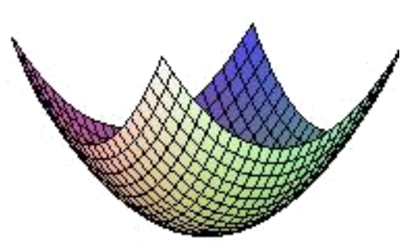
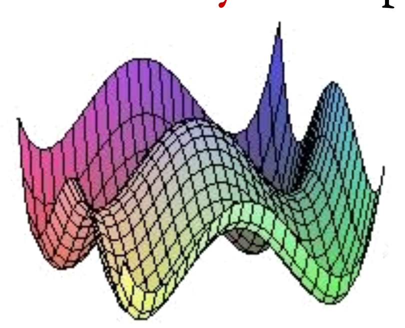
Beyond Log-concavity: Provable Guarantees for Sampling Multi-Modal Distributions using Simulated Tempering Langevin Monte Carlo

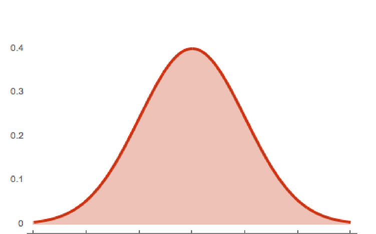
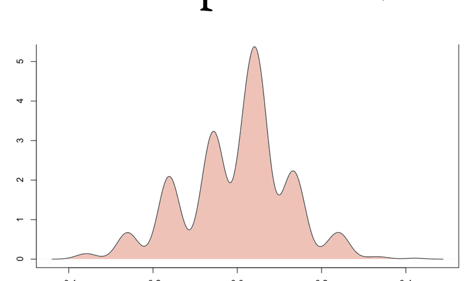
Rong Ge[1], Holden Lee[2], Andrej Risteski[3]

Problem

Sample from distribution $p(x) \propto e^{-f(x)}, x \in \mathbb{R}^d$ given access to $f(x), \nabla f(x)$. (*e.g. sampling posteriors*)

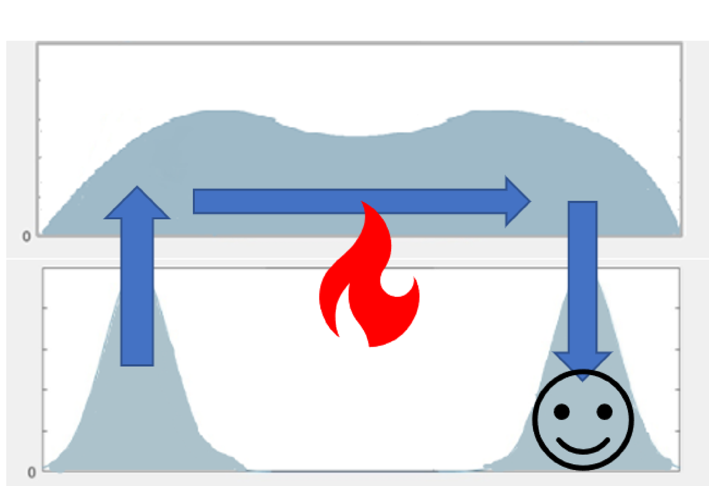
Background

The great divide of optimization	
Convex optimization	**Non-convex optimization**
Local minima = global minima	Possibly bad local minima
Gradient descent finds global min.	Gradient descent can be bad
Provable algorithms, beautiful math	NP-hard in the worst-case (messy?)
ML problems often non-convex	Works remarkably well in practice
	

The great divide of sampling	
Log-concave distribution	**Non-log-concave distribution**
Unimodal	Potentially multimodal
Natural algorithm: Langevin diffusion	Langevin can mix exponentially slowly
Provable algorithms, beautiful math	#P-hard in the worst-case (messy?)
ML problems often non-log-concave	Works well in practice w/ temperature
	

Fixing Langevin?

A Markov chain with local moves such as Langevin diffusion gets stuck in a local mode.



