Polynomial Mixing Times of Simulated Tempering for Mixture Target Distributions

Presenter: Quan Zhou

Department of Statistics
Texas A&M University

Roadmap

- Background: why simulated tempering is useful for multimodal targets.
- Main results: polynomial mixing for both RWM and MALA and a near-optimal ladder.
- Proof architecture: construct proper auxiliary chains for comparison and then decompose s-conductance (or restricted spectral gap).
- Optimality and open problems: what the manuscript now settles and what remains open.

Background: MCMC Samplers and Mixture Targets

MCMC Sampling

Suppose we want to sample from a probability density on $\mathbb{R}^d$ of the form

$$\pi(x) \propto e^{-f(x)}, \qquad x \in \mathbb{R}^d.$$

- $x = (x_1, \ldots, x_d)$ is the state vector;
- d is the problem dimension;
- $f : \mathbb{R}^d \to \mathbb{R}$ is the potential function.

The standard Markov chain Monte Carlo (MCMC) strategy is:

1. at the current state $X_t = x$, propose a nearby point Y from a proposal kernel $Q(x, \cdot)$;
2. accept or reject Y so that the resulting chain has stationary density π.

Random-Walk Metropolis (RWM)

RWM uses $Q(x, \cdot) = N(x, 2hI_d)$; i.e., it proposes

$$Y = X_t + \sqrt{2h}\, Z, \qquad Z \sim N(0, I_d).$$

Because the proposal is symmetric, the acceptance probability is

$$\alpha_{\mathrm{RWM}}(x, y) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} = \min\left\{1, e^{-f(y)+f(x)}\right\}.$$

RWM uses only function values of f.

Metropolis-Adjusted Langevin Algorithm (MALA)

MALA uses first-order information through the gradient $\nabla f(x)$ and proposes

$$Y = X_t - h\nabla f(X_t) + \sqrt{2h}\, Z, \qquad Z \sim N(0, I_d),$$

so $Q(x, \cdot) = N\big(x - h\nabla f(x), 2hI_d\big)$.

Since the proposal is no longer symmetric, the acceptance probability becomes

$$\alpha_{\mathrm{MALA}}(x, y) = \min\left\{1, \frac{\pi(y)\, q(y, x)}{\pi(x)\, q(x, y)}\right\},$$

where $q(x, y)$ is the Gaussian proposal density.

- The drift term $-h\nabla f(x)$ pushes the proposal toward regions of higher density.
- For smooth log-concave targets, MALA often mixes faster than RWM.

RWM vs MALA

Strongly Log-Concave Targets

Consider a target distribution

$$\pi(x) \propto e^{-f(x)}, \qquad mI_d \preceq \nabla^2 f(x) \preceq LI_d,$$

for some constants $0 < m \leq L$. This means that f is both m-strongly convex and L-smooth.

- RWM and MALA mix in polynomial time as functions of dimension d and log-accuracy (in TV distance) $\log \epsilon^{-1}$ [Dwivedi et al., 2019, Chewi et al., 2021, Wu et al., 2022].
- The geometry is favorable: π only has one mode.
- What happens when the target is a *mixture* of such components?

Location Mixture of Log-concave Components

Our target of interest is

$$\pi^*(x) \propto \sum_{j=1}^{K} w_j e^{-f(x-\mu_j)}, \text{ where}$$

- K is the number of mixture components,
- $w_1, \ldots, w_K$ are positive weights with $\sum_{j=1}^{K} w_j = 1$,
- $\mu_1, \ldots, \mu_K \in \mathbb{R}^d$ are component locations,
- the same function f defines the shape of each component,
- and we still assume $mI_d \preceq \nabla^2 f(x) \preceq LI_d$.

The difficulty grows with the separation parameter

$$D = \max_{j \in [K]} \|\mu_j\|.$$

We will examine how the sampling complexity scales with d, D and ϵ.

