Computational Statistics & Machine Learning

Lecture 12

Langevin Markov chain Monte Carlo

Mark Girolami
mag92@cam.ac.uk

Department of Engineering
University of Cambridge

July 25, 2025

Overview

- Langevin SDEs and their stationary measures
- Metropolis-adjusted Langevin Algorithm (MALA)
- Unadjusted Langevin Algorithm (ULA)
- An Overview of Modern Bayesian ML Applications

The Langevin SDE

In many branches of sciences, the following Langevin SDE is of special interest:

$$\mathrm{d}X_t = -\nabla U(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,$$

where $U(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is called a potential (e.g. negative log-target measure) and $(B_t)_{t\geq 0}$ is the Brownian motion.

- Derived by Paul Langevin in 1908 to model the random movement of a particle in a fluid.

The Langevin SDE

The Langevin SDE is very popular in the sampling literature as for the SDE

$$\mathrm{d}X_t = -\nabla U(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,$$

the stationary measure is

$$\pi \propto \exp(-U(x)).$$

In other words, in order to sample from π, one can just set

$$U(x) = -\log \pi(x) \qquad \text{and} \qquad \nabla U(x) = -\nabla \log \pi(x).$$

From now on, we use the notation

$$\mathrm{d}X_t = \nabla \log \pi(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t.$$

The Langevin SDE

The measure of $X_t \sim \rho_t$ satisfies the following Fokker-Planck equation

$$\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t \nabla \log \pi) + \Delta \rho_t.$$

The asymptotic solution is

$$\rho_\infty = \pi.$$

To see this, check with direct substitution (i. e. $\partial \rho_t = 0$). Simulating this SDE will provide samples from π in the limit.

- Use it as an MCMC scheme.

The Langevin SDE

Can we say anything more specific about the convergence?

- We can – if we limit the class of targets we are interested in.
- We restrict ourselves to the broad class of log-concave distributions.

Definition

A probability measure with density π is log-concave if

$$\log \pi(\lambda x + (1 - \lambda y) \geq \lambda \log \pi(x) + (1 - \lambda) \log \pi(y).$$

for $0 < \lambda < 1$.

More compactly $\pi(\lambda x + (1 - \lambda)y) \geq \pi(x)^{\lambda} \pi(y)^{1-\lambda}$.

The Langevin SDE

We also focus further on m-strongly log-concave measures as they will allow us to obtain precise convergence rates.

Definition

A probability measure with density π is m-strongly log-concave if

$$\log \pi(\lambda x + (1-\lambda)y) \geq \lambda \log \pi(x) + (1-\lambda) \log \pi(y) + \frac{m\lambda(1-\lambda)}{2}\|x-y\|^2,$$

for $0 < \lambda < 1$.

This is a stronger notion, satisfied by e.g. Gaussians (log-Gaussian = quadratic function).

- It basically means there are no flat regions and the maximum is unique. So for one dimension $\nabla^2 \log \pi(x) \leq -m < 0 \; \forall \; x$.

The Langevin SDE

The Langevin SDE has many nice properties

- For an m-strongly-log-concave π, the convergence rate to the stationary measure is exponentially fast[1]:

$$W_2(\pi_t, \pi) \leq e^{-mt} \left\{ \|x - x^\star\| + (d/m)^{1/2} \right\},$$

  for $\pi_0 = \delta_x$ and $x^\star = \arg\max_{x \in \mathbb{R}^d} \log \pi(x)$.
- It has convergence guarantees under even more general assumptions.
- Convergence properties are often reflected in practical approximations of the SDE.

[1] W_2 is a distance on the space of probability measures.

Metropolis-adjusted Langevin Algorithm

- However, the continuous-time SDE can't be simulated exactly:
  - Discretisation schemes are needed which introduce bias.
  - The bias can be corrected via a Metropolis step.
- Among many candidates, the first obvious one is the first-order Euler discretisation, i.e., sample $X_0 \sim \pi_0$ and

$$X_{k+1} = X_k + \gamma \nabla \log \pi(X_k) + \sqrt{2\gamma} W_{k+1},$$

where $(W_k) \geq 0$ are standard Normal and $\gamma > 0$ is the step-size.

Metropolis-adjusted Langevin Algorithm

- The bias can be corrected by "Metropolising" the discretisation,
  - Given a state of the Markov chain X_k, propose

$$\overline{X}_{k+1} = X_k + \gamma \nabla \log \pi(X_k) + \sqrt{2\gamma} W_{k+1}.$$

  - Accept the sample with probability

$$\alpha := \min \left\{ 1, \frac{\pi(\overline{X}_{k+1}) q(X_k | \overline{X}_{k+1})}{\pi(X_k) q(\overline{X}_{k+1} | X_k)} \right\},$$

where

$$q(x' | x) \propto \exp\left(-\frac{1}{4\gamma} \|x' - x - \gamma \nabla \log \pi(x)\|^2 \right).$$

Metropolis-adjusted Langevin Algorithm

- The MALA targets the measure π exactly as it satisfies detailed balance.
- Well studied theoretically:
  - Step-size has to be tuned so that $\alpha = 0.574$.
- But may suffer in high dimension (d)
  - Acceptance probabilities may get exponentially slower in d, which results in poor mixing
- Can we skip the Metropolis step?

Unadjusted Langevin Algorithm

We can run a Langevin SDE discretisation without the Metropolis steps, i.e.

$$\overline{X}_{k+1} = X_k + \gamma \nabla \log \pi(X_k) + \sqrt{2\gamma} W_{k+1}.$$

Some facts:

- This chain converges to a stationary measure π_γ which is **not** the target measure π (bias).
- On the other hand, it avoids a Metropolis step and performs well in high dimensions.

Relevant questions:

- How bad is the (nonasymptotic) bias?
- How fast does this method converge?

Unadjusted Langevin Algorithm

Let us look at the Gaussian target

$$\pi(x) = \mathcal{N}(x; \mu, \Sigma).$$

The Langevin SDE in this case is the Ornstein-Uhlenbeck (OU) process

$$\mathrm{d}X_t = -\Sigma^{-1}(X_t - \mu)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t.$$

The ULA scheme is given by

$$X_{k+1} = X_k - \gamma\Sigma^{-1}(X_k - \mu) + \sqrt{2\gamma}W_{k+1}.$$

Under appropriate conditions, this chain has the invariant measure

$$\pi_\gamma = \mathcal{N}\left(\mu, \Sigma(I - \frac{\gamma}{2}\Sigma^{-1})^{-1}\right).$$

Unadjusted Langevin Algorithm

What happens more generally?

- Results for a broad class of measures can be characterised if we look into *log-concave* measures.
- The convergence rate of the ULA with $X_0 \sim \pi_0$, for $\gamma \leq 2/(m+L)$,

$$W_2(\pi_k, \pi) \leq (1 - m\gamma)^k W_2(\pi_0, \pi) + 1.65(L/m)(\gamma d)^{1/2},$$

  for m-strongly-log-concave π with L-Lipschitz gradients.
- The asymptotic bias is of order $\mathcal{O}(\gamma^{1/2})$.
- It can be made arbitrarily small with small step-sizes.

These guarantees can be extended to non-log-concave distributions.