Creating a meta-Markov chain which changes the temperature (simulated tempering) can exponentially speed up mixing.

Our question: Can we give **provable guarantees** for such an algorithm in natural "non-log-concave" settings?

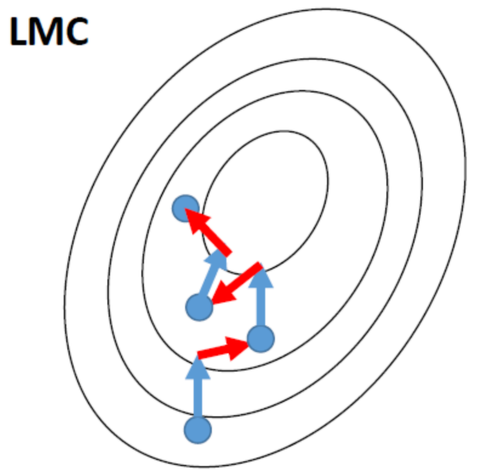
Main Theorem

Let $p(x) \propto e^{-f(x)}$ on $\mathbb{R}^d$ be s.t. $f(x) = -\log\left(\sum_{j=1}^n w_j e^{-\frac{\|x-\mu_j\|^2}{2\sigma^2}}\right)$ and we can query $f(x), \nabla f(x)$. There is an algorithm (based on Langevin diffusion + simulated tempering) running in time $\text{poly}\left(\frac{1}{w_{\min}}, \frac{1}{\sigma^2}, \frac{1}{\varepsilon}, d, \max\|\mu_j\|\right)$ that samples from a distribution q with $\|p-q\|_1 \le \varepsilon$. A L^∞ perturbation of Δ multiplies time by a factor $\text{poly}(e^\Delta)$.

Algorithmic tools

1. **Langevin diffusion** (gradient flow + Brownian motion, or in discrete form, gradient descent + gaussian noise)
2. **Simulated tempering**: heuristic for speeding up MCs on multimodal distributions

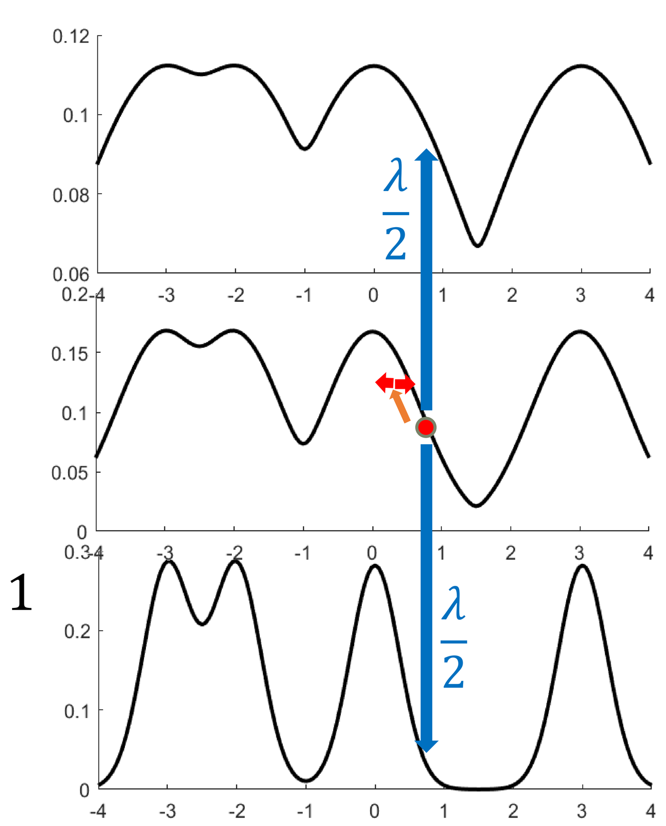
LMC



Simulated tempering + Langevin diffusion

At point (i, x),

- Evolve according to Langevin with inverse temperature β_i: $dx_t = -\beta_i \nabla f(x_t)dt + \sqrt{2}dW_t$.
- Propose swaps with rate λ. When a swap is proposed, pick $i' = i \pm 1$ with probability ½. Set next point to (i', x) with probability $\min\left(\frac{p_{i'}(x)}{p_i(x)}, 1\right)$.



