Langevin and MALA Problems

ELG 5218 – Uncertainty Evaluation in Engineering Measurements and Machine Learning

Instructor: Miodrag Bolić, University of Ottawa

Date: January 28, 2026

Instructions

PART A: DRIFT–DIFFUSION BASICS

A1. Pure diffusion SDE: mean and variance

Consider the pure diffusion SDE in one dimension

$$dX_t = \sigma\, dW_t, \quad X_0 = 0, \quad \sigma > 0.$$

(a) Write down the solution X_t in terms of W_t.

(b) Derive $\mathbb{E}[X_t]$ and $\mathbb{V}(X_t)$.

(c) Give a short intuitive explanation of what the sample paths look like (in words).

A2. Pure drift SDE: no uncertainty

Consider the pure drift SDE

$$dX_t = \mu\, dt, \quad X_0 = x_0, \quad \sigma = 0.$$

(a) Solve for X_t.

(b) What is $\mathbb{V}(X_t)$?

(c) Intuitively, how do these sample paths differ from pure diffusion?

PART B: LANGEVIN DYNAMICS AND STATIONARITY

B1. Langevin SDE for a standard Gaussian

Let the target density be $p(x) = \mathcal{N}(0, 1)$, so

$$\log p(x) = -\tfrac{1}{2}x^2 + \text{const}, \quad \nabla_x \log p(x) = -x.$$

Consider the Langevin SDE

$$dX_t = \frac{1}{2}\nabla_x \log p(X_t)\, dt + dW_t.$$

(a) Write the SDE explicitly for this $p(x)$.

(b) Interpret the drift and diffusion terms.

(c) Intuition: Explain why you expect trajectories to "spend more time" near $x = 0$ than far in the tails.

B2. Stationary distribution (conceptual)

For the same SDE

$$dX_t = -\frac{1}{2} X_t \, dt + dW_t,$$

we know (from theory) that the stationary distribution is $\mathcal{N}(0, 1)$.

(a) In words, what does it mean for $\mathcal{N}(0, 1)$ to be the stationary distribution of this SDE?

(b) How does this connect to the idea of using Langevin dynamics for Monte Carlo sampling from $p(x)$?

PART C: FROM SDE TO DISCRETE-TIME: ULA

C1. Unadjusted Langevin Algorithm (ULA)

We discretize the Langevin SDE

$$dX_t = -\frac{1}{2} \nabla U(X_t) \, dt + dW_t,$$

using a step size Δt:

$$\theta_{k+1} = \theta_k - \frac{\Delta t}{2} \nabla U(\theta_k) + \sqrt{\Delta t} \, Z_k, \quad Z_k \sim \mathcal{N}(0, I).$$

(a) Explain why, for finite Δt, the stationary distribution of this Markov chain is not exactly $\pi(\theta) \propto e^{-U(\theta)}$.

(b) What happens if we take Δt very small? Discuss the bias–mixing trade-off qualitatively.

PART D: MALA – DERIVATION AND INTUITION

D1. MALA proposal for a Gaussian target

Let $\pi(\theta) = \mathcal{N}(0, 1)$ with potential $U(\theta) = \frac{1}{2}\theta^2 + \text{const}$, so $\nabla U(\theta) = \theta$. MALA uses the proposal

$$\theta^* = \theta - \frac{\Delta t}{2} \nabla U(\theta) + \sqrt{\Delta t} \, Z, \quad Z \sim \mathcal{N}(0, 1).$$

(a) Write the proposal distribution $q(\theta^* \mid \theta)$ in Gaussian form (mean and variance).

(b) State the Metropolis–Hastings acceptance probability in terms of π and q.

(c) Intuition: Compared to random-walk MH with $\theta^* = \theta + \epsilon$, why is this proposal better aligned with the target in this Gaussian case?

PART E: DATA-BASED MALA FOR LOGISTIC REGRESSION

In this section, you analyze synthetic data from a Bayesian logistic regression model and MALA sampling results.

Setup

We consider a binary outcome $y_i \in \{0,1\}$ with features $x_i \in \mathbb{R}^2$, $i = 1, \dots, n$. The model is

$$P(y_i = 1 \mid x_i, w) = \sigma(x_i^\top w), \quad \sigma(z) = \frac{1}{1+e^{-z}},$$

with prior

$$w \sim \mathcal{N}(0, \lambda^{-1} I_2), \quad \lambda = 1.$$

A simulated dataset with $n = 200$ observations is used; you do not need to reproduce the data. The potential energy is

$$U(w) = -\sum_{i=1}^{n} \Big[y_i \log \sigma(x_i^\top w) + (1-y_i) \log(1 - \sigma(x_i^\top w)) \Big] + \frac{\lambda}{2} \|w\|_2^2.$$

A MALA sampler is run with step size $\Delta t = 0.01$. Four independent chains of length $N = 4000$ (after burn-in) are obtained for each coefficient w_1, w_2.

Selected diagnostics (for w_1):

- Posterior mean (across all chains): $\hat{w}_1 \approx 1.35$.
- Posterior standard deviation: $\widehat{\mathrm{sd}}(w_1) \approx 0.20$.
- Gelman–Rubin $\hat{R}$ for w_1: $\hat{R} \approx 1.01$.
- Effective sample size (ESS) per chain: $\mathrm{ESS}_{\text{per chain}} \approx 1600$.
- ACF (single chain) for lags 0–30:

$$\rho(0) = 1.0, \quad \rho(1) \approx 0.25, \quad \rho(5) \approx 0.05, \quad \rho(10) \approx 0.01, \quad \rho(\ell) \approx 0 \text{ for } \ell > 10.$$

- Trace plots show "hairy," stationary paths across all four chains, exploring similar ranges.

E1. Convergence and mixing for w_1

(a) Based on $\hat{R}$ and ESS, does w_1 appear to have converged? Justify.

(b) Using the ACF information, comment on the mixing quality for w_1.

(c) Intuition: Why is MALA expected to mix better than a random-walk MH sampler on this logistic regression posterior?