Markov Chain Monte Carlo

Statistics · Module 9 — Computational Statistics

Node 9.3

Position in Knowledge Graph

System: Statistics

Structural Domain: Computational Statistics

Node: 9.3

MCMC extends Monte Carlo integration (Node 9.2) to settings where direct i.i.d. sampling from the target distribution is infeasible. By constructing a Markov chain whose stationary distribution is the target, MCMC converts the problem of sampling into the problem of simulating a chain long enough for ergodic averages to converge. This node develops the Metropolis–Hastings algorithm, proves that detailed balance implies stationarity, derives Gibbs sampling as a special case, and establishes the convergence diagnostic framework.

Formal Definition

Let $\pi(\theta)$ be a target density on $\Theta \subseteq \mathbb{R}^k$, known up to a normalizing constant. A **Markov chain Monte Carlo** method constructs a Markov chain $(\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \dots)$ with transition kernel $K(\theta, \theta')$ such that:

1. π is a **stationary distribution**: $\int K(\theta, \theta')\pi(\theta)\, d\theta = \pi(\theta')$.
2. The chain is **ergodic** (irreducible, aperiodic, positive recurrent).

Under these conditions, the **ergodic theorem** guarantees: for any h with $\mathbb{E}_\pi[|h|] < \infty$,

$$\frac{1}{T}\sum_{t=1}^{T} h(\theta^{(t)}) \xrightarrow{a.s.} \mathbb{E}_\pi[h(\theta)].$$

A sufficient condition for stationarity is **detailed balance**:

$$\pi(\theta)K(\theta, \theta') = \pi(\theta')K(\theta', \theta) \quad \text{for all } \theta, \theta'.$$

Intuition

MCMC trades the requirement of independent samples for the ability to sample from distributions known only up to a normalizing constant. The chain wanders through parameter space, spending time in regions proportional to their probability under π. The art of MCMC is designing chains that explore the target distribution efficiently – avoiding both slow random walks and proposals that are rejected too often. The burn-in period discards early samples that are influenced by the initialization, and convergence diagnostics assess whether the chain has reached stationarity.

Structural Role

MCMC is the computational engine of modern Bayesian statistics. It extends posterior computation beyond conjugate families (Node 6.2) to arbitrary posteriors, including hierarchical models, nonparametric models, and latent variable models. It depends on Monte Carlo integration (Node 9.2) for the ergodic averaging framework and enables variational inference (Node 9.6) as a faster alternative when MCMC mixing is prohibitively slow. The bootstrap (Node 9.4) is a parallel computational strategy that uses resampling rather than Markov chains.

Mathematical Development

Metropolis–Hastings Algorithm

Let $q(\theta' \mid \theta)$ be a proposal kernel. The **Metropolis–Hastings** algorithm generates the chain as follows:

1. Given $\theta^{(t)}$, draw $\theta' \sim q(\cdot \mid \theta^{(t)})$.
2. Compute the acceptance probability:

$$\alpha(\theta^{(t)}, \theta') = \min\left(1,\ \frac{\pi(\theta')\, q(\theta^{(t)} \mid \theta')}{\pi(\theta^{(t)})\, q(\theta' \mid \theta^{(t)})}\right).$$

3. With probability α, set $\theta^{(t+1)} = \theta'$; otherwise, set $\theta^{(t+1)} = \theta^{(t)}$.

Theorem (Detailed Balance). The Metropolis–Hastings chain satisfies detailed balance with respect to π.

Proof. The transition kernel is

$$K(\theta, \theta') = q(\theta' \mid \theta)\alpha(\theta, \theta') + r(\theta)\delta_\theta(\theta'),$$

where $r(\theta) = 1 - \int q(\theta' \mid \theta)\alpha(\theta, \theta')\, d\theta'$ is the rejection probability. For the continuous part, we must show $\pi(\theta)q(\theta' \mid \theta)\alpha(\theta, \theta') = \pi(\theta')q(\theta \mid \theta')\alpha(\theta', \theta)$.

Without loss of generality, assume $\pi(\theta')q(\theta \mid \theta') \leq \pi(\theta)q(\theta' \mid \theta)$. Then $\alpha(\theta, \theta') = \pi(\theta')q(\theta \mid \theta')/[\pi(\theta)q(\theta' \mid \theta)]$ and $\alpha(\theta', \theta) = 1$. Substituting:

$$\pi(\theta)q(\theta' \mid \theta) \cdot \frac{\pi(\theta')q(\theta \mid \theta')}{\pi(\theta)q(\theta' \mid \theta)} = \pi(\theta')q(\theta \mid \theta') \cdot 1. \quad \square$$

Since π is the stationary distribution and the chain is typically irreducible and aperiodic (given mild conditions on q), the ergodic theorem applies.

