401-4634-24L: Diffusion Models, Sampling and Stochastic Localization

Lecture 3 – Langevin algorithms

Lecturer: Yuansi Chen *Spring 2024*

Key concepts:

- Sampling from a smooth density
- Langevin diffusion
- Unadjusted Langevin Algorithm (ULA) and Metropolis-adjusted Langevin Algorithm (MALA)
- Convergence of continuous Langevin diffusion
  - in Wasserstein distance, using strong-logconcavity via coupling
  - in χ^2 distance, using Poincaré inequality via Fokker Planck equation
- (note is here but only discussed in the 4th lecture) Convergence of ULA

The material of this lecture is based on Chapter 1 and 4 of [Che23].

3.1 Introduction

In the previous lecture, we saw that the corners of the convex body cause a lot of problems for the sampling algorithm Ball walk: Ball walk has to choose a smaller step-size otherwise it would have close-to-zero acceptance rate in many places. It is not always the case in practice that we encounter distributions that are as nonsmooth as the uniform distribution on a convex body. In this lecture, we completely avoid the nonsmooth problem by making a simplifying assumption that we are dealing with smooth densities of the form

$$\mu \propto e^{-f}$$

where f is twice continuously differentiable. We would like to know whether there exist sampling algorithms better than Ball walk.

3.1.1 Langevin diffusion

Given $f : \mathbb{R}^n \to \mathbb{R}$ a twice-differentiable function, the **Langevin diffusion** is the following stochastic differential equation (SDE)

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t, \tag{3.1}$$

where B_t is the Brownian motion in $\mathbb{R}^n$.

Brownian motion. We define Brownian motion in $\mathbb{R}^n$, denoted by $\{B_t\}_{t\geq 0}$, to be a stochastic process, i.e. a sequence of random variables in $\mathbb{R}^n$ indexed by $t \geq 0$, satisfying the following four properties

1. $B_0 = 0$
2. $\{B_t\}_{t\geq 0}$ is continuous with probability 1
3. (independent increment) For any $k \in \mathbb{N}$ and $\{t_i\}_{i=0}^k$ with $t_0 = 0 < t_1 < \cdots < t_k$, the random variables $B_{t_{i+1}} - B_{t_i}$ for $0 \leq i \leq k-1$, are mutually independent
4. (Gaussian increment) for any $0 \leq s < t$, $B_t - B_s$ is distributed as $\mathcal{N}(0, (t-s)\mathbb{I}_n)$.

Intuitively, we may think dB_t in Eq. (3.1) is Gaussian noise with mean 0 and variance dt. Then Eq. (3.1) may be thought of as a noisy gradient descent, with a deterministic gradient step $-\nabla f(X_t)dt$ and a diffusion component $\sqrt{2}dB_t$. For a rigorous treatment of Brownian motion and stochastic calculus, readers are referred to [Pro04].

3.1.2 Sampling algorithms connected to Langevin diffusion

Langevin diffusion in Eq (3.1) is a continuous process. To simulate it in practice, we need to discretize it.

Unadjusted Langevin Algorithm (ULA). Unadjusted Langevin Algorithm is the outcome of Euler discretization of the Langevin diffusion. Starting for X_0 drawn from an initial distribution, it iterates as follows: from the current state X_k, it produces the next state by

$$X_{k+1} = X_k - h\nabla f(X_k) + \sqrt{2h}\xi_k \tag{3.2}$$

where $h > 0$ is the step-size (or the discretization size) to be chosen by the user and $\xi_k \sim \mathcal{N}(0, \mathbb{I}_n)$ is independent Gaussian noise. Intuitively, taking the limit $h \to 0$ in ULA would get us back to the Langevin diffusion in Eq (3.1). For small step-size and large k, we expect the distribution of X_k to be close to the stationary measure of the Langevin diffusion (hopefully the target measure μ, but we haven't proved it yet) with an error that depends on h.

