Lecture 8: Markov Chain Monte Carlo

STATS305B: Applied Statistics II

Scott Linderman

January 29, 2025

Announcements

- HW2 posted online. Due Wed, Feb 12.
- **Midterm next Wednesday in MCCULL 115** See practice test online.

Recap

Last Time...

- Bayesian Inference
- Conjugate Priors
- Laplace Approximation
- Monte Carlo
- Start on MCMC

Outline

Today...

- Markov Chains
- Metropolis Hastings Algorithm
- Gradient-based Proposals (MALA and HMC)
- Gibbs Sampling
- Augmentation Schemes (with Demo)

Approximating Posterior Expectations

Generally, we can't analytically compute posterior expectations. In these cases, we need to resort to approximations. For example, we could use *quadrature methods* like Simpson's rule or the trapezoid rule to numerically approximate the integral over Θ.

Roughly,

$$\mathbb{E}_{p(\theta|x)}[f(\theta)] \approx \sum_{m=1}^{M} p(\theta_m \mid x) f(\theta_m) \Delta_m$$

where $\theta_m \subset \Theta$ is a grid of points and Δ_m is a volume around that point.

This works for low-dimensional problems (say, up to 5 dimensions), but the number of points (M) needed to get a good estimate grows exponentially with the parameter dimension.

Monte Carlo Approximations

Idea: approximate the expectation via sampling,

$$\mathbb{E}_{p(\theta|x)}[f(\theta)] \approx \frac{1}{M}\sum_{m=1}^{M} f(\theta_m) \quad \text{where} \quad \theta_m \sim p(\theta \mid x).$$

Let $\hat{f} = \frac{1}{M}\sum_{m=1}^{M} f(\theta_m)$ denote the Monte Carlo estimate. It is a random variable, since it's a function of random samples θ_m. As such, we can reason about its mean and variance.

Unbiasedness

Clearly,

$$\mathbb{E}[\hat{f}] = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{p(\theta|x)}[f(\theta)] = \mathbb{E}_{p(\theta|x)}[f(\theta)].$$

Thus, $\hat{f}$ is an *unbiased* estimate of the desired expectation.

Monte Carlo Variance

What about its variance?

$$\mathrm{Var}[\hat{f}] = \mathrm{Var}\left(\frac{1}{M}\sum_{m=1}^{M} f(\theta_m)\right) = \frac{1}{M^2}\left(\sum_{m=1}^{M}\mathrm{Var}[f(\theta)] + 2\sum_{1\le m<m'\le M}\mathrm{Cov}[f(\theta_m), f(\theta_{m'})]\right)$$

Comparison to Numerical Quadrature

- If the samples are not only identically distributed but also *uncorrelated*, then $\mathrm{Var}[\hat{f}] = \frac{1}{M}\mathrm{Var}[f(\theta)]$.
- In this case, the *root mean squared error* (RMSE) of the estimate is $\sqrt{\mathrm{Var}[\hat{f}]} = O(M^{-\frac{1}{2}})$.
- Compare this to Simpson's rule, which for smooth 1D problems has an error rate of $O(M^{-4})$. That's roughly 8 times better than Monte Carlo!
- However, for multidimensional problems, Simpson's rule is $O(M^{-\frac{4}{D}})$, whereas the **error rate of Monte Carlo does not depend on the dimensionality!**

The Catch

So far so good: we'll just draw a lot of samples to drive down our Monte Carlo error. **Here's the catch!** How do you draw samples from the posterior $p(\theta \mid x)$? We're interested in Monte Carlo for cases where the posterior does not admit a simple closed form! In general, sampling the posterior is as hard as computing the marginal likelihood.

Markov Chains

A *Markov chain* is a joint distribution of a sequence of variables, $\pi(\theta_1, \theta_2, \ldots, \theta_M)$. (To avoid confusion with the model p, we denote the densities associated with the Markov chain by π.) The Markov chain factorizes so that each variable is drawn conditional on the previous variable,

$$\pi(\theta_1, \theta_2, \ldots, \theta_M) = \pi_1(\theta_1) \prod_{m=2}^{M} \pi(\theta_m \mid \theta_{m-1}).$$

This is called the *Markov property*.

