Improved dimension dependence of a proximal algorithm for sampling

Jiaojiao Fan* JIAOJIAOFAN@GATECH.EDU
Georgia Institute of Technology

Bo Yuan* BYUAN48@GATECH.EDU
Georgia Institute of Technology

Yongxin Chen YONGCHEN@GATECH.EDU
Georgia Institute of Technology

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We propose a sampling algorithm that achieves superior complexity bounds in all the classical settings (strongly log-concave, log-concave, Logarithmic-Sobolev inequality (LSI), Poincaré inequality) as well as more general settings with semi-smooth or composite potentials. Our algorithm is based on the proximal sampler introduced in Lee et al. (2021). The performance of this proximal sampler is determined by that of the restricted Gaussian oracle (RGO), a key step in the proximal sampler. The main contribution of this work is an inexact realization of RGO based on approximate rejection sampling. To bound the inexactness of RGO, we establish a new concentration inequality for semi-smooth functions over Gaussian distributions, extending the well-known concentration inequality for Lipschitz functions. Applying our RGO implementation to the proximal sampler, we achieve state-of-the-art complexity bounds in almost all settings. For instance, for strongly log-concave distributions, our method has complexity bound $\tilde{\mathcal{O}}(\kappa d^{1/2})$ without warm start, better than the minimax bound for MALA. For distributions satisfying the LSI, our bound is $\tilde{\mathcal{O}}(\hat{\kappa} d^{1/2})$ where $\hat{\kappa}$ is the ratio between smoothness and the LSI constant, better than all existing bounds.

Keywords: Sampling, MCMC, non-asymptotic analysis, concentration inequality, proximal algorithm, semi-smooth functions

1. Introduction

The task of sampling from a target distribution $\nu \propto \exp(-f)$ on $\mathbb{R}^d$ plays an instrumental role in Bayesian inference (Ghosal and Van der Vaart, 2017), scientific computing (Pulido and van Leeuwen, 2019), and machine learning (Murphy, 2012; Liu and Wang, 2016; Fan et al., 2021). Myriad works have been devoted to the theoretical analysis of sampling, ranging from the smooth strongly log-concave setting (Dalalyan, 2017; Vempala and Wibisono, 2019; Durmus et al., 2019) to non-log-concave (Chewi et al., 2022a) or non-smooth settings (Durmus et al., 2018; Salim and Richtárik, 2020; Fan et al., 2022).

In this work we make inroads towards better non-asymptotic complexity bound for sampling by focusing on the proximal sampler (Lee et al., 2021). The proximal sampler is essentially a Gibbs sampler over an augmented distribution based on the target distribution. The difficulty of implementing the proximal sampler comes from restricted Gaussian oracle (RGO) – a task of sampling from $\exp(-f(\cdot) - \frac{1}{2\eta}\|x-y\|^2)$ for some given step size $\eta > 0$ and $y \in \mathbb{R}^d$. Given that RGO is implementable and exact, the proximal sampler can converge exponentially fast to the target distribution exponentially under mild assumptions (Lee et al., 2021; Chen et al., 2022). However, the

* Equal contribution

total complexity of the algorithm heavily depends on the implementation of RGO. Except for some special settings (Gopi et al., 2022), the best-known dimension dependence of the proximal sampler is $\tilde{\mathcal{O}}(d)$ (Chen et al., 2022; Liang and Chen, 2022c). This is worse than the best-known bounds of other sampling methods such as underdamped Langevin Monte Carlo (ULMC) (Shen and Lee, 2019) or Metropolis-Adjusted Langevin Algorithm (MALA) (Wu et al., 2021).

In this paper, we aim to improve the dimension dependence of the RGO and thus the proximal sampler. We first introduce an inexact RGO algorithm based on approximate rejection sampling. The underpinning of our analysis for this RGO algorithm is a novel Gaussian concentration inequality for semi-smooth functions, which extends the Gaussian concentration inequality for Lipschitz functions (Boucheron et al., 2013, Theorem 5.6). Our proof is based on the argument of Maurey and Pisier. Our RGO algorithm is an inexact algorithm, and the crux is to bound the step size such that the output of our RGO is close enough to the exact RGO. The RGO step size is then processed to only depend on the order of the aforementioned concentration inequality.

Contribution First, we propose a novel realization of RGO. Our RGO implementation is inexact but can achieve a better step size pertaining to the dimension. Next, we prove a Gaussian concentration inequality for semi-smooth functions, which could be of independent interest. It can recover the order of the well-known Gaussian concentration inequality for Lipschitz functions. This concentration inequality is the underpinning of our RGO analysis. Third, we control the accumulated error from our inexact RGO algorithm in terms of both total variation and Wasserstein metric. This, combined with the existing proximal sampler convergence results, gives state-of-the-art sampling convergence results under various conditions (see Table 1). Finally, we extend all the results for semi-smooth potential (1) to composite potential (4) (see Table 2).

Table 1: Complexity bounds for sampling from semi-smooth potentials satisfying (1). Here $L_1, \delta, C_{\mathrm{LSI}}, C_{\mathrm{PI}}, \mathcal{M}_2, \mathcal{M}_4$ denote the smoothness constant, accuracy, LSI constant, Poincaré inequality constant, second moment, and fourth moment of the target distribution.

