Torpid Mixing of Simulated Tempering on the Potts Model
by Bhatnagar and Randall

Presentation Notes by John Turner and John Stewart

I. Abstract and Intro

The purpose of this paper is two fold. First, it proves that temperature-based sampling algorithms (specifically simulated tempering and swapping), which have successfully been used to sample bimodal distributions such as the low-temperature mean-field Ising model, can fail to converge rapidly on the more general Potts model. This is due to the nature of the phase change with three states or greater. Secondly, it presents a variant of the swapping algorithm, the Flat-Swap algorithm, and proves that it is rapidly mixing for any bimodal mean-field model.

II. Model

The primary model used in the paper is the 3-state mean-field Potts model, where mean-field denotes that the underlying graph is complete. Each vertex in this graph represents particles on a lattice with a particular spin, represented by a color, and each edge between vertices represents mutual influence between two particles on their respective spin values. The mapping of a spin assignment for every particle (a color for each vertex) is called a configuration, where 3-state denotes that there are three different spins. The energy of a particular configuration determines that configuration's likelihood, and is defined as a function of the similarity of colors of all edge-connected pairs of vertices in the graph. This energy is used with a term encoding the inverse of the temperature as an exponent to derive the Gibbs/Boltzmann's distribution which relates the state energy with that state's probability.

$$H(\sigma) = \sum_{(i,j)\in E(G)} J \cdot \delta(q_i, q_j), \quad \pi_\beta(\sigma) = \frac{e^{\beta H(\sigma)}}{Z(\beta)},$$

Drawing 1: Potts Model Distribution

III. Algorithms

The **Metropolis-Hastings algorithm** is considered useful when sampling from non-uniform distributions. A graph (the Markov kernel) is derived that connects the state space of the underlying model, where every vertex is a state in that state space, and every edge is a possible single-step transition between adjacent states and is assigned some non-zero probability proportional to the ratio of the probabilities of being in either state.

In the case of the mean-field Potts model, the Markov kernel is built such that every vertex represents a configuration of the Potts model, while every edge is a single-step transition between two configurations that differ at only one location (in other words, in each vertex of an edge-connected pair in the Markov kernel, only a single vertex of the underlying Potts model configuration differs in color).

Despite this algorithm being a common approach, it has been proven for some models that the existence of a bad cut in the underlying model can cause it to converge exponentially slowly to an appropriate stationary distribution. The Potts model is known to be one where Metropolis-Hastings converges slowly. This is due to the Potts model's preference for monochromatic configurations at low temperature, which conflicts with the chain's need to move through very unlikely configurations (where all colors are represented somewhat equally) to cover the state space (to transition from a configuration that is predominantly one color, to another configuration that is predominantly another color, configurations of no predominant color need to be transitioned through, and these are exponentially unlikely to arise).

Furthermore, algorithms that have been shown to work efficiently in overcoming this bottleneck for two-state models (i.e. the Ising model), such as Simulated Tempering and Swapping, are proven in this paper (for the first time) to not be able to reliably mix in polynomial time for three-state models. This is due to the nature of the phase-transition exhibited by all Potts models – for two-state models the phase transition is second order (and continuous in the derivative of the energy function), but for for three-state models, the phase-transition is first-order and discontinuous.

Temperature Algorithms

Temperature-based algorithms, specifically tempering and swapping, are algorithms which modify the temperature (used to derive the Gibbs distribution) of the Potts Model in order to transition among different distributions in inverse-temperature space. Whereas the Metropolis-Hastings algorithm will transition through the state space at a fixed temperature, these algorithms will extend such state spaces across a series of temperatures.

They have been shown to work well with the mean field Ising model, and were considered natural choices to try with higher order Potts models.

Tempering

The **simulated tempering** algorithm introduces a set of *M+1* inverse temperature values β, and the state space of the tempering chain is then defined to be the union of all *M+1* original state spaces derived for each inverse temperature, where $\beta_0 = 0$ is infinite temperature and a uniform stationary distribution, and β_m represents the inverse temperature of the distribution we wish to sample. The algorithm then interpolates these inverse temperatures geometrically (linear in the log) to get all the *M-1* remaining values of temperature. It uses each temperature to derive each distribution, which provides the stationary distribution for the tempering chain.

The chain has two types of moves : fixed-temperature **level moves** (where a single step transition between configurations is made at a fixed temperature), and fixed-configuration **temperature moves**, which links two configurations at neighboring temperatures. The tempering algorithm calculates the ratio of distributions at different temperatures in order to derive the probability of transitioning to a new temperature. This involves summing the distribution over exponentially many configurations at a fixed temperature, in order to find the normalizing factor. This is a computational cost that is avoided by using the swapping algorithm.

Swapping

The **swapping** algorithm is a variant of tempering, with the state space being defined instead as the product of *M+1* versions of the original Markov chain, where each chain corresponds to one of the *M+1* inverse temperature values. A configuration in the swapping chain is then defined as the mapping from each of the *M+1* temperatures to a particular configuration, so that there would be a Markov chain configuration for each temperature. The stationary distribution is then the product of the stationary distributions across all inverse temperatures.

The swapping chain also has two moves, a **level move** and a **swap move**. The level move connects two swapping chain configurations when they differ by all but a one-step Metropolis-driven transition in a single component. The swap move interchanges neighboring configurations in inverse-temperature space. The normalizing constants required to calculate these transition probabilities cancel out, rendering moves of the swapping chain simpler to calculate than for the tempering chain.

