Annealing and Tempering Markov Chains

Scott C. Schmidler

Stat 863: Advanced Statistical Computing
Duke University
Fall 2018

Last time

- Ising model
- Gibbs sampler/heatbath for Ising model
- Review basic convergence theory Markov chains

Bottlenecks in Gibbs sampler for Ising model

exponentially slow convergence

Markov chains for optimization: Simulated annealing

We can construct MC to sample from $\pi(x)$ known up to constant

We can also use this for *optimization*.

Suppose we wish to find mode $x^* = \arg\max_{x \in \mathcal{X}} \pi(x)$.
Define

$$\pi_t(x) \propto \exp\left(\frac{\ln(\pi(x))}{t}\right) = \pi(x)^{\frac{1}{t}}$$

with potential $U(x) = \ln \pi(x)$.

- $\pi_0(x) = \delta_{x^*}(x)$ is delta function at mode(s)
- $\pi_\infty(x)$ is uniform distn

Can we use MCMC to sample π_0?
Initializing w/ pos density equivalent to finding mode! (And may be reducible if multi-modal.)

Simulated annealing

Take a sequence of decreasing temperatures *(annealing schedule)*

$$t_{\max} = t_0 > t_1 > \ldots > t_n = 0$$

where $t_{\max} \gg 1$.

Construct a time-*inhomogeneous* MCMC chain to simulate

$$\begin{aligned} x_1, \ldots, x_k &\sim \pi_{t_0} \\ x_{k+1}, \ldots, x_{2k} &\sim \pi_{t_1} \\ &\vdots \\ x_{(n-1)k+1}, \ldots, x_{nk} &\sim \pi_{t_n} \end{aligned}$$

Lowering temp concentrates stationary distribution near mode(s) (convergence of MC generally slower at lower t's)

Simulated annealing

In limit as $nk \to \infty$ and $\Delta t \to 0$, guaranteed to obtain global min of $U(x)$

In practice, must empirically determine annealing schedule & restart many times.

Tempering

Can we use this idea to speed up *sampling*?

Suppose we our MCMC chain targets $\pi(x)$ but mixes slowly

"Energy barriers"/low density regions hard to cross at low temps; can be easier at high temps:

$$U(x) = -\ln(\pi(x)) \qquad \text{vs} \qquad \frac{U(x)}{t}$$

Tempering raises the temperature to allow faster exploration of the state space.

Simulated tempering

Define *augmented* state space $(\mathcal{X}, T)$ with

$$\pi(x,t) \propto \pi_t(x)\gamma(t)$$

for finite set T, i.e. $t \in \{1 = t_0 < t_1 < \ldots < t_{\max}\}$

Note that the conditional dist $\pi(x \mid t = 1) = \pi(x)$.

Construct MCMC chain to sample $\pi(x,t)$.
Then $\{(x^{(i)}, t^{(i)}) : t^{(i)} = 1\}$ form a sample from $\pi(x)$

Mixing can be dramatically increased.

Simulated tempering

Where does $\gamma(x)$ come from?

Look more closely:

$$\pi_i(x) = Z_i^{-1} e^{-\frac{\ln(\pi(x))}{t_i}} \qquad Z_i = \int_{\mathcal{X}} e^{-\frac{\ln(\pi(x))}{t_i}} = \int_{\mathcal{X}} \pi(x)^{\frac{1}{t_i}}$$

Consider $\gamma(t) \propto 1$ so sampling from $\pi(x, t_i) \propto \pi(x)^{\frac{1}{t_i}}$. Then marginal

$$\pi(t_i) = \frac{\int_{\mathcal{X}} \pi(x)^{\frac{1}{t_i}}}{\sum_j \int_{\mathcal{X}} \pi(x)^{\frac{1}{t_j}}} = \frac{Z_i}{\sum_j Z_j}$$

What will $\pi(1)$ be? Likely very *small*.

Simulated tempering

Instead prefer $\pi(x, t_i) \propto Z_i^{-1}\pi(x)^{\frac{1}{t_i}}$, so $\pi(t)$ uniform.

