- MH algorithm : 1 Propose a new state using Q, say y. Initial state $= x$

  2 (i) With probability $\alpha(x,y)$, we accept the move.

  (ii) Otherwise, we stay at the same state x.

  3 Repeat 1, 2.

ef^n 1 (MH algorithm M_1) : Given a target distribution π and proposal chain with generator $Q = (Q(x,y))$,

the MH algorithm is the Markov chain with generator M_1, where

$$M_1(x,y) := \begin{cases} \alpha(x,y)\,Q(x,y) = \cancel{\min\{1, \frac{\pi(y)}{\pi(x)}\}\,Q(x,y)} & x \neq y \\ \qquad = \min\{1, \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\}\,Q(x,y) & \\ -\sum_{y : y \neq x} M_1(x,y) & x = y \end{cases}$$

- Useful when we don't know how to compute the normalizing constant of π.

. 2 Basic properties

mma 1 : 1. M_1 is reversible with respect to π.

2. (Geometric interpretation of M_1) (Diaconis and Billera '01)

$$d_\pi(Q, M_1) = \inf_{R \in \mathcal{R}(\pi)} d_\pi(Q, R)$$

where $d_\pi(A,B) := \sum_x \sum_{y : x \neq y} \pi(x)\,|A(x,y) - B(x,y)|$

is the distance between two Markov generators A, B, (3)
and $\mathcal{R}(\pi)$ is the set of π-reversible generators.

1.3 Accelerated MH algorithm M_2

- Many variants of MH with improved convergence, e.g. lifting (Chen et al. '99), non-reversible MH (Bierkens '16), ...

- Today we will focus on a variant that we call M_2 (Choi '18, Choi and Huang '18).

Def^n 2 (Accelerated MH M_2): Given π: proposal distribution
Q: generator of proposal chain,

$$M_2(x,y) := \begin{cases} \max\left\{1, \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right\} Q(x,y), & x \neq y \\ -\sum_{y: y\neq x} M_2(x,y), & x = y \end{cases}$$

1.4 Comparison of M_1 and M_2

Hilbert space $\ell^2(\pi)$ with inner product $\langle f, g\rangle_\pi := \sum_{x\in\mathcal{X}} f(x)g(x)\pi(x)$
for $f, g: \mathcal{X} \to \mathbb{R}$.

(Peskun ordering): Suppose that there are two Markov generators A, B which are reversible with respect to π. B is said to dominate A off-diagonally, written as

$$B \overset{\text{Peskun}}{\geq} A \quad , \text{ if } \quad B(x,y) \geq A(x,y) \quad \forall x \neq y.$$

Consequently, $\langle Bf, f \rangle_\pi \leq \langle Af, f \rangle_\pi$ and

$$\lambda_2(B) \geq \lambda_2(A) \, , \text{ where}$$

$$\lambda_2(B) = \inf_{\substack{\langle 1, f \rangle_\pi = 0 \\ \langle f, f \rangle_\pi \leq 1}} \langle -Bf, f \rangle_\pi$$ is the spectral gap of B

(or the second smallest eigenvalue of $-B$)

Lemma 2 (Comparison of M_1 and M_2):

1. M_2 is reversible w.r.t. π (equivalently M_2 is a self-adjoint operator in $\ell^2(\pi)$)

2. $M_1 \overset{\text{Peskun}}{\leq} M_2$, which implies $\forall f \in \ell^2(\pi)$,

$$\langle M_2 f, f \rangle_\pi \leq \langle M_1 f, f \rangle_\pi$$

$$\lambda_2(M_2) \geq \lambda_2(M_1)$$

3. $d_\pi(Q, M_1) = d_\pi(Q, M_2) = d_\pi(Q, \alpha M_1 + (1-\alpha) M_2)$ for $\alpha \in [0,1]$. In words, $\alpha M_1 + (1-\alpha) M_2$ is the "closest" reversible generator (w.r.t. π) to Q.

(The geometric interpretation: Gives a sense why M_1, M_2 are natural transformation to study)

<u>Proof</u>: 1. $\pi(x)M_2(x,y) = \max\{\pi(x)Q(x,y), \pi(y)Q(y,x)\}$

$$= \pi(y)M_2(y,x).$$

2. $$M_2(x,y) = \max\left\{1, \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right\}Q(x,y)$$
$$\geq \min\left\{1, \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right\}Q(x,y) = M_1(x,y)$$

3. Omitted.

1.5 Simulated annealing and its variant

- Simulated annealing: ~~a time~~ non-homogeneous MH.
- Introduce $T(t)$, the temperature at time t with $T(t) \downarrow 0$ as $t \to \infty$.
- Given a target function U to minimize, a μ-reversible proposal chain with generator Q, we take

$$\pi_{T(t)}(x) = \frac{e^{-\frac{U(x)}{T(t)}}\mu(x)}{Z_{T(t)}}$$ as the target distribution in MH.

