Improved analysis for a proximal algorithm for sampling

Yongxin Chen YONGCHEN@GATECH.EDU
Georgia Institute of Technology

Sinho Chewi SCHEWI@MIT.EDU
Massachusetts Institute of Technology

Adil Salim ADILSALIM@MICROSOFT.COM
Microsoft Research

Andre Wibisono ANDRE.WIBISONO@YALE.EDU
Yale University

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We study the proximal sampler of Lee et al. (2021a) and obtain new convergence guarantees under weaker assumptions than strong log-concavity: namely, our results hold for (1) weakly log-concave targets, and (2) targets satisfying isoperimetric assumptions which allow for non-log-concavity. We demonstrate our results by obtaining new state-of-the-art sampling guarantees for several classes of target distributions. We also strengthen the connection between the proximal sampler and the proximal method in optimization by interpreting the proximal sampler as an entropically regularized Wasserstein proximal method, and the proximal point method as the limit of the proximal sampler with vanishing noise.

Keywords: functional inequality, isoperimetry, optimization, proximal point method, proximal sampler, sampling

1. Introduction

The problem of sampling from a target density $\pi^X \propto \exp(-f)$ on $\mathbb{R}^d$ has seen a resurgence of interest due to its staple role in scientific computing (Robert and Casella, 2004), as well as its surprising and deep connections with the field of optimization. Indeed, the standard Langevin algorithm can be viewed as a gradient flow of the Kullback–Leibler (KL) divergence on the space of probability measures equipped with the geometry of optimal transport, a perspective which has led to new analyses (Durmus et al., 2019; Salim and Richtarik, 2020; Ahn and Chewi, 2021) and algorithms (Pereyra, 2016; Zhang et al., 2020; Ding and Li, 2021; Ma et al., 2021) inspired by the theory of convex optimization.

Among the algorithms in the optimization toolkit, we focus on *proximal* methods. Classically, proximal methods are used to minimize composite objectives of the form $f + g$, where g is smooth and convex and f is non-smooth but simple enough to allow for evaluation of the proximal map $\mathrm{prox}_f : y \mapsto \arg\min_{x \in \mathbb{R}^d} \{f(x) + \frac{1}{2\eta} \|x - y\|^2\}$. However, the setting of our investigation is more closely related to the minimization of a non-composite objective f, for which the proximal method is known as the *proximal point algorithm* (Martinet, 1970; Rockafellar, 1976).

As a natural first step towards developing a proximal point algorithm for sampling, one can combine the proximal map with the standard Langevin algorithm, leading to the *proximal Langevin algorithm*. This algorithm was introduced in Pereyra (2016) and analyzed in the papers Bernton

(2018); Wibisono (2019); Salim and Richtarik (2020). Although these results are encouraging, the analogy between optimization methods and Langevin-based algorithms is imperfect because the discretization of the latter leads to asymptotic *bias*, a feature which is typically not present in optimization (see Wibisono (2018) for a thorough discussion).

Remarkably, a new proximal algorithm for sampling was proposed recently in Lee et al. (2021a) which overcomes this issue via a novel Gibbs sampling approach. Briefly, the *proximal sampler* is a sampling algorithm which assumes access to samples from an oracle distribution, known as the *restricted Gaussian oracle* (RGO); the RGO is a sampling analogue of the proximal map from optimization. Under this assumption, as well as the additional assumption that the target π^X is strongly log-concave, Lee et al. (2021a) proved[1] that the proximal sampler converges exponentially fast to π^X in total variation distance. In their paper, the proximal sampler was used as a *reduction framework* to improve the condition number dependence of other sampling algorithms. Indeed, the RGO is a better conditioned distribution than the target distribution, so that implementing the RGO is easier than solving the original sampling task. In turn, the reduction framework allowed them to establish improved complexity results for a variety of structured log-concave sampling problems. We review the proximal sampler and its implementability in Section 3.

Our contributions. Prior to our work, the convergence of the proximal sampler was only known in the case when $\pi^X \propto \exp(-f)$ is strongly log-concave. In this paper, we greatly expand the classes of targets to which the proximal sampler is applicable by providing new convergence guarantees.

First, we consider the case when f is weakly convex. We show that after k iterations, the proximal sampler outputs a distribution whose KL divergence to the target is $O(1/k)$. Our proof is analogous to, and is inspired by, the corresponding guarantee for minimizing a weakly convex function (in particular, the $O(1/k)$ rate matches the optimization result).

Next, we assume that π^X satisfies a *functional inequality*, e.g., a Poincaré inequality or a log-Sobolev inequality. Such functional inequalities have been employed in the sampling literature as tractable settings for non-log-concave sampling; see Vempala and Wibisono (2019); Chewi et al. (2021a). For these distributions, we show that the proximal sampler converges to the target in Rényi divergence (or any other weaker metric, such as KL divergence) with a rate that matches the known convergence rates for the continuous-time Langevin diffusion under the same assumptions.

In each of these settings, if we additionally assume that ∇f is Lipschitz, then the RGO is implementable, as it becomes a smooth strongly log-concave distribution. Hence, we obtain new sampling guarantees for gradient Lipschitz potentials when the target is weakly log-concave or satisfies a functional inequality. In all cases, our results are *stronger* than known results in the literature. Subsequent works have also considered implementability of the RGO under weaker smoothness conditions (Liang and Chen, 2021; Gopi et al., 2022; Liang and Chen, 2022a,b).

Finally, we clarify the connection between the proximal sampler and the proximal point algorithm in optimization in the following ways: (1) We show that convergence proofs for the proximal sampler can be translated to yield convergence proofs for the proximal point algorithm. As a consequence, we obtain a new convergence guarantee for the proximal point method under a gradient domination condition with optimal rate, which is (to the best of our knowledge) a new result. (2) We show that the RGO can be interpreted as a proximal mapping on Wasserstein space, and that the proximal sampler can be interpreted as an entropically regularized Wasserstein proximal method (i.e., JKO

1. There is an error in the conference version of the paper which is fixed in the arXiv version (Lee et al., 2021b).

scheme). The latter perspective allows us to recover the proximal point algorithm as a certain limit of the proximal sampler as the "noise level" (corresponding to the entropic regularization) tends to zero.

Other related work. Sampling algorithms which are conceptually similar or directly related to the proximal sampler have been previously proposed in the literature (Girolami and Calderhead, 2011; Marnissi et al., 2016; Titsias and Papaspiliopoulos, 2018; Vono et al., 2022). The RGO has also been considered as an adjoint of the heat semigroup in Klartag and Putterman (2021), which was then used in the recent breakthrough on the KLS conjecture in Klartag and Lehec (2022). After the first version of our work appeared online, our result under LSI (Theorem 3) was recovered via the framework of localization schemes in Chen and Eldan (2022).

Organization. The rest of the paper is organized as follows. We begin with background on distances between probability measures in Section 2 and on the proximal sampler in Section 3.

We give our main results in Section 4. In particular, we state our new convergence guarantees for the proximal sampler in Section 4.1, and we give applications of our results in Section 4.2. We then describe the connections between the proximal sampler and the proximal point method in Section 4.3. All proofs are given in Section A.

Finally, we conclude and list open directions in Section 5.

2. Background and notation

Throughout the paper, we abuse notation by identifying a probability measure with its density w.r.t. Lebesgue measure. For a probability measure $\rho \ll \pi$, we define the *KL divergence*, the *chi-squared divergence*, and the *Rényi divergence* of order $q \geq 1$ respectively via

$$H_\pi(\rho) \coloneqq \int \rho \log \frac{\rho}{\pi}\,, \qquad \chi^2_\pi(\rho) \coloneqq \int \frac{\rho^2}{\pi} - 1\,, \qquad R_{q,\pi}(\rho) \coloneqq \frac{1}{q-1} \log \int \frac{\rho^q}{\pi^{q-1}}\,,$$

with $R_{1,\pi} = H_\pi$. We recall that for $1 \leq q \leq q' < \infty$, we have the monotonicity property $R_{q,\pi} \leq R_{q',\pi}$, and that $R_{2,\pi} = \log(1 + \chi^2_\pi)$.

We also define the 2-Wasserstein distance between ρ and π to be

$$W_2^2(\rho, \pi) \coloneqq \inf_{\gamma \in \mathcal{C}(\rho,\pi)} \int \|x - y\|^2 \, d\gamma(x, y)\,,$$

where $\mathcal{C}(\rho, \pi)$ is the set of *couplings* of ρ and π, i.e., joint distributions on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are ρ and π. We refer readers to Villani (2003) for an introduction to optimal transport, and to Ambrosio et al. (2008) for a detailed treatment of Wasserstein calculus.

3. The proximal sampler

Our goal is to sample from a target probability distribution π^X on $\mathbb{R}^d$ with density $\pi^X \propto \exp(-f)$ and finite second moment, where $f \colon \mathbb{R}^d \to \mathbb{R}$ is the *potential*.

Following Lee et al. (2021a), we define the joint target distribution π on $\mathbb{R}^d \times \mathbb{R}^d$ with density

$$\pi(x, y) \propto \exp\Big(-f(x) - \frac{1}{2\eta} \|x - y\|^2\Big)\,,$$

where $\eta > 0$ is the *step size* of the algorithm.

Observe that the X-marginal of π is equal to the original target distribution π^X, whereas the conditional distribution of Y given X is Gaussian: $\pi^{Y|X}(\cdot \mid x) = \mathcal{N}(x, \eta I)$. Therefore, the Y-marginal is the convolution of π^X with a Gaussian, $\pi^Y = \pi^X * \mathcal{N}(0, \eta I)$. The perspective that we adopt in our proofs is that π^Y is obtained by evolving π^X along the heat flow for time η.

The conditional distribution of X given Y is the "regularized" distribution

$$\pi^{X|Y}(x \mid y) \propto_x \exp\Big(-f(x) - \frac{1}{2\eta}\|x - y\|^2\Big)\,.$$

The *restricted Gaussian oracle* (RGO) is defined as an oracle that, given $y \in \mathbb{R}^d$, outputs a random variable distributed according to $\pi^{X|Y}(\cdot \mid y)$. We also write $\pi^{X|Y}(\cdot \mid y) = \pi^{X|Y=y}$.

Proximal Sampler: The proximal sampler is initialized at a point $x_0 \in \mathbb{R}^d$ and performs Gibbs sampling on the joint target π. That is, the proximal sampler iterates the following two steps:

1. From x_k, sample $y_k \mid x_k \sim \pi^{Y|X}(\cdot \mid x_k) = \mathcal{N}(x_k, \eta I)$.
2. From y_k, sample $x_{k+1} \mid y_k \sim \pi^{X|Y}(\cdot \mid y_k)$.

The first step consists in sampling a Gaussian random variable centered at x_k, and is therefore easy to implement. The second step calls the RGO at the point y_k.

As is well-known from the theory of Gibbs sampling, the iterates $(x_k, y_k)_{k\in\mathbb{N}}$ form a reversible Markov chain with stationary distribution π. That is, the proximal sampler is an *unbiased* sampling algorithm, unlike algorithms based on discretizations of stochastic processes such as the unadjusted Langevin algorithm. This is because the proximal sampler is an idealized algorithm in which we assume *exact* access to the RGO. For our applications, we implement the RGO via rejection sampling; see Section 4.2 for details and Section 4.4 for an explicit example in the Gaussian case.

4. Results

4.1. New convergence results for the proximal sampler

In this section, we describe our new convergence results for the proximal sampler under various assumptions, beginning with the strongly log-concave and weakly log-concave cases, and then proceeding to targets satisfying functional inequalities which allow for non-log-concavity.

4.1.1. STRONG LOG-CONCAVITY

We start by recalling the W_2 contraction result from Lee et al. (2021b) for the proximal sampler under strong log-concavity.

Theorem 1 (Lee et al. (2021b, Lemma 2)) *Assume that $\pi^X \propto \exp(-f)$ is α-strongly log-concave (i.e., f is α-strongly convex), where $\alpha \geq 0$. For any $\eta > 0$ and for any two initial distributions ρ_0^X, $\bar{\rho}_0^X$, after k iterations of the proximal sampler with step size η, the respective distributions ρ_k^X, $\bar{\rho}_k^X$ satisfy the bound*

$$W_2(\rho_k^X, \bar{\rho}_k^X) \leq \frac{W_2(\rho_0^X, \bar{\rho}_0^X)}{(1+\alpha\eta)^k}\,. \tag{1}$$

Although this result was stated in Lee et al. (2021b) as a convergence result rather than a contraction, the latter is implicit in the proof. From the proof of Lee et al. (2021b), one can also read off a convergence guarantee in KL divergence, although this will be a corollary of our result in Section 4.1.3.

We revisit Theorem 1 in Section A.2 and provide a proof which more closely resembles a classical convergence proof of the proximal point algorithm. We use Wasserstein subdifferential calculus.

We note that this is the sampling analogue of the classical fact that the proximal map for an α-strongly convex function with step size η is a $\frac{1}{1+\alpha\eta}$-contraction. In Appendix B.1, we give a new proof of this fact by translating the proof of Lee et al. (2021b) into optimization.

