CSC/DSCC 574 Continuous Algorithms for Optimization and Sampling Lecture 8

Langevin Monte Carlo

Lecturer: Jiaming Liang *February 29, 2024*

1 Langevin dynamics

The Langevin dynamics for sampling from $\nu \propto e^{-f}$ is a stochastic process $\{X_t\}_{t\geq 0}$ where $X_t \in \mathbb{R}^d$ evolves following the stochastic differential equation:

$$\mathrm{d}X_t = -\nabla f(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \tag{1}$$

where $\{B_t\}_{t\geq 0}$ is the standard Brownian motion in $\mathbb{R}^d$ starting from $B_0 = 0$. If $X_t \in \mathbb{R}^d$ evolves following the Langevin dynamics (1), then its probability density function $\rho_t \in \mathcal{P}_2(\mathbb{R}^d)$ evolves following the Fokker-Planck equation:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\nu}\right) = \nabla \cdot (\rho_t \nabla f) + \Delta \rho_t. \tag{2}$$

It is known that the Fokker-Planck equation has an optimization interpretation as the gradient flow for minimizing the relative entropy $\mathsf{KL}(\cdot \parallel \nu)$ in the space of probability distributions with the Wasserstein $\mathcal{W}_2$ metric. Conditions on ν such as log-concavity or log-Sobolev inequality (LSI) can be interpreted as convexity-type conditions on the objective function (relative entropy) that guarantee fast convergence of the gradient flow (Fokker-Planck equation). Since the Fokker-Planck equation (2) is the continuity equation of Langevin dynamics (1), the latter is well suited for sampling.

It is easy to check that $\frac{\partial \rho_t}{\partial t} = 0$ in (2) with $\rho_t = \nu$, so ν is the stationary/invariant distribution of Fokker-Planck equation (2).

1.1 Continuity equation

There are two complementary perspectives on fluid flows: the Lagrangian perspective which emphasizes particle trajectories, and the Eulerian perspective which tracks the evolution of the fluid density. Since (1) describes the evolution of the particle trajectory, it is the Lagrangian description of the dynamics. The corresponding Eulerian description is the continuity equation (2).

Suppose $v(x,t)$ is a vector field and $\rho(x,t)$ is the density of some material q. Then, we define the flux $J = \rho v$ to be the amount of q flowing per unit time through a unit volume. Te continuity equation is

$$\frac{\partial \rho}{\partial t} + \nabla \cdot J = \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v) = \sigma(x,t).$$

If $\sigma > 0$, then it is a source that generates q. If $\sigma < 0$, then it is a sink that removes q. If $\sigma = 0$, then the continuity equation is indeed the conservation law. It appears in fluid mechanics (Navier-Stokes equations), electromagnetism (Maxwell's equations), statistical mechanics (Fokker-Planck equation), and thermodynamics (heat equation). It states the conservation of mass, electric charge, probability mass, and energy.

In fluid dynamics, we define material derivative

$$\frac{\mathrm{D}\rho}{\mathrm{D}t} \equiv \frac{\partial \rho}{\partial t} + \nabla \rho \cdot v,$$

which can be understood using the chain rule.

$$\frac{\mathrm{d}\rho(x,t)}{\mathrm{d}t} = \frac{\partial \rho}{\partial t} + \nabla \rho \cdot v = \frac{\mathrm{D}\rho}{\mathrm{D}t}$$

If the flow is incompressible flow (e.g., water but not air), i.e., $\rho(x,t) \equiv \rho$ is a constant over space and time, then the continuity equation becomes

$$\nabla \cdot v = 0.$$

This means divergence is zero. It applies to hydrodynamics but not aerodynamics.

Theorem 1. *Let v_t be a vector field and consider the evolution of particle X_t following $\mathrm{d}X_t = v_t(X_t)\mathrm{d}t$. Then, the law ρ_t of X_t evolves according to the continuity equation*

$$\frac{\partial \rho_t}{\partial t} + \nabla \cdot (\rho_t v_t) = 0.$$

Proof. For any given test function ϕ, we have

$$\mathbb{E}_{X \sim \rho_t}[\phi(X)] = \mathbb{E}[\phi(X_t)].$$

Then, we have

$$\begin{aligned}\int \phi \frac{\partial \rho_t}{\partial t} \mathrm{d}x &= \frac{\partial}{\partial t} \int \phi \rho_t \mathrm{d}x = \frac{\partial}{\partial t} \mathbb{E}_{X \sim \rho_t}[\phi(X)] = \frac{\partial}{\partial t} \mathbb{E}[\phi(X_t)] \\ &= \mathbb{E}[\langle \nabla \phi(X_t), \dot{X}_t \rangle] = \int \langle \nabla \phi, v_t \rangle \rho_t \mathrm{d}x = -\int \phi \nabla \cdot (\rho_t v_t) \mathrm{d}x,\end{aligned}$$

where we use the integration by parts in the last equality. Since the above identity holds for an arbitrary ϕ, the continuity equation holds. □