Metropolis-adjusted Langevin Algorithm (MALA). To ensure that a Markov chain has the correct stationary measure, we can always add a Metropolis-Hastings filter (or accept-reject step) to it. This is what Metropolis-adjusted Langevin Algorithm does in addition to ULA. It iterates as follows: from the current state X_k, it has a proposal step and an accept-reject step

- Proposal step: same as in ULA

$$Z_{k+1} = X_k - h\nabla f(X_k) + \sqrt{2h}\xi_k$$

- Accept-reject step: go to

$$X_{k+1} = \begin{cases} Z_{k+1} & \text{with probability } \min\left\{1, \frac{\mu(Z_{k+1})\mathcal{P}_{Z_{k+1}}(X_k)}{\mu(X_k)\mathcal{P}_{X_k}(Z_{k+1})}\right\} \\ X_k & \text{with the remaining probability.} \end{cases}$$

Note that conditioned on X_k, the proposal step boils down to drawing a Gaussian with mean $X_k - h\nabla f(X_k)$ and covariance $2h\mathbb{I}_n$. Hence, the proposal kernel has an explicit form

$$\mathcal{P}_z(x) = \frac{1}{(2\pi \cdot 2h)^{\frac{n}{2}}} \exp\left(-\frac{\|x - (z - h\nabla f(z))\|_2^2}{4h}\right).$$

Then, the acceptance rate also has an explicit form

$$\begin{aligned} &\min\left\{1, \frac{\mu(z)\mathcal{P}_z(x)}{\mu(x)\mathcal{P}_x(z)}\right\} \\ &= \min\left\{1, \exp\left(-f(z) - \frac{1}{4h}\|x - (z - h\nabla f(z))\|_2^2 + f(x) + \frac{1}{4h}\|z - (x - h\nabla f(x))\|_2^2\right)\right\}. \end{aligned}$$

In addition to one gradient evaluation step in the proposal step, MALA requires two more gradient evaluation steps and two function evaluation steps per iteration.

Metropolized random walk (MRW). We can always introduce a Ball-walk-like sampling algorithm for sampling a smooth density. In each iteration, it has a Gaussian proposal followed by an accept-reject step.

- Proposal step:

$$Z_{k+1} = X_k + \sqrt{2h}\xi_k.$$

- Accept-reject step: go to

$$X_{k+1} = \begin{cases} Z_{k+1} & \text{with probability } \min\left\{1, \frac{\mu(Z_{k+1})\mathcal{P}^{\mathrm{MRW}}_{Z_{k+1}}(X_k)}{\mu(X_k)\mathcal{P}^{\mathrm{MRW}}_{X_k}(Z_{k+1})}\right\} \\ X_k & \text{with the remaining probability.} \end{cases}$$

Here, because of the symmetry of the proposal $\mathcal{P}^{\text{MRW}}$, $\mathcal{P}^{\text{MRW}}_{Z_{k+1}}(X_k)$ cancels with $\mathcal{P}^{\text{MRW}}_{X_k}(Z_{k+1})$. The acceptance rate boils down to $\min\left\{1, \frac{\mu(Z_{k+1})}{\mu(X_k)}\right\}$.

Main questions. We are interested in the convergence of the continuous Langevin diffusion and the three sampling algorithms for sampling a smooth density. In this lecture, we ask the following three main questions, and we try to answer in the next section

1. What is the stationary measure of Langevin diffusion (3.1)? We hope it to be $\mu \propto e^{-f}$.
2. How fast does Langevin diffusion converge to its stationary measure?
3. What is the mixing time of ULA?

3.2 Convergence of Langevin diffusion

Because both sampling algorithms ULA and MALA are closely related to the Langevin diffusion, it is natural to make use of the convergence of the Langevin diffusion in continuous time to analyze the two discrete-time algorithms. We call it SDE-based mixing proof technique, in contrast to the conductance-based mixing proof technique in Lecture 2.

We first introduce the Fokker-Planck equation associated with the Langevin diffusion in Eq. (3.1), assume its correctness, and then analyze the Langevin diffusion based on it. Once we have a good understanding of the convergence of Langevin diffusion, the mixing time analysis of ULA follows from a careful discretization analysis.

3.2.1 Fokker-Planck equation

Consider a drift-diffusion process $\{X_t\}_{t\geq 0}$ on $\mathbb{R}$ driven by a drift term $a : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and a diffusion term $b : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, and characterized by the following SDE

$$dX_t = a(X_t, t)dt + b(X_t, t)dB_t, \tag{3.3}$$

where B_t is the Brownian motion in $\mathbb{R}$. We assume the following fact without proving it.

