Characterizing Dependence of Samples Along the Langevin Dynamics and Algorithms via Contraction of Φ-Mutual Information

Jiaming Liang, Siddharth Mitra, and Andre Wibisono

***Abstract*—The mixing time of a Markov chain determines how fast the iterates of the Markov chain converge to the stationary distribution; however, it does not control the dependencies between samples along the Markov chain. In this paper, we study the question of how fast the samples become approximately independent along popular Markov chains for continuous-space sampling: the Langevin dynamics in continuous time, and the Unadjusted Langevin Algorithm and the Proximal Sampler in discrete time. We measure the dependence between samples via Φ-mutual information, which is a broad generalization of the standard mutual information, and which is equal to 0 if and only if the samples are independent. We show that along these Markov chains, the Φ-mutual information between the first and the k-th iterate decreases to 0 exponentially fast in k when the target distribution is strongly log-concave. Our proof technique is based on showing the Strong Data Processing Inequalities (SDPIs) hold along the Markov chains. To prove fast mixing of the Markov chains, we only need to show the SDPIs hold for the stationary distribution. In contrast, to prove the contraction of Φ-mutual information, we need to show the SDPIs hold along the entire trajectories of the Markov chains; we prove this when the iterates along the Markov chains satisfy the corresponding Φ-Sobolev inequality, which is implied by the strong log-concavity of the target distribution.**

***Index Terms*—Markov chain, Langevin dynamics, unadjusted Langevin algorithm, proximal sampler, Φ-mutual information, Φ-Sobolev inequality, strong data processing inequality.**

I. Introduction

CONSIDER the task of sampling from a target probability distribution $\nu \propto \exp(-f)$ supported on $\mathbb{R}^d$. This is a fundamental algorithmic question appearing in many fields including machine learning, statistics, and Bayesian inference [1], [2], [3]. In settings where exact sampling from ν is not possible, a common approach is to construct a Markov chain with ν as its stationary distribution, and output the samples after an initial waiting ("burn-in") period when the chain has approximately mixed. When implementing a Markov chain to draw samples, there are many considerations to make in order to obtain strong statistical guarantees, for instance, how many chains to run, how long to wait before outputting a sample, and where to start the chains from [4] and [5].

The mixing time of a Markov chain tracks how fast the iterate along the Markov chain converges to the stationary distribution, and thus it controls the burn-in period, i.e., how long to wait before obtaining a useful sample [6], [7], [8]. The mixing time can be defined with respect to a statistical divergence between probability distributions that we use to measure the error; this includes for example the Total Variation (TV) distance, Kullback-Leibler (KL) divergence, and chi-squared divergence, all of which are instances of a general family of Φ-divergences induced by convex function Φ (see Definition 1, and [9], [10]). Existing results in the literature have established good mixing time guarantees for Markov chains in various divergences, see e.g., [8], [11] for discrete-space Markov chains, and [12] for continuous-space Markov chains.

The mixing time of a Markov chain only tracks how close the last iterate is from the stationary distribution; however, it does not characterize the dependency between iterates along the Markov chain. Even if each iterate along the Markov chain has the correct distribution (e.g., when we initialize the Markov chain from the stationary distribution), the dependencies between successive iterates can be large. In many applications, we may want to generate multiple samples from the target distribution which are *approximately independent*. If we have the ability to run multiple independent chains, then we can produce multiple independent samples. However, if we can only run *one* Markov chain, then a natural strategy is to output a subsequence of the iterates along the Markov chain, by dropping several successive iterates until the next sample that we output is approximately independent from the current sample. For this strategy to work, we need an estimate of how fast the dependence between the iterates is decreasing, or equivalently, how fast the iterates along a Markov chain become approximately independent. This is the central question that we study in this paper.

There are various ways to quantify the dependence between samples. A simple way is to control the covariance or the correlation between the samples; this may be sufficient for

Received 24 July 2025; revised 12 November 2025; accepted 8 December 2025. Date of publication 24 December 2025; date of current version 22 January 2026. The work of Siddharth Mitra and Andre Wibisono was supported by NSF under Award CCF-2403391 and Award CCF-2443097. An earlier version of this paper was presented in part at the 38th Annual Conference on Learning Theory (COLT), 2025. *(Corresponding author: Siddharth Mitra.)*

Jiaming Liang is with the Goergen Institute for Data Science and Artificial Intelligence and the Department of Computer Science, University of Rochester, Rochester, NY 14609 USA (e-mail: jiaming.liang@rochester.edu).

Siddharth Mitra and Andre Wibisono are with the Department of Computer Science, Yale University, New Haven, CT 06520 USA (e-mail: siddharth.mitra@yale.edu; andre.wibisono@yale.edu).

Communicated by C. Rush, Associate Editor for Signal Processing and Source Coding.

Digital Object Identifier 10.1109/TIT.2025.3648262

applications where we want to approximate the expectation of some function via the empirical average from the samples. It is well-known that to control the correlation decay along a Markov chain, a spectral gap or a Poincaré inequality is sufficient; see e.g., [13], [14], see also Appendix F for a review. However, in other applications where we need to estimate more complicated properties of the distribution, such as for entropy or density estimation, we may need to have a stronger control on the dependence between the samples.

A natural measure of dependence in information theory is *mutual information*, which is the KL divergence between the joint distribution of the samples and the product of the marginal distributions; thus, mutual information is non-negative, and it is 0 when the samples are independent. In some applications such as density estimation, we need a control not in the standard mutual information (induced by the KL divergence), but in a generalized notion including the Hellinger mutual information (induced by the Hellinger divergence) [15, Section 4.3] or the chi-squared mutual information (induced by the chi-squared divergence) [16]. This motivates our study in this paper where we measure the dependence between samples via the Φ-*mutual information* induced by the Φ-divergence functional arising from any (twice-differentiable) strictly convex function Φ; see Definition 2. The Φ-mutual information is a stronger measure of dependence than covariance, in that a small Φ-mutual information between samples implies that their covariance is small; see Lemma 10 in Appendix E.

We study three Markov chains which are popular for continuous-space sampling: the (overdamped) Langevin dynamics (7) in continuous time, and its time discretizations – the Unadjusted Langevin Algorithm (ULA) (8) and the Proximal Sampler (10) algorithm; see Section II-D for the definitions. Mixing time guarantees for these Markov chains in various divergences have been well-studied, for the continuous-time Langevin dynamics in e.g., [9], [17], [18], [19]; for the ULA in e.g., [20], [21], [22], [23], [24]; and for the Proximal Sampler in e.g., [24], [25], [26], [27], [28]. We provide additional references and discussion on related works in Appendix A.

In this paper, we study the question of when the iterates become approximately independent for these Markov chains, as measured in Φ-mutual information. We note that in the special case of the standard mutual information, *if* we have a mixing time guarantee for the Markov chain from *any* point mass initialization, then we can deduce a bound on the dependence between the samples; see Lemma 13 in Appendix I. However, this requires mixing from any point mass initialization, which (for continuous-space sampling) is a strong assumption and only known to hold for a few results; furthermore, this argument does not seem to hold for more general Φ-mutual information, see Appendix I for further discussion. We provide guarantees in general Φ-mutual information.

Contributions: We show that when the target distribution is strongly log-concave, the Φ-mutual information converges exponentially fast to 0 along all the Markov chains we study: along the Langevin dynamics (7) in continuous time (see Theorem 1), along the ULA (8) in discrete time under an additional smoothness assumption (see Theorem 3), and along the Proximal Sampler (10) in discrete time (see Theorem 5). Our proof technique proceeds via establishing the strong data processing inequalities along the iterates of the Markov chains, which we describe in Section I-A.

A. Strong Data Processing Inequalities

We briefly discuss our main technique via strong data processing inequalities (SDPIs); we provide a more detailed discussion on SDPIs in Appendix J.

Data processing inequality (DPI) is a fundamental concept in information theory, which states that information cannot increase along a noisy channel or a Markov chain. Concretely, let $\mathbf{P}$ be a Markov kernel or transition operator representing the noisy channel, and let D_Φ be the Φ-divergence (Definition 1) induced by any convex function Φ. Then the DPI in Φ-divergence states that for any probability distributions μ and π, when we apply the same Markov kernel $\mathbf{P}$ to both μ and π, the Φ-divergence between them cannot increase: $\mathsf{D}_\Phi(\mu\mathbf{P} \,\|\, \pi\mathbf{P}) \le \mathsf{D}_\Phi(\mu \,\|\, \pi)$.

Strong data processing inequality (SDPI) [10], [29], [30] is a strengthening of DPI which quantifies the rate at which information is decreasing, and it is typically a function of both the Markov chain and one of the input distributions. For a Markov kernel $\mathbf{P}$ and a probability distribution π, we can define the *contraction coefficient* of $(\mathbf{P}, \pi)$ in Φ-divergence as:

$$\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}, \pi) := \sup_{\mu} \frac{\mathsf{D}_\Phi(\mu\mathbf{P} \,\|\, \pi\mathbf{P})}{\mathsf{D}_\Phi(\mu \,\|\, \pi)} \tag{1}$$

where the supremum is over all probability distributions μ such that $0 < \mathsf{D}_\Phi(\mu \,\|\, \pi) < \infty$. Note $\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}, \pi) \le 1$ by the (weak) DPI. We say that $(\mathbf{P}, \pi)$ satisfies an *SDPI in Φ-divergence* if $\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}, \pi) < 1$. If we can bound $\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}, \nu)$ when ν is stationary for $\mathbf{P}$ (i.e., $\nu\mathbf{P} = \nu$), then we immediately obtain a bound on the mixing time in Φ-divergence along the Markov chain defined by $\mathbf{P}$; see eq. (34) in Appendix J. In practice, bounding the contraction coefficient $\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}, \nu)$ is typically the challenging step to prove mixing guarantees via SDPI; see e.g., [10] for the discrete state space setting, and [24] for the continuous state space setting.

We can equivalently describe DPIs and SDPIs in terms of the decay of mutual information of the iterates along a Markov chain. Let $\mathbf{P}$ be a Markov chain and suppose that the iterates along the chain are $X_i \sim \rho_i$ for $i \ge 0$. The data processing inequality in terms of mutual information states that for any $i \ge 0$, $\mathsf{MI}_\Phi(X_i; X_{i+2}) \le \mathsf{MI}_\Phi(X_i; X_{i+1})$, i.e., that the Φ-mutual information cannot increase along the chain. SDPIs in Φ-mutual information follow similarly by defining appropriate contraction coefficients (Definition 6). After minor calculations, one can relate the drop in Φ-mutual information between X_0 and X_k by the product of contraction coefficients in Φ-divergence along the chain (Lemma 15). That is, $\mathsf{MI}_\Phi(X_0; X_k) \le \prod_{i=\ell}^{k} \varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}, \rho_i)\, \mathsf{MI}_\Phi(X_0; X_\ell)$ for any $\ell \ge 1$ and $k \ge \ell$. Thus, controlling the contraction coefficients also

helps us control the decay of mutual information; however, note that now we need to control the contraction coefficients for distributions *along the trajectory* of the Markov chain (whereas for mixing time, we only need to control the contraction coefficient for the stationary distribution). This task of controlling the contraction coefficients along the trajectory makes studying the information contraction more subtle and challenging than the task for bounding the mixing time. We bound the contraction coefficients along the trajectory for the Langevin dynamics, ULA, and Proximal Sampler in Lemmas 18, 3(a), and 4(a) respectively. These lemmas show that the contraction coefficients are strictly less than 1 so long as the distributions along the trajectory satisfy a Φ-Sobolev inequality (Definition 3). Ensuring that the distributions along the trajectory satisfy a Φ-Sobolev inequality is the crucial part where we need the strong log-concavity of the target distribution.

II. Preliminaries

We say a probability distribution $\nu \propto \exp(-f)$ on $\mathbb{R}^d$ with a twice continuously differentiable potential function $f\colon \mathbb{R}^d \to \mathbb{R}$ is *α-strongly log-concave* (α-SLC) for some $\alpha > 0$ if $\nabla^2 f(x) \succeq \alpha I$ for all $x \in \mathbb{R}^d$; when $\alpha = 0$, we call ν *weakly log-concave*. We say $\nu \propto \exp(-f)$ is *L-smooth* for some $0 < L < \infty$ if $-LI \preceq \nabla^2 f(x) \preceq LI$ for all $x \in \mathbb{R}^d$.

Throughout, we take $\Phi\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ to be a twice-differentiable strictly convex function with $\Phi(1) = 0$. The twice-differentiability ensures that the Φ-Fisher information (5) is well defined, and the strict convexity ensures that the Φ-divergence (Definition 1) is 0 if and only if both arguments are the same. Consequently, this ensures that the Φ-mutual information (Definition 2) is 0 if and only if its arguments are independent.

Let $\mathcal{P}(\mathbb{R}^d)$ denote the space of probability distributions supported on $\mathbb{R}^d$. Let $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ denote the space of distributions $\rho \in \mathcal{P}(\mathbb{R}^d)$ with finite second moment and which are absolutely continuous with respect to the Lebesgue measure. In this paper, we will consider distributions in $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, with a minor exception for Dirac distributions. We let δ_x denote the Dirac distribution (point mass) at point $x \in \mathbb{R}^d$; note $\delta_x \in \mathcal{P}(\mathbb{R}^d)$ but $\delta_x \notin \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. When $\rho \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, we will identify ρ via its density function with respect to the Lebesgue measure, which we also denote by $\rho\colon \mathbb{R}^d \to \mathbb{R}$. The notation $\mu \ll \nu$ denotes that μ is absolutely continuous with respect to ν. Throughout the paper, we assume that the density functions satisfy the required regularity conditions for the statements and results to be well-defined.

A. Φ-Divergence

The Φ-divergence or f-divergence [29], [31] functional is a generalization of many popular statistical divergences such as KL divergence $\left(\mathsf{KL}(\mu \,\|\, \nu) = \mathbb{E}_\mu\left[\log \frac{\mu}{\nu}\right]\right)$ and chi-squared divergence $\left(\chi^2(\mu \,\|\, \nu) = \mathbb{E}_\nu\left[\left(\frac{\mu}{\nu} - 1\right)^2\right]\right)$. The family of Φ-divergences are defined via a convex function $\Phi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ with $\Phi(1) = 0$ as follows. We further assume that Φ is twice-differentiable and strictly convex.

Definition 1: The Φ-divergence between probability distributions μ and ν with $\mu \ll \nu$ is defined by

$$\mathsf{D}_\Phi(\mu \,\|\, \nu) := \mathop{\mathbb{E}}_{\nu}\left[\Phi\left(\frac{\mu}{\nu}\right)\right] = \int_{\mathbb{R}^d} \Phi\left(\frac{\mu(x)}{\nu(x)}\right) \nu(x)\,\mathrm{d}x. \tag{2}$$

If $\mu \not\ll \nu$, then $\mathsf{D}_\Phi(\mu \,\|\, \nu) := +\infty$.

Since we assume Φ is strictly convex, the Φ-divergence functional is non-negative, and it is equal to 0 if and only if both arguments are the same [29, Theorem 7.5]. The KL divergence corresponds to the case $\Phi(x) = x \log x$, and the chi-squared divergence corresponds to the case $\Phi(x) = (x-1)^2$; see Table I in Appendix B for further examples. The family of Φ-divergences share many common properties such as satisfying data processing inequalities, and are therefore natural to study in a unified manner.

B. Φ-Mutual Information

Each Φ-divergence induces a Φ-mutual information, in the same way that the KL divergence induces the standard mutual information. Recall for a joint random variable $(X, Y) \sim \rho^{XY}$, the (classical) mutual information functional is defined as $\mathsf{MI}(X;Y) \equiv \mathsf{MI}(\rho^{XY}) = \mathsf{KL}(\rho^{XY} \,\|\, \rho^X \otimes \rho^Y)$. We can generalize this to define Φ-mutual information in terms of Φ-divergence as follows.

Definition 2: Given two random variables X and Y on $\mathbb{R}^d$ with joint law ρ^{XY}, the Φ-mutual information functional is given by

$$\begin{aligned}\mathsf{MI}_\Phi(X;Y) \equiv \mathsf{MI}_\Phi(\rho^{XY}) &:= \mathsf{D}_\Phi\left(\rho^{XY} \,\|\, \rho^X \otimes \rho^Y\right) \\ &= \mathop{\mathbb{E}}_{x\sim\rho^X}\left[\mathsf{D}_\Phi\left(\rho^{Y|X=x} \,\|\, \rho^Y\right)\right].\end{aligned} \tag{3}$$

By the property of Φ-divergence, we see that the Φ-mutual information functional is always non-negative, and it is equal to 0 if and only if X and Y are independent, i.e., $\rho^{XY} = \rho^X \otimes \rho^Y$. Note Φ-mutual information is symmetric, i.e., $\mathsf{MI}_\Phi(X;Y) = \mathsf{MI}_\Phi(Y;X)$.

C. Φ-Sobolev Inequalities

We now describe Φ-Sobolev inequalities [9], [10], a family of isoperimetric inequalities which include as special cases popular inequalities such as the log-Sobolev inequality and the Poincaré inequality. The family of Φ-Sobolev inequalities also form a natural condition for the exponential convergence of Φ-divergence along the Langevin dynamics, as we review in Lemma 9 in Appendix D; this generalizes the convergence guarantees of KL divergence along the dynamics under a log-Sobolev inequality, and of chi-squared divergence under a Poincaré inequality.

As stated in Section I-A, we bound the contraction coefficients arising in the SDPIs as long as the distributions along the trajectory of the Markov chain satisfy Φ-Sobolev inequalities. Hence, studying these families of inequalities and how they relate to the distributions along the Markov chain is crucial in our approach.

TABLE I
COMMON Φ FUNCTIONS ALONG WITH CORRESPONDING Φ-DIVERGENCES (DEFINITION 1) AND Φ-SOBOLEV INEQUALITIES (DEFINITION 3)

$\Phi(x)$	$\mathsf{D}_\Phi(\mu \,\|\, \nu)$	$\mathsf{D}_\Phi(\mu \,\|\, \nu)$ name	Φ-Sobolev inequality
$x \log x$	$\int \mathrm{d}\mu \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}$	KL divergence	log-Sobolev inequality
$(x-1)^2$	$\int \frac{(\mathrm{d}\mu - \mathrm{d}\nu)^2}{\mathrm{d}\nu}$	chi-squared divergence	Poincaré inequality
$\frac{1}{2}(\sqrt{x}-1)^2$	$\frac{1}{2}\int (\sqrt{\mathrm{d}\mu} - \sqrt{\mathrm{d}\nu})^2$	squared Hellinger distance	–
$\frac{1}{2}\lvert x-1\rvert$	$\frac{1}{2}\int \lvert \mathrm{d}\mu - \mathrm{d}\nu\rvert$	TV distance	–
$-\log x$	$\int \mathrm{d}\nu \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$	reverse KL divergence	–
$\frac{1}{x} - x$	$2 + \int \frac{(\mathrm{d}\mu - \mathrm{d}\nu)^2}{\mathrm{d}\mu}$	reverse chi-squared divergence	–

Definition 3: A probability distribution ν satisfies a Φ-Sobolev inequality (ΦSI) with constant $\alpha > 0$ if for all probability distributions $\mu \ll \nu$, we have

$$2\alpha\, \mathsf{D}_\Phi(\mu \,\|\, \nu) \le \mathsf{FI}_\Phi(\mu \,\|\, \nu), \tag{4}$$

where $\mathsf{D}_\Phi(\mu\|\nu)$ is defined in (2) and $\mathsf{FI}_\Phi(\mu\|\nu)$ is the Φ-Fisher information functional defined as

$$\begin{aligned}\mathsf{FI}_\Phi(\mu \,\|\, \nu) &:= \mathop{\mathbb{E}}_{\nu}\left[\left\|\nabla \frac{\mu}{\nu}\right\|^2 \Phi''\left(\frac{\mu}{\nu}\right)\right] \\ &= \int_{\mathbb{R}^d} \left\|\nabla \frac{\mu(x)}{\nu(x)}\right\|^2 \Phi''\left(\frac{\mu(x)}{\nu(x)}\right) \nu(x)\,\mathrm{d}x. \end{aligned} \tag{5}$$

We define the Φ-Sobolev constant of ν to be the optimal (largest) constant α such that (4) holds, i.e.,

$$\alpha_{\Phi\mathsf{SI}}(\nu) := \inf_{\mu} \frac{\mathsf{FI}_\Phi(\mu \,\|\, \nu)}{2\mathsf{D}_\Phi(\mu \,\|\, \nu)} \tag{6}$$

where the infimum is taken over all probability distributions μ with $0 < \mathsf{D}_\Phi(\mu \,\|\, \nu) < \infty$.

The Φ-Sobolev inequality recovers the log-Sobolev inequality when $\Phi(x) = x\log x$, and the Poincaré inequality when $\Phi(x) = (x-1)^2$. The Poincaré inequality is the weakest Φ-Sobolev inequality in that it is implied by any other Φ-Sobolev inequality [9, Section 2.2].

To study how the Φ-Sobolev constant (6) evolves along the Markov chain, we require further properties of these constants such as how they change under convolutions and pushforwards. We discuss these properties in Appendix C.

D. Langevin Dynamics, Unadjusted Langevin Algorithm, and the Proximal Sampler

Here we formally introduce the Markov chains we will study. References for mixing times of these Markov chains can be found in Section I and Appendix A.

a) Langevin Dynamics: The overdamped Langevin dynamics to sample from $\nu \propto \exp(-f)$ is given by the following stochastic differential equation (SDE)

$$\mathrm{d}X_t = -\nabla f(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \tag{7}$$

where W_t is standard Brownian motion on $\mathbb{R}^d$. The Langevin dynamics has ν as the stationary distribution [32], and hence it is a natural process to study for sampling. However, this dynamics needs to be discretized in time to implement in practice. We will focus on two discrete-time algorithms.

b) Unadjusted Langevin Algorithm: A forward Euler discretization of these dynamics gives rise to the Unadjusted Langevin Algorithm (ULA), explicitly given as

$$X_{k+1} = X_k - \eta\nabla f(X_k) + \sqrt{2\eta}Z_k, \tag{8}$$

where $\eta > 0$ is the step-size and $Z_k \sim \mathcal{N}(0, I)$. Note that as $\eta \to 0$ and $\eta k \to t$, the ULA update (8) recovers the Langevin dynamics (7). However, for each fixed $\eta > 0$, the ULA is biased, i.e., its stationary distribution is $\nu^\eta \neq \nu$. For mixing time analysis, this results in a low-accuracy guarantee where the iteration complexity to reach an error ϵ in any specified divergence scales polynomially in ϵ^{-1}. Here, we show that the ULA still decreases the Φ-mutual information exponentially fast, despite the biased limit; see Theorem 3.

c) Proximal Sampler: The Proximal Sampler is a discrete-time Gibbs sampling-based algorithm. The Proximal Sampler considers an augmented (X, Y) space $\mathbb{R}^d \times \mathbb{R}^d$ and alternatively samples from the conditional distributions. When referring to the Proximal Sampler, we denote the target distribution as $\nu^X \propto \exp(-f)$ on $\mathbb{R}^d$ and in general, use superscripts to denote the space supporting the distribution. The Proximal Sampler then considers the following joint target distribution

$$\nu^{XY}(x, y) \propto \exp\left(-f(x) - \frac{\|x-y\|^2}{2\eta}\right), \tag{9}$$

for step-size $\eta > 0$. The Proximal Sampler, initialized from $X_0 \sim \rho_0^X$, is the following two-step algorithm

Step 1 (forward step):
Sample $Y_k \mid X_k \sim \nu^{Y|X=X_k} = \mathcal{N}(X_k, \eta I)$,
Step 2 (backward step):
Sample $X_{k+1} \mid Y_k \sim \nu^{X|Y=Y_k}$. (10)

Observe that ν^{XY} (9) has the desired target distribution ν^X as the X-marginal. The forward step is easy to implement as

it corresponds to drawing a Gaussian random variable. The backward step can be implemented given access to a Restricted Gaussian Oracle (RGO). A RGO is an oracle that, given any $y \in \mathbb{R}^d$, outputs a sample from $\nu^{X|Y=y}$, i.e., from

$$\nu^{X|Y}(x \mid y) \propto_x \exp\left(-f(x) - \frac{\|x-y\|^2}{2\eta}\right). \tag{11}$$

Similar to [24], [25], and [26], we consider the ideal Proximal Sampler which assumes access to a perfect RGO for our main result. We mention further details and background on the Proximal Sampler including a rejection sampling-based RGO implementation in Appendix G.

III. Convergence of Φ-Mutual Information Along Langevin Dynamics

We show that along the continuous-time Langevin dynamics (7) for strongly log-concave target distributions, the Φ-mutual information converges exponentially fast to 0 as soon as we have an iterate which satisfies a Φ-Sobolev inequality.