Fact 1 (Integration by parts)**.** *Given differentiable functions $f, g : \mathbb{R}^d \to \mathbb{R}$ and a smooth vector field $v : \mathbb{R}^d \to \mathbb{R}^d$ that approaches 0 at infinity, we have the following* integration by parts *formulas:*

$$\int \langle v(x), \nabla f(x) \rangle \mathrm{d}x = -\int f(x)(\nabla \cdot v)(x) \mathrm{d}x,$$

and

$$\int f(x) \Delta g(x) \mathrm{d}x = -\int \langle \nabla f(x), \nabla g(x) \rangle \mathrm{d}x = \int g(x) \Delta f(x) \mathrm{d}x.$$

Lemma 1 (Itô's lemma)**.** *For any process $x_t \in \mathbb{R}^d$ satisfying $\mathrm{d}x_t = b(x_t)\mathrm{d}t + \sigma(x_t)\mathrm{d}B_t$ where $b(x_t) \in \mathbb{R}^d$ is the drift term and $\sigma(x_t) \in \mathbb{R}^{d\times m}$ is the diffusion term, we have*

$$\mathrm{d}\phi\left(x_t\right) = \nabla\phi\left(x_t\right)^\top b\left(x_t\right)\mathrm{d}t + \nabla\phi\left(x_t\right)^\top \sigma\left(x_t\right)\mathrm{d}B_t + \frac{1}{2}\operatorname{tr}\left(\sigma\left(x_t\right)^\top \nabla^2\phi\left(x_t\right)\sigma\left(x_t\right)\right)\mathrm{d}t.$$

With the Itô's lemma above, we are able to extend Theorem (1) to the continuity equation of a general Itô's diffusion process (SDE).

Theorem 2. *Consider an Itô's diffusion process, i.e., X_t follows $dX_t = b(X_t)dt + \sigma(X_t)dB_t$ then the law ρ_t of X_t evolves according to*

$$\frac{\partial\rho_t}{\partial t} = -\nabla\cdot(\rho_t b) + \frac{1}{2}\left\langle\nabla^2, \rho_t\sigma\sigma^\top\right\rangle \tag{3}$$

where

$$\left\langle\nabla^2, A(x)\right\rangle = \sum_i\sum_j \frac{\partial^2}{\partial x_i\partial x_j}(A(x))_{ij}$$

Proof. For any given test function ϕ, we have

$$\mathbb{E}_{X\sim\rho_t}[\phi(X)] = \mathbb{E}[\phi(X_t)].$$

Taking differential on both sides and using Lemma 1, we have

$$\begin{aligned}\int \phi\mathrm{d}\rho_t(x)\mathrm{d}x &= \mathbb{E}_{X\sim\rho_t}[\mathrm{d}\phi(X)] = \mathbb{E}[\mathrm{d}\phi(X_t)]\\ &= \mathbb{E}\left[\nabla\phi\left(x_t\right)^\top f\left(x_t\right)\mathrm{d}t + \nabla\phi\left(x_t\right)^\top \sigma\left(x_t\right)\mathrm{d}B_t + \frac{1}{2}\operatorname{tr}\left(\sigma\left(x_t\right)^\top \nabla^2\phi\left(x_t\right)\sigma\left(x_t\right)\right)\mathrm{d}t\right]\\ &= \mathbb{E}\left[\nabla\phi\left(x_t\right)^\top f\left(x_t\right)\mathrm{d}t + \frac{1}{2}\operatorname{tr}\left(\nabla^2\phi\left(x_t\right)D\left(x_t\right)\right)\mathrm{d}t\right]\end{aligned}$$

where we use $\mathbb{E}[\mathrm{d}B_t] = 0$ and $\operatorname{tr}(ABC) = \operatorname{tr}(BCA)$ in the last identity. Using $X_t \sim \rho_t$, we have

$$\int \phi\frac{\partial\rho_t}{\partial t}\mathrm{d}x = \int \nabla\phi\left(x\right)^\top f\left(x\right)\rho_t(x)\mathrm{d}x + \frac{1}{2}\int\left\langle\nabla^2\phi(x), \rho_t(x)\sigma(x)\sigma(x)^\top\right\rangle\mathrm{d}x. \tag{4}$$

Using integrating by parts on the first integral in (4), we have

$$\int \nabla\phi\left(x\right)^\top f\left(x\right)\rho_t(x)\mathrm{d}x = -\int \phi(x)\nabla\cdot(\rho_t b)\mathrm{d}x.$$

Using integrating by parts twice on the second integral in (4) gives

$$\int\left\langle\nabla^2\phi(x), \rho_t(x)\sigma(x)\sigma(x)^\top\right\rangle\mathrm{d}x = \int \phi(x)\left\langle\nabla^2, \rho_t\sigma(x)\sigma^\top(x)\right\rangle\mathrm{d}x.$$

Plugging the above identities into (4), we have

$$\int \phi\frac{\partial\rho_t}{\partial t}\mathrm{d}x = -\int \phi(x)\nabla\cdot(\rho_t b)\mathrm{d}x + \frac{1}{2}\int \phi(x)\left\langle\nabla^2, \rho_t\sigma(x)\sigma^\top(x)\right\rangle\mathrm{d}x.$$

Since the above identity holds for an arbitrary ϕ, the continuity equation (3) holds. □

Now, we are ready to prove that (2) is the continuity equation of (1).

