Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices

Santosh S. Vempala
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332
vempala@gatech.edu

Andre Wibisono
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332
wibisono@gatech.edu

Abstract

We study the Unadjusted Langevin Algorithm (ULA) for sampling from a probability distribution $\nu = e^{-f}$ on $\mathbb{R}^n$. We prove a convergence guarantee in Kullback-Leibler (KL) divergence assuming ν satisfies log-Sobolev inequality and f has bounded Hessian. Notably, we do not assume convexity or bounds on higher derivatives. We also prove convergence guarantees in Rényi divergence of order $q > 1$ assuming the limit of ULA satisfies either log-Sobolev or Poincaré inequality.

1 Introduction

Sampling is a fundamental algorithmic task. Many applications require sampling from probability distributions in high-dimensional spaces, and in modern applications the probability distributions are complicated and non-logconcave. While the setting of logconcave functions is well-studied, it is important to have efficient sampling algorithms with good convergence guarantees beyond the logconcavity assumption. There is a close interplay between sampling and optimization, either via optimization as a limit of sampling (annealing) [34, 55], or via sampling as optimization in the space of distributions [36, 62]. Motivated by the widespread use of non-convex optimization and sampling, there is resurgent interest in understanding non-logconcave sampling.

In this paper we study a simple algorithm, the Unadjusted Langevin Algorithm (ULA), for sampling from a target probability distribution $\nu = e^{-f}$ on $\mathbb{R}^n$. ULA requires oracle access to the gradient ∇f of the log density $f = -\log \nu$. In particular, ULA does not require knowledge of f, which makes it applicable in practice where we often only know ν up to a normalizing constant.

As the step size $\epsilon \to 0$, ULA recovers the Langevin dynamics, which is a continuous-time stochastic process in $\mathbb{R}^n$ that converges to ν. We recall the optimization interpretation of the Langevin dynamics for sampling as the gradient flow of the Kullback-Leibler (KL) divergence with respect to ν in the space of probability distributions with the Wasserstein metric [36]. When ν is strongly logconcave, the KL divergence is a strongly convex objective function, so the Langevin dynamics as gradient flow converges exponentially fast [6, 60]. From the classical theory of Markov chains and diffusion processes, there are several known conditions milder than logconcavity that are sufficient for rapid convergence *in continuous time*. These include isoperimetric inequalities such as Poincaré inequality or log-Sobolev inequality (LSI). Along the Langevin dynamics in continuous time, Poincaré inequality implies an exponential convergence rate in χ^2-divergence, while LSI—which is stronger—implies an exponential convergence rate in KL divergence (as well as in Rényi divergence).

However, in discrete time, sampling under Poincaré inequality or LSI is a more challenging problem. ULA is an inexact discretization of the Langevin dynamics, and it converges to a biased limit $\nu_\epsilon \neq \nu$. When ν is strongly logconcave and smooth, it is known how to control the bias and prove a convergence guarantee on KL divergence along ULA [17, 21, 22, 24]. When ν is strongly

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

logconcave, there are many sampling algorithms with provable rapid convergence; these include the ball walk and hit-and-run [37, 43, 44, 42] (which give truly polynomial algorithms), various discretizations of the overdamped or underdamped Langevin dynamics [21, 22, 24, 8, 26] (which have polynomial dependence on smoothness parameters but low dependence on dimension), and the Hamiltonian Monte Carlo [47, 48, 25, 39, 16]. It is of great interest to extend these results to non-logconcave densities ν, where existing results require strong assumptions with bounds that grow exponentially with the dimension or other parameters [2, 18, 45, 49]. There are recent works that analyze convergence of sampling using various techniques such as reflection coupling [28], kernel methods [29], and higher-order integrators [40], albeit still under some strong conditions such as distant dissipativity, which is similar to strong logconcavity outside a bounded domain.

In this paper we study the convergence along ULA under minimal (and necessary) isoperimetric assumptions, namely, LSI and Poincaré inequality. These are sufficient for fast convergence in continuous time; moreover, in the case of logconcave distribution, the log-Sobolev and Poincaré constants can be bounded and lead to convergence guarantees for efficient sampling in discrete time. However, do they suffice on their own without the assumption of logconcavity?

We note that LSI and Poincaré inequality apply to a wider class of measures than logconcave distributions. In particular, LSI and Poincaré inequality are preserved under bounded perturbation and Lipschitz mapping, whereas logconcavity is destroyed. Given these properties, it is easy to exhibit examples of non-logconcave distributions satisfying LSI or Poincaré inequality. For example, we can take a small perturbation of a convex body to make it nonconvex but still satisfies isoperimetry; then the uniform probability distribution (or a smooth version of it) on the body is not logconcave but satisfies LSI and Poincaré inequality. Similarly, we can start with a strongly logconcave distribution and make bounded perturbations; then the resulting (normalized) probability distribution is not logconcave, but it satisfies LSI and Poincaré inequality. See Figure 1 for an illustration.

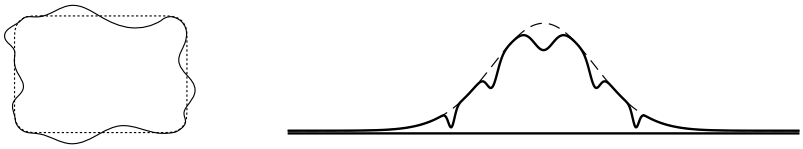

Figure 1: Illustrations of non-logconcave distributions satisfying isoperimetry: uniform distribution on a nonconvex set (left) and a perturbation of a logconcave distribution (right).

We measure the mode of convergence using KL divergence and Rényi divergence of order $q \geq 1$, which is stronger. Our first main result says the only further assumption we need is smoothness. We say $\nu = e^{-f}$ is L-smooth if ∇f is L-Lipschitz. Here $H_\nu(\rho)$ is the KL divergence between ρ and ν. See Theorem 2 in Section 3.1 for more detail.

Theorem 2. *Assume $\nu = e^{-f}$ satisfies log-Sobolev inequality with constant $\alpha > 0$ and is L-smooth. ULA with step size $0 < \epsilon \leq \frac{\alpha}{4L^2}$ satisfies*

$$H_\nu(\rho_k) \leq e^{-\alpha\epsilon k} H_\nu(\rho_0) + \tfrac{8\epsilon n L^2}{\alpha}.$$

For $0 < \delta < 4n$, ULA with $\epsilon \leq \frac{\alpha\delta}{16L^2 n}$ reaches error $H_\nu(\rho_k) \leq \delta$ after $k \geq \frac{1}{\alpha\epsilon} \log \frac{2H_\nu(\rho_0)}{\delta}$ iterations.

For example, if we start with a Gaussian $\rho_0 = \mathcal{N}(x^*, \frac{1}{L} I)$ where x^* is a stationary point of f (which we can find, e.g., via gradient descent), then $H_\nu(\rho_0) = \tilde{O}(n)$ (see Lemma 1), and Theorem 2 gives an iteration complexity of $k = \tilde{\Theta}\big(\frac{L^2 n}{\alpha^2 \delta}\big)$ to achieve $H_\nu(\rho_k) \leq \delta$ using ULA with step size $\epsilon = \Theta(\frac{\alpha\delta}{L^2 n})$.

The result above matches previous known bounds for ULA when ν is strongly logconcave [17, 21, 22, 24]. Our result complements the recent work of Ma et al. [45] who study the underdamped version of the Langevin dynamics under LSI and show an iteration complexity for the discrete-time algorithm that has better dependence on the dimension ($\sqrt{\frac{n}{\delta}}$ in place of $\frac{n}{\delta}$ above for ULA), but under an additional smoothness assumption (f has bounded third derivatives) and with higher polynomial dependence on other parameters. Our result also complements the work of Mangoubi and Vishnoi [49] who study the Metropolis-adjusted version of ULA (MALA) for non-logconcave ν and show a $\log(\frac{1}{\delta})$ iteration complexity from a warm start, under the additional assumption that f has bounded third and fourth derivatives in an appropriate ∞-norm.

We note that in general some isoperimetry condition is needed for rapid mixing of Markov chains (such as Langevin dynamics and ULA), otherwise there are bad regions in the state space from which the chains take arbitrarily long to escape. Smoothness or bounded Hessian is a common assumption needed for the analysis of discrete-time algorithms (such as gradient descent or ULA).

In the second part of this paper, we study the convergence of Rényi divergence of order $q > 1$ along ULA. Rényi divergence is a family of generalizations of KL divergence [56, 59, 11], which becomes stronger as the order q increases. There are physical and operational interpretations of Rényi divergence [31, 3]. Rényi divergence has been useful in many applications, including for the exponential mechanism in differential privacy [27, 1, 12, 52], lattice-based cryptography [4], information-theoretic encryption [35], variational inference [41], machine learning [32, 50], information theory and statistics [20, 53], and black hole physics [23].

Our second result proves a convergence bound for the Rényi divergence of order $q > 1$. While Rényi divergence is a stronger measure of convergence than KL divergence, the situation is more complicated. First, we can hope to converge to the biased limit ν_ϵ only for finite q for any step size ϵ (as we illustrate with an example). Second, it is unclear how to bound the Rényi divergence between ν_ϵ and ν. We first show the following convergence guarantees of Rényi divergence along the continuous-time Langevin dynamics under LSI or Poincaré inequality; see Theorem 3 and Theorem 5. Here $R_{q,\nu}(\rho)$ is the Rényi divergence of order q between ρ and ν.

Theorem 3. *Suppose ν satisfies LSI with constant $\alpha > 0$. Let $q \geq 1$. Along the Langevin dynamics,*

$$R_{q,\nu}(\rho_t) \leq e^{-\frac{2\alpha t}{q}} R_{q,\nu}(\rho_0).$$

Theorem 5. *Suppose ν satisfies Poincaré inequality with constant $\alpha > 0$. Let $q \geq 2$. Along the Langevin dynamics,*

$$R_{q,\nu}(\rho_t) \leq \begin{cases} R_{q,\nu}(\rho_0) - \frac{2\alpha t}{q} & \text{if } R_{q,\nu}(\rho_0) \geq 1 \text{ and as long as } R_{q,\nu}(\rho_t) \geq 1, \\ e^{-\frac{2\alpha t}{q}} R_{q,\nu}(\rho_0) & \text{if } R_{q,\nu}(\rho_0) \leq 1. \end{cases}$$

Notice that under Poincaré inequality, compared to LSI, the convergence is slower in the beginning before it becomes exponential. For a reasonable starting distribution (such as a Gaussian centered at a stationary point), this leads to an extra factor of n compared to the convergence under LSI. We then turn to discrete time and show the convergence of Rényi divergence along ULA to the biased limit ν_ϵ under the assumption that ν_ϵ itself satisfies either LSI or Poincaré inequality. We combine this with a decomposition result on Rényi divergence to derive a convergence guarantee for Rényi divergence to ν along ULA; see Theorem 4 and Theorem 6.

In what follows, we review KL divergence and its properties along the Langevin dynamics in Section 2, and prove a convergence guarantee for KL divergence along ULA under LSI in Section 3. We provide a review of Rényi divergence and its properties along the Langevin dynamics in Section 4. We then prove the convergence guarantee for Rényi divergence along ULA under LSI in Section 5, and under Poincaré inequality in Section 6. We conclude with a discussion in Section 7.

2 Review of KL divergence along Langevin dynamics

In this section we review the definition of Kullback-Leibler (KL) divergence, log-Sobolev inequality, and the convergence of KL divergence along the Langevin dynamics in continuous time under log-Sobolev inequality. See Appendix A.1 for a review on notation.

2.1 KL divergence

Let ρ, ν be probability distributions on $\mathbb{R}^n$, represented via their probability density functions with respect to the Lebesgue measure on $\mathbb{R}^n$. We assume ρ, ν have full support and smooth densities.

Recall the **Kullback-Leibler (KL) divergence** of ρ with respect to ν is

$$H_\nu(\rho) = \int_{\mathbb{R}^n} \rho(x) \log \frac{\rho(x)}{\nu(x)} \, dx. \tag{1}$$

KL divergence is the relative form of *Shannon entropy* $H(\rho) = -\int_{\mathbb{R}^n} \rho(x) \log \rho(x) \, dx$. Whereas Shannon entropy can be positive or negative, KL divergence is nonnegative and minimized at ν:

$H_\nu(\rho) \geq 0$ for all ρ, and $H_\nu(\rho) = 0$ if and only if $\rho = \nu$. Therefore, KL divergence serves as a measure of (albeit asymmetric) "distance" of a probability distribution ρ from a base distribution ν. KL divergence is a relatively strong measure of distance; for example, Pinsker's inequality implies that KL divergence controls total variation distance. Furthermore, under log-Sobolev (or Talagrand) inequality, KL divergence also controls the quadratic Wasserstein W_2 distance, as we review below. We say $\nu = e^{-f}$ is L-**smooth** if f has bounded Hessian: $-LI \preceq \nabla^2 f(x) \preceq LI$ for all $x \in \mathbb{R}^n$. We provide the proof of Lemma 1 in Appendix B.1.1.

Lemma 1. *Suppose $\nu = e^{-f}$ is L-smooth. Let $\rho = \mathcal{N}(x^*, \frac{1}{L} I)$ where x^* is a stationary point of f. Then $H_\nu(\rho) \leq f(x^*) + \frac{n}{2} \log \frac{L}{2\pi}$.*

2.2 Log-Sobolev inequality

Recall we say ν satisfies the **log-Sobolev inequality (LSI)** with a constant $\alpha > 0$ if for all smooth function $g \colon \mathbb{R}^n \to \mathbb{R}$ with $\mathbb{E}_\nu[g^2] < \infty$,

$$\mathbb{E}_\nu[g^2 \log g^2] - \mathbb{E}_\nu[g^2] \log \mathbb{E}_\nu[g^2] \leq \frac{2}{\alpha} \mathbb{E}_\nu[\|\nabla g\|^2]. \tag{2}$$

Recall the **relative Fisher information** of ρ with respect to ν is

$$J_\nu(\rho) = \int_{\mathbb{R}^n} \rho(x) \left\| \nabla \log \frac{\rho(x)}{\nu(x)} \right\|^2 dx. \tag{3}$$

LSI is equivalent to the following relation between KL divergence and Fisher information for all ρ:

$$H_\nu(\rho) \leq \frac{1}{2\alpha} J_\nu(\rho). \tag{4}$$

Indeed, to obtain (4) we choose $g^2 = \frac{\rho}{\nu}$ in (2); conversely, to obtain (2) we choose $\rho = \frac{g^2 \nu}{\mathbb{E}_\nu[g^2]}$ in (4).

LSI is an isoperimetry condition and implies, among others, concentration of measure and sub-Gaussian tail property [38]. LSI was first shown by Gross [30] for the case of Gaussian ν. It was extended by Bakry and Émery [6] to strongly log-concave ν; namely, when $f = -\log \nu$ is α-strongly convex, then ν satisfies LSI with constant α. However, LSI applies more generally. For example, the classical perturbation result by Holley and Stroock [33] states that LSI is stable under bounded perturbation. Furthermore, LSI is preserved under a Lipschitz mapping. In one dimension, there is an exact characterization of when a probability distribution on $\mathbb{R}$ satisfies LSI [9]. Moreover, LSI satisfies a tensorization property [38]: If ν_1, ν_2 satisfy LSI with constants $\alpha_1, \alpha_2 > 0$, respectively, then $\nu_1 \otimes \nu_2$ satisfies LSI with constant $\min\{\alpha_1, \alpha_2\} > 0$. Thus, there are many examples of non-logconcave distributions ν on $\mathbb{R}^n$ satisfying LSI (with a constant independent of dimension). There are also Lyapunov function criteria and exponential integrability conditions that can be used to verify when a probability distribution satisfies LSI; see for example [14, 15, 51, 61, 7].

2.2.1 Talagrand inequality

Recall the **Wasserstein distance** between ρ and ν is

$$W_2(\rho, \nu) = \inf_{\Pi} \mathbb{E}_\Pi[\|X - Y\|^2]^{\frac{1}{2}} \tag{5}$$

where the infimum is over joint distributions Π of (X, Y) with the correct marginals $X \sim \rho, Y \sim \nu$. Recall we say ν satisfies **Talagrand inequality** with a constant $\alpha > 0$ if for all ρ:

$$\frac{\alpha}{2} W_2(\rho, \nu)^2 \leq H_\nu(\rho). \tag{6}$$

Talagrand's inequality implies concentration of measure of Gaussian type. It was first studied by Talagrand [58] for Gaussian ν, and extended by Otto and Villani [54] to all ν satisfying LSI; namely, if ν satisfies LSI with constant $\alpha > 0$, then ν also satisfies Talagrand's inequality with the same constant [54, Theorem 1]. Therefore, under LSI, KL divergence controls the Wasserstein distance. Moreover, when ν is log-concave, LSI and Talagrand's inequality are equivalent [54, Corollary 3.1]. We recall in Appendix A.2 the geometric interpretation of LSI and Talagrand's inequality from [54].

