# Lecture Notes for STAT546: Computational Statistics

## —Lecture 7: Monte Carlo

Faming Liang

Purdue University

September 11, 2024

# Sequential parallel tempering

With the development of science and technology, we more and more need to deal with high dimensional systems. For example, we need to align a group of protein or DNA sequences to infer their homology, identify a single-nucleotide polymorphism (SNP) associated with certain disease from millions of SNPs, estimate the volatility of asset returns to understand the price trend of the option market, simulate from spin systems to understand their physical properties, and etc. In these problems, the dimensions of the systems often range from several hundreds to several thousands or even higher, and the solution space are so huge that Monte Carlo has been an indispensable tool for making inference for them. How to efficiently sample from these high dimensional systems puts a great challenge on the existing MCMC methods.

For many problems the slow convergence is not due to the multimodality, but the curse of dimensionality; that is, the number of samples increase exponentially with dimension to maintain a given level of accuracy. For example, the witch's hat distribution (Matthews, 1993) has only a single mode, but the convergence time of the Gibbs sampler on it increases exponentially with dimension. For this kind of problems, although the convergence can be improved by the tempering-based or importance weight-based algorithms to some extent, the curse of dimensionality cannot be much reduced, as these samplers always work in the same sample space.

To eliminate the curse of dimensionality, Liang (2003) provided a sequential parallel tempering (SPT) algorithm, which makes use of the sequential structure of high dimensional systems. As an extension of parallel tempering, SPT works by simulating from a sequence of systems of different dimensions. The idea is to use the information provided by the simulation of low dimensional systems as a clue for simulating high dimensional systems.

A buildup ladder (Wong and Liang, 1997) comprises a sequence of systems of different dimensions. Consider $m$ systems with density $f_i(x_i)$, where $x_i \in \mathcal{X}_i$ for $i = 1, \ldots, m$. Typically,

$$dim(\mathcal{X}_1) < dim(\mathcal{X}_2) < \cdots < dim(\mathcal{X}_m),$$

The principle of buildup ladder construction is to approximate the original system by a system with a reduced dimension, the reduced system is again approximated by a system with a further reduced dimension, until a system of a manageable dimension is reached; that is, the corresponding system can be easily sampled using a local updating algorithm, such as the MH algorithm or the Gibbs sampler. The solution of the reduced system is then extrapolated level by level until the target system is reached. For many problems, the buildup ladder can be constructed in a simple way. For example, for both the traveling salesman problem (Wong and Liang, 1997) and the phylogenetic tree reconstruction problem (Cheon and Liang, 2008), the build-up ladders were constructed by the method of marginalization.

# *Sequential Parallel Tempering*

1. *Local Updating*: Update each $x_i$ independently by a local updating algorithm for a few steps.
2. *Between-level transition*: Try between-level transitions for $N$ pairs of neighboring levels, with $i$ being sampled uniformly on $\{1, 2, \cdots, N\}$ and $j = i \pm 1$ with probability $q_e(i, j)$, where $q_e(i, i+1) = q_e(i, i-1) = 0.5$ for $1 < i < N$ and $q_e(1, 2) = q_e(N, N-1) = 1$.

The between-level transition involves two operations, namely, projection and extrapolation. Two different levels, say, $i$ and $j$, are proposed to make the between level transition. Without loss of generality, we assume that $\mathcal{X}_i \subset \mathcal{X}_j$. The transition is to extrapolate $x_i$ ($\in \mathcal{X}_i$) to $x_j'$ ($\in \mathcal{X}_j$), and simultaneously project $x_j$ ($\in \mathcal{X}_j$) to $x_i'$ ($\in \mathcal{X}_i$). The extrapolation and projection operators are chosen such that the pairwise move $(x_i, x_j)$ to $(x_i', x_j')$ is reversible. The transition is accepted with probability

$$\min\left\{1, \frac{f_i(x_i')f_j(x_j')}{f_i(x_i)f_j(x_j)} \frac{T_e(x_i' \to x_j)\,T_p(x_j' \to x_i)}{T_e(x_i \to x_j')\,T_p(x_j \to x_i')}\right\}, \qquad (1)$$

where $T_e(\cdot \to \cdot)$ and $T_p(\cdot \to \cdot)$ denote the extrapolation and projection probabilities, respectively. For simplicity, the between level transitions are only restricted to neighboring levels, i.e., $|i - j| = 1$.

# Witch's Hat distribution

The witch's hat distribution has the density

$$\pi_d(x) = (1-\delta)\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^d \exp\left\{-\frac{\sum_{i=1}^d(\tilde{x}_i-\theta_i)^2}{2\sigma^2}\right\} + \delta I_{x\in C},$$

where $x = (\tilde{x}_1, \ldots, \tilde{x}_d)$, $d$ is dimension, $C$ denotes the open $d$-dimensional hypercube $(0,1)^d$, and $\delta$, $\sigma$ and $\theta_i$'s are known constants. In the case of $d = 2$, the density shapes like a witch's hat with a broad flat brim and a high conical peak, so the distribution is called the witch's hat distribution. This distribution is constructed by Matthews (1993) as a counterexample to the Gibbs sampler, on which the mixing time of the Gibbs sampler increases exponentially with dimension.