Review on Simulated Tempering

Simulated Tempering

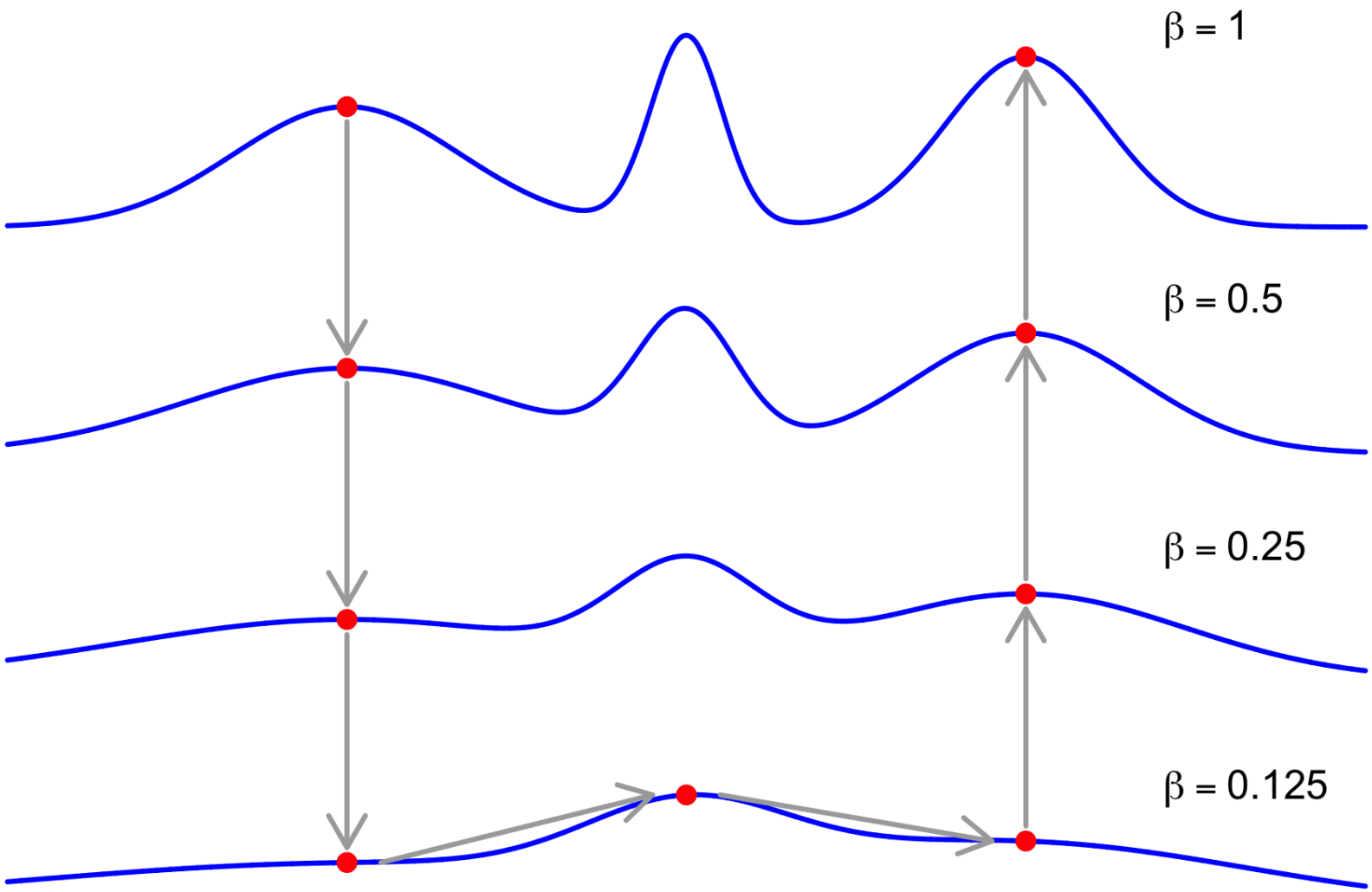
Simulated tempering [Marinari and Parisi, 1992] augments the state space to $[T] \times \mathbb{R}^d$ with stationary density

$$\pi^*(i, x) = \frac{r_i \, \pi^*(x)^{\beta_i}}{\int_{\mathbb{R}^d} \pi^*(y)^{\beta_i} \mathrm{d}y}, \qquad 0 < \beta_1 \leq \cdots \leq \beta_T = 1.$$

Here $[T] = \{1, \ldots, T\}$ indexes the temperature levels, β_i are inverse temperatures, and $r_i > 0$ are weights with $\sum_i r_i = 1$.

- With probability $\alpha \in (0, 1)$, propose changing the temperature level: $i \to i'$.
- With probability $1 - \alpha$, fix i and update x using RWM or MALA at level β_i.
- Small β_i flatten the landscape, so the chain can move between components more easily.
- Returning to $\beta_T = 1$ recovers the original target.

How Tempering Helps



How Tempering Helps

Local buses explore one city well, regional trains connect nearby cities, and flights are needed for very distant cities.

Known Results about Simulated Tempering

In Woodard–Schmidler–Huber (2009) , "Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions", spectral gap decomposition was used to establish the polynomial mixing time of simulated tempering.