Proof outline

1. Markov chain decomposition theorem

2. Mixing for each component

3. Mixing for "projected" chain

=> Mixing for "approximate" heating: $\tilde{p}_i \propto \sum_{j=1}^m w_j \exp\left(-\frac{\beta_i\|x-\mu_j\|^2}{2}\right)$

4. Mixing for "actual" heating $p_i \propto \left[\sum_{j=1}^m w_j \exp\left(-\frac{\|x-\mu_j\|^2}{2}\right)\right]^{\beta_i}$

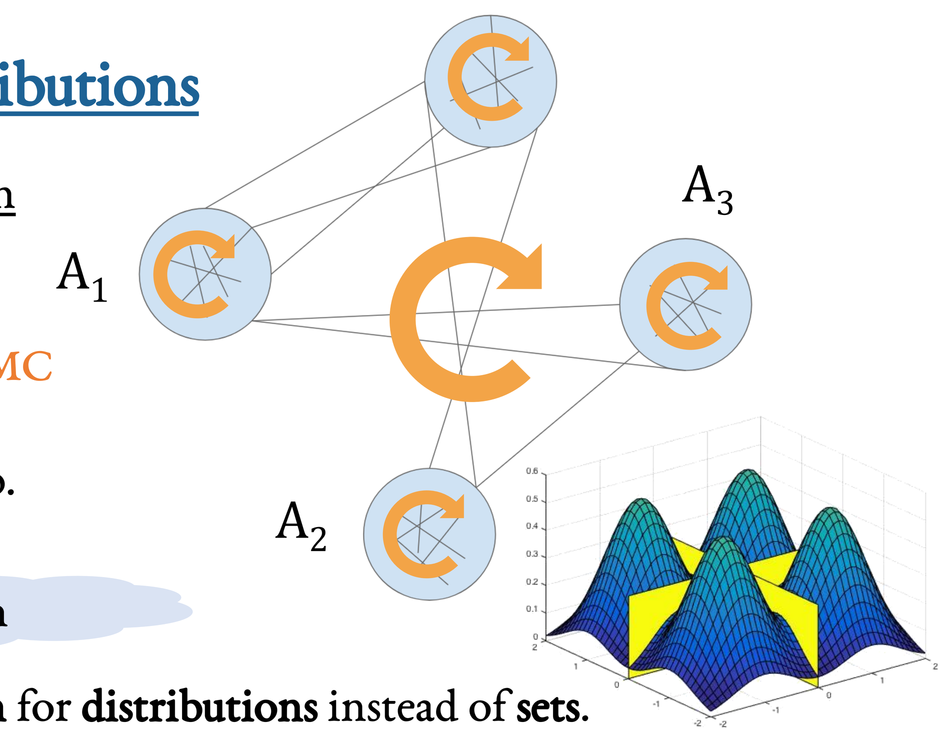
Main theorem

Decomposing using distributions

Inspiration: MC decomposition theorem (Madras, Randall 2002)

If MC mixes rapidly when restricted to each set of a partition, and "projected" MC mixes rapidly => MC mixes rapidly.

(Transition in projected chain: avg. prob. flow between sets.)



Soft partition

We prove a **new decomposition theorem** for **distributions** instead of **sets**.

Soft decomposition theorem

Let tempering chain be made up of Markov chains M_i. Suppose there is a **decomposition** $M_i(x,y) = \sum_{j=1}^m w_{ij} M_{ij}(x,y)$ where M_{ij} has stationary distribution p_{ij}. If each M_{ij} mixes in time C and projected chain mixes in time $\bar{C}$ => simulated tempering chain mixes in time $O(C(\bar{C}+1))$.