Corollary 1. *If X_t follows (1), then the law ρ_t of X_t evolves according to (2).*

Proof. This claim immediately follows from Theotem 2 with $b(x) = -\nabla f(x)$ and $\sigma = \sqrt{2}I$. □

1.2 Divergences

Throughout the notes, we abuse notation by identifying a probability measure with its density w.r.t. Lebesgue measure. For a probability measure $\rho \ll \nu$ (i.e., ρ is absolutely continuous w.r.t. ν), we define the total variation (TV) distance, the Kullback–Leibler (KL) divergence, and the chi-squared (χ^2) divergence, respectively, as

$$\|\rho - \nu\|_{\mathsf{TV}} = \sup_{A \in \mathcal{F}} |\rho(A) - \mu(A)|, \quad \mathsf{KL}(\rho \,\|\, \nu) = \int \rho \log \frac{\rho}{\nu} dx, \quad \chi^2(\rho \,\|\, \nu) = \int \left(\frac{\rho}{\nu} - 1\right)^2 \nu dx. \tag{5}$$

In general, for a convex function ϕ, we can define the ϕ-divergence

$$D(\rho \,\|\, \nu) = \mathbb{E}_\nu \left[\phi\left(\frac{\rho}{\nu}\right)\right].$$

The three divergences in (5) are all instances of the ϕ-divergence.

1. TV distance, $\phi(x) = \frac{1}{2}|x - 1|$

   $$\|\rho - \nu\|_{\mathsf{TV}} = \frac{1}{2}\int |\rho - \nu| dx = \frac{1}{2}\int \left|\frac{\rho}{\nu} - 1\right| \nu dx.$$

   Note that $\mathbb{E}_\nu[\rho/\nu] = 1$. The TV distance is half of the L^1 distance between the probability measures.

2. KL divergence, $\phi(x) = x \log x$

   $$\mathsf{KL}(\rho \,\|\, \nu) = \int \rho \log \frac{\rho}{\nu} dx = \int \frac{\rho}{\nu} \log \frac{\rho}{\nu} \nu dx.$$

   Pinker's inequality

   $$2\|\rho - \nu\|_{\mathsf{TV}}^2 = \frac{1}{2}\|\rho - \nu\|_1^2 \le \mathsf{KL}(\rho \,\|\, \nu).$$

3. χ^2 divergence, $\phi(x) = (x - 1)^2$

   $$\chi^2(\rho \,\|\, \nu) = \int \left(\frac{\rho}{\nu} - 1\right)^2 \nu dx = \int \left(\frac{\rho}{\nu} - \mathbb{E}\left[\frac{\rho}{\nu}\right]\right)^2 \nu dx = \mathsf{Var}_\nu\left(\frac{\rho}{\nu}\right) = \int \frac{\rho^2}{\nu} dx - 1.$$

Lemma 2. *If ϕ is convex and $\phi(1) = 0$, then the ϕ-divergence is always nonnegative, i.e.,*

$$D(\rho \,\|\, \nu) \ge 0, \quad \forall \rho, \nu \in \mathcal{P}(\mathbb{R}^d).$$

Proof. It follows from the convexity of ϕ and the Jensen's inequality that

$$\mathbb{E}_\nu\left[\phi\left(\frac{\rho}{\nu}\right)\right] \geq \phi\left(\mathbb{E}_\nu\left[\frac{\rho}{\nu}\right]\right) = \phi(1) = 0.$$

□

Note that $\phi(1) = 0$ holds for each corresponding ϕ used in TV distance, KL divergence, and χ^2 divergence. Hence, the three divergences are all nonnegative.

Definition 1. *The Fisher information functional* $\mathsf{FI}\colon \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ *is defined by:*

$$\mathsf{FI}(\rho) \coloneqq \mathbb{E}_\rho\left[\|\nabla \log \rho\|^2\right] = \int_{\mathbb{R}^d} \rho(x)\|\nabla \log \rho(x)\|^2 \,\mathrm{d}x = \int_{\mathbb{R}^d} \frac{\|\nabla \rho(x)\|^2}{\rho(x)} \,\mathrm{d}x\,.$$

The relative Fisher information with respect to $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ *is a functional* $\mathsf{FI}(\cdot \parallel \nu)\colon \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ *defined by:*

$$\mathsf{FI}(\rho \parallel \nu) \coloneqq \mathbb{E}_\rho\left[\left\|\nabla \log \frac{\rho}{\nu}\right\|^2\right] = \int_{\mathbb{R}^d} \rho(x)\left\|\nabla \log \frac{\rho(x)}{\nu(x)}\right\|^2 \,\mathrm{d}x\,.$$

If $\rho \not\ll \nu$*, then* $\mathsf{FI}(\rho \parallel \nu) \coloneqq +\infty$*.*

1.3 Transport inequalities

Definition 2. *A probability distribution* ν *is said to be* α*-strongly log-concave (*α*-SLC) with constant* $\alpha > 0$ *if* $\alpha I \preceq -\nabla^2 \log \nu$*. When* $\alpha = 0$*, we say that the* ν *is log-concave.*

It is easy to see that $\nu \propto \exp(-f)$ is α-SLC if and only if f is α-strongly convex.

