# Accelerated simulated annealing with fast cooling

Michael Choi

The Chinese University of Hong Kong, Shenzhen
Institute for Data and Decision Analytics (iDDA)

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

数据运筹科学研究院
Institute for Data and Decision Analytics

July 2019

# Introduction

- Simulated annealing is a stochastic optimization algorithm that has found huge success in statistics and image processing.

# Introduction

- Simulated annealing is a stochastic optimization algorithm that has found huge success in statistics and image processing.
- As we shall see, it involves studying the convergence of a non-homogeneous Markov chain/process.

# Introduction

- Simulated annealing is a stochastic optimization algorithm that has found huge success in statistics and image processing.
- As we shall see, it involves studying the convergence of a non-homogeneous Markov chain/process.
- Our focus today is an accelerated version of simulated annealing proposed by Choi.
- Reference: "An accelerated variant of simulated annealing that converges under fast cooling" arXiv:1901.10269

# Introduction

- Let $\pi$ be a discrete or continuous distribution.
  **Goal**: Sample from $\pi$ 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 $\pi$ be a discrete or continuous distribution.
  **Goal**: Sample from $\pi$ 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 $\pi$ 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 $\pi$ be a discrete or continuous distribution.
  **Goal**: Sample from $\pi$ 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 $\pi$ 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 $\pi$, which only depends on the ratio
  $$\frac{\pi(y)}{\pi(x)}.$$
  Thus there is no need to know $Z$.

# The Metropolis-Hastings algorithm

- In our talk today, we will focus on **continuous-time** Metropolis-Hastings algorithm.
- Two ingredients:
  **(i). Target distribution**: $\pi$
  **(ii). Proposal chain** with generator $Q = (Q(x, y))_{x,y}$.

# The Metropolis-Hastings algorithm

**Algorithm 1:** The Metropolis-Hastings algorithm

**Input:** Proposal chain $Q$, target distribution $\pi$

1 (Generate the proposal): Given $X_t$, propose the next jump $Y_{t+s} \sim Q(X_t, \cdot)$ according to Q, say at time $t+s$

2 (Acceptance-rejection): Take

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

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 $\pi$, is a Markov chain $X = (X_t)_{t \geq 0}$ with generator

$$M_1(x,y) = \begin{cases} \alpha(x,y)Q(x,y), & \text{for } x \neq y, \\ -\sum_{y;\ y \neq x} M_1(x,y), & \text{for } x = y. \end{cases}$$

# The Metropolis-Hastings (MH) algorithm

**Theorem**

*Given target distribution $\pi$ and proposal chain $Q$, the Metropolis-Hastings chain is*

- ***reversible*** *with respect to $\pi$, that is, for all $x, y$,*

$$\pi(x)M_1(x,y) = \pi(y)M_1(y,x).$$

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

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t f(X_s)\, ds = \pi(f).$$

# Simulated annealing

- **Goal**: Find the global minimizers of a target function $U$.

# Simulated annealing

- **Goal**: Find the global minimizers of a target function $U$.
- **Idea of simulated annealing**: Construct a **non-homogeneous** Metropolis-Hastings Markov chain that converges to $\pi_\infty$, which is supported on the set of global minima of $U$.

# Simulated annealing

- **Goal**: Find the global minimizers of a target function $U$.
- **Idea of simulated annealing**: Construct a **non-homogeneous** Metropolis-Hastings Markov chain that converges to $\pi_\infty$, 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 $U_{min}$:

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

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

# Simulated annealing

**Algorithm 2:** Simulated annealing

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

1 (Generate the proposal): Given $X_t$, propose the next jump $Y_{t+s} \sim Q(X_t, \cdot)$ according to Q, say at time $t+s$

2 (Acceptance-rejection): Take

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

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.