Fokker-Planck equation. Let $\{X_t\}_{t\geq 0}$ be a drift-diffusion process following SDE (3.3), starting from $X_0 \sim \mu_0$. Then for all $t \geq 0$, denoting the law of X_t by μ_t, we have

$$\frac{\partial}{\partial t}\mu_t(x) = -\frac{\partial}{\partial x}\left[a(x,t)\mu_t(x)\right] + \frac{\partial^2}{\partial^2 x}\left[D(x,t)\mu_t(x)\right], \forall x \in \mathbb{R}, \tag{3.4}$$

where $D(x,t) = b(x,t)^2/2$. The above equation is called the Fokker-Planck equation associated to the drift-diffusion process $\{X_t\}_{t\geq 0}$. The Fokker-Planck equation describes the time evolution of the probability density function via a partial differential equation (PDE). Unlike Eq. (3.3), the Fokker-Planck equation in Eq. (3.4) is completely deterministic.

In general, there are two main approaches to interpret a drift-diffusion process in Eq. (3.3) as illustrated in Figure 3.1. The first approach is the pathwise view: given a random draw of the Brownian motion B_t, Eq. (3.3) becomes an ordinary differential equation and, it generates a continuous path in $\mathbb{R}$. Each random draw of the Brownian motion generates a path. The collection of all paths describes the SDE. The second approach is the density evolution view: since we don't really care about the identity of each path, we can focus on the evolution of the law of the density of X_t at any time $t > 0$. Fokker-Planck equation enables this second approach via a PDE. The two approaches are complimentary and are related via Markov semigroup theory and Kolmogorov's forward and backward equations. For a detailed exposition and a proof of the Fokker-Planck equation, see Chapter 1.2 of [Che23].

Example 1 (heat equation)**.** *Taking $a = 0, b = 1$ in Eq. (3.4), we obtain the heat equation*

$$\frac{\partial}{\partial t}\mu_t = \frac{1}{2}\frac{\partial^2}{\partial x^2}\mu_t.$$

Starting from 0, the PDE has a closed-form solution

$$\mu_t = \frac{1}{\sqrt{2\pi t}}\exp\left(-\frac{x^2}{2t}\right).$$

The above density is exactly the law of X_t defined via $dX_t = dB_t$.

Finally, one can also introduce a higher dimensional formulation of the Fokker-Planck equation. Consider a drift-diffusion process $\{X_t\}_{t\geq 0}$ on $\mathbb{R}^n$ driven by a drift term $\mathbf{a} : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ and a diffusion term $\mathbf{b} : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^{n\times m}$, characterized by the following SDE

$$dX_t = \mathbf{a}(X_t,t)dt + \mathbf{b}(X_t,t)dB_t, \tag{3.5}$$

where B_t is the Brownian motion in $\mathbb{R}^m$.

Fokker-Planck equation in n-dimension. Let $\{X_t\}_{t\geq 0}$ be a drift-diffusion process following SDE (3.5), starting from $X_0 \sim \mu_0$. Then for all $t \geq 0$, denoting the law of X_t by μ_t, we have

$$\frac{\partial}{\partial t}\mu_t(x) = -\sum_{i=1}^{n}\frac{\partial}{\partial x_i}\left[\mathbf{a}_i(x,t)\mu_t(x)\right] + \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2}{\partial x_i \partial x_j}\left[\mathbf{D}_{ij}(x,t)\mu_t(x)\right], \forall x \in \mathbb{R}^n, \tag{3.6}$$

where $\mathbf{D} = \frac{1}{2}\mathbf{b}\mathbf{b}^\top$.

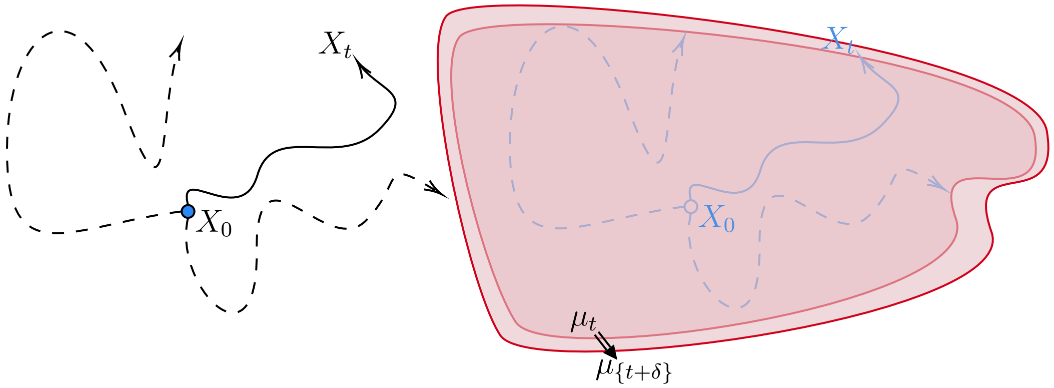

Figure 3.1. Two interpretations of a drift-diffusion process. Left: pathwise view. Right: density evolution view.