Definition 3. *A probability distribution* ν *satisfies* α*-log-Sobolev inequality (*α*-LSI) with constant* $\alpha > 0$ *if for any* ρ*,*

$$2\alpha \mathsf{KL}(\rho \parallel \nu) \leq \mathsf{FI}(\rho \parallel \nu).$$

Definition 4. *A probability distribution* ν *satisfies* α*-Poincareé inequality (*α*-PI) with constant* $\alpha > 0$ *if for any smooth* $g : \mathbb{R}^d \mapsto \mathbb{R}$*,*

$$\mathsf{Var}_\nu(g) \leq \frac{1}{\alpha}\mathbb{E}_\nu[\|\nabla g\|^2],$$

where $\mathsf{Var}_\nu(g) = \mathbb{E}_\nu[g^2] - \mathbb{E}_\nu[g]^2$ *is the variance of* g *under* ν*.*

It is known α-SLC implies α-LSI and α-LSI implies α-PI. Indeed, PI is the linearization of LSI (taking $\rho = (1 + \varepsilon g)\nu$ for $\varepsilon > 0$).

Definition 5. *For any two distributions* μ *and* ν*, the Wasserstein-2 distance between them is defiened as*

$$\mathcal{W}_2^2(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \int \|x - y\|^2 \mathrm{d}\gamma(x, y),$$

where Γ *is the set of couplings* γ *of* μ *and* ν*, i.e., the marginal distributions of* γ *are* μ *and* ν*.*

Definition 6. *A probability distribution ν satisfies α-Talagrand's inequality (α-TI) with constant $\alpha > 0$ if for any ρ,*

$$\frac{\alpha}{2}\mathcal{W}_2^2(\rho, \nu) \le \mathsf{KL}(\rho \,\|\, \nu).$$

Note that α-LSI implies α-TI and PI is also the linearization of TI with the same constant.

1.4 Convergence

The following result presents the rate of change of KL divergence along Langevin dynamics (1), which is a variant of de Bruijn's identity.

Lemma 3. *Let X_t evolves following Langevin dynamics (1). Then, the law ρ_t of X_t satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t \,\|\, \nu) = -\mathsf{FI}(\rho_t \,\|\, \nu).$$

Proof. The time derivative of KL divergence along any flow is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t \,\|\, \nu) = \frac{\mathrm{d}}{\mathrm{d}t}\int \rho_t \log\frac{\rho_t}{\nu}\mathrm{d}x = \int \frac{\partial \rho_t}{\partial t}\log\frac{\rho_t}{\nu}\mathrm{d}x \tag{6}$$

since the second part of the chain rule is zero:

$$\int \rho_t \frac{\partial}{\partial t}\log\frac{\rho_t}{\nu}\mathrm{d}x = \int \frac{\partial \rho_t}{\partial t}\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}t}\int \rho_t \mathrm{d}x = 0.$$

It follows from Corollary 1 that ρ_t satisfies the Fokker-Planck equation (2). Therefore, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t \,\|\, \nu) &= \int \nabla \cdot \left(\rho_t \nabla \log\frac{\rho_t}{\nu}\right)\log\frac{\rho_t}{\nu_t}\mathrm{d}x \\
&= -\int \rho_t \left\|\nabla \log\frac{\rho_t}{\nu}\right\|^2 \mathrm{d}x \\
&= -\mathsf{FI}(\rho_t \,\|\, \nu).
\end{aligned}$$

where the second identity follows from integration by parts. □

Theorem 3. *Suppose ν satisfies α-LSI. Along the Langevin dynamics (1), we have*

$$\mathsf{KL}(\rho_t \,\|\, \nu) \le e^{-2\alpha t}\mathsf{KL}(\rho_0 \,\|\, \nu). \tag{7}$$

Moreover, we have

$$\mathcal{W}_2(\rho_t, \nu) \le e^{-\alpha t}\sqrt{\frac{2}{\alpha}\mathsf{KL}(\rho_0 \,\|\, \nu)}.$$

Proof. It follows from Lemma 3 and α-LSI that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t \,\|\, \nu) = -\mathsf{FI}(\rho_t \,\|\, \nu) \le -2\alpha \mathsf{KL}(\rho_t \,\|\, \nu).$$

Integrating gives the desired bound (7). Moreover, since α-LSI implies α-TI, we have

$$\frac{\alpha}{2}\mathcal{W}_2^2(\rho_t, \nu) \le \mathsf{KL}(\rho_t \,\|\, \nu) \le e^{-2\alpha t}\mathsf{KL}(\rho_0 \,\|\, \nu).$$

□

Similar to Lemma 3, we have the rate of change of χ^2 divergence along Langevin dynamics (1).