But then to accept/reject temp change $(x, t_i) \to (x, t_j)$ we have

$$1 \wedge \frac{Z_i \pi(x)^{\frac{1}{t_j}}}{Z_j \pi(x)^{\frac{1}{t_i}}}$$

now *two* unknown norm constants, which don't cancel

One soln: instead of $\gamma(i) \propto c$, adaptively *estimate* Z_i's and set $\gamma(i) = \hat{Z}_i^{-1}$. (we'll revisit this)

Alternative: *parallel* tempering

Parallel tempering *(aka replica-exchange)*

- Simulate parallel MCMC chains at temps $t_0 = 1 < t_1 < \ldots < t_k$
- Intermittently attempt to swap configurations:

$$\min\left\{1, \frac{\pi_{t_j}(x_i)\pi_{t_i}(x_j)}{\pi_{t_i}(x_i)\pi_{t_j}(x_j)}\right\}$$

Note: preserves the *joint* measure on product space:

$$\pi(\mathbf{x}) = \prod_{i=1}^{k} \pi_{t_i}(x_i) \qquad \mathbf{x} \in \mathcal{X}^{k+1}$$

Easily seen that marginal distribution of X_1 is π.

Unlike simulated tempering, only one unknown norm constant.

Parallel tempering

Note: each chain performing something like random walk on t

(FIGURE)

Practical issues:

- Choose $t_{\max}$ s.t. $\pi_{t_{\max}} \approx$ Unif, or at least mixes rapidly
- *Overlap* between neighboring t_i, t_{i+1} cruicial
  - Heuristic: space temps exponentially $t_i = t_0 \exp(\ln(\frac{t_{\max}}{t_{\min}})\frac{i}{k})$
  - *Monitor* acceptance rates of *all* neighboring pair swaps
  - Automated/adaptive temp placement methods exist

Important note: theory shows adequate swapping rates *nec*, but not *sufficient*[1]

Works well for many high-dimensional, multimodal problems.

[1] see Woodard, Huber, Schmidler

Parallel tempering speedups

Intuition: chains move to higher temperatures, cross energy barriers between modes, return to lower temperatures. This should speed up convergence.

Question: Is this intuition correct? Is parallel tempering better than running a single chain kn steps?

Aside: generally speaking, researchers spend a lot of time coming up with "new" MC schemes, but little can be said about relative performance.

Answer: Yes. Compare convergence *rates*.
(this is hard; let's talk a bit about how to do it)

MCMC can be slow

When $X_0, X_1, X_2, \ldots, X_n$ come from a Markov chain, convergence of ergodic averages

$$\hat{\mu}_h = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

can converge very slowly.

Mixing time

$$\tau_\epsilon = \sup_{\pi_0} \min\{n : \|\pi_{n'} - \pi\|_{\mathrm{TV}} < \epsilon \quad \forall n' \geq n\}.$$

where

$$\|\pi_n - \pi\|_{\mathrm{TV}} = \sup_{A \subset \mathcal{X}} |\pi_n(A) - \pi(A)|$$

In problems with multimodality, high dimensions, or simply strong dependence, mixing times can be very long.

Rapid and slow mixing

One way to characterize this is *rapid mixing*.

Let $(\mathcal{X}^{(d)}, \mathcal{F}^{(d)}, \lambda^{(d)})$ a sequence of measure spaces, and $\pi^{(d)}$ densities wrt $\lambda^{(d)}$ for $d \in \mathbb{N}$ the *problem size*.

P is *rapidly mixing* if $\tau_\epsilon(d)$ is bounded above by a polynomial in d.

P is *torpidly mixing* if $\tau_\epsilon(d)$ bounded below by an exponential in d.

Even if the chain is "rapidly" mixing, τ_ϵ may be impractically large.

Example

Mean-field Ising model

$$\pi(x) = \frac{1}{Z}\exp\left\{\frac{\alpha}{2M}\left(\sum_{i=1}^{M} x_i\right)^2\right\} \qquad \mathcal{X} = \{-1, +1\}^M$$

Thm: Gibbs sampler (Glauber dynamics) is slowly mixing.
Thm (WSH07a): Parallel tempering is rapidly mixing (see also).

Mean-field Potts

With $k \geq 3$ colors.

Thm (WSH07b): Tempering is torpidly mixing (see also BR06).

Examples

Mixtures of normals

$$\pi(z) = \frac{1}{2}N_d(z; -1_d, \sigma_1^2 \mathrm{I}_d) + \frac{1}{2}N_d(z; 1_d, \sigma_2^2 \mathrm{I}_d)$$

Upper/lower bounds on spectral gap (WSH07a,b) yield:

Thm: RW-MH is torpidly mixing.
Thm: Tempering is rapidly mixing for $\sigma_1 = \sigma_2$.
Thm: Tempering is torpidly mixing for $\sigma_1 \neq \sigma_2$.