# Simulated annealing

## Definition

Simulated annealing, with proposal chain $Q$, target distribution $\pi_{T(t)}$ and temperature schedule $T(t)$, is a non-homogeneous Markov chain with generator at time $t$ to be

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

# 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 $\pi_\infty$ in total variation distance if and only if for any $\epsilon > 0$,*

$$T(t) = \frac{c_{M_1} + \epsilon}{\ln(t+1)},$$

*where $c_{M_1}$ is known as the optimal hill-climbing constant that depends on the target function $U$ and proposal chain $Q$.*

# What is $c_{M_1}$?

- $c_{M_1}$ is the highest hill one need to climb from a local minimum to a global minimum.
- **A path $\gamma$ from $x$ to $y$**: any sequence of points starting from $x_0 = x, x_1, x_2, \ldots, x_n = y$ such that $Q(x_{i-1}, x_i) > 0$ for $i = 1, 2, \ldots, n$.

# What is $c_{M_1}$?

- $c_{M_1}$ is the highest hill one need to climb from a local minimum to a global minimum.
- **A path $\gamma$ from $x$ to $y$**: any sequence of points starting from $x_0 = x, x_1, x_2, \ldots, x_n = y$ such that $Q(x_{i-1}, x_i) > 0$ for $i = 1, 2, \ldots, n$.
- $\Gamma^{x,y}$ := set of paths from $x$ to $y$.

# What is $c_{M_1}$?

- $c_{M_1}$ is the highest hill one need to climb from a local minimum to a global minimum.
- **A path $\gamma$ from $x$ to $y$**: any sequence of points starting from $x_0 = x, x_1, x_2, \ldots, x_n = y$ such that $Q(x_{i-1}, x_i) > 0$ for $i = 1, 2, \ldots, n$.
- $\Gamma^{x,y}$ := set of paths from $x$ to $y$.
- $\text{Elev}(\gamma)$ := highest elevation along a path $\gamma \in \Gamma^{x,y} = \max\{U(\gamma_i); \gamma_i \in \gamma\}$

# What is $c_{M_1}$?

- $c_{M_1}$ is the highest hill one need to climb from a local minimum to a global minimum.
- **A path $\gamma$ from $x$ to $y$**: any sequence of points starting from $x_0 = x, x_1, x_2, \ldots, x_n = y$ such that $Q(x_{i-1}, x_i) > 0$ for $i = 1, 2, \ldots, n$.
- $\Gamma^{x,y}$ := set of paths from $x$ to $y$.
- $\text{Elev}(\gamma)$ := highest elevation along a path $\gamma \in \Gamma^{x,y} = \max\{U(\gamma_i); \gamma_i \in \gamma\}$
- $H(x, y) := \min\{\text{Elev}(\gamma);\ \gamma \in \Gamma^{x,y}\}$.

**Definition**

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

# Accelerated Metropolis-Hastings $M_2$

- There are many variants of Metropolis-Hastings with improved convergence, e.g. lifting (Chen et al. '99), non-reversible MH (Hwang et al. 93, Bierkens '16), ...

# Accelerated Metropolis-Hastings $M_2$

- There are many variants of Metropolis-Hastings with improved convergence, e.g. lifting (Chen et al. '99), non-reversible MH (Hwang et al. 93, Bierkens '16), ...
- Today we will focus on a variant that we call $M_2$ (Choi SPA '19+, Choi and Huang JTP '19+)

# Accelerated Metropolis-Hastings $M_2$

## Definition

With proposal chain $Q$ and target distribution $\pi$, the accelerated MH is a Markov chain with generator $M_2$ given by

$$M_2(x,y) = \begin{cases} \max\left\{ \dfrac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}, 1 \right\} Q(x,y), & \text{for } x \neq y, \\ -\sum_{y;\ y \neq x} M_2(x,y), & \text{for } x = y. \end{cases}$$

# Accelerated Metropolis-Hastings $M_2$

**Definition**

With proposal chain $Q$ and target distribution $\pi$, the accelerated MH is a Markov chain with generator $M_2$ given by

$$M_2(x,y) = \begin{cases} \max\left\{\dfrac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}, 1\right\}Q(x,y), & \text{for } x \neq y, \\ -\sum_{y;\ y\neq x} M_2(x,y), & \text{for } x = y. \end{cases}$$

- Recall that $M_1(x,y) = \min\left\{\dfrac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}, 1\right\}Q(x,y)$ for $x \neq y$.