2.3 Langevin dynamics

The **Langevin dynamics** for target distribution $\nu = e^{-f}$ is a continuous-time stochastic process $(X_t)_{t\geq 0}$ in $\mathbb{R}^n$ that evolves following the stochastic differential equation:

$$dX_t = -\nabla f(X_t)\, dt + \sqrt{2}\, dW_t \tag{7}$$

where $(W_t)_{t\geq 0}$ is the standard Brownian motion in $\mathbb{R}^n$ with $W_0 = 0$.

If $(X_t)_{t\geq 0}$ evolves following the Langevin dynamics (7), then their probability density function $(\rho_t)_{t\geq 0}$ evolves following the **Fokker-Planck equation**:

$$\frac{\partial \rho_t}{\partial t} \;=\; \nabla \cdot (\rho_t \nabla f) + \Delta \rho_t \;=\; \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\nu}\right). \tag{8}$$

Here $\nabla\cdot$ is the divergence and Δ is the Laplacian operator. We provide a derivation in Appendix A.3. From (8), if $\rho_t = \nu$, then $\frac{\partial \rho_t}{\partial t} = 0$, so ν is the stationary distribution for the Langevin dynamics (7). Moreover, the Langevin dynamics brings any distribution $X_t \sim \rho_t$ closer to the target distribution ν, as the following lemma shows.

Lemma 2. *Along the Langevin dynamics* (7) *(or equivalently, the Fokker-Planck equation* (8)*),*

$$\frac{d}{dt} H_\nu(\rho_t) = -J_\nu(\rho_t). \tag{9}$$

We provide the proof of Lemma 2 in Appendix B.1.2. Since $J_\nu(\rho) \geq 0$, the identity (9) shows KL divergence is decreasing along the Langevin dynamics, so indeed the distribution ρ_t converges to ν.

2.3.1 Exponential convergence of KL divergence along Langevin dynamics under LSI

When ν satisfies LSI, KL divergence converges exponentially fast along the Langevin dynamics.

Theorem 1. *Suppose ν satisfies LSI with constant $\alpha > 0$. Along the Langevin dynamics* (7)*,*

$$H_\nu(\rho_t) \leq e^{-2\alpha t} H_\nu(\rho_0). \tag{10}$$

Furthermore, $W_2(\rho_t, \nu) \leq \sqrt{\frac{2}{\alpha} H_\nu(\rho_0)}\, e^{-\alpha t}$.

We provide the proof of Theorem 1 in Appendix B.1.3. We also recall the optimization interpretation of Langevin dynamics as the gradient flow of KL divergence in the space of distributions with the Wasserstein metric [36, 60, 54]. Then the exponential convergence rate in Theorem 1 is a manifestation of the general fact that gradient flow converges exponentially fast under gradient domination condition. This provides a justification for using the Langevin dynamics for sampling from ν, as a natural steepest descent flow that minimizes the KL divergence H_ν.

3 Unadjusted Langevin Algorithm

Suppose we wish to sample from a smooth target probability distribution $\nu = e^{-f}$ in $\mathbb{R}^n$. The **Unadjusted Langevin Algorithm (ULA)** with step size $\epsilon > 0$ is the discrete-time algorithm

$$x_{k+1} = x_k - \epsilon \nabla f(x_k) + \sqrt{2\epsilon}\, z_k \tag{11}$$

where $z_k \sim \mathcal{N}(0, I)$ is an independent standard Gaussian random variable in $\mathbb{R}^n$. Let ρ_k denote the probability distribution of x_k that evolves following ULA.

As $\epsilon \to 0$, ULA recovers the Langevin dynamics (7) in continuous-time. However, for fixed $\epsilon > 0$, ULA converges to a biased limiting distribution $\nu_\epsilon \neq \nu$. Therefore, KL divergence $H_\nu(\rho_k)$ does not tend to 0 along ULA, as it has an asymptotic bias $H_\nu(\nu_\epsilon) > 0$.

Example 1. *Let $\nu = \mathcal{N}(0, \frac{1}{\alpha} I)$. The ULA iteration is $x_{k+1} = (1 - \epsilon\alpha) x_k + \sqrt{2\epsilon} z_k$. For $0 < \epsilon < \frac{2}{\alpha}$, the limit is $\nu_\epsilon = \mathcal{N}\big(0, \frac{1}{\alpha(1-\frac{\epsilon\alpha}{2})}\big)$ and the bias is $H_\nu(\nu_\epsilon) = \frac{n}{2}\big(\frac{\epsilon\alpha}{2(1-\frac{\epsilon\alpha}{2})} + \log(1 - \frac{\epsilon\alpha}{2})\big)$. In particular, $H_\nu(\nu_\epsilon) \leq \frac{n\epsilon^2\alpha^2}{16(1-\frac{\epsilon\alpha}{2})^2} = O(\epsilon^2)$.*

3.1 Convergence of KL divergence along ULA under LSI

When ν satisfies LSI and a smoothness condition, we can prove a convergence guarantee in KL divergence along ULA. Recall we say $\nu = e^{-f}$ is L-smooth if $-LI \preceq \nabla^2 f(x) \preceq LI$ for all $x \in \mathbb{R}^n$. A key part in our analysis is the following lemma which bounds the decrease in KL divergence along one iteration of ULA. Here $x_{k+1} \sim \rho_{k+1}$ is the output of one step of ULA (11) from $x_k \sim \rho_k$.

Lemma 3. *Suppose ν satisfies LSI with constant $\alpha > 0$ and is L-smooth. If $0 < \epsilon \leq \frac{\alpha}{4L^2}$, then along each step of ULA* (11)*,*

$$H_\nu(\rho_{k+1}) \leq e^{-\alpha\epsilon} H_\nu(\rho_k) + 6\epsilon^2 n L^2. \tag{12}$$

We provide the proof of Lemma 3 in Appendix B.2.1. The proof of Lemma 3 compares the evolution of KL divergence along one step of ULA with the evolution along Langevin dynamics in continuous time (which converges exponentially fast under LSI), and bounds the discretization error; see Figure 2 for an illustration. This comparison technique has been used in many papers. Our proof structure is similar to that of Cheng and Bartlett [17], whose analysis needs ν to be strongly logconcave.

With Lemma 3, we can prove our main result on the convergence rate of ULA under LSI. We provide the proof of Theorem 2 in Appendix B.2.2.

Theorem 2. *Assume $\nu = e^{-f}$ satisfies log-Sobolev inequality with constant $\alpha > 0$ and is L-smooth. ULA with step size $0 < \epsilon \leq \frac{\alpha}{4L^2}$ satisfies*

$$H_\nu(\rho_k) \leq e^{-\alpha\epsilon k} H_\nu(\rho_0) + \tfrac{8\epsilon n L^2}{\alpha}.$$

For $0 < \delta < 4n$, ULA with $\epsilon \leq \frac{\alpha\delta}{16L^2 n}$ reaches error $H_\nu(\rho_k) \leq \delta$ after $k \geq \frac{1}{\alpha\epsilon} \log \frac{2H_\nu(\rho_0)}{\delta}$ iterations.

In particular, suppose $\delta < 4n$ and we choose the largest permissible step size $\epsilon = \Theta\big(\frac{\alpha\delta}{L^2 n}\big)$. Suppose we start with a Gaussian $\rho_0 = \mathcal{N}(x^*, \frac{1}{L} I)$, where x^* is a stationary point of f (which we can find, e.g., via gradient descent), so $H_\nu(\rho_0) \leq f(x^*) + \frac{n}{2} \log \frac{L}{2\pi} = \tilde{O}(n)$ by Lemma 1. Theorem 2 states that to achieve $H_\nu(\rho_k) \leq \delta$, ULA has iteration complexity $k = \tilde{\Theta}\left(\frac{L^2 n}{\alpha^2 \delta}\right)$. Since LSI implies Talagrand's inequality, Theorem 2 also yields a convergence guarantee in Wasserstein distance. As $k \to \infty$, Theorem 2 implies the following bound on the bias of ULA under LSI. However, we note the bound $O(\epsilon)$ may be loose, since from Example 1 we see $H_\nu(\nu_\epsilon) = \Theta(\epsilon^2)$ in Gaussian case.

Corollary 1. *Suppose ν satisfies LSI with constant $\alpha > 0$ and is L-smooth. For $0 < \epsilon \leq \frac{\alpha}{4L^2}$, the biased limit ν_ϵ of ULA with step size ϵ satisfies $H_\nu(\nu_\epsilon) \leq \frac{8nL^2\epsilon}{\alpha}$ and $W_2(\nu, \nu_\epsilon)^2 \leq \frac{16nL^2\epsilon}{\alpha^2}$.*

4 Review of Rényi divergence along Langevin dynamics

4.1 Rényi divergence

Rényi divergence [56] is a family of generalizations of KL divergence. See [59, 11] for properties of Rényi divergence.

For $q > 0$, $q \neq 1$, the **Rényi divergence** of order q of a probability distribution ρ with respect to ν is

$$R_{q,\nu}(\rho) = \frac{1}{q-1} \log F_{q,\nu}(\rho) \tag{13}$$

where

$$F_{q,\nu}(\rho) = \mathbb{E}_\nu \left[\left(\frac{\rho}{\nu} \right)^q \right] = \int_{\mathbb{R}^n} \nu(x) \frac{\rho(x)^q}{\nu(x)^q} \, dx = \int_{\mathbb{R}^n} \frac{\rho(x)^q}{\nu(x)^{q-1}} dx. \tag{14}$$

Rényi divergence is the relative form of *Rényi entropy* [56]: $H_q(\rho) = \frac{1}{q-1} \log \int \rho(x)^q \, dx$. The case $q = 1$ is defined via limit, and recovers the KL divergence (1): $R_{1,\nu}(\rho) = \lim_{q \to 1} R_{q,\nu}(\rho) = H_\nu(\rho)$. Rényi divergence has the property that $R_{q,\nu}(\rho) \geq 0$ for all ρ, and $R_{q,\nu}(\rho) = 0$ if and only if $\rho = \nu$. Furthermore, the map $q \mapsto R_{q,\nu}(\rho)$ is increasing (see Section B.3.1). Therefore, Rényi divergence provides an alternative measure of "distance" of ρ from ν, which becomes stronger as q increases. In particular, $R_{\infty,\nu}(\rho) = \log \left\| \frac{\rho}{\nu} \right\|_\infty = \log \sup_x \frac{\rho(x)}{\nu(x)}$ is finite if and only if ρ is *warm* relative to ν. It is possible that $R_{q,\nu}(\rho) = \infty$ for large enough q, as the following example shows.

Example 2. *Let $\rho = \mathcal{N}(0, \sigma^2 I)$ and $\nu = \mathcal{N}(0, \lambda^2 I)$. If $\sigma^2 > \lambda^2$ and $q \geq \frac{\sigma^2}{\sigma^2 - \lambda^2}$, then $R_{q,\nu}(\rho) = \infty$. Otherwise, $R_{q,\nu}(\rho) = \frac{n}{2} \log \frac{\lambda^2}{\sigma^2} - \frac{n}{2(q-1)} \log \left(q - (q-1)\frac{\sigma^2}{\lambda^2}\right)$.*

The following is analogous to Lemma 1. We provide the proof of Lemma 4 in Appendix B.3.2.

Lemma 4. *Suppose $\nu = e^{-f}$ is L-smooth. Let $\rho = \mathcal{N}(x^*, \frac{1}{L} I)$ where x^* is a stationary point of f. Then for all $q \geq 1$, $R_{q,\nu}(\rho) \leq f(x^*) + \frac{n}{2} \log \frac{L}{2\pi}$.*

4.1.1 Log-Sobolev inequality

For $q > 0$, we define the **Rényi information** of order q of ρ with respect to ν as

$$G_{q,\nu}(\rho) = \mathbb{E}_\nu \left[\left(\frac{\rho}{\nu}\right)^q \left\| \nabla \log \frac{\rho}{\nu} \right\|^2 \right] = \mathbb{E}_\nu \left[\left(\frac{\rho}{\nu}\right)^{q-2} \left\| \nabla \frac{\rho}{\nu} \right\|^2 \right] = \frac{4}{q^2} \mathbb{E}_\nu \left[\left\| \nabla \left(\frac{\rho}{\nu}\right)^{\frac{q}{2}} \right\|^2 \right]. \tag{15}$$

The case $q = 1$ recovers relative Fisher information (3): $G_{1,\nu}(\rho) = \mathbb{E}_\nu \left[\frac{\rho}{\nu} \left\| \nabla \log \frac{\rho}{\nu} \right\|^2 \right] = J_\nu(\rho)$. We have the following relation under log-Sobolev inequality. Note the case $q = 1$ recovers LSI (4). We provide the proof of Lemma 5 in Appendix B.3.3.

Lemma 5. *Suppose ν satisfies LSI with constant $\alpha > 0$. Let $q \geq 1$. For all ρ,*

$$\frac{G_{q,\nu}(\rho)}{F_{q,\nu}(\rho)} \geq \frac{2\alpha}{q^2} R_{q,\nu}(\rho). \tag{16}$$

4.2 Langevin dynamics

Along the Langevin dynamics (7) for ν, we can compute the rate of change of the Rényi divergence.

Lemma 6. *For all $q > 0$, along the Langevin dynamics* (7)*,*

$$\frac{d}{dt} R_{q,\nu}(\rho_t) = -q \frac{G_{q,\nu}(\rho_t)}{F_{q,\nu}(\rho_t)}. \tag{17}$$

We provide the proof of Lemma 6 in Appendix B.3.4. In particular, $\frac{d}{dt} R_{q,\nu}(\rho_t) \leq 0$, so Rényi divergence is always decreasing along the Langevin dynamics. Furthermore, analogous to how the Langevin dynamics is the gradient flow of KL divergence under the Wasserstein metric, one can also show that the Langevin dynamics is the the gradient flow of Rényi divergence with respect to a suitably defined metric (which depends on ν) on the space of distributions; see [13].

4.2.1 Convergence of Rényi divergence along Langevin dynamics under LSI

When ν satisfies LSI, Rényi divergence converges exponentially fast along the Langevin dynamics. Note the case $q = 1$ recovers the exponential convergence rate of KL divergence from Theorem 1.

Theorem 3. *Suppose ν satisfies LSI with constant $\alpha > 0$. Let $q \geq 1$. Along the Langevin dynamics,*

$$R_{q,\nu}(\rho_t) \leq e^{-\frac{2\alpha t}{q}} R_{q,\nu}(\rho_0).$$

We provide the proof of Theorem 3 in Appendix B.3.5. Theorem 3 shows that if the initial Rényi divergence is finite, then it converges exponentially fast. However, even if initially the Rényi divergence is ∞, it will be finite along the Langevin dynamics, after which time Theorem 3 applies. This is because when ν satisfies LSI, the Langevin dynamics satisfies a *hypercontractivity* property [30, 10, 60]; see Section B.3.6. Furthermore, as shown in [13], we can combine the exponential convergence rate above with the hypercontractivity property to improve the exponential rate to be 2α, independent of q, at the cost of some initial waiting time; here we leave the rate as above for simplicity.

5 Rényi divergence along ULA

In this section we prove a convergence guarantee for Rényi divergence along ULA under the assumption that the biased limit satisfies LSI. As before, let $\nu = e^{-f}$, and let ν_ϵ denote the biased limit of ULA (11) with step size $\epsilon > 0$. We note that the bias $R_{q,\nu}(\nu_\epsilon)$ may be ∞ for large enough q.