About their set-up:

- A mixture of two Gaussian distributions with means $-\mathbf{1}$ and $\mathbf{1}$.
- At each temperature level, RWM is used.
- $T = d + 1$ and $\beta_i = d^{-(d-i+1)/d}$. In particular,

$$\beta_1 = \frac{1}{d}, \quad \frac{\beta_{i+1}}{\beta_i} = d^{1/d} \approx 1 + \frac{\log d}{d}.$$

Known Results about Simulated Tempering

Ge–Lee–Risteski (2018) studied simulated tempering Langevin diffusion with Euler discretization. They considered general location mixtures of strongly log-concave components and obtained a mixing time polynomial in d, D, ϵ^{-1} (rather than $\log \epsilon^{-1}$). They required

$$\frac{\beta_{i+1}}{\beta_i} = 1 + O\left(\frac{1}{d}\right).$$

Atchadé–Roberts–Rosenthal (2011) used optimal scaling techniques to show that, for certain targets including Gaussian, the asymptotically optimal choice of the temperature spacing is

$$\frac{\beta_{i+1}}{\beta_i} = 1 + O\left(\frac{1}{\sqrt{d}}\right).$$

What Was Missing

- Polynomial-time guarantee for simulated tempering + MALA.
- Polynomial dependence on d, D, and $\log \epsilon^{-1}$ for general location mixtures of strongly log-concave components.
- Characterization of the optimal choice of $(\beta_i)_{1 \le i \le T}$ for mixture targets.

Main Results

Target Distribution and Complexity Parameters

Assume

$$\pi^*(x) \propto e^{-U(x)}, \qquad U(x) = -\log \sum_{j=1}^{K} w_j e^{-f(x-\mu_j)},$$

where U is the mixture potential and f is twice differentiable, L-smooth, and m-strongly convex, with $f(0) = 0$. Recall that $r_i = \int \pi^*(i, x)\mathrm{d}x$ is the marginal probability of the i-th temperature level in simulated tempering.

Model Complexity Parameters

$$\kappa = \frac{L}{m}, \quad D = \max_{j\in[K]} \|\mu_j\|, \quad w_{\min} = \min_j w_j, \quad r_{\min} = \min_i r_i, \quad \tilde{r} = \min_{i,i'} \frac{r_{i'}}{r_i}.$$

Proposal Kernels and Temperature Ladder

Within-level proposals

$$Q_i^{\mathrm{RWM}}(x,\cdot) = N\big(x, (2h\beta_i^{-1})I_d\big),$$
$$Q_i^{\mathrm{MALA}}(x,\cdot) = N\big(x - h\nabla U(x), (2h\beta_i^{-1})I_d\big).$$

Here Q_i is the proposal kernel used at temperature level i, and the factor β_i^{-1} keeps the effective step size comparable across temperatures.

Ladder choice

$$T = \left\lceil (\kappa\sqrt{d}+1)\log(4LD^2+1) \right\rceil,$$
$$\frac{\beta_{i+1}}{\beta_i} = 1 + \frac{1}{\kappa\sqrt{d}},$$
$$\beta_T = 1, \text{ which implies } \beta_1 \le \frac{1}{4LD^2}.$$

Main Theorem

Under a warm start, the mixing time of simulated tempering with RWM or MALA at each level is polynomial in $d, \log \epsilon^{-1}$ and all model complexity parameters (including D), if the step size h satisfies

$$h = \frac{1}{Ld}, \qquad \text{for RWM,}$$
$$h \approx \frac{c}{L^2(D+\sqrt{d})^2 d}, \qquad \text{for MALA.}$$

Further, the dependence of our temperature ladder on d is asymptotically optimal up to a logarithmic factor.

Complexity Takeaways

Simulated Tempering + RWM

Assume $r_{\min} = \Theta(T^{-1})$. The number of iterations required to reach ϵ-accuracy in TV distance is

$$t_{\mathrm{RWM}} = O\left(\frac{d^2\kappa^3 \log^2(LD)}{w_{\min}^2} \log \frac{\tilde{\eta}}{2\epsilon}\right),$$

where $\tilde{\eta}$ is a χ^2-warm start parameter.

Simulated Tempering + MALA

$$t_{\mathrm{MALA}} \approx O\left(\frac{d^3(d+mD^2)\kappa^6 \log^4(LD)}{w_{\min}^4} \log \frac{2\eta}{\epsilon}\right),$$

where η is another warm start parameter.

