

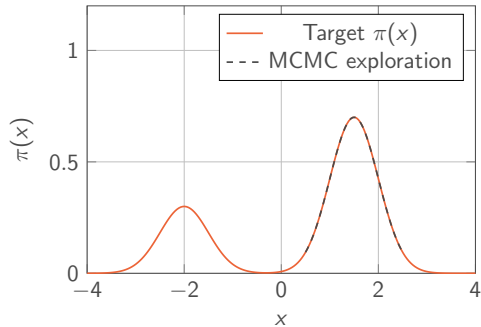
The Challenge: Multimodal Distributions

Standard MCMC Problems:

- ▶ Gets trapped in local modes
- ▶ Exponentially slow mixing times
- ▶ Poor exploration of state space
- ▶ Fails to discover all modes

Common in:

- ▶ Mixture models
- ▶ Bayesian model selection
- ▶ Phase transitions in physics
- ▶ Protein folding simulations



The Temperature Mechanism

Key Idea: Tempered Distributions

Define a family of distributions indexed by inverse temperature

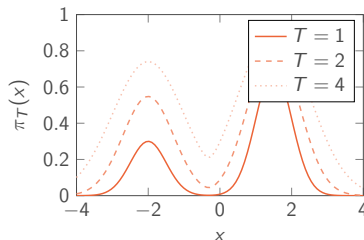
$0 < \gamma_1 < \gamma_2 < \dots < \gamma_N = 1$:

$$\pi_{\gamma_n}(x) \propto \pi(x)^{\gamma_n}$$

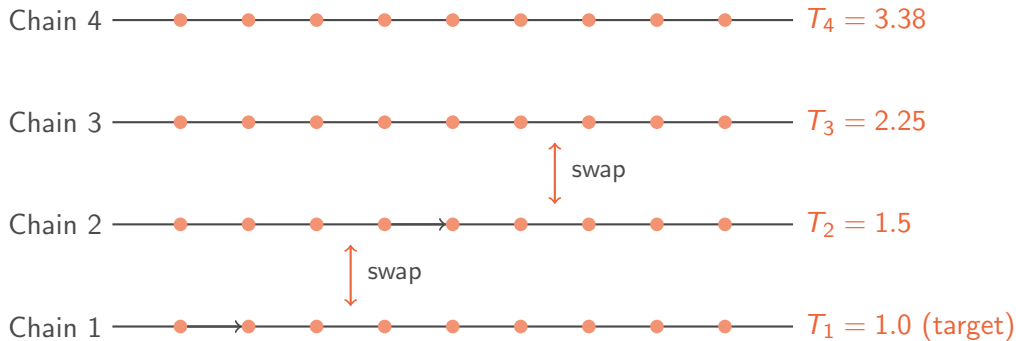
where $n = 1, \dots, N$ and $\pi(x)$ is our target distribution.

Properties:

- ▶ $T = 1$: Original target distribution
- ▶ $T > 1$: "Heated" - flatter distribution
- ▶ $T \rightarrow \infty$: Approaches uniform
- ▶ $T < 1$: "Cooled" - more peaked



Parallel Chains Architecture



Two Types of Moves

1. **Within-chain updates:** Standard MCMC at each temperature
2. **Between-chain swaps:** Exchange states between adjacent temperatures

Parallel Tempering Algorithm Prerequisites

- ▶ **Target Distribution:** $\pi(x)$
- ▶ **Proposal Distribution:** $q(x'|x)$ - could potentially depend on temperature
- ▶ **Initialization:** $x_n^{(0)}$ for $n = 1, \dots, N$
- ▶ **Standard MCMC Step:** Any MCMC kernel (e.g., RWM, MALA)
- ▶ **Number of Chains:** N
- ▶ **Number of Samples per Chain:** T
- ▶ **Swapping Interval:** s
- ▶ **Temperature Schedule:** $\{\gamma_n\}_{n=1}^N$ with $\gamma_N = 1$

Parallel Tempering Algorithm

Algorithm 1 Parallel Tempering MCMC

```
1: for  $t = 1$  to  $T$  do
2:   for all  $n \in \{1, \dots, N\}$  in parallel do
3:     Sample  $x_n^{(t)}$  using a standard MCMC step targeting  $\pi^{\gamma_n}$ 
4:   end for
5:   if  $t \bmod s = 0$  then
6:      $k \sim \text{Uniform}\{1, \dots, N-1\}$ 
7:      $\alpha_{\text{swap}} = \min \left\{ 1, \left( \frac{\pi(x_{k+1}^{(t)})}{\pi(x_k^{(t)})} \right)^{\gamma_k - \gamma_{k+1}} \right\}$ 
8:     Swap  $(x_k^{(t)}, x_{k+1}^{(t)})$  with probability  $\alpha_{\text{swap}}$ 
9:   end if
10: end for
11: return  $\{x_N^{(t)}\}_{t=1}^T$ 
```