Example 2 (Langevin diffusion)**.** *Taking* $\mathbf{a}(x,t) = -\nabla f(x)$ *and* $\mathbf{b}(x,t) = \sqrt{2}\mathbb{I}_n$*, the SDE corresponds to Langevin diffusion*

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t.$$

The associated Fokker-Planck equation is

$$\frac{\partial}{\partial t}\mu_t = \nabla \cdot (\mu_t \nabla f) + \Delta \mu_t. \tag{3.7}$$

Differential operator notation.

- The *divergence* of a continuously differentiable vector function $F : \mathbb{R}^n \to \mathbb{R}^n$ is

  $$\nabla \cdot F = \sum_{i=1}^{n} \frac{\partial}{\partial x_i} F_i,$$

  where $F_i : \mathbb{R}^n \to \mathbb{R}$ is the i-th coordinate output of F.

- The *Laplacian* of a twice-differentiable function $g : \mathbb{R}^n \to \mathbb{R}$ is

  $$\Delta g = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2} g.$$

  Note that it is also the divergence of the gradient (∇g), i.e., $\Delta g = \nabla \cdot \nabla g$.

3.2.2 The stationary measure of Langevin diffusion

Assuming the correctness of the Fokker-Planck equation (3.6), we are ready to show that μ is a stationary measure of Langevin diffusion. To show that μ is a stationary measure, it suffices to show that

$$\frac{\partial}{\partial t}\mu_t$$

vanishes pointwise when μ_t is evaluated at μ. We already know the Fokker-Planck equation of Langevin diffusion in Eq. (3.7). It remains to show that

$$0 \overset{?}{=} \nabla \cdot (\mu \nabla f) + \Delta \mu.$$

We have by definition of the divergence

$$\nabla \cdot (\mu \nabla f) = \sum_{i=1}^{n} \partial_i (\mu \cdot \partial_i f),$$

and

$$\begin{aligned} \Delta \mu &= \sum_{i=1}^{n} \partial_i^2 \mu \\ &\overset{(i)}{=} \sum_{i=1}^{n} \partial_i (-\mu \cdot \partial_i f) \\ &= -\sum_{i=1}^{n} \partial_i (\mu \cdot \partial_i f), \end{aligned}$$

where ∂_i is used as a shorthand for $\frac{\partial}{\partial x_i}$, (i) used the assumption $\mu = ce^{-f}$ with c a constant. So, the two terms above sum to 0. And we prove μ is a stationary measure.

3.2.3 Convergence of Langevin diffusion in Wasserstein distance

We prove the convergence of Langevin diffusion in Wasserstein distance under strong logconcavity.

Wasserstein distance. Let μ, ν be two measures on $\mathbb{R}^n$ with finite second moments, i.e., $\mathbb{E}_{X\sim\nu}[\|X\|_2^2] < \infty$ and $\mathbb{E}_{X\sim\mu}[\|X\|_2^2] < \infty$. We define the Wasserstein-2 distance between μ and ν by

$$W_2(\mu, \nu) := \left(\inf_{\gamma \in \mathcal{C}(\mu,\nu)} \int \|x - y\|_2^2 \, \gamma(x, y) dx dy \right)^{\frac{1}{2}},$$

where $\mathcal{C}(\mu, \nu)$ is the set of all couplings of μ and ν. We say γ is a *coupling* of μ and ν, if its marginal on the first variable is μ and its marginal on the second is ν.

Strong logconcavity. We say a measure μ is m-strongly logconcave if $\mu \propto \exp(-f)$ with f being m-strongly convex, i.e., $m\mathbb{I}_n \preceq \nabla^2 f$.

Theorem 3.2.1 (Convergence of Langevin diffusion in Wasserstein distance)**.** *Let $\{X_t\}_{t\geq 0}$ be generated according to the Langevin diffusion (3.1) with initialization $X_0 \sim \mu_0$ and stationary measure $\mu \propto e^{-f}$. Assume μ is m-strongly logconcave. Let μ_t denote the law of X_t, then*

$$W_2^2(\mu_t, \mu) \leq \exp(-2mt) W_2^2(\mu_0, \mu).$$

Proof. The main proof strategy is to construct a coupling between μ_t and μ by taking advantage of the Langevin SDE (3.1), and then prove that it is bounded. We construct a coupling as follows. Let γ_0 be an optimal coupling of (μ_0, μ) which achieves $W_2^2(\mu_0, \mu)$. Draw $(X_0, X_0^*) \sim \gamma_0$. Let X_0 and X_0^* evolve through the Langevin SDE with the same copy of Brownian motion $\{B_s\}_{s\geq 0}$. Let γ_t denote the law of the resulting (X_t, X_t^*). γ_t is a coupling of (μ_t, μ) because

- Marginally, we just followed the Langevin SDE. So the law of X_t is μ_t
- μ is a stationary measure, so the law of X_t^* remains μ.