4.1.2. LOG-CONCAVITY

The preceding result does not yield convergence when $\alpha = 0$. We provide a new convergence guarantee for the weakly convex case which mirrors a Lyapunov analysis of gradient flows for convex functions.

Theorem 2 *Assume that $\pi^X \propto \exp(-f)$ is log-concave (i.e., f is convex). For the k-th iterate ρ_k^X of the proximal sampler,*

$$H_{\pi^X}(\rho_k^X) \leq \frac{W_2^2(\rho_0^X, \pi^X)}{k\eta}\,.$$

Proof Section A.3. ■

4.1.3. LOG-SOBOLEV INEQUALITY

Recall that a probability distribution π satisfies the log-Sobolev inequality (LSI) with constant $\alpha > 0$ (α-LSI) if for any probability distribution ρ, the following inequality holds:

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

Here $J_\pi(\rho)$ is the Fisher information of ρ w.r.t. π; see Section A.4. Recall that strong log-concavity implies LSI, and that LSI is equivalent to the gradient domination condition for relative entropy H_π (Otto and Villani, 2000); see also Section 4.3.1.

Theorem 3 *Assume that $\pi^X \propto \exp(-f)$ satisfies α-LSI. For any $\eta > 0$ and any initial distribution ρ_0^X, the k-th iterate ρ_k^X of the proximal sampler with step size η satisfies*

$$H_{\pi^X}(\rho_k^X) \leq \frac{H_{\pi^X}(\rho_0^X)}{(1+\alpha\eta)^{2k}}\,. \tag{3}$$

Furthermore, for all $q \geq 1$:

$$R_{q,\pi^X}(\rho_k^X) \leq \frac{R_{q,\pi^X}(\rho_0^X)}{(1+\alpha\eta)^{2k/q}}\,. \tag{4}$$

Proof Section A.4. ■

4.1.4. POINCARÉ INEQUALITY

Recall that a probability distribution π satisfies the Poincaré inequality (PI) with constant $\alpha > 0$ (α-PI) if for any smooth bounded function $\psi : \mathbb{R}^d \to \mathbb{R}$, the following inequality holds:

$$\mathrm{var}_\pi(\psi) \leq \frac{1}{\alpha}\,\mathbb{E}_\pi[\|\nabla\psi\|^2]\,. \tag{5}$$

Recall also that α-LSI implies α-PI.

Theorem 4 *Assume $\pi^X \propto \exp(-f)$ satisfies α-PI. For any $\eta > 0$ and any initial distribution ρ_0^X, the k-th iterate ρ_k^X of the proximal sampler with step size η satisfies*

$$\chi^2_{\pi^X}(\rho_k^X) \leq \frac{\chi^2_{\pi^X}(\rho_0^X)}{(1+\alpha\eta)^{2k}}\,. \tag{6}$$

Furthermore, for all $q \geq 2$,

$$R_{q,\pi^X}(\rho_k^X) \leq \begin{cases} R_{q,\pi^X}(\rho_0^X) - \frac{2k\log(1+\alpha\eta)}{q}\,, & \text{if } k \leq \frac{q}{2\log(1+\alpha\eta)}\,(R_{q,\pi^X}(\rho_0^X)-1)\,, \\ 1/(1+\alpha\eta)^{2(k-k_0)/q}\,, & \text{if } k \geq k_0 \coloneqq \lceil \frac{q}{2\log(1+\alpha\eta)}\,(R_{q,\pi^X}(\rho_0^X)-1)\rceil\,. \end{cases} \tag{7}$$

Proof Section A.5. ■

4.1.5. LATAŁA–OLESZKIEWICZ INEQUALITY

We next consider a family of functional inequalities which interpolate between PI and LSI. A probability distribution π satisfies the Latała–Oleszkiewicz inequality (LOI) of order $r \in [1,2]$ and constant $\alpha > 0$ ((r,α)-LOI) if for any smooth bounded function $\psi : \mathbb{R}^d \to \mathbb{R}_+$, the following inequality holds:

$$\sup_{p\in(1,2)} \frac{\mathrm{var}_{p,\pi}(\psi)}{(2-p)^{2\,(1-1/r)}} \coloneqq \sup_{p\in(1,2)} \frac{\mathbb{E}_\pi[\psi^2] - \mathbb{E}_\pi[\psi^p]^{2/p}}{(2-p)^{2\,(1-1/r)}} \leq \frac{1}{\alpha}\,\mathbb{E}_\pi[\|\nabla\psi\|^2]\,.$$

This inequality was introduced in Latała and Oleszkiewicz (2000), and sampling guarantees for the Langevin algorithm under LOI were given in Chewi et al. (2021a). The LOI for $r = 1$ is equivalent to PI and the LOI for $r = 2$ is equivalent to LSI, up to absolute constants. Generally speaking, (r,α)-LOI captures targets $\pi \propto \exp(-f)$ such that the tails of f grow as $\|\cdot\|^r$.

Theorem 5 *Assume $\pi^X \propto \exp(-f)$ satisfies (r,α)-LOI with $r \in [1,2)$. For any $\eta > 0$, $q \geq 2$, and any initial distribution ρ_0^X, the k-th iterate ρ_k^X of the proximal sampler with step size η satisfies*

$$R_{q,\pi^X}(\rho_k^X) \leq \begin{cases} \left(R_{q,\pi^X}(\rho_0^X)^{2/r-1} - \frac{(2/r-1)\,k\log(1+\alpha\eta)}{68q}\right)^{r/(2-r)}\,, & \text{if } k \leq c_0\,, \\ 1/(1+\alpha\eta)^{(k-\lceil c_0\rceil)/(68q)}\,, & \text{if } k \geq \lceil c_0\rceil\,. \end{cases} \tag{8}$$

where

$$c_0 \coloneqq \frac{68q}{(2/r-1)\log(1+\alpha\eta)}\left(R_{q,\pi^X}(\rho_0^X)^{2/r-1} - 1\right).$$

(For $r = 2$, we can instead use Theorem 3.)

Proof Section A.6. ■

To interpret the result, suppose that $R_{q,\pi^X}(\rho_0^X) = O(d)$ at initialization and that $\eta \ll 1/\alpha$. Then, the theorem states that after an initial waiting period of $\lceil c_0 \rceil = O(d^{2/r-1}/\eta)$ iterations, in which the Rényi divergence decays to $O(1)$, the Rényi divergence decays exponentially thereafter. This interpolates between a waiting time of $O(d/\eta)$ under PI ($r = 1$; Theorem 4) and a waiting time of $O((\log d)/\eta)$ under LSI ($r = 2$; Theorem 3).

4.2. Applications of the convergence results

We start with a corollary of Theorem 2. Suppose that f is β-smooth, i.e., ∇f is β-Lipschitz. Then, provided $\frac{1}{\eta} \geq \beta$, the RGO $\pi^{X|Y}$ is strongly-log-concave, with condition number $(1+\beta\eta)/(1-\beta\eta)$. We can implement the RGO via rejection sampling.

Rejection Sampling: Given a target distribution $\tilde{\pi} \propto \exp(-\tilde{f})$, where $\tilde{f}$ is $\tilde{\alpha}$-strongly convex, perform the following steps.

1. Compute the minimizer $x^\star$ of $\tilde{f}$.
2. Repeat until acceptance: draw a random variable $Z \sim \mathcal{N}(x^\star, \tilde{\alpha}^{-1} I)$ and accept it with probability $\exp(-\tilde{f}(Z) + \tilde{f}(x^\star) + \frac{\tilde{\alpha}}{2}\|Z - x^\star\|^2)$.

The resulting sample is distributed according to $\tilde{\pi}$, and one can show that the expected number of iterations of the algorithm is bounded by $\tilde{\kappa}^{d/2}$ with $\tilde{\kappa} \coloneqq \tilde{\beta}/\tilde{\alpha}$ and $\tilde{\beta}$ is the smoothness of $\tilde{f}$; see, e.g., Chewi et al. (2021b, Theorem 7).

We apply this to $\tilde{f}$ given by $\tilde{f}(x) = f(x) + \frac{1}{2\eta}\|x - y\|^2$. The algorithm above requires exact minimization of $\tilde{f}$, which we assume for simplicity (since it is well-known how to efficiently minimize a strongly convex and smooth function). With the choice $\eta \asymp \frac{1}{\beta d}$, the expected number of iterations is $O(1)$. Combining this implementation of the RGO with Theorem 2, we obtain:

Corollary 6 *Suppose $\pi^X \propto \exp(-f)$ where f is convex and β-smooth. Take $\eta \asymp \frac{1}{\beta d}$ and implement the RGO with rejection sampling as described above. Then, the proximal sampler outputs ρ_k^X with $H_{\pi^X}(\rho_k^X) \leq \varepsilon$ and the expected number of calls to an oracle for f is $O(\beta d\, W_2^2(\rho_0^X, \pi^X)/\varepsilon)$.*

More precisely, our algorithm requires access to an oracle of f which can evaluate f and compute the proximity operator for f.

We now compare this rate with others in the literature. Let $\mathfrak{m}_2$ denote the second moment of π^X. For example, $\mathfrak{m}_2 = O(d)$ for a product measure, and $\mathfrak{m}_2 = O(d^2)$ when $f(x) = \sqrt{1+\|x\|^2}$. It is reasonable to assume that the Poincaré constant α of π^X is $\Omega(d/\mathfrak{m}_2)$ and that $W_2^2(\rho_0^X, \pi^X) = O(\mathfrak{m}_2)$. With these simplifications, our complexity is $O(\beta d \mathfrak{m}_2/\varepsilon)$; averaged LMC achieves $\widetilde{O}(\beta d \mathfrak{m}_2/\varepsilon^2)$ (Durmus et al., 2019); MALA achieves $\widetilde{O}(\beta^{3/2} d^{1/2} \mathfrak{m}_2^{3/2}/\varepsilon^{3/4})$ albeit in TV^2 (Dwivedi et al., 2019; Chen et al., 2020); and LMC achieves $\widetilde{O}(\beta^2 \mathfrak{m}_2^2/\varepsilon)$ in the stronger Rényi metric (Chewi et al., 2021a). Since all these complexity results also hold in terms of the squared total variation distance, our result has arguably the state-of-the-art complexity for this setting (at least, if dimension dependence is the primary consideration).

Similarly, implementing the RGO with rejection sampling in Theorem 5 yields:

Corollary 7 *Suppose $\pi^X \propto \exp(-f)$ where f is β-smooth and π^X satisfies (r, α)-LOI. Take $\eta \asymp \frac{1}{\beta d}$ and implement the RGO with rejection sampling as described above. Then, the proximal sampler outputs ρ_k^X with $R_{q,\pi^X}(\rho_k^X) \leq \varepsilon$ and the expected number of calls to an oracle for f is $\widetilde{O}(\frac{\beta dq}{\alpha}(R_{q,\pi^X}(\rho_0^X)^{2/r-1} \vee \log(1/\varepsilon)))$.*

Even for the special case of a Poincaré inequality and smoothness, the first sampling guarantee under these assumptions is quite recent (Chewi et al., 2021a). Let us write $\hat{\kappa} := \beta/\alpha$ for the "condition number" and assume $R_{q,\pi^X}(\rho_0^X) = O(d)$ (see, e.g., Chewi et al., 2021a, Appendix A). Then, our complexity is $\widetilde{O}(\hat{\kappa} dq\,(d^{2/r-1} \vee \log(1/\varepsilon)))$, whereas Chewi et al. (2021a, Theorem 7) gives a complexity bound for LMC of order $\widetilde{O}(\hat{\kappa}^2 d^{4/r-1} q^3/\varepsilon)$. We note that our result is the *first* high-accuracy guarantee for this setting (i.e., the complexity depends polylogarithmically on ε). Moreover, even in the low-accuracy regime $\varepsilon \asymp 1$, our complexity of $\widetilde{O}(\hat{\kappa} d^{2/r} q)$ is always better (e.g., in the Poincaré case $r = 1$, our rate is $\widetilde{O}(\hat{\kappa} d^2 q)$ whereas Chewi et al. (2021a) yields $\widetilde{O}(\hat{\kappa}^2 d^3 q^3)$), although we note that Chewi et al. (2021a) handles the more general weakly smooth case.

