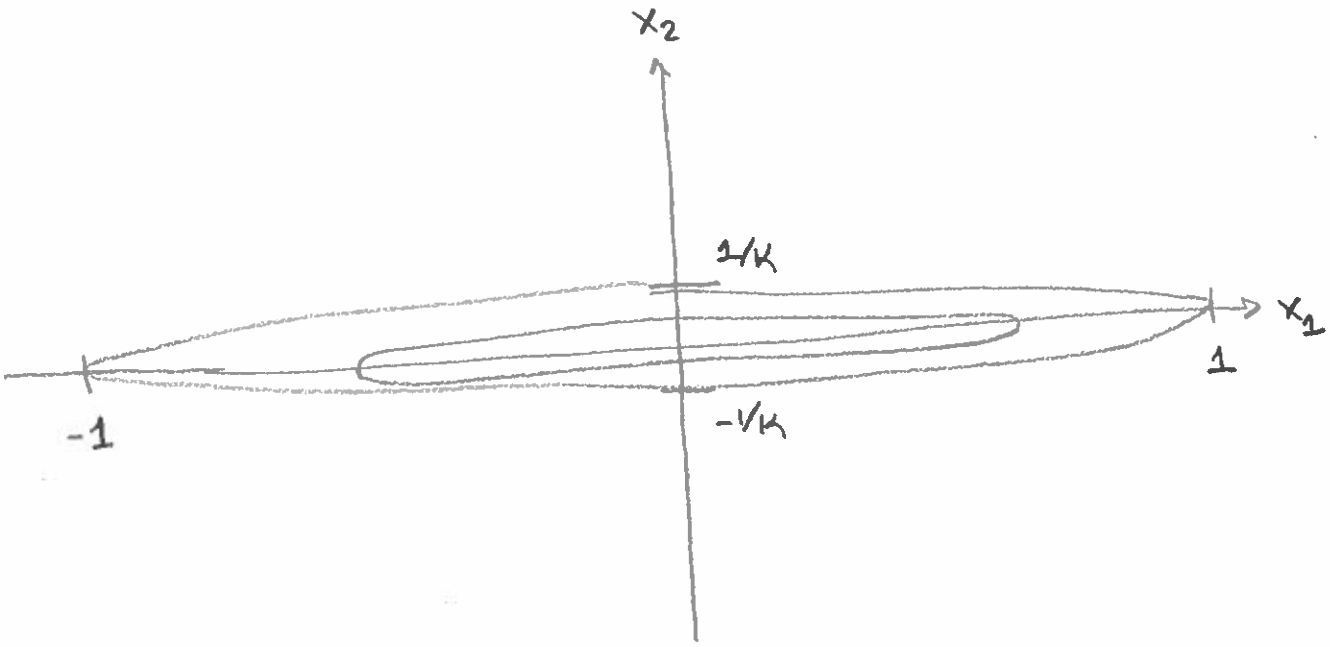
Lecture 8 : All Killer, no filler

- ☐ Review of MALA + HMC
- ☐ Goodman-Weare sampler
- ☐ Parallel tempering

MALA + HMC Review

- Imagine we are sampling a Gaussian $\eta(0, \begin{pmatrix} 1 & \\ & \frac{1}{K} \end{pmatrix})$ where $K \gg 1$ is the "condition number".



- Target distribution is wide in x_1 direction, narrow in x_2 direction.

MALA uses the update

$$x \longleftarrow x + \delta \nabla \log \pi(x) + \sqrt{2\delta}\, v,$$

where v is a random Gaussian velocity $v \sim \eta(0, I)$

Q what is $\nabla \log \pi(x)$?