Next, we control the $\mathbb{E}_{(X_t, X_t^*)\sim\gamma_t}\left[\|X_t - X_t^*\|_2^2\right]$. We have

$$\begin{aligned} d\|X_t - X_t^*\|_2^2 &= 2\langle X_t - X_t^*, dX_t - dX_t^*\rangle \\ &\overset{(i)}{=} -2\langle X_t - X_t^*, \nabla f(X_t) - \nabla f(X_t^*)\rangle\, dt \\ &\overset{(ii)}{\leq} -2m\|X_t - X_t^*\|_2^2\, dt. \end{aligned} \tag{3.8}$$

(i) uses the fact that X_t and X_t^* shared the same Brownian motion. (ii) uses the mean value theorem in the following way:

$$\begin{aligned} \langle y - x, \nabla f(y) - \nabla f(x)\rangle &= \langle y - x, \nabla f(\omega_t) - \nabla f(\omega_0)\rangle\,|_{t=1} \\ &\overset{(iii)}{=} \left\langle y - x, \nabla^2 f(\omega_\tau)(y - x)\right\rangle \\ &\geq m\|y - x\|_2^2, \end{aligned}$$

where $\omega_t = (1-t)x + ty$, (iii) uses the mean value theorem for the function $t \to \langle y - x, \nabla f(\omega_t) - \nabla f(\omega_0)\rangle$ with the derivative $\langle y - x, \nabla^2 f(\omega_t)(y - x)\rangle$. So there exists $\tau \in [0,1]$ such that (iii) holds. The last step follows from m-strong concavity.

Solving the ODE inequality (3.8) or applying Grönwall's inequality, we obtain

$$\|X_t - X_t^*\|_2^2 \leq \exp(-2mt)\|X_0 - X_0^*\|_2^2.$$

Taking expectation on both sides, we obtain

$$\mathbb{E}_{\gamma_t} \|X_t - X_t^*\|_2^2 \le \exp(-2mt)\mathbb{E}_{\gamma_0} \|X_0 - X_0^*\|_2^2 = \exp(-2mt)W_2^2(\mu_0, \mu).$$

We complete the proof by noticing that γ_t is one coupling and $W_2^2(\mu_t, \mu)$ takes the infimum over all couplings. □

The following results show that in sampling a strongly log-concave density, it is not hard to have a reasonable control of the initial Wasserstein distance.

Lemma 1. *Let $\mu \propto e^{-f}$, where f is m-strongly convex and minimized at x^*. Then*

$$\mathbb{E}_{X\sim\mu} \|X - x^*\|_2^2 \le \frac{2n}{m}.$$

Remark that when f satisfies $m\mathbb{I}_n \preceq \nabla^2 f \preceq L\mathbb{I}_n$, x^* can be obtained up to ϵ-error in $\frac{L}{m}\log(1/\epsilon)$ iterations via gradient descent method (see e.g., [B+15]).

Proof. Let $\mu = c\exp(-f)$, where c is a constant. We have

$$\begin{aligned}\mathbb{E}_{X\sim\mu} \|X - x^*\|_2^2 &= c\int \|x - x^*\|_2^2 \exp(-f(x))dx \\ &\overset{(i)}{\le} \frac{2c}{m}\int \langle \nabla f(x), x - x^*\rangle \exp(-f(x))dx \\ &\overset{(ii)}{=} \frac{2c}{m}\int \operatorname{trace}(\mathbb{I}_n)\exp(-f(x))dx \\ &= \frac{2n}{m}.\end{aligned}$$

(i) follows from the strong convexity of f: $\langle \nabla f(x) - \nabla f(x^*), x - x^*\rangle \ge \frac{m}{2}\|x - x^*\|_2^2$ and $\nabla f(x^*) = 0$. (ii) follows from integration by parts: for a differentiable function $g : \mathbb{R}^n \to \mathbb{R}$ and vector field $v : \mathbb{R}^n \to \mathbb{R}^n$ with sufficiently fast decay at infinity, we have

$$\int \langle v(x), \nabla g(x)\rangle\, dx = -\int g(x)(\nabla \cdot v)(x)dx. \tag{3.9}$$

We integrated over ∇g, differentiated over v and the boundary term is 0 because of the decay at infinity.