- The distribution $\pi_1(\theta_1)$ is called the *initial distribution*.
- The distribution $\pi(\theta_m \mid \theta_{m-1})$ is called the *transition distribution*. If the transition distribution is the same for each m, the Markov chain is *homogeneous*.

Stationary distributions

Let $\pi_m(\theta_m)$ denote the marginal distribution of sample θ_m. It can be obtained recursively as,

$$\pi_m(\theta_m) = \int \pi_{m-1}(\theta_{m-1})\, \pi(\theta_m \mid \theta_{m-1})\, \mathrm{d}\theta_{m-1}.$$

We are interested in the asymptotic behavior of the marginal distributions as $m \to \infty$.

A distribution $\pi^\star(\theta)$ is a **stationary distribution** if,

$$\pi^\star(\theta) = \int \pi^\star(\theta')\, \pi(\theta \mid \theta')\, \mathrm{d}\theta'.$$

That is, suppose the marginal of sample θ' is $\pi^\star(\theta)$. Then the marginal of the next time point is also $\pi^\star(\theta)$.

Detailed balance

How can we relate transition distributions and stationary distributions? A sufficient (but not necessary) condition for $\pi^\star(\theta)$ to be a stationary distribution is that it satisfies *detailed balance*,

$$\pi^\star(\theta')\pi(\theta \mid \theta') = \pi^\star(\theta)\pi(\theta' \mid \theta).$$

In words, the probability of starting at θ' and moving to θ is the same as that of starting at θ and moving to θ', if you draw the starting point from the stationary distribution.

To see that detailed balance is sufficient, integrate both sides to get,

$$\int \pi^\star(\theta')\pi(\theta \mid \theta') \,\mathrm{d}\theta' = \int \pi^\star(\theta)\pi(\theta' \mid \theta) \,\mathrm{d}\theta' = \pi^\star(\theta).$$

Thus, $\pi^\star(\theta)$ is a stationary distribution of the Markov chain with transitions $\pi(\theta \mid \theta')$.

Ergodicity

Detailed balance can be used to show that $\pi^\star(\theta)$ is *a* stationary distribution, but not that it is *the unique* one. This is where *ergodicity* comes in. A Markov chain is ergodic if $\pi_m(\theta_m) \to \pi^\star(\theta)$ regardless of $\pi_1(\theta_1)$. An ergodic chain has only one stationary distribution, $\pi^\star(\theta)$.

The easiest way to prove ergodicity is to show that it is possible to reach any θ' from any other θ. E.g. this is trivially so if $\pi(\theta' \mid \theta) > 0$.

Note: A more technical definition is that all pairs of sets *communicate*, in which case the chain is *irreducible*, and that each state is *aperiodic*. The definitions can be a bit overwhelming.

Markov Chain Monte Carlo (MCMC)

Finally, we come to our **main objective**: designing a Markov chain for which *the posterior is the unique stationary distribution.* That is, we want $\pi^\star(\theta) = p(\theta \mid x)$.

Recall our **constraint**: we can only compute the joint probability (the numerator in Bayes' rule), not the marginal likelihood (the denominator). Fortunately, that still allows us to compute ratios of posterior densities! We have,

$$\frac{p(\theta \mid x)}{p(\theta' \mid x)} = \frac{p(\theta, x)}{p(x)} \frac{p(x)}{p(\theta', x)} = \frac{p(\theta, x)}{p(\theta', x)}.$$

Now rearrange the detailed balance condition to relate ratios of transition probabilities to ratios of joint probabilities,

$$\frac{\pi(\theta \mid \theta')}{\pi(\theta' \mid \theta)} = \frac{\pi^\star(\theta)}{\pi^\star(\theta')} = \frac{p(\theta \mid x)}{p(\theta' \mid x)} = \frac{p(\theta, x)}{p(\theta', x)}$$

The Metropolis-Hastings algorithm

To construct such a transition distribution $\pi(\theta \mid \theta')$, break it down into two steps.

1. Sample a proposal θ from a *proposal distribution* $q(\theta \mid \theta')$,
2. Accept the proposal with *acceptance probability* $a(\theta' \to \theta)$. (Otherwise, set $\theta = \theta'$.)