Parallel Tempering: Swap Move Acceptance

Setup

- ▶ **Joint target:** $\pi^{\gamma_1} \otimes \pi^{\gamma_2} \otimes \dots \otimes \pi^{\gamma_N}$ where $\gamma_i = 1/T_i$ (inverse temperature)
- ▶ Chain i at state x_i with temperature T_i (cold: $T_N = 1$, hot: $T_1 > 1$)

Metropolis-Hastings Derivation

Propose swap: Exchange states $x_{k_1} \leftrightarrow x_{k_2}$ between chains k_1 and k_2

MH acceptance ratio:

$$\alpha = \frac{\text{Probability of proposed state}}{\text{Probability of current state}} = \frac{\pi^{\gamma_{k_1}}(x_{k_2})\pi^{\gamma_{k_2}}(x_{k_1})}{\pi^{\gamma_{k_1}}(x_{k_1})\pi^{\gamma_{k_2}}(x_{k_2})}$$

Accept with probability: $\min(1, \alpha)$

This ensures detailed balance w.r.t. the joint distribution!

Swap Acceptance Ratio Derivation

Detailed Balance Requirement

For the extended state space with joint distribution $\pi(x_1, \dots, x_K) = \prod_{i=1}^K \pi_i(x_i)$

Consider swapping states between chains i and j :

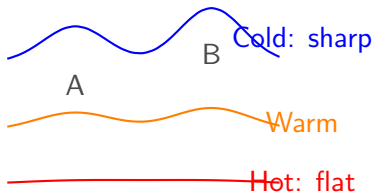
$$\begin{aligned}\alpha_{\text{swap}} &= \min \left\{ 1, \frac{\pi^{\gamma_{k_1}}(x_{k_2}) \cdot \pi^{\gamma_{k_2}}(x_{k_1})}{\pi^{\gamma_{k_1}}(x_{k_1}) \cdot \pi^{\gamma_{k_2}}(x_{k_2})} \right\} \\ &= \min \left\{ 1, \frac{[\pi(x_{k_2})]^{\gamma_{k_1}}}{[\pi(x_{k_2})]^{\gamma_{k_2}}} \cdot \frac{[\pi(x_{k_1})]^{\gamma_{k_2}}}{[\pi(x_{k_1})]^{\gamma_{k_1}}} \right\} \\ &= \min \left\{ 1, [\pi(x_{k_2})]^{\gamma_{k_1} - \gamma_{k_2}} \cdot [\pi(x_{k_1})]^{\gamma_{k_2} - \gamma_{k_1}} \right\} \\ &= \min \left\{ 1, \frac{[\pi(x_{k_2})]^{\gamma_{k_1} - \gamma_{k_2}}}{[\pi(x_{k_1})]^{\gamma_{k_1} - \gamma_{k_2}}} \right\} = \min \left\{ 1, \left(\frac{\pi(x_{k_2})}{\pi(x_{k_1})} \right)^{\gamma_{k_1} - \gamma_{k_2}} \right\}\end{aligned}$$

Note: Normalizing constants cancel out!

The Problem and Solution

The Problem:

- ▶ Standard MCMC gets stuck in local modes
- ▶ Can't explore separated peaks
- ▶ Dilemma: small steps (stuck) vs. large steps (rejected)



The Solution: Temperature Ladder

- ▶ Run N chains targeting π^{γ_n}
- ▶ $0 < \gamma_1 < \dots < \gamma_N = 1$
- ▶ **Hot** ($\gamma \approx 0$): Explores freely
- ▶ **Cold** ($\gamma = 1$): Exploits peaks



Key Insight

Different temperatures see the same distribution differently - hot chains explore, cold

Parallel Tempering MCMC: Normalization Requirements

Target Distribution

NOT required to be normalized

Tempered Distributions

Also NOT normalized

Why It Works: Acceptance Ratios

Within-chain moves:

$$\alpha = \min \left(1, \frac{\pi_i(\mathbf{x}')}{\pi_i(\mathbf{x})} \right)$$

Normalizing constants cancel!

Between-chain swaps:

$$\alpha = \min \left(1, \frac{\pi_i(\mathbf{x}_j)\pi_j(\mathbf{x}_i)}{\pi_i(\mathbf{x}_i)\pi_j(\mathbf{x}_j)} \right)$$

Constants cancel again!