Theorem 1: Assume ν is α-SLC for some $\alpha > 0$. Let $X_t \sim \rho_t$ evolve following the Langevin dynamics (7) to ν from $X_0 \sim \rho_0$, and let $\rho_{0,t}$ be the joint law of (X_0, X_t). If for some $s > 0$ we know that ρ_s satisfies a Φ-Sobolev inequality with constant $\alpha_{\Phi\mathsf{SI}}(\rho_s)$, then for all $t \geq s$:

$$\mathsf{MI}_\Phi(\rho_{0,t}) \leq e^{-2\alpha(t-s)} \max\left\{1, \frac{\alpha}{\alpha_{\Phi\mathsf{SI}}(\rho_s)}\right\} \mathsf{MI}_\Phi(\rho_{0,s}). \tag{12}$$

We provide two proofs of Theorem 1. The first proof via our primary strategy of SDPIs follows by taking the appropriate limits of the discrete-time ULA analysis. We present the SDPI-based proof of Theorem 1 in Appendix L-C.2. The second proof of Theorem 1 following a direct time derivative approach is described in Section III-A, and the full proof is presented in Appendix L-B.3.

This rate of convergence of Φ-mutual information matches the rate of convergence of Φ-divergence along the dynamics (Lemma 9). The condition that ρ_s satisfy a Φ-Sobolev inequality (Definition 3) can be ensured by initializing the process from ρ_0 which satisfies a Φ-Sobolev inequality, such as from a strongly log-concave distribution, for example a Gaussian distribution. The fact that strongly log-concave distributions satisfy a Φ-Sobolev inequality is mentioned in Lemma 8, and the fact that if ρ_0 satisfies a Φ-Sobolev inequality then ρ_s (for $s > 0$) does too is a consequence of Lemma 17.

Observe how the right-hand side of (12) has dependence on time s. For the special case of $\Phi(x) = x \log x$, we obtain bounds on the convergence of mutual information along the Langevin dynamics which do not possess this dependence and which do not require ρ_s to satisfy a Φ-Sobolev inequality. This result follows by exploiting the regularity properties of the dynamics and we state it in Theorem 2. We discuss the regularity based approach in Appendix K and prove Theorem 2 in Appendix K-A.

Theorem 2: Let $X_t \sim \rho_t$ evolve following the Langevin dynamics (7) from $X_0 \sim \rho_0$ and let the joint law of (X_0, X_t) be $\rho_{0,t}$. The mutual information $\mathsf{MI}(\rho_{0,t})$ satisfies the following:

(a) if ν is weakly log-concave, then for all $t > 0$

$$\mathsf{MI}(\rho_{0,t}) \leq \frac{1}{2t}\mathsf{Var}(X_0).$$

(b) if ν is α-strongly log-concave for some $\alpha > 0$, then for all $t > 0$

$$\mathsf{MI}(\rho_{0,t}) \leq \frac{\alpha}{e^{2\alpha t} - 1}\mathsf{Var}(X_0). \tag{13}$$

We mention the tightness of Theorems 1 and 2 for the special case of $\Phi(x) = x \log x$ by explicitly computing the mutual information along the Langevin dynamics for ν being a Gaussian distribution; we describe this in Appendix O-A. We also present, in Lemma 5, a bound on the contraction of mutual information along Langevin dynamics under a log-Sobolev inequality assumption. It retains dependence on s but is able to go beyond the strong log-concavity present in all other results.

A. Proof Sketch Based on Time Derivative

We present the direct proof of Theorem 1 which is based on taking the time derivative of the Φ-mutual information functional. The classical de Bruijn's identity [33] shows that the time derivative of the entropy functional along the heat flow is the Fisher information; this has been extended to computing the time derivative of the relative entropy functional between simultaneous evolutions along the heat flow [34] and to general SDEs [23], [24], [26]. Here we show that the time derivative of the Φ-mutual information functional along the Langevin dynamics is given by the Φ-mutual Fisher information, which we define below. We prove Lemma 1 in Appendix L-B.1.

Lemma 1: Suppose $X_t \sim \rho_t$ evolves following the Langevin dynamics (7) and let the joint law of (X_0, X_t) be $\rho_{0,t}$. Then for $t > 0$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{MI}_\Phi(\rho_{0,t}) = -\mathsf{FI}^{\mathsf{M}}_\Phi(\rho_{0,t}), \tag{14}$$

where $\mathsf{FI}^{\mathsf{M}}_\Phi$ is the Φ-mutual Fisher information defined for a joint distribution ρ^{XY} as

$$\mathsf{FI}^{\mathsf{M}}_\Phi(\rho^{XY}) := \mathop{\mathbb{E}}_{x\sim\rho^X}\left[\mathsf{FI}_\Phi\left(\rho^{Y|X=x} \,\|\, \rho^Y\right)\right]. \tag{15}$$

Next, we show that the Φ-mutual Fisher information can be lower bounded in terms of the Φ-mutual information, under a Φ-Sobolev inequality assumption on one of the marginal distributions. We prove Lemma 2 in Appendix L-B.2.

Lemma 2: Let the joint law of (X, Y) be ρ^{XY} and suppose ρ^Y satisfies Φ-Sobolev inequality with optimal constant $\alpha_{\Phi\mathsf{SI}}(\rho^Y)$. Then

$$2\alpha_{\Phi\mathsf{SI}}(\rho^Y)\,\mathsf{MI}_\Phi(\rho^{XY}) \leq \mathsf{FI}^{\mathsf{M}}_\Phi(\rho^{XY}).$$

Theorem 1 follows by combining Lemma 2 with (14) and integrating; we provide this proof in Appendix L-B.3. However, note that combining Lemma 2 with (14) requires ρ_t along the Langevin dynamics to satisfy a Φ-Sobolev inequality. This challenge is consistent with the SDPI approach (discussed in Section I-A), and this is where we crucially require the strong log-concavity of ν.

IV. Convergence of Φ-Mutual Information Along ULA

We show the following result on the exponential convergence of Φ-mutual information along the ULA (8) for

smooth and strongly log-concave target distribution, provided an iterate along ULA satisfies a Φ-Sobolev inequality. We present a proof sketch of Theorem 3 in Section IV-A, and we provide the full proof in Appendix M-B.

Theorem 3: Suppose ν is α-strongly log-concave and L-smooth for some $0 < \alpha \le L < \infty$. Let $X_k \sim \rho_k$ evolve following ULA (8) with step-size $\eta \le 1/L$ from $X_0 \sim \rho_0$, and let the joint law of (X_i, X_j) be $\rho_{i,j}$. If ρ_ℓ satisfies a Φ-Sobolev inequality with optimal constant $\alpha_{\Phi\mathsf{SI}}(\rho_\ell)$ for some $\ell \ge 1$. Then for all $k \ge \ell$, we have

$$\mathsf{MI}_\Phi(\rho_{0,k}) \le (1-\alpha\eta)^{2(k-\ell)} \max\left\{1, \frac{\alpha}{\alpha_{\Phi\mathsf{SI}}(\rho_\ell)}\right\} \mathsf{MI}_\Phi(\rho_{0,\ell}). \quad (16)$$

Theorem 3 implies the following corollary regarding the iteration complexity of ULA to output approximately independent samples. We provide the proof of Corollary 1 in Appendix M-C.

Corollary 1: Under the same assumptions as Theorem 3 and given any error threshold $\epsilon > 0$, ULA (8) outputs a sample X_k such that $\mathsf{MI}_\Phi(X_0; X_k) \le \epsilon$ as long as

$$k \ge \ell + \frac{1}{2\alpha\eta} \log\left(\epsilon^{-1} \max\left\{1, \frac{\alpha}{\alpha_{\Phi\mathsf{SI}}(\rho_\ell)}\right\} \mathsf{MI}_\Phi(\rho_{0,\ell})\right).$$

The condition in Theorem 3 that ρ_ℓ satisfy a Φ-Sobolev inequality can be met by choosing ρ_0 which satisfies a Φ-Sobolev inequality, for example a strongly log-concave distribution. In this case, ρ_ℓ will satisfy a Φ-Sobolev inequality for all $\ell \ge 1$ (consequence of Lemma 3(b)). The fact that strongly log-concave distribution satisfy a Φ-Sobolev inequality is mentioned in Lemma 8. The exponential rate of convergence matches the rate of convergence of Φ-divergence along the ULA [24, Theorem 1]. Also note that the additional smoothness assumption on ν in Theorem 3 is standard in the analysis of ULA.

For the special case of $\Phi(x) = x\log x$, Theorem 4 below studies the convergence of mutual information along ULA as a consequence of the regularity properties of the algorithm. Theorem 4 does not require the distributions along the trajectory to satisfy a Φ-Sobolev inequality and the resulting bound does not depend on $\alpha_{\Phi\mathsf{SI}}(\rho_\ell)$. We discuss the regularity based approach to bound the mutual information in Appendix K and prove Theorem 4 in Appendix K-B.

Theorem 4: Suppose ν is α-strongly log-concave and L-smooth for some $0 < \alpha \le L < \infty$. Let $X_k \sim \rho_k$ evolve following ULA (8) with step-size $\eta \le 1/L$ from $X_0 \sim \rho_0$. Define $\gamma = 1 - \alpha\eta \in (0,1)$ and let the joint law of (X_0, X_k) be $\rho_{0,k}$. Then for all $k \ge 1$, we have

$$\mathsf{MI}(\rho_{0,k}) \le \frac{\alpha\gamma^{2k}}{1-\gamma^{2k}} \mathsf{Var}(X_0).$$

The tightness of Theorems 3 and 4 for the special case of $\Phi(x) = x\log x$ follows by explicitly computing $\mathsf{MI}(\rho_{0,k})$ along the ULA when ν is a Gaussian distribution; we describe this in Appendix O-B.

A. Proof Sketch of Theorem 3

In order to analyze the ULA (8) via SDPIs, it will be helpful to view the update in the space of distributions. Letting $X_k \sim \rho_k$, the update of ρ_k as X_k evolves following (8) is

$$\rho_{k+1} = \rho_k \mathbf{P}_{\mathsf{ULA}} = F_\# \rho_k * \mathcal{N}(0, 2\eta I), \quad (17)$$

where $F(x) = x - \eta\nabla f(x)$ and $\mathbf{P}_{\mathsf{ULA}}$ denotes the Markov kernel of ULA. This two-step interpretation of a pushforward followed by a Gaussian convolution will be crucial in the SDPI analysis. We will denote $\mathbf{P}_{\mathsf{ULA}}$ by $\mathbf{P}$ when the Markov chain is clear.

As described briefly in Section I-A and in detail in Appendix J, studying the Φ-mutual information contraction along a Markov chain involves bounding each of the contraction coefficients along the trajectory. The contraction coefficient in terms of Φ-divergence, which we bound for the ULA in Lemma 3(a), are defined in (1) and in Definition 4. Lemma 3(a) bounds the contraction coefficient along ULA so long as the distribution along the chain satisfies a Φ-Sobolev inequality. Ensuring that successive iterates along ULA satisfy a Φ-Sobolev inequality is given in Lemma 3(b). We prove Lemma 3 in Appendix M-A.

Lemma 3: Consider the ULA kernel $\mathbf{P}$ defined in (17) with $\|F\|_{\mathsf{Lip}} = \gamma < 1$ and $\eta > 0$. Let ρ be a distribution that satisfies a Φ-Sobolev inequality with optimal constant $\alpha_{\Phi\mathsf{SI}}(\rho)$. Then we have:

(a) If F is also bijective, then the contraction coefficient satisfies
$$\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}, \rho) \le \frac{\gamma^2}{\gamma^2 + 2\eta\,\alpha_{\Phi\mathsf{SI}}(\rho)}.$$

(b) The Φ-Sobolev inequality constants $\alpha_{\Phi\mathsf{SI}}(\rho)$ and $\alpha_{\Phi\mathsf{SI}}(\rho\mathbf{P})$ satisfy
$$\alpha_{\Phi\mathsf{SI}}(\rho\mathbf{P}) \ge \frac{\alpha_{\Phi\mathsf{SI}}(\rho)}{\gamma^2 + 2\eta\,\alpha_{\Phi\mathsf{SI}}(\rho)}. \quad (18)$$

Theorem 3 then follows by combining both parts of Lemma 3 across multiple steps of ULA.

V. Convergence of Φ-Mutual Information Along Proximal Sampler

We show the following result on the exponential convergence of Φ-mutual information along the Proximal Sampler (10) for strongly log-concave target distributions, provided an iterate along the Proximal Sampler satisfies a Φ-Sobolev inequality. We provide a proof sketch of Theorem 5 in Section V-A, and present the complete proof in Appendix N-C.2.

Theorem 5: Suppose $\nu^X \propto \exp(-f)$ is α-strongly log-concave for some $\alpha > 0$. Let $X_k \sim \rho_k^X$ evolve following the Proximal Sampler (10) with step-size $\eta > 0$ from $X_0 \sim \rho_0^X$, and let the joint law of (X_i, X_j) be $\rho_{i,j}^X$. If ρ_ℓ^X satisfies a Φ-Sobolev inequality with optimal constant $\alpha_{\Phi\mathsf{SI}}(\rho_\ell^X)$ for some $\ell \ge 1$. Then for all $k \ge \ell$, we have

$$\mathsf{MI}_\Phi(\rho_{0,k}^X) \le \frac{\mathsf{MI}_\Phi(\rho_{0,\ell}^X)}{(1+\eta\min\{\alpha, \alpha_{\Phi\mathsf{SI}}(\rho_\ell^X)\})^{2(k-\ell)}}. \quad (19)$$

Theorem 5 implies the following iteration complexity for the Proximal Sampler, stated in Corollary 2. We provide the proof of Corollary 2 in Appendix N-C.3.

Corollary 2: Under the same assumptions as Theorem 5 and given any error threshold $\epsilon > 0$, the Proximal Sampler (10) outputs a sample X_k such that $\mathsf{MI}_\Phi(X_0; X_k) \le \epsilon$ as long as

$$k \ge \ell + \left[\frac{1}{2} + \frac{1}{2\eta \min\{\alpha, \alpha_{\Phi\mathsf{SI}}(\rho_\ell^X)\}}\right] \log\left(\epsilon^{-1}\mathsf{MI}_\Phi(\rho_{0,\ell}^X)\right).$$

The condition in Theorem 5 that ρ_ℓ^X satisfy a Φ-Sobolev inequality can be met by choosing a ρ_0^X which satisfies a Φ-Sobolev inequality, such as a strongly log-concave distribution. The fact that when ρ_0^X satisfies a Φ-Sobolev inequality, ρ_ℓ^X does as well (for all $\ell \ge 1$) is a consequence of Lemma 4(b). Additionally, Lemma 8 states that strongly log-concave distributions satisfy Φ-Sobolev inequalities. Note that the rate of convergence of Φ-mutual information matches that of Φ-divergence; see [24, Theorem 2]. Further note that there is no smoothness assumption required in Theorem 5, since the result is for the ideal Proximal Sampler which assumes access to a perfect RGO. Implementing the RGO typically requires smoothness assumptions; see Appendix G for a review. In Corollary 3 in Appendix G, we show the expected oracle complexity when we combine the convergence guarantee from Theorem 5 with the standard implementation of RGO via rejection sampling.

For the special case of $\Phi(x) = x \log x$ and $\ell = 1$, we are able to bound the initial mutual information $\mathsf{MI}(\rho_{0,1}^X)$ in terms of the variance of ρ_0. We state this in Appendix K-C and present the corresponding iteration complexity in Corollary 4. The tightness of Theorem 5 for the special case of $\Phi(x) = x \log x$ can be seen by doing explicit calculations for the case when ν is Gaussian, and we present this in Appendix O-C.

A. Proof Sketch of Theorem 5

Recall the Proximal Sampler (10). We denote the Proximal Sampler by $\mathbf{P}_{\text{prox}}$, i.e., $\rho_k^X \mathbf{P}_{\text{prox}} = \rho_{k+1}^X$ where $\rho_k^X := \mathsf{law}(X_k)$. The proof of Theorem 5 follows the template mentioned in Section I-A (and mentioned in detail in Appendix J) where the key step is in bounding the contraction coefficient along the trajectory of the Proximal Sampler. The bound on the contraction coefficient for the Proximal Sampler is presented in Lemma 4(a), which holds under the distribution satisfying a Φ-Sobolev inequality. To apply Lemma 4(a) across multiple steps of the Proximal Sampler, we need to ensure that the Φ-Sobolev inequality assumption is maintained along the trajectory of the chain. This is guaranteed by Lemma 4(b), which crucially requires the strong log-concavity of ν^X. We prove Lemma 4 in Appendix N-C.1.

Lemma 4: Suppose $\nu^X \propto \exp(-f)$ is α-strongly log-concave for some $\alpha > 0$, and consider the Proximal Sampler $\mathbf{P}_{\text{prox}}$ (10) with step-size $\eta > 0$ for sampling from ν^X. Let ρ be a distribution that satisfies a Φ-Sobolev inequality with optimal constant $\alpha_{\Phi\mathsf{SI}}(\rho)$. Then we have the following:

(a) The contraction coefficient satisfies

$$\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}_{\text{prox}}, \rho) \le \frac{1}{1 + 2\eta\alpha_{\Phi\mathsf{SI}}(\rho) + \eta^2\alpha\alpha_{\Phi\mathsf{SI}}(\rho)}.$$

(b) The Φ-Sobolev inequality constants $\alpha_{\Phi\mathsf{SI}}(\rho)$ and $\alpha_{\Phi\mathsf{SI}}(\rho\mathbf{P}_{\text{prox}})$ satisfy

$$\frac{1}{\alpha_{\Phi\mathsf{SI}}(\rho\mathbf{P}_{\text{prox}})} \le \frac{1 + \alpha_{\Phi\mathsf{SI}}(\rho)\eta}{\alpha_{\Phi\mathsf{SI}}(\rho)(1+\alpha\eta)^2} + \frac{\eta}{1+\alpha\eta}.$$

Suitably combining both parts of Lemma 4 across multiple steps of the Proximal Sampler proves Theorem 5, which we present in Appendix N-C.2.

1) Proof Sketch of Lemma 4: Recall that the Proximal Sampler (10) is composed of a forward step and a backward step. We denote by $\mathbf{P}_{\text{prox}}^+$ and $\mathbf{P}_{\text{prox}}^-$ the Markov kernel corresponding to the forward and the backward step, respectively, so we can write the Proximal Sampler as the composition $\mathbf{P}_{\text{prox}} = \mathbf{P}_{\text{prox}}^+ \mathbf{P}_{\text{prox}}^-$.

We use the SDE interpretations of the forward step $\mathbf{P}_{\text{prox}}^+$ and the backward step $\mathbf{P}_{\text{prox}}^-$ to control the contraction coefficient and the evolution of the Φ-Sobolev constant along each step. We combine these estimates to to prove the estimates for the Proximal Sampler claimed in Lemma 4.

a) Forward Step: Suppose we start from $X_0 \sim \rho_0^X$. Along the forward step of the Proximal Sampler (10), $Y_0 \mid X_0 \sim \mathcal{N}(X_0, \eta I)$, so in particular, $\rho_0^Y = \rho_0^X * \mathcal{N}(0, \eta I)$. Therefore, the action of the forward step $\mathbf{P}_{\text{prox}}^+$ is via a Gaussian convolution: $\rho\mathbf{P}_{\text{prox}}^+ = \rho * \mathcal{N}(0, \eta I)$, which can be interpreted as the solution to the heat flow (generated by the Brownian motion SDE $\mathrm{d}X_t = \mathrm{d}W_t$) at time $\eta > 0$.

b) Backward Step: For the backward step $\mathbf{P}_{\text{prox}}^-$, it will be helpful to think of the corresponding SDE as the time reversal of the forward step SDE, i.e., the time reversal of the heat flow, which is known as the *backward heat flow*; see [26] and [12, Chapter 8.3].

Fixing a step-size $\eta > 0$, we have along the forward step ($\mathrm{d}X_t = \mathrm{d}W_t$) that if $X_0 \sim \nu^X$, then $X_\eta \sim \nu^Y$. The backward heat flow SDE is defined by

$$\mathrm{d}Y_t = \nabla \log\left(\nu^X * \mathcal{N}_{\eta - t}\right)(Y_t)\,\mathrm{d}t + \mathrm{d}W_t. \tag{20}$$

By construction, if we start the SDE (20) from $Y_0 \sim \mu_0 = \nu^Y$, then for any $t \in [0, \eta]$, the distribution of $Y_t \sim \mu_t$ along (20) is given by $\mu_t = \nu^X * \mathcal{N}_{\eta - t}$, and in particular, at time $t = \eta$, $Y_\eta \sim \mu_\eta = \nu^X$. For the Proximal Sampler, we start the backward SDE (20) from $Y_0 \sim \rho_0^Y$, to obtain $X_1 := Y_\eta \sim \rho_1^X$.

We note that computing the contraction coefficient for the backward step (Lemma 20) and analyzing the evolution of the Φ-Sobolev constant along the backward step (Lemma 21) is challenging due to the time-varying drift in the backward step SDE (20), and it is the key difficulty when studying the Proximal Sampler. We provide the details of the computations above in Appendix N.

VI. Discussion

We study the convergence of Φ-mutual information along the Langevin dynamics (7), ULA (8), and Proximal Sampler (10) assuming the strong log-concavity of ν. Our primary proof strategy is based on SDPIs and we explain how studying the contraction of information along a Markov chain requires controlling the SDPI contraction coefficients along the trajectory. We require the strong log-concavity of ν as this implies that the distributions along the trajectory

satisfy Φ-Sobolev inequalities, under which the contraction coefficients are strictly less than 1. In addition to the SDPI based proof strategy, we have a direct time derivative approach for the continuous-time Langevin dynamics as well as an approach based on the regularity of the Langevin dynamics and ULA kernel – both of these approaches also require the strong log-concavity of the target distribution to obtain exponentially fast contraction of Φ-mutual information and mutual information respectively. The reliance on strong-log concavity for contraction of information is distinct from mixing time results, where we have a rich understanding beyond strong log-concavity and under only isoperimetric assumptions on ν such as a log-Sobolev inequality, a Poincaré inequality, or in general a Φ-Sobolev inequality. We now present an attempt at going beyond strong log-concavity and show contraction of (classical) mutual information along Langevin dynamics under a log-Sobolev inequality assumption on ν.

Lemma 5: Suppose ν satisfies a log-Sobolev inequality with optimal constant $\alpha > 0$. Let $X_t \sim \rho_t$ evolve along the Langevin dynamics (7) to ν from $X_0 \sim \rho_0$ and let $\rho_{0,t}$ be the joint law of (X_0, X_t). Then for any $t \geq s > 0$,

$$\mathsf{MI}(\rho_{0,t}) \leq e^{-2\alpha(t-s)} \left[\mathsf{MI}(\rho_{0,s}) + \mathsf{KL}(\rho_s \,\|\, \nu)\right] .$$

We prove Lemma 5 in Appendix P. It would be interesting to extend Lemma 5 and explore the extent to which we can obtain bounds on the contraction of information in both continuous and discrete time. It would also be fascinating to study regimes where contraction of information is possibly faster than mixing time.

Additional directions for future work include studying the contraction of mutual information and Φ-mutual information for other Markov chains such as the underdamped Langevin dynamics and Hamiltonian Monte Carlo. Furthermore, just as the Langevin dynamics is the Wasserstein gradient flow for the relative entropy functional, one can also study the gradient flow of Φ-mutual information directly and it would be interesting to explore algorithmic implications of such a flow.

Appendix A
Additional Related Works

The Langevin Markov chains we consider in this paper are widely used for continuous-space sampling; see [12] for an overview. The mixing time of the Langevin dynamics in various divergences has been studied in many works, including [9], [17], [18], [19], [35], [36]. We review the convergence guarantee of the Φ-divergence along the Langevin dynamics under a Φ-Sobolev inequality in Appendix D. The mixing time of the ULA has been studied in [20], [21], [22], [23], [24], [37], [38], [39], [40], [41], and [42], and the mixing time of the Proximal Sampler has been studied in [24], [25], [26], [27], [28], [43], [44], and [45].

The family of Φ-divergences [29], [31] is broad and includes many applications in addition to mixing time of Markov chains [9], [10], [24]. They have been used for hypothesis and distribution testing [46], [47], [48], reinforcement learning [49], [50], neuroscience [51], [52], and differential privacy [53], [54], among other applications. The induced Φ-mutual informations have also been used in varied applications such as density estimation [15], [16], contrastive learning [55], generalization [56], and investment and portfolio theory [10], [57], [58].

SDPIs have been popular for proving mixing time of Markov chains [10] and other general networks and processes [30]. They have been frequent in the context of Langevin-type Markov chains as well, although much of the connection between SDPIs and mixing times there has been implicit. [23] use SDPIs to study the mixing time of ULA in Rényi divergence, [26] use them for the Proximal Sampler, [27] use them for the Proximal Sampler on graphs, and recently, [24] use them explicitly to study the mixing time of the ULA and Proximal Sampler in Φ-divergence.