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

Define the set of global minima:

$$U_{min} \triangleq \{x \; ; \; U(x) \leq U(y) \; \forall y\}.$$

$$m \triangleq \min_x U(x)$$

$$\lim_{t\to\infty} \pi_{T(t)}(x) = \begin{cases} \dfrac{\mu(x)}{\mu(U_{min})}, & x \in U_{min} \\ 0, & x \notin U_{min} \end{cases}$$

$$\pi_{T(t)}(x) = \frac{e^{-\left(\frac{U(x)-m}{T(t)}\right)} \mu(x)}{\mu(H) + \sum_{y \notin H} e^{-\left(\frac{U(y)-m}{T(t)}\right)}}$$

$$\rightarrow \begin{cases} \dfrac{\mu(x)}{\mu(H)}, & x \in U_{min} \\ 0, & x \notin U_{min} \end{cases}$$

Def^n 3 (Simulated annealing): U: target function

Q: proposal chain generator, reversible w.r.t. μ.

$T(t)$: temperature at time t

SA is a non-homogeneous CTMC with generator

$$M_{1,t} = Q(x,y)\min\left\{1, \frac{\pi_{T(t)}(y)\,Q(y,x)}{\pi_{T(t)}(x)\,Q(x,y)}\right\}$$

$$= Q(x,y)\min\left\{1, e^{\frac{U(x)-U(y)}{T(t)}}\right\} = Q(x,y)\, e^{-\frac{(U(y)-U(x))_+}{T(t)}}, \quad x \neq y$$

depends on time t.

As $t \to \infty$, $T(t) \downarrow 0$ "slowly" such that the Markov chain with generator $M_{1,t}$ converges to

$$\pi_\infty := \lim_{t\to\infty} \pi_{T(t)}$$

How slow? (Cannot be too slow in practice, it takes too long to converge).

A path from x to y = any sequence of points starting from $x_0 = x, x_1, x_2, \dots, x_n = y$ such that $Q(x_{i-1}, x_i) > 0$ for $i = 1, 2, \dots, n$.

$$\Gamma^{x,y} \triangleq \text{set of path from } x \text{ to } y$$

$$\text{Elev}(\gamma) \triangleq \text{highest elevation along a path } \gamma \in \Gamma^{x,y}$$

$$= \max\{U(\gamma_i);\ \gamma_i \in \gamma\}.$$

$$H(x,y) \triangleq \min\{\text{Elev}(\gamma);\ \gamma \in \Gamma^{x,y}\}.$$

$$C_{M_1} = C_{M_1}(Q,U) \triangleq \max_{x,y}\{H(x,y) - U(x) - U(y)\}.$$

Convergence guarantee of SA (Holley and Stroock '88).

Thm 1 For any $\varepsilon > 0$, if $T(t) = \frac{C_{M_1} + \varepsilon}{\ln(t+1)}$ (logarithmic cooling), then SA is strongly ergodic and converges to π_∞.

(Hajek '88): SA is strong ergodic iff $T(t) = \frac{C_{M_1}}{\ln(t+1)}$

That is $\| P_t^{M_2}(x, \cdot) - \pi_\infty \|_{TV} \to 0$ as $t \to \infty$ (8)

for any x.

M_2 variant of simulated annealing :

Def^n 4 :

$$M_{2,t}(x,y) = Q(x,y) \max\left\{1, e^{\frac{U(x)-U(y)}{T(t)}}\right\}$$

$$= Q(x,y) e^{\frac{(U(x)-U(y))_+}{T(t)}}, \quad x \neq y.$$

Lemma 3 (extended Lemma 2) :

1. $M_{1,t}$ and $M_{2,t}$ are reversible w.r.t. the Gibbs distribution $\pi_{T(t)}$.

2. $M_{2,t} \overset{\text{Peskun}}{\geq} M_{1,t}$

(i). $\langle M_{2,t} f, f \rangle_{\pi_{T(t)}} \leq \langle M_{1,t} f, f \rangle_{\pi_{T(t)}}$

(ii). $\lambda_2(M_{2,t}) \geq \lambda_2(M_{1,t})$.