Lemma 4. *Let X_t evolves following Langevin dynamics (1). Then, the law ρ_t of X_t satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}t}\chi^2(\rho_t \,\|\, \nu) = -2\mathbb{E}_\nu\left[\left\|\nabla \frac{\rho_t}{\nu}\right\|^2\right].$$

Proof. The time derivative of χ^2 divergence along any flow is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\chi^2(\rho_t \,\|\, \nu) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\int \frac{\rho_t^2}{\nu} dx - 1\right) = 2\int \frac{\rho_t}{\nu}\frac{\partial \rho_t}{\partial t}\mathrm{d}x.$$

Therefore, along the Fokker-Planck equation (2), we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\chi^2(\rho_t \,\|\, \nu) &= 2\int \frac{\rho_t}{\nu_t}\nabla\cdot\left(\rho_t \nabla \log\frac{\rho_t}{\nu}\right)\mathrm{d}x \\
&= -2\int \langle \nabla\frac{\rho_t}{\nu}, \rho_t \nabla\log\frac{\rho_t}{\nu}\rangle \mathrm{d}x \\
&= -2\int \left\|\nabla\frac{\rho_t}{\nu}\right\|^2 \nu \mathrm{d}x \\
&= -2\mathbb{E}_\nu\left[\left\|\nabla\frac{\rho_t}{\nu}\right\|^2\right]
\end{aligned}$$

where the second identity follows from integration by parts. □

Theorem 4. *Suppose ν satisfies α-PI. Along the Langevin dynamics (1), we have*

$$\chi^2(\rho_t \,\|\, \nu) \le e^{-2\alpha t}\chi^2(\rho_0 \,\|\, \nu). \tag{8}$$

Proof. It follows from Lemma 4 and α-PI that

$$\frac{\mathrm{d}}{\mathrm{d}t}\chi^2(\rho_t \,\|\, \nu) = -2\mathbb{E}_\nu\left[\left\|\nabla\frac{\rho_t}{\nu}\right\|^2\right] \le -2\alpha \mathsf{Var}_\nu\left(\frac{\rho_t}{\nu}\right) = -2\alpha\chi^2(\rho_t \,\|\, \nu).$$

Integrating gives (8). □

We conclude this section by giving another proof of convergence in terms of $\mathcal{W}_2$ distance based on the technique of coupling.

Theorem 5. *Suppose ν is α-SLC. Assume $(X_0, Y_0) \sim \gamma(x, y)$ is sampled from the optimal coupling of ρ_0 and π_0. Let X_t and Y_t evolve along the same Langevin dynamics (1), then the laws ρ_t and π_t of X_t and Y_t, respectively, satisfy*

$$\mathcal{W}_2^2(\rho_t, \pi_t) \leq e^{-2\alpha t} \mathcal{W}_2^2(\rho_0, \pi_0). \tag{9}$$

In particular, we have

$$\mathcal{W}_2^2(\rho_t, \nu) \leq e^{-2\alpha t} \mathcal{W}_2^2(\rho_0, \nu). \tag{10}$$

Proof. Along the Langevin dynamics (1), using the same Brownian motion $\mathrm{d}B_t$ for both processes, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(x_t - y_t) = \nabla f(y_t) - \nabla f(x_t).$$

Hence,

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|x_t - y_t\|^2 = 2\langle \nabla f(y_t) - \nabla f(x_t), x_t - y_t \rangle.$$

Next, since f is α-strongly convex, we have

$$f(y_t) - f(x_t) \geq \nabla f(x_t)^\top (y_t - x_t) + \frac{\alpha}{2}\|x_t - y_t\|^2,$$
$$f(x_t) - f(y_t) \geq \nabla f(y_t)^\top (x_t - y_t) + \frac{\alpha}{2}\|x_t - y_t\|^2.$$

Adding two equations together, we have

$$(\nabla f(x_t) - \nabla f(y_t))^\top (x_t - y_t) \geq \alpha \|x_t - y_t\|^2.$$

Therefore,

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|x_t - y_t\|^2 \leq -\alpha \|x_t - y_t\|^2.$$

Integrating from 0 to t, we have

$$\mathbb{E}\|x_t - y_t\|^2 \leq e^{-2\mu t}\|x_0 - y_0\|^2.$$

Since (X_0, Y_0) is generated from the optimal coupling of ρ_0 and π_0, we have

$$\mathbb{E}\|x_t - y_t\|^2 \leq e^{-2\mu t}\|x_0 - y_0\|^2 = e^{-2\mu t}\mathcal{W}_2^2(\rho_0, \pi_0).$$