Hint $\pi(x) = \frac{\sqrt{K}}{2\pi} \exp\left(-\frac{x_1^2}{2} - \frac{K}{2}x_2^2\right)$

A $\nabla \log \pi(x) = \begin{pmatrix} -x_1 \\ -Kx_2 \end{pmatrix}$

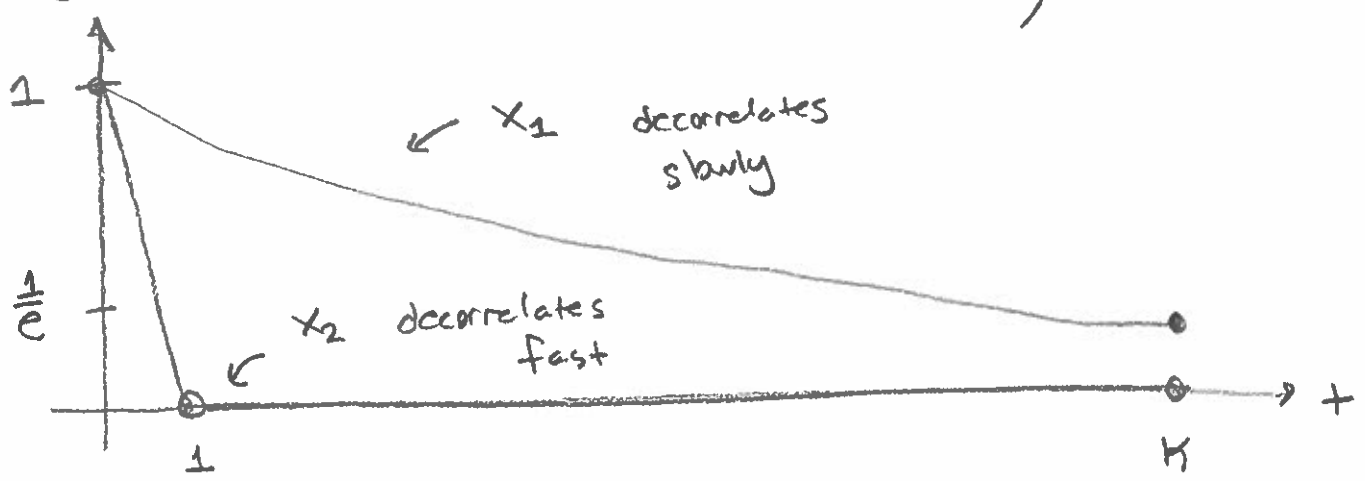
- Update formula is

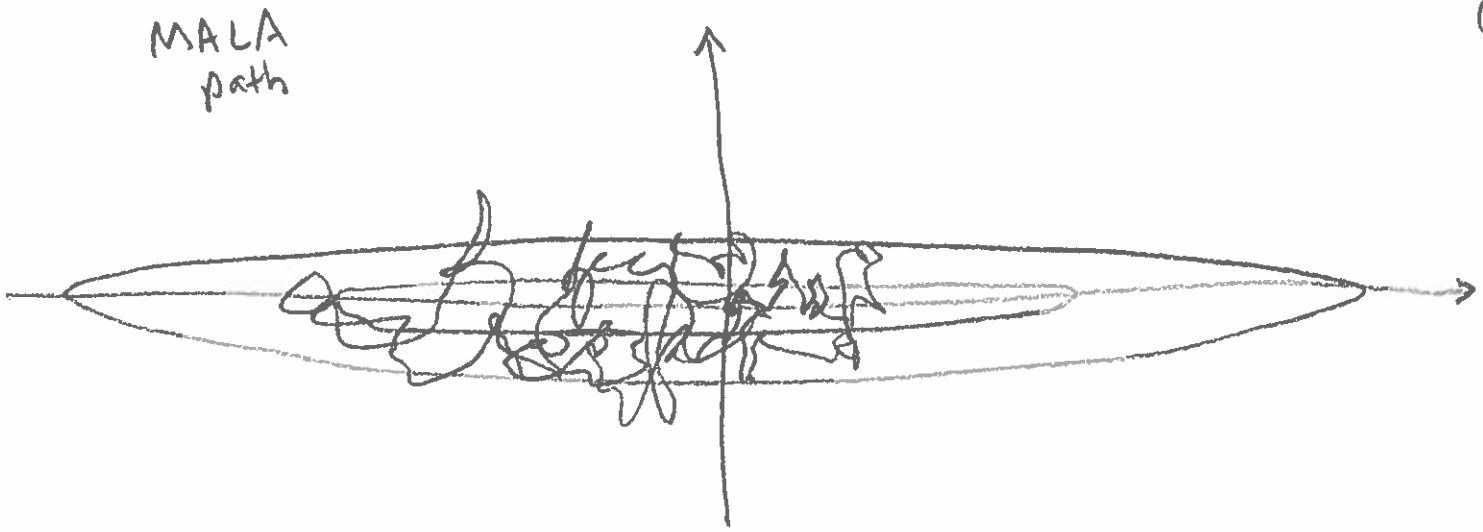
$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \longleftarrow \begin{pmatrix} x \\ x_2 \end{pmatrix} + \delta \begin{pmatrix} -x_1 \\ -Kx_2 \end{pmatrix} + \sqrt{2\delta} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