Identically distributed and dependent random variables are also studied more generally under *mixing* [59], [60], where different definitions of dependency between sigma-algebras such as α, β, and ρ-mixing [60, (1.1) to (1.5)] are compared [60, (1.11) to (1.18)] and studied. In this paper, we focus specifically on data coming from Langevin-type Markov chains, and on Φ-mutual information as our functional to measure dependency. Additionally, we do not assume the data to be stationary (i.e., identically distributed), which is helpful for modern Markov Chain Monte Carlo (MCMC) with short and unmixed chains [61], [62].

The behaviors of mutual information along the evolution of a stochastic process have attracted interests in the information theory literature. Along the Gaussian channel or the heat flow, the rate of change of mutual information is related to the minimum mean-square error (MMSE), known as the I-MMSE relationship [63]; this relationship can be generalized to the Poisson channel [64] and to the family of *Fokker-Planck* channels induced by a stochastic process driven by a Brownian motion [65], or the jump-diffusion process [66]. Furthermore, the I-MMSE relationship along the Gaussian channel was recently used [34] to obtain lower bounds on the mutual information between the input and output random variables in terms of the Poincaré constant of the input distribution. The *convexity* of how the mutual information is decreasing has been studied along the heat flow [67], the Ornstein-Uhlenbeck process [68], and the general Fokker-Planck channel [69]. In this work, we quantify the rate of decrease of the mutual information along the channel induced by the Langevin dynamics and its discrete-time implementations, the ULA and the Proximal Sampler.

Appendix B
Examples of Φ-Divergences

Table I includes examples of common Φ-divergences. Further examples include Jensen-Shannon divergence and Le Cam divergence and can be found in [29, Chapter 7] and [31].

Appendix C
Properties of Φ-Sobolev Inequalities

Recall the Φ-Sobolev inequality defined in Definition 3. Note that the inequality (4) is equivalent to saying that for all smooth functions $g : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ with $\mathbb{E}_\nu[g] = 1$,

$$2\alpha\, \mathsf{Ent}^\nu_\Phi(g) \leq \mathcal{E}^\nu_\Phi(g) ,$$

where

$$\mathsf{Ent}^{\nu}_{\Phi}(g) := \mathop{\mathbb{E}}_{\nu}[\Phi(g)] - \Phi(\mathop{\mathbb{E}}_{\nu}[g]) \text{ and } \mathcal{E}^{\nu}_{\Phi}(g) := \mathop{\mathbb{E}}_{\nu}\left[\|\nabla g\|^2 \Phi''(g)\right].$$

This can be seen by taking g to be the density function of μ with respect to ν.

Further recall the optimal Φ-Sobolev constant defined in (6). Understanding how this constant evolves along the various operations that constitute the Markov chains we study such as pushforward and convolution is crucial in our approach. The following properties from [9] describe how the Φ-Sobolev inequality constant evolves along these operations.

Lemma 6 ([9, remark 7]): Assume ν satisfies a Φ-Sobolev inequality with optimal constant $\alpha_{\Phi\mathsf{SI}}(\nu)$. Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a γ-Lipschitz map. Then the pushforward $\tilde{\nu} = T_{\#}\nu$ satisfies a Φ-Sobolev inequality with optimal constant

$$\alpha_{\Phi\mathsf{SI}}(\tilde{\nu}) \geq \frac{\alpha_{\Phi\mathsf{SI}}(\nu)}{\gamma^2}. \tag{21}$$

The next lemma describes the change of the $\Phi\mathsf{SI}$ constant after convolution.

Lemma 7 ([9, corollary 3.1]): Assume μ and ν satisfy a Φ-Sobolev inequality with optimal constants $\alpha_{\Phi\mathsf{SI}}(\mu)$ and $\alpha_{\Phi\mathsf{SI}}(\nu)$ respectively. Then the convolution $\mu * \nu$ satisfies the Φ-Sobolev inequality with constant

$$\frac{1}{\alpha_{\Phi\mathsf{SI}}(\mu * \nu)} \leq \frac{1}{\alpha_{\Phi\mathsf{SI}}(\mu)} + \frac{1}{\alpha_{\Phi\mathsf{SI}}(\nu)}. \tag{22}$$

The following lemma tells us that when ν is α-strongly log-concave, it also satisfies a Φ-Sobolev inequality with the same constant.

Lemma 8 ([9, corollary 2.1]): If ν is α-strongly log-concave for some $\alpha > 0$, then ν satisfies a Φ-Sobolev inequality with constant

$$\alpha_{\Phi\mathsf{SI}}(\nu) \geq \alpha.$$

Appendix D
Review of the Fast Mixing of Langevin Dynamics in Φ-Divergence

Recall the Langevin dynamics (7) to sample from $\nu \propto \exp(-f)$ on $\mathbb{R}^d$

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

where W_t is Brownian motion on $\mathbb{R}^d$. These dynamics can equivalently be viewed as the gradient flow for the KL divergence or relative entropy functional $\mathsf{KL}(\cdot \,\|\, \nu)$ in the space of distributions with the Wasserstein metric [70], [71]. A log-Sobolev inequality corresponds to the gradient-domination condition for the relative entropy objective functional, under which there is rapid convergence of $\mathsf{KL}(\cdot\|\nu)$ along the dynamics. This affirms that the Langevin dynamics are well-suited for sampling from ν.

The fast convergence of KL divergence under a log-Sobolev inequality assumption on ν can be extended to showing rapid convergence of Φ-divergence under a Φ-Sobolev inequality assumption [9], [36]. As the Poincaré inequality is the Φ-Sobolev inequality for $\Phi(x) = (x-1)^2$, this also includes the convergence of chi-squared divergence under a Poincaré inequality as a special case. The convergence of Φ-divergence under a Φ-Sobolev inequality follows easily from Lemma 12 and we mention it below.

Lemma 9: Suppose $X_t \sim \rho_t$ evolves along the Langevin dynamics (7) to sample from $\nu \propto \exp(-f)$, and let ν satisfy a Φ-Sobolev inequality with optimal constant $\alpha > 0$. Then,

$$\mathsf{D}_{\Phi}(\rho_t \,\|\, \nu) \leq e^{-2\alpha t}\mathsf{D}_{\Phi}(\rho_0 \,\|\, \nu).$$

Proof: Applying Lemma 12 (with $\mu_t = \rho_t$, $\nu_t = \nu$, $b_t(x) = -\nabla f(x)$ and $c = 1$) along with the Φ-Sobolev inequality of ν (Definition 3) yields,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{D}_{\Phi}(\mu_t \,\|\, \nu) = -\mathsf{FI}_{\Phi}(\mu_t \,\|\, \nu) \leq -2\alpha \mathsf{D}_{\Phi}(\mu_t \,\|\, \nu).$$

Using Grönwall's lemma, we conclude the desired bound and thus complete the proof. □

Appendix E
Covariance and Φ-Mutual Information

Here we show that the covariance between two random variables can be upper bounded by their Φ-mutual information. This illustrates that Φ-mutual information is a stronger notion of independence than covariance or correlation.

For random variables $X, Y \in \mathbb{R}^d$, let $\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^\top] \in \mathbb{R}^{d\times d}$ denote the covariance matrix of X and Y. Recall that the convex conjugate of a function $f : \mathbb{R} \to \mathbb{R}$ is $f^*(y) = \sup_{x\in\mathbb{R}} xy - f(x)$.

Lemma 10: Let $(X,Y) \sim \rho^{XY}$ be a joint random variable. Let Φ_{ext} be an extension of Φ, defined as $\Phi_{\mathrm{ext}}(x) = \Phi(x)$ for $x \geq 0$, and $\Phi_{\mathrm{ext}}(x) = \infty$ for $x < 0$. Further denote Φ^*_{ext} to be its convex conjugate. Assume $Y \sim \rho^Y$ satisfies the following:

$$\exists\, 0 < \xi < \infty, \text{such that } \forall\, \theta \in \mathbb{R}^d, \tag{23}$$
$$\inf_{a\in\mathbb{R}} \mathop{\mathbb{E}}_{\rho^Y}\left[\Phi^*_{\mathrm{ext}}(\langle\theta, Y - \mathbb{E}[Y]\rangle - a)\right] + a \leq \frac{\|\theta\|^2\xi^2}{2}.$$

Then,

$$\|\mathrm{Cov}(X,Y)\|_{\mathsf{op}} \leq \xi\sqrt{2\,\|\mathrm{Cov}(X,X)\|_{\mathsf{op}}\,\mathsf{MI}_{\Phi}(\rho^{XY})}. \tag{24}$$

Proof: Let $u, v \in \mathbb{R}^d$ with $\|u\| = \|v\| = 1$. We can bound

$$\begin{aligned} &u^\top \mathrm{Cov}(X,Y)v \\ &= \mathop{\mathbb{E}}_{\rho^{XY}}\left[(u^\top X - \mathop{\mathbb{E}}_{\rho^X}[u^\top X])(v^\top Y - \mathop{\mathbb{E}}_{\rho^Y}[v^\top Y])\right] \\ &= \mathop{\mathbb{E}}_{\rho^X}\left[(u^\top X - \mathop{\mathbb{E}}_{\rho^X}[u^\top X])(v^\top \mathop{\mathbb{E}}_{\rho^{Y|X}}[Y] - \mathop{\mathbb{E}}_{\rho^Y}[v^\top Y])\right] \\ &\leq \mathop{\mathbb{E}}_{\rho^X}\left[\left|u^\top X - \mathop{\mathbb{E}}_{\rho^X}[u^\top X]\right| \cdot \left|v^\top \mathop{\mathbb{E}}_{\rho^{Y|X}}[Y] - \mathop{\mathbb{E}}_{\rho^Y}[v^\top Y]\right|\right]. \end{aligned} \tag{25}$$

The variational representation of Φ-divergence [29, Theorem 7.26] for $\pi_1 \ll \pi_2$ states that:

$$\mathsf{D}_{\Phi}(\pi_1 \,\|\, \pi_2) = \sup_{g:\,\mathbb{R}^d\to\mathbb{R}} \mathop{\mathbb{E}}_{Y\sim\pi_1}\left[g(Y)\right] - \psi^*_{\pi_2}(g), \tag{26}$$

where

$$\psi^*_{\pi_2}(g) := \inf_{a\in\mathbb{R}} \mathop{\mathbb{E}}_{Y\sim\pi_2}\left[\Phi^*_{\mathrm{ext}}(g(Y) - a)\right] + a. \tag{27}$$

For any $\lambda > 0$, take

$$g(y) = \lambda\left(v^\top y - \mathop{\mathbb{E}}_{\rho^Y}[v^\top Y]\right), \quad \pi_1 = \rho^{Y|X}, \quad \pi_2 = \rho^Y$$

for a fixed $X \in \mathbb{R}^d$, and substitute in (26) to obtain,

$$\lambda v^\top \mathop{\mathbb{E}}_{\rho^{Y|X}}[Y] - \lambda \mathop{\mathbb{E}}_{\rho^Y}[v^\top Y] = \mathop{\mathbb{E}}_{\rho^{Y|X}}[g(Y)] \le \mathsf{D}_\Phi(\rho^{Y|X} \,\|\, \rho^Y) + \psi^*_{\rho^Y}(g).$$

Therefore, we have that

$$\begin{aligned} v^\top \mathop{\mathbb{E}}_{\rho^{Y|X}}[Y] - \mathop{\mathbb{E}}_{\rho^Y}[v^\top Y] &\le \frac{1}{\lambda}\mathsf{D}_\Phi(\rho^{Y|X} \,\|\, \rho^Y) + \frac{1}{\lambda}\psi^*_{\rho^Y}(g), \\ &= \frac{1}{\lambda}\mathsf{D}_\Phi(\rho^{Y|X} \,\|\, \rho^Y) + \\ &\quad \frac{1}{\lambda}\left[\inf_{a\in\mathbb{R}} \mathop{\mathbb{E}}_{\rho^Y}\left[\Phi^*_{\text{ext}}\left(\lambda(v^\top Y - \mathop{\mathbb{E}}_{\rho^Y}[v^\top Y]) - a\right)\right] + a\right], \end{aligned}$$

where the identity is due to (27). Using the assumption (23) on ρ^Y along with the fact that $\|v\| = 1$, we have that,

$$v^\top \mathop{\mathbb{E}}_{\rho^{Y|X}}[Y] - \mathop{\mathbb{E}}_{\rho^Y}[v^\top Y] \le \frac{1}{\lambda}\mathsf{D}_\Phi(\rho^{Y|X} \,\|\, \rho^Y) + \frac{\lambda}{2}\xi^2.$$

Choosing the optimal $\lambda = \frac{1}{\xi}\sqrt{2\mathsf{D}_\Phi(\rho^{Y|X} \,\|\, \rho^Y)}$ gives the bound,

$$v^\top \mathop{\mathbb{E}}_{\rho^{Y|X}}[Y] - \mathop{\mathbb{E}}_{\rho^Y}[v^\top Y] \le \xi\sqrt{2\mathsf{D}_\Phi(\rho^{Y|X} \,\|\, \rho^Y)}.$$

Since the right-hand side does not depend on v, the same argument applied to $-v$ yields the bound

$$\left| v^\top \mathop{\mathbb{E}}_{\rho^{Y|X}}[Y] - \mathop{\mathbb{E}}_{\rho^Y}[v^\top Y] \right| \le \xi\sqrt{2\mathsf{D}_\Phi(\rho^{Y|X} \,\|\, \rho^Y)}.$$

We then plug this in to (25) and get

$$\begin{aligned} &u^\top \mathrm{Cov}(X,Y)v \\ &\le \mathop{\mathbb{E}}_{\rho^X}\left[\left|u^\top X - \mathop{\mathbb{E}}_{\rho^X}[u^\top X]\right| \cdot \xi\sqrt{2\mathsf{D}_\Phi(\rho^{Y|X} \,\|\, \rho^Y)}\right] \\ &\le \xi\sqrt{2}\left(\mathop{\mathbb{E}}_{\rho^X}\left[(u^\top X - \mathop{\mathbb{E}}_{\rho^X}[u^\top X])^2\right]\right)^{\frac{1}{2}}\left(\mathop{\mathbb{E}}_{\rho^X}[\mathsf{D}_\Phi(\rho^{Y|X} \,\|\, \rho^Y)]\right)^{\frac{1}{2}} \\ &= \xi\sqrt{2}\ \sqrt{u^\top \mathrm{Cov}_{\rho^X}(X,X)u}\ \sqrt{\mathsf{MI}_\Phi(\rho^{XY})} \\ &\le \xi\sqrt{2}\ \sqrt{\|\mathrm{Cov}(X,X)\|_{\mathsf{op}}\,\mathsf{MI}_\Phi(\rho^{XY})}, \end{aligned}$$

where the second inequality follows by applying the Cauchy-Schwarz inequality, and the last inequality is because $\|\mathrm{Cov}(X,Y)\|_{\mathsf{op}} = \sup\{u^\top \mathrm{Cov}(X,Y)v : \|u\| = \|v\| = 1\}$. Finally, the conclusion (24) of the lemma follows from this argument as well. □

The assumption on ρ^Y in (23) in Lemma 10 is a generalization of a sub-Gaussianity assumption on ρ^Y, which (23) simplifies to for $\Phi(x) = x\log x$. Indeed, for $\Phi(x) = x\log x$, $\Phi^*_{\text{ext}}(y) = e^{y-1}$ and $\psi^*_{\pi_2}(g) = \log \mathbb{E}_{\pi_2}[e^{g(Y)}]$ (defined in (27)), which means that ρ^Y satisfying (23) means that it is ξ sub-Gaussian. For $\Phi(x) = x\log x$, (26) corresponds to the Donsker-Varadhan variational formula for KL divergence.

Lemma 10 strengthens [72, Lemma 1] to (a) only requiring a marginal sub-Gaussian condition, and (b) extending the result to any Φ-mutual information.

Appendix F
Review of the Covariance Decay Under Poincaré Inequality

We review the classical fact that for any reversible Markov semigroup $\mathbf{P}_t$ satisfying a Poincaré inequality, there is exponential convergence of covariance when measured against functions in $L^2(\nu)$ (where ν is stationary for $\mathbf{P}_t$); see also [13], [14].

Lemma 11: Let $\mathbf{P}_t$ be a reversible Markov semigroup with stationary distribution ν and associated stochastic process $X_t \sim \rho_t$. Suppose $\mathbf{P}_t$ satisfies a Poincaré inequality with constant $\alpha > 0$ and is at stationarity, i.e., $X_0 \sim \nu$. Then for all functions $f \in L^2(\nu)$ and all $t \ge 0$

$$\mathrm{Cov}(f(X_t), f(X_0)) \le \exp(-\alpha t)\mathsf{Var} f(X_0).$$

Proof: Let f be an arbitrary function in $L^2(\nu)$. Denote $\mathbb{E}_\nu[f] = \hat{f}$ and $g := f - \hat{f}$. Then $g \in L^2(\nu)$ and $\mathbb{E}_\nu[g] = 0$. Also denote $\rho_{0,t} := \mathsf{law}(X_0, X_t)$. Recall (see e.g., [12, Theorem 1.2.21]) that a reversible semigroup $\mathbf{P}_t$ satisfies a Poincaré inequality with constant $\alpha > 0$ if for all functions $g \in L^2(\nu)$ with $\mathbb{E}_\nu[g] = 0$,

$$\|\mathbf{P}_t g\|^2_{L^2(\nu)} \le \exp(-2\alpha t)\,\|g\|^2_{L^2(\nu)}. \tag{28}$$

Therefore, keeping in mind that the process is at stationarity, we have the following.

$$\begin{aligned} \mathrm{Cov}(f(X_t), f(X_0)) &= \mathbb{E}\left[(f(X_t) - \hat{f})(f(X_0) - \hat{f})\right] \\ &= \mathbb{E}\left[(f - \hat{f})(X_t)\,(f - \hat{f})(X_0)\right] \\ &= \mathop{\mathbb{E}}_{\rho_{0,t}}\left[g(X_t)\,g(X_0)\right] \\ &= \mathop{\mathbb{E}}_{\rho_0}\left[g(X_0)\ \mathop{\mathbb{E}}_{\rho_{t|0}}[g(X_t)]\right] \\ &= \langle g, \mathbf{P}_t g\rangle_\nu \\ &\le \sqrt{\|\mathbf{P}_t g\|^2_{L^2(\nu)}\|g\|^2_{L^2(\nu)}} \\ &\le \exp(-\alpha t)\|g\|^2_{L^2(\nu)} \\ &= \exp(-\alpha t)\|f - \hat{f}\|^2_{L^2(\nu)} \\ &= \exp(-\alpha t)\mathsf{Var}_\nu f \\ &= \exp(-\alpha t)\mathsf{Var} f(X_0). \end{aligned}$$

Here, the third equality is by definition of g, the fifth equality is by the definition of a Markov semigroup and the $L^2(\nu)$ inner product, the first inequality is by Cauchy-Schwarz inequality, and the last inequality is due to Poincaré inequality (28). □

Appendix G
Further Background on Proximal Sampler and RGO Implementation via Rejection Sampling

The Proximal Sampler (10) is a discrete-time continuous-space Gibbs sampling algorithm which has been studied for both unbounded space [24], [25], [26] and constrained space [45], [73]. Our main result for the Proximal Sampler, Theorem 5, considers the ideal Proximal Sampler which assumes access to an RGO which outputs a sample from (11) without any bias. This is assumed in many works [24], [25], [26]. Improved analysis of the Proximal Sampler and better RGO

implementations is an active area of research [28], [43], [44], [74], [75].

For clarity, we provide a brief review of a basic RGO implementation via rejection sampling in the smooth case, and state the corresponding oracle complexity (i.e., the expected number of calls to a first-order oracle of f) when using the Proximal Sampler with a rejection sampling-based RGO implementation in Corollary 3. RGO implementations in more general setups, where the potential function f can be nonsmooth, weakly-smooth, and even nonconvex, have been extensively studied in [28], [43], [44], and [74]. In these general settings, related proximal optimization problems are also efficiently solved within the RGO implementations.

Rejection Sampling: Suppose $\pi \propto \exp(-V)$ on $\mathbb{R}^d$ is β-strongly log-concave and M-smooth. The rejection sampling method to sample from π is the following:

1) Compute the minimizer x^* of V, so that for any $z \in \mathbb{R}^d$, $V(z) \geq V(x^*) + \frac{\beta}{2}\|z - x^*\|^2$.
2) Draw $Z \sim \mathcal{N}\left(x^*, \frac{1}{\beta} I\right)$ and accept it with probability

$$\exp\left(-V(Z) + V(x^*) + \frac{\beta}{2}\|Z - x^*\|^2\right).$$

Repeat this until acceptance.

The output of this method is distributed according to π and the expected number of iterations is $\left(\frac{M}{\beta}\right)^{d/2}$ [76, Theorem 7].

We can use rejection sampling to implement a RGO as follows. Recall the conditional distribution we seek to sample from (11):

$$\nu^{X|Y}(x \mid y) \propto_x \exp\left(-f(x) - \frac{\|x - y\|^2}{2\eta}\right).$$

Define $g_y(x) := f(x) + \frac{\|x-y\|^2}{2\eta}$ so that for any fixed $y \in \mathbb{R}^d$, the target distribution for the RGO is $\tilde{\nu}_y(x) \propto \exp(-g_y(x))$. Suppose the potential function f is L-smooth and that $\eta < \frac{1}{L}$. In this case, $\tilde{\nu}_y$ is strongly log-concave with condition number $\frac{1+L\eta}{1-L\eta}$.

Using rejection sampling to sample from $\tilde{\nu}_y$ with $\eta \asymp \frac{1}{Ld}$ gives a valid implementation of the RGO under smoothness of f with $\mathcal{O}(1)$ many iterations in expectation. Specifically, $M = L + \frac{1}{\eta}$ and $\beta = -L + \frac{1}{\eta}$ and therefore, $\frac{M}{\beta} = \frac{1+L\eta}{1-L\eta}$. So if $\eta = \frac{1}{Ld}$, $\left(\frac{M}{\beta}\right)^{d/2} = \left(1 + \frac{2}{d-1}\right)^{d/2} = \mathcal{O}(1)$.

We therefore have the following corollary describing the oracle complexity of the Proximal Sampler with a rejection sampling-based RGO. For simplicity, in Corollary 3 we assume ρ_1^X satisfies a Φ-Sobolev inequality with optimal constant $\alpha_{\Phi\mathsf{SI}}(\rho_1^X)$ as opposed to some ρ_ℓ^X for $\ell \geq 1$ as we state in Theorem 5. Changing this only corresponds to an additive constant ℓ term in the bound below.

Corollary 3: Suppose $\nu^X \propto \exp(-f)$ is α-strongly log-concave and L-smooth. Let the joint law of (X_i, X_j) be $\rho_{i,j}^X$ and suppose ρ_1^X satisfies a Φ-Sobolev inequality with optimal constant $\alpha_{\Phi\mathsf{SI}}(\rho_1^X)$. Then for any $\epsilon > 0$, the Proximal Sampler (10) with $\eta \asymp \frac{1}{Ld}$ and with a rejection sampling-based RGO implementation (as described above) outputs $X_k \sim \rho_k^X$ with $\mathsf{MI}_\Phi(\rho_{0,k}^X) \leq \epsilon$ as long as $k \geq \frac{Ld}{2\min\{\alpha, \alpha_{\Phi\mathsf{SI}}(\rho_1^X)\}} \log \frac{\mathsf{MI}_\Phi(\rho_{0,1}^X)}{\epsilon}$. The expected number of oracle calls to ∇f is

$$\mathcal{O}\left(\frac{Ld}{\min\{\alpha, \alpha_{\Phi\mathsf{SI}}(\rho_1^X)\}} \log \frac{\mathsf{MI}_\Phi(\rho_{0,1}^X)}{\epsilon}\right).$$

Appendix H
Rate of Change of Divergence Between Simultaneous Evolutions

Here we describe the rate of change of Φ-divergence along simultaneous evolutions of the same SDE. The following lemma is crucial in the analyses of all Langevin-type Markov chains considered in this paper. It is identical to [24, Lemma 8] and presented below for completeness. Similar results can be found in [26, Lemmas 12 and 15] and [12, Theorem 8.3.1].