Thus,

$$\pi(\theta \mid \theta') = \begin{cases} q(\theta \mid \theta')\, a(\theta' \to \theta) & \text{if } \theta' \neq \theta \\ \int q(\theta'' \mid \theta')\,(1 - a(\theta' \to \theta''))\,\mathrm{d}\theta'' & \text{if } \theta' = \theta \end{cases}$$

Detailed balance is trivially satisfied when $\theta = \theta'$. When $\theta \neq \theta'$, we need

$$\frac{\pi(\theta \mid \theta')}{\pi(\theta' \mid \theta)} = \frac{q(\theta \mid \theta')\, a(\theta' \to \theta)}{q(\theta' \mid \theta)\, a(\theta \to \theta')} = \frac{p(\theta, x)}{p(\theta', x)} \Rightarrow \frac{a(\theta' \to \theta)}{a(\theta \to \theta')} = \underbrace{\frac{p(\theta, x)\, q(\theta' \mid \theta)}{p(\theta', x)\, q(\theta \mid \theta')}}_{\triangleq A(\theta' \to \theta)}$$

The Metropolis-Hastings algorithm

WLOG, assume $A(\theta' \to \theta) \leq 1$. (If it's not, its inverse $A(\theta \to \theta')$ must be.) A simple way to ensure detailed balance is to set $a(\theta' \to \theta) = A(\theta' \to \theta)$ and $a(\theta \to \theta') = 1$.

We can succinctly capture both cases with,

$$a(\theta' \to \theta) = \min\left\{1, A(\theta' \to \theta)\right\} = \min\left\{1, \frac{p(\theta, x)\, q(\theta' \mid \theta)}{p(\theta', x)\, q(\theta \mid \theta')}\right\}.$$

The Metropolis algorithm

Now consider the special case in which the proposal distribution is symmetric;
i.e. $q(\theta \mid \theta') = q(\theta' \mid \theta)$. Then the proposal densities cancel in the acceptance probability and,

$$a(\theta' \to \theta) = \min\left\{1, \frac{p(\theta, x)}{p(\theta', x)}\right\}.$$

In other words, you accept any proposal that moves "uphill," and only accept "downhill" moves with some probability.

This is called the *Metropolis algorithm* and it has close connections to *simulated annealing*.

Smarter Proposals with Gradients

- Metropolis-Hastings with a symmetric Gaussian proposal behaves (kind of) like a random walk.
- Neal (2012) argues that in D dimensions, random walk MH needs $O(D^2)$ iterations to get an independent sample.
- Can we develop more efficient transition distributions? **Yes!** If we have more information about the log probability.
- For example, suppose that the log probability $\log p(\theta)$ is differentiable. We can use the gradient to make proposals that move farther and are more likely to be accepted.

Metropolis-Adjusted Langevin Algorithm (MALA)

The *Metropolis-Adjusted Langevin Algorithm* uses the gradient of the log probability to make asymmetric proposals,

$$q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} + \tau \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}, \boldsymbol{X}),\, 2\tau^2 \boldsymbol{I})$$

Note: $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) \neq q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')$! To calculate the acceptance probability, you need the gradient at both points.

MALA can be motivated as a discrete-time approximation to the *Langevin* diffusion, a continuous-time stochastic differential equation for modeling molecular dynamics.

In high dimensions, the extra information provided by the gradient can lead to much more efficient chains. Neal argues that MALA needs $O(D^{4/3})$ computation to produce an independent sample.

But why stop at one gradient step?

Hamiltonian Monte Carlo (HMC)

Idea: Think of negative log probability as an *energy landscape*. Now imagine a puck sliding around on this bumpy surface. Give it random kicks; it will tend to slide downhill toward points of low potential energy (high probability). Each kick can displace the puck by a large amount. Done properly, the puck will visit points with probability proportional to the posterior probability.