Example 3. *As in Examples 1 and 2, let $\nu = \mathcal{N}(0, \frac{1}{\alpha} I)$, so $\nu_\epsilon = \mathcal{N}\big(0, \frac{1}{\alpha(1-\frac{\epsilon\alpha}{2})} I\big)$. The bias is*

$$R_{q,\nu}(\nu_\epsilon) = \begin{cases} \frac{n}{2(q-1)} \left(q \log\left(1 - \frac{\epsilon\alpha}{2}\right) - \log\left(1 - \frac{q\epsilon\alpha}{2}\right)\right) & \text{if } 1 < q < \frac{2}{\epsilon\alpha}, \\ \infty & \text{if } q \geq \frac{2}{\epsilon\alpha}. \end{cases}$$

Thus, for each fixed $q > 1$, there is an asymptotic bias $R_{q,\nu}(\nu_\epsilon)$ which is finite for small enough ϵ. In Example 3 we have $R_{q,\nu}(\nu_\epsilon) = O(\epsilon^2)$. In general, we assume for each $q > 1$ there is a **growth function** $g_q(\epsilon)$ that controls the bias: $R_{q,\nu}(\nu_\epsilon) \leq g_q(\epsilon)$ for small $\epsilon > 0$, and $\lim_{\epsilon \to 0} g_q(\epsilon) = 0$.

5.1 Decomposition of Rényi divergence

For order $q > 1$, we have the following decomposition of Rényi divergence.

Lemma 7. *Let $q > 1$. For all probability distribution ρ,*

$$R_{q,\nu}(\rho) \leq \left(\frac{q - \frac{1}{2}}{q-1}\right) R_{2q,\nu_\epsilon}(\rho) + R_{2q-1,\nu}(\nu_\epsilon). \tag{18}$$

We provide the proof of Lemma 7 in Appendix B.4.1. The first term in the bound above is the Rényi divergence with respect to the biased limit, which converges exponentially fast under LSI (see Lemma 8). The second term in (18) is the bias, which is controlled by the growth function $g_{2q-1}(\epsilon)$.

5.2 Rapid convergence of Rényi divergence with respect to ν_ϵ along ULA

We show Rényi divergence with respect to the biased limit ν_ϵ converges exponentially fast along ULA, assuming ν_ϵ itself satisfies LSI.

Assumption 1. *The probability distribution ν_ϵ satisfies LSI with a constant $\beta \equiv \beta_\epsilon > 0$.*

We can verify Assumption 1 in the Gaussian case. However, it is unclear how to verify Assumption 1 in general. One might hope to prove that if ν satisfies LSI, then Assumption 1 holds.

Example 4. *Let $\nu = \mathcal{N}(0, \frac{1}{\alpha} I)$, so $\nu_\epsilon = \mathcal{N}\big(0, \frac{1}{\alpha(1-\frac{\epsilon\alpha}{2})} I\big)$ satisfies LSI with $\beta = \alpha(1 - \frac{\epsilon\alpha}{2})$.*

Under Assumption 1, we can prove an exponential convergence rate to the biased limit ν_ϵ.

Lemma 8. *Assume Assumption 1. Suppose $\nu = e^{-f}$ is L-smooth, and let $0 < \epsilon \leq \min\left\{\frac{1}{3L}, \frac{1}{9\beta}\right\}$. For $q \geq 1$, along ULA* (11)*,*

$$R_{q,\nu_\epsilon}(\rho_k) \leq e^{-\frac{\beta\epsilon k}{q}} R_{q,\nu_\epsilon}(\rho_0). \tag{19}$$

We provide the proof of Lemma 8 in Appendix B.4.2. In the proof of Lemma 8, we decompose each step of ULA as a sequence of two operations; see Figure 3 for an illustration. In the first part we take a gradient step, which is a deterministic bijective map, so it preserves Rényi divergence. In the second part we add an independent Gaussian, which is evolution along the heat flow, and we derive a formula on the decrease in Rényi divergence (which is similar to (17) along the Langevin dynamics).

5.3 Convergence of Rényi divergence along ULA under LSI

We combine Lemma 7 and Lemma 8 to obtain the following characterization of the convergence of Rényi divergence along ULA under LSI. We provide the proof of Theorem 4 in Appendix B.4.3.

Theorem 4. *Assume Assumption 1. Suppose $\nu = e^{-f}$ is L-smooth, and let $0 < \epsilon \leq \min\left\{\frac{1}{3L}, \frac{1}{9\beta}\right\}$. Let $q > 1$, and suppose $R_{2q,\nu_\epsilon}(\rho_0) < \infty$. Then along ULA* (11)*,*

$$R_{q,\nu}(\rho_k) \leq \left(\frac{q - \frac{1}{2}}{q-1}\right) R_{2q,\nu_\epsilon}(\rho_0) e^{-\frac{\beta\epsilon k}{2q}} + g_{2q-1}(\epsilon). \tag{20}$$

For $\delta > 0$, let $g_q^{-1}(\delta) = \sup\{\epsilon > 0 \colon g_q(\epsilon) \leq \delta\}$. Theorem 4 states that to achieve $R_{q,\nu}(\rho_k) \leq \delta$, it suffices to run ULA with step size $\epsilon = \Theta\left(\min\left\{\frac{1}{L}, g_{2q-1}^{-1}\left(\frac{\delta}{2}\right)\right\}\right)$ for $k = O\big(\frac{1}{\beta\epsilon} \log \frac{R_{2q,\nu_\epsilon}(\rho_0)}{\delta}\big)$ iterations. Suppose δ is small so $g_{2q-1}^{-1}\left(\frac{\delta}{2}\right) < \frac{1}{L}$. Note ν_ϵ is $\frac{1}{2\epsilon}$-smooth, so if we choose ρ_0

to be a Gaussian with covariance $2\epsilon I$, we have $R_{2q,\nu_\epsilon}(\rho_0) = \tilde{O}(n)$ by Lemma 4. Therefore, Theorem 4 yields an iteration complexity of $k = \tilde{O}\big(\frac{1}{\beta g_{2q-1}^{-1}(\delta/2)}\big)$. For example, if $g_q(\epsilon) = O(\epsilon)$, then $g_q^{-1}(\delta) = \Omega(\delta)$, so the iteration complexity is $k = \tilde{O}\big(\frac{1}{\beta\delta}\big)$ with $\epsilon = \Theta(\delta)$. If $g_q(\epsilon) = O(\epsilon^2)$, as in Example 3, then $g_q^{-1}(\delta) = \Omega(\sqrt{\delta})$, so the iteration complexity is $k = \tilde{O}\big(\frac{1}{\beta\sqrt{\delta}}\big)$ with $\epsilon = \Theta(\sqrt{\delta})$.

6 Poincaré inequality

We recall ν satisfies **Poincaré inequality (PI)** with a constant $\alpha > 0$ if for all smooth $g\colon \mathbb{R}^n \to \mathbb{R}$,

$$\mathrm{Var}_\nu(g) \le \tfrac{1}{\alpha}\mathbb{E}_\nu[\|\nabla g\|^2] \tag{21}$$

where $\mathrm{Var}_\nu(g) = \mathbb{E}_\nu[g^2] - \mathbb{E}_\nu[g]^2$ is the variance of g under ν. Poincaré inequality is an isoperimetry condition which is weaker than LSI. LSI implies PI with the same constant; in fact, PI is a linearization of LSI (4), i.e., when $\rho = (1+\eta g)\nu$ as $\eta \to 0$ [57, 60]. Furthermore, it is known Talagrand's inequality implies PI with the same constant, and PI is also a linearization of Talagrand's inequality [54]. Poincaré inequality is better behaved than LSI [15], and there are various Lyapunov criteria and integrability conditions to verify when a distribution satisfies Poincaré inequality [5, 51, 19].

6.1 Convergence of Rényi divergence along Langevin dynamics under Poincaré inequality

When ν satisfies Poincaré inequality, Rényi divergence converges along the Langevin dynamics. The convergence is initially linear, then becomes exponential once Rényi divergence falls below 1.

Theorem 5. *Suppose ν satisfies Poincaré inequality with constant $\alpha > 0$. Let $q \ge 2$. Along the Langevin dynamics,*

$$R_{q,\nu}(\rho_t) \le \begin{cases} R_{q,\nu}(\rho_0) - \frac{2\alpha t}{q} & \text{if } R_{q,\nu}(\rho_0) \ge 1 \text{ and as long as } R_{q,\nu}(\rho_t) \ge 1, \\ e^{-\frac{2\alpha t}{q}} R_{q,\nu}(\rho_0) & \text{if } R_{q,\nu}(\rho_0) \le 1. \end{cases}$$

We provide the proof of Theorem 5 in Appendix B.5.2. Theorem 5 states that starting from $R_{q,\nu}(\rho_0) \ge 1$, the Langevin dynamics reaches $R_{q,\nu}(\rho_t) \le \delta$ in $t \le O\left(\frac{q}{\alpha}\left(R_{q,\nu}(\rho_0) + \log\frac{1}{\delta}\right)\right)$ time.

6.2 Rapid convergence of Rényi divergence with respect to ν_ϵ along ULA

We assume the biased limit ν_ϵ satisfies Poincaré inequality.

Assumption 2. *The distribution ν_ϵ satisfies Poincaré inequality with a constant $\beta \equiv \beta_\epsilon > 0$.*

Under Assumption 2 we can show Rényi divergence with respect to ν_ϵ converges at a rate similar to the Langevin dynamics; see Lemma 18 in Appendix B.5.3.

6.3 Convergence of Rényi divergence along ULA under Poincaré inequality

We combine Lemma 7 and Lemma 18 to obtain the following convergence of Rényi divergence along ULA under Poincaré inequality. We provide the proof of Theorem 6 in Appendix B.5.4.

Theorem 6. *Assume Assumption 2. Suppose $\nu = e^{-f}$ is L-smooth, and let $0 < \epsilon \le \min\left\{\frac{1}{3L}, \frac{1}{9\beta}\right\}$. Let $q > 1$ and assume $1 \le R_{2q,\nu_\epsilon}(\rho_0) < \infty$. Along ULA (11), for $k \ge k_0 := \frac{2q}{\beta\epsilon}(R_{2q,\nu_\epsilon}(\rho_0) - 1)$,*

$$R_{q,\nu}(\rho_k) \le \left(\frac{q-\frac{1}{2}}{q-1}\right) e^{-\frac{\beta\epsilon(k-k_0)}{2q}} + g_{2q-1}(\epsilon). \tag{22}$$

This yields an iteration complexity for ULA under Poincaré which is a factor of n larger than the complexity under LSI; see Appendix B.5.5.

7 Discussion

In this paper we proved convergence guarantees on KL and Rényi divergence along ULA under isoperimetry and bounded Hessian, without assuming convexity or bounds on higher derivatives. It would be interesting to verify when Assumptions 1 and 2 hold or whether they follow from isoperimetry and bounded Hessian of the target density. Another intriguing question is whether there is an affine-invariant version of the Langevin dynamics. This might lead to a sampling algorithm with logarithmic dependence on smoothness parameters, rather than the current polynomial dependence.

Acknowledgment

The first author was supported in part by NSF awards CCF-1563838 and CCF-1717349. The authors would like to thank Kunal Talwar for explaining the application of Rényi divergence to data privacy. The authors thank Yu Cao, Jianfeng Lu, and Yulong Lu for alerting us to their work [13] on Rényi divergence. The authors also thank Xiang Cheng and Peter Bartlett for helpful comments on an earlier version of this paper.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.

[2] David Applegate and Ravi Kannan. Sampling and integration of near log-concave functions. In *Proceedings of the Twenty-third Annual ACM Symposium on Theory of Computing*, STOC '91, pages 156–163, New York, NY, USA, 1991. ACM.

[3] John C Baez. Rényi entropy and free energy. *arXiv preprint arXiv:1102.2098*, 2011.

[4] Shi Bai, Tancrède Lepoint, Adeline Roux-Langlois, Amin Sakzad, Damien Stehlé, and Ron Steinfeld. Improved security proofs in lattice-based cryptography: using the rényi divergence rather than the statistical distance. *Journal of Cryptology*, 31(2):610–640, 2018.

[5] Dominique Bakry, Franck Barthe, Patrick Cattiaux, Arnaud Guillin, et al. A simple proof of the Poincaré inequality for a large class of probability measures. *Electronic Communications in Probability*, 13:60–66, 2008.

[6] Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84*, pages 177–206. Springer, 1985.

[7] Jean-Baptiste Bardet, Nathaël Gozlan, Florent Malrieu, and Pierre-André Zitt. Functional inequalities for Gaussian convolutions of compactly supported measures: Explicit bounds and dimension dependence. *Bernoulli*, 24(1):333–353, 2018.

[8] Espen Bernton. Langevin Monte Carlo and JKO splitting. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, pages 1777–1798, 2018.

[9] Sergej G Bobkov and Friedrich Götze. Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 163(1):1–28, 1999.

[10] Sergey G Bobkov, Ivan Gentil, and Michel Ledoux. Hypercontractivity of Hamilton–Jacobi equations. *Journal de Mathématiques Pures et Appliquées*, 80(7):669–696, 2001.

[11] SG Bobkov, GP Chistyakov, and Friedrich Götze. Rényi divergence and the central limit theorem. *The Annals of Probability*, 47(1):270–323, 2019.

[12] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[13] Yu Cao, Jianfeng Lu, and Yulong Lu. Exponential decay of Rényi divergence under Fokker–Planck equations. *Journal of Statistical Physics*, pages 1–13, 2018.

[14] Djalil Chafaï. Entropies, convexity, and functional inequalities: On ϕ-entropies and ϕ-Sobolev inequalities. *Journal of Mathematics of Kyoto University*, 44(2):325–363, 2004.

[15] Djalil Chafai and Florent Malrieu. On fine properties of mixtures with respect to concentration of measure and Sobolev type inequalities. In *Annales de l'IHP Probabilités et statistiques*, volume 46, pages 72–96, 2010.

[16] Zongchen Chen and Santosh S. Vempala. Optimal convergence rate of Hamiltonian Monte Carlo for strongly logconcave distributions. *arXiv preprint arXiv:1905.02313*, 2019.

[17] Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. In Firdaus Janoos, Mehryar Mohri, and Karthik Sridharan, editors, *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 186–211. PMLR, 07–09 Apr 2018.

[18] Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

[19] Thomas A Courtade. Bounds on the Poincaré constant for convolution measures. *arXiv preprint arXiv:1807.00027*, 2018.

[20] Imre Csiszár. Generalized cutoff rates and Rényi's information measures. *IEEE Transactions on information theory*, 41(1):26–34, 1995.

[21] Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 678–689. PMLR, 07–10 Jul 2017.

[22] Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 2019.

[23] Xi Dong. The gravity dual of Rényi entropy. *Nature communications*, 7:12472, 2016.

[24] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *arXiv preprint arXiv:1802.09188*, 2018.

[25] Alain Durmus, Eric Moulines, and Eero Saksman. On the convergence of Hamiltonian Monte Carlo. *arXiv preprint arXiv:1705.00166*, 2017.

[26] Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, pages 793–797, 2018.

[27] Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

[28] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability*, 47(4):1982–2010, 2019.

[29] Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1292–1301, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[30] Leonard Gross. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

[31] Peter Harremoës. Interpretations of Rényi entropies and divergences. *Physica A: Statistical Mechanics and its Applications*, 365(1):57–62, 2006.

[32] Yun He, A Ben Hamza, and Hamid Krim. A generalized divergence measure for robust image registration. *IEEE Transactions on Signal Processing*, 51(5):1211–1220, 2003.

[33] Richard Holley and Daniel Stroock. Logarithmic Sobolev inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46(5):1159–1194, 1987.

[34] Richard Holley and Daniel Stroock. Simulated annealing via Sobolev inequalities. *Communications in Mathematical Physics*, 115(4):553–569, 1988.

[35] Mitsugu Iwamoto and Junji Shikata. Information theoretic security for encryption based on conditional Rényi entropies. In *International Conference on Information Theoretic Security*, pages 103–121. Springer, 2013.

[36] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, January 1998.

[37] R. Kannan, L. Lovász, and M. Simonovits. Random walks and an $O^*(n^5)$ volume algorithm for convex bodies. *Random Structures and Algorithms*, 11:1–50, 1997.

[38] Michel Ledoux. Concentration of measure and logarithmic Sobolev inequalities. *Séminaire de probabilités de Strasbourg*, 33:120–216, 1999.

[39] Yin Tat Lee and Santosh S Vempala. Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1115–1121. ACM, 2018.

[40] Xuechen Li, Denny Wu, Lester Mackey, and Murat A Erdogdu. Stochastic Runge–Kutta accelerates Langevin Monte Carlo and beyond. *arXiv preprint arXiv:1906.07868*, 2019.

[41] Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1073–1081. Curran Associates, Inc., 2016.