Lemma 12: Suppose $X_t \sim \mu_t$ and $X_t \sim \nu_t$ with initial conditions μ_0 and ν_0 are two solutions of the following SDE:

$$\mathrm{d}X_t = b_t(X_t)\,\mathrm{d}t + \sqrt{2c}\,\mathrm{d}W_t\,, \tag{29}$$

where $b_t : \mathbb{R}^d \to \mathbb{R}^d$ is a time-varying drift function, c is a positive constant, and W_t is the standard Brownian motion on $\mathbb{R}^d$. Then for all $t \geq 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{D}_\Phi(\mu_t \,\|\, \nu_t) = -c\,\mathsf{FI}_\Phi(\mu_t \,\|\, \nu_t).$$

Proof: Begin by recalling that if $X_t \sim \rho_t$ where $\mathrm{d}X_t = b_t(X_t)\,\mathrm{d}t + \sqrt{2c}\,\mathrm{d}W_t$, then $\rho_t : \mathbb{R}^d \to \mathbb{R}$ satisfies the Fokker-Planck equation, given by:

$$\partial_t \rho_t = -\nabla \cdot (\rho_t\, b_t) + c\Delta\rho_t\,.$$

Also note that using the identity $\Delta\rho = \nabla\cdot(\rho\nabla\log\rho)$, the above can be written as:

$$\partial_t \rho_t = -\nabla \cdot (\rho_t\, b_t) + c\nabla\cdot(\rho_t \nabla \log \rho_t)\,. \tag{30}$$

We identify μ_t and ν_t with their densities with respect to Lebesgue measure, and further denote their relative density as $h_t = \frac{\mu_t}{\nu_t}$. We also assume enough regularity to take the differential under the integral sign and use $\int fg$ as a shorthand for $\int f(x)g(x)\,\mathrm{d}x$. Throughout the proof, we use integration by parts in various steps, denoted by (IBP). With all of this in mind, we have the following:

$$\begin{aligned}
&\partial_t\, \mathsf{D}_\Phi(\mu_t \,\|\, \nu_t)\\
&\quad= \partial_t \int \nu_t\, \Phi(h_t)\\
&\quad= \int (\partial_t \nu_t)\Phi(h_t) + \int \nu_t \left(\partial_t\, \Phi(h_t)\right)\\
&\quad= \int (\partial_t \nu_t)\Phi(h_t) + \int \nu_t \Phi'(h_t)\frac{\nu_t\,\partial_t\mu_t - \mu_t\,\partial_t\nu_t}{\nu_t^2}\\
&\quad= \underbrace{\int (\partial_t \nu_t)\Phi(h_t)}_{T_1} + \underbrace{\int \Phi'(h_t)(\partial_t\mu_t)}_{T_2} - \underbrace{\int \Phi'(h_t)\frac{\mu_t}{\nu_t}(\partial_t\nu_t)}_{T_3}.
\end{aligned}$$

We will now handle each of these terms separately. We have:

$$T_1 = \int (\partial_t \nu_t)\Phi(h_t)$$

$$\overset{(30)}{=} \int \left(-\nabla\cdot(\nu_t b_t) + c\nabla\cdot(\nu_t \nabla \log \nu_t)\right)\Phi(h_t)$$
$$= -\int \nabla\cdot(\nu_t b_t)\Phi(h_t) + c\int \nabla\cdot(\nu_t\nabla\log\nu_t)\Phi(h_t)$$
$$\overset{(IBP)}{=} \int \langle \nu_t b_t, \nabla(\Phi(h_t))\rangle - c\int \langle \nu_t\nabla\log\nu_t, \nabla(\Phi(h_t))\rangle$$
$$= \int \langle \nu_t b_t, \Phi'(h_t)\nabla\frac{\mu_t}{\nu_t}\rangle - c\int \langle \nu_t\nabla\log\nu_t, \Phi'(h_t)\nabla\frac{\mu_t}{\nu_t}\rangle .$$

We also have:

$$T_2 = \int \Phi'(h_t)(\partial_t\mu_t)$$
$$\overset{(30)}{=} \int \Phi'(h_t)\left(-\nabla\cdot(\mu_t b_t) + c\nabla\cdot(\mu_t\nabla\log\mu_t)\right)$$
$$\overset{(IBP)}{=} \int \langle \mu_t b_t, \nabla(\Phi'(h_t))\rangle - c\int \langle \mu_t\nabla\log\mu_t, \nabla(\Phi'(h_t))\rangle$$
$$= \int \langle \mu_t b_t, \Phi''(h_t)\nabla\frac{\mu_t}{\nu_t}\rangle - c\int \langle \mu_t\nabla\log\mu_t, \Phi''(h_t)\nabla\frac{\mu_t}{\nu_t}\rangle .$$

We also have:

$$T_3$$
$$= \int \Phi'(h_t)\frac{\mu_t}{\nu_t}(\partial_t\nu_t)$$
$$\overset{(30)}{=} \int \Phi'(h_t)\frac{\mu_t}{\nu_t}\left(-\nabla\cdot(\nu_t b_t) + c\nabla\cdot(\nu_t\nabla\log\nu_t)\right)$$
$$\overset{(IBP)}{=} \int \left\langle \nu_t b_t, \nabla\left(\Phi'(h_t)\frac{\mu_t}{\nu_t}\right)\right\rangle$$
$$-c\int \left\langle \nu_t\nabla\log\nu_t, \nabla\left(\Phi'(h_t)\frac{\mu_t}{\nu_t}\right)\right\rangle$$
$$= \int \langle \mu_t b_t, \Phi''(h_t)\nabla\frac{\mu_t}{\nu_t}\rangle + \int \langle \nu_t b_t, \Phi'(h_t)\nabla\frac{\mu_t}{\nu_t}\rangle$$
$$- c\int \left\langle \mu_t\nabla\log\nu_t, \Phi''(h_t)\nabla\frac{\mu_t}{\nu_t}\right\rangle$$
$$-c\int \left\langle \nu_t\nabla\log\nu_t, \Phi'(h_t)\nabla\frac{\mu_t}{\nu_t}\right\rangle .$$

Therefore, combining the above, we see many terms cancel and we have the following:

$$\partial_t \mathsf{D}_\Phi(\mu_t\,\|\,\nu_t) = T_1 + T_2 - T_3$$
$$= c\int \left\langle \mu_t\nabla\log\nu_t, \Phi''(h_t)\nabla\frac{\mu_t}{\nu_t}\right\rangle$$
$$- c\int \left\langle \mu_t\nabla\log\mu_t, \Phi''(h_t)\nabla\frac{\mu_t}{\nu_t}\right\rangle$$
$$= -c\int \left\langle \nabla\log\frac{\mu_t}{\nu_t}, \Phi''(h_t)\nabla\frac{\mu_t}{\nu_t}\right\rangle \mu_t$$
$$= -c\,\mathop{\mathbb{E}}_{\mu_t}\left[\left\langle \nabla\log\frac{\mu_t}{\nu_t}, \Phi''(h_t)\nabla\frac{\mu_t}{\nu_t}\right\rangle\right]$$
$$= -c\,\mathop{\mathbb{E}}_{\nu_t}\left[\frac{\mu_t}{\nu_t}\left\langle \nabla\log\frac{\mu_t}{\nu_t}, \Phi''(h_t)\nabla\frac{\mu_t}{\nu_t}\right\rangle\right]$$
$$= -c\,\mathop{\mathbb{E}}_{\nu_t}\left[\left\langle \nabla\frac{\mu_t}{\nu_t}, \Phi''(h_t)\nabla\frac{\mu_t}{\nu_t}\right\rangle\right]$$
$$= -c\,\mathsf{FI}_\Phi(\mu_t\,\|\,\nu_t),$$

which proves the desired statement. □

Appendix I
From Mixing to Independence for Mutual Information

We show that in the special case of $\Phi(x) = x\log x$, i.e., the standard mutual information, the contraction of mutual information can be bounded in terms of the mixing time in KL divergence. Let $X_i \sim \rho_i$ for $i \geq 0$ be iterates along a Markov chain $\mathbf{P}$ starting from $X_0 \sim \rho_0$ and denote the joint law of (X_0, X_k) by $\rho_{0,k}$. If $\mathbf{P}$ mixes in KL divergence from initial distributions taken to be point masses, i.e., δ_x for all $x \in \mathbb{R}^d$, then $\mathsf{MI}(\rho_{0,k})$ decreases at the same rate as the KL divergence to the stationary distribution.

Lemma 13: Let $\mathbf{P}$ be a Markov chain with stationary distribution ν and iterates $X_i \sim \rho_i$ for $i \geq 0$. Further denote the joint law of (X_i, X_j) to be $\rho_{i.j}$. Given any error threshold $\epsilon \geq 0$, suppose there exists $k \in \mathbb{N}$ such that $\mathsf{KL}(\delta_x\mathbf{P}^k \,\|\, \nu) \leq \epsilon$ for all $x \in \mathbb{R}^d$. Then, we have

$$\mathsf{MI}(\rho_{0,k}) \leq \epsilon .$$

Proof: Let $X_0 \sim \rho_0$. For any $i \geq 1$, we have $X_i \sim \rho_i = \rho_0\mathbf{P}^i$. It follows from Definition 2 and direct calculation that

$$\mathsf{MI}(\rho_{0,k}) \overset{(3)}{=} \mathop{\mathbb{E}}_{x\sim\rho_0}[\mathsf{KL}(\rho_{k|0=x}\,\|\,\rho_k)] = \mathop{\mathbb{E}}_{x\sim\rho_0}[\mathsf{KL}(\delta_x\mathbf{P}^k\,\|\,\rho_k)]$$
$$= \mathop{\mathbb{E}}_{x\sim\rho_0}\left[\mathop{\mathbb{E}}_{\delta_x\mathbf{P}^k}\left[\log\frac{\delta_x\mathbf{P}^k}{\rho_k}\right]\right]$$
$$= \mathop{\mathbb{E}}_{x\sim\rho_0}\left[\mathop{\mathbb{E}}_{\delta_x\mathbf{P}^k}\left[\log\frac{\delta_x\mathbf{P}^k}{\nu} - \log\frac{\rho_k}{\nu}\right]\right]$$
$$= \mathop{\mathbb{E}}_{x\sim\rho_0}[\mathsf{KL}(\delta_x\mathbf{P}^k\,\|\,\nu)] - \mathop{\mathbb{E}}_{\rho_k}\left[\log\frac{\rho_k}{\nu}\right]$$
$$= \mathop{\mathbb{E}}_{x\sim\rho_0}[\mathsf{KL}(\delta_x\mathbf{P}^k\,\|\,\nu)] - \mathsf{KL}(\rho_k\,\|\,\nu)$$
$$\leq \mathop{\mathbb{E}}_{x\sim\rho_0}[\mathsf{KL}(\delta_x\mathbf{P}^k\,\|\,\nu)]$$
$$\leq \epsilon ,$$

where the first inequality is by the non-negativity of KL divergence and the second inequality is due to the assumption that $\mathsf{KL}(\delta_x\mathbf{P}^k\,\|\,\nu) \leq \epsilon$ for all$x \in \mathbb{R}^d$. □

We emphasize that the proof of Lemma 13 does not extend to general Φ-divergences and Φ-mutual informations. The key step in the proof of Lemma 13, of expressing

$$\mathop{\mathbb{E}}_{x\sim\rho_0}[\mathsf{KL}(\delta_x\mathbf{P}^k\,\|\,\rho_k)] = \mathop{\mathbb{E}}_{x\sim\rho_0}[\mathsf{KL}(\delta_x\mathbf{P}^k\,\|\,\nu)] - \mathsf{KL}(\rho_k\,\|\,\nu), \quad (31)$$

need not hold for general Φ-divergences. Indeed, (31) can alternatively be derived from the three-point identity of Bregman divergence. It is known that KL divergence is the Bregman divergence of the entropy functional $\mathsf{H}(\rho) = -\mathbb{E}_\rho[\log\rho]$, that is $\mathsf{KL}(\rho\|\nu) = D_{-\mathsf{H}}(\rho;\nu)$. So, KL divergence is special as it is both a Φ-divergence and a Bregman divergence. In general, a Φ-divergence need not be a Bregman divergence.

We would also like to mention that the requirement of having $\mathbf{P}$ to mix from Dirac initializations is common for discrete-space chains [11] but non-trivial in the setting of continuous-space Markov chains. Mixing guarantees usually have dependence on the initial divergence to the target distribution, which becomes undefined in continuous-space (for

example in KL divergence) when the initial distribution is not absolutely continuous with respect to the stationary distribution. Hence most mixing time guarantees for Langevin-type Markov chains implicitly require some regularity of the initial distribution. Challenges pertaining to the regularity of the initial distribution are prevalent in continuous-space sampling [77], [78], with corresponding results therefore restricted to TV distance or Wasserstein metric, which remain finite for Dirac distributions.

Appendix J
Strong Data Processing Inequalities

Recall the discussion on Strong Data Processing Inequalities (SDPIs) in Section I-A. Here, we provide a more comprehensive introduction to these inequalities. When studying SDPIs in Φ-divergence, the key quantity to bound is the contraction coefficient in Φ-divergence (1). This is formally defined as follows.

Definition 4: Let π be a probability distribution and $\mathbf{P}$ be a Markov kernel. Then the D_Φ-contraction coefficient $\varepsilon_{\mathsf{D}_\Phi}$ is defined as follows

$$\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P},\pi) := \sup_{\mu\,:\,0<\mathsf{D}_\Phi(\mu\|\pi)<\infty} \frac{\mathsf{D}_\Phi(\mu\mathbf{P}\,\|\,\pi\mathbf{P})}{\mathsf{D}_\Phi(\mu\,\|\,\pi)}\,. \tag{32}$$

The fact that (32) is ≤ 1 is guaranteed by the data processing inequality. We say that $(\mathbf{P},\pi)$ satisfies an SDPI in Φ-divergence when $\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P},\pi)<1$. Using Definition 4, we state the definition of an SDPI in Φ-divergence.

Definition 5: Let π be a probability distribution and $\mathbf{P}$ be a Markov kernel, and further define $\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P},\pi)$ as in (32). Then we say that $(\mathbf{P},\pi)$ satisfies a strong data processing inequality in Φ-divergence when $\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P},\pi)<1$. In particular, we have,

$$\mathsf{D}_\Phi(\mu\mathbf{P}\,\|\,\pi\mathbf{P}) \le \varepsilon_{\mathsf{D}_\Phi}(\mathbf{P},\pi)\mathsf{D}_\Phi(\mu\,\|\,\pi)\,, \tag{33}$$

where μ is any distribution such that $\mathsf{D}_\Phi(\mu\,\|\,\pi)<\infty$.

Observe that SDPIs immediately yield mixing guarantees for the Markov chain. Suppose ν is stationary for $\mathbf{P}$, i.e., $\nu\mathbf{P}=\nu$, and μ is the initial distribution for the Markov chain. Then repeated application of (33) yields that

$$\mathsf{D}_\Phi(\mu\mathbf{P}^k\,\|\,\nu) \le \varepsilon^k_{\mathsf{D}_\Phi}(\mathbf{P},\nu)\,\mathsf{D}_\Phi(\mu\,\|\,\nu)\,. \tag{34}$$

Therefore, so long as $\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P},\nu)<1$ (where ν is stationary for $\mathbf{P}$), we can get quantitative mixing time bounds for $\mathbf{P}$ in Φ-divergence. Bounding the contraction coefficient $\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P},\nu)$ is the key challenge in SDPI-based mixing time analyses.

SDPIs can also be stated in terms of information. Suppose $U\to X\to Y$ forms a Markov chain, then the data processing inequality states that $\mathsf{MI}_\Phi(U;Y)\le\mathsf{MI}_\Phi(U;X)$. SDPIs in Φ-mutual information are stated in terms of the contraction coefficient in Φ-mutual information. This is defined as follows.

Definition 6: Let π be a probability distribution and $\mathbf{P}$ be a Markov kernel. Furthermore, let $X\sim\pi$ and $Y\sim\pi\mathbf{P}$. Then the MI_Φ-contraction coefficient $\varepsilon_{\mathsf{MI}_\Phi}$ is defined as follows

$$\varepsilon_{\mathsf{MI}_\Phi}(\mathbf{P},\pi) := \sup_{\rho^{U|X}} \frac{\mathsf{MI}_\Phi(U;Y)}{\mathsf{MI}_\Phi(U;X)}\,, \tag{35}$$

where $U\to X\to Y$ forms a Markov chain and the sup is over all conditional distributions of $U\mid X$, denoted as $\rho^{U|X}$. Another way to think of the sup in (35) is over all Markov chains $U\to X\to Y$ with fixed joint distribution $\rho^{XY}(x,y)=\pi(x)\mathbf{P}(y\,|\,x)$. Using Definition 6, we state SDPIs in Φ-mutual information.

Definition 7: Let π be a probability distribution and $\mathbf{P}$ be a Markov kernel, and further define $\varepsilon_{\mathsf{MI}_\Phi}(\mathbf{P},\pi)$ as in (35). Then we say that $(\mathbf{P},\pi)$ satisfies a strong data processing inequality in Φ-mutual information when $\varepsilon_{\mathsf{MI}_\Phi}(\mathbf{P},\pi)<1$. In particular, we have,

$$\mathsf{MI}_\Phi(U;Y) \le \varepsilon_{\mathsf{MI}_\Phi}(\mathbf{P},\pi)\mathsf{MI}_\Phi(U;X)\,, \tag{36}$$

where $U\to X\to Y$ is any Markov chain where $X\sim\pi$, and $Y\sim\pi\mathbf{P}$.

In Lemma 14, we show that $\varepsilon_{\mathsf{MI}_\Phi}$ can always be upper bounded by $\varepsilon_{\mathsf{D}_\Phi}$. The converse also holds when Φ is bounded in a neighbourhood of 1 [10, Theorem 5.2]). Hence, although we are interested in SDPIs in information, we will bound the contraction coefficients in divergence out of convenience.

Lemma 14: For any probability distribution π and Markov kernel $\mathbf{P}$, we have that

$$\varepsilon_{\mathsf{MI}_\Phi}(\mathbf{P},\pi) \le \varepsilon_{\mathsf{D}_\Phi}(\mathbf{P},\pi)\,.$$

Proof: Consider a Markov chain $U\to X\to Y$ where $U\sim\rho^U$, $X\sim\rho^X=\pi$ and $Y\sim\rho^Y=\pi\mathbf{P}$. It follows from Definitions 2 and 4 that

$$\begin{aligned}
\mathsf{MI}_\Phi(\rho^{UY}) &\overset{(3)}{=} \mathop{\mathbb{E}}_{u\sim\rho^U}\left[\mathsf{D}_\Phi(\rho^{Y|U=u}\,\|\,\rho^Y)\right],\\
&= \mathop{\mathbb{E}}_{u\sim\rho^U}\left[\mathsf{D}_\Phi(\rho^{X|U=u}\mathbf{P}\,\|\,\rho^X\mathbf{P})\right],\\
&\overset{(32)}{\le} \mathop{\mathbb{E}}_{u\sim\rho^U}\left[\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P},\pi)\mathsf{D}_\Phi(\rho^{X|U=u}\,\|\,\rho^X)\right],\\
&\overset{(3)}{=} \varepsilon_{\mathsf{D}_\Phi}(\mathbf{P},\pi)\mathsf{MI}_\Phi(\rho^{UX})\,.
\end{aligned}$$

Hence, the claim immediately follows from Definition 6. □

To study the contraction of Φ-mutual information along multiple steps of a Markov chain, we have the following lemma.

Lemma 15: Let $X_i\sim\rho_i$ be iterates along a Markov chain $\mathbf{P}$ starting from $X_0\sim\rho_0$. Then for any $\ell\ge1$ and $k\ge\ell$

$$\mathsf{MI}_\Phi(X_0;X_k) \le \prod_{i=\ell}^{k-1}\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P},\rho_i)\,\mathsf{MI}_\Phi(X_0;X_\ell)\,.$$

Proof: Applying (36)$k-\ell$ times reveals that

$$\mathsf{MI}_\Phi(X_0;X_k) \le \prod_{i=\ell}^{k-1}\varepsilon_{\mathsf{MI}_\Phi}(\mathbf{P},\rho_i)\,\mathsf{MI}_\Phi(X_0;X_\ell)\,.$$

From Lemma 14 we know that $\varepsilon_{\mathsf{MI}_\Phi}(\mathbf{P},\rho_i)\le\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P},\rho_i)$. Therefore, we get the desired claim. □

Therefore, we can see from Lemma 15 that to compute the contraction of information, we need to compute the contraction coefficients along the trajectory of the Markov chain $\varepsilon_{\mathsf{MI}_\Phi}(\mathbf{P},\rho_i)$ as opposed to just the coefficient for the stationary distribution $\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P},\nu)$ for mixing time (34). This analysis along the trajectory makes studying the information functional more challenging.

APPENDIX K
REGULARITY-BASED BOUNDS FOR THE STANDARD MUTUAL INFORMATION

We now consider the special case of $\Phi(x) = x\log x$. We discuss an alternate set of results studying the convergence of mutual information by exploiting the regularity properties of the Markov chains. Regularity results for a Markov chain seek to bound the change in the initial data after the Markov chain is applied to it. For example for $x, y \in \mathbb{R}^d$, Markov kernel $\mathbf{P}$, and $C > 0$, bounds of the form

$$\mathsf{KL}(\delta_x\mathbf{P} \,\|\, \delta_y\mathbf{P}) \le C\|x-y\|_2^2,$$

arise frequently in the analysis of Hamiltonian Monte Carlo [79] and other Markov chains [80], [81]. Our goal is to study the convergence of the mutual information and we do so by relating it to the KL divergence via $\mathsf{MI}(\rho^{XY}) = \mathbb{E}_{\rho^X}[\mathsf{KL}(\rho^{Y|X} \,\|\, \rho^Y)]$ (Definition 2) and using regularity bounds for these Markov chains in KL divergence. Before proceeding, we define the Wasserstein-2 distance, which arises in the following analyses.

Definition 8: The $\mathcal{W}_2$ distance between probability distributions μ and ν is given by

$$\mathcal{W}_2(\mu,\nu) := \inf_{\gamma\in\mathcal{C}(\mu,\nu)} \left(\mathop{\mathbb{E}}_{(x,y)\sim\gamma}\left[\|x-y\|_2^2\right]\right)^{1/2},$$

where $\mathcal{C}(\mu,\nu)$ denotes the set of couplings of μ and ν.

We study the Langevin dynamics, ULA, and Proximal Sampler in Appendices K-A, K-B, and K-C respectively. In Appendix K-D we present Theorem 7 describing the regularity properties of the Langevin dynamics for strongly log-concave targets. We de not utilize Theorem 7 when studying mutual information but include it as it might be of independent interest.

A. Regularity-Based Bounds for Mutual Information Along Langevin Dynamics

Theorem 2 describes the decay of mutual information along the Langevin dynamics as a consequence of the regularity properties of the dynamics. Observe that Theorem 2 does not require a Φ-Sobolev inequality assumption on ρ_s and also works for the weakly log-concave case. Theorem 2(b) with $t = s$ can be used to bound $\mathsf{MI}(\rho_{0,s})$ from Theorem 1 (for the case of $\Phi(x) = x\log x$). We now prove Theorem 2.

Proof of Theorem 2: We first consider case when ν is α-strongly log-concave as the weakly log-concave case follows by taking the limit $\alpha \to 0$. Let $\mathbf{P}_t$ be the Markov kernel associated with the Langevin dynamics. Note that $\rho_{t|0}(\cdot \mid x_0) = \delta_{x_0}\mathbf{P}_t$ is the law at time t if we start the Langevin dynamics from initial distribution δ_{x_0}, for any fixed $x_0 \in \mathbb{R}^d$. It then follows from [81, Corollary 1][1] that

$$\begin{aligned}\mathsf{KL}(\delta_{x_0}\mathbf{P}_t \,\|\, \rho_0\mathbf{P}_t) &\le \frac{\alpha}{2(e^{2\alpha t}-1)}\mathcal{W}_2^2(\delta_{x_0},\rho_0)\\ &= \frac{\alpha}{2(e^{2\alpha t}-1)}\mathop{\mathbb{E}}_{X\sim\rho_0}\left[\|X-x_0\|^2\right].\end{aligned}$$

[1]Also note that [80, Eq. (4.5)] is equivalent to [81, Corollary 1] via the duality between log-Harnack and reverse transport inequalities as explained in [81, Section VI.B].

Using Definition 2 and the above inequality, we have

$$\begin{aligned}\mathsf{MI}(\rho_{0,t}) &= \mathop{\mathbb{E}}_{x_0\sim\rho_0}[\mathsf{KL}(\delta_{x_0}\mathbf{P}_t \,\|\, \rho_0\mathbf{P}_t)]\\ &\le \frac{\alpha}{2(e^{2\alpha t}-1)}\mathop{\mathbb{E}}_{X,x_0\sim\rho_0}\left[\|X-x_0\|^2\right]\\ &= \frac{\alpha}{e^{2\alpha t}-1}\mathsf{Var}(X_0),\end{aligned}$$

where in the above, $X, x_0 \sim \rho_0$ are independent, and we have used the variance formula

$$\mathsf{Var}(X_0) = \frac{1}{2}\mathop{\mathbb{E}}_{X,x_0\sim\rho_0}\left[\|X-x_0\|^2\right].$$

Taking $\alpha \to 0$ yields the weakly log-concave result. □

B. Regularity-Based Bounds for Mutual Information Along ULA

In this section we prove Theorem 4, which bounds the mutual information along ULA be leveraging the regularity properties of the ULA kernel. Theorem 4 with $k = \ell$ provides a bound on $\mathsf{MI}(\rho_{0,\ell})$ from Theorem 3 and also does not require ρ_ℓ to satisfy a Φ-Sobolev inequality assumption.