Practical Advantage

Why Swapping Works

The Relay Race Mechanism:

1. **Hot chain** randomly discovers new mode
2. Swap propagates discovery downward
3. **Cold chain** thoroughly explores it
4. Information flows both ways

Swap Acceptance (adjacent chains):

$$\alpha = \min \left\{ 1, \left(\frac{\pi(x_{k+1})}{\pi(x_k)} \right)^{\gamma_k - \gamma_{k+1}} \right\}$$

- Favors moving high-prob states to cold
- Favors moving low-prob states to hot
- Adjacent swaps \rightarrow high acceptance

Example: Two Islands



Without PT: Stuck on Island A forever

With PT: Both islands sampled!

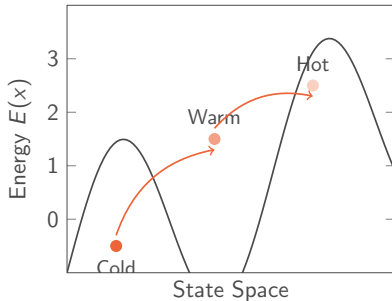
Intuition: Why Swaps Work

Energy Landscape Analogy:

- ▶ Low energy = high probability
- ▶ High temperature = less selective
- ▶ Low temperature = more selective

Swap Success Scenarios:

- ▶ Good state to cold chain
- ▶ Bad state to hot chain
- ▶ Bad state to cold chain
- ▶ Good state to hot chain



Hot chains explore broadly
Cold chains exploit locally

Choosing Temperatures: The Critical Decision

The Temperature Selection Problem

- ▶ Too few temperatures \Rightarrow poor communication between chains
- ▶ Too many temperatures \Rightarrow computational waste
- ▶ Poor spacing \Rightarrow inefficient mixing

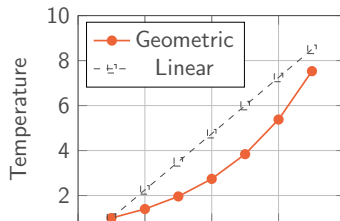
Common Strategies:

1. Geometric spacing:

$$T_i = T_1 \cdot \rho^{i-1}$$

2. Optimal for Gaussians:

$$T_{i+1}/T_i \approx 1 + \sqrt{\frac{2\alpha}{n}}$$



Optimal Number of Temperatures

Theorem (Atchadé et al., 2011)

For a d -dimensional problem, the optimal number of temperatures scales as:

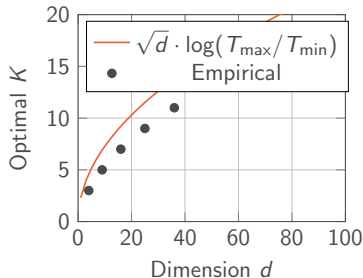
$$K_{opt} \propto \sqrt{d} \cdot \log \left(\frac{T_{\max}}{T_{\min}} \right)$$

Exchange Acceptance Rate:

- ▶ Target: 20-40% (problem-dependent)
- ▶ Kone & Kofke (2005): 23.4% optimal
- ▶ Monitor during runtime

Adaptive Algorithm:

$$\log T_i^{(n+1)} = \log T_i^{(n)} + \gamma_n (\alpha_{i,i+1} - \alpha^*)$$



Detailed Balance and Ergodicity

Proposition (Detailed Balance)

The parallel tempering algorithm satisfies detailed balance with respect to the joint distribution:

$$\pi(x_1, \dots, x_K) = \prod_{i=1}^K \frac{1}{Z_i} \pi(x_i)^{1/T_i}$$

Proof Sketch:

1. Within-chain moves: Standard MCMC detailed balance
2. Swap moves: Show $\pi(\mathbf{x})P(\mathbf{x} \rightarrow \mathbf{x}') = \pi(\mathbf{x}')P(\mathbf{x}' \rightarrow \mathbf{x})$
3. Symmetry of proposal + Metropolis ratio ensures balance

Ergodicity Conditions

- ▶ Each chain must be irreducible
- ▶ Temperature set must include $T = 1$

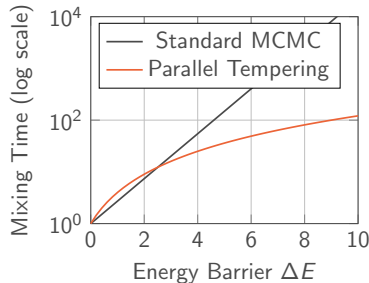
Mixing Time Improvements

Theorem (Woodard et al., 2009)

For certain multimodal distributions, parallel tempering reduces mixing time from exponential to polynomial in problem size.

Example: Double-well potential

- ▶ Standard MCMC: $\tau_{\text{mix}} \sim e^{\beta\Delta E}$
- ▶ Parallel Tempering: $\tau_{\text{mix}} \sim K^2$
- ▶ Where K = number of temperatures

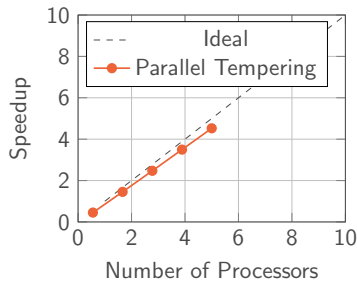


Exponential \rightarrow Polynomial speedup!