$$\Rightarrow \begin{cases} x_1 \leftarrow (1-\delta)x_1 + \sqrt{2\delta}\, v_1 \\ x_2 \longleftarrow (1-K\delta)x_2 + \sqrt{2\delta}\, v_2 \end{cases}$$

- For stability, we need $\delta \in \left(0, \frac{2}{K}\right)$.
- If we set $\delta = \frac{1}{K}$, correlations decay as

$$\mathbb{E}\left[x^{(t)} \mid x^{(0)} = x \right] = \begin{pmatrix} 1 - \frac{1}{K} & & \\ & & \\ & & 0 \end{pmatrix}^t x$$





$\Rightarrow$ It takes $t = \Theta(\kappa)$ timesteps for the MALA sampler to explore the distribution.

$\Rightarrow$ $\boxed{t \approx 10^6 \text{ if } \kappa \approx 10^6 \quad \ddot{\frown}}$

HMC uses the update

$$v \sim \mathcal{N}(0, I)$$

For $t = 1, \dots, L$:

$$v \leftarrow v + \frac{\delta}{2} \nabla \log \pi(x), \quad x \leftarrow x + \delta v, \quad v \leftarrow v + \frac{\delta}{2} \nabla \log \pi(x).$$

Q: what is the update formula for sampling a Gaussian $\mathcal{N}\left(0, \begin{pmatrix} 1 & \\ & 1/\kappa \end{pmatrix}\right)$?

A: $v_1 \leftarrow v_1 - \frac{\delta}{2} x_1, \quad x_1 \leftarrow x_1 + \delta v_1, \quad v_1 \leftarrow v_1 - \frac{\delta}{2} x_1$

$$v_2 \leftarrow v_2 - \frac{\delta\kappa}{2} x_2, \quad x_2 \leftarrow x_2 + \delta v_2, \quad v_2 \leftarrow v_2 - \frac{\delta\kappa}{2} x_2$$