Surprisingly, the same strategy of rejection sampling also applies to non-smooth potentials. In Liang and Chen (2021), it was shown that when the above rejection sampling is applied to $\tilde{f}(x) = f(x) + \frac{1}{2\eta}\|x - y\|^2$ with $f(x)$ being a convex and M-Lipschitz function, if $\eta \leq 1/(16M^2 d)$, the expected number of iterations of the algorithm is bounded above by 2. Moreover, the result is insensitive to the inexactness of the minimizer of $\tilde{f}$ (Liang and Chen, 2021). Combining it with Theorem 2 and Theorem 4 we establish:

Corollary 8 *Suppose $\pi^X \propto \exp(-f)$ where f is convex and M-Lipschitz. Take $\eta \asymp \frac{1}{M^2 d}$ and implement the RGO with rejection sampling as described above.*

1. *Applying Theorem 2, we deduce that the proximal sampler outputs ρ_k^X with $H_{\pi^X}(\rho_k^X) \leq \varepsilon$ and the expected number of calls to an oracle for f is $O(M^2 d\, W_2^2(\rho_0^X, \pi^X)/\varepsilon)$.*

2. *Applying Theorem 4 (using the fact that log-concave measures satisfy α-PI for some $\alpha > 0$), we deduce that the proximal sampler outputs ρ_k^X with $R_{q,\pi^X}(\rho_k^X) \leq \varepsilon$ and the expected number of calls to an oracle for f is $O(\frac{M^2 dq}{\alpha}(R_{q,\pi^X}(\rho_0^X) \vee \log(1/\varepsilon)))$.*

We make the same simplifications as above to compare the rates. Our complexity (from the second part of Corollary 8 is $O(M^2 \mathfrak{m}_2\,(d \vee \log(1/\varepsilon)))$, whereas Durmus et al. (2019) achieves $O(M^2 \mathfrak{m}_2/\varepsilon^2)$ in KL divergence and Liang and Chen (2021) achieves $\widetilde{O}(M^2 d \mathfrak{m}_2/\varepsilon^{1/2})$ in squared total variation distance. In particular, when $\mathfrak{m}_2 = O(d)$, our result is the state-of-the-art.

We summarize the ways in which the proximal sampler improves upon the standard discretized Langevin algorithm.

1. Under weaker assumptions on the target π^X, such as a Poincaré inequality, the analysis of the Langevin algorithm is affected in two ways: first, the continuous-time convergence of the diffusion is slower; and second, the discretization analysis becomes much more challenging. In contrast, although the ideal proximal sampler also converges more slowly under weaker assumptions, the second issue is no longer present. In particular, regardless of the isoperimetric assumption on π^X, as soon as ∇f is Lipschitz we can implement the RGO via rejection sampling, yielding a simple analysis with strong convergence guarantees.

2. Related to the first point, it is currently not known how to perform a discretization analysis of the Langevin algorithm with linear dependence on the condition number $\kappa = \frac{\beta}{\alpha}$ under α-LSI or α-PI. Our results therefore constitute the first $O(\kappa)$ guarantees for such distributions.

3. When implemented via rejection sampling, the proximal sampler provides a new approach to obtaining high-accuracy guarantees for sampling (i.e., complexity guarantees with dependence $\mathrm{polylog}(1/\varepsilon)$ on the accuracy ε). The simplicity of the analysis makes it an attractive alternative to Metropolis–Hastings algorithms, whose analysis is often involved.

4. Finally, we mention that when the RGO is implemented via the Metropolized random walk (Dwivedi et al., 2019), the resulting algorithm only uses *zeroth-order* queries to f, which is crucial for certain applications (e.g., Bayesian inverse problems).

4.3. On the relation between the proximal sampler and the proximal point algorithm

The proximal sampler is motivated by the proximal point method in optimization. Recall that in optimization, the proximal point method for minimizing f is the iteration of the proximal mapping

$$\mathrm{prox}_{\eta f}(y) \coloneqq \arg\min_{x \in \mathbb{R}^d} \left\{ f(x) + \frac{1}{2\eta} \|x - y\|^2 \right\} \tag{9}$$

with some step size $\eta > 0$. Formally, using the correspondence $f \leftrightarrow \exp(-f)$ between optimization and sampling, the RGO can be viewed as the sampling analogue of the proximal mapping.

In this section, we establish a more precise correspondence between the proximal sampler algorithm (for sampling from $\exp(-f)$) and the proximal point method (for minimizing f).

4.3.1. Convergence under LSI/PL

We recall that LSI for $\pi \propto \exp(-f)$ is equivalent to the statement that the relative entropy H_π satisfies the gradient domination condition (or the Polyak–Łojasiewicz (PL) inequality) in the Wasserstein metric (Otto and Villani, 2000). Thus, in the optimization setting, the analogous assumption to LSI is that f satisfies PL.

We recall f satisfies the PL inequality with constant $\alpha > 0$ (α-PL) if for all x,

$$\|\nabla f(x)\|^2 \geq 2\alpha \, (f(x) - f^\star),$$

where $f^* = \inf f$. The PL inequality allows for mild non-convexity of f, yet still implies exponential convergence of gradient flow or proximal point method for minimizing f; see for example (Karimi et al., 2016).

In light of our convergence guarantee for the proximal sampler under LSI in Theorem 3, it is natural to ask whether there is an analogous result for the proximal point method under PL. We answer this affirmatively via the following theorem. We note that a less careful proof of the argument gives the suboptimal contraction factor $\frac{1}{1+\alpha\eta}$; to the best of our knowledge, we are not aware of another reference which obtains the optimal contraction factor under PL (Attouch and Bolte, 2009).[2]

2. The optimality of our bound can be obtained by considering $f(x) = \frac{\alpha}{2}\|x\|^2$.

Theorem 9 *Suppose that $f : \mathbb{R}^d \to (-\infty, +\infty]$ is differentiable and satisfies α-PL and let $x' \in \operatorname{prox}_{\eta f}(x)$. Also, write $f^\star = \inf f$. Then, it holds that*

$$f(x') - f^\star \leq \frac{1}{(1+\alpha\eta)^2} \{f(x) - f^\star\}.$$

Proof Section B.2. ■

4.3.2. RGO AS A PROXIMAL OPERATOR ON WASSERSTEIN SPACE

Consider $y \in \mathbb{R}^d$. Noting that $\pi^{X|Y=y}(dx) \propto_x \exp(-\frac{1}{2\eta}\|x-y\|^2)\, \pi^X(dx)$ and using Ambrosio et al. (2008, Remark 9.4.2) we have

$$H_{\pi^X}(\rho^X) = H_{\pi^{X|Y=y}}(\rho^X) - \int \frac{1}{2\eta} \|x-y\|^2 \, d\rho^X(x) + C(y),$$

where $C(y)$ is a constant depending only on y. Using $\arg\min H_{\pi^{X|Y=y}}(\cdot) = \pi^{X|Y=y}$, the RGO can be expressed as

$$\begin{aligned} \pi^{X|Y=y} &= \operatorname*{arg\,min}_{\rho^X \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ H_{\pi^X}(\rho^X) + \frac{1}{2\eta} \int \|x-y\|^2 \, d\rho^X(x) \right\} \\ &= \operatorname*{arg\,min}_{\rho^X \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ H_{\pi^X}(\rho^X) + \frac{1}{2\eta} W_2^2(\rho^X, \delta_y) \right\}. \end{aligned} \tag{10}$$

Thus, by replacing the Euclidean distance by the Wasserstein distance, $\pi^{X|Y=y} = \operatorname{prox}_{\eta H_{\pi^X}}(\delta_y)$. We use this fact in Section A.2 to provide a new proof of the contraction of the proximal sampler under strong log-concavity (Theorem 1). The proximal operator over the Wasserstein space is also known as the JKO scheme (Jordan et al., 1998), which we describe further in the next section.

4.3.3. PROXIMAL SAMPLER AS ENTROPY-REGULARIZED JKO SCHEME

The Wasserstein gradient flow models the steepest descent dynamics of a functional F over the space of probability distributions with respect to the 2-Wasserstein distance W_2. One strategy to approximate the Wasserstein gradient flow in discrete time is the JKO scheme (Jordan et al., 1998), which follows the iterations

$$\mu_{k+1} = \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ F(\mu) + \frac{1}{2\eta} W_2^2(\mu_k, \mu) \right\}, \tag{11}$$

where $\eta > 0$ is the step size. Note that this is a Wasserstein analogue of the proximal point method. A variant of the JKO scheme with an extra entropic regularization term was developed in Peyré (2015) to improve the computational efficiency. In this entropy-regularized Wasserstein gradient flow algorithm, one instead follows the update

$$\mu_{k+1} = \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ F(\mu) + \frac{1}{2\eta} W_{2,\epsilon}^2(\mu_k, \mu) \right\}, \tag{12}$$

where $W_{2,\epsilon}$ is the entropy-regularized 2-Wasserstein distance defined as

$$W_{2,\epsilon}^2(\mu,\nu) \coloneqq \min_{\gamma\in\mathcal{C}(\mu,\nu)} \left\{ \int \|x-y\|^2 \, d\gamma(x,y) + \epsilon H(\gamma) \right\}, \tag{13}$$

where $H(\gamma) = \int \gamma \log \gamma$ denotes the negative entropy.

We show that proximal sampler can be viewed as an entropy-regularized JKO scheme in the following result.

Theorem 10 *Let $\rho_k^X, \rho_k^Y, \rho_{k+1}^X$ be the distributions of x_k, y_k, x_{k+1}, respectively, in one iteration of the proximal sampler algorithm. Then, they follow the entropy-regularized JKO scheme*

$$\rho_k^Y = \operatorname*{arg\,min}_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \frac{1}{2\eta} W_{2,2\eta}^2(\rho_k^X, \mu), \tag{14}$$

and

$$\rho_{k+1}^X = \operatorname*{arg\,min}_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \left\{ \int f \, d\mu + \frac{1}{2\eta} W_{2,2\eta}^2(\rho_k^Y, \mu) \right\}. \tag{15}$$

Proof Section A.7. ■

4.3.4. PROXIMAL POINT METHOD AS THE LIMIT OF THE PROXIMAL SAMPLER

The interpretation of the proximal sampler algorithm above provides some insights on its connections to optimization. We can define a more general family of proximal sampler algorithm with a different level of entropy regularization. The forward step is

$$\rho_k^Y = \operatorname*{arg\,min}_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \frac{1}{2\eta} W_{2,2\eta\epsilon}^2(\rho_k^X, \mu), \tag{16}$$

and the backward step reads

$$\rho_{k+1}^X = \operatorname*{arg\,min}_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \left\{ \int f \, d\mu + \frac{1}{2\eta} W_{2,2\eta\epsilon}^2(\rho_k^Y, \mu) \right\}. \tag{17}$$

Theorem 11 *As $\epsilon \searrow 0$, (16)–(17) reduces to the proximal point algorithm in optimization.*

Proof Section A.8. ■

Indeed, when $\epsilon = 0$, with $\rho_k^X = \delta_{x_k}$, we have

$$\rho_k^Y = \rho_k^X = \delta_{x_k}$$

and furthermore, $\rho_{k+1}^X = \delta_{x_{k+1}}$ with

$$x_{k+1} = \operatorname*{arg\,min}_{x\in\mathbb{R}^d} \left\{ f(x) + \frac{1}{2\eta} \|x - x_k\|^2 \right\}.$$

This is exactly the proximal point method. In fact, even if ρ_k^X is not a Dirac distribution, (16)–(17) with $\epsilon = 0$ can be viewed as a parallel implementation of the proximal point method with many different initial points. See Section A.8 for more discussion.

4.4. Example: Gaussian case

Suppose that the target distribution is a Gaussian $\mathcal{N}(0, \Sigma)$, i.e., $f(x) = \frac{1}{2}\langle x, \Sigma^{-1}x\rangle$. In this case we can compute the iterations of the proximal sampler explicitly.

If we initialize the proximal sampler at

$$\rho_0^X = \mathcal{N}(m_0, \Sigma_0)\,,$$

then some calculations show that

$$\begin{aligned}\rho_k^Y &= \mathcal{N}(m_k, \Sigma_k + \eta I)\,,\\ \rho_{k+1}^X &= \mathcal{N}(m_{k+1}, \Sigma_{k+1})\,,\end{aligned}$$

where[3]

$$\begin{aligned}m_{k+1} &\coloneqq \Sigma\,(\Sigma + \eta I)^{-1}\,m_k\,,\\ \Sigma_{k+1} &\coloneqq \Sigma\,(\Sigma + \eta I)^{-1}\,(\Sigma_k + \eta I)\,(\Sigma + \eta I)^{-1}\,\Sigma + \eta\,\Sigma\,(\Sigma + \eta I)^{-1}\,.\end{aligned}$$

Specializing to the case where $\Sigma = I$, $\eta = 1$, and we initialize at $\mathcal{N}(0, \sigma_0^2 I)$, we obtain

$$|\sigma_k^2 - 1| = \frac{|\sigma_0^2 - 1|}{4^k}\,. \tag{18}$$

In particular, this shows that the contraction factor $\frac{1}{(1+\alpha\eta)^2}$ in Theorem 3 is sharp.

5. Conclusion and open directions

In this paper, we have studied in detail the proximal sampler of Lee et al. (2021a). In particular, we have given new convergence proofs under weaker assumptions than what were previously considered, allowing for a much wider class of distributions beyond log-concavity. In some cases, our proofs are inspired by convex optimization; in others, they show a remarkable parallel with the continuous-time theory of the Langevin diffusion under isoperimetry. Additionally, we have drawn more precise links between the proximal sampler and the proximal point method in optimization.

We conclude by listing a few directions for future study.