[42] L. Lovász and S. Vempala. Fast algorithms for logconcave functions: sampling, rounding, integration and optimization. In *FOCS*, pages 57–68, 2006.

[43] L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Struct. Algorithms*, 30(3):307–358, 2007.

[44] László Lovász and Santosh S. Vempala. Hit-and-run from a corner. *SIAM Journal on Computing*, 35(4):985–1005, 2006.

[45] Yi-An Ma, Niladri Chatterji, Xiang Cheng, Nicolas Flammarion, Peter Bartlett, and Michael I Jordan. Is there an analog of Nesterov acceleration for MCMC? *arXiv preprint arXiv:1902.00996*, 2019.

[46] Michael C. Mackey. *Time's Arrow: The Origins of Thermodynamics Behavior*. Springer-Verlag, 1992.

[47] Oren Mangoubi and Aaron Smith. Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. *arXiv preprint arXiv:1708.07114*, 2017.

[48] Oren Mangoubi and Nisheeth Vishnoi. Dimensionally tight bounds for second-order Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems 31*, pages 6027–6037. Curran Associates, Inc., 2018.

[49] Oren Mangoubi and Nisheeth K Vishnoi. Nonconvex sampling with the Metropolis-adjusted Langevin algorithm. *arXiv preprint arXiv:1902.08452*, 2019.

[50] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the Rényi divergence. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 367–374. AUAI Press, 2009.

[51] Georg Menz and André Schlichting. Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape. *Ann. Probab.*, 42(5):1809–1884, 09 2014.

[52] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

[53] D Morales, L Pardo, and I Vajda. Rényi statistics in directed families of exponential experiments. *Statistics: A Journal of Theoretical and Applied Statistics*, 34(2):151–174, 2000.

[54] Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

[55] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

[56] Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[57] OS Rothaus. Diffusion on compact Riemannian manifolds and logarithmic Sobolev inequalities. *Journal of functional analysis*, 42(1):102–109, 1981.

[58] M Talagrand. Transportation cost for Gaussian and other product measures. *Geometric and Functional Analysis*, 6:587–600, 01 1996.

[59] Tim Van Erven and Peter Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[60] Cédric Villani. *Topics in optimal transportation*. Number 58 in Graduate Studies in Mathematics. American Mathematical Society, 2003.

[61] Feng-Yu Wang and Jian Wang. Functional inequalities for convolution of probability measures. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 52, pages 898–914. Institut Henri Poincaré, 2016.

[62] Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, pages 2093–3027, 2018.

A Review

A.1 Review on notation and basic properties

Throughout, we represent a probability distribution ρ on $\mathbb{R}^n$ via its probability density function with respect to the Lebesgue measure, so $\rho\colon \mathbb{R}^n \to \mathbb{R}$ with $\int_{\mathbb{R}^n} \rho(x)dx = 1$. We typically assume ρ has full support and smooth density, so $\rho(x) > 0$ and $x \mapsto \rho(x)$ is differentiable. Given a function $f\colon \mathbb{R}^n \to \mathbb{R}$, we denote the expected value of f under ρ by

$$\mathbb{E}_\rho[f] = \int_{\mathbb{R}^n} f(x)\rho(x)\,dx.$$

We use the Euclidean inner product $\langle x, y\rangle = \sum_{i=1}^n x_i y_i$ for $x = (x_i)_{1\le i\le n}, y = (y_i)_{1\le i\le n} \in \mathbb{R}^n$. For symmetric matrices $A, B \in \mathbb{R}^{n\times n}$, let $A \preceq B$ denote that $B - A$ is positive semidefinite. For $\mu \in \mathbb{R}^n$, $\Sigma \succ 0$, let $\mathcal{N}(\mu, \Sigma)$ denote the Gaussian distribution on $\mathbb{R}^n$ with mean μ and covariance matrix Σ.

Given a smooth function $f\colon \mathbb{R}^n \to \mathbb{R}$, its **gradient** $\nabla f\colon \mathbb{R}^n \to \mathbb{R}^n$ is the vector of partial derivatives:

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \dots, \frac{\partial f(x)}{\partial x_n}\right).$$

The **Hessian** $\nabla^2 f\colon \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is the matrix of second partial derivatives:

$$\nabla^2 f(x) = \left(\frac{\partial^2 f(x)}{\partial x_i x_j}\right)_{1\le i,j\le n}.$$

The **Laplacian** $\Delta f\colon \mathbb{R}^n \to \mathbb{R}$ is the trace of its Hessian:

$$\Delta f(x) = \mathrm{Tr}(\nabla^2 f(x)) = \sum_{i=1}^n \frac{\partial^2 f(x)}{\partial x_i^2}.$$

Given a smooth vector field $v = (v_1, \dots, v_n)\colon \mathbb{R}^n \to \mathbb{R}^n$, its **divergence** $\nabla \cdot v\colon \mathbb{R}^n \to \mathbb{R}$ is

$$(\nabla \cdot v)(x) = \sum_{i=1}^n \frac{\partial v_i(x)}{\partial x_i}.$$

In particular, the divergence of gradient is the Laplacian:

$$(\nabla \cdot \nabla f)(x) = \sum_{i=1}^n \frac{\partial^2 f(x)}{\partial x_i^2} = \Delta f(x).$$

For any function $f\colon \mathbb{R}^n \to \mathbb{R}$ and vector field $v\colon \mathbb{R}^n \to \mathbb{R}^n$ with sufficiently fast decay at infinity, we have the following **integration by parts** formula:

$$\int_{\mathbb{R}^n} \langle v(x), \nabla f(x)\rangle dx = -\int_{\mathbb{R}^n} f(x)(\nabla \cdot v)(x)dx.$$

Furthermore, for any two functions $f, g\colon \mathbb{R}^n \to \mathbb{R}$,

$$\int_{\mathbb{R}^n} f(x)\Delta g(x)dx = -\int_{\mathbb{R}^n} \langle \nabla f(x), \nabla g(x)\rangle dx = \int_{\mathbb{R}^n} g(x)\Delta f(x)dx.$$

When the argument is clear, we omit the argument (x) in the formulae for brevity. For example, the last integral above becomes

$$\int f\,\Delta g\,dx = -\int \langle \nabla f, \nabla g\rangle\,dx = \int g\,\Delta f\,dx. \tag{23}$$

A.2 Geometric interpretation of LSI and Talagrand's inequality

In the space of probability distributions with the Riemannian metric defined by the Wasserstein W_2 distance, the Fisher information (3) is the squared norm of the gradient of KL divergence (1). Therefore, LSI (4) is the gradient dominated condition (also known as the Polyak-Łojaciewicz (PL) inequality) for KL divergence. On the other hand, Talagrand's inequality (6) is the quadratic growth condition for KL divergence. In general, the gradient dominated condition implies the quadratic growth condition [54, Proposition 1']. Therefore, LSI implies Talagrand's inequality.

A.3 Derivation of the Fokker-Planck equation

Consider a stochastic differential equation

$$dX = v(X)\,dt + \sqrt{2}\,dW \tag{24}$$

where $v\colon \mathbb{R}^n \to \mathbb{R}^n$ is a smoth vector field and $(W_t)_{t\geq 0}$ is the Brownian motion on $\mathbb{R}^n$ with $W_0 = 0$.

We will show that if X_t evolves following (24), then its probability density function $\rho_t(x)$ evolves following the Fokker-Planck equation:

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot (\rho v) + \Delta \rho. \tag{25}$$

We can derive this heuristically as follows; we refer to standard textbooks for rigorous derivation [46].

For any smooth test function $\phi\colon \mathbb{R}^n \to \mathbb{R}$, let us compute the time derivative of the expectation

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

On the one hand, we can compute this as

$$\dot{A}(t) = \frac{d}{dt}A(t) = \frac{d}{dt}\int_{\mathbb{R}^n} \rho_t(x)\phi(x)\,dx = \int_{\mathbb{R}^n} \frac{\partial \rho_t(x)}{\partial t}\phi(x)\,dx. \tag{26}$$

On the other hand, by (24), for small $\epsilon > 0$ we have

$$\begin{aligned} X_{t+\epsilon} &= X_t + \int_t^{t+\epsilon} v(X_s)ds + \sqrt{2}(W_{t+\epsilon} - W_t) \\ &= X_t + \epsilon v(X_t) + \sqrt{2}(W_{t+\epsilon} - W_t) + O(\epsilon^2) \\ &\overset{d}{=} X_t + \epsilon v(X_t) + \sqrt{2\epsilon}Z + O(\epsilon^2) \end{aligned}$$

where $Z \sim \mathcal{N}(0, I)$ is independent of X_t, since $W_{t+\epsilon} - W_t \sim \mathcal{N}(0, \epsilon I)$. Then by Taylor expansion,

$$\begin{aligned} \phi(X_{t+\epsilon}) &\overset{d}{=} \phi\left(X_t + \epsilon v(X_t) + \sqrt{2\epsilon}Z + O(\epsilon^2)\right) \\ &= \phi(X_t) + \epsilon\langle \nabla\phi(X_t), v(X_t)\rangle + \sqrt{2\epsilon}\langle \nabla\phi(X_t), Z\rangle + \frac{1}{2}2\epsilon\langle Z, \nabla^2\phi(X_t)Z\rangle + O(\epsilon^{\frac{3}{2}}). \end{aligned}$$

Now we take expectation on both sides. Since $Z \sim \mathcal{N}(0, I)$ is independent of X_t,

$$\begin{aligned} A(t+\epsilon) &= \mathbb{E}[\phi(X_{t+\epsilon})] \\ &= \mathbb{E}\left[\phi(X_t) + \epsilon\langle \nabla\phi(X_t), v(X_t)\rangle + \sqrt{2\epsilon}\langle \nabla\phi(X_t), Z\rangle + \epsilon\langle Z, \nabla^2\phi(X_t)Z\rangle\right] + O(\epsilon^{\frac{3}{2}}) \\ &= A(t) + \epsilon\left(\mathbb{E}[\langle \nabla\phi(X_t), v(X_t)\rangle] + \mathbb{E}[\Delta\phi(X_t)]\right) + O(\epsilon^{\frac{3}{2}}). \end{aligned}$$

Therefore, by integration by parts, this second approach gives

$$\begin{aligned} \dot{A}(t) &= \lim_{\epsilon\to 0}\frac{A(t+\epsilon) - A(t)}{\epsilon} \\ &= \mathbb{E}[\langle \nabla\phi(X_t), v(X_t)\rangle] + \mathbb{E}[\Delta\phi(X_t)] \\ &= \int_{\mathbb{R}^n}\langle \nabla\phi(x), \rho_t(x)v(x)\rangle dx + \int_{\mathbb{R}^n}\rho_t(x)\Delta\phi(x)\,dx \\ &= -\int_{\mathbb{R}^n}\phi(x)\nabla\cdot(\rho_t v)(x)\,dx + \int_{\mathbb{R}^n}\phi(x)\Delta\rho_t(x)\,dx \\ &= \int_{\mathbb{R}^n}\phi(x)\left(-\nabla\cdot(\rho_t v)(x) + \Delta\rho_t(x)\right)\,dx. \end{aligned} \tag{27}$$

Comparing (26) and (27), and since ϕ is arbitrary, we conclude that

$$\frac{\partial \rho_t(x)}{\partial t} = -\nabla\cdot(\rho_t v)(x) + \Delta\rho_t(x)$$

as claimed in (25).

When $v = -\nabla f$, the stochastic differential equation (24) becomes the Langevin dynamics (7) from Section 2.3, and the Fokker-Planck equation (25) becomes (8).

In the proof of Lemma 3, we also apply the Fokker-Planck equation (25) when $v = -\nabla f(x_0)$ is a constant vector field to derive the evolution equation (30) for one step of ULA.

B Proofs and details

B.1 Proofs for §2: KL divergence along Langevin dynamics

B.1.1 Proof of Lemma 1

Proof of Lemma 1. Since f is L-smooth and $\nabla f(x^*) = 0$, we have the bound

$$f(x) \le f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \frac{L}{2}\|x - x^*\|^2 = f(x^*) + \frac{L}{2}\|x - x^*\|^2.$$

Let $X \sim \rho = \mathcal{N}(x^*, \frac{1}{L}I)$. Then

$$\mathbb{E}_\rho[f(X)] \le f(x^*) + \frac{L}{2}\mathrm{Var}_\rho(X) = f(x^*) + \frac{n}{2}.$$

Recall the entropy of ρ is $H(\rho) = -\mathbb{E}_\rho[\log \rho(X)] = \frac{n}{2}\log\frac{2\pi e}{L}$. Therefore, the KL divergence is

$$H_\nu(\rho) = \int \rho\,(\log\rho + f)\,dx = -H(\rho) + \mathbb{E}_\rho[f] \le f(x^*) + \frac{n}{2}\log\frac{L}{4\pi e}.$$

□

B.1.2 Proof of Lemma 2

Proof of Lemma 2. Recall the time derivative of KL divergence along any flow is given by

$$\frac{d}{dt}H_\nu(\rho_t) = \frac{d}{dt}\int_{\mathbb{R}^n} \rho_t \log\frac{\rho_t}{\nu}\,dx = \int_{\mathbb{R}^n} \frac{\partial \rho_t}{\partial t}\log\frac{\rho_t}{\nu}\,dx$$

since the second part of the chain rule is zero: $\int \rho_t \frac{\partial}{\partial t}\log\frac{\rho_t}{\nu}\,dx = \int \frac{\partial \rho_t}{\partial t}\,dx = \frac{d}{dt}\int \rho_t\,dx = 0$. Therefore, along the Fokker-Planck equation (8) for the Langevin dynamics (7),

$$\begin{aligned}\frac{d}{dt}H_\nu(\rho_t) &= \int \nabla\cdot\left(\rho_t \nabla\log\frac{\rho_t}{\nu}\right)\log\frac{\rho_t}{\nu}\,dx \\ &= -\int \rho_t \left\|\nabla\log\frac{\rho_t}{\nu}\right\|^2 dx \\ &= -J_\nu(\rho_t)\end{aligned}$$

where in the second equality we have applied integration by parts. □

B.1.3 Proof of Theorem 1

Proof of Theorem 1. From Lemma 2 and the LSI assumption (4),

$$\frac{d}{dt}H_\nu(\rho_t) = -J_\nu(\rho_t) \le -2\alpha H_\nu(\rho_t).$$

Integrating implies the desired bound $H_\nu(\rho_t) \le e^{-2\alpha t}H_\nu(\rho_0)$.

Furthermore, since ν satisfies LSI with constant α, it also satisfies Talagrand's inequality (6) with constant α [54, Theorem 1]. Therefore, $W_2(\rho_t,\nu)^2 \le \frac{2}{\alpha}H_\nu(\rho_t) \le \frac{2}{\alpha}e^{-2\alpha t}H_\nu(\rho_0)$, as desired. □

B.2 Proofs for §3: Unadjusted Langevin Algorithm

B.2.1 Proof of Lemma 3

We will use the following auxiliary results.

Lemma 9. *Assume $\nu = e^{-f}$ is L-smooth. Then*

$$\mathbb{E}_\nu[\|\nabla f\|^2] \le nL.$$

Proof. Since $\nu = e^{-f}$, by integration by parts we can write

$$\mathbb{E}_\nu[\|\nabla f\|^2] = \mathbb{E}_\nu[\Delta f].$$

Since ν is L-smooth, $\nabla^2 f(x) \preceq L\,I$, so $\Delta f(x) \le nL$ for all $x \in \mathbb{R}^n$. Therefore, $\mathbb{E}_\nu[\|\nabla f\|^2] = \mathbb{E}_\nu[\Delta f] \le nL$, as desired. □

Lemma 10. *Suppose ν satisfies Talagrand's inequality with constant $\alpha > 0$ and is L-smooth. For any ρ,*

$$\mathbb{E}_\rho[\|\nabla f\|^2] \le \frac{4L^2}{\alpha} H_\nu(\rho) + 2nL.$$

Proof. Let $x \sim \rho$ and $x^* \sim \nu$ with an optimal coupling (x, x^*) so that $\mathbb{E}[\|x - x^*\|^2] = W_2(\rho, \nu)^2$. Since $\nu = e^{-f}$ is L-smooth, ∇f is L-Lipschitz. By triangle inequality,

$$\begin{aligned}\|\nabla f(x)\| &\le \|\nabla f(x) - \nabla f(x^*)\| + \|\nabla f(x^*)\| \\ &\le L\|x - x^*\| + \|\nabla f(x^*)\|.\end{aligned}$$

Squaring, using $(a+b)^2 \le 2a^2 + 2b^2$, and taking expectation, we get

$$\begin{aligned}\mathbb{E}_\rho[\|\nabla f(x)\|^2] &\le 2L^2\, \mathbb{E}[\|x - x^*\|^2] + 2\mathbb{E}_\nu[\|\nabla f(x^*)\|^2] \\ &= 2L^2\, W_2(\rho, \nu)^2 + 2\mathbb{E}_\nu[\|\nabla f(x^*)\|^2].\end{aligned}$$

By Talagrand's inequality (6), $W_2(\rho, \nu)^2 \le \frac{2}{\alpha} H_\nu(\rho)$. By Lemma 9 we have $\mathbb{E}_\nu[\|\nabla f(x^*)\|^2] \le nL$. Plugging these to the bound above gives the desired result. □

We are now ready to prove Lemma 3. See Figure 2 for an illustration.