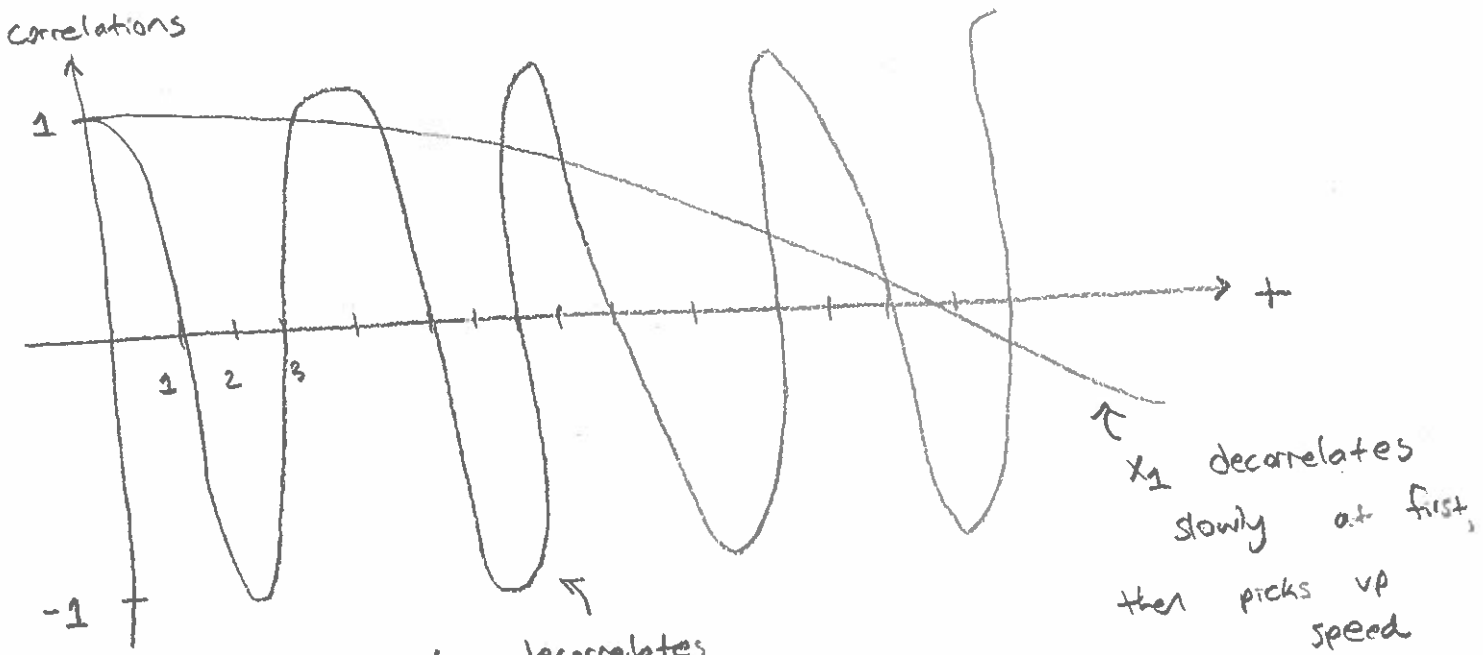
- For stability, we need to take

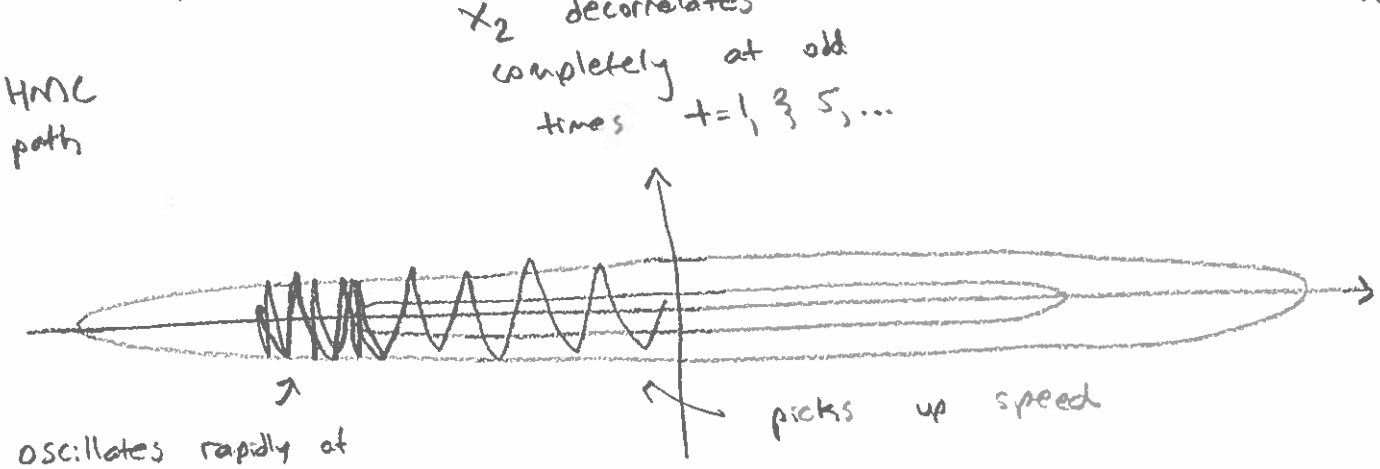
$$\delta \in \left(0, \frac{2}{\sqrt{K}}\right) \Rightarrow \text{Much larger steps!}$$