Assumption	Source	Complexity	Metric
β-strongly log-concave	Chen et al. (2022)	$\tilde{\mathcal{O}}(L_1 d/\beta)$	Rényi
	Wu et al. (2021)	$\tilde{\Omega}(L_1\sqrt{d}/\beta)$ (Warm start)	TV
	Shen and Lee (2019)	$\tilde{\mathcal{O}}\left((\frac{L_1}{\beta})^{7/6}(\frac{2}{\delta}\sqrt{\frac{d}{\beta}})^{1/3} + \frac{L_1}{\beta}(\frac{2}{\delta}\sqrt{\frac{d}{\beta}})^{2/3}\right)$	W_2
	Proposition 10, 30	$\tilde{\mathcal{O}}(L_1\sqrt{d}/\beta)$	TV/W_2/χ^2
log-concave	Liang and Chen (2022c)	$\tilde{\mathcal{O}}(\sqrt{\mathcal{M}_4} L_\alpha^{\frac{2}{\alpha+1}} d/\delta)$	TV
	Proposition 11	$\tilde{\mathcal{O}}(\mathcal{M}_2 L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}/\delta)$	TV
log-Sobolev	Chen et al. (2022)	$\tilde{\mathcal{O}}(L_1 d/C_{\mathrm{LSI}})$	Rényi
	Proposition 13, 31	$\tilde{\mathcal{O}}(L_1\sqrt{d}/C_{\mathrm{LSI}})$	TV/χ^2
Poincaré	Liang and Chen (2022a)	$\tilde{\mathcal{O}}(L_\alpha^{\frac{2}{\alpha+1}} d^2/C_{\mathrm{PI}})$	Rényi
	Proposition 14, 31	$\tilde{\mathcal{O}}(L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{2\alpha+1}{\alpha+1}}/C_{\mathrm{PI}})$	TV/χ^2

Related works Our RGO algorithm is inspired by a rejection sampling-based RGO (Gopi et al., 2022) for distributions with Lipschitz potentials. Compared to their algorithm, we add a linear function to the potential to have a delicate stationary point, and we simplify the rejection rule by eliminating part of the randomness. The convergence of the proximal sampler is established for strongly log-concave distributions in Lee et al. (2021). Chen et al. (2022) then extended the class of target distributions to a much wilder range, including Poincaré inequality. Some other works (Liang and Chen, 2022b,c,a; Gopi et al., 2022) consider the convergence under weaker smoothness conditions, e.g. semi-smooth potential or composite potential. We also mention a concurrent work (Altschuler and Chewi, 2023) that achieves similar complexity bounds as ours for smooth potentials with a very different RGO implementation based on MALA and ULMC. Other than the analysis for the proximal sampler, there exist numerous works for other sampling methods. To name a few, Wu et al. (2021); Shen and Lee (2019) study strongly-log-concave and smooth potential, Chewi et al. (2022a); Erdogdu and Hosseinzadeh (2021); Erdogdu et al. (2022) study the non-log-concave potential, and Nguyen et al. (2021); Durmus et al. (2018, 2019); Salim and Richtárik (2020); Bernton (2018) study the composite potential. More detailed discussions appear in §4.

Comparison to the cocurrent work (Altschuler and Chewi, 2023) **Algorithm**: We both use the proximal sampler but with different implementations of RGO. Our RGO is based on approximate rejection sampling and has complexity $\mathcal{O}(1)$. In contrast, their RGO is based on MALA with a ULMC warm start and has complexity $\tilde{\mathcal{O}}(\sqrt{d})$. Our RGO is easier in implementation and parameter tuning. **Results**: For log-smooth distribution, we share the same high-accuracy complexity results in χ^2. Our results cover the semi-smooth and composite potentials, but they only cover log-smooth distributions. Thus, our complexity results are more general and include theirs as a special case. **Contributions**: The main contribution of our work includes a new RGO implementation and a new concentration inequality. The main contribution of Altschuler and Chewi (2023) is a warm start result using ULMC. One disadvantage of our result is that it is not clear how our method can be used outside the proximal sampler framework, but their warm start method is applicable in various algorithms.

Table 2: Complexity bounds for sampling from composite potentials satisfying (4).