Notation

Following Neal (2012), let

- $\boldsymbol{q} \in \mathbb{R}^D$ denote the *position*; i.e. the current parameters (previously θ)
- $\boldsymbol{p} \in \mathbb{R}^D$ denote the *momentum*; auxiliary variables that we don't care about, but which are necessary for HMC.
- $\boldsymbol{z} = [\boldsymbol{q}, \boldsymbol{p}]^\top \in \mathbb{R}^{2D}$ denote the combined *state of the system*.
- $\boldsymbol{M}$ denote the *mass matrix*, another artificial construct. Typically, this will be $m\boldsymbol{I}$
- $U(\boldsymbol{q})$ denote the *potential energy*
- $K(\boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{M}^{-1}\boldsymbol{p}$ denote the *kinetic energy*

Hamiltonian Dynamics

The *Hamiltonian* is the sum of the potential $H(\boldsymbol{q}, \boldsymbol{p}) = U(\boldsymbol{q}) + K(\boldsymbol{p}) = U(\boldsymbol{q}) + \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{M}^{-1}\boldsymbol{p}$.

The partial derivatives determine how the state evolves over time,

$$\frac{\mathrm{d}q_d}{\mathrm{d}t} = \frac{\partial H}{\partial p_d} = [\boldsymbol{M}^{-1}\boldsymbol{p}]_d$$
$$\frac{\mathrm{d}p_d}{\mathrm{d}t} = -\frac{\partial H}{\partial q_d} = -\frac{\partial U}{\partial q_d}$$

for $d = 1, \ldots, D$.

Compactly,

$$\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t} = \boldsymbol{J}\nabla H(z) \qquad \text{where} \qquad \boldsymbol{J} = \begin{bmatrix} \boldsymbol{0}, \boldsymbol{I} \\ -\boldsymbol{I}, \boldsymbol{0} \end{bmatrix}$$

Using Hamiltonian Dynamics for Posterior Inference

Define a joint distribution on positions and momenta as,

$$p(\boldsymbol{q}, \boldsymbol{p}) \propto \exp\left\{-H(\boldsymbol{q}, \boldsymbol{p})\right\} \propto \exp\left\{-U(\boldsymbol{q}) - K(\boldsymbol{p})\right\}.$$

Now let $U(\boldsymbol{q}) = -\log p(\boldsymbol{\theta} = \boldsymbol{q}, \boldsymbol{X})$ be the *negative* log joint probability. Then,

$$p(\boldsymbol{q}, \boldsymbol{p}) = p(\boldsymbol{\theta} = \boldsymbol{q} \mid \boldsymbol{X}) \times p(\boldsymbol{p})$$

Samples of $\boldsymbol{q}$ will be marginally distributed according to the posterior $p(\boldsymbol{\theta} = \boldsymbol{q} \mid \boldsymbol{X})$.

Samples of $\boldsymbol{p}$ will be marginally distributed $p(\boldsymbol{p}) = \frac{\exp\{-K(\boldsymbol{p})\}}{\int_{\mathbb{R}^D} \exp\{-K(\boldsymbol{p})\}\, \mathrm{d}\boldsymbol{p}}$. These are *auxiliary variables* that we don't really care about—they're just there to help us construct MH proposals.

We choose $K(\boldsymbol{p})$ so $p(\boldsymbol{p})$ is convenient; e.g. if $K(\boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{M}^{-1} \boldsymbol{p}$ then

$$p(\boldsymbol{p}) = \mathcal{N}(\boldsymbol{p} \mid \boldsymbol{0}, \boldsymbol{M}).$$

Algorithm

Hamiltonian Monte Carlo (HMC) is Metropolis-Hastings on the joint distribution of $(\boldsymbol{q}, \boldsymbol{p})$ with proposals based on Hamiltonian dynamics.

Starting at point $(\boldsymbol{q}', \boldsymbol{p}')$, sample the proposal distribution:

1. Throw away $\boldsymbol{p}'$ and sample new momenta from their marginal distribution $\boldsymbol{p} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{M})$.
2. Approximate Hamiltonian dynamics for Δt time to get to a new point $(\boldsymbol{q}, \boldsymbol{p})$. (Use $L = \Delta t / \epsilon$ Leapfrog integration steps each of size ϵ.)
3. Flip the momentum $\boldsymbol{p} \leftarrow -\boldsymbol{p}$ to make the proposal symmetric.