□

3.2.4 Convergence of Langevin diffusion in χ^2-divergence

To show that the above convergence is not mainly due to the choice of Wasserstein distance, we prove the convergence of Langevin diffusion in χ^2-divergence under Poincaré inequality.

χ^2-divergence. Let ν, μ be two measures on $\mathbb{R}^n$. We define the χ^2-divergence between ν and μ by

$$\chi^2(\nu \parallel \mu) := \mathrm{Var}_\mu \left[\frac{\nu}{\mu}\right] = \int \left(\frac{\nu(x)}{\mu(x)}\right)^2 \mu(x)dx - 1.$$

The χ^2-divergence is an upper bound of the total variation distance because of Cauchy-Schwarz inequality.

Poincaré inequality. We say a measure μ satisfies a Poincaré inequality with constant C_{PI} if for all differentiable and square integrable function with respect to ν, we have

$$\mathrm{Var}_\mu[g] \leq C_{\mathrm{PI}} \mathbb{E}_\mu \left\|\nabla g(x)\right\|_2^2.$$

Here $\mathbb{E}_\mu$ and Var_μ denote the expectation with respect to μ and variance with respect to μ respectively

$$\mathbb{E}_\mu[g] := \int g(x)\mu(x)dx,$$
$$\mathrm{Var}_\mu[g] := \mathbb{E}_\mu[g^2] - \left(\mathbb{E}_\mu[g]\right)^2.$$

Similar to the isoperimetry in Lecture 2, Poincaré inequality is also an intrinsic property of the measure μ, and this definition has nothing to do with the sampling algorithm. Intuitively, the Poincaré constant being large also indicates that the measure μ has a bottleneck (see Figure 3.2). Additionally, the isoperimetric constant and Poincaré constant are related as $\psi \leq \frac{2}{C_{\mathrm{PI}}}$ according to [Maz60] and [Che69].

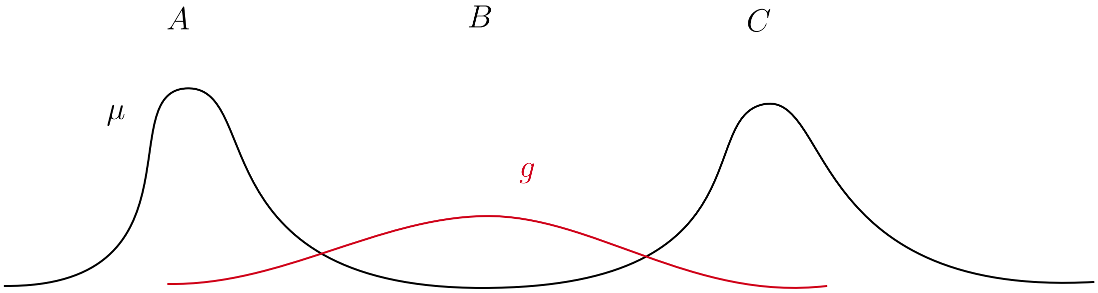

Figure 3.2. Illustration of large Poincaré constant. μ is bimodal with one mode in region A and the other mode in region C. It has a bottleneck in region B where density is close to 0. When the two modes are far away, it becomes possible to design a g with small gradient, hiding inside region B and has a large variance. In this case, the Poincaré constant of μ has to be large.

Theorem 3.2.2 (Convergence of Langevin diffusion in χ^2-divergence)**.** *Let $\{X_t\}_{t\geq 0}$ be generated according to the Langevin diffusion (3.1) with initialization $X_0 \sim \mu_0$ and stationary measure $\mu \propto e^{-f}$. Assume μ is m-strongly logconcave. Let μ_t denote the law of X_t, then*

$$\chi^2(\mu_t \parallel \mu) \leq \exp\left(-\frac{2t}{C_{PI}}\right)\chi^2(\mu_0 \parallel \mu).$$