Review on Spectral Gap and Conductance

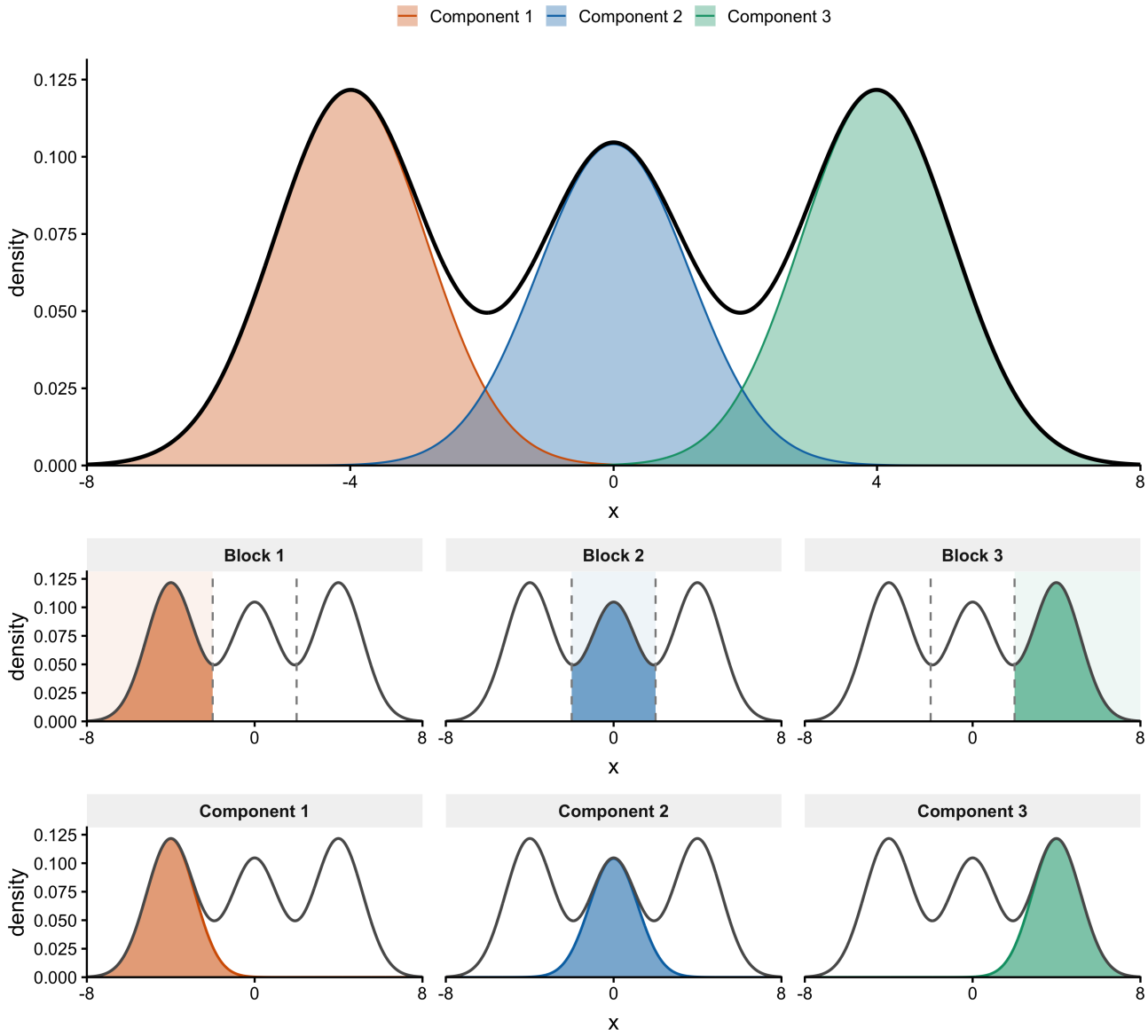
Spectral Gap Decomposition: Two Approaches

Let $(I, X) \sim \pi^*(i, x)$. Spectral gap decomposition begins by applying the law of total variance:

$$\mathrm{Var}(g(I, X)) = \mathsf{E}[\mathrm{Var}(g(I, X) \mid Z)] + \mathrm{Var}(\mathsf{E}[g(I, X) \mid Z]).$$

- Woodard et al. [2009] considered state space decomposition: letting $A_1, \ldots, A_n$ be a disjoint partition of $\mathbb{R}^d$, define $Z = k$ if $X \in A_k$.
- Ge et al. [2018] proposed the Markov chain decomposition, where Z denotes the mixture membership of X.

Spectral Gap Decomposition: Two Approaches



Spectral Gap and Conductance

Let P be a reversible transition kernel with state space $\mathcal{X}$ and stationary density μ. Let $Y_0 \sim \mu$, $Y_1 \,|\, Y_0 \sim P(Y_0, \cdot)$.

Spectral gap of P:

$$\mathrm{Gap}(P) = \inf_{g:\mathrm{Var}(g(Y_0))>0} \frac{\mathsf{E}\left[(g(Y_1) - g(Y_0))^2\right]}{2\mathrm{Var}(g(Y_0))}.$$

Conductance of P:

$$\begin{aligned}\Phi(P) &= \inf_{A\subset\mathcal{X}:\ \mu(A)\in(0,1)} \frac{P(A, A^c)}{\mu(A)\mu(A^c)} \\ &= \inf_{g=\mathbb{1}_A:\ \mu(A)\in(0,1)} \frac{\mathsf{E}\left[(g(Y_1) - g(Y_0))^2\right]}{2\mathrm{Var}(g(Y_0))},\end{aligned}$$

