Torpid Mixing of Simulated Tempering on the Potts Model

Nayantara Bhatnagar
Dana Randall

Presentation by John Stewart and John Turner

Temperature Algorithms And Bimodal Distributions

Simulated Tempering and Swapping are sampling algorithms similar to simulated annealing - vary a temperature variable in the hopes of avoiding bad cuts that are present in the state space of bimodal distributions.

Shown to work efficiently on the mean-field (complete graph) Ising model. (Madras,Zheng; 2003), where purely local MC's fail.

What about other models?

Q-state mean-field Potts Model is generalization of Ising model (where q = 2). Natural to consider these algorithms for other mean-field Potts Models where q >2.

Paper shows that for q >= 3 they will not mix rapidly due to the nature of phase transition.

- q=2 Potts (Ising) has 2nd order phase transition (continuous in derivative of energy)
- q>=3 has 1st order phase transition (discontinuous in derivative)

Paper also provides a modified Swapping algorithm that provably samples efficiently from the mean-field Potts model.

Potts Model

- Q-state mean-field Potts Model
  - Q : # of particle spins = # of vertex colors
  - Edges connect particles that affect each other
  - Mean-field : complete graph

- Configuration σ is assignment of colors to ever vertex.

- Energy of configuration is function of Hamiltonian

$$H(\sigma) = \sum_{(i,j)\in E(G)} J \cdot \delta(q_i, q_j)$$

where $\delta(q_i, q_j)$ is 1 if $q_i = q_j$ and 0 otherwise. High energy implies a high level of monocromaticity

Potts Model 2

- State space Ω is the space of all q^n colorings.
- Inverse Temperature : $\beta = 1/(kT)$
- Gibbs Distribution (probability of a particular configuration) :

$$\pi_\beta(\sigma) = \frac{e^{\beta H(\sigma)}}{Z(\beta)}$$

- Partition function : $Z(\beta) = \Sigma\ e^{\beta H(\sigma)}$; for all $\sigma \in \Omega$
- Paper uses q = 3

Markov Chains

- Ergodic, reversible, finite state space
- Metropolis Hastings - define Markov Kernel as a graph that
  - Connects Ω
  - Vertices are configurations and edges are 1-step transitions.
- Potts Model $\rightarrow$ use Hamming distance of 1.
  - Metropolis then converges slowly because most probable states are monochromatic, and to go from one color dominant to another need to pass through exponentially unlikely “transition” states.
- Temperature Chains use temperature moves to try to move around bad cuts.

Simulated Tempering

- Expanded State Space $\widehat{\Omega}$ to include M+1 different inverse temperatures:
  - $\widehat{\Omega}$ = union of M+1 copies of Ω, for each inverse temperature
  - $\beta_i = \beta_M * i/M$; $\beta_0 = 0$, β_M corresponds to desired distribution.
- Chain configuration : (x, i) : i → index of β
- Conditional distributions : $\widehat{\pi}(x, i) = \frac{1}{M+1}\pi_i(x), \quad x \in \Omega$
- Two moves for Simulated Tempering chain :
  - **Level Move** : Metropolis Hastings at a fixed β_i

    w/p: $\frac{1}{2(M+1)} \min\left(1, \frac{\pi_i(x')}{\pi_i(x)}\right)$
  - **Temperature Move** : Move from i to i +/- 1 in temp space

    w/p: $\frac{1}{2(M+1)} \min\left(1, \frac{Z(\beta_i)}{Z(\beta_{i\pm1})} e^{(\beta_{i\pm1}-\beta_i)H(x)}\right)$
- Partition functions expensive to calculate → Swapping

Swapping

- Chain configuration : $x = (x_0,\ldots,x_M)$: across inverse temperature
- M+1 different inverse temperatures:
  - $\widehat{\Omega}$ = product of M+1 copies of Ω, for each inverse temperature
  - Configuration is M+1-tuple of configurations chosen at each inverse temperature
- Conditional distributions : $\widehat{\pi}(x) = \prod_{i=0}^{M} \pi_i(x_i)$
- Two moves for Simulated Tempering chain :
  - **Level Move** : Metropolis Hastings at a fixed temperature :
    $x = (x_0,\ldots x_i,\ldots,x_M)$ to $x' = (x_0,\ldots x'_i,\ldots,x_M)$ where x and x' only differ at i, and at i they differ only by one-step Metropolis. (same probability)
  - **Swap Move** : $x = (x_0,\ldots x_i,x_{i+1},\ldots,x_M)$ to $x' = (x_0,\ldots x_{i+1},x'_i,\ldots,x_M)$
    w/p : $$\frac{1}{2(M+1)} \min\left(1, e^{(\beta_{i+1}-\beta_i)(H(x_i)-H(x_{i+1}))}\right)$$