Proof. Taking derivative with respect to t, we have

$$\begin{aligned}
\frac{d}{dt}\chi^2(\mu_t \parallel \mu) &= \frac{d}{dt}\int \left(\frac{\mu_t^2(x)}{\mu^2(x)} - 1\right)\mu(x)dx \\
&\overset{(i)}{=} 2\int \left(\frac{\mu_t(x)}{\mu(x)}\right)\left(\frac{\partial}{\partial t}\frac{\mu_t(x)}{\mu(x)}\right)\mu(x)dx \\
&\overset{(ii)}{=} 2\int \left(\frac{\mu_t(x)}{\mu(x)}\right)\left(\frac{-\nabla\cdot\left(\mu\nabla\left(\frac{\mu_t}{\mu}\right)\right)}{\mu(x)}\right)\mu(x)dx \\
&\overset{(iii)}{=} -2\int \left\|\nabla\frac{\mu_t}{\mu}\right\|_2^2 \mu(x)dx \\
&\overset{(iv)}{\leq} -\frac{2}{C_{\mathrm{PI}}}\chi^2(\mu_t \parallel \mu)
\end{aligned}$$

In (i) we switched the order of derivative and integral, which can be done after verifying conditions for dominated convergence. (ii) follows from the Fokker-Planck equation for μ_t in Eq. (3.7) and the observation that

$$\nabla\cdot(\mu_t\nabla f) + \Delta\mu_t = -\nabla\cdot\left(\mu\nabla\left(\frac{\mu_t}{\mu}\right)\right).$$

(iii) follows from integration by parts (3.9). (iv) follows from Poincaré inequality. Solving the ODE for χ^2 divergence or apply Grönwall's inequality, we obtain the desired result. □

3.3 Mixing time of ULA

Recall the iteration of ULA from Eq. (3.2),

$$X_{k+1} = X_k - h\nabla f(X_k) + \sqrt{2h}\xi_k. \tag{3.10}$$

Let μ^k denote the law of X_k. Then we have the following mixing time result.

Theorem 3.3.1. *Assume that the target measure $\mu \propto \exp(-f)$ satisfies $m\mathbb{I}_n \preceq \nabla^2 f \preceq L\mathbb{I}_n$. Let $\kappa := \frac{L}{m}$. Then, given $h \lesssim \frac{1}{L\kappa}$, for $K \geq 1$,*

$$W_2(\mu^K, \mu) \leq \exp\left(-\frac{mhK}{2}\right) W_2(\mu_0, \mu) + ch^{\frac{1}{2}}n^{\frac{1}{2}}\kappa,$$

where c is a universal constant.

A few remarks

- If we set the initial measure μ_0 to be point mass at x^* which is the mode of μ, then
$$W_2(\mu_0, \mu)^2 = \mathbb{E}_{X\sim\mu} \|X - x^*\|_2^2 \leq \frac{n}{m},$$
as a result of integration by parts and strong log-concavity.
- For mixing, we want to achieve $\sqrt{m}W_2 \leq \epsilon$. It is more convenient to use the metric $\sqrt{m}W_2$ instead of W_2 because the former is scale-invariant.
- In order for $\sqrt{m}W_2 \leq \epsilon$ in Theorem 3.3.1, we need both terms to be less than ϵ. It results in the step-size choice
$$h \lesssim \frac{\epsilon^2}{L\kappa n},$$
and the number of steps K choice
$$K \gtrsim \frac{\kappa^2 n}{\epsilon^2} \log\left(\frac{\sqrt{m}W_2(\mu_0, \mu)}{\epsilon}\right).$$

Proof sketch. Since ULA is the Euler discretization of the continuous Langevin diffusion in Eq. (3.1), it is natural to analyze the convergence of ULA by comparing it to the continuous Langevin diffuison. We know in Section 3.2.3 that the continuous Langevin diffusion convergences exponentially fast to the target measure μ with a rate that depends on the strong log-concavity m. It remains to analyze the discretization error and how it accumulates as a function of the total number of steps K.

Given the above intuition, the main problem becomes how to write $W_2(\mu_{k+1}, \mu)$ as a function of $W_2(\mu_k, \mu)$. In other words, we want to upper bound $\mathbb{E}\left\|X^{k+1} - X_{(k+1)h}\right\|_2^2$ as a function of $\mathbb{E}\left\|X^k - X_{kh}\right\|_2^2$. This analysis is separate into two parts:

- The one-step discretization error if both the discrete process and the continuous process are started at the same distribution. The distance between X^{k+1} and $\bar{X}_{(k+1)h}$ in Figure 3.3.