Proof of Theorem 4: The proof is similar to that of Theorem 2. Let $\mathbf{P}^k$ denote the kernel for k-step of ULA. It follows from [81, Theorem 6] with $L = 1-\eta\alpha$ that

$$\begin{aligned}\mathsf{KL}(\delta_{x_0}\mathbf{P}^k \,\|\, \rho_0\mathbf{P}^k) &\le \frac{1-L^2}{4\eta(L^{-2k}-1)}\mathcal{W}_2^2(\delta_{x_0},\rho_0)\\ &\le \frac{\alpha}{2[(1-\eta\alpha)^{-2k}-1]}\mathop{\mathbb{E}}_{X\sim\rho_0}\left[\|X-x_0\|^2\right].\end{aligned}$$

Using Definition 2 and the above inequality, we have

$$\begin{aligned}\mathsf{MI}(\rho_{0,k}) &= \mathop{\mathbb{E}}_{x_0\sim\rho_0}[\mathsf{KL}(\delta_{x_0}\mathbf{P}^k \,\|\, \rho_0\mathbf{P}^k)]\\ &\le \frac{\alpha}{2[(1-\eta\alpha)^{-2k}-1]}\mathop{\mathbb{E}}_{X,x_0\sim\rho_0}\left[\|X-x_0\|^2\right]\\ &= \frac{\alpha}{[(1-\eta\alpha)^{-2k}-1]}\mathsf{Var}(X_0),\end{aligned}$$

where in the above, $X, x_0 \sim \rho_0$ are independent, and we use the variance formula $\mathsf{Var}(X_0) = \frac{1}{2}\mathbb{E}_{X,x_0\sim\rho_0}\left[\|X-x_0\|^2\right]$. □

C. Regularity-Based One-Step Bounds for the Proximal Sampler

We wish to bound $\mathsf{MI}(\rho_{0,1}^X)$ appearing in Theorem 5 when $\ell = 1$.

Lemma 16: Consider the Proximal Sampler starting from $X_0 \sim \rho_0^X$ with step-size $\eta > 0$. Then we have that

$$\mathsf{MI}(\rho_{0,1}^X) \le \frac{1}{2\eta}\mathsf{Var}(X_0).$$

Proof: Note that $\mathsf{MI}(\rho_{0,1}^X) = \mathsf{MI}(X_0, X_1) \le \mathsf{MI}(X_0; Y_0)$ due to the data processing inequality. By the Proximal Sampler (10), we know that the law of $Y_0 \mid X_0$ is obtained via evolving ρ_0^X along the heat flow for time η. Given this, it follows from Theorem 2(a) with $t = \eta$ that$\mathsf{MI}(X_0; Y_0) \le \frac{1}{2\eta}\mathsf{Var}(X_0)$. □

This yields the following corollary.

Corollary 4: Consider the same conditions as Theorem 5 with $\ell = 1$, i.e., suppose ρ_1^X satisfies a Φ-Sobolev inequality with optimal constant $\alpha_{\Phi\mathsf{SI}}(\rho_1)$. Then

$$\mathsf{MI}(\rho_{0,k}^X) \le \frac{\mathsf{Var}(X_0)}{2\eta(1+\eta\min\{\alpha, \alpha_{\Phi\mathsf{SI}}(\rho_1^X)\})^{2(k-1)}}.$$

D. Regularity of Langevin Dynamics

So far in this appendix we have studied bounds on the mutual information via the regularity properties of the Langevin dynamics (Appendix K-A), the Unadjusted Langevin Algorithm (Appendix K-B), and the Proximal Sampler (Appendix K-C). We now focus on the continuous-time Langevin dynamics.

Recall from the proof of Theorem 2 (in Appendix K-A) how the regularity properties of the Langevin dynamics (as studied in [81, Corollary 1] for strongly log-concave targets and simultaneous Dirac initializations) result in guarantees for the contraction of mutual information. In this appendix we provide an alternate regularity theorem for the Langevin dynamics (Theorem 7) for strongly log-concave targets and smooth initializations. Although we do not utilize Theorem 7 to study mutual information contraction, we include it here as it might be of independent interest.

Before describing Theorem 7 we state results from [82] which we build upon and which are for the log-concave case.

Theorem 6 ([82, Theorem & Corollary 2]): Let ν be a log-concave distribution and let $X_t \sim \rho_t$ evolve along the Langevin dynamics (7) from any $X_0 \sim \rho_0$ such that ρ_0 is smooth and $\mathcal{W}_2(\rho_0, \nu) < \infty$. Then, for any $t > 0$, we have

$$t^2\, \mathsf{FI}(\rho_t \,\|\, \nu) + 2t\, \mathsf{KL}(\rho_t \,\|\, \nu) + \mathcal{W}_2^2(\rho_t, \nu) \;\le\; \mathcal{W}_2^2(\rho_0, \nu)\,.$$

In particular, for any $t > 0$,

$$\mathsf{KL}(\rho_t \,\|\, \nu) \le \frac{\mathcal{W}_2^2(\rho_0, \nu)}{4t}\,. \tag{37}$$

Although the smoothness assumption on ρ_0 may be restrictive, this bound is useful in many settings where $\mathsf{KL}(\rho_0\|\nu) = \infty$ but $\mathcal{W}_2(\rho_0, \nu)$ is finite, such as the case where ρ_0 is Cauchy, and ν is a Gaussian. We extend the analysis of [82] to SLC target in Theorem 7. Note that this provides an alternate proof to [80, Lemma 4.2].

Theorem 7: Let ν be α-SLC for some $\alpha > 0$ and let $X_t \sim \rho_t$ evolve along the Langevin dynamics (7) from any $X_0 \sim \rho_0$ such that ρ_0 is smooth and $\mathcal{W}_2(\rho_0, \nu) < \infty$. Then, for any $t > 0$, we have

$$\frac{(e^{\alpha t} - 1)^2}{\alpha^2}\mathsf{FI}(\rho_t \,\|\, \nu) + \frac{2(e^{\alpha t} - 1)}{\alpha}\mathsf{KL}(\rho_t \,\|\, \nu) + e^{\alpha t}\mathcal{W}_2^2(\rho_t, \nu) \le \mathcal{W}_2^2(\rho_0, \nu)\,. \tag{38}$$

In particular, for any $t > 0$,

$$\mathsf{KL}(\rho_t \,\|\, \nu) \le \frac{\alpha}{2(e^{2\alpha t} - 1)}\mathcal{W}_2^2(\rho_0, \nu)\,. \tag{39}$$

Proof: Similar to the proof of Theorem 6 given in [82], we seek to find a Lyapunov functional of the form

$$\psi(t) = A_t\, \mathsf{FI}(\rho_t \,\|\, \nu) + B_t\, \mathsf{KL}(\rho_t \,\|\, \nu) + \mathcal{W}_2^2(\rho_t, \nu). \tag{40}$$

for some $A_t, B_t \ge 0$ that will be determined later to ensure that $\psi(t)$ is decreasing exponentially fast. From Lemma 12 for the Langevin dynamics (i.e., with $\mu_t = \rho_t$, $\nu_t = \nu$, $b_t(x) = -\nabla f(x)$ and $c = 1$), we get that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t \,\|\, \nu) = -\mathsf{FI}(\rho_t \,\|\, \nu)\,. \tag{41}$$

Recalling [17, Formula 15.7] [68, eq.(10)-(12)], we can characterize the rate of change of relative Fisher information,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{FI}(\rho_t \,\|\, \nu) &= -2\mathcal{K}_\nu(\rho_t) - 2\mathop{\mathbb{E}}_{\rho_t}\left[\left\langle \nabla \log \frac{\rho_t}{\nu}, (\nabla^2 f)\nabla \log \frac{\rho_t}{\nu}\right\rangle\right] \\ &\le -2\alpha \mathop{\mathbb{E}}_{\rho_t}\left[\left\|\nabla \log \frac{\rho_t}{\nu}\right\|^2\right] = -2\alpha\mathsf{FI}(\rho_t \,\|\, \nu).\end{aligned} \tag{42}$$

where $\mathcal{K}_\nu(\rho) := \mathbb{E}_\rho\left[\left\|\nabla^2 \log \frac{\rho}{\nu}\right\|_{\mathrm{HS}}^2\right]$ is the second-order relative Fisher information and the inequality follows from the facts that $\mathcal{K}_\nu(\rho) \ge 0$ and ν is α-SLC, i.e., $\nabla^2 f \ge \alpha I$. Also, recall the lemma from [82], which shows that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\bigg|^+ \mathcal{W}_2^2(\rho_t, \nu) &\le -2\mathop{\mathbb{E}}_{\rho_t}\left[\left\langle x - \nabla\varphi_t(x),\, \nabla \log \frac{\rho_t(x)}{\nu(x)}\right\rangle\right] \\ &\le -2\mathsf{KL}(\rho_t \,\|\, \nu) - \alpha\mathcal{W}_2^2(\rho_t, \nu)\end{aligned} \tag{43}$$

where $(\mathrm{d}/\mathrm{d}t)^+$ is the upper derivative, and $\nabla\varphi_t$ is the (unique) gradient of the convex function that pushes forward ρ_t to ν, i.e., $\nabla\varphi_t \# \rho_t = \nu$. The second inequality above follows since ν is α-SLC, which implies that relative entropy $\rho \mapsto \mathsf{KL}(\rho \,\|\, \nu)$ is α-strongly convex, which further means that

$$\begin{aligned}\mathsf{KL}(\nu \,\|\, \nu) - \mathsf{KL}(\rho_t \,\|\, \nu) - \mathop{\mathbb{E}}_{\rho_t}\left[\left\langle \nabla\varphi_t(x) - x,\, \nabla \log \frac{\rho_t(x)}{\nu(x)}\right\rangle\right] \\ \ge \frac{\alpha}{2}\mathcal{W}_2^2(\rho_t, \nu)\,.\end{aligned} \tag{44}$$

Combining bounds (41), (42), and (43), along with the Lyapunov functional (40), we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\bigg|^+ \psi(t) &= (\dot{A}_t - B_t)\mathsf{FI}(\rho_t \,\|\, \nu) + A_t\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{FI}(\rho_t \,\|\, \nu) \\ &\quad + \dot{B}_t\mathsf{KL}(\rho_t \,\|\, \nu) + \frac{\mathrm{d}}{\mathrm{d}t}\bigg|^+ \mathcal{W}_2^2(\rho_t, \nu) \\ &\le -\left(2\alpha A_t + B_t - \dot{A}_t\right)\mathsf{FI}(\rho_t \,\|\, \nu) \\ &\quad - (2 - \dot{B}_t)\,\mathsf{KL}(\rho_t \,\|\, \nu) - \alpha\mathcal{W}_2^2(\rho_t, \nu)\,.\end{aligned}$$

We want to choose A_t and B_t so that the Lyapunov functional decays exponentially fast with rate α along the Langevin dynamics. To this end, we set $2 - \dot{B}_t = \alpha B_t$ with $B_0 = 0$, which implies that

$$B_t = \frac{2(1 - e^{-\alpha t})}{\alpha}\,.$$

We also set $2\alpha A_t + B_t - \dot{A}_t = \alpha A_t$ with $A_0 = 0$, which can be solved to yield

$$A_t = \frac{1}{\alpha^2}(e^{\alpha t} + e^{-\alpha t} - 2) = \frac{e^{\alpha t}(1 - e^{-\alpha t})^2}{\alpha^2}\,.$$

With these choices, we have $\frac{\mathrm{d}}{\mathrm{d}t}|^+\psi(t) \le -\alpha\psi(t)$, so indeed,

$$\psi(t) \le e^{-\alpha t}\psi(0) = e^{-\alpha t}\mathcal{W}_2^2(\rho_0, \nu)\,, \tag{45}$$

which, after proper normalization, is the claim in (38).

Next, we prove (39). Inequality (44) together with the Cauchy-Schwarz inequality actually implies the HWI inequality [71, Section 9.4],

$$\sqrt{\mathsf{FI}(\rho_t \,\|\, \nu)}\,\mathcal{W}_2(\rho_t, \nu) \;\ge\; \mathsf{KL}(\rho_t \,\|\, \nu) + \frac{\alpha}{2}\mathcal{W}_2^2(\rho_t, \nu)\,.$$

Using the fact that $a^2+b^2 \geq 2ab$ and the above relation, for any $C_t \geq 0$, we have

$$\begin{aligned} A_t \,\mathsf{FI}(\rho_t \,\|\, \nu) + C_t \,\mathcal{W}_2(\rho_t,\nu)^2 &\geq 2\,\sqrt{A_t C_t}\;\sqrt{\mathsf{FI}(\rho_t \,\|\, \nu)}\,\mathcal{W}_2(\rho_t,\nu) \\ &\geq 2\,\sqrt{A_t C_t}\,\left(\mathsf{KL}(\rho_t \,\|\, \nu) + \frac{\alpha}{2}\mathcal{W}_2^2(\rho_t,\nu)\right). \end{aligned}$$

We can then decompose and bound the Lyapunov functional $\psi(t)$ as follows

$$\begin{aligned} \psi(t) &= A_t \,\mathsf{FI}(\rho_t \,\|\, \nu) + C_t \,\mathcal{W}_2^2(\rho_t,\nu) + B_t \,\mathsf{KL}(\rho_t \,\|\, \nu) \\ &\quad + (1-C_t)\,\mathcal{W}_2^2(\rho_t,\nu) \\ &\geq \left(B_t + 2\,\sqrt{A_t C_t}\right)\,\mathsf{KL}(\rho_t \,\|\, \nu) \\ &\quad + \left(1 - C_t + \alpha\,\sqrt{A_t C_t}\right)\,\mathcal{W}_2^2(\rho_t,\nu). \end{aligned} \tag{46}$$

We now choose C_t such that $1 - C_t + \alpha\,\sqrt{A_t C_t} = 0$. Choosing the positive solution, we have

$$\sqrt{C_t} = \frac{\alpha\,\sqrt{A_t} + \sqrt{\alpha^2 A_t + 4}}{2}. \tag{47}$$

This choice of C_t, the previous choices $B_t + \alpha A_t = \dot{A}_t = \frac{1}{\alpha}(e^{\alpha t} - e^{-\alpha t})$ and the subsequent calculation

$$\begin{aligned} A_t(\alpha^2 A_t + 4) &= \frac{e^{\alpha t}(1-e^{-\alpha t})^2}{\alpha^2}\left(e^{\alpha t} + e^{-\alpha t} + 2\right) \\ &= \frac{(e^{\alpha t} - e^{-\alpha t})^2}{\alpha^2}, \end{aligned}$$

yield that

$$B_t + 2\,\sqrt{A_t C_t} \overset{(47)}{=} B_t + \alpha A_t + \sqrt{A_t(\alpha^2 A_t + 4)} = \frac{2}{\alpha}(e^{\alpha t} - e^{-\alpha t}).$$

Thus, with this choice of C_t, we have

$$\psi_t \overset{(46)}{\geq} (B_t + 2\,\sqrt{A_t C_t})\,\mathsf{KL}(\rho_t \,\|\, \nu) = \frac{2}{\alpha}(e^{\alpha t} - e^{-\alpha t})\,\mathsf{KL}(\rho_t \,\|\, \nu).$$

Therefore, we conclude that

$$\begin{aligned} \mathsf{KL}(\rho_t \,\|\, \nu) &\overset{(45)}{\leq} \frac{\alpha}{2(e^{\alpha t} - e^{-\alpha t})}\, e^{-\alpha t}\,\mathcal{W}_2^2(\rho_0,\nu) \\ &= \frac{\alpha}{2(e^{2\alpha t} - 1)}\,\mathcal{W}_2^2(\rho_0,\nu), \end{aligned}$$

as claimed in (39). □

We note that [81, Corollary 1] follows a different proof to obtain a stronger version of Theorem 7, which applies for simultaneous Langevin dynamics and under no assumption on ρ_0. Also note that the dual form of the log-Harnack inequality [80, Equation (4.5)] yields a version of Theorem 7 which is applicable for simultaneous Langevin dynamics and from Dirac initializations. The duality between Harnack inequalities and reverse transport inequalities (such as (39)) is explained in [81, Section VI.B].

Appendix L
Proofs for Langevin Dynamics

Here we present the proofs of and related to Theorem 1. In Appendix L-A we study the evolution of the Φ-Sobolev constant along the Langevin dynamics. This approach is crucial to both of our proof strategies for Theorem 1. In Appendix L-B we present the direct time derivative-based approach for Theorem 1 and in Appendix L-C we discuss the SDPI approach.

A. Evolution of Φ-Sobolev Constant Along Langevin Dynamics

Lemma 17: Suppose $X_t \sim \rho_t$ evolves according to (7) where $\nabla^2 f \geq \alpha I$ with $\alpha > 0$. If ρ_s satisfies Φ-Sobolev inequality with optimal constant $\alpha_{\Phi\mathsf{SI}}(\rho_s)$ where $s \geq 0$. Then for $t \geq s$ we have that

$$\frac{1}{\alpha_{\Phi\mathsf{SI}}(\rho_t)} \leq \frac{e^{-2\alpha(t-s)}}{\alpha_{\Phi\mathsf{SI}}(\rho_s)} + \frac{1 - e^{-2\alpha(t-s)}}{\alpha}. \tag{48}$$

Proof: Consider the forward discretization of (7) (i.e., the ULA (8)) with step-size $\eta > 0$:

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

We will consider this discrete time update and then take the appropriate limit so that we recover the desired results for the continuous time dynamics. If $X_k \sim \rho_k$, then we have that along the ULA,

$$\rho_{k+1} = (I - \eta\nabla f)_{\#}\rho_k * \mathcal{N}(0, 2\eta I).$$

Under $\nabla^2 f \geq \alpha I$, we can see that the map $F(x) = x - \eta\nabla f(x)$ is $(1-\eta\alpha)$-Lipschitz. Using Lemmas 6 and 7, along with the shorthand α_i for $\alpha_{\Phi\mathsf{SI}}(\rho_i)$ and c_i for $1/\alpha_i$, we have that

$$c_{i+1} \overset{(22)}{\leq} \frac{1}{\alpha_{\Phi\mathsf{SI}}((I-\eta\nabla f)_{\#}\rho_i)} + 2\eta \overset{(21)}{\leq} (1-\eta\alpha)^2 c_i + 2\eta.$$

Recursing this from $i = j$ to $i = k$, we get that

$$\begin{aligned} c_{k+1} &\leq (1-\eta\alpha)^{2(k+1-j)} c_j \\ &\quad + 2\eta[1 + (1-\eta\alpha)^2 + \ldots + (1-\eta\alpha)^{2(k-j)}] \\ &= (1-\eta\alpha)^{2(k+1-j)} c_j + 2\eta\left[\frac{1 - (1-\eta\alpha)^{2(k+1-j)}}{\eta\alpha(2-\eta\alpha)}\right]. \end{aligned}$$

Taking $\eta k \to t$, $\eta j \to s$, and $\eta \to 0$, we get that

$$c_t \leq e^{-2\alpha(t-s)} c_s + \frac{1 - e^{-2\alpha(t-s)}}{\alpha},$$

which yields the desired claim. □

B. Proofs for Direct Time Derivative Analysis

1) Proof of Lemma 1:

Proof: Consider Lemma 12 with $\mu_t = \tilde{\rho}_t$, $\nu_t = \rho_t$, $b_t = -\nabla f$, and $c = 1$. Let $\tilde{\rho}_0 = \delta_{x_0}$ for some $x_0 \in \mathbb{R}^d$, so that $\tilde{\rho}_t = \rho_{t|0}(\cdot \mid x_0)$. It follows from Lemma 12 that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{D}_\Phi(\tilde{\rho}_t \,\|\, \rho_t) = -\mathsf{FI}_\Phi(\tilde{\rho}_t \,\|\, \rho_t).$$

It thus follows from Definition 2 and (15) that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathsf{MI}_\Phi(\rho_{0,t}) &= \frac{\mathrm{d}}{\mathrm{d}t}\mathop{\mathbb{E}}_{\rho_0}\left[\mathsf{D}_\Phi(\rho_{t|0} \,\|\, \rho_t)\right] = \mathop{\mathbb{E}}_{\rho_0}\left[\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{D}_\Phi(\rho_{t|0} \,\|\, \rho_t)\right] \\ &= -\mathop{\mathbb{E}}_{\rho_0}[\mathsf{FI}_\Phi(\rho_{t|0} \,\|\, \rho_t)] \overset{(15)}{=} -\mathsf{FI}_\Phi^{\mathsf{M}}(\rho_{0,t}). \end{aligned}$$

We have thus completed the proof. □

2) Proof of Lemma 2:

Proof: As ρ^Y satisfies Φ-Sobolev inequality, it follows from Definition 3 that the following holds for each $x \sim \rho^X$,

$$2\alpha_{\Phi\mathsf{SI}}(\rho^Y)\mathsf{D}_\Phi(\rho^{Y|X=x} \,\|\, \rho^Y) \overset{(4)}{\leq} \mathsf{FI}_\Phi(\rho^{Y|X=x} \,\|\, \rho^Y).$$

Taking expectation over ρ^X and using Definition 2 and (15) completes the proof. □

3) Proof of Theorem 1 (Direct Time Derivative-Based Proof):

Proof: Fix an $s > 0$ such that ρ_s satisfies a Φ-Sobolev inequality with optimal constant $\alpha_{\Phi\mathsf{SI}}(\rho_s)$. Lemma 17 tells us that for any $t \geq s$, we have

$$\begin{aligned}\alpha_{\Phi\mathsf{SI}}(\rho_t) &\geq \frac{\alpha_{\Phi\mathsf{SI}}(\rho_s)\alpha}{e^{-2\alpha(t-s)}(\alpha - \alpha_{\Phi\mathsf{SI}}(\rho_s)) + \alpha_{\Phi\mathsf{SI}}(\rho_s)} \\ &= \frac{1}{e^{-2\alpha(t-s)}(\alpha_{\Phi\mathsf{SI}}(\rho_s)^{-1} - \alpha^{-1}) + \alpha^{-1}}. \end{aligned} \tag{49}$$

Now from Lemmas 1 and 2, we have that,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{MI}_\Phi(\rho_{0,t}) = -\mathsf{FI}^{\mathsf{M}}_\Phi(\rho_{0,t}) \leq -2\alpha_{\Phi\mathsf{SI}}(\rho_t)\,\mathsf{MI}_\Phi(\rho_{0,t}).$$

Integrating the above from s to t, and using (49) gives us

$$\mathsf{MI}_\Phi(\rho_{0,t}) \leq \exp(-2A_t)\,\mathsf{MI}_\Phi(\rho_{0,s}),$$

where

$$A_t := \int_s^t \frac{e^{2\alpha(r-s)}}{\alpha_{\Phi\mathsf{SI}}(\rho_s)^{-1} - \alpha^{-1} + \alpha^{-1}e^{2\alpha(r-s)}}\,\mathrm{d}r.$$

Upon simplifying, we get that

$$\begin{aligned}A_t = \frac{1}{2}\log\left(\frac{e^{-2\alpha(t-s)}(\alpha_{\Phi\mathsf{SI}}(\rho_s)^{-1} - \alpha^{-1}) + \alpha^{-1}}{\alpha_{\Phi\mathsf{SI}}(\rho_s)^{-1}}\right) \\ + \alpha(t-s).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathsf{MI}_\Phi(\rho_{0,t}) &\leq e^{-2A_t}\,\mathsf{MI}_\Phi(\rho_{0,s}) \\ &= \frac{\alpha e^{-2\alpha(t-s)}}{\alpha_{\Phi\mathsf{SI}}(\rho_s)(1-e^{-2\alpha(t-s)}) + \alpha e^{-2\alpha(t-s)}}\,\mathsf{MI}_\Phi(\rho_{0,s}).\end{aligned}$$

Observe that the denominator is a convex combination of α and $\alpha_{\Phi\mathsf{SI}}(\rho_s)$, and so we get that

$$\frac{\alpha}{\alpha_{\Phi\mathsf{SI}}(\rho_s)(1-e^{-2\alpha(t-s)}) + \alpha e^{-2\alpha(t-s)}} \leq \max\left\{1, \frac{\alpha}{\alpha_{\Phi\mathsf{SI}}(\rho_s)}\right\}$$

which proves the desired result. □

C. Proofs for SDPI Analysis for Langevin Dynamics

The key idea to apply the SDPI-based approach is to bound the contraction coefficient for the dynamics. We do so in Lemma 18.

We let $\mathbf{P}_t$ denote the map which takes as input a distribution μ and outputs ρ_t where ρ_t evolves following (7) from $\rho_0 = \mu$.

