Gibbs Sampling Is a Special Case of Metropolis–Hastings

Gibbs sampling is a computationally convenient Bayesian inference algorithm that is a special case of the Metropolis–Hastings algorithm. I discuss Gibbs sampling in the broader context of Markov chain Monte Carlo methods.

PUBLISHED

23 February 2020

I avoided learning about Gibbs sampling for a long time. Practically, it is straightforward to implement without understanding the theory, and I assumed that learning about it would be yet another hill to climb. I found that, in fact, understanding Gibbs sampling is easy if you already understand Metropolis–Hastings. It is a special case in which the acceptance ratio is always one. In this post, I'll outline the conceptual thread from Markov chain Monte Carlo (MCMC) methods to Metropolis–Hastings to Gibbs sampling, with links as needed for those who want more details.

Markov chain Monte Carlo

In Bayesian inference, given data X and parameters θ, the posterior

$$p(\theta \mid X) = \frac{p(X \mid \theta)P(\theta)}{\int p(X \mid \theta)P(\theta)\mathrm{d}\theta} \tag{1}$$

is generally unavailable in closed form, and we must rely on other methods to perform inference. MCMC methods approach this problem by simulating a Markov chain whose stationary distribution is the desired posterior, $P(\theta \mid X)$. This works because an ergodic Markov chain is one in which the long-term probability of being on each state is independent of the initial state. The random walk is fated. Thus, walking an ergodic Markov chain and recording states is, in the long-run, like sampling from its stationary distribution.

If a Markov chain is *irreducible*, meaning any state is reachable from any other state, and if it *aperiodic*, meaning the same state is not reached on a fixed frequency, then it is ergodic. But how do we perform a

random walk such that the implicitly constructed Markov chain's stationary distribution is the desired posterior? A sufficient condition is the *reversibility constraint* or the *detailed balance*: the probability of transitioning from one state to another must be equivalent to the probability of moving in the reverse direction. If we denote the stationary distribution as $\pi(\cdot)$ and the transition kernel, the function that computes the probability of moving from state θ to $\theta^\star$, as $K(\theta^\star \mid \theta)$, then the reversibility constraint is

$$K(\theta^\star \mid \theta)\pi(\theta) = K(\theta \mid \theta^\star)\pi(\theta^\star). \tag{2}$$

You can find a proof of why this works in my previous post on Metropolis–Hastings or (Chib & Greenberg, 1995). Intuitively, I think of the reversibility constraint as meaning that our proposed kernel is direction- and time-invariant; the only thing that matters is the probability of being on a given state, defined by $\pi(\cdot)$. In MCMC, we must propose a distribution $\pi(\cdot)$ and a kernel $K(\cdot \mid \cdot)$ such that the constraints above hold. Then we know we walking an ergodic Markov chain whose stationary distribution *is* $\pi(\cdot)$.

Metropolis–Hastings

The classic MCMC method is Metropolis–Hastings (Metropolis et al., 1953; Hastings, 1970). The idea is to use a *proposal* distribution $Q(\theta)$ from which one can sample new states. At each iteration, propose a new state $\theta^\star \sim Q(\theta)$ and accept it with probability

$$\alpha(\theta^\star \mid \theta) = \min\left\{1, \frac{P(\theta^\star \mid X)Q(\theta)}{P(\theta \mid X)Q(\theta^\star)}\right\}. \tag{3}$$

Note that since the chain stays on state θ with probability $1 - \alpha(\theta^\star \mid \theta)$, the chain is aperiodic. For a discrete state-space Markov chain, it is as if each state had a self-loop. If $Q(\cdot)$ assigns non-zero probability for each $\theta \in \Theta$, the chain is irreducible and therefore also ergodic. The detailed balance is achieved because

$$\begin{aligned}
\overbrace{P(\theta \mid X)Q(\theta^\star)\alpha(\theta^\star \mid \theta)}^{\pi(\cdot)\qquad K(\cdot\mid\cdot)} &= \min\left\{P(\theta \mid X)Q(\theta^\star), P(\theta^\star \mid X)Q(\theta)\right\} \\
&= \min\left\{P(\theta^\star \mid X)Q(\theta), P(\theta \mid X)Q(\theta^\star)\right\} \\
&= P(\theta^\star \mid X)Q(\theta)\alpha(\theta \mid \theta^\star).
\end{aligned} \tag{4}$$

In other words, the acceptance ratio in (3) was carefully constructed to ensure detailed balance. At this point you might ask: how does this help if we don't have access to $P(\theta \mid X)$? Notice that

$$\frac{P(\theta^\star \mid X)}{P(\theta \mid X)} = \frac{\frac{P(X|\theta^\star)P(\theta^\star)}{P(X)}}{\frac{P(X|\theta)P(\theta)}{P(X)}} = \frac{P(X \mid \theta^\star)P(\theta^\star)}{P(X \mid \theta)P(\theta)}. \tag{5}$$