- If we set $\delta = \sqrt{\frac{2}{K}}$, correlations decay like

$$\mathbb{E}\left[X^{(t)} \mid X^{(0)} = x \right] = \begin{pmatrix} \cos(\theta t) & & \\ & \ddots & \\ & & \cos\left(\frac{\pi}{2} t\right) \end{pmatrix} x$$

where $\theta = \arccos\left(1 - \frac{1}{K}\right) \approx \sqrt{\frac{2}{K}}$





$\Rightarrow$ It takes $t = O(\sqrt{\kappa})$ timesteps for the HMC sampler to explore the distribution

$\Rightarrow$ $t \approx 10^3$ if $\kappa \approx 10^6$;)

$\Rightarrow$ "Heavy ball rolling downhill"

HMC algorithm

Input: target density $\pi(x)$, initial condition X_0, stepsize δ, number of leapfrog steps L, terminal time T

Output: approximate samples $X_0, X_1, \dots, X_T$ from π

For $t = 1, 2, \dots, T$:

- Generate $V_t \sim N(0, I)$
- Set $X_t' = X_{t-1}$
- For $i = 1, 2, \dots, L$:
  - $V_t = V_t + \frac{\delta}{2} \nabla \log \pi(X_t')$
  - $X_t' = X_t' + \delta V_t$
  - $V_t = V_t + \frac{\delta}{2} \nabla \log \pi(X_t')$
- Generate $U_t \sim \text{Unif}(0,1)$
- Set $X_t = \begin{cases} X_t', & U_t \le \pi(X_t') / \pi(X_{t-1}) \\ X_{t-1}, & \text{otherwise.} \end{cases}$

Q: What can we do to remove the dependence on k entirely?

Goodman-Weare sampler

Input: target density $\pi(x)$, number of ensemble members N, initial conditions $X_0^{(1)}, X_0^{(2)}, \dots, X_0^{(N)} \in \mathbb{R}^d$, terminal time T

Output: approximate samples $(X_t^{(1)}, \dots, X_t^{(N)})_{0 \le t \le T}$, all from π.

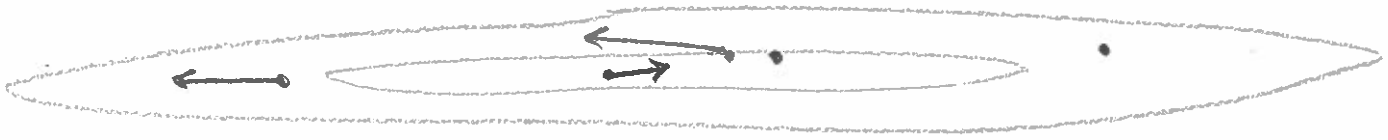
For $t = 1, 2, \dots, T$:

- Set $(X_t^{(1)}, \dots, X_t^{(N)}) = (X_{t-1}^{(1)}, \dots, X_{t-1}^{(N)})$.
- For $i = 1, 2, \dots, N$:

Stretch moves:

- Select $j \in \{1, \dots, N\} \setminus \{i\}$ uniformly at random.
- Sample $u \sim \mathrm{Unif}(2^{-3/2}, 2^{3/2})$
- Set $y_t^{(i)} = X_t^{(j)} + (1-u)(X_t^{(j)} - X_t^{(i)})$
- Sample $v \sim \mathrm{Unif}(0,1)$
- Set $X_t^{(i)} = y_t^{(i)}$ if $v \le u^{\frac{2}{3}(d-1)} \dfrac{\pi(y_t^{(i)})}{\pi(X_t^{(i)})}$.