1) Proof of Contraction Coefficient:

Lemma 18: Let $\mathbf{P}_t$ denote the Markov kernel corresponding to the Langevin dynamics (7) where $\nabla^2 f \geq \alpha I$ with $\alpha > 0$. Then, if ρ satisfies a Φ-Sobolev inequality with optimal constant $\alpha_{\Phi\mathsf{SI}}(\rho)$,

$$\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}_t, \rho) \leq \frac{\alpha e^{-2\alpha t}}{\alpha_{\Phi\mathsf{SI}}(\rho)(1-e^{-2\alpha t}) + \alpha e^{-2\alpha t}}.$$

Proof: Let π be an arbitrary distribution such that $\mathsf{D}_\Phi(\pi \,\|\, \rho) < \infty$. And let $\pi_t = \pi\mathbf{P}_t$ and $\rho_t = \rho\mathbf{P}_t$. Therefore, $\pi_0 = \pi$ and $\rho_0 = \rho$. Recall from Lemma 17 that ρ_t satisfies a Φ-Sobolev inequality with optimal constant $\alpha_{\Phi\mathsf{SI}}(\rho_t)$ as in (48). From Lemma 12 with $\mu_t = \pi_t$, $\nu_t = \rho_t$, $b_t = -\nabla f$, and $c = 1$, along with Definition 3, we have that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{D}_\Phi(\pi_t \,\|\, \rho_t) = -\mathsf{FI}_\Phi(\pi_t \,\|\, \rho_t) \overset{(4)}{\leq} -2\,\alpha_{\Phi\mathsf{SI}}(\rho_t)\mathsf{D}_\Phi(\pi_t \,\|\, \rho_t).$$

Plugging (48) into the above inequality, we have

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{D}_\Phi(\pi_t \,\|\, \rho_t) \\ &\quad\leq \frac{-2\alpha_{\Phi\mathsf{SI}}(\rho_s)\alpha}{e^{-2\alpha(t-s)}\alpha + [1-e^{-2\alpha(t-s)}]\alpha_{\Phi\mathsf{SI}}(\rho_s)}\mathsf{D}_\Phi(\pi_t \,\|\, \rho_t) \\ &\quad= \frac{-2}{e^{-2\alpha s}(\alpha_{\Phi\mathsf{SI}}(\rho_0)^{-1} - \alpha^{-1}) + \alpha^{-1}}\mathsf{D}_\Phi(\pi_t \,\|\, \rho_t)\end{aligned}$$

Applying Grönwall's inequality gives

$$\mathsf{D}_\Phi(\pi_t \,\|\, \rho_t) \leq e^{-2\int_0^t \frac{1}{e^{-2\alpha s}(\alpha_{\Phi\mathsf{SI}}(\rho_0)^{-1} - \alpha^{-1}) + \alpha^{-1}}\,\mathrm{d}s}\mathsf{D}_\Phi(\pi_0 \,\|\, \rho_0),$$

which simplifies to,

$$\mathsf{D}_\Phi(\pi_t \,\|\, \rho_t) \leq \frac{\alpha e^{-2\alpha t}}{\alpha_{\Phi\mathsf{SI}}(\rho_0)(1-e^{-2\alpha t}) + \alpha e^{-2\alpha t}}\mathsf{D}_\Phi(\pi_0 \,\|\, \rho_0).$$

Therefore, the desired bound immediately follows from the above inequality and Definition 4. □

2) Proof of Theorem 1 (SDPI-Based Proof):

Proof: Let $\mathbf{P}_t$ denote the Markov kernel for the Langevin dynamics (7). From Definition 7 and Lemma 14, we have that,

$$\frac{\mathsf{MI}_\Phi(\rho_{0,t})}{\mathsf{MI}_\Phi(\rho_{0,s})} \overset{(36)}{\leq} \varepsilon_{\mathsf{MI}_\Phi}(\mathbf{P}_{t-s}, \rho_s) \leq \varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}_{t-s}, \rho_s).$$

Now using Lemma 18 we can upper bound this by

$$\frac{\alpha e^{-2\alpha(t-s)}}{\alpha_{\Phi\mathsf{SI}}(\rho_s)(1-e^{-2\alpha(t-s)}) + \alpha e^{-2\alpha(t-s)}}.$$

Observe that the denominator is a convex combination of α and $\alpha_{\Phi\mathsf{SI}}(\rho_s)$, and so we get that

$$\frac{\alpha}{\alpha_{\Phi\mathsf{SI}}(\rho_s)(1-e^{-2\alpha(t-s)}) + \alpha e^{-2\alpha(t-s)}} \leq \max\left\{1, \frac{\alpha}{\alpha_{\Phi\mathsf{SI}}(\rho_s)}\right\}$$

which proves the desired result. □

APPENDIX M
PROOFS FOR ULA

A. Proof of Lemma 3

Proof: **Part (a):** Fix any μ such that $\mathsf{D}_\Phi(\mu\|\rho) < \infty$. Denote

$$F_\#\rho * \mathcal{N}(0, tI) = \rho_t, \quad F_\#\mu * \mathcal{N}(0, tI) = \mu_t,$$

where $F(x) = x - \eta\nabla f(x)$. It follows from Lemma 6 that

$$\alpha_{\Phi\mathsf{SI}}(\rho_0) \geq \frac{\alpha_{\Phi\mathsf{SI}}(\rho)}{\gamma^2} \tag{50}$$

where $\rho_0 = F_{\#}\rho$. Furthermore, it follows from Lemma 7 that

$$\alpha_{\Phi\mathsf{SI}}(\rho_t) \geq \frac{\alpha_{\Phi\mathsf{SI}}(\rho_0)}{1 + \alpha_{\Phi\mathsf{SI}}(\rho_0)t}\,. \tag{51}$$

Now using Lemma 12 with $\mu_t = \mu_t$, $\nu_t = \rho_t$, $b_t \equiv 0$, and $c = \frac{1}{2}$, along with Definition 3, we have the following

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{D}_\Phi(\mu_t \,\|\, \rho_t) = -\frac{1}{2}\mathsf{FI}_\Phi(\mu_t \,\|\, \rho_t) \leq -\alpha_{\Phi\mathsf{SI}}(\rho_t)\mathsf{D}_\Phi(\mu_t \,\|\, \rho_t).$$

Integrating this from $t = 0$ to $t = 2\eta$, and using (51) and (50), we get that

$$\begin{aligned}\frac{\mathsf{D}_\Phi(\mu_{2\eta} \,\|\, \rho_{2\eta})}{\mathsf{D}_\Phi(\mu_0 \,\|\, \rho_0)} &\leq \exp\left(-\int_0^{2\eta} \alpha_{\Phi\mathsf{SI}}(\rho_t)\,\mathrm{d}t\right)\\ &\overset{(51)}{\leq} \frac{1}{1 + 2\eta\,\alpha_{\Phi\mathsf{SI}}(\rho_0)} \overset{(50)}{\leq} \frac{\gamma^2}{\gamma^2 + 2\eta\,\alpha_{\Phi\mathsf{SI}}(\rho)}\,.\end{aligned} \tag{52}$$

Note that $\mathsf{D}_\Phi(\mu\,\|\rho) = \mathsf{D}_\Phi(\mu_0 \,\|\, \rho_0)$ as Φ-divergence is invariant to applying a bijective map to both arguments. It thus follows that

$$\begin{aligned}\frac{\mathsf{D}_\Phi(\mu_{2\eta} \,\|\, \rho_{2\eta})}{\mathsf{D}_\Phi(\mu_0 \,\|\, \rho_0)} &= \frac{\mathsf{D}_\Phi(\mu_0 * \mathcal{N}(0, 2\eta I) \,\|\, \rho_0 * \mathcal{N}(0, 2\eta I))}{\mathsf{D}_\Phi(\mu \,\|\, \rho)}\\ &= \frac{\mathsf{D}_\Phi(\mu\mathbf{P} \,\|\, \rho\mathbf{P})}{\mathsf{D}_\Phi(\mu \,\|\, \rho)}.\end{aligned}$$

Finally, the statement follows from the above observation, Definition 4, and (52).

Part (b): Recall the ULA update $\rho\mathbf{P} = F_{\#}\rho * \mathcal{N}(0, 2\eta I)$ mentioned in (17) and note that $\mathcal{N}(0, 2\eta I)$ is $\frac{1}{2\eta}$-strongly log-concave, and therefore satisfies Φ-Sobolev inequality with the same constant (Lemma 8). The result then follows from Lemma 6 and Lemma 7. □

B. Proof of Theorem 3

Proof: Recall the ULA update $\rho\mathbf{P} = F_{\#}\rho * \mathcal{N}(0, 2\eta I)$ mentioned in (17). Denote the Lipschitz constant of F by $\gamma > 0$. We begin by ensuring that F satisfies the conditions required by Lemma 3. The α-strong log-concavity and L-smoothness of ν implies $(1 - \eta L)I \preceq \nabla F \preceq (1 - \eta\alpha)I$ where note that $\nabla F(x) = I - \eta\nabla^2 f(x)$. Therefore, F is $\gamma = (1 - \eta\alpha)$-Lipschitz. For Lemma 3(a), it remains to check that F is bijective. The continuity of F ensures that it is surjective. For injectivity, note that, $F(x) = F(y) \implies x - y = \eta[\nabla f(x) - \nabla f(y)]$. Taking norm on both sides and using that ∇f is L-Lipschitz, along with $\eta \leq 1/L$, implies that F is injective. Therefore F meets all of the requirements set forth by Lemma 3.

Fix an $\ell \geq 1$ such that ρ_ℓ satisfies a Φ-Sobolev inequality. This is guaranteed by assumption. Then Lemma 15 states that

$$\mathsf{MI}_\Phi(\rho_{0,k}) \leq \prod_{i=\ell}^{k-1} \varepsilon_i\, \mathsf{MI}_\Phi(\rho_{0,\ell}), \quad \varepsilon_i = \varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}, \rho_i). \tag{53}$$

As ρ_ℓ satisfies a Φ-Sobolev inequality, Lemma 3(b) guarantees that ρ_j satisfies a Φ-Sobolev inequality too for all $j \geq \ell + 1$. Denote $\alpha_{\Phi\mathsf{SI}}(\rho_i)$ by α_i, then the claim of Lemma 3(b) can be rewritten as

$$1 + \frac{2\alpha_i\eta}{\gamma^2} \overset{(18)}{\geq} \frac{\alpha_i}{\gamma^2\alpha_{i+1}}\,. \tag{54}$$

It thus follows from Lemma 3(a) that

$$\varepsilon_i \leq \left(1 + \frac{2\alpha_i\eta}{\gamma^2}\right)^{-1} \overset{(54)}{\leq} \frac{\gamma^2\alpha_{i+1}}{\alpha_i}\,.$$

Plugging this bound into (53), we obtain

$$\mathsf{MI}_\Phi(\rho_{0,k}) \leq \frac{\gamma^{2(k-\ell)}\alpha_k}{\alpha_\ell}\mathsf{MI}_\Phi(\rho_{0,\ell})\,. \tag{55}$$

Our goal now is to derive a simple upper bound on α_k/α_ℓ. To that end, further denote $c_i = 1/\alpha_i$. Then, (54) is equivalent to $c_{i+1} \leq \gamma^2 c_i + 2\eta$. Subtracting $2\eta/(1 - \gamma^2)$ from both sides and applying the resulting inequality repeatedly, we have

$$c_k - \frac{2\eta}{1 - \gamma^2} \leq \gamma^{2(k-\ell)}\left(c_\ell - \frac{2\eta}{1 - \gamma^2}\right).$$

Recalling $\alpha_k = 1/c_k$ yields

$$\frac{1}{\alpha_k} = \frac{1 - \gamma^{2(k-\ell)}}{\alpha^*} + \frac{\gamma^{2(k-\ell)}}{\alpha_\ell} \geq \min\left\{\frac{1}{\alpha^*}, \frac{1}{\alpha_\ell}\right\} \geq \min\left\{\frac{1}{\alpha}, \frac{1}{\alpha_\ell}\right\} \tag{56}$$

where $\alpha^* = \alpha(1 - \eta\alpha/2) \leq \alpha$. Therefore, we obtain

$$\frac{\alpha_k}{\alpha_\ell} \leq \max\left\{1, \frac{\alpha}{\alpha_\ell}\right\},$$

and hence the claim of the theorem immediately follows from (55). □

C. Proof of Corollary 1

Proof: Note that $\tau \leq e^{\tau - 1}$ for every $\tau \in \mathbb{R}$. Using Theorem 3 and this observation with $\tau = 1 - \alpha\eta$, we have

$$\frac{\mathsf{MI}_\Phi(\rho_{0,k})}{\max\left\{1, \frac{\alpha}{\alpha_{\Phi\mathsf{SI}}(\rho_\ell)}\right\}\mathsf{MI}_\Phi(\rho_{0,\ell})} \overset{(16)}{\leq} (1 - \alpha\eta)^{2(k-\ell)} \leq e^{-2\alpha\eta(k-\ell)}.$$

In view of the above inequality, to bound $\mathsf{MI}_\Phi(X_0; X_k) \leq \epsilon$, it suffices to bound

$$e^{-2\alpha\eta(k-\ell)} \leq \frac{\epsilon}{\max\left\{1, \frac{\alpha}{\alpha_{\Phi\mathsf{SI}}(\rho_\ell)}\right\}\mathsf{MI}_\Phi(\rho_{0,\ell})},$$

which gives the desired complexity bound on k. □

Appendix N
Proofs for Proximal Sampler

In this section, our goal is to prove Theorem 5. As mentioned in Section V-A, the SDPI-based proof for the Proximal Sampler proceeds by studying the contraction coefficient and evolution of the Φ-Sobolev constant along the forward step and backward step separately, and then combining them to obtain the contraction coefficient and Φ-Sobolev constant evolution for the entire Proximal Sampler (Lemma 4).

We denote the Proximal Sampler (10) as $\mathbf{P}_{\text{prox}} = \mathbf{P}^+_{\text{prox}}\mathbf{P}^-_{\text{prox}}$ where $\mathbf{P}^+_{\text{prox}}$ and $\mathbf{P}^-_{\text{prox}}$ correspond to the forward and backward steps, respectively. As each step is an update of probability distributions on $\mathbb{R}^d$, this composition is justified. Regarding

notation, we denote $\rho_k^X := \mathsf{law}(X_k)$, $\rho_k^Y := \mathsf{law}(Y_k)$, and therefore $\rho_k^X \mathbf{P}_{\text{prox}} = \rho_{k+1}^X$, $\rho_k^X \mathbf{P}_{\text{prox}}^+ = \rho_k^Y$, and $\rho_k^Y \mathbf{P}_{\text{prox}}^- = \rho_{k+1}^X$.

We repeat the SDE interpretations of both the forward and backward steps mentioned in Section V.

a) *Forward Step:* Suppose we start from $X_0 \sim \rho_0^X$. Along the forward step of the Proximal Sampler (10), $Y_0 \mid X_0 \sim \mathcal{N}(X_0, \eta I)$, so in particular, $\rho_0^Y = \rho_0^X * \mathcal{N}(0, \eta I)$. Therefore, the action of the forward step $\mathbf{P}_{\text{prox}}^+$ is via a Gaussian convolution: $\rho \mathbf{P}_{\text{prox}}^+ = \rho * \mathcal{N}(0, \eta I)$, which can be interpreted as the solution to the heat flow (generated by the Brownian motion SDE $\mathrm{d}X_t = \mathrm{d}W_t$) at time $\eta > 0$.

b) *Backward Step:* For the backward step $\mathbf{P}_{\text{prox}}^-$, it will be helpful to think of the corresponding SDE as the time reversal of the forward step SDE, i.e., the time reversal of the heat flow, which is known as the *backward heat flow*; see [26] and [12, Chapter 8.3].

Fixing a step-size $\eta > 0$, we have along the forward step ($\mathrm{d}X_t = \mathrm{d}W_t$) that if $X_0 \sim \nu^X$, then $X_\eta \sim \nu^Y$. The backward heat flow SDE (20) is defined by

$$\mathrm{d}Y_t = \nabla \log (\nu^X * \mathcal{N}_{\eta - t})(Y_t)\, \mathrm{d}t + \mathrm{d}W_t\,.$$

By construction, if we start the SDE (20) from $Y_0 \sim \mu_0 = \nu^Y$, then for any $t \in [0, \eta]$, the distribution of $Y_t \sim \mu_t$ along (20) is given by $\mu_t = \nu^X * \mathcal{N}_{\eta - t}$, and in particular, at time $t = \eta$, $Y_\eta \sim \mu_\eta = \nu^X$. For the Proximal Sampler, we start the backward SDE (20) from $Y_0 \sim \rho_0^Y$, to obtain $X_1 := Y_\eta \sim \rho_1^X$.

We present the proofs corresponding to the forward step in Appendix N-A, the backward step in Appendix N-B, and the complete Proximal Sampler in Appendix N-C.

A. Proofs for the Forward Step of the Proximal Sampler

1) Contraction Coefficient for Forward Step:

Lemma 19: Let $\mathbf{P}_t$ denote the Markov kernel or heat semigroup corresponding to evolution along $\mathrm{d}X_t = \mathrm{d}W_t$ for time t. If ρ satisfies a Φ-Sobolev inequality with optimal constant $\alpha_{\Phi\mathsf{SI}}(\rho)$, then

$$\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}_\eta, \rho) \le \frac{1}{1 + \eta \alpha_{\Phi\mathsf{SI}}(\rho)}\,.$$

Proof: Let π be an arbitrary distribution such that $\mathsf{D}_\Phi(\pi \| \rho) < \infty$. And let $\pi_t = \pi \mathbf{P}_t$ and $\rho_t = \rho \mathbf{P}_t$. Therefore, $\pi_0 = \pi$ and $\rho_0 = \rho$. It follows from Definition 3 and Lemma 12 with $\mu_t = \pi_t$, $\nu_t = \rho_t$, $b_t \equiv 0$, and $c = \frac{1}{2}$ that

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathsf{D}_\Phi(\pi_t \,\|\, \rho_t) \overset{(29)}{=} -\frac{1}{2} \mathsf{FI}_\Phi(\pi_t \,\|\, \rho_t) \le -\alpha_{\Phi\mathsf{SI}}(\rho_t) \mathsf{D}_\Phi(\pi_t \,\|\, \rho_t)\,. \tag{57}$$

Since the forward step is the heat flow, i.e., $\rho_\eta = \rho \mathbf{P}_{\text{prox}}^+ = \rho * \mathcal{N}(0, \eta I)$, it follows from Lemma 7 that

$$\frac{1}{\alpha_{\Phi\mathsf{SI}}(\rho_t)} \le \frac{1}{\alpha_{\Phi\mathsf{SI}}(\rho_0)} + t\,.$$

Plugging the above inequality into (57) yields that

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathsf{D}_\Phi(\pi_t \,\|\, \rho_t) \le \frac{-\alpha_{\Phi\mathsf{SI}}(\rho_0)}{1 + t \alpha_{\Phi\mathsf{SI}}(\rho_0)} \mathsf{D}_\Phi(\pi_t \,\|\, \rho_t).$$

Applying Grönwall's inequality, we have

$$\mathsf{D}_\Phi(\pi_\eta \,\|\, \rho_\eta) \le e^{-\alpha_{\Phi\mathsf{SI}}(\rho_0) \int_0^\eta \frac{1}{1 + t\alpha_{\Phi\mathsf{SI}}(\rho_0)} \mathrm{d}t} \mathsf{D}_\Phi(\pi_0 \,\|\, \rho_0),$$

which simplifies to

$$\mathsf{D}_\Phi(\pi_\eta \,\|\, \rho_\eta) \le \frac{\mathsf{D}_\Phi(\pi_0 \,\|\, \rho_0)}{1 + \eta \alpha_{\Phi\mathsf{SI}}(\rho_0)}.$$

The conclusion of the lemma immediately follows from the above inequality and Definition 4. □

B. Proofs for the Backward Step of the Proximal Sampler

1) Contraction Coefficient for Backward Step:

Lemma 20: Let ν_0 be α-SLC for some $\alpha > 0$ and define $\nu_t = \nu_0 * \mathcal{N}(0, tI)$. Fix some positive constant $T > 0$. For $t \in [0, T]$, let $\mathbf{P}_t^T$ denote evolution along the following SDE for time t

$$\mathrm{d}X_t = \nabla \log \nu_{T-t}(X_t)\, \mathrm{d}t + \mathrm{d}W_t\,.$$

If ρ satisfies a Φ-Sobolev inequality with optimal constant $\alpha_{\Phi\mathsf{SI}}(\rho)$, then,

$$\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}_t^T, \rho) \le \frac{1 + \alpha T - \alpha t}{(1 + \alpha T)[1 + \alpha_{\Phi\mathsf{SI}}(\rho) t] - \alpha t}\,. \tag{58}$$

Proof: Let π be an arbitrary distribution such that $\mathsf{D}_\Phi(\pi \| \rho) < \infty$. And let $\pi_t = \pi \mathbf{P}_t^T$ and $\rho_t = \rho \mathbf{P}_t^T$. Therefore, $\pi_0 = \pi$ and $\rho_0 = \rho$. From Lemma 12 with $\mu_t = \pi_t$, $\nu_t = \rho_t$, $b_t = \nabla \log \nu_{T-t}$, and $c = \frac{1}{2}$, and Definition 3, we have that

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathsf{D}_\Phi(\pi_t \,\|\, \rho_t) = -\frac{1}{2} \mathsf{FI}_\Phi(\pi_t \,\|\, \rho_t) \overset{(4)}{\le} -\alpha_{\Phi\mathsf{SI}}(\rho_t) \mathsf{D}_\Phi(\pi_t \,\|\, \rho_t)\,. \tag{59}$$

Denote $\alpha_{\Phi\mathsf{SI}}(\rho_t)$ as α_t for $t \in [0, T]$. Then Lemma 21 tells us,

$$\alpha_t \ge \frac{\alpha_0 (1 + \alpha T)^2}{(1 + \alpha(T - t))^2 + \alpha_0 t [1 + \alpha(T - t)](1 + \alpha T)}\,.$$

To apply Grönwall's inequality, we need to evaluate the following

$$\begin{aligned}
A_t &= \int_0^t \frac{\alpha_0 (1 + \alpha T)^2}{(1 + \alpha(T - s))^2 + \alpha_0 s (1 + \alpha(T - s))(1 + \alpha T)}\, \mathrm{d}s \\
&= \int_0^t \frac{\alpha_0 (1 + \alpha T)^2 \ \mathrm{d}s}{\alpha(\alpha - \alpha_0(1 + \alpha T)) \left(s - \frac{1 + \alpha T}{\alpha - \alpha_0(1 + \alpha T)}\right) \left(s - \frac{1 + \alpha T}{\alpha}\right)} \\
&= \int_0^t \frac{1}{s - \frac{1 + \alpha T}{\alpha - \alpha_0(1 + \alpha T)}}\, \mathrm{d}s - \int_0^t \frac{1}{s - \frac{1 + \alpha T}{\alpha}}\, \mathrm{d}s \\
&= \log \frac{1 + \alpha T - t(\alpha - \alpha_0(1 + \alpha T))}{1 + \alpha T} - \log \frac{1 + \alpha T - \alpha t}{1 + \alpha T} \\
&= \log \frac{(1 + \alpha T)(1 + \alpha_0 t) - \alpha t}{1 + \alpha T - \alpha t}.
\end{aligned}$$

Applying Grönwall's inequality to (59) gives $\mathsf{D}_\Phi(\pi_t \,\|\, \rho_t) \le e^{-A_t} \mathsf{D}_\Phi(\pi_0 \,\|\, \rho_0)$, i.e.,

$$\mathsf{D}_\Phi(\pi_t \,\|\, \rho_t) \le \frac{1 + \alpha T - \alpha t}{(1 + \alpha T)(1 + \alpha_0 t) - \alpha t} \mathsf{D}_\Phi(\pi_0 \,\|\, \rho_0)\,.$$

The result follows by using Definition 4 and noting that the choice of π is arbitrary. □

2) Evolution of Φ-Sobolev Constant Along Backward Step:

Lemma 21: Let ν_0 be α-SLC for some $\alpha > 0$, and define $\nu_t = \nu_0 * \mathcal{N}(0, tI)$. Fix some $T > 0$. For $t \in [0, T]$, consider $X_t \sim \rho_t$ which evolves according to

$$\mathrm{d}X_t = \nabla \log \nu_{T-t}(X_t)\,\mathrm{d}t + \mathrm{d}W_t\,,$$

from $X_0 \sim \rho_0$ where ρ_0 satisfies a Φ-Sobolev inequality with optimal constant $\alpha_{\Phi\mathsf{SI}}(\rho_0)$. Then we have that,

$$\frac{1}{\alpha_{\Phi\mathsf{SI}}(\rho_t)} \le \frac{1}{\alpha_{\Phi\mathsf{SI}}(\rho_0)}\left(1 - \frac{\alpha t}{1+\alpha T}\right)^2 + \frac{t(1+\alpha(T-t))}{1+\alpha T}\,. \tag{60}$$