Implementation Considerations

Computational Aspects:

- ▶ **Parallelization:** Natural parallelism across chains
- ▶ **Communication:** Minimal (only for swaps)
- ▶ **Memory:** Linear in number of chains
- ▶ **Scaling:** Near-linear with processors



Software Packages:

- ▶ emcee (Python) - adaptive PT
- ▶ PyMC3 - Bayesian modeling
- ▶ PLUMED - molecular dynamics
- ▶ MCMCpack (R) - general purpose

Communication Pattern:



Diagnostics and Convergence

Key Diagnostics:

1. Exchange acceptance rates

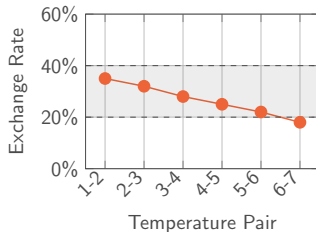
- ▶ Monitor between all pairs
- ▶ Target: 20-40%

2. Round-trip times

- ▶ Time for state to visit all temperatures
- ▶ Should be finite and reasonable

3. Temperature diffusion

- ▶ States should visit all temperatures
- ▶ Check histogram of visits



Convergence Criteria:

- ▶ Standard \hat{R} statistic
- ▶ Effective sample size (ESS)
- ▶ KL divergence between chains

Modern Extensions

1. Non-reversible PT

- ▶ Syed et al. (2022)
- ▶ Persistent direction of swaps
- ▶ Further reduces mixing time

2. Infinite Swapping

- ▶ Plattner et al. (2011)
- ▶ Continuous-time limit
- ▶ Optimal temperature schedules

3. PT with Normalizing Flows

- ▶ Learn optimal proposal distributions
- ▶ Adaptive temperature mappings
- ▶ Neural network augmentation

4. Simulated Tempering vs PT

Aspect	PT	ST
Chains	Multiple	Single
Memory	$O(K)$	$O(1)$
Parallel	Yes	No
Tuning	Easier	Harder

5. PT-based Model Selection

- ▶ Thermodynamic integration
- ▶ Model evidence estimation
- ▶ Bayes factor computation

Limitations and When Not to Use PT

Limitations:

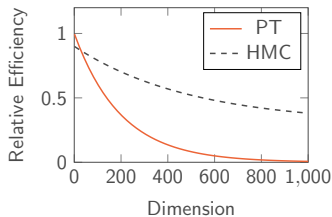
- ▶ Computational cost scales with K
- ▶ Memory requirements $\propto K \times d$
- ▶ Temperature tuning can be difficult
- ▶ Less effective in very high dimensions

When PT Struggles:

- ▶ Modes separated by vast low-probability regions
- ▶ Dimension $d > 1000$
- ▶ When modes have very different scales
- ▶ Real-time applications

Alternatives to Consider:

- ▶ **SMC**: For sequential problems
- ▶ **HMC**: For smooth, high-dim targets
- ▶ **Variational Inference**: When approximate is OK
- ▶ **Annealed Importance Sampling**: For evidence estimation



Key Takeaways

Strengths:

- ▶ Excellent for multimodal distributions
- ▶ Naturally parallel
- ▶ Theoretically rigorous
- ▶ Automatic diagnostics via exchange rates
- ▶ Wide applicability

Key Design Choices:

- ▶ Number of temperatures: $O(\sqrt{d})$
- ▶ Spacing: Geometric or adaptive
- ▶ Target exchange rate: 20-40%

Remember:

- ▶ Temperature = "exploration parameter"
- ▶ Hot chains explore, cold chains exploit
- ▶ Swaps enable global communication
- ▶ Detailed balance is preserved

The PT Philosophy

"Heat to explore, cool to exploit, swap to communicate"

Active Research Areas:

- ▶ Optimal temperature schedules

References

► Foundational:

- Geyer (1991). "Markov chain Monte Carlo maximum likelihood"
- Hukushima & Nemoto (1996). "Exchange Monte Carlo method"

► Reviews:

- Earl & Deem (2005). "Parallel tempering: Theory, applications, and new perspectives"
- Swendsen & Wang (2016). "Replica Monte Carlo simulation (revisited)"

► Theory:

- Atchadé et al. (2011). "Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo"
- Woodard et al. (2009). "Conditions for rapid mixing of parallel and simulated tempering"

► Recent Advances:

- Syed et al. (2022). "Non-reversible parallel tempering: A scalable highly parallel MCMC scheme"
- Vousden et al. (2016). "Dynamic temperature selection for parallel tempering"

Questions?

Thank you!