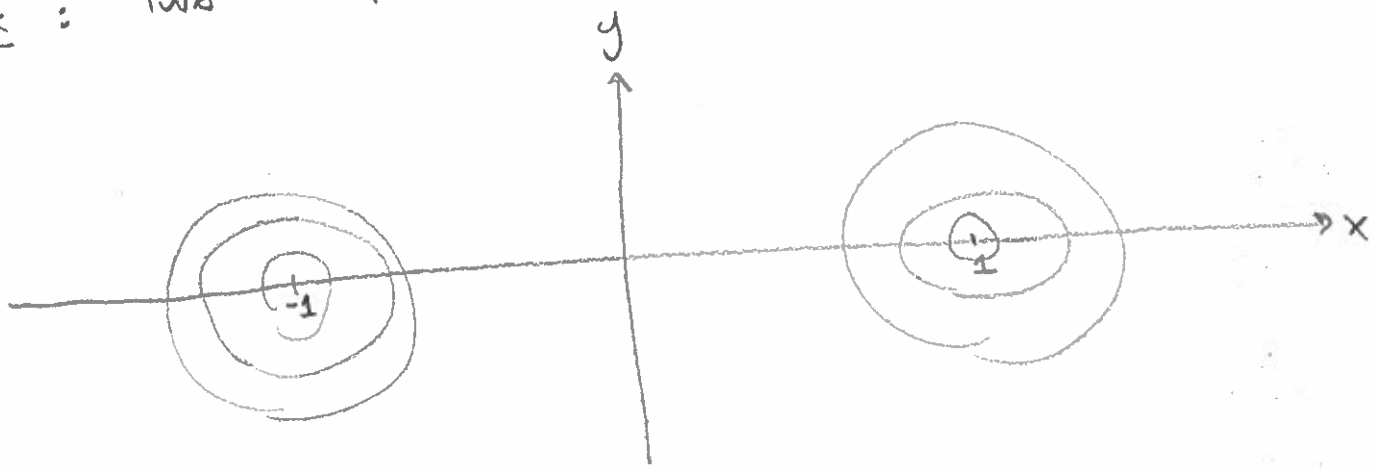
Intuition Randomly push one ensemble member toward or away from another ensemble member



- Goodman-Weare is "affine-invariant" $\Rightarrow$ correlations do not depend on K at all!
- Works well for $d \leq 10\text{-}20$.
- At each step, $X_+^{(i)}$ either halves its distance from $X_+^{(i)}$, doubles its distance from $X_+^{(i)}$, or goes somewhere in between.

Q: What can we do to address multi-modality?