Proof: As ν_0 is α-SLC and $\nu_t = \nu_0 * \mathcal{N}(0, tI)$, it follows from [83, Theorem 3.7(b)] that ν_t is α_t-SLC where $\alpha_t = \alpha/(1+t\alpha)$. Writing $\nu_t \propto \exp(-f_t)$, the backward heat flow SDE can be rewritten as

$$\mathrm{d}X_t = -\nabla f_{T-t}(X_t)\,\mathrm{d}t + \mathrm{d}W_t\,.$$

Now consider a discretization of this SDE with a sufficiently small step-size $\eta > 0$ to get

$$X_{k+1} = X_k - \eta \nabla f_{T-\eta k}(X_k) + \sqrt{\eta} Z_k\,,$$

where $Z_k \sim \mathcal{N}(0, I)$. We therefore know that $f_{T-\eta k}$ is β_k-strongly convex where

$$\beta_k = \frac{\alpha}{1+\alpha(T-\eta k)}. \tag{61}$$

Letting $X_k \sim \rho_k$, we therefore have that

$$\rho_{k+1} = (I - \eta \nabla f_{T-\eta k})_{\#}\rho_k * \mathcal{N}(0, \eta I), \tag{62}$$

and the mapping $I - \eta\nabla f_{T-\eta k}$ is $(1-\eta\beta_k)$-Lipschitz continuous. Denote $c_k = 1/\alpha_{\Phi\mathsf{SI}}(\rho_k)$. In view of (62), using Lemmas 6 and 7, we have the following recursion

$$c_{k+1} \overset{(22)}{\le} \frac{1}{\alpha_{\Phi\mathsf{SI}}((I-\eta\nabla f_{T-\eta k})_{\#}\rho_k)} + \eta \overset{(21)}{\le} (1-\eta\beta_k)^2 c_k + \eta\,.$$

Upon further iteration this yields

$$c_k \le \prod_{i=0}^{k}(1-\eta\beta_i)^2 c_0 + \eta\left[1 + \sum_{j=1}^{k-1}\prod_{i=j}^{k-1}(1-\eta\beta_i)^2\right].$$

Plugging in the expression for β_i in (61) and simplifying, we get

$$c_k \le \left(1 - \frac{\alpha\eta k}{1+\alpha T}\right)^2 c_0 + \eta\left[1 + \sum_{j=1}^{k-1}\left(\frac{1+\alpha(T-\eta k)}{1+\alpha(T-\eta j)}\right)^2\right].$$

Now taking the limit $\eta \to 0$, $\eta k \to t$, and $\eta j \to s$, we get

$$c_t \le \left(1 - \frac{\alpha t}{1+\alpha T}\right)^2 c_0 + (1+\alpha(T-t))^2 \int_0^t \frac{\mathrm{d}s}{(1+\alpha(T-s))^2}\,,$$

which simplifies to

$$c_t \le \left(1 - \frac{\alpha t}{1+\alpha T}\right)^2 c_0 + \frac{t(1+\alpha(T-t))}{1+\alpha T}\,.$$

The desired claim (60) now follows from the fact that $c_t = 1/\alpha_{\Phi\mathsf{SI}}(\rho_t)$.□

C. Proofs for the Complete Proximal Sampler

1) Proof of Lemma 4:

Proof: **Part (a):** It follows from Definition 4 and the fact that $\mathbf{P}_{\mathrm{prox}} = \mathbf{P}^+_{\mathrm{prox}}\mathbf{P}^-_{\mathrm{prox}}$ that

$$\begin{aligned}
&\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}_{\mathrm{prox}}, \rho)\\
&\overset{(32)}{:=} \sup_{\pi:\,0<\mathsf{D}_\Phi(\pi\|\rho)<\infty} \frac{\mathsf{D}_\Phi(\pi\,\mathbf{P}_{\mathrm{prox}}\,\|\,\rho\,\mathbf{P}_{\mathrm{prox}})}{\mathsf{D}_\Phi(\pi\,\|\,\rho)}\\
&= \sup_{\pi:\,0<\mathsf{D}_\Phi(\pi\|\rho)<\infty} \frac{\mathsf{D}_\Phi(\pi\,\mathbf{P}^+_{\mathrm{prox}}\mathbf{P}^-_{\mathrm{prox}}\,\|\,\rho\,\mathbf{P}^+_{\mathrm{prox}}\mathbf{P}^-_{\mathrm{prox}})}{\mathsf{D}_\Phi(\pi\,\mathbf{P}^+_{\mathrm{prox}}\,\|\,\rho\,\mathbf{P}^+_{\mathrm{prox}})}\\
&\qquad\qquad \times \frac{\mathsf{D}_\Phi(\pi\,\mathbf{P}^+_{\mathrm{prox}}\,\|\,\rho\,\mathbf{P}^+_{\mathrm{prox}})}{\mathsf{D}_\Phi(\pi\,\|\,\rho)}\\
&\overset{(32)}{\le} \varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}^+_{\mathrm{prox}}, \rho)\,\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}^-_{\mathrm{prox}}, \rho\,\mathbf{P}^+_{\mathrm{prox}}),
\end{aligned} \tag{63}$$

where the inequality is also due to Definition 4. Observe that $\mathbf{P}^+_{\mathrm{prox}}$ is the forward heat flow operation for time $\eta > 0$, and hence $\rho\mathbf{P}^+_{\mathrm{prox}} = \rho * \mathcal{N}(0, \eta I)$. It thus follows from Lemma 7 that

$$\alpha_{\Phi\mathsf{SI}}(\rho\,\mathbf{P}^+_{\mathrm{prox}}) \overset{(22)}{\ge} \frac{\alpha_{\Phi\mathsf{SI}}(\rho)}{1+\eta\alpha_{\Phi\mathsf{SI}}(\rho)}, \tag{64}$$

and from Lemma 19 that

$$\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}^+_{\mathrm{prox}}, \rho) \le \frac{1}{1+\eta\alpha_{\Phi\mathsf{SI}}(\rho)}. \tag{65}$$

In view of Lemma 20, $\mathbf{P}^-_{\mathrm{prox}}$ is the backward heat flow operation with $\nu_0 = \nu^X$, $T = \eta$, and for time η. Hence, using (63), (64), (65), and Lemma 20 with $T = t = \eta$, $\mathbf{P}^T_t = \mathbf{P}^-_{\mathrm{prox}}$, and $\rho = \rho\mathbf{P}^+_{\mathrm{prox}}$, we obtain

$$\begin{aligned}
&\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}_{\mathrm{prox}}, \rho)\\
&\overset{(63),(65)}{\le} \frac{1}{1+\eta\alpha_{\Phi\mathsf{SI}}(\rho)}\,\varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}^-_{\mathrm{prox}}, \rho\,\mathbf{P}^+_{\mathrm{prox}})\\
&\overset{(58)}{\le} \frac{1}{1+\eta\alpha_{\Phi\mathsf{SI}}(\rho)} \times \frac{1}{(1+\alpha\eta)(1+\alpha_{\Phi\mathsf{SI}}(\rho\mathbf{P}^+_{\mathrm{prox}})\eta) - \alpha\eta}\\
&\overset{(64)}{\le} \frac{1}{1+2\eta\alpha_{\Phi\mathsf{SI}}(\rho) + \eta^2\alpha\alpha_{\Phi\mathsf{SI}}(\rho)}.
\end{aligned}$$

This completes part (a).

Part (b): Observe that $\mathbf{P}^+_{\mathrm{prox}}$ is the forward heat flow operation for time $\eta > 0$, and hence $\rho\mathbf{P}^+_{\mathrm{prox}} = \rho * \mathcal{N}(0, \eta I)$. It thus follows from Lemma 7 that

$$\frac{1}{\alpha_{\Phi\mathsf{SI}}(\rho\mathbf{P}^+_{\mathrm{prox}})} \overset{(22)}{\le} \frac{1}{\alpha_{\Phi\mathsf{SI}}(\rho)} + \eta\,. \tag{66}$$

In view of Lemma 21, $\mathbf{P}^-_{\mathrm{prox}}$ is the backward heat flow operation with $\nu_0 = \nu^X$, $T = \eta$, and for time η. Hence, using (66) and Lemma 21 with $T = t = \eta$ and $\rho_0 = \rho\mathbf{P}^+_{\mathrm{prox}}$, we obtain

$$\begin{aligned}
\frac{1}{\alpha_{\Phi\mathsf{SI}}(\rho\,\mathbf{P}_{\mathrm{prox}})} &\overset{(60)}{\le} \frac{1}{\alpha_{\Phi\mathsf{SI}}(\rho\mathbf{P}^+_{\mathrm{prox}})}\frac{1}{(1+\alpha\eta)^2} + \frac{\eta}{1+\alpha\eta}\\
&\overset{(66)}{\le} \frac{1+\alpha_{\Phi\mathsf{SI}}(\rho)\eta}{\alpha_{\Phi\mathsf{SI}}(\rho)}\frac{1}{(1+\alpha\eta)^2} + \frac{\eta}{1+\alpha\eta}.
\end{aligned}$$

Therefore, the desired claim is proved. □

2) Proof of Theorem 5:

Proof: First, it follows from Lemma 15 for the Proximal Sampler that for any $\ell \geq 1$ and $k \geq \ell$

$$\mathsf{MI}_\Phi(\rho^X_{0,k}) \leq \prod_{i=\ell}^{k-1} \varepsilon_i \, \mathsf{MI}_\Phi(\rho^X_{0,\ell}), \quad \varepsilon_i = \varepsilon_{\mathsf{D}_\Phi}(\mathbf{P}_{\text{prox}}, \rho^X_i). \tag{67}$$

Throughout the proof, let α_i denote $\alpha_{\Phi\mathsf{SI}}(\rho^X_i)$. Recall from Lemma 4(b) that

$$\frac{1}{\alpha_{i+1}} \leq \frac{1+\alpha_i\eta}{\alpha_i(1+\alpha\eta)^2} + \frac{\eta}{1+\alpha\eta}. \tag{68}$$

Fix ℓ such that ρ_ℓ satisfies a Φ-Sobolev inequality. This holds by assumption.

a) Case 1: $\alpha \leq \alpha_\ell$: Using $\alpha_\ell \geq \alpha$ and (68) with $i = \ell$, we have

$$\frac{1}{\alpha_{\ell+1}} \overset{(68)}{\leq} \frac{1+\alpha_\ell\eta}{\alpha_\ell(1+\alpha\eta)^2} + \frac{\eta}{1+\alpha\eta} \leq \frac{1}{\alpha(1+\alpha\eta)} + \frac{\eta}{1+\alpha\eta} = \frac{1}{\alpha},$$

which implies that $\alpha_{\ell+1} \geq \alpha$. Repeating this argument on (68), we have that $\alpha_i \geq \alpha$ for all $i \geq \ell$. The contraction coefficient bound in Lemma 4(a) therefore simplifies to

$$\varepsilon_i \leq \frac{1}{1+2\eta\alpha_i+\eta^2\alpha\alpha_i} \leq \frac{1}{(1+\eta\alpha)^2},$$

for all $i \geq \ell$. Plugging this inequality into (67), we obtain

$$\mathsf{MI}_\Phi(\rho^X_{0,k}) \leq \frac{\mathsf{MI}_\Phi(\rho^X_{0,\ell})}{(1+\eta\alpha)^{2(k-\ell)}}. \tag{69}$$

b) Case 2: $\alpha > \alpha_\ell$: Using $\alpha > \alpha_\ell$ and (68), we have

$$\frac{1}{\alpha_{i+1}} \leq \frac{1+\alpha_i\eta}{\alpha_i(1+\alpha_\ell\eta)^2} + \frac{\eta}{1+\alpha_\ell\eta}, \tag{70}$$

for all $i \geq \ell$. We next prove by induction that $\alpha_i \geq \alpha_\ell$ for all $i \geq \ell$. This is of course true when $i = \ell$. Assume for some $i \geq \ell$ that $\alpha_i \geq \alpha_\ell$. Then, it follows from (70) that

$$\frac{1}{\alpha_{i+1}} \leq \frac{1+\alpha_\ell\eta}{\alpha_\ell(1+\alpha_\ell\eta)^2} + \frac{\eta}{1+\alpha_\ell\eta} = \frac{1}{\alpha_\ell},$$

and hence that the claim is true for the case $i+1$. Together with this conclusion, Lemma 4(a) simplifies to

$$\varepsilon_i \leq \frac{1}{1+2\eta\alpha_i+\eta^2\alpha\alpha_i} \leq \frac{1}{(1+\eta\alpha_\ell)^2},$$

for all $i \geq \ell$. Plugging this inequality into (67), we obtain

$$\mathsf{MI}_\Phi(\rho^X_{0,k}) \leq \frac{\mathsf{MI}_\Phi(\rho^X_{0,\ell})}{(1+\eta\alpha_\ell)^{2(k-\ell)}}. \tag{71}$$

Finally, the theorem follows by combining (69) and (71). □

3) Proof of Corollary 2:

Proof: Note that $\tau \leq e^{\tau-1}$ for every $\tau \in \mathbb{R}$. Using Theorem 5 and this observation with

$$\tau = \frac{1}{1+\eta\tilde{\alpha}}, \quad \tilde{\alpha} = \min\{\alpha, \alpha_{\Phi\mathsf{SI}}(\rho^X_\ell)\},$$

we have

$$\frac{\mathsf{MI}_\Phi(\rho^X_{0,k})}{\mathsf{MI}_\Phi(\rho^X_{0,\ell})} \overset{(19)}{\leq} \frac{1}{(1+\eta\tilde{\alpha})^{2(k-\ell)}} \leq \exp\left(-\frac{2\eta\tilde{\alpha}(k-\ell)}{1+\eta\tilde{\alpha}}\right).$$

In view of the above inequality, to bound $\mathsf{MI}_\Phi(X_0; X_k) \leq \epsilon$, it suffices to bound

$$\exp\left(-\frac{2\eta\tilde{\alpha}(k-\ell)}{1+\eta\tilde{\alpha}}\right) \leq \frac{\epsilon}{\mathsf{MI}_\Phi(\rho^X_{0,\ell})},$$

which gives the desired complexity bound on k. □

APPENDIX O
EXAMPLES FOR THE ORNSTEIN-UHLENBECK PROCESS AND THE HEAT FLOW

In this section, we fix $\Phi(x) = x\log x$ and discuss the convergence of mutual information for the Ornstein-Uhlenbeck (OU) process and the heat flow. The entropy functional of a distribution ρ is defined to be $\mathsf{H}(\rho) = -\mathbb{E}_\rho[\log\rho]$. Recall that the standard mutual information is

$$\begin{aligned}\mathsf{MI}(\rho^{XY}) &= \mathsf{KL}\left(\rho^{XY} \,\|\, \rho^X \otimes \rho^Y\right) \\ &= \mathsf{H}(\rho^Y) - \mathop{\mathbb{E}}_{x\sim\rho^X}[\mathsf{H}(\rho_{Y|X}(\cdot \mid x))].\end{aligned} \tag{72}$$

A. Convergence Rates in Continuous Time Along Langevin Dynamics

We provide two propositions showing that the convergence rates of mutual information along OU process and heat flow are $\Theta\left(e^{-2\alpha t}\right)$ and $\Theta(1/t)$, respectively. This shows the tightness of Theorems 1 and 2. Before proceeding, we mention the following fact which will be useful in the analysis.

Fact 1: [84, Theorem 17.7.3] The *entropy power* of a probability distribution ρ on $\mathbb{R}^d$ is x

$$\Lambda(\rho) = \frac{1}{2\pi e} e^{\frac{2}{d}\mathsf{H}(\rho)}.$$

The *entropy power inequality* states for independent random variables with distributions ρ and ν,

$$\Lambda(\rho * \nu) \geq \Lambda(\rho) + \Lambda(\nu).$$

1) Mutual Information Along the OU Process: The OU process is the Langevin dynamics for a Gaussian target distribution $\nu = \mathcal{N}(c, \Sigma)$ for some $c \in \mathbb{R}^d$ and $\Sigma \succ 0$. Here, for simplicity, we consider the case where $c = 0$ and $\Sigma = \frac{1}{\alpha}I$ for some $\alpha > 0$. In this case, (7) becomes

$$\mathrm{d}X_t = -\alpha X_t \, \mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t. \tag{73}$$

The solution to the OU process is

$$X_t = e^{-\alpha t}X_0 + \sqrt{\tau_\alpha(t)}Z, \tag{74}$$

where $Z \sim \mathcal{N}(0, I)$ and

$$\tau_\alpha(t) = \frac{1-e^{-2\alpha t}}{\alpha}. \tag{75}$$

We have the following theorem that describes the rate of convergence of mutual information along the OU process. Note when ρ_0 is a Gaussian, ρ_t is a Gaussian for all t, and we can compute the mutual information explicitly; however, our statement holds for more general initial distributions ρ_0.

Proposition 1: Let $X_t \sim \rho_t$ evolve along the OU process (73) from $X_0 \sim \rho_0$, where ρ_0 satisfies $\mathrm{Cov}(\rho_0) \preceq JI$, and let $\rho_{0,t}$ be the joint law of (X_0, X_t). Then, we have

$$\mathsf{MI}(\rho_{0,t}) \le \frac{d}{2}\log\left(1+\frac{\alpha J}{e^{2\alpha t}-1}\right) \le \frac{\alpha d J}{2(e^{2\alpha t}-1)}, \tag{76}$$

and,

$$\mathsf{MI}(\rho_{0,t}) \ge \frac{d}{2}\log\left(\frac{\alpha e^{\frac{2}{d}\mathsf{H}(\rho_0)}}{2\pi e(e^{2\alpha t}-1)}+1\right). \tag{77}$$

Therefore, along the OU process, as $t \to \infty$, we have

$$\mathsf{MI}(\rho_{0,t}) = \Theta\left(e^{-2\alpha t}\right). \tag{78}$$

Proof: It follows from (74) that

$$\rho_{t|0} = \mathcal{N}(e^{-\alpha t}X_0, \tau_\alpha(t)I), \quad \mathsf{H}(\rho_{t|0}) = \frac{d}{2}\log\left(2\pi e\tau_\alpha(t)\right),$$

and

$$\mathop{\mathbb{E}}_{\rho_0}[\mathsf{H}(\rho_{t|0})] = \frac{d}{2}\log\left(2\pi e\tau_\alpha(t)\right). \tag{79}$$

We next bound $\mathsf{H}(\rho_t)$ where,

$$\rho_t(y) = \int_{\mathbb{R}^d} \rho_0(x)\,\mathbb{P}_{\mathcal{N}(e^{-\alpha t}x,\tau_\alpha(t)I)}(y)\,\mathrm{d}x,$$

and $\mathbb{P}_{\mathcal{N}(c,\Sigma)}$ is the density of $\mathcal{N}(c,\Sigma)$. We first derive an upper bound on $\mathsf{H}(\rho_t)$. Using the fact that for a fixed covariance, the Gaussian distribution maximizes entropy, and the observation that

$$\mathrm{Cov}(\rho_t) = e^{-2\alpha t}\mathrm{Cov}(\rho_0) + \tau_\alpha(t)I,$$

we have

$$\begin{aligned}\mathsf{H}(\rho_t) &\le \mathsf{H}(\mathcal{N}(0,\mathrm{Cov}(\rho_t)))\\ &= \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det\left(e^{-2\alpha t}\mathrm{Cov}(\rho_0)+\tau_\alpha(t)I\right)\\ &\le \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det\left(e^{-2\alpha t}JI+\tau_\alpha(t)I\right)\\ &= \frac{d}{2}\log\left(2\pi e(e^{-2\alpha t}J+\tau_\alpha(t))\right),\end{aligned} \tag{80}$$

where the second inequality follows from the assumption that $\mathrm{Cov}(\rho_0) \preceq JI$. Next, we derive a lower bound on $\mathsf{H}(\rho_t)$. Considering the solution (74) and letting ρ denote the distribution of $e^{-\alpha t}X_0$ and $\nu = \mathcal{N}(0,\tau_\alpha(t)I)$, then we have $\rho_t = \rho * \nu$. By definition, we have

$$\mathsf{H}(\rho) = \mathsf{H}(\mathsf{law}(e^{-\alpha t}X_0)) = \mathsf{H}(\rho_0) - d\alpha t.$$

It thus follows from the entropy power inequality (i.e., Fact 1) that

$$\begin{aligned}\mathsf{H}(\rho_t) &\ge \frac{d}{2}\log\left(e^{\frac{2}{d}\mathsf{H}(\mathsf{law}(e^{-\alpha t}X_0))} + e^{\frac{2}{d}\mathsf{H}(\mathcal{N}(0,\tau_\alpha(t)I))}\right)\\ &= \frac{d}{2}\log\left(e^{\frac{2}{d}\mathsf{H}(\rho_0)-2\alpha t} + 2\pi e\tau_\alpha(t)\right).\end{aligned} \tag{81}$$

Now, (76) immediately follows (72), (75), (79), and (80),

$$\begin{aligned}\mathsf{MI}(\rho_{0,t}) &\overset{(72)}{=} \mathsf{H}(\rho_t) - \mathop{\mathbb{E}}_{\rho_0}[\mathsf{H}(\rho_{t|0})]\\ &\overset{(79),(81)}{\le} \frac{d}{2}\log\left(\frac{e^{-2\alpha t}J}{\tau_\alpha(t)}+1\right) \overset{(75)}{=} \frac{d}{2}\log\left(\frac{\alpha J}{e^{2\alpha t}-1}+1\right)\\ &\le \frac{\alpha d J}{2(e^{2\alpha t}-1)},\end{aligned}$$

where the last inequality is due to the fact that $\log(1+x) \le x$. Using (72), (75), (79), and (81), we have

$$\begin{aligned}\mathsf{MI}(\rho_{0,t}) &\overset{(72)}{=} \mathsf{H}(\rho_t) - \mathop{\mathbb{E}}_{\rho_0}[\mathsf{H}(\rho_{t|0})]\\ &\overset{(79),(81)}{\ge} \frac{d}{2}\log\left(\frac{e^{\frac{2}{d}\mathsf{H}(\rho_0)-2\alpha t}}{2\pi e\tau_\alpha(t)}+1\right)\\ &\overset{(75)}{=} \frac{d}{2}\log\left(\frac{\alpha e^{\frac{2}{d}\mathsf{H}(\rho_0)}}{2\pi e(e^{2\alpha t}-1)}+1\right).\end{aligned}$$

Hence, (77) is proved. It follows from the fact that $\log(1+x) \ge x/2$ for $x \in [0,1]$ and (77) that as $t \to \infty$,

$$\mathsf{MI}(\rho_{0,t}) \ge \frac{\alpha d e^{\frac{2}{d}\mathsf{H}(\rho_0)}}{8\pi e(e^{2\alpha t}-1)}.$$

Finally, (78) follows from the above conclusion and (76). □

Remark 1: Using Stam's inequality [33],

$$\Lambda(\rho)\mathsf{FI}(\rho) \ge d,$$

and the Blachman-Stam inequality [85],

$$\frac{1}{\mathsf{FI}(\rho*\nu)} \ge \frac{1}{\mathsf{FI}(\rho)} + \frac{1}{\mathsf{FI}(\nu)},$$

instead of using the entropy power inequality (i.e., Fact 1), we can derive an alternative lower bound,

$$\mathsf{MI}(\rho_{0,t}) \ge \frac{d}{2}\log\left(\frac{\alpha d}{(e^{2\alpha t}-1)\mathsf{FI}(\rho_0)}+1\right).$$

2) Mutual Information Along the Heat Flow: Recall the target distribution for the OU process discussed in Appendix O-A.1 is $\nu = \mathcal{N}\left(0, \frac{1}{\alpha}I\right)$. As $\alpha \to 0$, (73) simply reduces to

$$\mathrm{d}X_t = \sqrt{2}\,\mathrm{d}W_t, \tag{82}$$

which is the heat flow. It is the SDE corresponding to the heat equation: $\partial_t\rho_t = \Delta\rho_t$.

We have the following proposition that describes the rate of convergence of mutual information along the heat flow. Its proof follows similarly from that of Proposition 1 and hence is omitted.

Proposition 2: Let $X_t \sim \rho_t$ evolve along the heat flow (82) from $X_0 \sim \rho_0$ from some ρ_0 and let $\rho_{0,t}$ be the joint law of (X_0, X_t). Then,

$$\mathsf{MI}(\rho_{0,t}) \le \frac{1}{2}\sum_{i=1}^{d}\log\left(\frac{\lambda_i(\mathrm{Cov}(\rho_0))}{2t}+1\right),$$

where $\lambda_i(\cdot)$ is the i-th largest eigenvalue of a matrix, and

$$\mathsf{MI}(\rho_{0,t}) \ge \frac{d}{2}\log\left(\frac{e^{\frac{2}{d}\mathsf{H}(\rho_0)}}{4\pi et}+1\right).$$

Therefore, along the heat flow, as $t \to \infty$, we have

$$\mathsf{MI}(\rho_{0,t}) = \Theta\left(\frac{1}{t}\right).$$

B. Convergence Rates in Discrete Time Along ULA

Here we discuss ULA for the OU process, indicating the tightness of Theorems 3 and 4.