1. Is there an extension of the theory we have developed to the problem of sampling from composite potentials $\pi^X \propto \exp(-f - g)$?
2. Is there an accelerated version of the proximal sampler?

Acknowledgments

We would like to thank Ruoqi Shen and Kevin Tian for helpful conversations, and anonymous reviewers for useful comments and references. YC was supported in part by grants NSF CAREER ECCS-1942523, NSF CCF-2008513, and a Berkeley–Simons Research Fellowship. SC was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. AS was supported by a Berkeley–Simons Research Fellowship. This work was done while the authors were visiting the Simons Institute for the Theory of Computing.

3. We can also notice that $m_{k+1} = \mathrm{prox}_{\eta f}(m_k)$, i.e., the means of the distributions follow the proximal point algorithm for f. Moreover, $m_k \to 0 = \arg\min f$ which is the mean of the target distribution.

References

Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-Langevin algorithm. In M. Ranzato, A. Beygelzimer, K. Nguyen, P. S. Liang, J. W. Vaughan, and Y. Dauphin, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 28405–28418. Curran Associates, Inc., 2021.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

Luigi Ambrosio, Daniele Semola, and Elia Brué. Lectures on optimal transport, 2021.

Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1):5–16, 2009.

François L. Baccelli, Guy Cohen, Geert J. Olsder, and Jean-Pierre Quadrat. *Synchronization and linearity*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester, 1992. An algebra for discrete event systems.

Espen Bernton. Langevin Monte Carlo and JKO splitting. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1777–1798, Stockholm, Sweden, 06–09 Jul 2018. PMLR.

Adrien Blanchet and Jérôme Bolte. A family of functional inequalities: Łojasiewicz inequalities and displacement convex functions. *Journal of Functional Analysis*, 275(7):1650–1673, 2018.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

Yuansi Chen and Ronen Eldan. Localization schemes: a framework for proving mixing bounds for Markov chains. *arXiv e-prints*, art. arXiv:2203.04163, 2022.

Yuansi Chen, Raaz Dwivedi, Martin J. Wainwright, and Bin Yu. Fast mixing of Metropolized Hamiltonian Monte Carlo: benefits of multi-step gradients. *J. Mach. Learn. Res.*, 21:Paper No. 92, 71, 2020.

Sinho Chewi, Murat A. Erdogdu, Mufan B. Li, Ruoqi Shen, and Matthew Zhang. Analysis of Langevin Monte Carlo from Poincaré to log-Sobolev. *arXiv e-prints*, art. arXiv:2112.12662, 2021a.

Sinho Chewi, Patrik Gerber, Chen Lu, Thibaut Le Gouic, and Philippe Rigollet. The query complexity of sampling from strongly log-concave distributions in one dimension. *arXiv e-prints*, art. arXiv:2105.14163, 2021b.

Zhiyan Ding and Qin Li. Langevin Monte Carlo: random coordinate descent and variance reduction. *J. Mach. Learn. Res.*, 22:Paper No. 205, 51, 2021.

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *J. Mach. Learn. Res.*, 20:Paper No. 73, 46, 2019.

Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. Log-concave sampling: Metropolis–Hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183):1–42, 2019.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2): 123–214, 2011.

Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. Private convex optimization via exponential mechanism. *arXiv e-prints*, art. arXiv:2203.00263, 2022.

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

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak–Lojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

Bo'az Klartag and Joseph Lehec. Bourgain's slicing problem and KLS isoperimetry up to polylog. *arXiv e-prints*, art. arXiv:2203.15551, 2022.

Bo'az Klartag and Eli Putterman. Spectral monotonicity under Gaussian convolution. *arXiv e-prints*, art. arXiv:2107.09496, 2021.

Rafał Latała and Krzysztof Oleszkiewicz. Between Sobolev and Poincaré. In *Geometric aspects of functional analysis*, volume 1745 of *Lecture Notes in Math.*, pages 147–168. Springer, Berlin, 2000.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted Gaussian oracle. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2993–3050. PMLR, 8 2021a.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted Gaussian oracle. *arXiv e-prints*, art. arXiv:2010.03106, 2021b.

Jiaming Liang and Yongxin Chen. A proximal algorithm for sampling from non-smooth potentials. *arXiv preprint arXiv:2110.04597*, 2021.

Jiaming Liang and Yongxin Chen. A proximal algorithm for sampling. *arXiv e-prints*, art. arXiv:2202.13975, 2022a.

Jiaming Liang and Yongxin Chen. A proximal algorithm for sampling from non-convex potentials. *arXiv e-prints*, art. arXiv:2205.10188, 2022b.

Yi-An Ma, Niladri S. Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L. Bartlett, and Michael I. Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3): 1942 – 1992, 2021.

Yosra Marnissi, Emilie Chouzenoux, Jean-Christophe Pesquei, and Amel Benazza-Benyahia. An auxiliary variable method for Langevin based MCMC algorithms. In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pages 1–5. IEEE, 2016.

Bernard Martinet. Brève communication. régularisation d'inéquations variationnelles par approximations successives. *Revue française d'informatique et de recherche opérationnelle. Série rouge*, 4 (R3):154–158, 1970.

Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.*, 173(2):361–400, 2000.

Marcelo Pereyra. Proximal Markov chain Monte Carlo algorithms. *Stat. Comput.*, 26(4):745–760, 2016.

Gabriel Peyré. Entropic approximation of Wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

Christian P. Robert and George Casella. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004.

R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

Adil Salim and Peter Richtarik. Primal dual interpretation of the proximal stochastic gradient Langevin algorithm. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3786–3796. Curran Associates, Inc., 2020.

Simo Särkkä and Arno Solin. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019.

Adrien Saumard and Jon A. Wellner. Log-concavity and strong log-concavity: a review. *Stat. Surv.*, 8:45–114, 2014.

Michalis K. Titsias and Omiros Papaspiliopoulos. Auxiliary gradient-based sampling algorithms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):749–767, 2018.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

Maxime Vono, Daniel Paulin, and Arnaud Doucet. Efficient MCMC sampling with dimension-free convergence rate using ADMM-type splitting. *Journal of Machine Learning Research*, 23(25), 2022.

Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 2093–3027. PMLR, 06–09 Jul 2018.

Andre Wibisono. Proximal Langevin algorithm: Rapid convergence under isoperimetry. *arXiv e-prints*, art. arXiv:1911.01469, 2019.

Kelvin S. Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. Wasserstein control of mirror Langevin Monte Carlo. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3814–3841. PMLR, 09–12 Jul 2020.

Appendix A. Proofs for the proximal sampler

A.1. Techniques

At a high level, our proofs proceed by considering the change in KL divergence or Rényi divergence when we apply the following two operations to the law ρ_k^X of the iterate and the target π^X: (1) we simultaneously evolve the two measures along the heat flow for time η, and then (2) we apply the RGO to the resulting measures.

For the first step, we formulate a remarkably general lemma in Section A.1.1 which shows that the computation of the time derivative of *any* ϕ-divergence along the simultaneous heat flow is similar (in a precise sense) to the analogous computation when studying the continuous-time Langevin diffusion. It is this property that allows us to apply functional inequalities which are usually used for the Langevin diffusion, such as the Poincaré and log-Sobolev inequalities, in order to study the convergence of the proximal sampler.

In the second step, we are applying the same operation (of sampling from the RGO) to each measure, so the data-processing inequality implies that the KL divergence or Rényi divergence can only decrease. Combined with the previous step, it is sufficient to prove a convergence guarantee for the proximal sampler; however, the rate turns out to be suboptimal. In order to recover the optimal rate, we introduce an argument based on the *Doob h-transform* (described in Section A.1.2) to obtain contraction in the second step as well, using the backward version of our general lemma (see Section A.1.3). We summarize our technique in Section A.1.4.

A.1.1. LEMMA ON THE SIMULTANEOUS HEAT FLOW

Let Φ_π be a ϕ-divergence for some convex function ϕ, i.e.

$$\Phi_\pi(\rho) \coloneqq \mathbb{E}_\pi\left[\phi\left(\frac{\rho}{\pi}\right)\right].$$

We assume that ϕ is regular enough to justify the interchange of differentiation and integration and to perform integration by parts; this is satisfied for all of our applications.

We will use the following result in each forward step of the proximal sampler. This is a generalization of Vempala and Wibisono (2019, Lemma 16).

Lemma 12 *Let $(\mu_t^X)_{t\geq 0}$ be the law of the continuous-time Langevin diffusion with target distribution π^X, and define the* dissipation functional *D_{π^X} via the time derivative of Φ_{π^X} along the diffusion:*

$$D_{\pi^X}(\mu_t^X) \coloneqq -\partial_t \Phi_{\pi^X}(\mu_t^X) = \mathbb{E}_{\mu_t^X}\left\langle \nabla\left(\phi' \circ \frac{\mu_t^X}{\pi^X}\right), \nabla \log \frac{\mu_t^X}{\pi^X}\right\rangle.$$

If $(\rho^X Q_t)_{t\geq 0}$ and $(\pi^X Q_t)_{t\geq 0}$ evolve according to the simultaneous heat flow,

$$\partial_t \rho^X Q_t = \frac{1}{2}\Delta(\rho^X Q_t), \qquad \partial_t \pi^X Q_t = \frac{1}{2}\Delta(\pi^X Q_t),$$

then

$$\partial_t \Phi_{\pi^X Q_t}(\rho^X Q_t) = -\frac{1}{2}\, D_{\pi^X Q_t}(\rho^X Q_t)\,.$$

Proof On one hand, we know that $(\mu_t^X)_{t\geq 0}$ satisfies the Fokker-Planck equation

$$\partial_t \mu_t^X = \operatorname{div}\big(\mu_t^X \nabla \log \frac{\mu_t^X}{\pi^X}\big)$$

so that

$$\begin{aligned}\partial_t \Phi_{\pi^X}(\mu_t^X) &= \int \phi'\big(\frac{\mu_t^X}{\pi^X}\big)\, \partial_t \mu_t^X = \int \phi'\big(\frac{\mu_t^X}{\pi^X}\big)\, \operatorname{div}\big(\mu_t^X \nabla \log \frac{\mu_t^X}{\pi^X}\big) \\ &= -\int \Big\langle \nabla\big[\phi'\big(\frac{\mu_t^X}{\pi^X}\big)\big], \nabla \log \frac{\mu_t^X}{\pi^X} \Big\rangle\, \mu_t^X\,.\end{aligned}$$

On the other hand, writing $\rho_t^X \coloneqq \rho^X Q_t$ and $\pi_t^X \coloneqq \pi^X Q_t$ for brevity, along the simultaneous heat flow we compute

$$\begin{aligned}2\,\partial_t \Phi_{\pi_t^X}(\rho_t^X) &= 2\int \phi'\big(\frac{\rho_t^X}{\pi_t^X}\big)\, \Big(\partial_t \rho_t^X - \frac{\rho_t^X}{\pi_t^X}\, \partial_t \pi_t^X\Big) + 2\int \phi\big(\frac{\rho_t^X}{\pi_t^X}\big)\, \partial_t \pi_t^X \\ &= \int \phi'\big(\frac{\rho_t^X}{\pi_t^X}\big)\, \Big(\operatorname{div}(\rho_t^X \nabla \log \rho_t^X) - \frac{\rho_t^X}{\pi_t^X} \operatorname{div}(\pi_t^X \nabla \log \pi_t^X)\Big) \\ &\qquad + \int \phi\big(\frac{\rho_t^X}{\pi_t^X}\big)\, \operatorname{div}(\pi_t^X \nabla \log \pi_t^X) \\ &= -\int \Big\langle \nabla\big[\phi'\big(\frac{\rho_t^X}{\pi_t^X}\big)\big], \nabla \log \rho_t^X \Big\rangle\, \rho_t^X + \int \Big\langle \nabla\big[\phi'\big(\frac{\rho_t^X}{\pi_t^X}\big)\, \frac{\rho_t^X}{\pi_t^X}\big], \nabla \log \pi_t^X \Big\rangle\, \pi_t^X \\ &\qquad - \int \Big\langle \nabla\big[\phi\big(\frac{\rho_t^X}{\pi_t^X}\big)\big], \nabla \log \pi_t^X \Big\rangle\, \pi_t^X \\ &= -\int \Big\langle \nabla\big[\phi'\big(\frac{\rho_t^X}{\pi_t^X}\big)\big], \nabla \log \frac{\rho_t^X}{\pi_t^X} \Big\rangle\, \rho_t^X + \int \Big\langle \nabla \frac{\rho_t^X}{\pi_t^X}, \nabla \log \pi_t^X \Big\rangle\, \phi'\big(\frac{\rho_t^X}{\pi_t^X}\big)\, \pi_t^X \\ &\qquad - \int \Big\langle \nabla \frac{\rho_t^X}{\pi_t^X}, \nabla \log \pi_t^X \Big\rangle\, \phi'\big(\frac{\rho_t^X}{\pi_t^X}\big)\, \pi_t^X \\ &= -D_{\pi_t^X}(\rho_t^X)\,.\end{aligned}$$

■

Remark 13 *A similar statement holds if we replace the ϕ-divergence Φ_π with any function $\psi \circ \Phi_\pi$ of the ϕ-divergence. This allows us to cover the Rényi divergence introduced in Section 2.*

A.1.2. DOOB'S h-TRANSFORM

Doob's h-transform is a useful method to analyze the properties of a diffusion process conditioned on its value at some terminal time point. Consider a general diffusion process modeled by the stochastic differential equation (SDE)

$$dZ_t = b(t, Z_t)\, dt + \sigma(t, Z_t)\, dW_t\,, \qquad Z_0 \sim \mu_0\,, \tag{19}$$

where $(W_t)_{t\geq 0}$ denotes a standard Wiener process. Assume that $b(t,z)$ and $\sigma(t,z)$ are piecewise continuous with respect to t and Lipschitz continuous with respect to z so that the above SDE (19) has a unique solution. The Doob h-transform characterizes the process conditional on its terminal value Z_T, summarized in the following lemma (Särkkä and Solin, 2019).

