Misc Remarks about Markov Chains and Gibbs Sampling

Apurva Nakade

2025-03-12

Table of contents

Number of Samples and Burn-in

There are no precise ways of deciding how many samples to generate from a Markov chain or how many samples to *burn* at the start. The simple principle is to generate as many samples as possible. This guarantees the best possible results thanks to the following two facts.

Effective Sample Size

Given dependent N samples that are identically distributed, the effective sample size N_{eff} measures the number of independent samples with the same distribution that'll have the same variance.

More precisely, if $X_1, X_2, \dots, X_N$ are identically distributed random variables with variance σ^2, then

$$\text{var}\left(\frac{X_1 + X_2 + \cdots + X_N}{N}\right) = \frac{\sigma^2}{N_{eff}}.$$

Now consider the following two approaches to sample from a Markov chain:

1. We generate and keep N samples, and keep all the (correlated) samples.
2. We only keep every k-th sample so as to reduce correlation.

Then one can show that the effective sample size in the first case is at least as large as the effective sample size in the second case, even though the samples are correlated. Thus *thinning* doesn't achieve improved results (but might save computational time).

MC Central Limit Theorem

Consider an ergodic Markov chain $X_0, X_1, X_2, ...$ with stationary distribution Π. The random variables X_i are highly correlated and hence the standard central limit theorem does not apply. Nevertheless, there is a Markov chain version of the CLT.

Consider any function f and define the mean

$$f(\bar{X})_N = \frac{f(X_0) + f(X_1) + \cdots + f(X_{N-1})}{N}.$$

Then as $N \to \infty$ (and hence as $N_{eff} \to \infty$)

$$f(\bar{X})_N \sim \mathcal{N}\left(\mathbb{E}_\Pi(f), \frac{\sigma^2_\Pi(f)}{N_{eff}}\right).$$

Because of this result, we can estimate the mean of any function f using sufficiently large number of samples generated from the Markov chain.

Gibbs Sampling

Normalization Constant

One advantage of Gibbs sampling is that it doesn't require the normalization constant of the distribution. This is because the normalization constant cancels out when finding the conditional distribution of each variable. This is extremely useful in practice because computing the normalization constant is often intractable.

Example 0.1. For the simple case of a truncated exponential distribution in 2D,

$$f(x,y) = \frac{1}{c} e^{-\lambda xy} \text{ for } x, y \in [0, D_1] \times [0, D_2],$$

the normalization constant c is given by the double integral

$$\begin{aligned} c &= \int_0^{D_1} \int_0^{D_2} e^{-\lambda xy} \, dy \, dx \\ &= \int_0^{D_1} \left[-\frac{1}{\lambda x} e^{-\lambda xy} \right]_{y=0}^{y=D_2} dx \\ &= \int_0^{D_1} \left(-\frac{1}{\lambda x} e^{-\lambda x D_2} + \frac{1}{\lambda x} \right) dx \\ &= \int_0^{D_1} \frac{1 - e^{-\lambda x D_2}}{\lambda x} \, dx \end{aligned}$$

This integral does not have a closed form solution and hence needs to be approximated numerically. However, in Gibbs sampling, we don't need to compute this constant because it cancels out when we find the conditional distribution of x and y.

::: {#exm-uniform-distribution}

Another common scenario where the normalization constant is hard to compute is when the distribution is uniform over a region. For $\Omega \in \mathbb{R}^d$, the uniform distribution is given by

$$f(x) = \frac{1}{\text{vol}(\Omega)} \text{ for } x \in \Omega.$$

The normalization constant is the volume of the region Ω which is often hard to compute. Thankfully, the conditionals of the uniform distribution are also uniform and the only thing we need are the bounds of the region.

Lack of Parallelization

One of the biggest disadvantages of Gibbs sampling is the lack of parallelization. Because the sampling of each variable depends on the current values of all other variables, it is not possible to sample multiple variables in parallel. This can be a bottleneck in practice when the number of variables is large. There are some methods to try and mitigate this issue.

1. **Blocked Gibbs Sampling**: In this method, we form a "block" of variables and sample them together. This can be done when the variables in the block are conditionally independent given the other variables. This can speed up the sampling process because we can sample multiple variables in parallel.
2. **Parallel Chains**: We can run multiple chains in parallel and sample each chain independently. This can be useful when the chains are not highly correlated. One could, for example, run a preliminary chain to generate a set of initial values and then run multiple chains in parallel from these initial values.

These methods can help speed up the sampling process but they are not always applicable. In general, Gibbs sampling is not as parallelizable as other methods like Metropolis-Hastings.

Mixing Time

The mixing time of a Markov chain is the number of steps it takes for the chain to reach its stationary distribution. In Gibbs sampling, the mixing time can be quite large because the sampling of each variable depends on the current values of all other variables. This can lead to slow convergence and poor mixing.

Sampling in Large Dimensions

Where Gibbs sampling really shines is in high-dimensional spaces. In such spaces, traditional sampling methods like rejection sampling become inefficient because the probability of accepting a sample becomes very low. Gibbs sampling, on the other hand, can efficiently sample from high-dimensional distributions by breaking the problem into smaller, more manageable conditional distributions.

Example 0.2. Consider the simple problem of uniformly sampling over the unit sphere in $\mathbb{R}^d$. One naive method would be to use rejection sampling over the hypercube $[-1, 1]^d$ and accept the sample if it lies within the sphere. The acceptance probability is given by the ratio of the volumes of the sphere and the hypercube, which can be shown to be

$$\text{acceptance ratio in } 2d \text{ dimensions} = \frac{\pi^d}{4^d d!}.$$

This goes to 0 as d increases and hence rejection sampling becomes inefficient. For example, in 10 dimensions, the acceptance ratio is only 0.0026 and so Gibbs sampling would be much more efficient.

Suppose $X_1, X_2, \ldots, X_d$ are the coordinates of a point on the unit sphere. Then the conditional distribution of X_i given all other coordinates is uniform over the interval

$$\left[-\sqrt{1-X_1^2-X_2^2-\cdots-X_{i-1}^2}, \sqrt{1-X_1^2-X_2^2-\cdots-X_{i-1}^2}\right],$$

which is trivial to sample from.