As a test, Liang (2003) applied to SPT to simulate from $\pi_d(x)$, with $d = 5, 6, \ldots, 15$, $\delta = 0.05$, $\sigma = 0.05$, and $\theta_1 = \cdots = \theta_d = 0.5$. For each value of $d$, the build-up ladder was constructed by setting $f_i(x) = \pi_i(x)$ for $i = 1, 2, \cdots, d$, where $\pi_i(\cdot)$ is the $i$-dimensional witch's hat distribution, which has the same parameter as $\pi_d(\cdot)$ except for the dimension. In the local updating step, each $x_i$ is updated iteratively by the MH algorithm for $i$ steps. At each MH step, one coordinate is randomly selected and then proposed to be replaced by a random draw from uniform(0,1). The between-level transition, say, between level $i$ and level $i + 1$, proceeds as follows:

Table 1: Comparison of SPT and parallel tempering for the witch's hat distribution (Liang, 2003).

| | SPT | | | PT | | |
|---|---|---|---|---|---|---|
| $d^1$ | Time$^2$(s) | $\hat{\alpha}$ | SD($\times 10^{-4}$) | Time$^2$(s) | $\hat{\alpha}$ | SD($\times 10^{-4}$) |
| 5 | 72.5 | 0.6546 | 8.5 | 58.8 | 0.6529 | 9.7 |
| 6 | 94.9 | 0.6540 | 9.1 | 84.7 | 0.6530 | 10.5 |
| 7 | 118.6 | 0.6541 | 9.2 | 115.6 | 0.6525 | 11.2 |
| 8 | 145.8 | 0.6530 | 9.3 | 152.4 | 0.6530 | 13.2 |
| 9 | 174.6 | 0.6534 | 9.2 | 190.8 | 0.6538 | 15.8 |
| 10 | 206.0 | 0.6533 | 9.4 | 236.7 | 0.6517 | 20.5 |
| 11 | 239.3 | 0.6528 | 9.3 | 711.7 | 0.6531 | 17.7 |
| 12 | 275.5 | 0.6525 | 9.9 | 847.7 | 0.6530 | 21.3 |
| 13 | 312.9 | 0.6532 | 9.7 | 996.1 | 0.6527 | 33.8 |
| 14 | 353.7 | 0.6531 | 10.0 | 1156.4 | 0.6506 | 47.5 |
| 15 | 397.4 | 0.6532 | 10.4 | 1338.0 | 0.6450 | 84.5 |

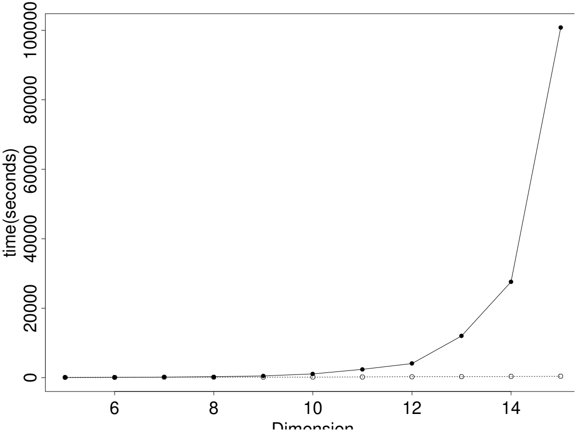

Figure 1: Estimated running times $T(SPT, d)$ (solid line) and $T(PT, d)$ (dotted line) for $d = 5, 6, \ldots, 15$. (Liang, 2003)

- Extrapolation: draw $u \sim \mathit{unif}(0,1)$ and set $x'_{i+1} = (x_i, u)$.
- Projection: set $x'_i$ to be the first $i$ coordinates of $x_{i+1}$.

The corresponding extrapolation and projection probabilities are $T_e(\cdot \to \cdot) = T_p(\cdot \to \cdot) = 1$.