Restricted Spectral Gap and s-conductance

Restricted spectral gap on $\mathcal{X}_0 \subset \mathcal{X}$ [Atchadé, 2021]:

$$\inf_{g:\mathrm{Var}(g(Y_0))>0} \frac{\mathsf{E}\left[(g(Y_1)-g(Y_0))^2 \mathbb{1}_{\mathcal{X}_0}(Y_0)\mathbb{1}_{\mathcal{X}_0}(Y_1)\right]}{2\mathsf{E}\left\{(g(Y_0)-\mathsf{E}[g(Y_0)])^2 \mathbb{1}_{\mathcal{X}_0}(Y_0)\right\}}.$$

For $s \in [0, 1/2)$, the s-conductance of P [Lovász and Simonovits, 1993]:

$$\Phi_s(P) = \inf_{A\subset\mathcal{X}:\, \mu(A)\in(s,1/2]} \frac{P(A, A^c)}{\mu(A)-s}.$$

- Under warm start condition, restricted spectral gap or s-conductance can be used to bound the mixing time of P.
- Decomposition techniques also work for restricted spectral gap, conductance, and s-conductance.

Proof Ideas: Conductance Decomposition

Step I: Build an Augmented Auxiliary Chain

Let P^* denote the transition kernel of the actual simulated tempering chain living on $(i, x) \in [T] \times \mathbb{R}^d$, where i is the temperature level and x is the position. Let $Q^*((i, x), \cdot)$ denote this proposal scheme on $[T] \times \mathbb{R}^d$.

Now expose the mixture label $j \in [K]$ and define

$$\pi(i, j, x) = \frac{r_i}{C_i} w_j e^{-\beta_i f(x-\mu_j)}, \qquad C_i = \int_{\mathbb{R}^d} e^{-\beta_i f(x)} \, \mathrm{d}x.$$

We construct an auxiliary chain P on $[T] \times [K] \times \mathbb{R}^d$.

Step I: Build an Augmented Auxiliary Chain

The chain P evolves as follows:

- with probability $1/2$, keep j fixed and update (i, x) via a Metropolis step with the same (i, x)-proposal Q^* as in P^*;
- with probability $1/2$, keep (i, x) fixed and resample $j' \sim \pi_{i,x}(\cdot)$, where

$$\pi_{i,x}(j) = \frac{\pi(i, j, x)}{\sum_{\ell=1}^{K} \pi(i, \ell, x)} \text{ denotes the full conditional.}$$

MALA remark. If $Q_i(x, \cdot) = N(x - h\nabla U(x), (2h/\beta_i)I_d)$ with $U(x) = -\log \sum_{\ell=1}^{K} w_\ell e^{-f(x-\mu_\ell)}$, then the drift is still computed from the *original mixture potential* U, not from the component potential $f(x - \mu_j)$.

Step II: Compare P^* with P

A Comparison Lemma

For every $s \in [0, 1/2)$,

$$\Phi_s(P^*) \geq 2w_{\min}^2 \Phi_{sw_{\min}}(P), \qquad \mathrm{Gap}(P^*) \geq 2w_{\min}^2 \mathrm{Gap}(P),$$

where $w_{\min} = \min_j w_j$.

This direct P^*-to-P comparison is new: the intuition is similar to the soft-decomposition viewpoint of Ge et al. [2018], but here the two chains live on different state spaces. Instead of a soft decomposition of P^*, we can apply a state decomposition to P.

Step III: State Decomposition Theorem for s-Conductance

Partition the state space as

$$\Theta = \bigcup_{(i,j)\in[T]\times[K]} \Theta_{i,j}, \qquad \Theta_{i,j} = \{i\} \times \{j\} \times \mathbb{R}^d.$$

Let $P_{i,j}$ be the restricted chain on $\Theta_{i,j}$ and $\overline{P}$ the projected chain on $[T] \times [K]$.

- The restricted chain $P_{i,j}$ handles local exploration. It lives on one temperature level and one component with stationary density

$$\pi_{i,j}(x) \propto e^{-\beta_i f(x-\mu_j)}.$$

- The projected chain $\overline{P}$ handles temperature moves and label changes. It only remembers the pair (i, j) and has stationary distribution

$$\pi(i, j) = r_i w_j.$$

- For MALA, $P_{i,j}$ is a *pseudo-MALA* chain: its stationary density is $\pi_{i,j}$, but the drift still uses the mixture potential U from the original target.

Step III: State Decomposition Theorem for s-Conductance

General Decomposition Bound

For the auxiliary chain P,

$$\Phi_s(P) \geq \frac{\mathrm{Gap}(\overline{P})}{8} \min_{i,j} \Phi_{(\mathrm{Gap}(\overline{P})/8)s}(P_{i,j}).$$

So the proof reduces to two tasks:

- show fast movement *between* blocks through $\mathrm{Gap}(\overline{P})$;
- show fast movement *within* each block through $\Phi_s(P_{i,j})$.