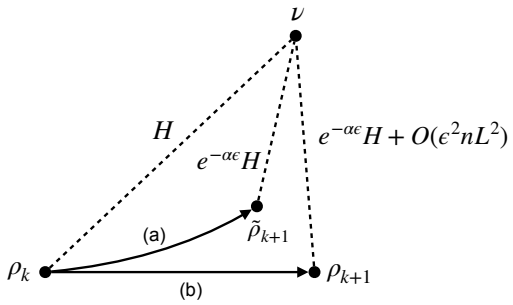

Figure 2: An illustration for the proof of Lemma 3. In each iteration, we compare the evolution of (a) the continuous-time Langevin dynamics for time ϵ, and (b) one step of ULA. If the current KL divergence is $H \equiv H_\nu(\rho_k)$, then after the Langevin dynamics (a) the KL divergence is $H_\nu(\tilde\rho_{k+1}) \le e^{-\alpha\epsilon} H$, and we show that after ULA (b) the KL divergence is $H_\nu(\rho_{k+1}) \le e^{-\alpha\epsilon} H + O(\epsilon^2 n L^2)$.

Proof of Lemma 3. For simplicity suppose $k = 0$, so we start at $x_0 \sim \rho_0$. We write one step of ULA

$$x_0 \mapsto x_0 - \epsilon \nabla f(x_0) + \sqrt{2\epsilon} z_0$$

as the output at time ϵ of the stochastic differential equation

$$dx_t = -\nabla f(x_0)\, dt + \sqrt{2}\, dW_t \tag{28}$$

where W_t is the standard Brownian motion in $\mathbb{R}^n$ starting at $W_0 = 0$. Indeed, the solution to (28) at time $t = \epsilon$ is

$$\begin{aligned}x_\epsilon &= x_0 - \epsilon \nabla f(x_0) + \sqrt{2}\, W_\epsilon \\ &\overset{d}{=}\ x_0 - \epsilon \nabla f(x_0) + \sqrt{2\epsilon}\, z_0.\end{aligned} \tag{29}$$

where $z_0 \sim \mathcal{N}(0, I)$, which is identical to the ULA update.

We derive the continuity equation corresponding to (28) as follows. For each $t > 0$, let $\rho_{0t}(x_0, x_t)$ denote the joint distribution of (x_0, x_t), which we write in terms of the conditionals and marginals as

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

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

$$\frac{\partial \rho_{t|0}(x_t \,|\, x_0)}{\partial t} = \nabla \cdot \big(\rho_{t|0}(x_t \,|\, x_0) \nabla f(x_0)\big) + \Delta \rho_{t|0}(x_t \,|\, x_0). \tag{30}$$

To derive the evolution of ρ_t, we take expectation over $x_0 \sim \rho_0$. Multiplying both sides of (30) by $\rho_0(x_0)$ and integrating over x_0, we obtain

$$\begin{aligned}
\frac{\partial \rho_t(x)}{\partial t} &= \int_{\mathbb{R}^n} \frac{\partial \rho_{t|0}(x \,|\, x_0)}{\partial t} \, \rho_0(x_0) \, dx_0 \\
&= \int_{\mathbb{R}^n} \big(\nabla \cdot \big(\rho_{t|0}(x \,|\, x_0) \nabla f(x_0)\big) + \Delta \rho_{t|0}(x \,|\, x_0)\big) \, \rho_0(x_0) \, dx_0 \\
&= \int_{\mathbb{R}^n} \left(\nabla \cdot \left(\rho_{t,0}(x, x_0) \nabla f(x_0)\right) + \Delta \rho_{t,0}(x, x_0)\right) dx_0 \\
&= \nabla \cdot \left(\rho_t(x) \int_{\mathbb{R}^n} \rho_{0|t}(x_0 \,|\, x) \nabla f(x_0) \, dx_0\right) + \Delta \rho_t(x) \\
&= \nabla \cdot \big(\rho_t(x) \mathbb{E}_{\rho_{0|t}}[\nabla f(x_0) \,|\, x_t = x]\big) + \Delta \rho_t(x).
\end{aligned} \tag{31}$$

Observe that the difference between the Fokker-Planck equations (31) for ULA and (8) for Langevin dynamics is in the first term, that the drift is now the conditional expectation $\mathbb{E}_{\rho_{0|t}}[\nabla f(x_0) \,|\, x_t = x]$, rather than the true gradient $\nabla f(x)$.

Recall the time derivative of relative entropy along any flow is given by

$$\frac{d}{dt} H_\nu(\rho_t) = \frac{d}{dt} \int_{\mathbb{R}^n} \rho_t \log \frac{\rho_t}{\nu} \, dx = \int_{\mathbb{R}^n} \frac{\partial \rho_t}{\partial t} \log \frac{\rho_t}{\nu} \, dx$$

since the second part of the chain rule is zero: $\int \rho_t \frac{\partial}{\partial t} \log \frac{\rho_t}{\nu} \, dx = \int \frac{\partial \rho_t}{\partial t} \, dx = \frac{d}{dt} \int \rho_t \, dx = 0$.

Therefore, the time derivative of relative entropy for ULA, using the Fokker-Planck equation (31) and integrating by parts, is given by:

$$\begin{aligned}
\frac{d}{dt} H_\nu(\rho_t) &= \int_{\mathbb{R}^n} \big(\nabla \cdot \big(\rho_t(x) \mathbb{E}_{\rho_{0|t}}[\nabla f(x_0) \,|\, x_t = x]\big) + \Delta \rho_t(x)\big) \, \log \frac{\rho_t(x)}{\nu(x)} \, dx \\
&= \int_{\mathbb{R}^n} \left(\nabla \cdot \left(\rho_t(x) \left(\nabla \log \frac{\rho_t(x)}{\nu(x)} + \mathbb{E}_{\rho_{0|t}}[\nabla f(x_0) \,|\, x_t = x] - \nabla f(x)\right)\right)\right) \log \frac{\rho_t(x)}{\nu(x)} \, dx \\
&= -\int_{\mathbb{R}^n} \rho_t(x) \left\langle \nabla \log \frac{\rho_t(x)}{\nu(x)} + \mathbb{E}_{\rho_{0|t}}[\nabla f(x_0) \,|\, x_t = x] - \nabla f(x), \, \nabla \log \frac{\rho_t(x)}{\nu(x)} \right\rangle dx \\
&= -\int_{\mathbb{R}^n} \rho_t(x) \left\| \nabla \log \frac{\rho_t}{\nu} \right\|^2 dx + \int_{\mathbb{R}^n} \rho_t(x) \left\langle \nabla f(x) - \mathbb{E}_{\rho_{0|t}}[\nabla f(x_0) \,|\, x_t = x], \, \nabla \log \frac{\rho_t(x)}{\nu(x)} \right\rangle dx \\
&= -J_\nu(\rho_t) + \int_{\mathbb{R}^n \times \mathbb{R}^n} \rho_{0t}(x_0, x) \left\langle \nabla f(x) - \nabla f(x_0), \nabla \log \frac{\rho_t(x)}{\nu(x)} \right\rangle dx_0 \, dx \\
&= -J_\nu(\rho_t) + \mathbb{E}_{\rho_{0t}} \left[\left\langle \nabla f(x_t) - \nabla f(x_0), \, \nabla \log \frac{\rho_t(x_t)}{\nu(x_t)} \right\rangle\right]
\end{aligned} \tag{32}$$

where in the last step we have renamed x as x_t. The first term in (32) is the same as in the Langevin dynamics. The second term in (32) is the discretization error, which we can bound as follows. Using $\langle a, b\rangle \le \|a\|^2 + \frac{1}{4}\|b\|^2$ and since ∇f is L-Lipschitz,

$$\begin{aligned}
\mathbb{E}_{\rho_{0t}} \left[\left\langle \nabla f(x_t) - \nabla f(x_0), \nabla \log \frac{\rho_t(x_t)}{\nu(x_t)} \right\rangle\right] &\le \mathbb{E}_{\rho_{0t}}[\|\nabla f(x_t) - \nabla f(x_0)\|^2] + \frac{1}{4} \mathbb{E}_{\rho_{0t}} \left[\left\| \nabla \log \frac{\rho_t(x_t)}{\nu(x_t)} \right\|^2\right] \\
&= \mathbb{E}_{\rho_{0t}}[\|\nabla f(x_t) - \nabla f(x_0)\|^2] + \frac{1}{4} J_\nu(\rho_t) \\
&\le L^2 \mathbb{E}_{\rho_{0t}}[\|x_t - x_0\|^2] + \frac{1}{4} J_\nu(\rho_t)
\end{aligned} \tag{33}$$

Recall from (29) the solution of ULA 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

$$\begin{aligned}\mathbb{E}_{\rho_{0t}}[\|x_t - x_0\|^2] &= \mathbb{E}_{\rho_{0t}}[\| - t\nabla f(x_0) + \sqrt{2t} z_0\|^2] \\ &= t^2 \mathbb{E}_{\rho_0}[\|\nabla f(x_0)\|^2] + 2tn \\ &\le \frac{4t^2L^2}{\alpha} H_\nu(\rho_0) + 2t^2 nL + 2tn\end{aligned}$$

where in the last inequality we have used Lemma 10. This bounds the discretization error by

$$\mathbb{E}_{\rho_{0t}}\left[\left\langle \nabla f(x_t) - \nabla f(x_0), \nabla \log \frac{\rho_t(x_t)}{\nu(x_t)} \right\rangle\right] \le \frac{4t^2L^4}{\alpha} H_\nu(\rho_0) + 2t^2 nL^3 + 2tnL^2 + \frac{1}{4} J_\nu(\rho_t).$$

Therefore, from (32), the time derivative of KL divergence along ULA is bounded by

$$\frac{d}{dt} H_\nu(\rho_t) \le -\frac{3}{4} J_\nu(\rho_t) + \frac{4t^2L^4}{\alpha} H_\nu(\rho_0) + 2t^2 nL^3 + 2tnL^2.$$

Then by the LSI (4) assumption,

$$\frac{d}{dt} H_\nu(\rho_t) \le -\frac{3\alpha}{2} H_\nu(\rho_t) + \frac{4t^2L^4}{\alpha} H_\nu(\rho_0) + 2t^2 nL^3 + 2tnL^2.$$

We wish to integrate the inequality above for $0 \le t \le \epsilon$. Using $t \le \epsilon$ and since $\epsilon \le \frac{1}{2L}$, we simplify the above to

$$\begin{aligned}\frac{d}{dt} H_\nu(\rho_t) &\le -\frac{3\alpha}{2} H_\nu(\rho_t) + \frac{4\epsilon^2L^4}{\alpha} H_\nu(\rho_0) + 2\epsilon^2 nL^3 + 2\epsilon nL^2 \\ &\le -\frac{3\alpha}{2} H_\nu(\rho_t) + \frac{4\epsilon^2L^4}{\alpha} H_\nu(\rho_0) + 3\epsilon nL^2.\end{aligned}$$

Multiplying both sides by $e^{\frac{3\alpha}{2}t}$, we can write the above as

$$\frac{d}{dt}\left(e^{\frac{3\alpha}{2}t} H_\nu(\rho_t)\right) \le e^{\frac{3\alpha}{2}t}\left(\frac{4\epsilon^2L^4}{\alpha} H_\nu(\rho_0) + 3\epsilon nL^2\right).$$

Integrating from $t = 0$ to $t = \epsilon$ gives

$$\begin{aligned}e^{\frac{3}{2}\alpha\epsilon} H_\nu(\rho_\epsilon) - H_\nu(\rho_0) &\le \frac{2(e^{\frac{3}{2}\alpha\epsilon} - 1)}{3\alpha}\left(\frac{4\epsilon^2L^4}{\alpha} H_\nu(\rho_0) + 3\epsilon nL^2\right) \\ &\le 2\epsilon\left(\frac{4\epsilon^2L^4}{\alpha} H_\nu(\rho_0) + 3\epsilon nL^2\right)\end{aligned}$$

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

$$H_\nu(\rho_\epsilon) \le e^{-\frac{3}{2}\alpha\epsilon}\left(1 + \frac{8\epsilon^3L^4}{\alpha}\right) H_\nu(\rho_0) + e^{-\frac{3}{2}\alpha\epsilon} 6\epsilon^2 nL^2.$$

Since $1 + \frac{8\epsilon^3L^4}{\alpha} \le 1 + \frac{\alpha\epsilon}{2} \le e^{\frac{1}{2}\alpha\epsilon}$ for $\epsilon \le \frac{\alpha}{4L^2}$, and using $e^{-\frac{3}{2}\alpha\epsilon} \le 1$, we conclude that

$$H_\nu(\rho_\epsilon) \le e^{-\alpha\epsilon} H_\nu(\rho_0) + 6\epsilon^2 nL^2.$$

This is the desired inequality, after renaming $\rho_0 \equiv \rho_k$ and $\rho_\epsilon \equiv \rho_{k+1}$. Note that the conditions $\epsilon \le \frac{1}{2L}$ and $\epsilon \le \frac{2}{3\alpha}$ above are also implied by the assumption $\epsilon \le \frac{\alpha}{4L^2}$ since $\alpha \le L$. □

B.2.2 Proof of Theorem 2

Proof of Theorem 2. Applying the recursion (12) from Lemma 3, we obtain

$$H_\nu(\rho_k) \;\le\; e^{-\alpha\epsilon k} H_\nu(\rho_0) + \frac{6\epsilon^2 nL^2}{1 - e^{-\alpha\epsilon}} \;\le\; e^{-\alpha\epsilon k} H_\nu(\rho_0) + \frac{8\epsilon nL^2}{\alpha}$$

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

Given $\delta > 0$, if we further assume $\epsilon \le \frac{\delta\alpha}{16nL^2}$, then the above implies $H_\nu(\rho_k) \le e^{-\alpha\epsilon k} H_\nu(\rho_0) + \frac{\delta}{2}$. This means for $k \ge \frac{1}{\alpha\epsilon} \log \frac{2H_\nu(\rho_0)}{\delta}$, we have $H_\nu(\rho_k) \le \frac{\delta}{2} + \frac{\delta}{2} = \delta$, as desired. □

B.3 Details for §4: Rényi divergence along Langevin dynamics

B.3.1 Properties of Rényi divergence

We show that Rényi divergence is increasing in the order.

Lemma 11. *For any probability distributions ρ, ν, the map $q \mapsto R_{q,\nu}(\rho)$ is increasing for $q > 0$.*

Proof. Let $0 < q \le r$. We will show that $R_{q,\nu}(\rho) \le R_{r,\nu}(\rho)$.

First suppose $q > 1$. We write $F_{q,\nu}(\rho)$ as an expectation over ρ and use power mean inequality:

$$F_{q,\nu}(\rho) = \mathbb{E}_\nu\left[\left(\frac{\rho}{\nu}\right)^q\right] = \mathbb{E}_\rho\left[\left(\frac{\rho}{\nu}\right)^{q-1}\right] \le \mathbb{E}_\rho\left[\left(\frac{\rho}{\nu}\right)^{r-1}\right]^{\frac{q-1}{r-1}} = \mathbb{E}_\nu\left[\left(\frac{\rho}{\nu}\right)^r\right]^{\frac{q-1}{r-1}} = F_{r,\nu}(\rho)^{\frac{q-1}{r-1}}.$$

Taking logarithm and dividing by $q - 1 > 0$ gives

$$R_{q,\nu}(\rho) = \frac{1}{q-1}\log F_{q,\nu}(\rho) \le \frac{1}{r-1}\log F_{r,\nu}(\rho) = R_{r,\nu}(\rho).$$

The case $q = 1$ follows by taking limit $q \to 1$.