# Accelerated Metropolis-Hastings $M_2$

We write $\langle f, g\rangle_\pi := \sum f(x)g(x)\pi(x)$ and $\lambda_2(M_i) := \inf_{\langle 1,f\rangle_\pi=0;\ \langle f,f\rangle_\pi\leq 1}\langle -M_i f, f\rangle_\pi$ is the spectral gap of $M_i$ for $i = 1, 2$.

**Theorem (Comparison between $M_1$ and $M_2$)**

*Given target distribution $\pi$ and proposal chain $Q$, $M_2$ is*

- ***reversible*** *with respect to $\pi$, that is, for all $x, y$,*

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

- $\langle M_2 f, f\rangle_\pi \leq \langle M_1 f, f\rangle_\pi$
- $\lambda_2(M_2) \geq \lambda_2(M_1)$.

# Accelerated Metropolis-Hastings $M_2$

We write $\langle f, g\rangle_\pi := \sum f(x)g(x)\pi(x)$ and $\lambda_2(M_i) := \inf_{\langle 1, f\rangle_\pi = 0;\ \langle f, f\rangle_\pi \leq 1} \langle -M_i f, f\rangle_\pi$ is the spectral gap of $M_i$ for $i = 1, 2$.

**Theorem (Comparison between $M_1$ and $M_2$)**

*Given target distribution $\pi$ and proposal chain $Q$, $M_2$ is*

- ***reversible*** *with respect to $\pi$, that is, for all $x, y$,*

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

- $\langle M_2 f, f\rangle_\pi \leq \langle M_1 f, f\rangle_\pi$
- $\lambda_2(M_2) \geq \lambda_2(M_1)$.

For more comparison results between $M_1$ and $M_2$, see Choi and Huang '19+.

# Accelerated simulated annealing

## Definition

Accelerated simulated annealing, with proposal chain $Q$, target distribution $\pi_{T(t)}$ (i.e. the Gibbs distribution) and temperature schedule $T(t)$, is a non-homogeneous Markov chain with generator at time $t$ to be

$$M_{2,t}(x,y) = \begin{cases} \max\left\{ \dfrac{\pi_{T(t)}(y)Q(y,x)}{\pi_{T(t)}(x)Q(x,y)}, 1 \right\} Q(x,y), & \text{for } x \neq y, \\ -\sum_{y;\ y \neq x} M_{2,t}(x,y), & \text{for } x = y. \end{cases}$$

# Accelerated simulated annealing

## Definition

Accelerated simulated annealing, with proposal chain $Q$, target distribution $\pi_{T(t)}$ (i.e. the Gibbs distribution) and temperature schedule $T(t)$, is a non-homogeneous Markov chain with generator at time $t$ to be

$$M_{2,t}(x,y) = \begin{cases} \max\left\{ \dfrac{\pi_{T(t)}(y)Q(y,x)}{\pi_{T(t)}(x)Q(x,y)}, 1 \right\} Q(x,y), & \text{for } x \neq y, \\ -\sum_{y;\ y \neq x} M_{2,t}(x,y), & \text{for } x = y. \end{cases}$$

- Recall the dynamics of classical simulated annealing:

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

# Main results

- The general message is that we can operate **faster** cooling schedule on $M_{2,t}$ than $M_{1,t}$!

# Main results

- The general message is that we can operate **faster** cooling schedule on $M_{2,t}$ than $M_{1,t}$!
- Similar to the classical case, the convergence behaviour of $M_{2,t}$ depends critically on a constant we call $c_{M_2}$.