For simplicity, as in Appendix O-A.1, we again consider the target to be $\nu = \mathcal{N}\left(0, \frac{1}{\alpha} I\right)$. In this case, the ULA update (8) becomes:

$$X_{k+1} = (1 - \eta\alpha)X_k + \sqrt{2\eta} Z_k, \tag{83}$$

and the solution to (83) is:

$$X_k = (1-\eta\alpha)^k X_0 + \sqrt{\frac{2(1-(1-\eta\alpha)^{2k})}{\alpha(2-\eta\alpha)}} Z, \tag{84}$$

where $Z \sim \mathcal{N}(0, I)$. To show a tightness result, we simply take ρ_0 to be $\mathcal{N}(0, I)$ in the proposition below.

Proposition 3: Let X_k evolve along (83) from $X_0 \sim \rho_0 = \mathcal{N}(0, I)$. Then

$$\mathsf{MI}(X_0; X_k) = \mathsf{MI}(\rho_{0,k}) = \frac{d}{2} \log\left[1 + \frac{(1-\eta\alpha)^{2k}(2\alpha - \eta\alpha^2)}{2(1-(1-\eta\alpha)^{2k})}\right].$$

Therefore, as $k \to \infty$, we have that $\mathsf{MI}(\rho_{0,k}) = \Theta(d\alpha(1-\eta\alpha)^{2k})$.

Proof: It follows from (84) that

$$\rho_{k|0} = \mathcal{N}\left((1-\eta\alpha)^k X_0, \frac{2(1-(1-\eta\alpha)^{2k})}{\alpha(2-\eta\alpha)} I\right), \tag{85}$$

which, together with the fact that $X_0 \sim \rho_0 = \mathcal{N}(0, I)$, implies that $X_k \sim \rho_k$ where

$$\rho_k = \mathcal{N}\left(0, \frac{2 + (1-\eta\alpha)^{2k}(2\alpha - \eta\alpha^2 - 2)}{\alpha(2-\eta\alpha)} I\right). \tag{86}$$

Using (72) and the formula for entropy of a Gaussian on (86) and (85), we therefore get that:

$$\begin{aligned}
\mathsf{MI}(\rho_{0,k}) &\overset{(72)}{=} \mathsf{H}(\rho_k) - \mathop{\mathbb{E}}_{\rho_0}[\mathsf{H}(\rho_{k|0})] \\
&= \frac{d}{2} \log\left[\frac{2 + (1-\eta\alpha)^{2k}(2\alpha - \eta\alpha^2 - 2)}{2(1-(1-\eta\alpha)^{2k})}\right] \\
&= \frac{d}{2} \log\left[1 + \frac{(1-\eta\alpha)^{2k}(2\alpha - \eta\alpha^2)}{2(1-(1-\eta\alpha)^{2k})}\right],
\end{aligned}$$

which, together with the fact that $\log(1+x) \geq x/2$ for $x \in [0, 1]$, proves the desired claim. □

This shows the tightness of Theorems 3 and 4.

C. Convergence Rates in Discrete Time Along the Proximal Sampler

Let the target distribution be $\nu^X = \mathcal{N}\left(0, \frac{1}{\alpha} I\right)$ for $\alpha > 0$ and $\rho_0^X = \mathcal{N}(m_0, c_0^2 I)$ for $m_0 \in \mathbb{R}^d$ and $c_0 > 0$. In this case, the conditional distribution $\nu^{X|Y}$ (11) is given by

$$\nu^{X|Y}(\,\cdot\,|\,y) = \mathcal{N}\left(\frac{y}{1+\alpha\eta}, \frac{\eta}{1+\alpha\eta} I\right).$$

Explicit computation for the Proximal Sampler for the Gaussian case [26, Section 4.4] reveals that we have $\rho_k^X = \mathcal{N}(m_k, c_k^2 I)$ for $k \geq 0$ where

$$m_{k+1} = \frac{m_k}{1+\alpha\eta} \text{ and } c_{k+1}^2 - \frac{1}{\alpha} = \frac{1}{(1+\alpha\eta)^2}\left(c_k^2 - \frac{1}{\alpha}\right).$$

Hence for all $k \geq 0$

$$m_k = \frac{m_0}{(1+\alpha\eta)^k} \text{ and } c_k^2 - \frac{1}{\alpha} = \frac{1}{(1+\alpha\eta)^{2k}}\left(c_0^2 - \frac{1}{\alpha}\right). \tag{87}$$

We use this explicit Gaussian solution to compute $\mathsf{MI}(X_0; X_k)$ along the Proximal Sampler.

Proposition 4: Let $X_k \sim \rho_k^X$ be iterates along the Proximal Sampler (10) with $\nu^X = \mathcal{N}\left(0, \frac{1}{\alpha} I\right)$ for $\alpha > 0$ and $\rho_0^X = \mathcal{N}(0, I)$. Then

$$\mathsf{MI}(X_0; X_k) = \frac{d}{2} \log\left[1 + \frac{\alpha}{(1+\alpha\eta)^{2k} - 1}\right].$$

Therefore, as $k \to \infty$, we have that $\mathsf{MI}(\rho_{0,k}) = \Theta(d\alpha(1+\eta\alpha)^{-2k})$.

Proof: Recall from (72) that $\mathsf{MI}(\rho_{0,k}) = \mathsf{H}(\rho_k) - \mathbb{E}_{\rho_0}[\mathsf{H}(\rho_{k|0})]$. It follows from (87) that $\rho_k = \mathcal{N}(m_k, c_k^2 I)$ with

$$m_k = 0 \qquad \text{and} \qquad c_k^2 - \frac{1}{\alpha} = \frac{1}{(1+\alpha\eta)^{2k}}\left(1 - \frac{1}{\alpha}\right). \tag{88}$$

It remains to evaluate $\rho_{k|0}$ in order to compute $\mathbb{E}_{x\sim\rho_0}[\mathsf{H}(\rho_{k|0=x})]$. Consider a fixed $x \in \mathbb{R}^d$ and suppose we wish to compute $\rho_{k|0=x}$. Then we can view this distribution as the solution along the Proximal Sampler when started from $\mathcal{N}(x, c^2 I)$ as $c^2 \to 0$. Using (87) again, we obtain

$$\rho_{k|0=x} = \mathcal{N}\left(\frac{x}{(1+\alpha\eta)^k}, \frac{1}{\alpha}\left(1 - \frac{1}{(1+\alpha\eta)^{2k}}\right) I\right). \tag{89}$$

Using (72) and the formula for entropy of a Gaussian on (88) and (89), we therefore get that

$$\begin{aligned}
\mathsf{MI}(X_0; X_k) &\overset{(72)}{=} \mathsf{H}(\rho_k) - \mathop{\mathbb{E}}_{\rho_0}[\mathsf{H}(\rho_{k|0})] \\
&\overset{(88),(89)}{=} \frac{d}{2} \log\left[\frac{(1+\alpha\eta)^{2k} + \alpha - 1}{(1+\alpha\eta)^{2k} - 1}\right] \\
&= \frac{d}{2} \log\left[1 + \frac{\alpha}{(1+\alpha\eta)^{2k} - 1}\right].
\end{aligned}$$

This proves the desired claim. □

This shows the tightness of Theorem 5.

APPENDIX P
GOING BEYOND STRONG-LOG CONCAVITY

In this section we prove Lemma 5, which bounds the contraction of mutual information along Langevin dynamics under a log-Sobolev inequality assumption.

Proof of Lemma 5: Beginning from Definition 2, we have the following

$$\begin{aligned}
\mathsf{MI}(\rho_{0,t}) &= \mathop{\mathbb{E}}_{x\sim\rho_0}[\mathsf{KL}(\rho_{t|0=x} \,\|\, \rho_t)] \\
&\overset{(31)}{=} \mathop{\mathbb{E}}_{x\sim\rho_0}\left[\mathsf{KL}(\rho_{t|0=x} \,\|\, \nu)\right] - \mathsf{KL}(\rho_t \,\|\, \nu) \\
&\leq \mathop{\mathbb{E}}_{x\sim\rho_0}\left[\mathsf{KL}(\rho_{t|0=x} \,\|\, \nu)\right] \\
&\leq e^{-2\alpha(t-s)} \mathop{\mathbb{E}}_{x\sim\rho_0}\left[\mathsf{KL}(\rho_{s|0=x} \,\|\, \nu)\right]
\end{aligned}$$

$$= e^{-2\alpha(t-s)} \mathop{\mathbb{E}}_{x\sim\rho_0} \left[\mathsf{KL}(\rho_{s|0=x} \,\|\, \nu) - \mathsf{KL}(\rho_s \,\|\, \nu) + \mathsf{KL}(\rho_s \,\|\, \nu) \right]$$
$$\overset{(31)}{=} e^{-2\alpha(t-s)} \left[\mathsf{MI}(\rho_{0,s}) + \mathsf{KL}(\rho_s \,\|\, \nu) \right],$$

where the first inequality follows from the non-negativity of KL divergence and the second inequality is from Lemma 9 with $\Phi(x) = x \log x$. □

ACKNOWLEDGMENT

The authors thank Vishwak Srinivasan for helpful comments, and to Sekhar Tatikonda and Sinho Chewi for feedback on an earlier version of the article.

REFERENCES

[1] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. Boca Raton, FL, USA: CRC Press, 1995.
[2] U. von Toussaint, "Bayesian inference in physics," *Rev. Mod. Phys.*, vol. 83, no. 3, pp. 943–999, 2011.
[3] M. Johannes and N. Polson, "MCMC methods for continuous-time financial econometric," in *Handbook of Financial Econometrics*, vol. 2. Amsterdam, The Netherlands: Elsevier, 2010.
[4] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. Boca Raton, FL, USA: CRC Press, 1995.
[5] R. E. Kass, B. P. Carlin, A. Gelman, and R. M. Neal, "Markov chain Monte Carlo in practice: A roundtable discussion," *Amer. Statistician*, vol. 52, no. 2, pp. 93–100, May 1998.
[6] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, "Introducing Markov chain Monte Carlo," in *Markov Chain Monte Carlo in Practice*, vol. 1. London, U.K.: Chapman & Hall, 1995.
[7] C. J. Geyer, "Introduction to Markov chain Monte Carlo," in *Handbook of Markov Chain Monte Carlo*, vol. 20116022. London, U.K.: Chapman & Hall, 2011, p. 22.
[8] D. A. Levin, Y. Peres, and E. Wilmer, *Markov Chains and Mixing Times*, vol. 107. Providence, RI, USA: American Mathematical Soc., 2017.
[9] D. Chafaï, "Entropies, convexity, and functional inequalities, On Φ-entropies and Φ-sobolev inequalities," *J. Math*, vol. 44, no. 2, pp. 325–363, 2004.
[10] M. Raginsky, "Strong data processing inequalities and Φ-sobolev inequalities for discrete channels," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3355–3389, Jun. 2016.
[11] R. Montenegro and P. Tetali, "Mathematical aspects of mixing times in Markov chains," *Found. Trends Theor. Comput. Sci.*, vol. 1, no. 3, pp. 237–354, 2005.
[12] S. Chewi. (2024). *Log-Concave Sampling*. [Online]. Available: https://chewisinho.github.io
[13] N. Madras and G. Slade, *The Self-Avoiding Walk*. Boston, MA, USA: Birkhäuser, 1993.
[14] N. Madras and D. Randall, "Markov chain decomposition for convergence rate analysis," *Ann. Appl. Probab.*, vol. 12, no. 2, pp. 581–606, May 2002.
[15] Y. Zhang, X. Cheng, and G. Reeves, "Convergence of Gaussian-smoothed optimal transport distance with sub-gamma distributions and dependent samples," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 2422–2430.
[16] Z. Goldfeld, "Estimating information flow in deep neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2299–2308.
[17] C. Villani, *Optimal Transport: Old and New*. Cham, Switzerland: Springer, 2009.
[18] J. Dolbeault and X. Li, "φ-entropies: Convexity, coercivity and hypocoercivity for Fokker–Planck and kinetic Fokker–Planck equations," *Math. Models Methods Appl. Sci.*, vol. 28, no. 13, pp. 2637–2666, Dec. 2018.
[19] Y. Cao, J. Lu, and Y. Lu, "Exponential decay of Rényi divergence under Fokker–Planck equations," *J. Stat. Phys.*, vol. 176, no. 5, pp. 1172–1184, Sep. 2019.
[20] A. S. Dalalyan, "Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent," in *Proc. Conf. Learn. Theory*, 2017, pp. 678–689.
[21] X. Cheng and P. L. Bartlett, "Convergence of Langevin MCMC in KL-divergence," in *Proc. Algorithmic Learn. Theory*, 2018, pp. 186–211.
[22] A. Durmus, S. Majewski, and B. Miasojedow, "Analysis of Langevin Monte Carlo via convex optimization," *J. Mach. Learn. Res.*, vol. 20, no. 73, pp. 1–46, 2018.
[23] S. Vempala and A. Wibisono, "Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8094–8106.
[24] S. Mitra and A. Wibisono, "Fast convergence of Φ-divergence along the unadjusted Langevin algorithm and proximal sampler," in *Proc. 36th Int. Conf. Algorithmic Learn. Theory*, 2024, pp. 846–869.
[25] Y. T. Lee, R. Shen, and K. Tian, "Structured logconcave sampling with a restricted Gaussian Oracle," in *Proc. Conf. Learn. Theory*, 2020, pp. 2993–3050.
[26] Y. Chen, S. Chewi, A. Salim, and A. Wibisono, "Improved analysis for a proximal algorithm for sampling," in *Proc. Conf. Learn. Theory*, 2022, pp. 2984–3014.
[27] B. Yuan, J. Fan, J. Liang, A. Wibisono, and Y. Chen, "On a class of Gibbs sampling over networks," in *Proc. Thirty 6th Conf. Learn. Theory*, 2023, pp. 5754–5780.
[28] J. Fan, B. Yuan, and Y. Chen, "Improved dimension dependence of a proximal algorithm for sampling," in *Proc. Thirty 6th Annu. Conf. Learn. Theory*, 2023, pp. 1473–1521.
[29] Y. Polyanskiy and Y. Wu, *Information Theory: From Coding to Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2025.
[30] Y. Polyanskiy and Y. Wu, "Strong data-processing inequalities for channels and Bayesian networks," in *Convexity and Concentration*. Cham, Switzerland: Springer, 2017, pp. 211–249.
[31] I. Sason and S. Verdú, "f -Divergence inequalities," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 5973–6006, Nov. 2016.
[32] D. Bakry, I. Gentil, and M. Ledoux, *Analysis and Geometry of Markov Diffusion Operators*, vol. 103. Cham, Switzerland: Springer, 2014.
[33] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Inf. Control*, vol. 2, no. 2, pp. 101–112, Jun. 1959.
[34] B. Klartag and O. Ordentlich, "The strong data processing inequality under the heat flow," *IEEE Trans. Inf. Theory*, vol. 71, no. 5, pp. 3317–3333, May 2025.
[35] F. Bolley and I. Gentil, "Phi-entropy inequalities for diffusion semigroups," *J. De Mathématiques Pures Et Appliquées*, vol. 93, no. 5, pp. 449–473, May 2010.
[36] F. Achleitner, A. Arnold, and D. Stürzer, "Large-time behavior in non-symmetric fokker-Planck equations," *Rivista di Matematica della Universit di Parma*, vol. 6, no. 1, pp. 1–68, 2015.
[37] G. O. Roberts and R. L. Tweedie, "Exponential convergence of Langevin distributions and their discrete approximations," *Bernoulli*, vol. 2, no. 4, pp. 341–363, Dec. 1996.
[38] G. O. Roberts and J. S. Rosenthal, "Optimal scaling of discrete approximations to Langevin diffusions," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 60, no. 1, pp. 255–268, Jan. 1998.
[39] A. S. Dalalyan, "Theoretical guarantees for approximate sampling from smooth and log-concave densities," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 79, no. 3, pp. 651–676, Jun. 2017.
[40] A. Durmus and É. Moulines, "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm," *Ann. Appl. Probab.*, vol. 27, no. 3, p. 1551, Jun. 2017.
[41] S. Chewi, M. A. Erdogdu, M. B. Li, R. Shen, and M. Zhang, "Analysis of Langevin Monte Carlo from Poincaré to log-Sobolev," in *Proc. 35th Conf. Learn. Theory*, 2022, pp. 1–2.
[42] J. M. Altschuler and K. Talwar, "Resolving the mixing time of the Langevin algorithm to its stationary distribution for log-concave sampling," in *Proc. 36th Conf. Learn. Theory*, 2022, pp. 993–1020.
[43] J. Liang and Y. Chen, "A proximal algorithm for sampling from non-smooth potentials," in *Proc. Winter Simulation Conf. (WSC)*, Dec. 2022, pp. 3229–3240.
[44] J. Liang and Y. Chen, "A proximal algorithm for sampling," *Trans. Mach. Learn. Res.*, 2022.
[45] Y. Kook, S. Vempala, and M. Zhang, "In-and-out: Algorithmic diffusion for sampling convex bodies," in *Proc. Adv. Neural Inf. Process. Syst. 37*, 2024, pp. 108354–108388.
[46] A. Pensia, V. Jog, and P. Loh, "The sample complexity of simple binary hypothesis testing," in *Proc. 37th Conf. Learn. Theory*, Jul. 2024, pp. 4205–4206.
[47] L. Györfi and I. Vajda, "Asymptotic distributions for goodness-of-fit statistics in a sequence of multinomial models," *Statist. Probab. Lett.*, vol. 56, no. 1, pp. 57–67, Jan. 2002.
[48] A. Gretton and L. Györfi, "Nonparametric independence tests: Space partitioning and kernel approaches," in *Proc. Int. Conf. Algorithmic Learn. Theory*, Oct. 2008, pp. 183–198.
[49] C. P. Ho, M. Petrik, and W. Wiesemann, "Robust Φ-divergence MDPs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 32680–32693.

[50] K. Panaganti, A. Wierman, and E. Mazumdar, "Model-free robust Φ-divergence reinforcement learning using both offline and online data," in *Proc. 41st Int. Conf. Mach. Learn.*, 2024, pp. 39324–39363.
[51] I. Nemenman, W. Bialek, and R. de Ruyter van Steveninck, "Entropy and information in neural spike trains: Progress on the sampling problem," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 5, May 2004, Art. no. 056111.
[52] A. Belitski et al., "Local field potentials and spiking activity in primary visual cortex convey independent information about natural stimuli," *J. Neurosci.*, vol. 28, no. 22, pp. 5696–5709, 2008.
[53] S. Asoodeh, M. Aliakbarpour, and F. P. Calmon, "Local differential privacy is equivalent to contraction of an f-divergence," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 545–550.
[54] B. Zamanlooy, S. Asoodeh, M. Diaz, and F. P. Calmon, "Eγ-mixing time," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2024, pp. 3474–3479.
[55] Y. Lu, G. Zhang, S. Sun, H. Guo, and Y. Yu, "f-MICL: Understanding and generalizing InfoNCE-based contrastive learning," *Trans. Mach. Learn. Res.*, 2024.
[56] A. R. Esposito, M. Gastpar, and I. Issa, "Robust generalization via f-Mutual information," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2723–2728.
[57] E. Erkip and T. M. Cover, "The efficiency of investment information," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1026–1040, May 1998.
[58] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by erkip and cover," 2013, *arXiv:1304.6133*.
[59] R. C. Bradley, "Basic properties of strong mixing conditions," in *Dependence in Probability and Statistics: A Survey of Recent Results*. New York, NY, USA: Springer, 1986, pp. 165–192.
[60] R. C. Bradley, "Basic properties of strong mixing Conditions. A survey and some open questions," *Probab. Surv.*, vol. 2, pp. 107–144, Jan. 2005.
[61] C. C. Margossian and A. Gelman, "For how many iterations should we run Markov chain Monte Carlo?" 2023, *arXiv:2311.02726*.
[62] C. C. Margossian, M. D. Hoffman, P. Sountsov, L. Riou-Durand, A. Vehtari, and A. Gelman, "Nested R: Assessing the convergence of Markov chain Monte Carlo when running many short chains," *Bayesian Anal.*, vol. 20, no. 4, pp. 1–28, Dec. 2025.
[63] D. Guo, S. Shamai (Shitz), and S. Verdu, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261–1282, Apr. 2005.
[64] R. Atar and T. Weissman, "Mutual information, relative entropy, and estimation in the Poisson channel," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1302–1318, Mar. 2012.
[65] A. Wibisono, V. Jog, and P.-L. Loh, "Information and estimation in fokker-Planck channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 2673–2677.
[66] L. Fan, J. Zou, J. Gao, and J. Wang, "Differential properties of information in jump-diffusion channels," 2025, *arXiv:2501.05708*.
[67] A. Wibisono and V. Jog, "Convexity of mutual information along the heat flow," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 1615–1619.
[68] A. Wibisono and V. Jog, "Convexity of mutual information along the Ornstein–Uhlenbeck flow," in *Proc. Int. Symp. Inf. Theory Appl. (ISITA)*, Oct. 2018, pp. 55–59.
[69] J. Zou, L. Fan, J. Gao, and J. Wang, "Convexity of mutual information along the fokker-Planck flow," 2025, *arXiv:2501.05094*.
[70] R. Jordan, D. Kinderlehrer, and F. Otto, "The variational formulation of the fokker-Planck equation," *SIAM J. Math. Anal.*, vol. 29, no. 1, pp. 1–17, Jan. 1998.
[71] C. Villani, *Topics in Optimal Transportation*, vol. 58. Providence, RI, USA: American Mathematical Soc., 2021.
[72] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 2524–2533.
[73] Y. Kook and S. S. Vempala, "Sampling and integration of logconcave functions by algorithmic diffusion," in *Proc. 57th Annu. ACM Symp. Theory Comput.*, Jun. 2025, pp. 924–932.
[74] J. Liang and Y. Chen, "Proximal Oracles for optimization and sampling," 2024, *arXiv:2404.02239*.
[75] J. M. Altschuler and S. Chewi, "Faster high-accuracy log-concave sampling via algorithmic warm starts," *J. ACM*, vol. 71, no. 3, pp. 1–55, Jun. 2024.
[76] S. Chewi, P. Gerber, C. Lü, T. L. Gouic, and P. Rigollet, "The query complexity of sampling from strongly log-concave distributions in one dimension," in *Proc. Conf. Learn. Theory*, 2021, pp. 2041–2059.
[77] J. Boursier, D. Chafaï, and C. Labbé, "Universal cutoff for Dyson Ornstein Uhlenbeck process," *Probab. Theory Rel. Fields*, vol. 185, nos. 1–2, pp. 449–512, Feb. 2023.
[78] F. Koehler, H. Lee, and T.-D. Vuong, "Efficiently learning and sampling multimodal distributions with data-based initialization," in *Proc. Conf. Learn. Theory*, 2024, pp. 3264–3326.
[79] N. Bou-Rabee and A. Eberle, "Mixing time guarantees for unadjusted Hamiltonian Monte Carlo," *Bernoulli*, vol. 29, no. 1, pp. 75–104, Feb. 2023.
[80] S. G. Bobkov, I. Gentil, and M. Ledoux, "Hypercontractivity of Hamilton–Jacobi equations," *J. De Mathématiques Pures Et Appliquées*, vol. 80, no. 7, pp. 669–696, 2001.
[81] J. M. Altschuler and S. Chewi, "Shifted composition I: Harnack and reverse transport inequalities," *IEEE Trans. Inf. Theory*, vol. 71, no. 1, pp. 90–113, Jan. 2025.
[82] F. Otto and C. Villani, "Comment on: 'Hypercontractivity of Hamilton–Jacobi equations, by S. Bobkov, I. Gentil and M. Ledoux," *J. De Mathématiques Pures Et Appliquées*, vol. 80, no. 7, pp. 697–700, 2001.
[83] A. Saumard and J. A. Wellner, "Log-concavity and strong log-concavity: A review," *Statist. Surveys*, vol. 8, no. none, p. 45, Jan. 2014.
[84] T. M. Cover, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1999.
[85] N. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. Inf. Theory*, vol. IT-11, no. 2, pp. 267–271, Apr. 1965.

Jiaming Liang received the B.S. degree in applied mathematics and ocean engineering from Shanghai Jiao Tong University in 2015 and the M.S. degree in computational science and engineering and the Ph.D. degree in operations research from Georgia Institute of Technology in 2017 and 2022, respectively. He was a Post-Doctoral Researcher with Yale University. He is currently an Assistant Professor of computer science and data science with the University of Rochester. His research interests include theory and algorithms for optimization, sampling, and optimal transport.

Siddharth Mitra received the B.Sc. degree in mathematics and computer science and the M.Sc. degree in computer science from Chennai Mathematical Institute. He is currently pursuing the Ph.D. degree in computer science with Yale University.

Andre Wibisono received the dual B.S. degree in mathematics and computer science from MIT, the master's degree in computer science from MIT, the master's degree in statistics from UC Berkeley, and the Ph.D. degree in computer science from UC Berkeley. He is currently an Assistant Professor with the Department of Computer Science, Yale University, with a secondary appointment at the Department of Statistics and Data Science. Before joining Yale, he has done postdoctoral research at UW Madison and Georgia Tech. His research interests include the design and analysis of algorithms for machine learning, in particular for problems in optimization, sampling, and game theory.