Now suppose $q \le r < 1$, so $1 - q \ge 1 - r > 0$. We again write $F_{q,\nu}(\rho)$ as an expectation over ρ and use power mean inequality:

$$F_{q,\nu}(\rho) = \mathbb{E}_\nu\left[\left(\frac{\rho}{\nu}\right)^q\right] = \mathbb{E}_\rho\left[\left(\frac{\nu}{\rho}\right)^{1-q}\right] \ge \mathbb{E}_\rho\left[\left(\frac{\nu}{\rho}\right)^{1-r}\right]^{\frac{1-q}{1-r}} = \mathbb{E}_\nu\left[\left(\frac{\rho}{\nu}\right)^r\right]^{\frac{1-q}{1-r}} = F_{r,\nu}(\rho)^{\frac{1-q}{1-r}}.$$

Taking logarithm and dividing by $q - 1 < 0$ (which flips the inequality) gives

$$R_{q,\nu}(\rho) = \frac{1}{q-1}\log F_{q,\nu}(\rho) \le \frac{1}{r-1}\log F_{r,\nu}(\rho) = R_{r,\nu}(\rho).$$

The case $q < 1 \le r$ follows since $R_{q,\nu}(\rho) \le R_{1,\nu}(\rho) \le R_{r,\nu}(\rho)$. □

B.3.2 Proof of Lemma 4

Proof of Lemma 4. Since f is L-smooth and x^* is a stationary point of f, for all $x \in \mathbb{R}^n$ we have

$$f(x) \le f(x^*) + \langle \nabla f(x^*), x - x^*\rangle + \frac{L}{2}\|x - x^*\|^2 = f(x^*) + \frac{L}{2}\|x - x^*\|^2.$$

Let $q > 1$. Then for $\rho = \mathcal{N}(x^*, \sigma^2 I)$ with $\frac{q}{\sigma^2} > (q-1)L$,

$$\begin{aligned}
F_{q,\nu}(\rho) &= \int_{\mathbb{R}^n} \frac{\rho(x)^q}{\nu(x)^{q-1}} dx \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{nq}{2}}} \int_{\mathbb{R}^n} e^{-\frac{q}{2\sigma^2}\|x-x^*\|^2 + (q-1)f(x)} dx \\
&\le \frac{1}{(2\pi\sigma^2)^{\frac{nq}{2}}} \int_{\mathbb{R}^n} e^{-\frac{q}{2\sigma^2}\|x-x^*\|^2 + (q-1)f(x^*) + \frac{(q-1)L}{2}\|x-x^*\|^2} dx \\
&= \frac{e^{(q-1)f(x^*)}}{(2\pi\sigma^2)^{\frac{nq}{2}}} \int_{\mathbb{R}^n} e^{-\frac{1}{2}\left(\frac{q}{\sigma^2} - (q-1)L\right)\|x-x^*\|^2} dx \\
&= \frac{e^{(q-1)f(x^*)}}{(2\pi\sigma^2)^{\frac{nq}{2}}} \left(\frac{2\pi}{\frac{q}{\sigma^2} - (q-1)L}\right)^{\frac{n}{2}} \\
&= \frac{e^{(q-1)f(x^*)}}{(2\pi)^{\frac{n}{2}(q-1)}(\sigma^2)^{\frac{nq}{2}}} \frac{1}{\left(\frac{q}{\sigma^2} - (q-1)L\right)^{\frac{n}{2}}}.
\end{aligned}$$

Therefore,

$$R_{q,\nu}(\rho) = \frac{1}{q-1}\log F_{q,\nu}(\rho) \le f(x^*) - \frac{n}{2}\log 2\pi - \frac{n}{2(q-1)}\log\sigma^{2q}\left(\frac{q}{\sigma^2} - (q-1)L\right).$$

In particular, if $\sigma^2 = \frac{1}{L}$, then $\frac{q}{\sigma^2} - (q-1)L = L > 0$, and the bound above becomes

$$R_{q,\nu}(\rho) \le f(x^*) + \frac{n}{2}\log\frac{L}{2\pi}.$$

The case $q = 1$ follows from Lemma 1, since $\frac{1}{4\pi e} < \frac{1}{2\pi}$. □

B.3.3 Proof of Lemma 5

Proof of Lemma 5. We plug in $h^2 = \left(\frac{\rho}{\nu}\right)^q$ to the LSI definition (2) to obtain

$$\begin{aligned}\frac{q^2}{2\alpha} G_{q,\nu}(\rho) &\geq q\mathbb{E}_\nu\left[\left(\frac{\rho}{\nu}\right)^q \log\frac{\rho}{\nu}\right] - F_{q,\nu}(\rho)\log F_{q,\nu}(\rho) \\ &= q\frac{\partial}{\partial q}F_{q,\nu}(\rho) - F_{q,\nu}(\rho)\log F_{q,\nu}(\rho).\end{aligned} \tag{34}$$

Therefore,

$$\begin{aligned}\frac{q^2}{2\alpha}\frac{G_{q,\nu}(\rho)}{F_{q,\nu}(\rho)} &\geq q\frac{\partial}{\partial q}\log F_{q,\nu}(\rho) - \log F_{q,\nu}(\rho) \\ &= q\frac{\partial}{\partial q}\left((q-1)R_{q,\nu}(\rho)\right) - (q-1)R_{q,\nu}(\rho) \\ &= qR_{q,\nu}(\rho) + q(q-1)\frac{\partial}{\partial q}R_{q,\nu}(\rho) - (q-1)R_{q,\nu}(\rho) \\ &= R_{q,\nu}(\rho) + q(q-1)\frac{\partial}{\partial q}R_{q,\nu}(\rho) \\ &\geq R_{q,\nu}(\rho)\end{aligned}$$

where in the last inequality we have used $q \geq 1$ and $\frac{\partial}{\partial q}R_{q,\nu}(\rho) \geq 0$ since $q \mapsto R_{q,\nu}(\rho)$ is increasing by Lemma 11. □

B.3.4 Proof of Lemma 6

Proof of Lemma 6. Let $q > 0$, $q \neq 1$. By the Fokker-Planck formula (8) and integration by parts,

$$\begin{aligned}\frac{d}{dt}F_{q,\nu}(\rho_t) &= \int_{\mathbb{R}^n} \nu\frac{\frac{\partial}{\partial t}(\rho_t^q)}{\nu^q}dx \\ &= q\int_{\mathbb{R}^n}\frac{\rho_t^{q-1}}{\nu^{q-1}}\frac{\partial\rho_t}{\partial t}dx \\ &= q\int_{\mathbb{R}^n}\left(\frac{\rho_t}{\nu}\right)^{q-1}\nabla\cdot\left(\rho_t\nabla\log\frac{\rho_t}{\nu}\right)dx \\ &= -q\int_{\mathbb{R}^n}\rho_t\left\langle\nabla\left(\frac{\rho_t}{\nu}\right)^{q-1}, \nabla\log\frac{\rho_t}{\nu}\right\rangle dx \\ &= -q(q-1)\int_{\mathbb{R}^n}\rho_t\left\langle\left(\frac{\rho_t}{\nu}\right)^{q-2}\nabla\frac{\rho_t}{\nu}, \left(\frac{\rho_t}{\nu}\right)^{-1}\nabla\frac{\rho_t}{\nu}\right\rangle dx \\ &= -q(q-1)\mathbb{E}_\nu\left[\left(\frac{\rho_t}{\nu}\right)^{q-2}\left\|\nabla\frac{\rho_t}{\nu}\right\|^2\right] \\ &= -q(q-1)G_{q,\nu}(\rho_t).\end{aligned} \tag{35}$$

Therefore,

$$\frac{d}{dt}R_{q,\nu}(\rho_t) = \frac{1}{q-1}\frac{\frac{d}{dt}F_{q,\nu}(\rho_t)}{F_{q,\nu}(\rho_t)} = -q\frac{G_{q,\nu}(\rho_t)}{F_{q,\nu}(\rho_t)}.$$

For $q = 1$, we have $R_{1,\nu}(\rho_t) = H_\nu(\rho_t)$, $G_{1,\nu}(\rho_t) = J_\nu(\rho_t)$, and $F_{1,\nu}(\rho_t) = 1$, and the claim (17) follows from Lemma 2. □

B.3.5 Proof of Theorem 3

Proof of Theorem 3. By Lemma 5 and Lemma 6,

$$\frac{d}{dt}R_{q,\nu}(\rho_t) = -q\frac{G_{q,\nu}(\rho_t)}{F_{q,\nu}(\rho_t)} \leq -\frac{2\alpha}{q}R_{q,\nu}(\rho_t).$$

Integrating gives

$$R_{q,\nu}(\rho_t) \leq e^{-\frac{2\alpha}{q}t}R_{q,\nu}(\rho_0)$$

as desired. □

B.3.6 Hypercontractivity

Lemma 12. *Suppose ν satisfies LSI with constant $\alpha > 0$. Let $q_0 > 1$, and suppose $R_{q_0,\nu}(\rho_0) < \infty$. Define $q_t = 1 + e^{2\alpha t}(q_0 - 1)$. Along the Langevin dynamics* (7)*, for all $t \geq 0$,*

$$\left(1 - \frac{1}{q_t}\right) R_{q_t,\nu}(\rho_t) \leq \left(1 - \frac{1}{q_0}\right) R_{q_0,\nu}(\rho_0). \tag{36}$$

In particular, for any $q \geq q_0$, we have $R_{q,\nu}(\rho_t) \leq R_{q_0,\nu}(\rho_0) < \infty$ for all $t \geq \frac{1}{2\alpha} \log \frac{q-1}{q_0-1}$.

Proof. We will show $\frac{d}{dt}\left\{\left(1 - \frac{1}{q_t}\right) R_{q_t,\nu}(\rho_t)\right\} \leq 0$, which implies the desired relation (36). Since $q_t = 1 + e^{2\alpha t}(q_0 - 1)$, we have $\dot{q}_t = \frac{d}{dt} q_t = 2\alpha(q_t - 1)$. Note that

$$\begin{aligned}
\frac{d}{dt} R_{q_t,\nu}(\rho_t) &= \frac{d}{dt}\left(\frac{\log F_{q_t,\nu}(\rho_t)}{q_t - 1}\right) \\
&\overset{(35)}{=} -\frac{\dot{q}_t \log F_{q_t,\nu}(\rho_t)}{(q_t-1)^2} + \frac{\dot{q}_t \mathbb{E}_\nu\left[\left(\frac{\rho_t}{\nu}\right)^{q_t} \log \frac{\rho_t}{\nu}\right] - q_t(q_t-1) G_{q_t,\nu}(\rho_t)}{(q_t-1) F_{q_t,\nu}(\rho_t)} \\
&= -2\alpha R_{q_t,\nu}(\rho_t) + 2\alpha \frac{\mathbb{E}_\nu\left[\left(\frac{\rho_t}{\nu}\right)^{q_t} \log \frac{\rho_t}{\nu}\right]}{F_{q_t,\nu}(\rho_t)} - q_t \frac{G_{q_t,\nu}(\rho_t)}{F_{q_t,\nu}(\rho_t)}.
\end{aligned}$$

In the second equality above we have used our earlier calculation (35) which holds for fixed q. Then by LSI in the form (34), we have

$$\begin{aligned}
\frac{d}{dt} R_{q_t,\nu}(\rho_t) &\leq -2\alpha R_{q_t,\nu}(\rho_t) + 2\alpha\left(\frac{q_t}{2\alpha}\frac{G_{q_t,\nu}(\rho_t)}{F_{q_t,\nu}(\rho_t)} + \frac{1}{q_t}\log F_{q_t,\nu}(\rho_t)\right) - q_t \frac{G_{q_t,\nu}(\rho_t)}{F_{q_t,\nu}(\rho_t)} \\
&= -2\alpha R_{q_t,\nu}(\rho_t) + 2\alpha\left(1 - \frac{1}{q_t}\right) R_{q_t,\nu}(\rho_t) \\
&= -\frac{2\alpha}{q_t} R_{q_t,\nu}(\rho_t).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{d}{dt}\left\{\left(1 - \frac{1}{q_t}\right) R_{q_t,\nu}(\rho_t)\right\} &= \frac{\dot{q}_t}{q_t^2} R_{q_t,\nu}(\rho_t) + \left(1 - \frac{1}{q_t}\right)\frac{d}{dt} R_{q_t,\nu}(\rho_t) \\
&\leq \frac{2\alpha(q_t - 1)}{q_t^2} R_{q_t,\nu}(\rho_t) - \left(1 - \frac{1}{q_t}\right)\frac{2\alpha}{q_t} R_{q_t,\nu}(\rho_t) \\
&= 0,
\end{aligned}$$

as desired.

Now given $q \geq q_0$, let $t_0 = \frac{1}{2\alpha}\log\frac{q-1}{q_0-1}$ so $q_{t_0} = q$. Then $R_{q,\nu}(\rho_{t_0}) \leq \frac{q}{(q-1)}\frac{(q_0-1)}{q_0} R_{q_0,\nu}(\rho_0) \leq R_{q_0,\nu}(\rho_0) < \infty$. For $t > t_0$, by applying Theorem 3 starting from ρ_{t_0}, we obtain $R_{q,\nu}(\rho_t) \leq e^{-\frac{2\alpha}{q}(t-t_0)} R_{q,\nu}(\rho_{t_0}) \leq R_{q,\nu}(\rho_{t_0}) \leq R_{q_0,\nu}(\rho_0) < \infty$. □

By combining Theorem 3 and Lemma 12, we obtain the following characterization of the behavior of Renyi divergence along the Langevin dynamics under LSI.

Corollary 2. *Suppose ν satisfies LSI with constant $\alpha > 0$. Suppose ρ_0 satisfies $R_{q_0,\nu}(\rho_0) < \infty$ for some $q_0 > 1$. Along the Langevin dynamics* (7)*, for all $q \geq q_0$ and $t \geq t_0 := \frac{1}{2\alpha}\log\frac{q-1}{q_0-1}$,*

$$R_{q,\nu}(\rho_t) \leq e^{-\frac{2\alpha}{q}(t-t_0)} R_{q_0,\nu}(\rho_0). \tag{37}$$

Proof. By Lemma 12, at $t = t_0$ we have $R_{q,\nu}(\rho_{t_0}) \leq R_{q_0,\nu}(\rho_0)$. For $t > t_0$, by applying Theorem 3 starting from ρ_{t_0}, we have $R_{q,\nu}(\rho_t) \leq e^{-\frac{2\alpha}{q}(t-t_0)} R_{q,\nu}(\rho_{t_0}) \leq e^{-\frac{2\alpha}{q}(t-t_0)} R_{q_0,\nu}(\rho_0)$. □

B.4 Proofs for §5: Rényi divergence along ULA

B.4.1 Proof of Lemma 7

Proof of Lemma 7. By Cauchy-Schwarz inequality,

$$\begin{aligned}
F_{q,\nu}(\rho) &= \int \frac{\rho^q}{\nu^{q-1}}\,dx \\
&= \int \nu_\epsilon \left(\frac{\rho}{\nu_\epsilon}\right)^q \left(\frac{\nu_\epsilon}{\nu}\right)^{q-1} dx \\
&\le \left(\int \nu_\epsilon \left(\frac{\rho}{\nu_\epsilon}\right)^{2q} dx\right)^{\frac{1}{2}} \left(\int \nu_\epsilon \left(\frac{\nu_\epsilon}{\nu}\right)^{2(q-1)} dx\right)^{\frac{1}{2}} \\
&= F_{2q,\nu_\epsilon}(\rho)^{\frac{1}{2}} F_{2q-1,\nu}(\nu_\epsilon)^{\frac{1}{2}}.
\end{aligned}$$

Taking logarithm gives

$$(q-1)R_{q,\nu}(\rho) \le \frac{(2q-1)}{2} R_{2q,\nu_\epsilon}(\rho) + \frac{(2q-2)}{2} R_{2q-1,\nu}(\nu_\epsilon).$$

Dividing both sides by $q-1>0$ gives the desired inequality (18). □

B.4.2 Proof of Lemma 8

We will use the following auxiliary results. Recall that given a map $T\colon \mathbb{R}^n \to \mathbb{R}^n$ and a probability distribution ρ, the pushforward $T_\#\rho$ is the distribution of $T(x)$ when $x \sim \rho$.