Special Case: Random Walk Metropolis

When $q(\theta' \mid \theta) = g(\theta' - \theta)$ for a symmetric density g, the acceptance ratio simplifies to

$$\alpha(\theta^{(t)}, \theta') = \min\left(1, \frac{\pi(\theta')}{\pi(\theta^{(t)})}\right).$$

The proposal variance controls the tradeoff: too small yields high acceptance but slow exploration; too large yields frequent rejection. The optimal acceptance rate for a k-dimensional Gaussian target is approximately 0.234 (Roberts, Gelman, and Gilks, 1997).

Gibbs Sampling

When $\theta = (\theta_1, \dots, \theta_k)$ and the full conditional distributions $\pi(\theta_j \mid \theta_{-j})$ are available, **Gibbs sampling** cycles through:

For $j = 1, \dots, k$: draw $\theta_j^{(t+1)} \sim \pi(\theta_j \mid \theta_1^{(t+1)}, \dots, \theta_{j-1}^{(t+1)}, \theta_{j+1}^{(t)}, \dots, \theta_k^{(t)})$.

Proposition. Gibbs sampling is a special case of Metropolis–Hastings with proposal $q(\theta' \mid \theta) = \pi(\theta'_j \mid \theta_{-j}) \prod_{i \neq j} \delta_{\theta_i}(\theta'_i)$ and acceptance probability identically 1.

Proof. The MH ratio for coordinate j is

$$\frac{\pi(\theta')\, q(\theta \mid \theta')}{\pi(\theta)\, q(\theta' \mid \theta)} = \frac{\pi(\theta'_j, \theta_{-j})\, \pi(\theta_j \mid \theta_{-j})}{\pi(\theta_j, \theta_{-j})\, \pi(\theta'_j \mid \theta_{-j})} = \frac{\pi(\theta_{-j})\pi(\theta'_j \mid \theta_{-j})\, \pi(\theta_j \mid \theta_{-j})}{\pi(\theta_{-j})\pi(\theta_j \mid \theta_{-j})\, \pi(\theta'_j \mid \theta_{-j})} = 1. \quad \square$$

Conjugate priors (Node 6.2) yield tractable full conditionals, making Gibbs the canonical algorithm for hierarchical Bayesian models.

Effective Sample Size

MCMC samples are correlated. The **effective sample size** is

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

where $\rho_k = \mathrm{Corr}(h(\theta^{(t)}), h(\theta^{(t+k)}))$ is the lag-k autocorrelation. The Monte Carlo standard error for $\bar{h}_T = T^{-1}\sum_t h(\theta^{(t)})$ is

$$\mathrm{SE}(\bar{h}_T) = \sqrt{\frac{\mathrm{Var}_\pi(h)}{\mathrm{ESS}}}.$$

In practice, the infinite sum is truncated at the first negative autocorrelation or estimated using spectral methods.

Convergence Diagnostics

Gelman–Rubin $\hat{R}$ statistic. Run $M \geq 2$ chains of length T. Let B be the between-chain variance and W the within-chain variance. Define

$$\hat{R} = \sqrt{\frac{(T-1)/T \cdot W + B/T}{W}}.$$

As $T \to \infty$, $\hat{R} \to 1$ if all chains have converged to the stationary distribution. Values $\hat{R} > 1.01$ indicate incomplete convergence.

Trace plots provide visual assessment: a well-mixing chain shows rapid, uncorrelated fluctuations around a stable level. Persistent trends or multimodal behavior indicate convergence failure.

Autocorrelation function (ACF): slow decay of ρ_k indicates poor mixing and low ESS per iteration.

MCMC Central Limit Theorem

Under regularity conditions (geometric ergodicity), the MCMC ergodic average satisfies a CLT:

$$\sqrt{T}(\bar{h}_T - \mathbb{E}_\pi[h]) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\text{MCMC}}),$$

where the **asymptotic variance** is

$$\sigma^2_{\text{MCMC}} = \text{Var}_\pi(h) \cdot \left(1 + 2\sum_{k=1}^{\infty} \rho_k\right) = \text{Var}_\pi(h) \cdot \frac{T}{\text{ESS}}.$$

The factor $1 + 2\sum \rho_k$ is the **integrated autocorrelation time** (IACT). A well-mixing chain has small IACT and large ESS; a poorly mixing chain has large IACT and ESS much smaller than the nominal sample size T.

Burn-in and Thinning

Burn-in discards the first T_0 samples to reduce sensitivity to initialization. Formally, if the chain converges geometrically, the bias from initialization decays as $O(\rho^{T_0})$, where $\rho < 1$ is the spectral gap.

Thinning retains every m-th sample, reducing storage but not improving ESS per unit of computation. Thinning is generally unnecessary for estimation but may be practical for storage-limited applications.