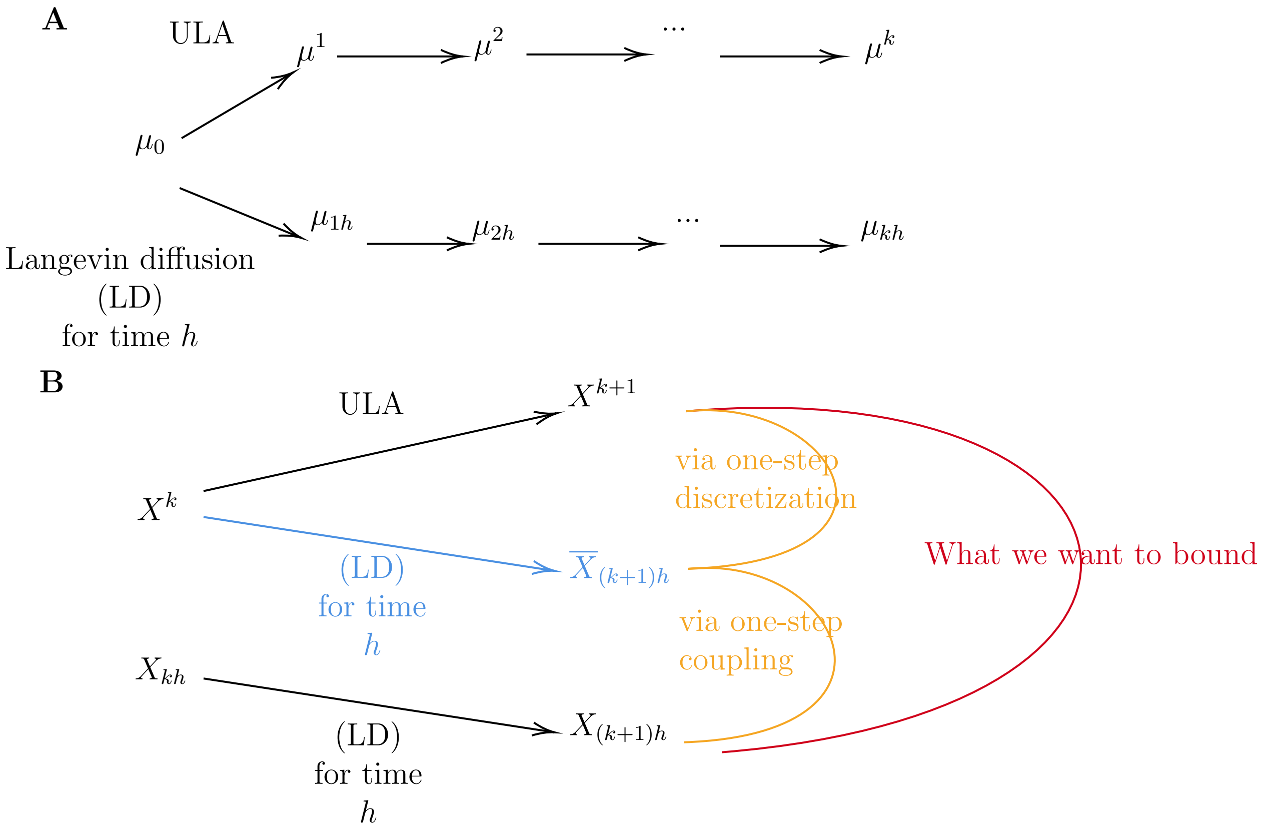

Figure 3.3: Illustration of ULA discretization analysis.

- The Wasserstein distance contraction result for continuous Langevin diffusion ran for time h, which we already know how to proceed in Section 3.2.3. The distance between $\bar{X}_{(k+1)h}$ and $X_{(k+1)h}$ in Figure 3.3.

See Section 4.1 of [Che23] for a full proof and other proof techniques in the following subsections. **YC — Or wait a bit for me to type it in**

Bibliography

[B$^+$15] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[Che69] Jeff Cheeger. A lower bound for the smallest eigenvalue of the laplacian. In *Proceedings of the Princeton conference in honor of Professor S. Bochner*, pages 195–199, 1969.

[Che23] Sinho Chewi. Log-concave sampling. *Book draft available at https://chewisinho. github. io*, 2023.

[Maz60] Vladimir Gilelevich Maz'ya. Classes of domains and imbedding theorems for function spaces. In *Doklady Akademii Nauk*, volume 133, pages 527–530. Russian Academy of Sciences, 1960.

[Pro04] Philip E Protter. Stochastic integration and differential equations. *Springer Mathematics and Statistics eBooks 2005 English/International*, 2004.