Size of Temp^{-1} Space

- M needs to be chosen carefully
  - Large enough for non-trivial temperature move probabilibilities.
  - Small enough for tractable running time
  - Paper chose $M = O(n)$

Proving Torpid mixing of Tempering on Potts

- Lower Bound on $\tau(\varepsilon)$ by showing poor conductance (bad cut)
  - High Temperature (Low β) : high entropy, “uniform-looking”
  - Low Temperature (High β) : high energy, “predominant color-looking”
  - Transition is discontinuous for 3-state mean-field Potts model – abrupt change in the size of the largest color class.

Slow Mixing proof setup

- $n = |V|$, $\Omega = 3^n$: all colorings.
- Ω_σ = Partition set of Ω such that $\sigma = (\sigma_1, \sigma_2, \sigma_3)$
- Partition set probability : $\pi_i(\Omega_\sigma) = \binom{n}{\sigma_1, \sigma_2, \sigma_3} \frac{e^{\beta_i(\sigma_1^2+\sigma_2^2+\sigma_3^2)}}{Z(\beta_i)}$
- Configuration sets of interest :
  - "Uniform looking" : $\Omega_{n/3} = \sigma = (n/3, n/3, n/3)$
  - "Color-dominant looking" : $\Omega_{2n/3} = \sigma = (2n/3, n/6, n/6)$
  - "Transition looking" : $\Omega_{n/2} = \sigma = (n/2, n/4, n/4)$
- Show that a temperature exists such that $\Omega_{n/3}$ and $\Omega_{2n/3}$ have large weight while $\Omega_{n/2}$ has exponentially small weight.

Proof

- **Lemma 1** : There exists β_c such that

  a) $\pi_{\beta c}(\Omega_{n/3}) = \pi_{\beta c}(\Omega_{2n/3}) + o(1)$: Uniform and Color-dominant are equally likely

  b) $\pi_{\beta c}(\Omega_{n/3}) >> \pi_{\beta c}(\Omega_{n/2})$: Transition sets are exponentially unlikely

  → **Shown by solving for β_c and then finding ratio of $\pi_{\beta c}(\Omega_{n/3})/\pi_{\beta c}(\Omega_{n/2})$**

- **Lemma 2** : Most likely $\Omega_{n/2}$ is $\sigma = (n/2,n/4,n/4)$

  → **Shown by solving for $d(\pi(n/2, xn, n/2 - xn))/dx$ to find critical point.**

- **Lemma 3** : For all $\beta_i <= \beta_c$ $\pi_{\beta i}(\Omega_{n/3}) >> \pi_{\beta i}(\Omega_{n/2})$

  → **Shown by extending proof of 1b to $\beta_i <= \beta_c$**

Theorem

- For large n, there exists α so that $\Phi_s <= e^{(-\alpha n + o(n))}$

  Shown by solving for conductance around region of bad cut (region bounded by $\sigma_1,\sigma_2,\sigma_3 = n/2$), showing conductance is bounded by result from O(n) * $\pi_{\beta c}(\Omega_{n/3})/\pi_{\beta c}(\Omega_{n/2})$ then finding α at β_c.

- (Zheng 1999) Result implies that with torpid tempering comes torpid swapping.

Bad Cuts

- Torpid mixing discovered for swapping
  - Due to bad cuts in the state space
- To subvert, choose an interpolation that does not preserve a bad cut
- The maxima and minima should be preserved throughout the interpolation

Bimodal Exponential Distribution

- $\pi(x) = \pi_C(x) = \frac{C^{|x|}}{Z}, \qquad x \in [-N, N'],$
- Bimodal with partition when x=0
- $\pi_i(x) = \frac{C^{\frac{i}{M}|x|}}{Z_i}, \ 0 \leq i \leq M, \ x \in [-N, N']$
- Unchanging maxima and minima over *i*
  - Unchanging basin of attraction
- Maps to a polynomial fraction of the uniform distribution
- Swapping chain over temperature
  - $\beta_i = \beta^* \cdot \frac{i}{M}$

Decomposition of Chain

- Partition swapping chain based on trace
  - Trace t = $(t_0,...,t_M)$: $t_i = 0$ if $x_i < 0$, else $ti = 1$
  - This defines a vector indicating the sign of each element
- Within a partition defined by fixed trace *t*, each state will have trace *t*

Bounding the restricted chain

- Ignoring swapping moves, each configuration is independent of the others
- The mixing time is the worst case mixing time over all temperatures
- A fixed trace restricts configurations to one side of the bimodal distribution, resulting in a unimodal distribution
- Unimodal distributions are rapidly mixing.