# Equi-energy sampler (Kou, Zhou and Wong, 2006)

Let $f(x) \propto \exp\{-H(x)\}$, $x \in \mathcal{X}$, denote the target distribution, where $H(x)$ is called the energy function. Let $E_0 < E_1 < \cdots < E_N < \infty$ denote a ladder of energy levels, and let $T_0 < T_1 < \cdots < T_N < \infty$ denote a ladder of temperatures, where $E_0 = -\infty$ and $T_0 = 1$. Based on the two ladders, the equi-energy sampler defines the trial distributions of the population by

$$f_i(x) \propto \exp\{-\max(H(x), E_i)/T_i\}, \quad i = 0, 1, 2 \ldots, N. \qquad (2)$$

Thus, $f_0(x)$ corresponds to the target distribution. Define $D_i = \{x : H(x) \in [E_i, E_{i+1})\}$ for $i = 0, \ldots, N$, where $E_{N+1} = \infty$. Thus, $D_0, D_1, \ldots, D_N$ form a partition of the sample space. Each $D_i$ is called an energy ring associated with the energy ladder. Let $I(x)$ denote the index of the energy ring that $x$ belong to; that is, $I(x) = k$ if and only if $H(x) \in [E_k, E_{k+1})$. The equi-energy sampler consists of $N + 1$ stages:

1. The equi-energy sampler begins with a MH chain $\{X_N\}$, which targets the highest order distribution $f_N(x)$. After a burn-in period of $B$ steps, the sampler starts to group its samples into energy rings $\widehat{D}_i^{(N)}$, $i = 0, \ldots, N$, where the sample $X_{N,j}$ is grouped into $\widehat{D}_i^{(N)}$ if $I(X_{N,j}) = i$.
2. After a period of $M$ steps, the equi-energy sampler starts to construct a parallel MH chain $\{X_{N-1}\}$, which targets the second highest order distribution $f_{N-1}(x)$, while keeping the chain $\{X_N\}$ running and updating. The chain $\{X_{N-1}\}$ is updated with two types of moves, local MH move and equi-energy jump, with respective probabilities $1 - p_{ee}$ and $p_{ee}$.

- Local MH move: Update the chain $\{X_{N-1}\}$ with a single MH transition.
- Equi-energy jump: Let $x_{N-1,t}$ denote the current state of the chain $\{X_{N-1}\}$, and let $k = I(x_{N-1,t})$. Choose $y$ uniformly from the energy ring $hD_k^{(N)}$, and accept $y$ as the next state of the chain $\{X_{N-1}\}$ with probability

  $$\min\left\{1, \frac{f_{N-1}(y)f_N(x_{N-1,t})}{f_{N-1}(x_{N-1,t})f_N(y)}\right\}.$$

  Otherwise, set $x_{N-1,t+1} = x_{N-1,t}$.

After a burn-in period of $B$ steps, the equi-energy sampler starts to group its samples into different energy rings $\widehat{D}_i^{(N-1)}$, $i = 0, 1, \ldots, N$.

3. After a period of another $M$ steps, the equi-energy sampler starts to construct another parallel MH chain $\{X_{N-2}\}$, which targets the third highest order distribution $f_{N-2}(x)$, while keeping the chains $\{X_N\}$ and $\{X_{N-1}\}$ running and updating. As for the chain $\{X_{N-1}\}$, the chain $\{X_{N-2}\}$ is updated with the local MH move and the equi-energy jump with respective probabilities $1 - p_{ee}$ and $p_{ee}$.

- ......

$N + 1$. Repeat the above procedure until the level 0 has been reached. Simulation at level 0 results in the energy rings $\widehat{D}_i^{(0)}$, which deposit samples for the target distribution $f(x)$.

Under the assumptions (i) the highest order chain $\{X_N\}$ is irreducible and aperiodic, (ii) for $i = 0, 1, \ldots, N-1$, the MH transition kernel $K_i$ of $\{X_i\}$ connects adjacent energy rings in the sense that for any $j$, there exist sets $A_1 \subset D_j$, $A_2 \subset D_j$, $B_1 \subset D_{j-1}$ and $B_2 \subset D_{j+1}$ with positive measure such that the transition probabilities $K_i(A_1, B_1) > 0$ and $K_i(A_2, B_2) > 0$, and (iii) the energy ring probabilities $p_{ij} = P_{f_i}(X \in D_j) > 0$ for all $i$ and $j$, Kou, Zhou and Wong (2006) showed that each chain $\{X_i\}$ is ergodic with $f_i$ as its invariant distribution.

It is clear that the equi-energy sampler stems from parallel tempering, but different from parallel tempering in two respects. The first difference is on the exchange operation, which is called the equi-energy jump in the equi-energy sampler. In parallel tempering, the current state of a lower order Markov chain is proposed to be exchanged with the current state of a higher order Markov chain. While, in the equi-energy sampler, the current state of a lower order Markov chain is proposed to be replaced by a past, energy-similar state of a higher order Markov chain. Since the two states have similar energy values, the equi-energy jump has usually a high acceptance rate, and this increases the interaction between different chains. The second difference is on distribution tempering.

Parallel tempering gradually tempers the target distribution, while the equi-energy sampler tempers a sequence of low-energy truncated distributions. It is apparent that simulation of the low-energy truncated distribution reduces the chance of getting trapped in local energy minima. As demonstrated by Kou, Zhou and Wong (2006), the equi-energy sampler is more efficient than parallel tempering. For example, for the multimodal example studied in EMC, the equi-energy sampler can sample all the modes within a reasonable CPU time, while parallel tempering fails to do so.