Lemma 13. *Let $T\colon \mathbb{R}^n \to \mathbb{R}^n$ be a differentiable bijective map. For any probability distributions ρ, ν, and for all $q > 0$,*

$$R_{q,T_\#\nu}(T_\#\rho) = R_{q,\nu}(\rho).$$

Proof. Let $\tilde\rho = T_\#\rho$ and $\tilde\nu = T_\#\nu$. By the change of variable formula,

$$\begin{aligned}
\rho(x) &= \det(\nabla T(x))\, \tilde\rho(T(x)), \\
\nu(x) &= \det(\nabla T(x))\, \tilde\nu(T(x)).
\end{aligned}$$

Since T is differentiable and bijective, $\det(\nabla T(x)) \neq 0$. Therefore,

$$\frac{\tilde\rho(T(x))}{\tilde\nu(T(x))} = \frac{\rho(x)}{\nu(x)}.$$

Now let $X \sim \nu$, so $T(X) \sim \tilde\nu$. Then for all $q > 0$.

$$F_{q,\tilde\nu}(\tilde\rho) = \mathbb{E}_{\tilde\nu}\left[\left(\frac{\tilde\rho}{\tilde\nu}\right)^q\right] = \mathbb{E}_{X\sim\nu}\left[\left(\frac{\tilde\rho(T(X))}{\tilde\nu(T(X))}\right)^q\right] = \mathbb{E}_{X\sim\nu}\left[\left(\frac{\rho(X)}{\nu(X)}\right)^q\right] = F_{q,\nu}(\rho).$$

Suppose $q \neq 1$. Taking logarithm on both sides and dividing by $q-1 \neq 0$ yields $R_{q,\tilde\nu}(\tilde\rho) = R_{q,\nu}(\rho)$, as desired. The case $q = 1$ follows from taking limit $q \to 1$, or by an analogous direct argument:

$$H_{\tilde\nu}(\tilde\rho) = \mathbb{E}_{\tilde\nu}\left[\frac{\tilde\rho}{\tilde\nu}\log\frac{\tilde\rho}{\tilde\nu}\right] = \mathbb{E}_{X\sim\nu}\left[\frac{\tilde\rho(T(X))}{\tilde\nu(T(X))}\log\frac{\tilde\rho(T(X))}{\tilde\nu(T(X))}\right] = \mathbb{E}_{X\sim\nu}\left[\frac{\rho(X)}{\nu(X)}\log\frac{\rho(X)}{\nu(X)}\right] = H_\nu(\rho).$$

□

We have the following result on how the LSI constant changes under a Lipschitz mapping. We recall that $T\colon \mathbb{R}^n \to \mathbb{R}^n$ is L-Lipschitz if $\|T(x) - T(y)\| \le L\|x - y\|$ for all $x, y \in \mathbb{R}^n$.

Lemma 14. *Suppose a probability distribution ν satisfies LSI with constant $\alpha > 0$. Let $T\colon \mathbb{R}^n \to \mathbb{R}^n$ be a differentiable L-Lipschitz map. Then $\tilde\nu = T_\#\nu$ satisfies LSI with constant α/L^2.*

Proof. Let $g\colon \mathbb{R}^n \to \mathbb{R}$ be a smooth function, and let $\tilde{g}\colon \mathbb{R}^n \to \mathbb{R}$ be the function $\tilde{g}(x) = g(T(x))$. Let $X \sim \nu$, so $T(X) \sim \tilde{\nu}$. Note that

$$\begin{aligned}\mathbb{E}_{\tilde{\nu}}[g^2] &= \mathbb{E}_{X\sim\nu}[g(T(X))^2] = \mathbb{E}_{\nu}[\tilde{g}^2],\\ \mathbb{E}_{\tilde{\nu}}[g^2 \log g^2] &= \mathbb{E}_{X\sim\nu}[g(T(X))^2 \log g(T(X))^2] = \mathbb{E}_{\nu}[\tilde{g}^2 \log \tilde{g}^2].\end{aligned}$$

Furthermore, we have $\nabla\tilde{g}(x) = \nabla T(x)\, \nabla g(T(x))$. Since T is L-Lipschitz, $\|\nabla T(x)\| \le L$. Then

$$\|\nabla\tilde{g}(x)\| \le \|\nabla T(x)\|\, \|\nabla g(T(x))\| \le L\|\nabla g(T(x))\|.$$

This implies

$$\mathbb{E}_{\tilde{\nu}}[\|\nabla g\|^2] = \mathbb{E}_{X\sim\nu}[\|\nabla g(T(X))\|^2] \ge \frac{\mathbb{E}_{\nu}[\|\nabla\tilde{g}\|^2]}{L^2}.$$

Therefore,

$$\frac{\mathbb{E}_{\tilde{\nu}}[\|\nabla g\|^2]}{\mathbb{E}_{\tilde{\nu}}[g^2 \log g^2] - \mathbb{E}_{\tilde{\nu}}[g^2] \log \mathbb{E}_{\tilde{\nu}}[g^2]} \ge \frac{1}{L^2}\, \frac{\mathbb{E}_{\nu}[\|\nabla\tilde{g}\|^2]}{\left(\mathbb{E}_{\nu}[\tilde{g}^2 \log \tilde{g}^2] - \mathbb{E}_{\nu}[\tilde{g}^2] \log \mathbb{E}_{\nu}[\tilde{g}^2]\right)} \ge \frac{\alpha}{2L^2}$$

where the last inequality follows from the assumption that ν satisfies LSI with constant α. This shows that $\tilde{\nu}$ satisfies LSI with constant α/L^2, as desired. □

We also recall the following result on how the LSI constant changes along Gaussian convolution.

Lemma 15. *Suppose a probability distribution ν satisfies LSI with constant $\alpha > 0$. For $t > 0$, the probability distribution $\tilde{\nu}_t = \nu * \mathcal{N}(0,\, 2tI)$ satisfies LSI with constant $\left(\frac{1}{\alpha} + 2t\right)^{-1}$.*

Proof. We recall the following convolution property of LSI [14]: If $\nu, \tilde{\nu}$ satisfy LSI with constants $\alpha, \tilde{\alpha} > 0$, respectively, then $\nu * \tilde{\nu}$ satisfies LSI with constant $\left(\frac{1}{\alpha} + \frac{1}{\tilde{\alpha}}\right)^{-1}$. Since $\mathcal{N}(0, 2tI)$ satisfies LSI with constant $\frac{1}{2t}$, the claim above follows. □

We now derive a formula for the decrease of Rényi divergence along simultaneous heat flow. We note the resulting formula (39) is similar to the formula (17) for the decrease of Rényi divergence along the Langevin dynamics.

Lemma 16. *For any probability distributions ρ_0, ν_0, and for any $t \ge 0$, let $\rho_t = \rho_0 * \mathcal{N}(0, 2tI)$ and $\nu_t = \nu_0 * \mathcal{N}(0, 2tI)$. Then for all $q > 0$,*

$$\frac{d}{dt} R_{q,\nu_t}(\rho_t) = -q \frac{G_{q,\nu_t}(\rho_t)}{F_{q,\nu_t}(\rho_t)}. \tag{39}$$

Proof. By definition, ρ_t and ν_t evolve following the simultaneous heat flow:

$$\frac{\partial \rho_t}{\partial t} = \Delta\rho_t, \qquad \frac{\partial \nu_t}{\partial t} = \Delta\nu_t. \tag{40}$$

We will use the following identity for any smooth function $h\colon \mathbb{R}^n \to \mathbb{R}$,

$$\Delta(h^q) = \nabla \cdot \left(q h^{q-1} \nabla h\right) = q(q-1) h^{q-2} \|\nabla h\|^2 + q h^{q-1} \Delta h.$$

We will also use the integration by parts formula (23). Then along the simultaneous heat flow (40),

$$\begin{aligned}
\frac{d}{dt}F_{q,\nu_t}(\rho_t) &= \frac{d}{dt}\int \frac{\rho_t^q}{\nu_t^{q-1}}\,dx \\
&= \int q\left(\frac{\rho_t}{\nu_t}\right)^{q-1}\frac{\partial\rho_t}{\partial t}\,dx - \int (q-1)\left(\frac{\rho_t}{\nu_t}\right)^{q}\frac{\partial\nu_t}{\partial t}\,dx \\
&= q\int \left(\frac{\rho_t}{\nu_t}\right)^{q-1}\Delta\rho_t\,dx - (q-1)\int\left(\frac{\rho_t}{\nu_t}\right)^{q}\Delta\nu_t\,dx \\
&= q\int \Delta\left(\left(\frac{\rho_t}{\nu_t}\right)^{q-1}\right)\rho_t\,dx - (q-1)\int\Delta\left(\left(\frac{\rho_t}{\nu_t}\right)^{q}\right)\nu_t\,dx \\
&= q\int\left((q-1)(q-2)\left(\frac{\rho_t}{\nu_t}\right)^{q-3}\left\|\nabla\frac{\rho_t}{\nu_t}\right\|^2 + (q-1)\left(\frac{\rho_t}{\nu_t}\right)^{q-2}\Delta\frac{\rho_t}{\nu_t}\right)\rho_t\,dx \\
&\quad - (q-1)\int\left(q(q-1)\left(\frac{\rho_t}{\nu_t}\right)^{q-2}\left\|\nabla\frac{\rho_t}{\nu_t}\right\|^2 + q\left(\frac{\rho_t}{\nu_t}\right)^{q-1}\Delta\frac{\rho_t}{\nu_t}\right)\nu_t\,dx \\
&= -q(q-1)\int\nu_t\left(\frac{\rho_t}{\nu_t}\right)^{q-2}\left\|\nabla\frac{\rho_t}{\nu_t}\right\|^2\,dx \\
&= -q(q-1)G_{q,\nu_t}(\rho_t). \qquad (41)
\end{aligned}$$

Note that the identity (41) above is analogous to the identity (35) along the Langevin dynamics. Therefore, for $q \neq 1$,

$$\frac{d}{dt}R_{q,\nu_t}\rho_t = \frac{1}{q-1}\frac{\frac{d}{dt}F_{q,\nu_t}(\rho_t)}{F_{q,\nu_t}(\rho_t)} = -q\frac{G_{q,\nu_t}(\rho_t)}{F_{q,\nu_t}(\rho_t)},$$

as desired.

The case $q = 1$ follows from taking limit $q \to 1$, or by an analogous direct calculation. We will use the following identity for $h\colon \mathbb{R}^n \to \mathbb{R}_{>0}$,

$$\Delta\log h = \nabla\cdot\left(\frac{\nabla h}{h}\right) = \frac{\Delta h}{h} - \|\nabla\log h\|^2.$$

Then along the simultaneous heat flow (40),

$$\begin{aligned}
\frac{d}{dt}H_{\nu_t}(\rho_t) &= \frac{d}{dt}\int\rho_t\log\frac{\rho_t}{\nu_t}\,dx \\
&= \int\frac{\partial\rho_t}{\partial t}\log\frac{\rho_t}{\nu_t}\,dx + \int\rho_t\frac{\nu_t}{\rho_t}\frac{\partial}{\partial t}\left(\frac{\rho_t}{\nu_t}\right)\,dx \\
&= \int\Delta\rho_t\,\log\frac{\rho_t}{\nu_t}\,dx + \int\nu_t\left(\frac{1}{\nu_t}\frac{\partial\rho_t}{\partial t}\,dx - \frac{\rho_t}{\nu_t^2}\frac{\partial\nu_t}{\partial t}\right)\,dx \\
&= \int\rho_t\,\Delta\log\frac{\rho_t}{\nu_t}\,dx - \int\frac{\rho_t}{\nu_t}\Delta\nu_t\,dx \\
&= \int\rho_t\left(\frac{\nu_t}{\rho_t}\Delta\left(\frac{\rho_t}{\nu_t}\right) - \left\|\nabla\log\frac{\rho_t}{\nu_t}\right\|^2\right)\,dx - \int\frac{\rho_t}{\nu_t}\Delta\nu_t\,dx \\
&= -J_{\nu_t}(\rho_t),
\end{aligned}$$

as desired. Note that this is also analogous to the identity (9) along the Langevin dynamics. □

We are now ready to prove Lemma 8. See Figure 3 for an illustration.

Proof of Lemma 8. We will prove that along each step of ULA (11) from $x_k \sim \rho_k$ to $x_{k+1} \sim \rho_{k+1}$, the Rényi divergence with respect to ν_ϵ decreases by a constant factor:

$$R_{q,\nu_\epsilon}(\rho_{k+1}) \le e^{-\frac{\beta\epsilon}{q}}R_{q,\nu_\epsilon}(\rho_k). \qquad (42)$$

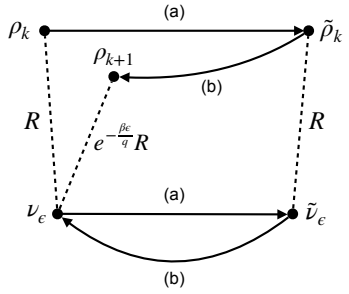

Figure 3: An illustration for the proof of Lemma 8. We decompose each step of ULA into two operations: (a) a deterministic gradient step, and (b) an evolution along the heat flow. If the current Rényi divergence is $R \equiv R_{q,\nu_\epsilon}(\rho_k)$, then the gradient step (a) does not change the Rényi divergence: $R_{q,\tilde{\nu}_\epsilon}(\tilde{\rho}_k) = R$, while the heat flow (b) decreases the Rényi divergence: $R_{q,\nu_\epsilon}(\rho_{k+1}) \le e^{-\alpha\epsilon}R$.

Iterating the bound above yields the desired claim (19).

We decompose each step of ULA (11) into a sequence of two steps:

$$\tilde{\rho}_k = (I - \epsilon\nabla f)_{\#}\rho_k, \tag{43a}$$

$$\rho_{k+1} = \tilde{\rho}_k * \mathcal{N}(0, 2\epsilon I). \tag{43b}$$

In the first step (43a), we apply a smooth deterministic map $T(x) = x - \epsilon\nabla f(x)$. Since ∇f is L-Lipschitz and $\epsilon < \frac{1}{L}$, T is a bijection. Then by Lemma 13,

$$R_{q,\nu_\epsilon}(\rho_k) = R_{q,\tilde{\nu}_\epsilon}(\tilde{\rho}_k) \tag{44}$$

where $\tilde{\nu}_\epsilon = (I - \epsilon\nabla f)_{\#}\nu_\epsilon$. Recall by Assumption 1 that ν_ϵ satisfies LSI with constant β. Since the map $T(x) = x - \epsilon\nabla f(x)$ is $(1+\epsilon L)$-Lipschitz, by Lemma 14 we know that $\tilde{\nu}_\epsilon$ satisfies LSI with constant $\frac{\beta}{(1+\epsilon L)^2}$.