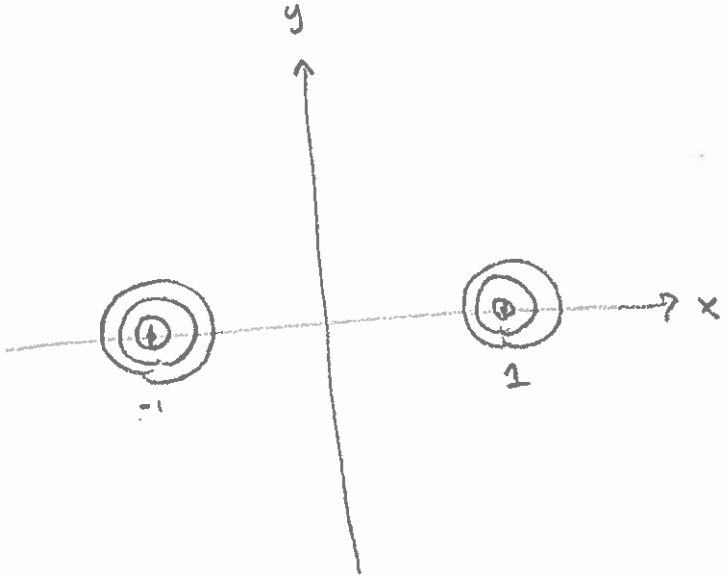
Ex: Two separated Gaussians



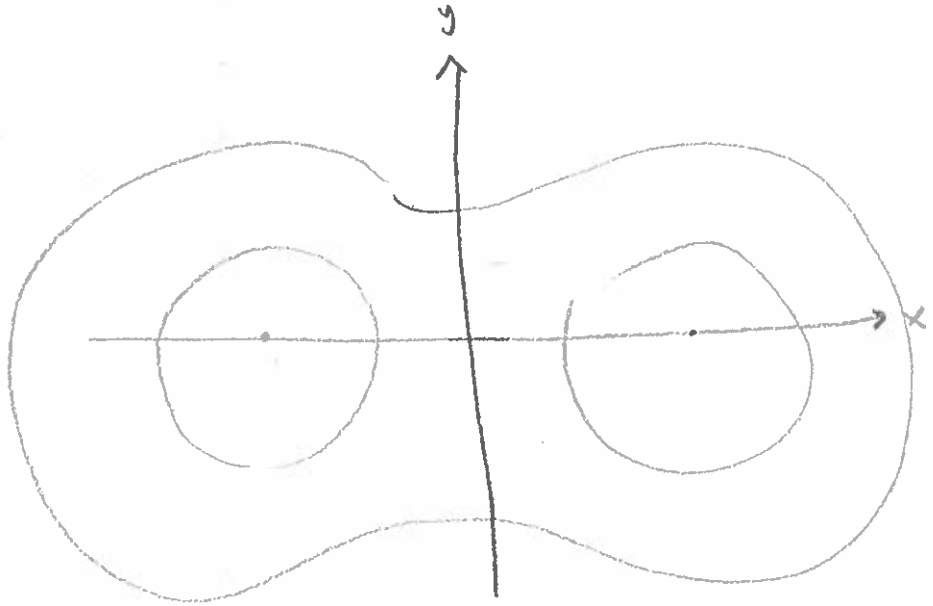
Idea: Why don't we increase the "temperature"?

$\pi(x) \propto \exp\left(\frac{-H(x)}{T}\right)$, T is the temperature.

Cold temperature $T \ll 1$:



Hot temperature $T \gg 1$:



Let's sample a range of temperatures

$$T_i = \frac{N}{i} \cdot T, \quad i = 1, 2, \ldots, N$$

$$\Rightarrow \pi_i(x) \propto \exp\left(-\frac{H(x)}{T_i}\right) \propto \pi(x)^{i/N}, \quad i = 1, 2, \ldots, N$$

Parallel tempering

Input: target density $\pi(x)$, number of ensemble members N, initial conditions $X_0^{(1)}, X_0^{(2)}, \dots, X_0^{(N)}$, terminal time T

Output: approximate samples $X_0^{(n)}, \dots, X_T^{(n)}$ from π.

For $t = 1, \dots, T$:

Parallel moves
- For $i = 1, \dots, N$:
  - Update $X_{t-1}^{(i)}$ to $X_t^{(i)}$ using an MCMC method that targets $\pi_i(x) \propto \pi(x)^{i/N}$

Swap moves
- For $i = 1, \dots, N-1$:
  - Sample $u \sim \text{Unif}(0,1)$
  - Set $(X_t^{(i)}, X_t^{(i+1)}) = (X_t^{(i+1)}, X_t^{(i)})$ if $u \le \left(\pi(X_t^{(i)}) / \pi(X_t^{(i+1)}) \right)^{1/N}$

Intuition We can jump between modes at high temperatures and information can filter down to low temperatures via a sequence of swaps.