Bounding the projection

- The projection is Markov chain defined by partitioning with the trace
  - This is a hypercube of dimension *M+1*
- The swapping transition on this projection results in the transpose of two neighboring bits
- The level transition may result in inverting a bit
  - Only probable at the lowest inverse temperatures

Bounding another Chain

- Consider the chain which involves selecting and inverting any of the *$M+1$* bits
- Each model configuration in the swapping configuration is independent of the other model configurations
  - Then when an bit of the trace is inverted, the distribution is uniform with respect to that bit
- Due to the Coupon Collector Theorem, this chain mixes rapidly. $G = O(M \log M)$

Path Comparison

- Compare the projection to the simple walk
  - For inverting a bit, consider this path
    - Transpose the bit successively to the lowest position
    - Invert the bit
    - Transpose the new bit back to the original position
  - Use the comparison theorem
    - $Gap(P) \geq \frac{1}{A} \cdot Gap(\widetilde{P}),$
    - $A = \max_{(z,w) \in E(P)} \left\{ \frac{1}{\pi(z)P(z,w)} \sum_{\Gamma(z,w)} |\gamma_{xy}| \widetilde{\pi}(x) \widetilde{P}(x,y) \right\}$
      - Restrict the probability of each transition in the chain

Bounding Path Probability

- Assume N ≤ N'
- The probability of each unit in the path is bounded by the transition in the simple walk
  - $\overline{\pi}(z)\,\overline{P}(z,z') \geq \overline{\pi}(t)\,\widetilde{P}(t,t'), = \dfrac{1}{2(M+1)}\min\left(\overline{\pi}(t),\overline{\pi}(t')\right) = \dfrac{\overline{\pi}(t^*)}{2(M+1)}$
- The probability of each state in the path
  - Partition t* into blocks contain 1s, separated by 0s

$$\prod_{l=k+1}^{i} \pi_l(z_l) \geq \prod_{l=k+1}^{i} \pi_l(t_l^*)$$

$$\pi_i(t_{i-1})\pi_{k+1}(0) \geq \pi_i(0)\pi_{k+1}(t_{k+1})$$

$$\pi_i(1)/\pi_i(0) \geq \pi_{k+1}(1)/\pi_{k+1}(0)$$

Partition Gap

- At each state, the total number of paths using a transition is M+1, since there are M+1 transitions in the simple walk
  - Further, the length of a path is O(M)
  - Coupled with the previous theorem
    - $A = O(M)$
    - $G = O(M^{-1})$

Bimodal Mean-field Spin Models

- Examples:
  - Consider the case where $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_q$
    - This constraint prevents more than a single maxima caused by preference for a predominate color
    - Entropy causes a single maxima emphasizing an equal distribution of colors
  - Alternatively, the Ising model under an external field
    - $\pi(x) = \pi_{(\beta,J)}(x) = \dfrac{e^{\beta\left(\sum_{i,j} \delta_{x_i = x_j} + J \sum_i \delta_{x_i = 1}\right)}}{Z(\beta, J)},$

Flat-Swap Algorithm

- Define swapping interpolation using an additional function
  - $\rho_i(x) = \frac{\pi_i(x) f_i(x)}{Z_i^!}, \qquad f_i(x) = \binom{n}{\sigma_1, \ldots, \sigma_q}^{\frac{i-M}{M}}$
- This interpolation directly counters the term provided by entropy

Flat-Swap Algorithm State Space

- For example: on the Ising model with an external field on the complete graph
  - $\rho_i(\Omega_{(k,n-k)}) = \binom{n}{k}\rho_i(x) = \frac{1}{Z_i'}\left(\rho_M(\Omega_{(k,n-k)})\right)^{\frac{i}{M}}$
- With respect to the total spin distributions at all temperatures
  - The distribution maintains the same relative shape
  - same maxima and minima
- The Ising model is bimodal
- The basin of attraction corresponds to a polynomial fraction of the total spin distributions
- Thus this algorithm mixes rapidly using swapping