Algorithm

Then accept the proposed point $(\boldsymbol{q}, \boldsymbol{p})$ with probability,

$$
\begin{aligned}
a((\boldsymbol{q}', \boldsymbol{p}') \to (\boldsymbol{q}, \boldsymbol{p})) &= \min\left\{1, \frac{\exp\{-H(\boldsymbol{q}, \boldsymbol{p})\}\, q(\boldsymbol{q}', \boldsymbol{p}' \mid \boldsymbol{q}, \boldsymbol{p})}{\exp\{-H(\boldsymbol{q}', \boldsymbol{p}')\} q(\boldsymbol{q}, \boldsymbol{p} \mid \boldsymbol{q}', \boldsymbol{p}')}\right\} \\
&= \min\left\{1, \frac{\exp\{-H(\boldsymbol{q}, \boldsymbol{p})\}}{\exp\{-H(\boldsymbol{q}', \boldsymbol{p}')\}}\right\}.
\end{aligned}
$$

If the Hamiltonian dynamics were simulated exactly, HMC would always accept. In practice, differences arise from numerical integration errors.

Gibbs Sampling

Gibbs is a special case of MH with proposals that always accept. Gibbs sampling updates one "coordinate" of $\theta \in \mathbb{R}^D$ at a time by sampling from its conditional distribution. The algorithm is:

Gibbs Sampling:

Input: Initial parameters $\theta^{(0)}$, observations x

- **For** $t = 1, \ldots, T$
  - **For** $d = 1, \ldots, D$
    - Sample $\theta_d^{(t)} \sim p(\theta_d \mid \theta_1^{(t)}, \ldots, \theta_{d-}^{(t)}, \theta_{d+1}^{(t-1)}, \ldots, \theta_D^{(t-1)}, x)$

Gibbs Sampling

Return samples $\{\theta^{(t)}\}_{t=1}^{T}$ ‘‘‘

You can think of Gibbs as cycling through D Metropolis-Hastings proposals, one for each coordinate $d \in 1, \ldots, D$,

$$q_d(\theta \mid \theta') = p(\theta_d \mid \theta'_{\neg d}, x)\, \delta_{\theta'_{\neg d}}(\theta_{\neg d}),$$

where $\theta_{\neg d} = (\theta_1, \ldots, \theta_{d-1}, \theta_{d+1}, \ldots, \theta_D)$ denotes all parameters except θ_d.

In other words, the proposal distribution q_d samples θ_d from its conditional distribution and leaves all the other parameters unchanged.

Gibbs Sampling

What is the probability of accepting this proposal?

$$
\begin{aligned}
a_d(\theta' \to \theta) &= \min\left\{1, \frac{p(\theta, x) q_d(\theta' \mid \theta)}{p(\theta', x) q_d(\theta \mid \theta')}\right\} \\
&= \min\left\{1, \frac{p(\theta, x) p(\theta'_d \mid \theta_{\neg d}, x) \delta_{\theta_{\neg d}}(\theta'_{\neg d})}{p(\theta', x) p(\theta_d \mid \theta'_{\neg d}, x) \delta_{\theta'_{\neg d}}(\theta_{\neg d})}\right\} \\
&= \min\left\{1, \frac{p(\theta_{\neg d}, x) p(\theta_d \mid \theta_{\neg d}, x) p(\theta'_d \mid \theta_{\neg d}, x) \delta_{\theta_{\neg d}}(\theta'_{\neg d})}{p(\theta'_{\neg d}, x) p(\theta'_d \mid \theta'_{\neg d}, x) p(\theta_d \mid \theta'_{\neg d}, x) \delta_{\theta'_{\neg d}}(\theta_{\neg d})}\right\} \\
&= \min\{1, 1\} = 1
\end{aligned}
$$

for all θ, θ' that differ only in their d-th coordinate.

The Gibbs proposal is an offer you cannot refuse.

Gibbs Sampling

Of course, if we only update one coordinate, the chain can't be ergodic. However, if we cycle through coordinates it generally will be.

Question: If Gibbs sampling always accepts, is it strictly better than other Metropolis-Hastings algorithms? :::

Conclusion

This was obviously a whirlwind of an introduction to Bayesian inference! There's plenty more to be said about Bayesian statistics — choosing a prior, subjective vs objective vs empirical Bayesian approaches, the role of the marginal likelihood in Bayesian model comparison, varieties of MCMC, and other approaches to approximate Bayesian inference. We'll dig into some of these topics as the course goes on, but for now, we have some valuable tools for developing Bayesian modeling and inference with discrete data!