Conveniently, the generally intractable integrals in the numerator and denominator,

$$P(X) = \int p(X \mid \theta) P(\theta) \mathrm{d}\theta, \tag{6}$$

cancel out, and the right-most term in (5) just requires computing the likelihood times period of our model under $\theta^\star$ and θ. This leads to a special case of Metropolis–Hastings that is commonly used to infer the posterior $P(\theta \mid X)$ in Bayesian inference:

> **Metropolis–Hastings:**
>
> for $t = 1, \ldots, T$ do
>
> 1. Draw $\theta^\star \sim Q(\theta)$.
> 2. Calculate
>
> $$r = \frac{P(X \mid \theta^\star) P(\theta^\star) Q(\theta^{(t)})}{P(X \mid \theta^{(t)}) P(\theta^{(t)}) Q(\theta^\star)}$$
>
> 1. Draw $u \sim \mathrm{Uniform}(0, 1)$.
> 2. If $u < r$ then $\theta^{(t+1)} := \theta^\star$. Otherwise, $\theta^{(t+1)} := \theta^{(t)}$.

Notice that in the special case that the proposal distribution is the prior, the ratio r reduces to a likelihood ratio,

$$r = \frac{P(X \mid \theta^\star) P(\theta^\star) P(\theta^{(t)})}{P(X \mid \theta^{(t)}) P(\theta^{(t)}) P(\theta^\star)} = \frac{P(X \mid \theta^\star)}{P(X \mid \theta^{(t)})}. \tag{7}$$

Gibbs sampling

Gibbs sampling is a special case of Metropolis–Hastings in which the newly proposed state is always accepted with probability one. It is fairly straightforward to see this once you know the algorithm. Consider a D-dimensional posterior with parameters $\theta = \{\theta_1, \ldots, \theta_D\}$. The basic idea of Gibbs sampling is to iterately sample from the conditional distribution $P(\theta_d \mid X, \theta_{-d})$ where θ_{-d} is θ without the dth parameter:

> **Gibbs sampling:**
>
> for $t = 1, \ldots T$ do
>
> $$\begin{aligned}
> \theta_1^{(t+1)} &:= \theta_1^\star \sim P(\theta_1^{(t)} \mid X, \theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_D^{(t)}) \\
> \theta_2^{(t+1)} &:= \theta_2^\star \sim P(\theta_2^{(t)} \mid X, \theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_D^{(t)}) \\
> &\vdots \\
> \theta_d^{(t+1)} &:= \theta_d^\star \sim P(\theta_d^{(t)} \mid X, \theta_1^{(t+1)}, \ldots, \theta_{d-1}^{(t+1)}, \theta_d^{(t)}, \ldots, \theta_D^{(t)})
> \end{aligned}$$

$$\vdots$$
$$\theta_D^{(t+1)} := \theta_D^\star \sim P(\theta_D^{(t)} \mid X, \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{D-1}^{(t+1)})$$

To see why this works, first note that

$$P(\theta \mid X) = P(\theta_d, \theta_{-d} \mid X) = P(\theta_d \mid X, \theta_{-d}) P(\theta_{-d} \mid X). \tag{8}$$

Ignoring the iterates' notation, the probability of a transition can be written as

$$\begin{aligned} \alpha(\theta^\star \mid \theta) &= \min\left\{1, \frac{P(\theta^\star \mid X) P(\theta_d \mid X, \theta_{-d})}{P(\theta \mid X) P(\theta_d^\star \mid X, \theta_{-d}^\star)}\right\} \\ &= \min\left\{1, \frac{P(\theta_d^\star \mid X, \theta_{-d}^\star) P(\theta_{-d}^\star \mid X) P(\theta_d \mid X, \theta_{-d})}{P(\theta_d \mid X, \theta_{-d}) P(\theta_{-d} \mid X) P(\theta_d^\star \mid X, \theta_{-d}^\star)}\right\} \\ &= 1. \end{aligned} \tag{9}$$

In (9), I have color-coded the terms that cancel. In particular, the terms in red cancel because $\theta_{-d}^\star = \theta_{-d}$. In other words, in each step of the Gibbs sampling algorithm, we are performing a Metropolis–Hastings-like random walk in which the proposed next state always adheres to the reversibility constraint.

The primary advantage of Gibbs sampling is simple: proposals are always accepted. The primary disadvantage is that we need to be able to derive the above conditional probability distributions. This is tractable when $P(\theta_d)$ is conjugate to the posterior in (10).

Conclusion

We can view Gibbs sampling as just a special case of the Metropolis–Hastings algorithm. In my mind, the conceptually hard part about both these algorithms is understanding the correctness of the reversibility constraint and how we can specify a distribution that is guaranteed to be the stationary distribution of an implicit Markov chain. Armed with this knowledge, we can see the correctness of Gibbs sampling with just a little algebra.

For a complete example of a Gibbs sampler, see my post on Pólya-gamma variable augmentation, particularly this section. The main idea of that work is to generate augmenting variables such that an intractable distribution becomes conditionally Gaussian. This allows for an efficient Gibb sampler by switching between sampling the augmenting variables and the main variables of interest.

1. Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, *49*(4), 327–335.

2. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.
3. Hastings, W. K. (1970). *Monte Carlo sampling methods using Markov chains and their applications*.