Lemma 14 *Let $(\hat{Z}_t)_{0\leq t\leq T}$ be the process (19) conditioned to satisfy $Z_T = z$. Then, the process satisfies the following SDE backwards in time:*

$$d\hat{Z}_t = [b(t,\hat{Z}_t) - \sigma(t,\hat{Z}_t)\,\sigma(t,\hat{Z}_t)^{\mathsf{T}}\,\nabla \log \mu_t(\hat{Z}_t)]\,dt + \sigma(t,\hat{Z}_t)\,dW_t\,,$$

where μ_t is the marginal distribution of Z_t in (19) and the SDE is started with $\hat{Z}_T = z$.

Equivalently, if we define the SDE

$$d\hat{Z}_t^- = [-b(T-t,\hat{Z}_t^-) + \sigma(T-t,\hat{Z}_t^-)\,\sigma(T-t,\hat{Z}_t^-)^{\mathsf{T}}\,\nabla \log \mu_{T-t}(\hat{Z}_t^-)]\,dt + \sigma(T-t,\hat{Z}_t^-)\,dW_t\,, \tag{20}$$

started at $\hat{Z}_0^- = z$, then at time T the law of $\hat{Z}_T^-$ is the conditional distribution of Z_0 given $Z_T = z$.

A.1.3. LEMMA ON THE SIMULTANEOUS BACKWARD HEAT FLOW

We present the following backward version of Lemma 12, which we use in each backward step of the proximal sampler. We assume the same set up as in Lemma 12: Let $\Phi_\pi(\rho) = \mathbb{E}_\pi[\phi(\frac{\rho}{\pi})]$ be a ϕ-divergence for some convex function ϕ, i.e.

$$\Phi_\pi(\rho) \coloneqq \mathbb{E}_\pi\big[\phi\big(\frac{\rho}{\pi}\big)\big]$$

and let

$$D_\pi(\rho) = \mathbb{E}_\rho\Big\langle \nabla\big(\phi' \circ \frac{\rho}{\pi}\big), \nabla \log \frac{\rho}{\pi}\Big\rangle$$

so that D_π is the dissipation of Φ_π along the Langevin dynamics with target π.

Lemma 15 *Let π^X be a probability distribution and let $\pi(x,y) = \pi^X(x)\,\mathcal{N}(y;x,\eta I)$ be a joint density for (X,Y) with Y obtained from X by running the heat flow for time η. Let $\pi^{X|Y}$ be the conditional distribution of X given Y under π, and let π^Y denote the marginal distribution of Y. Then, for each $t \in [0,\eta]$, there exists a channel Q_t^- that maps probability measures to probability measures, with the following properties: (1) Q_0^- is the identity channel; (2) Q_η^- maps a probability measure ρ^Y to the the measure $\rho^Y Q_\eta^-(x) = \int \pi^{X|Y}(x \mid y)\,\rho^Y(dy)$; (3) for every t, $\pi^Y Q_t^- = \pi * \mathcal{N}(0,(\eta-t)I)$; and (4) for every ρ^Y,*

$$\partial_t \Phi_{\pi^Y Q_t^-}(\rho^Y Q_t^-) = -\frac{1}{2}\,D_{\pi^Y Q_t^-}(\rho^Y Q_t^-)\,.$$

The channel is obtained from the Doob h-transform. To give intuition for the construction, consider the process $\mathrm{d}Z_t = \mathrm{d}B_t$ started at $Z_0 \sim \pi^X$, i.e., Brownian motion initialized from π^X. Then, the joint target distribution π of the proximal sampler can be expressed as $\pi = \mathrm{law}(Z_0, Z_\eta)$, and consequently $\pi^{X|Y=y} = \mathrm{law}(Z_0 \mid Z_\eta = y)$. If we define the time reversal $Z_t^- = Z_{\eta-t}$, then we can also express this as $\pi^{X|Y=y} = \mathrm{law}(Z_\eta^- \mid Z_0^- = y)$; moreover, the reversed process $(Z_t^-)_{t\in[0,\eta]}$

satisfies the SDE given in Lemma 14. Hence, we can take $\mu Q_t^- \coloneqq \mathrm{law}(Z_\eta^- \mid Z_0^- \sim \mu)$ and use calculus in order to prove the result.

Proof Let $\pi_t \coloneqq \pi^X * \mathcal{N}(0, tI)$. We define Q_t^- as follows: given ρ^Y, we set $\rho^Y Q_t^-$ to be the law at time t of the SDE

$$d\hat{Z}_t^- = \nabla \log \pi_{\eta-t}(\hat{Z}_t^-)\, dt + dW_t\,, \tag{21}$$

started at $\hat{Z}_0^- \sim \rho^Y$. According to Lemma 14 applied to the Brownian motion process (started at π^X), the channels $(Q_t^-)_{0\le t\le\eta}$ satisfy properties (1), (2), and (3). It remains to verify (4). In the proof, we write $\pi_t^- \coloneqq \pi^Y Q_t^-$ and $\rho_t^- \coloneqq \rho^Y Q_t^-$ for brevity. Note that $\pi_{\eta-t} = \pi_t^-$ by construction, and we have the Fokker–Planck equations:

$$\begin{aligned}
\partial_t \pi_t^- &= -\operatorname{div}(\pi_t^- \nabla \log \pi_t^-) + \frac{1}{2}\Delta \pi_t^- = -\frac{1}{2}\Delta \pi_t^-\,,\\
\partial_t \rho_t^- &= -\operatorname{div}(\rho_t^- \nabla \log \pi_t^-) + \frac{1}{2}\Delta \rho_t^- = \operatorname{div}\big(\rho_t^- \nabla \log \frac{\rho_t^-}{\pi_t^-}\big) - \frac{1}{2}\Delta \rho_t^-\,.
\end{aligned}$$

Hence,

$$\begin{aligned}
2\,\partial_t \Phi_{\pi_t^-}(\rho_t^-) &= 2\int \phi'\big(\tfrac{\rho_t^-}{\pi_t^-}\big)\Big(\partial_t \rho_t^- - \tfrac{\rho_t^-}{\pi_t^-}\,\partial_t \pi_t^-\Big) + 2\int \phi\big(\tfrac{\rho_t^-}{\pi_t^-}\big)\,\partial_t \pi_t^-\\
&= \int \phi'\big(\tfrac{\rho_t^-}{\pi_t^-}\big)\Big(2\operatorname{div}\big(\rho_t^- \nabla \log \tfrac{\rho_t^-}{\pi_t^-}\big) - \Delta \rho_t^- + \tfrac{\rho_t^-}{\pi_t^-}\Delta \pi_t^-\Big) - \int \phi\big(\tfrac{\rho_t^-}{\pi_t^-}\big)\Delta \pi_t^-\\
&= 2\int \phi'\big(\tfrac{\rho_t^-}{\pi_t^-}\big)\operatorname{div}\big(\rho_t^- \nabla \log \tfrac{\rho_t^-}{\pi_t^-}\big)\\
&\qquad \underbrace{- \int \phi'\big(\tfrac{\rho_t^-}{\pi_t^-}\big)\Big(\Delta \rho_t^- - \tfrac{\rho_t^-}{\pi_t^-}\Delta \pi_t^-\Big) + \int \phi\big(\tfrac{\rho_t^-}{\pi_t^-}\big)\Delta \pi_t^-}_{= -D_{\pi_t^-}(\rho_t^-) \text{ by Lemma 12}}\\
&= -2\int \Big\langle \nabla\big[\phi'\big(\tfrac{\rho_t^-}{\pi_t^-}\big)\big], \nabla \log \tfrac{\rho_t^-}{\pi_t^-}\Big\rangle \rho_t^- + D_{\pi_t^-}(\rho_t^-)\\
&= -2D_{\pi_t^-}(\rho_t^-) + D_{\pi_t^-}(\rho_t^-) = -D_{\pi_t^-}(\rho_t^-)\,.
\end{aligned}$$

■

A.1.4. GENERAL STRATEGY OF THE PROOFS

Suppose that we want to understand the change in the ϕ-divergence $\Phi_{\pi^X}(\rho_1^X)$ after one iteration of the proximal sampler, compared to the ϕ divergence $\Phi_{\pi^X}(\rho_0^X)$ at initialization. We split the analysis into two steps.

1. **Forward step:** In the first step, we draw $y_0 \mid x_0 \sim \pi^{Y|X=x_0} = \mathcal{N}(x_0, \eta I)$.

   This creates a joint distribution $\rho_0(x,y) = \rho_0^X(x)\,\mathcal{N}(y; x, \eta I)$ with the correct conditionals: $\rho_0^{Y|X} = \pi^{Y|X}$. Therefore, the ϕ-divergence of the joint distribution is equal to the initial ϕ-divergence of the X-marginal: $\Phi_\pi(\rho_0) = \Phi_{\pi^X}(\rho_0^X)$.

Consider the Y-marginal $y_0 \sim \rho_0^Y$. Observe that $\rho_0^Y = \rho_0^X * \mathcal{N}(0, \eta I)$ is the output $\rho_0^Y = \tilde{\rho}_\eta$ of the heat flow $\partial_t \tilde{\rho}_t = \frac{1}{2}\Delta \tilde{\rho}_t$ at time $t = \eta$ starting from $\tilde{\rho}_0 = \rho_0^X$. We denote this by $\rho_0^Y = \rho_0^X Q_\eta$, where $(Q_t)_{t \geq 0}$ denotes the heat semigroup.

Similarly, we can write the Y-marginal of the target as $\pi^Y = \pi^X * \mathcal{N}(0, \eta I) = \pi^X Q_\eta$.

In particular, $(\rho_0^X Q_t)_{t \geq 0}$ and $(\pi^X Q_t)_{t \geq 0}$ evolve following the simultaneous heat flow.

By Lemma 12, along the simultaneous heat flow,

$$\partial_t \Phi_{\pi^X Q_t}(\rho_0^X Q_t) = -\frac{1}{2} D_{\pi^X Q_t}(\rho_0^X Q_t)$$

where $D_\cdot(\cdot)$ denotes the dissipation functional for the ϕ-divergence along the Langevin dynamics. Hence, a lower bound on $D_{\pi^X Q_t}(\rho_0^X Q_t)$ leads to an upper bound on

$$\Phi_{\pi^Y}(\rho_0^Y) - \Phi_{\pi^X}(\rho_0^X) = \Phi_{\pi^X Q_\eta}(\rho_0^X Q_\eta) - \Phi_{\pi^X}(\rho_0^X)\,.$$

2. **Backward step:** In the second step, we draw $x_1 \mid y_0 \sim \pi^{X|Y=y_0}$.

   This time, we consider the backward heat flow and apply Lemma 15, which yields the Doob channels $(Q_t^-)_{0 \leq t \leq \eta}$ with $\rho_1^X = \rho_0^Y Q_\eta^-$ and $\pi^X = \pi^Y Q_\eta^-$. Lemma 15 implies that

$$\partial_t \Phi_{\pi^Y Q_t^-}(\rho_0^Y Q_t^-) = -\frac{1}{2} D_{\pi^Y Q_t^-}(\rho_0^Y Q_t^-)\,.$$

   Observe that this is almost symmetric with the forward step! In particular, a lower bound on $D_{\pi^Y Q_t^-}(\rho_0^Y Q_t^-)$ leads to an upper bound on

$$\Phi_{\pi^X}(\rho_1^X) - \Phi_{\pi^Y}(\rho_0^Y) = \Phi_{\pi^Y Q_\eta^-}(\rho_0^Y Q_\eta^-) - \Phi_{\pi^Y}(\rho_0^Y)\,.$$

Combining the two steps allows to understand each iteration of the proximal sampler.

