Langevin and MALA Problems

ELG 5218 – Uncertainty Evaluation in Engineering Measurements and Machine Learning

Instructor: Miodrag Bolić, University of Ottawa

Date: January 28, 2026

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).

Solution A1.

(a) Integrating,

$$X_t = X_0 + \sigma W_t = \sigma W_t,$$

since $X_0 = 0$.

(b) Because $W_t \sim \mathcal{N}(0, t)$,

$$\mathbb{E}[X_t] = \sigma\mathbb{E}[W_t] = 0, \qquad \mathbb{V}(X_t) = \sigma^2\mathbb{V}(W_t) = \sigma^2 t.$$

(c) Paths fluctuate randomly around zero with no preferred direction. The mean stays at zero, but the variance (spread) grows linearly in time, so trajectories wander further away as t increases.

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?

Solution A2.

(a) Integrating,

$$X_t = X_0 + \mu t = x_0 + \mu t.$$

(b) There is no noise, so $\mathbb{V}(X_t) = 0$ for all t.

(c) All paths are identical straight lines with slope μ; there is no randomness. In contrast, pure diffusion produces many distinct random paths with zero drift and increasing variance.

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.

Solution B1.

(a) Using $\nabla_x \log p(x) = -x$, we get

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

(b) The drift term $-(1/2)X_t\, dt$ pulls the process back towards zero: when X_t is positive, the drift is negative, and vice versa. The diffusion term dW_t injects random noise, allowing exploration of the state space.

(c) Because the drift always points back toward the origin, trajectories that wander into the tails are "pulled" back. Noise occasionally kicks them away from the mode, but the restoring drift reduces the time spent in the tails. As a result, paths tend to spend more time near $x = 0$, which is consistent with the peak of the Gaussian density.

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 $p(x)$?

Solution B2.

(a) Stationarity means that if X_0 is distributed as $\mathcal{N}(0,1)$, then X_t is also $\mathcal{N}(0,1)$ for all t. Moreover, regardless of the initial distribution, the distribution of X_t converges to $\mathcal{N}(0,1)$ as $t \to \infty$.

(b) If we simulate the SDE for a long time, the marginal distribution of X_t approaches $p(x) = \mathcal{N}(0,1)$. Thus, recording X_t at large times produces approximate samples from $p(x)$, which we can use for Monte Carlo estimation of expectations under p.

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.

Solution C1.

(a) The discrete-time update is an Euler–Maruyama approximation to the continuous-time SDE. This numerical scheme introduces discretization error: the transition kernel of ULA is only an approximation to that of the true Langevin diffusion. As a result, the stationary distribution of the Markov chain is $\pi_{\Delta t}(\theta)$, which differs from $\pi(\theta)$, so ULA is biased.

(b) As $\Delta t \to 0$, the discrete chain better approximates the continuous SDE, so the stationary distribution $\pi_{\Delta t}$ converges to the true π. However, smaller step size means smaller moves and slower exploration: mixing deteriorates, and more iterations are needed to cover the posterior. Larger Δt increases mixing speed but also increases bias, because the Euler approximation becomes cruder.

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?

Solution D1.

(a) Here $\nabla U(\theta) = \theta$, so

$$\theta^* = \theta - \frac{\Delta t}{2}\theta + \sqrt{\Delta t}\, Z = \left(1 - \frac{\Delta t}{2}\right)\theta + \sqrt{\Delta t}\, Z.$$

Thus

$$q(\theta^* \mid \theta) = \mathcal{N}\left(\theta^* \;\middle|\; \left(1 - \frac{\Delta t}{2}\right)\theta,\ \Delta t\right).$$

(b) The MH acceptance probability is

$$\alpha(\theta \to \theta^*) = \min\left(1,\ \frac{\pi(\theta^*)\, q(\theta \mid \theta^*)}{\pi(\theta)\, q(\theta^* \mid \theta)}\right).$$

(c) The drift term $-(\Delta t/2)\nabla U(\theta) = -(\Delta t/2)\theta$ pulls proposals toward the mode at 0. So proposals tend to move along directions of increasing posterior density (instead of proposing blindly). For a Gaussian target, this is well aligned with the true structure, leading to higher acceptance rates and longer, more directed moves in high-density regions than an isotropic random-walk proposal.

PART E: DATA-BASED MALA FOR LOGISTIC REGRESSION

E1. Convergence and mixing for w_1

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

(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?

Solution E1.

(a) $\hat{R} \approx 1.01$ indicates that between- and within-chain variability agree closely, a standard threshold for convergence. ESS per chain of about 1600 out of 4000 iterations suggests a high fraction of effectively independent draws. Together, these are strong evidence that the chains for w_1 have converged to the target posterior.

(b) The ACF decays quickly: $\rho(1) \approx 0.25$, and correlations are near zero by lag 10. This rapid decay implies a small integrated autocorrelation time and hence a large ESS. The trace plots show good exploration with no visible trends, consistent with efficient mixing.

(c) MALA uses the gradient $\nabla U(w)$ to propose updates that move along directions of increasing posterior density, combining gradient descent with noise. For logistic regression, the posterior is relatively smooth and differentiable; gradient information helps the sampler align proposals with the local curvature, avoiding the diffusive behavior of random-walk MH. This leads to larger, more informed moves with higher acceptance rates and lower autocorrelation.