Step IV: Projected-Chain Spectral Gap

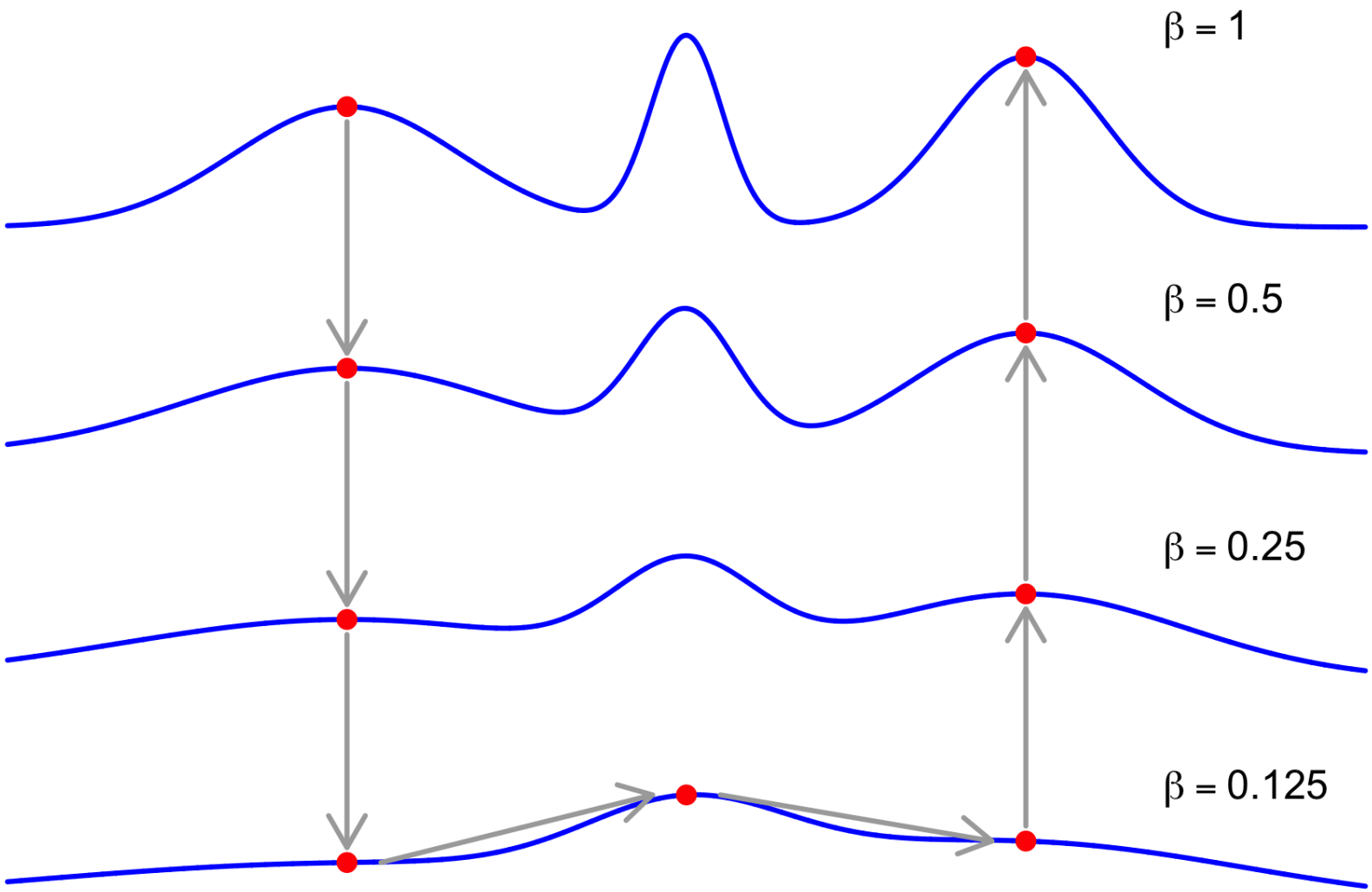
At the cold level $\beta_T = 1$, the chain should preserve the target. At the hot level β_1, different components overlap substantially.

So a canonical route from one cold component to another is

$$(T, j) \to (T-1, j) \to \cdots \to (1, j) \to (1, j') \to \cdots \to (T, j').$$

- Vertical motion in temperature keeps the same component label.
- Horizontal motion across labels happens at the hottest temperature.
- This is exactly where the separation parameter D enters the analysis.

