MCMC Sampling Problems

ELG 5218 Uncertainty Evaluation in Engineering Measurements and Machine Learning

Instructor: Miodrag Bolić, University of Ottawa

Date: January 28, 2026

PART A: CONCEPTUAL QUESTIONS

A1. Define a first-order Markov chain and stationary distribution.

Question.

Provide:

(a) The definition of a first-order Markov chain.

(b) The definition of a stationary distribution for a Markov chain with kernel $P(\theta' \mid \theta)$.

A2. State the ergodicity conditions and their consequence.

Question.

For a Markov chain with transition kernel $P(\theta' \mid \theta)$:

(a) Briefly define the concepts of irreducibility, aperiodicity, and positive recurrence.

(b) State the consequence of ergodicity for convergence to the stationary distribution.

A3. What is detailed balance and why is it useful in MCMC design?

Question.

State the detailed balance condition and explain why it is convenient when constructing MCMC algorithms such as Metropolis–Hastings.

PART B: METROPOLIS–HASTINGS DERIVATIONS

B1. Derive the Metropolis–Hastings acceptance ratio and evidence cancellation.

Question.

Let the target posterior be $\pi(\theta) = p(\theta \mid x)$, known up to a normalizing constant:

$$\pi(\theta) \propto \tilde{\pi}(\theta) = p(x \mid \theta)p(\theta).$$

Given a proposal density $q(\theta^* \mid \theta)$, the Metropolis–Hastings acceptance ratio is

$$r(\theta \to \theta^*) = \frac{\pi(\theta^*)q(\theta \mid \theta^*)}{\pi(\theta)q(\theta^* \mid \theta)}.$$

Show explicitly that the evidence term $p(x)$ cancels in this ratio.

B2. Random-walk Metropolis tuning intuition.

Question.
For a Gaussian random-walk Metropolis proposal

$$\theta^* = \theta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2_{\text{prop}}),$$

explain qualitatively:

(a) What happens when σ_{prop} is too small?

(b) What happens when σ_{prop} is too large?

PART C: CONVERGENCE DIAGNOSTICS

C1. Gelman–Rubin $\hat{R}$ statistic derivation (scalar case).

Question.
Suppose you run J independent chains, each of length L after burn-in, and obtain scalar draws $\theta_{j\ell}$, $j = 1, \dots, J$, $\ell = 1, \dots, L$. Define:

- Chain means: $\bar{\theta}_j = \frac{1}{L} \sum_{\ell=1}^{L} \theta_{j\ell}$.
- Grand mean: $\bar{\theta}_{\cdot} = \frac{1}{J} \sum_{j=1}^{J} \bar{\theta}_j$.
- Within-chain variances: $s_j^2 = \frac{1}{L-1} \sum_{\ell=1}^{L} (\theta_{j\ell} - \bar{\theta}_j)^2$.

Derive expressions for:

(a) Within-chain variance W.

(b) Between-chain variance B.

(c) The marginal variance estimate $\hat{V}^+ = \frac{L-1}{L} W + \frac{1}{L} B$.

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

C2. Interpreting trace plots and ACF.

Question.
Describe, qualitatively, what you expect to see in:

(a) A trace plot of a well-mixed chain.

(b) The autocorrelation function (ACF) of a poorly mixed chain.

PART D: MODERN SAMPLERS (HMC, NUTS) AND PRACTICAL ISSUES

D1. Intuition behind Hamiltonian Monte Carlo (HMC).

Question. Provide a concise explanation of how HMC reduces random-walk behavior compared to random-walk Metropolis (RWM).

D2. What problem does NUTS solve on top of HMC?

Question. Explain what key tuning parameter NUTS (No-U-Turn Sampler) eliminates compared to plain HMC, and how it does so conceptually.

PART E: OTHER PROBLEMS

E1. AR(1) Markov chain stationarity and convergence.

Consider the AR(1) process

$$x_{t+1} = \rho x_t + v_t, \quad v_t \sim \mathcal{N}(0, \sigma_v^2),$$

with $|\rho| < 1$ and x_0 arbitrary.

(a) Show that the stationary distribution is $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 = \sigma_v^2/(1-\rho^2)$.

(b) Briefly explain why, regardless of x_0, the distribution of x_t converges to this stationary distribution.

E2. MCMC for a biased coin (NUTS vs random-walk MH).

You flip a coin $n = 100$ times and observe $k = 90$ heads. Assume a Beta(α, β) prior with $\alpha = 5, \beta = 3$, and denote the coin bias by $\theta \in (0, 1)$.