A.2. Convergence under strong log-concavity

Suppose that A is a set-valued mapping on $\mathbb{R}^d$ which is strongly monotone, in the sense that

$$\langle A(x) - A(y), x - y \rangle \geq \alpha \|x - y\|^2 \qquad \text{for all } x, y \in \mathbb{R}^d\,.$$

Suppose that $x' \in x - \eta A(x')$ and $y' \in y - \eta A(y')$. Then, by expanding out the square, one can easily show that $\|x' - y'\|^2 \leq \frac{1}{(1+\alpha\eta)^2} \|x - y\|^2$. In particular, by applying this to the subdifferential $A = \partial f$, where f is α-strongly convex, one immediately obtains the fact that the proximal point algorithm is a $\frac{1}{1+\alpha\eta}$-contraction. In this section, we translate this proof to the sampling setting.

Recall from (10) that $\pi^{X|Y=y} = \mathrm{prox}_{\eta F}(\delta_y)$, where $F = H_{\pi^X}$ is α-geodesically strongly strongly convex (Ambrosio et al., 2008, Equation 10.1.8). Then, from the first-order optimality conditions on Wasserstein space (see Ambrosio et al., 2008, Lemma 10.1.2), we have

$$0 \in \partial F(\pi^{X|Y=y}) + \frac{1}{\eta}(\mathrm{id} - y)\,, \qquad \pi^{X|Y=y}\text{-a.s.}, \tag{22}$$

where ∂F denotes the Wasserstein subdifferential of F.

Proof [Proof of Theorem 1] First, let $y, \bar{y} \in \mathbb{R}^d$. Then, from (22):

$$\mathrm{id} \in y - \eta\, \partial F(\pi^{X|Y=y})\,, \qquad \pi^{X|Y=y}\text{-a.s.} \tag{23}$$

$$\mathrm{id} \in \bar{y} - \eta\, \partial F(\pi^{X|Y=\bar{y}})\,, \qquad \pi^{X|Y=\bar{y}}\text{-a.s.} \tag{24}$$

Let T be the optimal transport map from $\pi^{X|Y=y}$ to $\pi^{X|Y=\bar{y}}$. We can rewrite (24) as

$$T \in \bar{y} - \eta\, \partial F(\pi^{X|Y=\bar{y}}) \circ T\,, \qquad \pi^{X|Y=y}\text{-a.s.} \tag{25}$$

We now abuse notation and write $\partial F(\pi^{X|Y=y})$ for an element of the subdifferential. Then, using (23) and (25), $\pi^{X|Y=y}$-a.s.,

$$\begin{aligned}\|T - \mathrm{id}\|^2 = \|\bar{y} - y\|^2 &- 2\eta \left\langle \partial F(\pi^{X|Y=\bar{y}}) \circ T - \partial F(\pi^{X|Y=y}), T - \mathrm{id} \right\rangle \\ &- \eta^2 \left\| \partial F(\pi^{X|Y=\bar{y}}) \circ T - \partial F(\pi^{X|Y=y}) \right\|^2 .\end{aligned}$$

Integrating with respect to $\pi^{X|Y=y}$, and using the geodesic strong convexity of F (Ambrosio et al., 2008, Equation 10.1.8),

$$W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}}) \le \|y - \bar{y}\|^2 - 2\alpha\eta\, W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}}) - \alpha^2\eta^2\, W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}})\,.$$

Therefore,

$$W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}}) \le \frac{1}{(1+\alpha\eta)^2}\, \|y - \bar{y}\|^2.$$

The rest of the argument is concluded as in Lee et al. (2021b, Lemma 2). We provide the details here for completeness. First, along the proximal sampler, we have $W_2(\rho_0^Y, \bar{\rho}_0^Y) \le W_2(\rho_0^X, \bar{\rho}_0^X)$ because the heat flow is a Wasserstein contraction (see Section A.1.4 for the notation). Next, let γ denote an optimal coupling of ρ_0^Y and $\bar{\rho}_0^Y$, and for all $y, y \in \mathbb{R}^d$ let $\gamma_{y,\bar{y}}$ denote an optimal coupling of $\pi^{X|Y=y}$ and $\pi^{X|Y=\bar{y}}$. We check that the measure $\hat{\gamma}(dx, d\bar{x}) \coloneqq \gamma(dy, d\bar{y})\, \gamma_{y,\bar{y}}(dx, d\bar{x})$ is a valid coupling of ρ_1^X and $\bar{\rho}_1^X$. To check that, for instance, the first marginal of $\hat{\gamma}$ is ρ_1^X, we take a bounded measurable function $\psi : \mathbb{R}^d \to \mathbb{R}$ and calculate

$$\begin{aligned}\int \psi(x)\, \hat{\gamma}(dx, d\bar{x}) &= \iint \psi(x)\, \gamma(dy, d\bar{y})\, \gamma_{y,\bar{y}}(dx, d\bar{x}) = \iint \psi(x)\, \gamma(dy, d\bar{y})\, \pi^{X|Y=y}(dx) \\ &= \iint \psi(x)\, \rho_0^Y(dy)\, \pi^{X|Y=y}(dx) = \int \psi(x)\, \rho_1^X(dx)\,,\end{aligned}$$

and similarly the second marginal of $\hat{\gamma}$ is $\bar{\rho}_1^X$. Therefore,

$$\begin{aligned}W_2^2(\rho_1^X, \bar{\rho}_1^X) &\le \int \|x - \bar{x}\|^2\, \hat{\gamma}(dx, d\bar{x}) = \iint \|x - \bar{x}\|^2\, \gamma(dy, d\bar{y})\, \gamma_{y,\bar{y}}(dx, d\bar{x}) \\ &= \int W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}})\, \gamma(dy, d\bar{y}) \\ &\le \frac{1}{(1+\alpha\eta)^2} \int \|y - \bar{y}\|^2\, \gamma(dy, d\bar{y}) = \frac{1}{(1+\alpha\eta)^2}\, W_2^2(\rho_0^Y, \bar{\rho}_0^Y)\,,\end{aligned}$$

which completes the proof. ■

A.3. Convergence under log-concavity

For a probability distribution ρ with smooth relative density $\frac{\rho}{\pi}$, the *Fisher information* of ρ with respect to π is

$$J_\pi(\rho) \coloneqq \int \rho \left\| \nabla \log \frac{\rho}{\pi} \right\|^2 = \mathbb{E}_\pi \left[\frac{\pi}{\rho} \left\| \nabla \frac{\rho}{\pi} \right\|^2 \right]. \tag{26}$$

Recall that Fisher information is the dissipation of KL divergence along the Langevin dynamics.

Proof [Proof of Theorem 2] We follow the strategy and notation of Section A.1.4.

1. **Forward step:** By log-concavity of $\pi^X Q_t$ (since log-concavity is preserved by convolution (Saumard and Wellner, 2014)), the convexity of $H_{\pi^X Q_t}$ along Wasserstein geodesics (Ambrosio et al., 2008, Theorem 9.4.11) yields the inequality

$$\begin{aligned} 0 &= H_{\pi^X Q_t}(\pi^X Q_t) \\ &\geq H_{\pi^X Q_t}(\rho_0^X Q_t) + \mathbb{E}_{(X_t, Y_t) \sim \mathsf{OPT}(\rho_0^X Q_t, \pi^X Q_t)} \left\langle \nabla \log \frac{\rho_0^X Q_t}{\pi^X Q_t}(X_t), Y_t - X_t \right\rangle \end{aligned}$$

where $\mathsf{OPT}(\cdot, \cdot)$ is used to denote the optimal transport plan. Hence,

$$\underbrace{\mathbb{E}_{\rho_0^X Q_t} \left[\left\| \nabla \log \frac{\rho_0^X Q_t}{\pi^X Q_t} \right\|^2 \right]}_{= J_{\pi^X Q_t}(\rho_0^X Q_t)} W_2^2(\rho_0^X Q_t, \pi^X Q_t) \geq H_{\pi^X Q_t}(\rho_0^X Q_t)^2 . \tag{27}$$

So, by Lemma 12 and (27),

$$\partial_t H_{\pi^X Q_t}(\rho_0^X Q_t) = -\frac{1}{2} J_{\pi^X Q_t}(\rho_0^X Q_t) \leq -\frac{1}{2} \frac{H_{\pi^X Q_t}(\rho_0^X Q_t)^2}{W_2^2(\rho_0^X Q_t, \pi^X Q_t)} .$$

Also, observe that $t \mapsto W_2^2(\rho_0^X Q_t, \pi^X Q_t)$ is decreasing because the heat flow is a W_2 contraction (which can be proven directly quite easily). Solving this differential inequality yields

$$\frac{1}{H_{\pi^Y}(\rho_0^Y)} = \frac{1}{H_{\pi^X Q_\eta}(\rho_0^X Q_\eta)} \geq \frac{1}{H_{\pi^X}(\rho_0^X)} + \frac{\eta}{2 W_2^2(\rho_0^X, \pi^X)} .$$

2. **Backward step:** By Lemma 15 and (27),

$$\partial_t H_{\pi^Y Q_t^-}(\rho_0^Y Q_t^-) = -\frac{1}{2} J_{\pi^Y Q_t^-}(\rho_0^Y Q_t^-) \leq -\frac{1}{2} \frac{H_{\pi_Y Q_t^-}(\rho_0^Y Q_t^-)^2}{W_2^2(\rho_0^Y Q_t^-, \pi^Y Q_t^-)} .$$

By (20), the channels $(Q_t^-)_{t \geq 0}$ can be modeled by the diffusion

$$dZ_t = \nabla \log \pi_{\eta - t}(Z_t)\, dt + dW_t.$$

Since $\log \pi_{\eta - t}$ is concave, with a standard coupling argument, one can show that $t \mapsto W_2(\rho_0^Y Q_t^-, \pi^Y Q_t^-)$ is decreasing. Hence,

$$W_2(\rho_0^Y Q_t^-, \pi^Y Q_t^-) \leq W_2(\rho_0^Y Q_0^-, \pi^Y Q_0^-) = W_2(\rho_0^Y, \pi^Y) \leq W_2(\rho_0^X, \pi^X) .$$

Therefore, we deduce that

$$\frac{1}{H_{\pi^X}(\rho_1^X)} = \frac{1}{H_{\pi^Y Q_\eta^-}(\rho_0^Y Q_\eta^-)} \geq \frac{1}{H_{\pi^Y}(\rho_0^Y)} + \frac{\eta}{2W_2^2(\rho_0^X, \pi^X)}\,.$$

Finally, we iterate this inequality and recall that $W_2^2(\rho_k^X, \pi^X) \leq W_2^2(\rho_0^X, \pi^X)$ for all $k \in \mathbb{N}$ (see Theorem 1 for $\alpha = 0$). It quickly yields

$$\frac{1}{H_{\pi^X}(\rho_k^X)} \geq \frac{1}{H_{\pi^X}(\rho_0^X)} + \frac{k\eta}{W_2^2(\rho_0^X, \pi^X)}$$

or

$$H_{\pi^X}(\rho_k^X) \leq \frac{H_{\pi^X}(\rho_0^X)}{1 + k\eta\, H_{\pi^X}(\rho_0^X)/W_2^2(\rho_0^X, \pi^X)} \leq \frac{W_2^2(\rho_0^X, \pi^X)}{k\eta}\,.$$

■

The above proof can be compared to the $O(1/t)$ convergence of the objective gap for the gradient flow $t \mapsto x_t$ of a convex function $f : \mathbb{R}^d \to \mathbb{R}$, which follows from differentiating the Lyapunov function $t \mapsto 2t\,\{f(x_t) - f(x^\star)\} + \|x_t - x^\star\|^2$, where $x^\star = \arg\min f$.

A.4. Convergence under LSI

We recall the following definitions. For a probability distribution ρ with smooth relative density $\frac{\rho}{\pi}$, the *Rényi information* of ρ with respect to π of order $q \geq 1$ is

$$J_{q,\pi}(\rho) \coloneqq q\,\frac{\mathbb{E}_\pi\big[\big(\frac{\pi}{\rho}\big)^{q-2}\,\big\|\nabla\frac{\rho}{\pi}\big\|^2\big]}{\mathbb{E}_\pi\big[\big(\frac{\pi}{\rho}\big)^q\big]}\,.$$

Note that $J_{1,\pi}(\rho) = J_\pi(\rho)$, where J_π is the Fisher information (26). Recall that by definition, π satisfies α-LSI if for all ρ, $J_\pi(\rho) \geq 2\alpha H_\pi(\rho)$. One can show this also implies for all $q \geq 1$:

$$J_{q,\pi}(\rho) \geq \frac{2\alpha}{q}\,R_{q,\pi}(\rho)\,, \tag{28}$$

see for example Vempala and Wibisono (2019, Lemma 5). Just as Fisher information is the dissipation of KL divergence along the Langevin dynamics, Rényi information is the dissipation of Rényi divergence along the Langevin dynamics.

Proof [Proof of Theorem 3] We will prove the following one-step improvement lemma for Rényi divergence of order $q \geq 1$: For any initial distribution ρ_0^X, after one iteration of the proximal sampler with step size $\eta > 0$, the resulting distribution ρ_1^X satisfies

$$R_{q,\pi^X}(\rho_1^X) \leq \frac{R_{q,\pi^X}(\rho_0^X)}{(1+\alpha\eta)^{2/q}}\,. \tag{29}$$

Iterating this lemma for k iterations yields the desired convergence rate in the theorem. The result for KL divergence is the special case $q = 1$.