The number of temperatures needs to be chosen with care. It must be large enough so that distributions that neighbor each other in inverse-temperature space have sufficiently low variation distance that temperature/swap moves are accepted reasonably often, while it must be small enough so that it doesn't blow up the running time. The paper advises for M to be on the order of the number of vertices in the underlying model, to ensure that the ratio of neighboring inverse-temperature distributions is bounded above and below by a constant.

IV. Proof of Slow Mixing

The paper uses the spectral gap to provide lower bounds for the mixing time of the tempering chain on the mean-field Potts model. In turn, it uses the conductance of the model to bound the spectral gap, demonstrating that a bad cut in the state space is sufficient to show torpid mixing.

In order to bound the conductance and prove that the the 3-state Potts model mixes slowly, the cut used is a consequence of the phase transition, and separates high-temperature, high-entropy states where each color is relatively equally represented, and low-temperature states where a single color will dominate. Unlike the Ising model, where the change from predominantly chaotic states to predominantly ordered ones is gradual, the Potts model exhibits an abrupt change in the size of the largest color class as the temperature fluctuates near some critical temperature. Despite interpolating the state space across many temperatures, this discontinuity is shown to represent a bad cut in all temperatures and subsequently the mixing of the tempering chain is shown to be

slow.

This is shown by partitioning the state space into 3 different sets, where
a) all 3 colors are represented equally, (chaotic states)
b) one color is present on as many vertices as the sum of the other two, (transition states)
c) one color is present on twice as many vertices as the sum of the other two. (ordered states)

It is demonstrated that there is a critical temperature where (a) and (c) have substantial and relatively similar probabilistic weight while (b)'s weight is very small in proportion to the weight of (a). It is then shown that (b) is proportionally smaller than (a) or (c) at every temperature, proving that (b) describes a bad cut across inverse-temperature space.

Previous work shows that slow mixing of tempering implies slow mixing of the given swapping algorithm.

V. Examining Performance on Bimodal Distributions

The second contribution of the paper is a variant of the swapping algorithm intended to enable rapid mixing on bimodal distributions. This is a modification of the interpolation between inverse-temperature-space distributions so that the interpolation does not preserve bad cuts. Two theorems, Comparison and Decomposition, are used to prove fast mixing.

Comparison theorem is useful to bound mixing times of chains when the mixing time of similar chains is known. This method involves providing a canonical path in the original chain for every transition in the similar chain. When applied, it bounds the spectral gap of the unknown chain (and thus its mixing time) by a polynomial multiple of the spectral gap of the known chain, where this multiple is derived from the probability of the canonical paths.

Decomposition is used to break down a complicated Markov chain into more easily analyzed pieces by first separating the state space into disjoint partitions and then defining Markov chains on each partition whose transition matrix has **restrictions** proscribing intra-partition (within the same partition) transitions and **projections** proscribing inter-partition (partition-to-partition) transitions.

These theorems are applied to the swapping chain on the bimodal exponential distribution to prove that it is rapidly mixing. The **trace** function is defined to be a measure of the sign of each configuration in the swapping chain, so it is represented by a binary bit-string of *M+1* values. Then the restrictions simulate the swapping chain acting on regions of fixed trace, while the projections represent the transitions between configurations with different traces.

$$\pi(x) = \pi_C(x) = \frac{C^{|x|}}{Z}, \qquad x \in [-N, N],$$

Drawing 2: Bimodal Exponential Distribution

The restricted chain is then shown to mix rapidly. The spectral gap of the entire restricted chain with a particular trace is bounded by the minimum spectral gap of each temperature component in that chain. The state space of the partition restricted to any of the *M+1* temperatures is unimodal (since the trace restricts it to either the positive or negative half). This then shows that the restriction chain for that temperature is rapidly mixing, and by extension, the restricted chain for the entire partition is rapidly mixing.

The projection chain is shown to be rapidly mixing by first drawing comparison with a simpler random walk on an *M+1* hypercube where each move allows for either the transposition of two neighboring bits or the flipping of the lowest bit. These describe a restriction of the behavior of the projection chain which allow it to be more easily decomposed.

By picking an even simpler chain and showing that it mixes rapidly, comparison is used to show that the projection chain is rapidly mixing. At each step in this simpler chain, a single bit of the trace is chosen and updated according to the stationary distribution. This simpler chain's transitions are then translated into a canonical path of moves in the original projection chain. This path consists of iteratively swapping the chosen bit to the lowest position, inverting the bit, and then swapping it back to its original position. This proves both that the transition probabilities for any transitions in the canonical path are bounded by the transitions in the simpler chain, and that the number of paths using any particular transition is at most polynomial. The comparison theorem shows that the swapping chain acting on the bimodal exponential distribution is rapidly mixing.

Using the previous results, the mean field model is then examined for a way to modify the swapping algorithm so that it is rapidly mixing. Two special cases of bi-modal mean-field spin models are given to set the

groundwork for the flat-swap algorithm. Instead of using the typical temperature interpolations, the flat-swap algorithm uses a multiplicative function based on the particular inverse-temperature. The choice of function acts to interpolate to the distribution that is uniform with respect to the total spin distribution, which eliminates the bad cut that otherwise occurs with a constant interpolating function.

Using the results from the chain acting on the bimodal exponential distribution, the Flat-Swap algorithm is shown to be rapidly mixing for any bimodal mean-field model. In effect, the spin distributions are shown to retain the same shape (including the same maxima and minima) as the desired distribution, but get "flatter" as temperature is changed, dampening the change in entropy that would otherwise be a discontinuous jump.

$$\rho_i(x) = \frac{\pi_i(x) f_i(x)}{Z_i'}, f_i(x) = \binom{n}{\sigma_1, \ldots, \sigma_q}^{\frac{i-M}{M}}$$

Drawing 3: Flat-Swap Distribution