Continuous Tempering Through Path Sampling

Yuling Yao and Andrew Gelman

Correspond to Yuling Yao: yy2619@columbia.com
Department of Statistics, Columbia University, New York.

Main Results

Simulated tempering is aimed to sample from a multimodal distribution where the target density contains metastable regions separated by high energy barriers. We present a new tempering algorithm by allowing temperature moves continuously. **Path sampling** gives a low-variance estimation of normalizing constant, making the algorithm scalable to high dimensions. The adaptive procedure biases the partition function to shrink the gap between the proposal and the target.

Simulated tempering

Markov chain Monte Carlo is widely used for Bayesian computation. The central task to sample from a posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$, where $\boldsymbol{\theta} \in \Theta$ is unknown parameters.
Regardless of the theoretic guarantee of convergence, multi-modal distribution is hard to sample in finite time because of the energy barriers between modes.
(Discrete) simulated tempering expands the state space by an inverse temperature λ, and λ is restricted on a prespecified discrete grid

$$\lambda \in (0 = \lambda_1 < \lambda_2 < \cdots < \lambda_K = 1)\,.$$

The new joint density can be defined as a geometric bridge

$$p(\theta, \lambda_k) = \frac{1}{z(\lambda)} p(\theta)^{\lambda_k} \phi(\theta)^{1-\lambda_k}.$$

Discrete tempering is not desired

One challenge of simulated tempering is to estimate normalizing constant $z(\lambda_k)$, which is unknown but determines the joint distribution. We can start from an initial guess and adaptively update $z(\lambda_k)$ by **importance sampling**:

$$\hat{z}(\lambda_k) \leftarrow \hat{z}(\lambda_k)\hat{p}(\lambda_k)/c(\lambda_k)$$

where $c(\lambda_k)$ is the pre-specified marginal distribution.
However, $z(\lambda_k)$ can change dramatically for λ_k in order of magnitude. The initial guess can be far away from the true value, leaving some λ_k rarely sampled. Particularly, the variance of importance sampling grows exponentially.
Applying **Rao-Blackwellized** strategy to the identity

$$p(\lambda) = E_\theta\big[p(\lambda|\theta)\big]$$

yields a lower-variance estimation

$$\hat{p}_{RB}(\lambda = \lambda_k) \propto \frac{1}{n}\sum_{i=1}^{n} p(\lambda_k|\theta_i).$$

But the smooth transition of β needs

$$\mathrm{KL}\Big(\pi_\lambda, \pi_{\lambda+\delta\lambda}\Big) \approxeq \text{constant}$$

The optimal tuning of spacing is infeasible without knowing $z(\lambda)$. Thus, we will expect failure of IS if $z(\lambda)$ changes rapidly.
Another justification: under normal approximation, the theoretical requirement for K grows exponentially as $\dim(\Theta)$ grows. It is the same reason for inefficiency of importance sampling in high dimensions.

Path sampling

We define a link function $f : [0, 2] \to [0, 1]$ such that $f(a)$ is flat near $a = 0$ and 1, hence the base and target distribution can be obtained directly from the joint samples.

Figure 1: The link function $\lambda = f(a)$

Consider the geometry bridge between target $\phi(\theta)$ and the base distribution $\psi(\theta)$:

$$p(\theta|a) = \frac{1}{z(a)}\phi(\theta)^{f(a)}\psi(\theta)^{1-f(a)}.$$

where $\psi(\theta)$ is a known base distribution. By construction, normalization constant

$$z(a) = \int \psi(\theta)^{f(a)}\phi(\theta)^{1-f(a)}\mu(d\theta).$$

satisfies $z(0) = 1$.
Path sampling is based on the identity

$$\frac{d}{da}\log z(a) = E_a\left[\frac{d}{da}\log q(\theta|a)\right] \qquad (1)$$

which dose not depends on the the prior distribution $c(a)$.
Summary of proposed algorithm:

- Sample from the extended joint $p(\theta, a) = \frac{1}{c(a)}\phi(\theta)^{f(a)}\psi(\theta)^{1-f(a)}$;
- Estimate $\log z(a)$ based on numerical integration of (1);
- Update $c(a) \leftarrow z(a)$ adaptively;
- Repeat until the marignal of a is uniform.

We next compare the results with other tempering methods and show the proposed approach has a quicker convergence rate.

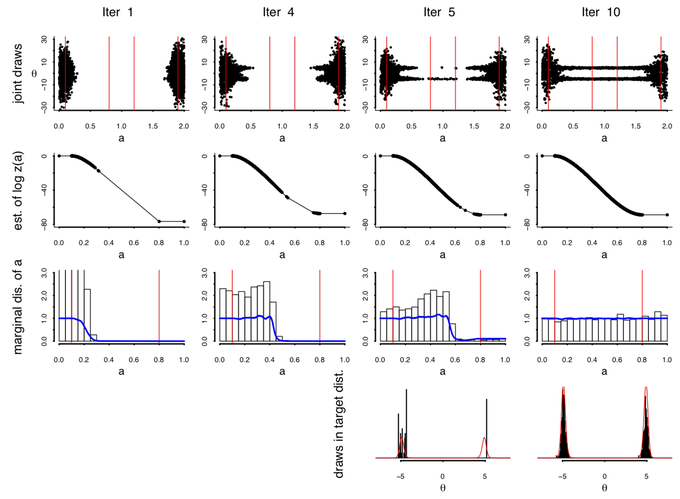

Figure 2: *Results of the continuous tempering with path sampling in a Cauchy example. The typical set is fully explored after iteration 5, and the marginal of a is nearly uniform after iteration 10.*

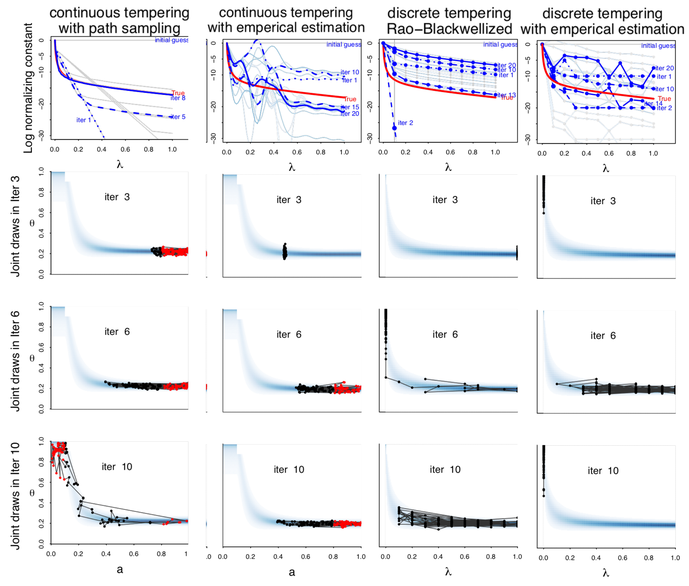

Figure 3: *Comparison of four tempering methods in a beta-binomial experiment. Starting with a uniform guess, only CPT converges to the the true value after 8 iterations, and fully explores the typical set efficiently with HMC jumps in the joint space.*