We follow the strategy and notation of Section A.1.4.

1. **Forward step:** By Lemma 12, along the simultaneous heat flow,

$$\partial_t R_{q,\pi^X Q_t}(\rho_0^X Q_t) = -\frac{1}{2} J_{q,\pi^X Q_t}(\rho_0^X Q_t) \leq -\frac{\alpha_t}{q} R_{q,\pi^X Q_t}(\rho_0^X Q_t)$$

where by (28), the last inequality holds if $\pi^X Q_t$ is α_t-LSI. Since π^X satisfies α-LSI by assumption, recall that $\pi^X Q_t = \pi^X * \mathcal{N}(0, tI)$ satisfies α_t-LSI with $\alpha_t = (\frac{1}{\alpha} + t)^{-1} = \frac{\alpha}{1+\alpha t}$. Integrating, we get

$$R_{q,\pi^X Q_t}(\rho_0^X Q_t) \leq \exp(-A_t)\, R_{q,\pi^X}(\rho_0^X)$$

where $A_t = \frac{1}{q}\int_0^t \alpha_s\, ds = \frac{1}{q}\int_0^t \frac{\alpha}{1+\alpha s}\, ds = \frac{1}{q}\log(1+\alpha t)$. Therefore, after the forward step,

$$R_{q,\pi^Y}(\rho_0^Y) = R_{q,\pi^X Q_\eta}(\rho_0^X Q_\eta) \leq \frac{R_{q,\pi^X}(\rho_0^X)}{(1+\alpha\eta)^{1/q}}\,.$$

2. **Backward step:** By Lemma 15, along the simultaneous backwards heat flow,

$$\partial_t R_{q,\pi^Y Q_t^-}(\rho_0^Y Q_t^-) = -\frac{1}{2} J_{q,\pi^Y Q_t^-}(\rho_0^Y Q_t^-) \leq -\frac{\alpha_{\eta-t}}{q} R_{q,\pi^Y Q_t^-}(\rho_0^Y Q_t^-)$$

where the last inequality holds since $\pi^Y Q_t^- = \pi * \mathcal{N}(0, (\eta - t)I)$ is $\alpha_{\eta-t}$-LSI. Therefore, just as in the forward step, integration yields

$$R_{q,\pi^X}(\rho_1^X) = R_{q,\pi^Y Q_\eta^-}(\rho_0^Y Q_\eta^-) \leq \frac{R_{q,\pi^Y}(\rho_0^Y)}{(1+\alpha\eta)^{1/q}}\,.$$

Combining the two steps above yields the desired contraction rate in (29). ■

A.5. Convergence under PI

The dissipation of the chi-squared divergence along the Langevin dynamics is

$$J_{\chi^2,\pi}(\rho) \coloneqq 2\,\mathbb{E}_\pi\left[\left\|\nabla\frac{\rho}{\pi}\right\|^2\right].$$

Proof [Proof of Theorem 4] We follow the strategy and notation of Section A.1.4.

1. **Forward step:** Along the simultaneous heat flow, Lemma 12 yields

$$\partial_t \chi^2_{\pi^X Q_t}(\rho_0^X Q_t) = -\frac{1}{2} J_{\chi^2,\pi^X Q_t}(\rho_0^X Q_t)\,,$$
$$\partial_t R_{q,\pi^X Q_t}(\rho_0^X Q_t) = -\frac{1}{2} J_{q,\pi^X Q_t}(\rho_0^X Q_t)\,.$$

Since π^X satisfies α-PI, then $\pi^X Q_t$ satisfies α_t-PI with $\alpha_t = \frac{\alpha}{1+\alpha t}$. Applying this yields

$$\partial_t \chi^2_{\pi^X Q_t}(\rho_0^X Q_t) = -\frac{1}{2} J_{\chi^2,\pi^X Q_t}(\rho_0^X Q_t) \leq -\alpha_t\, \chi^2_{\pi^X Q_t}(\rho_0^X Q_t)$$

and therefore

$$\chi^2_{\pi^Y}(\rho_0^Y) = \chi^2_{\pi^X Q_\eta}(\rho_0^X Q_\eta) \leq \frac{\chi^2_{\pi^X}(\rho_0^X)}{1+\alpha\eta}$$

upon integration.

Next, from Vempala and Wibisono (2019, Lemma 17), α_t-PI implies

$$\partial_t R_{q,\pi^X Q_t}(\rho_0^X Q_t) = -\frac{1}{2} J_{q,\pi^X Q_t}(\rho_0^X Q_t) \leq -\frac{2\alpha_t}{q}\left\{1 - \exp(-R_{q,\pi^X Q_t}(\rho_0^X Q_t))\right\}.$$

We split into two cases. If $R_{q,\pi^X}(\rho_0^X) \geq 1$, then as long as $R_{q,\pi^X Q_t}(\rho_0^X Q_t) \geq 1$ we can use the inequality $1 - \exp(-x) \geq \frac{1}{2}$ for $x \geq 1$, so that

$$\partial_t R_{q,\pi^X Q_t}(\rho_0^X Q_t) \leq -\frac{\alpha_t}{q}.$$

Integrating, we obtain

$$R_{q,\pi^Y}(\rho_0^Y) = R_{q,\pi^X Q_\eta}(\rho_0^X Q_\eta) \leq \left(R_{q,\pi^X}(\rho_0^X) - \frac{\log(1+\alpha\eta)}{q}\right) \vee 1.$$

In the second case, if $R_{q,\pi^X}(\rho_0^X) \leq 1$, then we use $1 - \exp(-x) \geq \frac{x}{2}$ for $x \in [0,1]$ to obtain

$$\partial_t R_{q,\pi^X Q_t}(\rho_0^X Q_t) \leq -\frac{\alpha_t}{q} R_{q,\pi^X Q_t}(\rho_0^X Q_t).$$

Integrating,

$$R_{q,\pi^Y}(\rho_0^Y) = R_{q,\pi^X Q_\eta}(\rho_0^X Q_\eta) \leq \frac{R_{q,\pi^X}(\rho_0^X)}{(1+\alpha\eta)^{1/q}}.$$

2. **Backward step:** Along the simultaneous backwards heat equation, Lemma 15 yields

$$\partial_t \chi^2_{\pi^Y Q_t^-}(\rho_0^Y Q_t^-) = -\frac{1}{2} J_{\chi^2,\pi^Y Q_t^-}(\rho_0^Y Q_t^-),$$
$$\partial_t R_{q,\pi^Y Q_t^-}(\rho_0^Y Q_t^-) = -\frac{1}{2} J_{q,\pi^Y Q_t^-}(\rho_0^Y Q_t^-).$$

Using entirely analogous arguments as in the forward step, we obtain

$$\chi^2_{\pi^X}(\rho_1^X) = \chi^2_{\pi^Y Q_\eta^-}(\rho_0^Y Q_\eta^-) \leq \frac{\chi^2_{\pi^Y}(\rho_0^Y)}{1+\alpha\eta}$$

for the chi-squared divergence,

$$R_{q,\pi^X}(\rho_1^X) = R_{q,\pi^Y Q_\eta^-}(\rho_0^Y Q_\eta^-) \leq \left(R_{q,\pi^Y}(\rho_0^Y) - \frac{\log(1+\alpha\eta)}{q}\right) \vee 1$$

for the Rényi divergence if $R_{q,\pi^Y}(\rho_0^Y) \geq 1$, and

$$R_{q,\pi^X}(\rho_1^X) = R_{q,\pi^Y Q_\eta^-}(\rho_0^Y Q_\eta^-) \leq \frac{R_{q,\pi^Y}(\rho_0^Y)}{(1+\alpha\eta)^{1/q}}$$

if $R_{q,\pi^Y}(\rho_0^Y) \leq 1$.

■

A.6. Convergence under LOI

Before giving the convergence proof under LOI, we recall the following property of the behavior of LOI under convolution.

Lemma 16 *Suppose that μ_0 satisfies (r, α_0)-LOI and μ_1 satisfies (r, α_1)-LOI. Then, $\mu_0 * \mu_1$ satisfies $(r, (\alpha_0^{-1} + \alpha_1^{-1})^{-1})$-LOI.*

Proof Let $X_0 \sim \mu_0$ and $X_1 \sim \mu_1$ be independent. Then, we can write

$$\mathrm{var}_{p,\mu_0*\mu_1}(\psi) = \mathbb{E}[\Phi(\psi^p(X_0 + X_1))] - \Phi(\mathbb{E}[\psi^p(X_0 + X_1)])$$

where $\Phi(x) \coloneqq x^{2/p}$. One can then deduce the conclusion of the lemma easily from the subadditivity of the Φ-entropy (Boucheron et al., 2013, Theorem 14.1). ■

Proof [Proof of Theorem 5] We follow the strategy and notation of Section A.1.4.

1. **Forward step:** Along the simultaneous heat flow, Lemma 12 yields

$$\partial_t R_{q,\pi^X Q_t}(\rho_0^X Q_t) = -\frac{1}{2} J_{q,\pi^X Q_t}(\rho_0^X Q_t)\,.$$

Since π^X satisfies (r, α)-LOI and $\mathcal{N}(0, tI)$ satisfies (r', t^{-1})-LOI for any $r' \in [1, 2]$ (see Latała and Oleszkiewicz, 2000, Corollary 1), then by Lemma 16, $\pi^X Q_t$ satisfies (r, α_t)-LOI with $\alpha_t = \frac{\alpha}{1+\alpha t}$.

Next, from Chewi et al. (2021a, Theorem 2), (r, α_t)-LOI implies

$$\begin{aligned}\partial_t R_{q,\pi^X Q_t}(\rho_0^X Q_t) &= -\frac{1}{2} J_{q,\pi^X Q_t}(\rho_0^X Q_t) \\ &\le -\frac{\alpha_t}{136q} \begin{cases} R_{q,\pi^X Q_t}(\rho_0^X Q_t)^{2-2/r}\,, & R_{q,\pi^X Q_t}(\rho_0^X Q_t) \ge 1\,, \\ R_{q,\pi^X Q_t}(\rho_0^X Q_t)\,, & R_{q,\pi^X Q_t}(\rho_0^X Q_t) \le 1\,. \end{cases}\end{aligned}$$

We split into two cases. If $R_{q,\pi^X}(\rho_0^X) \ge 1$, then as long as $R_{q,\pi^X Q_t}(\rho_0^X Q_t) \ge 1$,

$$\partial_t\, R_{q,\pi^X Q_t}(\rho_0^X Q_t)^{2/r-1} = (\frac{2}{r} - 1)\, \frac{\partial_t R_{q,\pi^X Q_t}(\rho_0^X Q_t)}{R_{q,\pi^X Q_t}(\rho_0^X Q_t)^{2-2/r}} \le -\frac{\alpha_t}{136q}\,(\frac{2}{r} - 1)$$

and therefore

$$\begin{aligned} R_{q,\pi^Y}(\rho_0^Y)^{2/r-1} &= R_{q,\pi^X Q_\eta}(\rho_0^X Q_\eta)^{2/r-1} \\ &\le \Big(R_{q,\pi^X}(\rho_0^X)^{2/r-1} - \frac{(2/r-1)\log(1+\alpha\eta)}{136q}\Big) \vee 1\,. \end{aligned}$$

In the second case, if $R_{q,\pi^X}(\rho_0^X) \le 1$, then

$$\partial_t R_{q,\pi^X Q_t}(\rho_0^X Q_t) \le -\frac{\alpha_t}{136q}\, R_{q,\pi^X Q_t}(\rho_0^X Q_t)\,.$$

Integrating,

$$R_{q,\pi^Y}(\rho_0^Y) = R_{q,\pi^X Q_\eta}(\rho_0^X Q_\eta) \le \frac{R_{q,\pi^X}(\rho_0^X)}{(1+\alpha\eta)^{1/(136q)}}\,.$$

2. **Backward step:** Along the simultaneous backwards heat equation, Lemma 15 yields

$$\partial_t R_{q,\pi^Y Q_t^-}(\rho_0^Y Q_t^-) = -\frac{1}{2} J_{q,\pi^Y Q_t^-}(\rho_0^Y Q_t^-)\,.$$

Using entirely analogous arguments as in the forward step, we obtain

$$\begin{aligned} R_{q,\pi^X}(\rho_1^X)^{2/r-1} &= R_{q,\pi^Y Q_\eta^-}(\rho_0^Y Q_\eta^-)^{2/r-1} \\ &\le \Big(R_{q,\pi^Y}(\rho_0^Y)^{2/r-1} - \frac{(2/r-1)\log(1+\alpha\eta)}{136q}\Big) \vee 1 \end{aligned}$$

if $R_{q,\pi^Y}(\rho_0^Y) \ge 1$, and

$$R_{q,\pi^X}(\rho_1^X) = R_{q,\pi^Y Q_\eta^-}(\rho_0^Y Q_\eta^-) \le \frac{R_{q,\pi^Y}(\rho_0^Y)}{(1+\alpha\eta)^{1/(136q)}}$$

if $R_{q,\pi^Y}(\rho_0^Y) \le 1$.