The proof of (9) is concluded upon noting that $\mathcal{W}_2^2(\rho_t, \pi_t) \leq \mathbb{E}\|x_t - y_t\|^2$. Finally, (10) follows from (9) and the fact that ν is the invariant distribution. □

2 Langevin Monte Carlo

In discrete time, a simple discretization of the Langevin dynamics (1) is the Langevin Monte Carlo (LMC) which updates the current iterate $X_k \in \mathbb{R}^d$ to

$$X_{k+1} = X_k - \eta \nabla f(X_k) + \sqrt{2\eta} Z_k, \tag{11}$$

where $\eta > 0$ is step size and $Z_k \sim \mathcal{N}(0, I)$ is an independent Gaussian random variable. It is known that LMC is biased, which means that for any fixed $\eta > 0$ and as $k \to \infty$, the law ρ_k of X_k converges to a limiting distribution ν_η which is different from the true target ν. We study the convergence rate of LMC to ν_η in this section.

We first observe that one step of LMC (11) can be interpreted as the output at time $t = \eta$ of the SDE

$$\mathrm{d}X_t = -\nabla f(X_0)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t. \tag{12}$$

Lemma 5. *Let ρ_t be the law of the process (12). Then, its Fokker-Planck equation is*

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \mathbb{E}_{\rho_{t|0}}[\nabla f(X_0)|X_t]) + \Delta \rho_t.$$

Proof. We first note that

$$\rho_{0t}(x_0, x_t) = \rho_0(x_0)\rho_{t|0}(x_t|x_0) = \rho_t(x_t)\rho_{0|t}(x_0|x_t).$$

Conditioning on x_0, the drift vector field $-\nabla f(x_0)$ is a constant. Following (2), the Fokker-Planck equation for the conditional density $\rho_{t|0}(x_t \mid x_0)$ is

$$\frac{\partial \rho_{t|0}(x_t \mid x_0)}{\partial t} = \nabla \cdot (\rho_{t|0}(x_t \mid x_0)\nabla f(x_0)) + \Delta \rho_{t|0}(x_t \mid x_0).$$

To derive the evolution of ρ_t, we take expectation over $x_0 \sim \rho_0$. Multiplying both sides of the above Fokker-Planck equation by $\rho_0(x_0)$ and integrating over x_0, we obtain

$$\begin{aligned}
\frac{\partial \rho_t(x)}{\partial t} &= \int \frac{\partial \rho_{t|0}(x \mid x_0)}{\partial t} \rho_0(x_0)\mathrm{d}x_0 \\
&= \int \left(\nabla \cdot \left(\rho_{t|0}(x \mid x_0)\nabla f(x_0)\right) + \Delta \rho_{t|0}(x \mid x_0)\right) \rho_0(x_0)\mathrm{d}x_0 \\
&= \int \left(\nabla \cdot (\rho_{t,0}(x, x_0)\nabla f(x_0)) + \Delta \rho_{t,0}(x, x_0)\right) \mathrm{d}x_0 \\
&= \nabla \cdot \left(\rho_t(x) \int \rho_{0|t}(x_0 \mid x)\nabla f(x_0) dx_0\right) + \Delta \rho_t(x) \\
&= \nabla \cdot \left(\rho_t(x)\mathbb{E}_{\rho_{0|t}}\left[\nabla f(x_0) \mid x_t\right]\right) + \Delta \rho_t(x).
\end{aligned}$$

□

Lemma 6. *Let ρ_t be the law of the process (12). Then, it satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t \parallel \nu) \leq -\frac{3}{4}\mathsf{FI}(\rho_t \parallel \nu) + \mathbb{E}_{\rho_{0t}}\left[\|\nabla f(x_t) - \nabla f(x_0)\|^2\right]. \tag{13}$$

Proof. It follows from (6) and Lemma 5 that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t \parallel \nu) = \int \left[\nabla \cdot \left(\rho_t \mathbb{E}_{\rho_{0|t}}\left[\nabla f(x_0) \mid x_t\right]\right) + \Delta \rho_t\right] \log \frac{\rho_t}{\nu} \mathrm{d}x.$$

Hence, using integration by parts, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t \parallel \nu) &= \int \left[\nabla \cdot \left(\rho_t \left(\nabla \log \frac{\rho_t}{\nu} + \mathbb{E}_{\rho_{0|t}}\left[\nabla f(x_0) \mid x_t\right] - \nabla f(x)\right)\right)\right] \log \frac{\rho_t}{\nu} \mathrm{d}x \\
&= -\int \rho_t \left\langle \nabla \log \frac{\rho_t}{\nu} + \mathbb{E}_{\rho_{0|t}}\left[\nabla f(x_0) \mid x_t\right] - \nabla f(x), \nabla \log \frac{\rho_t}{\nu} \right\rangle \mathrm{d}x \\
&= -\int \rho_t \left\| \nabla \log \frac{\rho_t}{\nu} \right\|^2 dx + \int \rho_t \left\langle \nabla f(x) - \mathbb{E}_{\rho_{0|t}}\left[\nabla f(x_0) \mid x_t\right], \nabla \log \frac{\rho_t}{\nu} \right\rangle \mathrm{d}x \\
&= -\mathsf{FI}(\rho_t \parallel \nu) + \int \rho_{0t}(x_0, x) \left\langle \nabla f(x) - \nabla f(x_0), \nabla \log \frac{\rho_t}{\nu} \right\rangle \mathrm{d}x_0 \mathrm{d}x \\
&= -\mathsf{FI}(\rho_t \parallel \nu) + \mathbb{E}_{\rho_{0t}}\left[\left\langle \nabla f(x_t) - \nabla f(x_0), \nabla \log \frac{\rho_t(x_t)}{\nu(x_t)} \right\rangle\right].
\end{aligned}$$