How Tempering Helps



Step IV: Projected-Chain Spectral Gap

Projected-Chain Bound

$$\mathrm{Gap}(\overline{P}) \geq \frac{C\tilde{r}r_{\min}}{T} \min\left\{1-\sqrt{\Delta},\, H^2\right\},$$

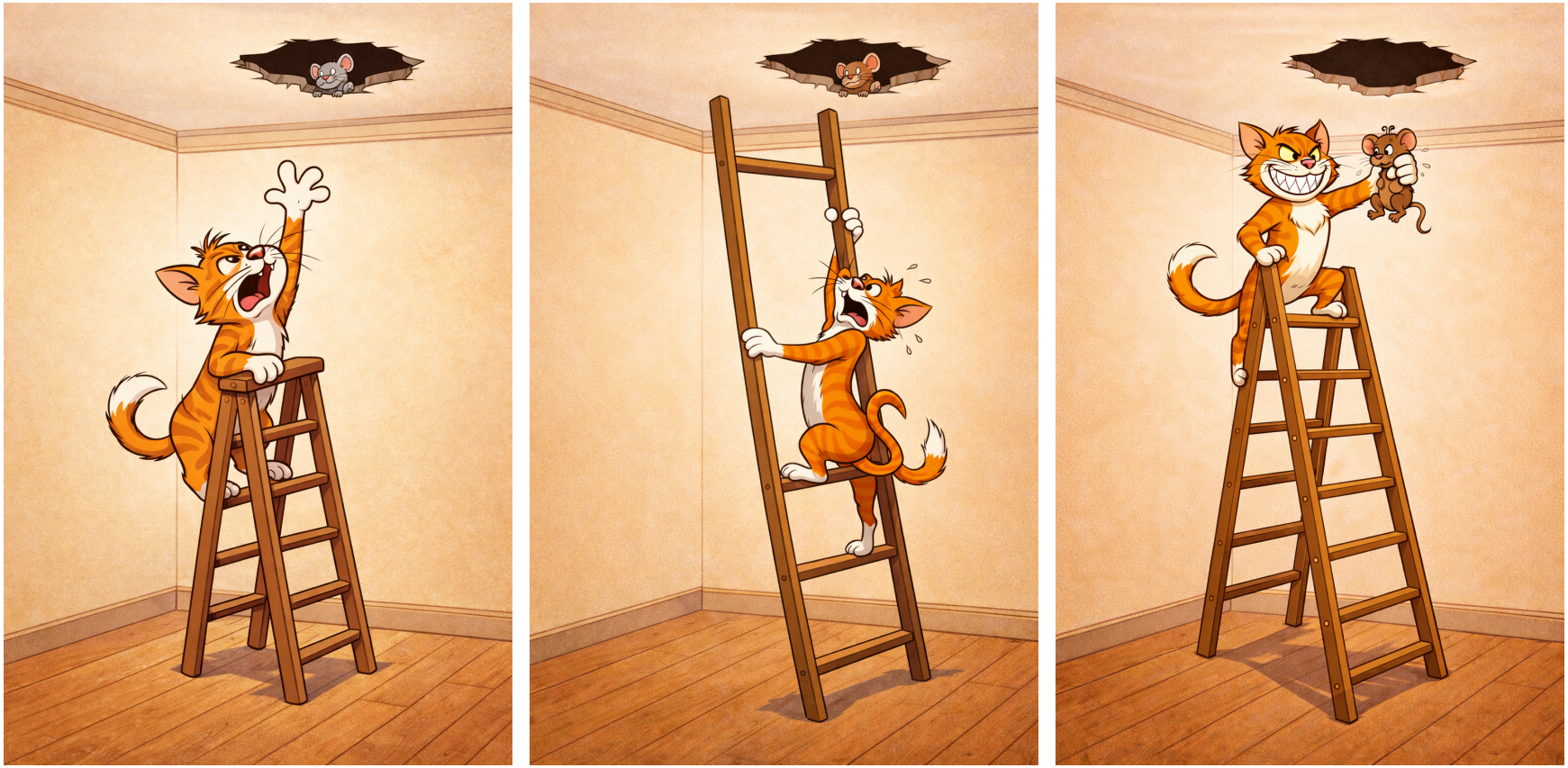
$$\text{where } \Delta = \frac{1}{2} \max_{i\in[T-1],\, j\in[K]} \mathrm{KL}(\Pi_{i,j} \,\|\, \Pi_{i+1,j}),$$

$$H = \min_{j,j'\in[K]} \int_{\mathbb{R}^d} \sqrt{\pi_{1,j}(x)\pi_{1,j'}(x)}\,\mathrm{d}x.$$

- Δ measures how compatible adjacent temperatures are.
- H measures how much overlap remains between components at level β_1.
- We show that to ensure $\min\{1-\sqrt{\Delta},\, H^2\} \geq 3/4$, it is enough to impose

$$\beta_1 \leq \frac{1}{4LD^2}, \qquad \frac{\beta_{i+1}}{\beta_i} \leq 1 + \frac{1}{\kappa\sqrt{d}}.$$

Why the Ladder Choice Works



Step V: Restricted-Chain Conductance for RWM

For the restricted chain $P_{i,j}$, if

$$Q_i(x,\cdot) = N(x,(2h\beta_i^{-1})I_d), \qquad h = \frac{1}{Ld},$$

then

$$\Phi_0(P_{i,j}) \geq \frac{c(1-\alpha)}{\sqrt{d\kappa}}, \qquad \mathrm{Gap}(P_{i,j}) \geq \frac{c(1-\alpha)}{d\kappa}.$$

- Each restricted target is genuinely strongly log-concave.
- So the recent sharp theory for RWM on log-concave targets applies directly [Andrieu et al., 2024].
- This leads to the cleaner final complexity bound in the RWM corollary.

