Monte Carlo Markov Chain

Søren Lund Pedersen

In this article, I want to discuss the motivation, intuition, and mathematics behind the Monte Carlo Markov Chain (MCMC) framework. The MCMC framework consists of two parts: Monte Carlo, a simulation technique used to generate random numbers, and Markov Chains, a mathematical framework describing transitions between states where the next state depends only on the current state. MCMC methods are particularly valuable in Bayesian inference, where exact computation of the posterior distribution is often infeasible due to difficult normalization constants.

Why MCMC is Necessary: Exact Bayesian inference typically requires integrating over high-dimensional parameter spaces to compute normalization constants or expectations, which is analytically intractable and often prohibitive for numerical integration methods. While Maximum a Posteriori (MAP) estimation provides a point estimate by maximizing the posterior,

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta \mid x),$$

it does not capture the full uncertainty about parameters. MCMC circumvents these challenges by generating samples from the posterior distribution, enabling approximate numerical integration and comprehensive uncertainty quantification without the need for exact analytical solutions.

Numerical Integration and Bayesian CLT: Numerical integration over complex, high-dimensional spaces can be computationally expensive and inaccurate. MCMC offers an alternative by approximating integrals through averages over samples drawn from the posterior. As the number of samples increases, the Bayesian Central Limit Theorem (CLT) assures that the posterior distribution tends toward a normal distribution under regularity conditions, which justifies the use of MCMC methods and certain approximations like the Laplace Approximation near the mode.

1 Bayesian CLT and Laplace Approximation

The **Bayesian Central Limit Theorem (CLT)** explains that, with a large enough sample size, the posterior distribution approximates a normal distribution centered around the true parameter value, and the influence of the prior diminishes. This result provides theoretical support for using Gaussian approximations in Bayesian inference as sample size grows.

The **Laplace Approximation** leverages this idea by approximating the posterior distribution near its mode using a second-order Taylor expansion. It is particularly useful when the posterior is unimodal and roughly Gaussian, but it may be inadequate for multimodal or highly skewed posteriors, where MCMC sampling becomes essential.

2 General Idea of MCMC, Markov Chains, and Stationary Distribution

MCMC methods generate a sequence of samples $\{x^{(1)}, x^{(2)}, \ldots\}$ from the target posterior distribution $p(\theta \mid x)$. This sequence is constructed as a Markov chain, where the probability of transitioning to the next state depends solely on the current state. The key property of these chains is that they are designed to have the target distribution as their **stationary distribution**. Once the chain has converged (reached stationarity), samples drawn from it can be treated as approximately independent draws from the posterior, enabling estimation of various statistics.

3 Gibbs Sampling

Gibbs Sampling is an MCMC technique that iteratively samples each parameter from its conditional distribution, holding others fixed. This is especially useful for hierarchical models with conjugate priors, where these conditionals take simple forms. Despite its simplicity and ease of implementation, Gibbs sampling can converge slowly if parameters are highly correlated or if the full conditional distributions are complex to sample from.

4 Metropolis-Hastings Algorithm

The **Metropolis-Hastings (MH) algorithm** generates a Markov chain by proposing new states and accepting them with a probability that ensures convergence to the target distribution. Its steps are:

- Start with a current state x_t.
- Propose a new state x' using a proposal distribution $q(x_t \to x')$.
- Compute the acceptance probability:

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

- Accept x' with probability $\alpha(x \to x')$; otherwise, retain x_t.

This mechanism ensures the Markov chain has $p(x)$ as its stationary distribution.

Gibbs within MH Sampling: Sometimes, full conditional distributions required for Gibbs sampling are not available in closed form. In such cases, we embed Metropolis-Hastings steps within Gibbs sampling: use MH to sample from difficult conditionals while using Gibbs for simpler ones. This hybrid approach maintains the structure of Gibbs sampling while handling complex conditional distributions flexibly.

5 Diagnosing and Improving Convergence

Assessing whether an MCMC chain has converged to its stationary distribution is crucial:

- **Trace Plots:** Plot parameter values over iterations. A well-mixed chain will appear as a random scatter without trends, indicating stationarity.

- **Autocorrelation Analysis:** High autocorrelation indicates slow mixing. Reducing autocorrelation (via thinning or using better proposals) improves effective sample size.
- **Multiple Chains:** Running several chains from different starting points and checking if they converge to the same distribution helps detect convergence issues.

6 Effective Sample Size

Even with many MCMC samples, high autocorrelation reduces the amount of independent information. The **Effective Sample Size (ESS)** quantifies the equivalent number of independent samples:

$$ESS = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho_k},$$

where n is the total number of samples and ρ_k is the autocorrelation at lag k. A higher ESS indicates more reliable estimates. Techniques to improve ESS include reparameterization, using more sophisticated proposal distributions, and running chains longer.