■

A.7. The proximal sampler as an entropy-regularized Wasserstein gradient flow

Proof [Proof of Theorem 10] Plugging (13) into (14) yields $\rho_k^Y = \gamma^Y$ with γ being the solution to

$$\min_{\substack{\gamma\in\mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d)\\ \gamma^X=\rho_k^X}} \Big\{\int \frac{1}{2\eta}\|x-y\|^2\,d\gamma(x,y) + H(\gamma)\Big\}\,,$$

which is clearly $\gamma(x,y) \propto \rho_k^X(x)\exp(-\frac{1}{2\eta}\|x-y\|^2)$. Thus, $\rho_k^Y = \gamma^Y = \rho_k^X * \mathcal{N}(0,\eta I)$.

Similarly, plugging (13) into (15) yields $\rho_{k+1}^X = \gamma^X$ with γ being the solution to

$$\min_{\substack{\gamma\in\mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d)\\ \gamma^Y=\rho_k^Y}} \Big\{\int \big[f(x) + \frac{1}{2\eta}\|x-y\|^2\big]\,d\gamma(x,y) + H(\gamma)\Big\}\,,$$

which is clearly $\gamma(x,y) \propto \rho_k^Y(y)\exp(-f(x) - \frac{1}{2\eta}\|x-y\|^2)$. Thus, ρ_{k+1}^X is induced by the conditional $\pi^{X|Y}(x\mid y) \propto_x \exp(-f(x) - \frac{1}{2\eta}\|x-y\|^2)$ from marginal distribution $Y\sim\rho_k^Y$. ■

A.8. The proximal point method as a limit of the proximal sampler

Proof [Proof of Theorem 11] With a general ϵ, following similar argument as in Section A.7, we can show that the updates (16)–(17) correspond to the sampling algorithm

$$y_k \sim \pi_\epsilon^{Y|X=x_k} = \mathcal{N}(x_k, \epsilon\eta I)\,, \tag{30a}$$

$$x_{k+1} \sim \pi_\epsilon^{X|Y=y_k} \propto \exp\Big[-\frac{1}{\epsilon}\Big(f(x) + \frac{1}{2\eta}\|x-y_k\|^2\Big)\Big]\,. \tag{30b}$$

As $\epsilon \searrow 0$, we see that (30a) converges to $y_k = x_k$, whereas (30b) converges to the proximal mapping $x_{k+1} = \arg\min_{x \in \mathbb{R}^d} \{f(x) + \frac{1}{2\eta} \|x - y_k\|^2\}$. Combining the two gives exactly the proximal point update $x_{k+1} = \mathrm{prox}_{\eta f}(x_k)$. In addition, the invariant distribution of this algorithm is $\pi_\epsilon^X \propto \exp(-f/\epsilon)$, which converges to a Dirac distribution concentrating on the minimizer of f (or a uniform distribution over the minimizer set of f). ■

It turns out that under some assumptions, the convergence rate of the updates (30) is independent of the entropy regularization level ϵ. We state and prove the result below for KL divergence only, but the result also holds for Rényi divergence and χ^2-divergence.

Theorem 17 *When f is α-strongly convex, the updates of the generalized proximal sampler algorithm converge to the stationary distribution $\pi_\epsilon^X \propto \exp(-f/\epsilon)$ with rate*

$$H_{\pi_\epsilon^X}(\rho_k^X) \le \frac{1}{(1+\alpha\eta)^{2k}} H_{\pi_\epsilon^X}(\rho_0^X)\,. \tag{31}$$

Proof The forward step (30a) can be modeled by the scaled diffusion

$$\partial_t \rho_t = \frac{\epsilon}{2} \Delta \rho_t \tag{32}$$

over the time interval $[0, \eta]$. Let $(Q_t^\epsilon)_{t \ge 0}$ denote the heat semigroup corresponding to (32). It follows from Lemma 12 that

$$\partial_t H_{\pi^X Q_t^\epsilon}(\rho_0^X Q_t^\epsilon) = -\frac{\epsilon}{2} J_{\pi^X Q_t^\epsilon}(\rho_0^X Q_t^\epsilon)\,. \tag{33}$$

Apparently, $\pi^X Q_t^\epsilon = \pi_\epsilon^X * \mathcal{N}(0, \epsilon t I)$. Thus, $\pi^X Q_t^\epsilon$ satisfies α_t-LSI with

$$\alpha_t = \frac{1}{\frac{\epsilon}{\alpha} + \epsilon t} = \frac{\alpha}{\epsilon\,(1+\alpha t)}\,, \tag{34}$$

where in the above we have used the fact that $\exp(-f/\epsilon)$ satisfies $\frac{\alpha}{\epsilon}$-LSI when f is α-strongly convex. Plugging (34) into (33) yields

$$\partial_t H_{\pi^X Q_t^\epsilon}(\rho_0^X Q_t^\epsilon) \le -\alpha_t\, H_{\pi^X Q_t^\epsilon}(\rho_0^X Q_t^\epsilon)\,. \tag{35}$$

Thus, as before,

$$H_{\pi_\epsilon^Y}(\rho_0^Y) = H_{\pi^X Q_\eta^\epsilon}(\rho_0^X Q_\eta^\epsilon) \le \frac{1}{1+\alpha\eta} H_{\pi^X}(\rho_0^X)\,. \tag{36}$$

The contraction rate in the backward direction is the same and the proof is similar to that of Theorem 3. This completes the proof. ■

Theorem 17 is true as long as $\exp(-f/\epsilon)$ satisfies (α/ϵ)-LSI. The latter is ensured when f is α-strongly convex; we ask whether it remains true under a weaker condition on f (such as α-PL).

Appendix B. Optimization proofs inspired by the proximal sampler

B.1. Alternative proof of the contractivity of the proximal map

The following theorem is well-known in optimization.

Theorem 18 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be α-strongly convex and differentiable. Then, the proximal mapping*

$$\mathrm{prox}_{\eta f}(y) \coloneqq \underset{x \in \mathbb{R}^d}{\arg\min} \left\{ f(x) + \frac{1}{2\eta} \|x - y\|^2 \right\}$$

is a $\frac{1}{1+\alpha\eta}$-contraction.

Here, we give a new proof of the theorem which translates the convergence proof of the proximal sampler in Lee et al. (2021b) to optimization.

We recall that α-strong convexity implies the α-PL inequality (or gradient domination inequality)

$$\|\nabla f(x)\|^2 \geq 2\alpha \{f(x) - \min f\} \qquad \text{for all } x \in \mathbb{R}^d \,,$$

which in turn implies the α-quadratic growth inequality

$$f(x) - \min f \geq \frac{\alpha}{2} \|x - x^\star\|^2 \qquad \text{for all } x \in \mathbb{R}^d \,,$$

with $x^\star = \arg\min f$, see Otto and Villani (2000); Blanchet and Bolte (2018).

Proof [Proof of Theorem 18] Let $f_x(z) \coloneqq f(z) + \frac{1}{2\eta} \|x - z\|^2$, and define f_y similarly. Then,

$$\begin{aligned} x' &\coloneqq \mathrm{prox}_{\eta f}(x) = \arg\min f_x \,, \\ y' &\coloneqq \mathrm{prox}_{\eta f}(y) = \arg\min f_y \,. \end{aligned}$$

Since f_x is $(\alpha + \frac{1}{\eta})$-strongly convex, then by applying the quadratic growth and PL inequalities,

$$\begin{aligned} \|x' - y'\|^2 &\leq \frac{2}{\alpha + 1/\eta} \{f_x(y') - f_x(x')\} \leq \frac{1}{(\alpha + 1/\eta)^2} \|\nabla f_x(y')\|^2 \\ &= \frac{1}{(\alpha + 1/\eta)^2} \left\|\nabla f(y') + \frac{1}{\eta}(y' - x)\right\|^2 \\ &= \frac{1}{(\alpha + 1/\eta)^2} \left\|-\frac{1}{\eta}(y' - y) + \frac{1}{\eta}(y' - x)\right\|^2 = \frac{1}{(1 + \alpha\eta)^2} \|x - y\|^2 \end{aligned}$$

where the last line uses the optimality condition $\nabla f(y') + \frac{1}{\eta}(y' - y) = 0$ from the definition of y'. ■

By comparing with the proof of Lee et al. (2021b, Lemma 2), we see that f_y is analogous to $H_{\pi^{X|Y=y}}$ for the proximal sampler.

At first glance, it may appear that the proof above only requires a PL inequality, and not strong convexity. However, this is not the case, as it in fact requires that f_x satisfies $(\alpha + 1/\eta)$-PL, which does not follow from (for example) the assumption that f satisfies α-PL.

B.2. Optimal contraction factor for the proximal point method under PL

Our proof uses the Hopf–Lax semigroup, guided by the following intuition. There is an analogy between the standard algebra $(+, \times)$ and the tropical algebra $(\inf, +)$; see, e.g., Baccelli et al. (1992, Section 9.4) or Ambrosio et al. (2021, Lecture 16). The following table describes these analogies.

$(+, \times)$	$(\inf, +)$
convolution	inf-convolution
Fourier transform	convex conjugate
diffusion	gradient flow
heat equation	Hamilton–Jacobi equation
heat semigroup	Hopf–Lax semigroup

As described in Section A.1.4, our proofs for the proximal sampler involve computing the time derivative of $t \mapsto H_{\pi^X Q_t}(\rho_0^X Q_t)$ where $(\pi^X Q_t)_{t \geq 0}$, $(\rho_0^X Q_t)_{t \geq 0}$ are simultaneously evolving according to the heat flow. In what follows, we will consider the time derivative of $t \mapsto f_t(x)$, where f_t is the Moreau envelope of f.

Proof [Proof of Theorem 9] Let us define, for $t > 0$,

$$f_{t,x}(z) \coloneqq f(z) + \frac{1}{2t} \|z - x\|^2 , \qquad x_t \coloneqq \arg\min f_{t,x} . \tag{37}$$

Then $x_t = \mathrm{prox}_{tf}(x)$ and $x \mapsto f_{t,x}(x_t)$ is the Moreau envelope of f. Recall the optimality condition

$$\nabla f(x_t) + \frac{1}{t} (x_t - x) = 0 .$$

The Moreau envelope satisfies the Hamilton–Jacobi equation

$$\partial_t f_{t,x}(x_t) = \langle \underbrace{\nabla f_{t,x}(x_t)}_{=0}, \dot{x}_t \rangle - \frac{1}{2t^2} \|x_t - x\|^2 .$$

Using the PL inequality,

$$\begin{aligned}
\partial_t f_{t,x}(x_t) &= -\frac{\alpha}{2t\,(1+\alpha t)} \|x_t - x\|^2 - \frac{1}{2t^2\,(1+\alpha t)} \|x_t - x\|^2 \\
&= -\frac{\alpha}{2t\,(1+\alpha t)} \|x_t - x\|^2 - \frac{1}{2\,(1+\alpha t)} \|\nabla f(x_t)\|^2 \\
&\leq -\frac{\alpha}{2t\,(1+\alpha t)} \|x_t - x\|^2 - \frac{\alpha}{1+\alpha t} \{f(x_t) - f^\star\}
\end{aligned}$$

which yields

$$\partial_t \{f_{t,x}(x_t) - f^\star\} \leq -\frac{\alpha}{1+\alpha t} \{f_{t,x}(x_t) - f^\star\} .$$

Integrating this yields[4]

$$f_{\eta,x}(x_\eta) - f^\star \leq \{f(x) - f^\star\} \exp\Big(-\int_0^\eta \frac{\alpha}{1+\alpha t} \,\mathrm{d}t\Big) = \frac{1}{1+\alpha\eta} \{f(x) - f^\star\} .$$

4. Denote by $(Q_t^{\mathrm{HL}})_{t \geq 0}$ the Hopf–Lax semigroup defined by $Q_t^{\mathrm{HL}} f(x) = f_{t,x}(x_t)$. One can check that $Q_t^{\mathrm{HL}} f(x^\star) = f(x^\star)$ where $x^\star = \arg\min f$. So, we can rewrite this inequality as $Q_t^{\mathrm{HL}} f(x) - Q_t^{\mathrm{HL}} f(x^\star) \leq \frac{1}{(1+\alpha t)} \{f(x) - f(x^\star)\}$.

Hence,

$$\begin{aligned}\frac{1}{1+\alpha\eta}\left\{f(x)-f^\star\right\} &\geq f(x')-f^\star+\frac{1}{2\eta}\left\|x'-x\right\|^2 = f(x')-f^\star+\frac{\eta}{2}\left\|\nabla f(x')\right\|^2 \\ &\geq f(x')-f^\star+\alpha\eta\left\{f(x')-f^\star\right\} = (1+\alpha\eta)\left\{f(x')-f^\star\right\}.\end{aligned}$$

This completes the proof. ■