# Main results

**Theorem (Choi '19)**

- ***Case 1:*** $c_{M_2} > 0$
  *The Markov chain generated by* $M_{2,t}$ *converges to* $\pi_\infty$ *in total variation distance if for any* $\epsilon > 0$,

$$T(t) = \frac{c_{M_2} + \epsilon}{\ln(t+1)}.$$

# Main results

## Theorem (Choi '19)

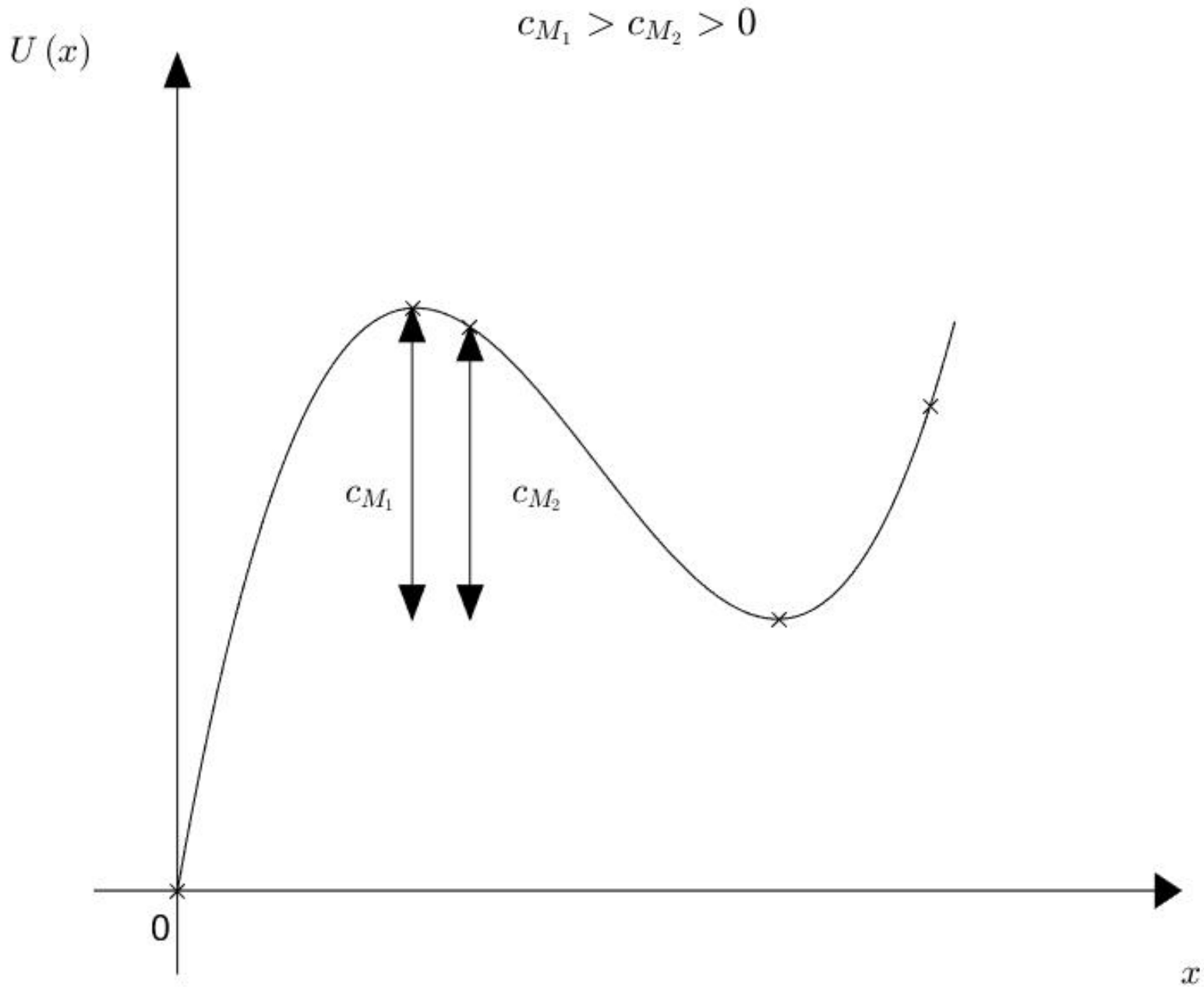
- ***Case 2:*** $c_{M_2} \leq 0$
  *The Markov chain generated by $M_{2,t}$ converges to $\pi_\infty$ in total variation distance if $T(t)$ satisfies*