Using the fact that $\langle a, b\rangle \leq \|a\|^2 + \|b\|^2/4$ for any $a, b \in \mathbb{R}^d$, we have

$$\begin{aligned}
\mathbb{E}_{\rho_{0t}}\left[\left\langle \nabla f(x_t) - \nabla f(x_0), \nabla \log \frac{\rho_t(x_t)}{\nu(x_t)} \right\rangle\right] &\leq \mathbb{E}_{\rho_{0t}}\left[\|\nabla f(x_t) - \nabla f(x_0)\|^2\right] + \frac{1}{4}\mathbb{E}_{\rho_{0t}}\left[\left\| \nabla \log \frac{\rho_t(x_t)}{\nu(x_t)} \right\|^2\right] \\
&= \mathbb{E}_{\rho_{0t}}\left[\|\nabla f(x_t) - \nabla f(x_0)\|^2\right] + \frac{1}{4}\mathsf{FI}(\rho_t \parallel \nu).
\end{aligned}$$

Finally, (13) follows from combining the above two relations. □

Lemma 7. *Assume $\nu = \exp(-f)$ and f is L-smooth. Then, the following statements hold:*

(a) $\mathbb{E}_\nu[\|\nabla f\|^2] \leq dL$;

(b) if ν also satisfies α-TI, then for any ρ,

$$\mathbb{E}_\rho[\|\nabla f\|^2] \leq \frac{4L^2}{\alpha}\mathsf{KL}(\rho \parallel \nu) + 2dL.$$

Proof. (a) Using integration by parts, we have

$$\mathbb{E}_\nu[\|\nabla f\|^2] = \int \langle \nabla f, \nabla f\rangle \exp(-f)\mathrm{d}x = -\int \langle \nabla \exp(-f), \nabla f\rangle \mathrm{d}x = \int \exp(-f)\nabla\cdot(\nabla f)\mathrm{d}x = \mathbb{E}_\nu[\Delta f].$$

Since f is L-smooth, $\nabla^2 f \preceq LI$, and hence $\Delta f \leq dL$.Therefore, (a) holds.

(b) Let (x, x^*) be generated from an optimal coupling of (ρ, ν), then $\mathbb{E}[\|x - x^*\|^2] = \mathcal{W}_2^2(\rho, \nu)$. Since f is L-smooth,

$$\|\nabla f(x)\| \le \|\nabla f(x) - \nabla f(x^*)\| + \|\nabla f(x^*)\| \le L\|x - x^*\| + \|\nabla f(x^*)\|.$$

Taking expectation gives

$$\mathbb{E}_\rho[\|\nabla f(x)\|^2] \le 2L^2\mathbb{E}[\|x - x^*\|^2] + 2\mathbb{E}_\nu[\|\nabla f(x^*)\|^2] = 2L^2\mathcal{W}_2^2(\rho, \nu) + 2\mathbb{E}_\nu[\|\nabla f(x^*)\|^2].$$

Now, statement (b) immediately follows from α-TI and (a). □

Lemma 8. *Assume ν satisfies α-LSI and f is L-smooth. Then, for $0 \le t \le \eta$, we have*

$$\mathbb{E}_{\rho_{0t}}\left[\|\nabla f(x_t) - \nabla f(x_0)\|^2\right] \le \frac{4\eta^2 L^4}{\alpha}\mathsf{KL}(\rho_0 \,\|\, \nu) + 2\eta^2 dL^3 + 2\eta dL^2.$$

Proof. Since the solution to (12) is

$$X_t \overset{d}{=} X_0 - t\nabla f(X_0) + \sqrt{2t}Z_0,$$

where $Z_0 \sim \mathcal{N}(0, I)$ is independent of X_0. Then,

$$\begin{aligned}
\mathbb{E}_{\rho_{0t}}\left[\|x_t - x_0\|^2\right] &= \mathbb{E}_{\rho_{0t}}\left[\left\|-t\nabla f(x_0) + \sqrt{2t}z_0\right\|^2\right] \\
&= t^2\mathbb{E}_{\rho_0}\left[\|\nabla f(x_0)\|^2\right] + 2td \\
&\le \frac{4t^2L^2}{\alpha}\mathsf{KL}(\rho_0 \,\|\, \nu) + 2t^2dL + 2td,
\end{aligned}$$

where the inequality is due to Lemma 7(b) and the fact that α-LSI implies α-TI. The lemma finally follows from the facts that f is L-smooth and $t \le \eta$. □