3. $$\langle -M_{2,t} f, f \rangle_{\pi_{T(t)}} = \frac{1}{2Z_{T(t)}} \sum_{x,y} (f(y)-f(x))^2 e^{-\frac{\min\{U(x),U(y)\}}{T(t)}} Q(x,y)\mu(x)$$

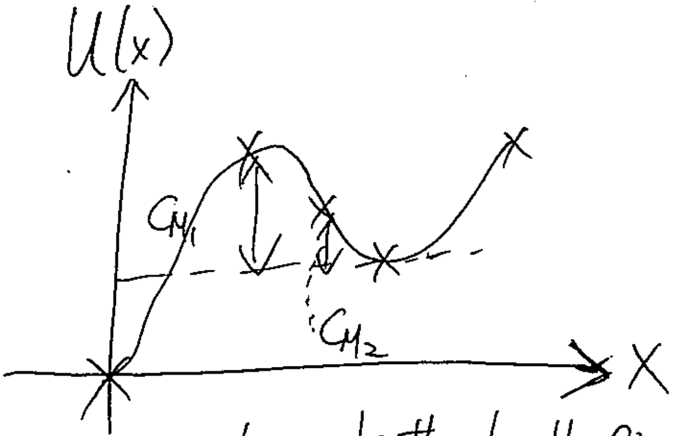
$$\langle -M_{1,t} f, f \rangle_{\pi_{T(t)}} = \frac{1}{2Z_{T(t)}} \sum_{x,y} (f(y)-f(x))^2 e^{-\frac{\max\{U(x),U(y)\}}{T(t)}} Q(x,y)\mu(x)$$

2. Main results

$$C_{M_1} = \max_{x,y} \left\{ H(x,y) - U(x) - U(y) \right\}$$

$$C_{M_2} = C_{M_2}(Q, U) \triangleq \max_{x,y} \left\{ \max_{\substack{z,w \in \gamma^{x,y},\ z=\gamma_i^{x,y},\ w=\gamma_{i+1}^{x,y} \\ \mathrm{Elev}(\gamma^{x,y}) = H(x,y)}} U(z) \wedge U(w) - U(x) - U(y) \right\}$$

Lemma 4: 1. $C_{M_1} \geq C_{M_2}$. In particular, when U does not have repeated values, $C_{M_1} > C_{M_2}$.



under birth-death Q.

- C_{M_1}: largest hill to climb from a local minimum to global minimum.
- $C_{M_2} \approx$ second largest hill to climb from a local minimum to global minimum.

- $C_{M_1} \geq 0$ while C_{M_2} can be negative.

Thm 2 (Convergence guarantee of $M_{2,t}$ when $C_{M_2} > 0$)

When $T(t) = \frac{C_{M_2} + \varepsilon}{\ln(t+1)}$, the non-homogeneous CTMC with generator $M_{2,t}$ is strongly ergodic and converges to π_∞, i.e. $\| P_t^{M_2}(x,\cdot) - \pi_\infty \|_{TV} \to 0$ as $t \to \infty$.

Thm 3 (Convergence guarantee of $M_{2,t}$ when $C_{M_2} \leq 0 < C_{M_1}$)

When $- \lim_{t\to\infty} \left(\frac{d}{dt} T(t)\right) \frac{e^{\frac{C_{M_2}}{T(t)}}}{T(t)^2} = 0$, then the CTMC with generator $M_{2,t}$ is strongly ergodic and converges to π_∞. Examples of $T(t)$ are

(i). $T(t) = (t+1)^{-\alpha}$, $\alpha \in (0,1)$.

2.1 Proofs

Lemma 5: For ~~all~~ any $t > 0$,

$$\lambda_2(M_{2,t}) \geq A e^{-\frac{C_{M_2}}{T(t)}}$$

where A is ~~some~~ a positive constant.

Lemma 6: (Gidas 85) If

(1). $\int_0^\infty \lambda_2(M_{2,t})\,dt = \infty$.

(2). $\lim_{t\to\infty} \frac{\beta(t)}{\lambda_2(M_{2,t})} = 0$

where $\left| \frac{d}{dt} \pi_{T(t)}(x) \right| \leq \beta(t)\, \pi_{T(t)}(x)$,

$$\beta(t) \stackrel{\Delta}{=} -\left(\frac{d}{dt}T(t)\right)\frac{1}{T(t)^2}\left(\max_x U(x) - \min_y U(y)\right)$$

then the CTMC with generator $M_{2,t}$ is strongly ergodic.

Assume that we have Lemma 5 and Lemma 6, then we can prove Theorem 2 and Theorem 3.

Theorem 3: When $C_{M_2} \leq 0$, $\lambda_2(M_{2,t}) \geq A$

so (1) in Lemma 6 is satisfied.

(2) is just ☆.