$$\lim_{t\to\infty} \left(\frac{d}{dt}T(t)\right) \frac{e^{\frac{c_{M_2}}{T(t)}}}{T(t)^2} = 0.$$

  *Examples of fast cooling schedule that satisfy the above requirement are*

  1. *(power law cooling)* $T(t) = (t+1)^{-\alpha}$, *where* $\alpha \in (0,1)$.
  2. *(powers of logarithmic cooling)* $T(t) = (\log(t+1))^{-k}$, *where* $k > 1$.
  3. $T(t) = (t+1)^{-\alpha}\,(\log(t+1))^{-1}$, *where* $\alpha \in (0,1)$.

# What is $c_{M_2}$?



# What is $c_{M_2}$?

- $c_{M_1} := \max_{x,y \in \mathcal{X}} \{H(x,y) - U(x) - U(y)\}$ = largest hill to climb from a local minimum to a global minimum.

# What is $c_{M_2}$?

- $c_{M_1} := \max_{x,y \in \mathcal{X}} \{H(x,y) - U(x) - U(y)\}$ = largest hill to climb from a local minimum to a global minimum.
- $c_{M_2} := \max_{x,y \in \mathcal{X}} \left\{ \max_{\substack{z,w \in \gamma^{x,y},\ z=\gamma_i^{x,y}, w=\gamma_{i+1}^{x,y} \text{ for some } i \\ \text{Elev}(\gamma^{x,y}) = H(x,y)}} U(z) \wedge U(w) - U(x) - U(y) \right\} \approx$ second largest hill to climb from a local minimum to a global minimum
- $c_{M_1} \geq c_{M_2}$. When $U$ has distinct values, $c_{M_1} > c_{M_2}$.

# Main results

**Theorem ($X^{M_2}$ effectively escapes local minimum while $X^{M_1}$ may get trapped under fast cooling)**

*Suppose that $x$ is a local minimum of $U$ and under any cooling schedule,*

$$\mathbb{P}_x(X_t^{M_2} = x \ \forall t \geq 0) = 0.$$

*Under cooling schedule of the form*

$$T(t) = \frac{d}{\log(t+1)},$$

*where $d < c_{M_1}$, then*

$$\mathbb{P}_x(X_t^{M_1} = x \ \forall t \geq 0) > 0.$$

# Summary

- Propose an accelerated simulated annealing $M_{2,t}$
  - ✓ Convergence guarantee under **fast** cooling that depends on $c_{M_2}$. The optimal cooling schedule can be faster than logarithmic cooling depending on $Q$ and $U$.
  - ✓ With probability 1 it will escape local minimum even under fast cooling
  - × Relatively hard to simulate

# Summary

- Propose an accelerated simulated annealing $M_{2,t}$
  - ✓ Convergence guarantee under **fast** cooling that depends on $c_{M_2}$. The optimal cooling schedule can be faster than logarithmic cooling depending on $Q$ and $U$.
  - ✓ With probability 1 it will escape local minimum even under fast cooling
  - × Relatively hard to simulate
- Classical simulated annealing $M_{1,t}$
  - × Convergence guarantee under **slow** cooling that depends on $c_{M_1}$
  - × With positive probability it can get stuck in local minimum under fast cooling
  - ✓ Relatively easy to simulate

Thank you! Question(s)?