Theorem 6. *Suppose ν satisfies α-LSI and f is L-smooth. If $\eta \le \alpha/(4L^2)$, then along the Langevin Monte Carlo (11),*

$$\mathsf{KL}(\rho_{k+1} \,\|\, \nu) \le e^{-\alpha\eta}\mathsf{KL}(\rho_k \,\|\, \nu) + 6\eta^2 dL^2. \tag{14}$$

Furthermore,

$$\mathsf{KL}(\rho_k \,\|\, \nu) \le e^{-\alpha\eta k}\mathsf{KL}(\rho_0 \,\|\, \nu) + \frac{8\eta dL^2}{\alpha}.$$

For $0 < \eta < 4d$, LMC with $\eta \le \frac{\alpha\eta}{16L^2d}$ reaches error $\mathsf{KL}(\rho_k \,\|\, \nu) \le \varepsilon$ after $k \ge \frac{1}{\alpha\eta}\log\frac{2\mathsf{KL}(\rho_0\|\nu)}{\varepsilon}$ iterations.

Proof. It follows from Lemmas 6 and 8 that for $t \le \eta$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t \,\|\, \nu) \le -\frac{3}{4}\mathsf{FI}(\rho_t \,\|\, \nu) + \frac{4\eta^2L^4}{\alpha}\mathsf{KL}(\rho_0 \,\|\, \nu) + 2\eta^2dL^3 + 2\eta dL^2.$$

Noting that $\eta \le 1/(2L)$ and using the fact that ν satisfies α-LSI, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t \,\|\, \nu) \le -\frac{3\alpha}{2}\mathsf{KL}(\rho_t \,\|\, \nu) + \frac{4\eta^2 L^4}{\alpha}\mathsf{KL}(\rho_0 \,\|\, \nu) + 3\eta d L^2.$$

Multiplying both sides by $\exp(3\alpha t/2)$, we rewrite the above inequality as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{\frac{3\alpha t}{2}}\mathsf{KL}(\rho_t \,\|\, \nu)\right) \le e^{\frac{3\alpha t}{2}}\left(\frac{4\eta^2 L^4}{\alpha}\mathsf{KL}(\rho_0 \,\|\, \nu) + 3\eta d L^2\right).$$

Integrating from 0 to η gives

$$\begin{aligned} e^{\frac{3\alpha\eta}{2}}\mathsf{KL}(\rho_\eta \,\|\, \nu) - \mathsf{KL}(\rho_0 \,\|\, \nu) &\le \frac{2}{3\alpha}\left(e^{\frac{3\alpha\eta}{2}} - 1\right)\left(\frac{4\eta^2 L^4}{\alpha}\mathsf{KL}(\rho_0 \,\|\, \nu) + 3\eta d L^2\right) \\ &\le 2\eta\left(\frac{4\eta^2 L^4}{\alpha}\mathsf{KL}(\rho_0 \,\|\, \nu) + 3\eta d L^2\right), \end{aligned}$$

where in the last step we have used the inequality $e^c \le 1 + 2c$ for $0 < c = \frac{3}{2}\alpha\eta \le 1$, which holds because $0 < \eta < \frac{2}{3\alpha}$. Rearranging the inequality above gives

$$\mathsf{KL}(\rho_\eta \,\|\, \nu) \le e^{\frac{-3\alpha\eta}{2}}\left(1 + \frac{8\eta^3 L^4}{\alpha}\right)\mathsf{KL}(\rho_0 \,\|\, \nu) + e^{\frac{-3\alpha\eta}{2}} 6\eta^2 d L^2.$$

Since $1 + \frac{8\eta^3 L^4}{\alpha} \le 1 + \frac{\alpha\eta}{2} \le e^{\frac{1}{2}\alpha\eta}$ for $\eta \le \frac{\alpha}{4L^2}$, and using $e^{-\frac{3}{2}\alpha\eta} \le 1$, we conclude that (14) holds after renaming

$$\rho_0 \equiv \rho_k, \quad \rho_\eta \equiv \rho_{k+1}.$$

Applying (14), we obtain

$$\mathsf{KL}(\rho_k \,\|\, \nu) \le e^{-\alpha\eta k}\mathsf{KL}(\rho_0 \,\|\, \nu) + \frac{6\eta^2 d L^2}{1 - e^{-\alpha\eta}} \le e^{-\alpha\eta k}\mathsf{KL}(\rho_0 \,\|\, \nu) + \frac{8\eta d L^2}{\alpha},$$

where in the last step we have used the inequality $1 - e^{-c} \ge \frac{3}{4}c$ for $0 < c = \alpha\eta \le \frac{1}{4}$, which holds since $\eta \le \frac{\alpha}{4L^2} \le \frac{1}{4\alpha}$. □