Theorem 2: $T(t) = \frac{C_{M_2}+\varepsilon}{\ln(t+1)}$

(1): $\int_0^\infty \lambda_2(M_{2,t})\,dt \geq \int_0^\infty A e^{-\frac{C_{M_2}}{T(t)}}\,dt$

$$= A\int_0^\infty (t+1)^{-\frac{C_{M_2}}{C_{M_2}+\varepsilon}}\,dt$$

$$\geq A\int_0^\infty \frac{1}{t+1}\,dt = \infty$$

(2):

$$\lim_{t\to\infty} \frac{\beta(t)}{\lambda_2(M_{2,t})} \leq \frac{A\left(\max_x U(x) - \min_x U(x)\right)}{C_{M_2}+\varepsilon} \lim_{t\to\infty} \frac{1}{(t+1)^{\frac{\varepsilon}{C_{M_2}+\varepsilon}}}$$

$$= 0$$

We will now prove Lemma 5.

Proof of Lemma 5: WLOG we assume $\min_x U(x) = 0$. We will prove that $\forall f \in \ell^2(\pi_{T(t)})$

$$\frac{\langle -M_{2,t} f, f\rangle_{\pi_{T(t)}}}{\langle f, f\rangle_{\pi_{T(t)}}} \geq A e^{-\frac{C_{M_2}}{T(t)}}$$

For $x, y \in \mathcal{X}$, we pick $\gamma^{x,y}$ such that $\mathrm{Elev}(\gamma^{x,y}) = H(x,y)$.

Let $n(x,y)$ be the length of the path $\gamma^{x,y}$ and

$$N \triangleq \max_{x,y} n(x,y).$$

Denote the indicator function $\chi_{z,w}(x,y)$ to be

$$\chi_{z,w}(x,y) = \begin{cases} 1, & \text{for some } 0 \leq i < n(x,y),\ \gamma_i^{x,y} = z,\ \gamma_{i+1}^{x,y} = w. \\ 0, & \text{otherwise.} \end{cases}$$

$$2\langle f, f\rangle_{\pi_{T(t)}} = \sum_{x,y}\big(f(y)-f(x)\big)^2 \pi_{T(t)}(x)\,\pi_{T(t)}(y)$$

$$= \sum_{x,y}\left(\sum_{i=1}^{n(x,y)} f(\gamma_i^{x,y}) - f(\gamma_{i-1}^{x,y})\right)^2 \pi_{T(t)}(x)\,\pi_{T(t)}(y)$$

$$\leq \sum_{x,y} n(x,y) \sum_{i=1}^{n(x,y)} \left(f(\gamma_i^{x,y}) - f(\gamma_{i-1}^{x,y})\right)^2 \pi_{T(t)}(x)\,\pi_{T(t)}(y)$$

$$\leq N \sum_{x,y}\sum_{w,z} \chi_{z,w}(x,y)\big(f(z)-f(w)\big)^2 \frac{\mu(z)Q(z,w)}{Z_{T(t)}} e^{-\frac{U(z)\wedge U(w)}{T(t)}} \frac{\pi_{T(t)}(x)\pi_{T(t)}(y)Z}{\mu(z)Q(z,w)\, e^{-\frac{U(z)\wedge U(w)}{T(t)}}}$$

$$\leq N \left(\max_{z,w}\left(\sum_{x,y} \chi_{z,w}(x,y) \frac{\pi_{T(t)}(x)\,\pi_{T(t)}(y)\,Z_{T(t)}}{\mu(z)\,Q(z,w)\,e^{-\frac{U(z)\wedge U(w)}{T(t)}}}\right)\right) \times \underbrace{\sum_{z,w}\big(f(z)-f(w)\big)^2 \frac{\mu(z)Q(z,w)}{Z_{T(t)}} e^{-\frac{U(z)\wedge U(w)}{T(t)}}}_{2\langle -M_{2,t}f,\, f\rangle_{\pi_{T(t)}}}$$

$$\chi_{z,w}(x,y)\,\frac{\pi_{T(t)}(x)\,\pi_{T(t)}(y)\,Z_{T(t)}}{\mu(z)\,Q(z,w)\,e^{-\frac{U(z)\wedge U(w)}{T(t)}}}$$

$$= \frac{\chi_{z,w}(x,y)}{\mu(z)\,Q(z,w)}\;\underbrace{\frac{\mu(x)\,\mu(y)}{Z_{T(t)}}}_{\le \frac{\mu(x)\mu(y)}{\mu(U_{min})}}\;\underbrace{e^{\frac{U(z)\wedge U(w)-U(x)-U(y)}{T(t)}}}_{\le e^{\frac{C_{M_2}}{T(t)}}}$$

So we can take

$$A^{-1} = N\left(\max_{z,w}\sum_{x,y}\frac{\chi_{z,w}(x,y)}{\mu(z)\,Q(z,w)}\,\frac{\mu(x)\cdot\mu(y)}{\mu(U_{min})}\right)$$

□.