Worked Examples

Example 1: Metropolis for a Beta Posterior

Let $X \sim \text{Binomial}(20, \theta)$ with $x = 13$ observed, and $\theta \sim \text{Beta}(1, 1)$ (uniform prior). The posterior is $\theta \mid x \sim \text{Beta}(14, 8)$. Use random walk Metropolis with proposal $\theta' = \theta^{(t)} + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, 0.05^2)$, truncated to $[0, 1]$. The acceptance ratio is $\alpha = \min(1, \theta'^{13}(1-\theta')^7 / [\theta^{(t)}]^{13}(1-\theta^{(t)})^7)$. After 10,000 iterations with 1,000 burn-in, the ergodic mean $\bar{\theta} \approx 0.636$ agrees with the exact posterior mean $14/22 \approx 0.636$, and a histogram of the chain matches the Beta$(14, 8)$ density.

Example 2: Gibbs Sampler for a Bivariate Normal

Let $(X, Y) \sim \mathcal{N}_2(0, \Sigma)$ with $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. The full conditionals are $X \mid Y = y \sim \mathcal{N}(\rho y, 1 - \rho^2)$ and $Y \mid X = x \sim \mathcal{N}(\rho x, 1 - \rho^2)$. The Gibbs sampler alternates between these. For $\rho = 0.9$, the lag-1 autocorrelation of the X-chain is $\rho^2 = 0.81$, so ESS $\approx T(1 - 0.81)/(1 + 0.81) = T \cdot 0.105$. For $T = 10{,}000$, ESS $\approx 1{,}050$. High correlation in the target distribution produces high autocorrelation in the chain, reducing effective sample size.

Cross-Links

- **Linear Algebra:** The mixing rate of a Metropolis chain on a discrete state space is governed by the spectral gap of the transition matrix; fast mixing requires the second-largest eigenvalue to be bounded away from 1.
- **AI:** Langevin MCMC (gradient of log-density added to random walk proposal) is used for sampling from energy-based models; stochastic gradient Langevin dynamics combines MCMC with mini-batch gradient estimation for scalable Bayesian deep learning.
- **Economics:** MCMC is used to estimate posterior distributions of structural parameters in DSGE models and Bayesian VARs where the likelihood is available only numerically.

- **Quantum:** Quantum Monte Carlo methods (path-integral Monte Carlo, diffusion Monte Carlo) use MCMC to sample from the Boltzmann distribution of quantum systems; the sign problem is a fundamental barrier analogous to multimodality in classical MCMC.

Structural Compression

Invariant: A Markov chain satisfying detailed balance with respect to π has π as its stationary distribution; ergodic averages converge almost surely to π-expectations.

Mechanism: The Metropolis acceptance probability is constructed so that detailed balance holds by construction. The ratio form $\pi(\theta')/\pi(\theta)$ cancels the normalizing constant, enabling sampling from distributions known only up to proportionality.

Cross-System Echo: MCMC converts integration into simulation, just as Monte Carlo integration converts integration into averaging. The structural pattern – replacing an analytically intractable global computation with an iterative local computation that converges to the global answer – appears throughout computational science: relaxation methods in PDE solvers, message passing in graphical models, and simulated annealing in combinatorial optimization are all instances of the same principle.

Exercises

1. Prove that detailed balance $\pi(\theta)K(\theta, \theta') = \pi(\theta')K(\theta', \theta)$ implies π is a stationary distribution of K, by integrating both sides over θ.
2. For a random walk Metropolis targeting $\pi(\theta) \propto e^{-\theta^4}$ on $\mathbb{R}$ with Gaussian proposal variance σ^2, describe qualitatively how the acceptance rate and mixing behavior change as σ^2 varies from 0.01 to 100.
3. Derive the full conditional distributions for a Bayesian normal model $X_i \mid \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$ with priors $\mu \sim \mathcal{N}(0, \tau^2)$ and $\sigma^2 \sim \text{Inv-Gamma}(a, b)$. Write the Gibbs sampling algorithm.
4. Show that the effective sample size satisfies $\text{ESS} \leq T$, with equality iff the chain produces independent samples (all autocorrelations are zero).
5. For the Gibbs sampler on a bivariate normal with correlation ρ, prove that the lag-1 autocorrelation of each coordinate is ρ^2, and derive the ESS as a function of T and ρ.
6. Implement the Gelman–Rubin diagnostic for 4 chains of length 5,000 targeting a $\text{Beta}(2, 5)$ distribution. Report $\hat{R}$ and compare it to the threshold 1.01.
7. Argue, structurally, why the normalizing-constant cancellation in the Metropolis ratio is the same structural principle as the sufficiency reduction in classical statistics: both allow inference to proceed without computing a quantity (the marginal likelihood or the normalizing constant) that depends on the full model but not on the parameter of interest.

Statistics · Module 9 — Computational Statistics · Node 9.3
Concepts: markov-chains, monte-carlo-methods, convergence, prior-posterior
Prerequisites: 9.2, 6.1
Enables: 9.4, 9.6

hbar.university — own.your.cognition