Intuition:

(1) mixing time is equal to Poincaré constant $\max_g[\text{Var}_p(g)/\mathcal{E}(g,g)]$ where $\mathcal{E}(g,h) = -\langle g, \mathcal{L}h\rangle_p$ is Dirichlet (bilinear) form and $\mathcal{L}$ is the generator of MC.

(2) Dirichlet form "decomposes" into Langevin chains for components, and variance decomposes as

$$\text{Var}_p(g) = \sum_{i=1}^L \sum_{j=1}^m \frac{1}{L} w_{ij} \Big[\underbrace{\text{Var}_{p_{ij}}(g_i)}_{\text{Use Poincare inequality for } p_{ij}\text{, get factor of } C.} + \underbrace{\left(\mathbb{E}_{p_{ij}} g_i - \mathbb{E}g\right)^2}_{\text{Use Poincare inequality for } \bar{p}\text{, get factor of } \bar{C}.}\Big]$$

Again, for intuition, 2 extreme cases.

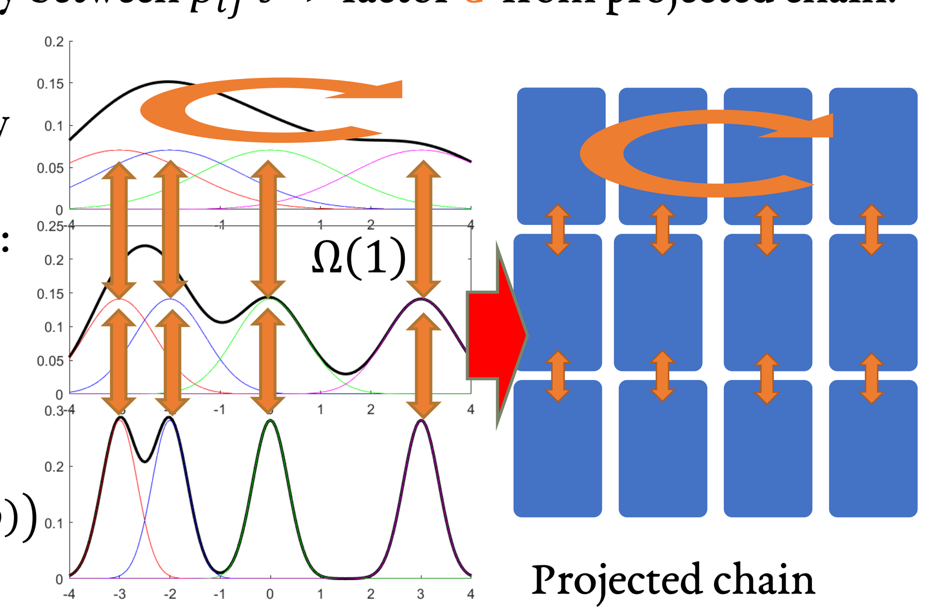
1. If all expectations $\mathbb{E}_{p_{ij}} g_i$ are equal => factor C from the component chains.
2. If g_i's constant on each p_{ij}, only vary between p_{ij}'s => factor $\bar{C}$ from projected chain.

The projected chain has large probability flow between (i,j) in the same or adjacent levels with similar distributions:

$$\bar{L}((i,j),(i,j')) = \frac{w_{ij'}}{\chi^2_{\text{sym}}(p_{ij}, p_{ij'})}$$

$$\bar{L}((i,j),(i' = i \pm 1, j)) = \frac{1}{\chi^2_{\text{sym}}(p_{ij}, p_{i'j})}$$

where $\chi^2_{\text{sym}}(p,q) = \max(\chi^2(p\|q), \chi^2(q\|p))$



Projected chain

Using the decomposition theorem:

(1) Apply Langevin for "approximately" heated distributions (Langevin on individual components $p_{ij} \propto \exp\left(-\frac{\beta_i\|x-\mu_j\|^2}{2}\right)$ mixes rapidly).

(2) Compare to "actually" heated distributions, losing factors of $w_{\min}$.

1. Duke University
2. Princeton University
3. Massachusetts Institute of Technology

More info: tiny.cc/glr17