(a) Write down the posterior $p(\theta \mid k, n)$.

(b) Suppose you run NUTS in NumPyro for θ and obtain a posterior mean $\hat{\theta}_{\text{NUTS}} = 0.88$ with ESS per chain ≈ 3700 (out of 2500 samples per chain across 4 chains, grouped). Comment on this.

(c) You also run a simple random-walk MH on the logit scale with acceptance rate ≈ 0.70, ESS per chain ≈ 1200. Both have $\hat{R} \approx 1.00$. Compare the two samplers.

E3. Single-chain convergence and ACF

You are given MCMC output for a scalar parameter θ from a single chain of length $N = 4000$ (after burn-in). The chain was generated by a random-walk Metropolis algorithm.

Summary diagnostics:

- Sample mean: $\hat{\theta} = 0.03$.
- Effective sample size (ESS): ESS ≈ 2660.
- Empirical autocorrelation function (ACF) for lags $\ell = 0, \ldots, 40$ (first few values):

$$\begin{aligned} \rho(0) &= 1.00, \\ \rho(1) &\approx 0.19, \\ \rho(2) &\approx 0.04, \\ \rho(3) &\approx 0.01, \\ \rho(\ell) &\approx 0 \quad \text{for } \ell \gtrsim 5. \end{aligned}$$

- A trace plot shows a stationary, "hairy" path around 0 with no visible trends.

Answer the following:

(a) Based on the trace and ACF, does the chain appear to have converged to its stationary distribution? Justify your answer.

(b) Comment on the mixing quality using the ACF and ESS.

(c) Would you recommend thinning this chain? Why or why not?

E4. Comparing chains via ACF and ESS

Two different MCMC algorithms (Algorithm G and Algorithm B) were used to sample from the same scalar posterior $p(\theta \mid x)$. Each produced a single chain of length $N = 4000$ after burn-in.

Diagnostics:

- Effective sample sizes:

$$\begin{aligned} \mathrm{ESS}_G &\approx 2660, \\ \mathrm{ESS}_B &\approx 128. \end{aligned}$$

- Empirical ACFs for lags 0–40 (selected values):
  - Algorithm G:

$$\begin{aligned} \rho_G(0) &= 1.00, \\ \rho_G(1) &\approx 0.19, \\ \rho_G(5) &\approx 0. \end{aligned}$$

  - Algorithm B:

$$\begin{aligned} \rho_B(0) &= 1.00, \\ \rho_B(1) &\approx 0.95, \\ \rho_B(10) &\approx 0.60, \\ \rho_B(20) &\approx 0.35, \\ \rho_B(40) &\approx 0.10. \end{aligned}$$

- Trace plots:
  - Algorithm G: rapid oscillations around the posterior bulk, no visible trend.
  - Algorithm B: very smooth, slowly drifting path; long stretches where the chain moves gradually in one direction.

Answer the following:

(a) Based on the ACFs and ESS, which chain mixes better? Explain in terms of autocorrelation and effective sample size.

(b) For Algorithm B, what can you conclude from the ACF plot about the correlation structure of the chain? How does this affect Monte Carlo error?

(c) Suppose both algorithms have similar posterior means and variances (i.e., no obvious bias). Which algorithm would you trust more for uncertainty quantification, and why?

E5. Metropolis–Hastings - Gamma proposal

For π the density of an inverse normal distribution with parameters $\theta_1 = 3/2$ and $\theta_2 = 2$,

$$\pi(x) \propto x^{-3/2} \exp\left(-\frac{3}{2}x - \frac{2}{x}\right) \mathbf{1}_{\{x>0\}},$$

write down and implement an independence Metropolis–Hastings sampler with a Gamma proposal with parameters $(\alpha, \beta) = (4/3, 1)$ and $(\alpha, \beta) = (0.5\sqrt{4/3}, 0.5)$.

E6. Metropolis–Hastings - Gaussian random walk

Consider the non-standard target distribution

$$p(x) \propto \exp\bigl(-x^4 + 3x^2 - x\bigr), \qquad x \in \mathbb{R},$$

and a Gaussian random-walk proposal

$$q(x'|x) = \mathcal{N}(x, \sigma_q^2).$$

Tasks:

1. Implement the Metropolis–Hastings algorithm.
2. Compare acceptance rates for $\sigma_q \in \{0.1, 0.5, 1, 2\}$.
3. Use the rule of thumb that the optimal acceptance rate in 1D is about 0.234.
4. Estimate the mean and variance from the samples.