In the second step (43b), we convolve with a Gaussian distribution, which is the result of evolving along the heat flow at time ϵ. For $0 \le t \le \epsilon$, let $\tilde{\rho}_{k,t} = \tilde{\rho}_k * \mathcal{N}(0, 2tI)$ and $\tilde{\nu}_{\epsilon,t} = \tilde{\nu}_\epsilon * \mathcal{N}(0, 2tI)$, so $\tilde{\rho}_{k,\epsilon} = \tilde{\rho}_{k+1}$ and $\tilde{\nu}_{\epsilon,\epsilon} = \nu_\epsilon$. By Lemma 16,

$$\frac{d}{dt}R_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{k,t}) = -q\frac{G_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{k,t})}{F_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{k,t})}.$$

Since $\tilde{\nu}_\epsilon$ satisfies LSI with constant $\frac{\beta}{(1+\epsilon L)^2}$, by Lemma 15 we know that $\tilde{\nu}_{\epsilon,t}$ satisfies LSI with constant $\left(\frac{(1+\epsilon L)^2}{\beta} + 2t\right)^{-1} \ge \left(\frac{(1+\epsilon L)^2}{\beta} + 2\epsilon\right)^{-1}$ for $0 \le t \le \epsilon$. In particular, since $\epsilon \le \min\{\frac{1}{3L}, \frac{1}{9\beta}\}$, the LSI constant is $\left(\frac{(1+\epsilon L)^2}{\beta} + 2\epsilon\right)^{-1} \ge \left(\frac{16}{9\beta} + \frac{2}{9\beta}\right)^{-1} = \frac{\beta}{2}$. Then by Lemma 5,

$$\frac{d}{dt}R_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{k,t}) = -q\frac{G_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{k,t})}{F_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{k,t})} \le -\frac{\beta}{q}R_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{\epsilon,t}).$$

Integrating over $0 \le t \le \epsilon$ gives

$$R_{q,\nu_\epsilon}(\rho_{k+1}) = R_{q,\tilde{\nu}_{\epsilon,\epsilon}}(\tilde{\rho}_{k,\epsilon}) \le e^{-\frac{\beta\epsilon}{q}}R_{q,\tilde{\nu}_\epsilon}(\tilde{\rho}_k). \tag{45}$$

Combining (44) and (45) gives the desired inequality (42). □

B.4.3 Proof of Theorem 4

Proof of Theorem 4. This follows directly from Lemma 7 and Lemma 8, and using the definition of the growth function $R_{2q-1,\nu}(\nu_\epsilon) \le g_{2q-1}(\epsilon)$. □

B.5 Details for §6: Poincaré inequality

B.5.1 A bound on Rényi information

Lemma 17. *Suppose ν satisfies Poincaré inequality with constant $\alpha > 0$. Let $q \geq 2$. For all ρ,*

$$\frac{G_{q,\nu}(\rho)}{F_{q,\nu}(\rho)} \geq \frac{4\alpha}{q^2}\left(1 - e^{-R_{q,\nu}(\rho)}\right).$$

Proof. We plug in $g^2 = \left(\frac{\rho}{\nu}\right)^q$ to Poincaré inequality (21) and use the monotonicity condition from Lemma 11 to obtain

$$\begin{aligned}
\frac{q^2}{4\alpha} G_{q,\nu}(\rho) &\geq F_{q,\nu}(\rho) - F_{\frac{q}{2},\nu}(\rho)^2 \\
&= e^{(q-1)R_{q,\nu}(\rho)} - e^{(q-2)R_{\frac{q}{2},\nu}(\rho)} \\
&\geq e^{(q-1)R_{q,\nu}(\rho)} - e^{(q-2)R_{q,\nu}(\rho)} \\
&= F_{q,\nu}(\rho)\left(1 - e^{-R_{q,\nu}(\rho)}\right).
\end{aligned}$$

Dividing both sides by $F_{q,\nu}(\rho)$ and rearranging yields the desired inequality. □

B.5.2 Proof of Theorem 5

Proof of Theorem 5. By Lemma 6 and Lemma 17,

$$\frac{d}{dt} R_{q,\nu}(\rho_t) = -q\frac{G_{q,\nu}(\rho_t)}{F_{q,\nu}(\rho_t)} \leq -\frac{4\alpha}{q}\left(1 - e^{-R_{q,\nu}(\rho_t)}\right).$$

We now consider two possibilities:

1. If $R_{q,\nu}(\rho_0) \geq 1$, then as long as $R_{q,\nu}(\rho_t) \geq 1$, we have $1 - e^{-R_{q,\nu}(\rho_t)} \geq 1 - e^{-1} > \frac{1}{2}$, so $\frac{d}{dt} R_{q,\nu}(\rho_t) \leq -\frac{2\alpha}{q}$, which implies $R_{q,\nu}(\rho_t) \leq R_{q,\nu}(\rho_0) - \frac{2\alpha t}{q}$.
2. If $R_{q,\nu}(\rho_0) \leq 1$, then $R_{q,\nu}(\rho_t) \leq 1$, and thus $\frac{1-e^{-R_{q,\nu}(\rho_t)}}{R_{q,\nu}(\rho_t)} \geq \frac{1}{1+R_{q,\nu}(\rho_t)} \geq \frac{1}{2}$. Thus, in this case $\frac{d}{dt} R_{q,\nu}(\rho_t) \leq -\frac{2\alpha}{q} R_{q,\nu}(\rho_t)$, and integrating gives $R_{q,\nu}(\rho_t) \leq e^{-\frac{2\alpha t}{q}} R_{q,\nu}(\rho_0)$, as desired.

□

B.5.3 Convergence of Rényi divergence to the biased limit along ULA under Poincaré

We show Rényi divergence to the biased limit converges along ULA under Poincaré inequality. Thus, starting from $R_{q,\nu_\epsilon}(\rho_0) \geq 1$, ULA reaches $R_{q,\nu_\epsilon}(\rho_k) \leq \delta$ in $k = \tilde{O}\left(\frac{q}{\epsilon\beta} R_{q,\nu_\epsilon}(\rho_0)\right)$ iterations.

Lemma 18. *Assume Assumption 2. Suppose $\nu = e^{-f}$ is L-smooth, and let $0 < \epsilon \leq \min\left\{\frac{1}{3L}, \frac{1}{9\beta}\right\}$. For $q \geq 2$, along ULA (11),*

$$R_{q,\nu_\epsilon}(\rho_k) \leq \begin{cases} R_{q,\nu_\epsilon}(\rho_0) - \frac{\beta\epsilon k}{q} & \text{if } R_{q,\nu_\epsilon}(\rho_0) \geq 1 \text{ and as long as } R_{q,\nu_\epsilon}(\rho_k) \geq 1, \\ e^{-\frac{\beta\epsilon k}{q}} R_{q,\nu_\epsilon}(\rho_0) & \text{if } R_{q,\nu_\epsilon}(\rho_0) \leq 1. \end{cases} \tag{46}$$

Proof. Following the proof of Lemma 8, we decompose each step of ULA (11) into two steps:

$$\tilde{\rho}_k = (I - \epsilon\nabla f)_{\#}\rho_k, \tag{47a}$$

$$\rho_{k+1} = \tilde{\rho}_k * \mathcal{N}(0, 2\epsilon I). \tag{47b}$$

The first step (47a) is a deterministic bijective map, so it preserves Rényi divergence by Lemma 13: $R_{q,\nu_\epsilon}(\rho_k) = R_{q,\tilde{\nu}_\epsilon}(\tilde{\rho}_k)$, where $\tilde{\nu}_\epsilon = (I - \epsilon\nabla f)_{\#}\nu_\epsilon$. Recall by Assumption 2 that ν_ϵ satisfies Poincaré inequality with constant β. Since the map $T(x) = x - \epsilon\nabla f(x)$ is $(1 + \epsilon L)$-Lipschitz, by Lemma 19 we know that $\tilde{\nu}_\epsilon$ satisfies Poincaré inequality with constant $\frac{\beta}{(1+\epsilon L)^2}$.

The second step (47b) is convolution with a Gaussian distribution, which is the result of evolving along the heat flow at time ϵ. For $0 \le t \le \epsilon$, let $\tilde{\rho}_{k,t} = \tilde{\rho}_k * \mathcal{N}(0, 2tI)$ and $\tilde{\nu}_{\epsilon,t} = \tilde{\nu}_\epsilon * \mathcal{N}(0, 2tI)$, so $\tilde{\rho}_{k,\epsilon} = \tilde{\rho}_{k+1}$ and $\tilde{\nu}_{\epsilon,\epsilon} = \nu_\epsilon$. Since $\tilde{\nu}_\epsilon$ satisfies Poincaré inequality with constant $\frac{\beta}{(1+\epsilon L)^2}$, by Lemma 20 we know that $\tilde{\nu}_{\epsilon,t}$ satisfies Poincaré inequality with constant $\left(\frac{(1+\epsilon L)^2}{\beta} + 2t\right)^{-1} \ge \left(\frac{(1+\epsilon L)^2}{\beta} + 2\epsilon\right)^{-1}$ for $0 \le t \le \epsilon$. In particular, since $\epsilon \le \min\{\frac{1}{3L}, \frac{1}{9\beta}\}$, the Poincaré constant is $\left(\frac{(1+\epsilon L)^2}{\beta} + 2\epsilon\right)^{-1} \ge \left(\frac{16}{9\beta} + \frac{2}{9\beta}\right)^{-1} = \frac{\beta}{2}$. Then by Lemma 16 and Lemma 17,

$$\frac{d}{dt} R_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{k,t}) = -q \frac{G_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{k,t})}{F_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{k,t})} \le -\frac{2\beta}{q}\left(1 - e^{-R_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{k,t})}\right).$$

We now consider two possibilities, as in Theorem 5:

1. If $R_{q,\nu_\epsilon}(\rho_k) = R_{q,\tilde{\nu}_{\epsilon,0}}(\tilde{\rho}_{k,0}) \ge 1$, then as long as $R_{q,\nu_\epsilon}(\rho_{k+1}) = R_{q,\tilde{\nu}_{\epsilon,\epsilon}}(\tilde{\rho}_{k,\epsilon}) \ge 1$, we have $1 - e^{-R_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{k,t})} \ge 1 - e^{-1} > \frac{1}{2}$, so $\frac{d}{dt} R_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{k,t}) \le -\frac{\beta}{q}$, which implies $R_{q,\nu_\epsilon}(\rho_{k+1}) \le R_{q,\nu_\epsilon}(\rho_k) - \frac{\beta\epsilon}{q}$. Iterating this step, we have that $R_{q,\nu_\epsilon}(\rho_k) \le R_{q,\nu_\epsilon}(\rho_0) - \frac{\beta\epsilon k}{q}$ if $R_{q,\nu_\epsilon}(\rho_0) \ge 1$ and as long as $R_{q,\nu_\epsilon}(\rho_k) \ge 1$.

2. If $R_{q,\nu_\epsilon}(\rho_k) = R_{q,\tilde{\nu}_{\epsilon,0}}(\tilde{\rho}_{k,0}) \le 1$, then $R_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{k,t}) \le 1$, and thus $\frac{1-e^{-R_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{k,t})}}{R_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{k,t})} \ge \frac{1}{1+R_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{k,t})} \ge \frac{1}{2}$. Thus, in this case $\frac{d}{dt} R_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{k,t}) \le -\frac{\beta}{q} R_{q,\tilde{\nu}_{\epsilon,t}}(\tilde{\rho}_{k,t})$. Integrating over $0 \le t \le \epsilon$ gives $R_{q,\nu_\epsilon}(\rho_{k+1}) = R_{q,\tilde{\nu}_{\epsilon,\epsilon}}(\tilde{\rho}_{k,\epsilon}) \le e^{-\frac{\beta\epsilon}{q}} R_{q,\tilde{\nu}_{\epsilon,0}}(\tilde{\rho}_{k,0}) = e^{-\frac{\beta\epsilon}{q}} R_{q,\nu_\epsilon}(\rho_k)$. Iterating this step gives $R_{q,\nu_\epsilon}(\rho_k) \le e^{-\frac{\beta\epsilon k}{q}} R_{q,\nu_\epsilon}(\rho_0)$ if $R_{q,\nu_\epsilon}(\rho_0) \le 1$, as desired.

□

In the proof above we use the following results, which are analogous to Lemma 14 and Lemma 15.

Lemma 19. *Suppose a probability distribution ν satisfies Poincaré inequality with constant $\alpha > 0$. Let $T\colon \mathbb{R}^n \to \mathbb{R}^n$ be a differentiable L-Lipschitz map. Then $\tilde{\nu} = T_\# \nu$ satisfies Poincaré inequality with constant α/L^2.*

Proof. Let $g\colon \mathbb{R}^n \to \mathbb{R}$ be a smooth function, and let $\tilde{g}\colon \mathbb{R}^n \to \mathbb{R}$ be the function $\tilde{g}(x) = g(T(x))$. Let $X \sim \nu$, so $T(X) \sim \tilde{\nu}$. Note that

$$\mathrm{Var}_{\tilde{\nu}}(g) = \mathrm{Var}_{X\sim\nu}(g(T(X))) = \mathrm{Var}_\nu(\tilde{g}).$$

Furthermore, we have $\nabla\tilde{g}(x) = \nabla T(x)\, \nabla g(T(x))$. Since T is L-Lipschitz, $\|\nabla T(x)\| \le L$. Then

$$\|\nabla\tilde{g}(x)\| \le \|\nabla T(x)\|\, \|\nabla g(T(x))\| \le L\|\nabla g(T(x))\|.$$

This implies

$$\mathbb{E}_{\tilde{\nu}}[\|\nabla g\|^2] = \mathbb{E}_{X\sim\nu}[\|\nabla g(T(X))\|^2] \ge \frac{\mathbb{E}_\nu[\|\nabla\tilde{g}\|^2]}{L^2}.$$

Therefore,

$$\frac{\mathbb{E}_{\tilde{\nu}}[\|\nabla g\|^2]}{\mathrm{Var}_{\tilde{\nu}}(g)} \ge \frac{1}{L^2} \frac{\mathbb{E}_\nu[\|\nabla\tilde{g}\|^2]}{\mathrm{Var}_\nu(\tilde{g})} \ge \frac{\alpha}{L^2}$$

where the last inequality follows from the assumption that ν satisfies Poincaré inequality with constant α. This shows that $\tilde{\nu}$ satisfies Poincaré inequality with constant α/L^2, as desired. □

Lemma 20. *Suppose a probability distribution ν satisfies Poincaré inequality with constant $\alpha > 0$. For $t > 0$, the probability distribution $\tilde{\nu}_t = \nu * \mathcal{N}(0, 2tI)$ satisfies Poincaré inequality with constant $\left(\frac{1}{\alpha} + 2t\right)^{-1}$.*

Proof. We recall the following convolution property of Poincaré inequality [19]: If $\nu, \tilde{\nu}$ satisfy Poincaré inequality with constants $\alpha, \tilde{\alpha} > 0$, respectively, then $\nu * \tilde{\nu}$ satisfies Poincaré inequality with constant $\left(\frac{1}{\alpha} + \frac{1}{\tilde{\alpha}}\right)^{-1}$. Since $\mathcal{N}(0, 2tI)$ satisfies Poincaré inequality with constant $\frac{1}{2t}$, the claim above follows. □

B.5.4 Proof of Theorem 6

Proof of Theorem 6. By Lemma 18 (which applies since $2q > 2$), after k_0 iterations we have $R_{2q,\nu_\epsilon}(\rho_{k_0}) \le 1$. Applying the second case of Lemma 18 starting from k_0 gives $R_{2q,\nu_\epsilon}(\rho_k) \le e^{-\frac{\beta\epsilon(k-k_0)}{2q}} R_{2q,\nu_\epsilon}(\rho_{k_0}) \le e^{-\frac{\beta\epsilon(k-k_0)}{2q}}$. Then by Lemma 7 and recalling the definition of the growth function,

$$R_{q,\nu}(\rho_k) \le \left(\frac{q-\frac{1}{2}}{q-1}\right) R_{2q,\nu_\epsilon}(\rho_k) + R_{2q-1,\nu}(\nu_\epsilon) \le \left(\frac{q-\frac{1}{2}}{q-1}\right) e^{-\frac{\beta\epsilon(k-k_0)}{2q}} + g_{2q-1}(\epsilon)$$

as desired. □

B.5.5 Iteration complexity of ULA under Poincaré

By Theorem 6, to achieve $R_{q,\nu}(\rho_k) \le \delta$, it suffices to run ULA with $\epsilon = \Theta\left(\min\left\{\frac{1}{L}, g_{2q-1}^{-1}\left(\frac{\delta}{2}\right)\right\}\right)$ for $k = O\left(\frac{1}{\beta\epsilon}(R_{2q,\nu_\epsilon}(\rho_0) + \log\frac{1}{\delta})\right)$ iterations, where $g_q^{-1}(\delta) = \sup\{\epsilon > 0 \colon g_q(\epsilon) \le \delta\}$. Suppose δ is small so $g_{2q-1}^{-1}(\frac{\delta}{2}) < \frac{1}{L}$. Since ν_ϵ is $\frac{1}{2\epsilon}$-smooth, we can choose ρ_0 to be a Gaussian with covariance $2\epsilon I$, so $R_{2q,\nu_\epsilon}(\rho_0) = \tilde{O}(n)$ by Lemma 4. Then Theorem 6 yields an iteration complexity of $k = \tilde{O}\Big(\frac{n}{\beta g_{2q-1}^{-1}(\delta/2)}\Big)$. Note the additional dependence on dimension n compared to the LSI case in Section 5.3.

For example, if $g_q(\epsilon) = O(\epsilon)$, then $g_q^{-1}(\delta) = \Omega(\delta)$, so the iteration complexity is $k = \tilde{O}\big(\frac{n}{\beta\delta}\big)$ with $\epsilon = \Theta(\delta)$. If $g_q(\epsilon) = O(\epsilon^2)$, then $g_q^{-1}(\delta) = \Omega(\sqrt{\delta})$, so the iteration complexity is $k = \tilde{O}\big(\frac{n}{\beta\sqrt{\delta}}\big)$ with $\epsilon = \Theta(\sqrt{\delta})$.