Step V: Restricted-Chain Conductance for MALA

Now let

$$Q_i(x, \cdot) = N(x - h\nabla U(x), (2h/\beta_i)I_d).$$

This is *not* the same as true MALA for the restricted target $\pi_{i,j}$, because the drift uses the mixture potential U, not the component potential $f(x - \mu_j)$.

Pseudo-MALA Bound

If

$$h \approx \frac{c}{L^2(D + \sqrt{d})^2 d},$$

for some sufficiently small $c > 0$, then

$$\Phi_s(P_{i,j}) \geq c'\sqrt{hm}.$$

This is enough for polynomial mixing, but the bound is probably not sharp.

Optimal Temperature Ladder

Necessary Conditions on the Temperature Ladder

To derive necessary conditions, we simply consider a two-Gaussian counterexample with modes at $\pm(D, 0, \ldots, 0)$. We show that to achieve polynomial complexity in d and D, we need

- β_1 must be small enough:

$$\beta_1 = O\left(\frac{\log D}{D^2}\right).$$

- The ladder spacing needs to be dense enough:

$$\frac{\beta_{i+1}}{\beta_i} = 1 + O\left(\sqrt{\frac{\log d}{d}}\right).$$

- So the sufficient conditions we require for our polynomial mixing result are optimal in d up to logarithmic factors.

Concluding Remarks

Discussion and Open Problems

- The MALA result is polynomial, but the current complexity bound is rougher than the RWM bound.
- It remains open whether MALA truly improves on RWM for simulated tempering with mixture targets.
- The conductance decomposition is multiplicative; that may still lose powers of T and hence powers of d.
- The auxiliary-chain framework is flexible and may extend to other samplers:
  - proximal samplers,
  - parallel tempering,
  - sequential Monte Carlo,
  - non-reversible tempering variants.

Takeaways

- Simulated tempering achieves polynomial mixing for general location mixtures of strongly log-concave components.
- The manuscript gives the first non-asymptotic polynomial guarantee covering simulated tempering with MALA.
- The temperature ladder

$$\beta_1 \asymp D^{-2}, \qquad \frac{\beta_{i+1}}{\beta_i} = 1 + O(d^{-1/2})$$

  is both sufficient and essentially necessary (up to logarithmic factors).

Acknowledgment

Thanks to Jhanvi Garg, Krishnakumar Balasubramanian, and many others for helpful discussion.
All pictures were made by ChatGPT.

The research presented in this talk is supported by NSF DMS-2245591, DMS-2311307.

- J. Garg, K. Balasubramanian, Q. Zhou. "Restricted spectral gap decomposition for simulated tempering targeting mixture distributions." *NeurIPS*, 2025.
- Q. Zhou. "Polynomial mixing time of simulated tempering by conductance decomposition."

References I

Christophe Andrieu, Anthony Lee, Sam Power, and Andi Q Wang. Explicit convergence bounds for Metropolis Markov chains: Isoperimetry, spectral gaps and profiles. *The Annals of Applied Probability*, 34(4):4022–4071, 2024.

Yves F Atchadé. Approximate spectral gaps for Markov chain mixing times in high dimensions. *SIAM Journal on Mathematics of Data Science*, 3(3):854–872, 2021.

Yves F Atchadé, Gareth O Roberts, and Jeffrey S Rosenthal. Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo. *Statistics and Computing*, 21(4):555–568, 2011. doi: 10.1007/s11222-010-9192-1.

Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm. In *Conference on Learning Theory*, pages 1260–1300. PMLR, 2021.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183):1–42, 2019.

Rong Ge, Holden Lee, and Andrej Risteski. Simulated tempering Langevin Monte Carlo II: An improved proof using soft Markov chain decomposition. *arXiv preprint arXiv:1812.00793*, 2018.

László Lovász and Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random structures & algorithms*, 4(4):359–412, 1993.

Enzo Marinari and Giorgio Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhysics letters*, 19(6):451, 1992.

Dawn B Woodard, Scott C Schmidler, and Mark Huber. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *The Annals of Applied Probability*, 19(2):617–640, 2009.

Keru Wu, Scott Schmidler, and Yuansi Chen. Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling. *Journal of Machine Learning Research*, 23(270):1–63, 2022.