Assumption	Source	Complexity	Metric
log-Sobolev	Nguyen et al. (2021)	$\tilde{\mathcal{O}}(n\max\{L_{\alpha_j}^2\}d/\delta)^{\max\{1/\alpha_j\}}/C_{\mathrm{LSI}}^{1+\max\{1/\alpha_j\}})$	KL
	Liang and Chen (2022a)	$\tilde{\mathcal{O}}(\sum_{j=1}^n L_{\alpha_j}^{2/(\alpha_j+1)}d/C_{\mathrm{LSI}})$	KL
	Proposition 19	$\tilde{\mathcal{O}}\left(\left(\sum_{j=1}^n L_{\alpha_j}^{1/(\alpha_j+1)}d^{\alpha_j/(2(\alpha_j+1))}\right)^2/C_{\mathrm{LSI}}\right)$	TV
Poincaré	Liang and Chen (2022a)	$\tilde{\mathcal{O}}(\sum_{j=1}^n L_{\alpha_j}^{2/(\alpha_j+1)}d^2/C_{\mathrm{PI}})$	Rényi
	Proposition 19	$\tilde{\mathcal{O}}\left(\left(\sum_{j=1}^n L_{\alpha_j}^{1/(\alpha_j+1)}d^{\alpha_j/(2(\alpha_j+1))}\right)^2/C_{\mathrm{PI}}\right)$	TV

Organization The paper is organized as follows. In §2, we review the proximal sampler. We then introduce our RGO implementation and present the improved dimension dependence of RGO in §3. We establish the sampling convergence under different conditions in §4. In §5, we extend all the

analysis (§3-4) for semi-smooth potentials to the composite potentials. We conclude in §6 with a discussion of future research directions.

2. Background: the proximal sampler

We consider sampling from the distribution $\nu \propto \exp(-f(x))$ where the potential f is bounded from below and is L_α-α-semi-smooth, i.e., f satisfies

$$\|f'(u) - f'(v)\| \le L_\alpha \|u - v\|^\alpha, \ \ \forall u, v \in \mathbb{R}^d \tag{1}$$

for $L_\alpha > 0$ and $\alpha \in [0, 1]$. Here f' represents a subgradient of f. When $\alpha > 0$, this subgradient can be replaced by the gradient. The condition (1) implies f is L_1-smooth when $\alpha = 1$ and a Lipschitz function satisfies (1) with $\alpha = 0$.

Algorithm 1: The proximal Sampler (Lee et al., 2021)

1 **Input:** Target distribution $\exp(-f(x))$, step size $\eta > 0$, initial point x_0
2 **for** $t = 1, \dots, T$ **do**
3 Sample $y_t \sim \pi^{Y|X}(y|x_{t-1}) \propto \exp(-\frac{1}{2\eta}\|x_{t-1} - y\|^2)$
4 Sample $x_t \sim \pi^{X|Y}(x|y_t) \propto \exp(-f(x) - \frac{1}{2\eta}\|x - y_t\|^2)$
5 **end**
6 **Return** x_T

The algorithm we adopt is the proximal sampler (or the alternating sampler) proposed by Lee et al. (2021), shown in Algorithm 1. It is essentially a Gibbs sampling method for the joint distribution $\pi^{XY}(x, y) \propto \exp\left(-f(x) - \frac{1}{2\eta}\|x - y\|^2\right)$. The target distribution ν is the X-marginal distribution of π^{XY}. The proximal sampler alternates between two sampling steps. The first one is to sample from the conditional distribution of Y given x_{t-1}; it is a Gaussian distribution $\pi^{Y|X}(y|x_{t-1}) = \mathcal{N}(x_{t-1}, \eta\mathbf{I})$ and thus trivial to implement. The paramount part of this method is the second sub-step, which is the restricted Gaussian oracle for f to sample from the conditional distribution

$$\pi^{X|Y}(x|y_t) \propto \exp\left(-f(x) - \frac{1}{2\eta}\|x - y_t\|^2\right),$$

which is not always easy to implement. To simplify the notation, we often use $\pi^{X|Y}$ to represent $\pi^{X|Y}(x|y)$. If line 4 can be implemented exactly, then the proximal sampler is unbiased because the iterates $\{(x_t, y_t)\}_{t\in\mathbb{N}}$ form a reversible Markov chain with the stationary distribution π^{XY}.

2.1. Convergence of proximal sampler given exact RGO

Next we briefly discuss the convergence property of Algorithm 1. Recall that a probability distribution ν satisfies log-Sobolev inequality (LSI) with constant $C_{\text{LSI}} > 0$ (C_{LSI}-LSI) if for all smooth functions $u : \mathbb{R}^d \to \mathbb{R}$, the following holds:

$$\mathbb{E}_\nu[u^2 \log u^2] - \mathbb{E}_\nu[u^2] \log \mathbb{E}_\nu[u^2] \le \frac{2}{C_{\text{LSI}}} \mathbb{E}_\nu[\|\nabla u\|^2]. \tag{LSI}$$

Denote Wasserstein-2 distance as $W_2(\nu,\mu) := \inf_{\gamma\in\Pi(\nu,\mu)} \int \|x-y\|^2 \mathrm{d}\gamma(x,y)$, where $\Pi(\nu,\mu)$ is the set of joint distributions of marginal distributions ν and μ. For a probability measure $\mu \ll \nu$, we define the KL divergence $H_\nu(\mu) := \int \mu \log \frac{\mu}{\nu}$, and the chi-squared divergence $\chi^2_\nu(\mu) := \int \frac{\mu^2}{\nu} - 1$.

A distribution ν satisfies (LSI) implies that it also satisfies Talagrand inequality (Otto and Villani, 2000), i.e., $W_2(\nu,\mu) \le \sqrt{\frac{2}{C_{\text{LSI}}} H_\nu(\mu)}$ for any probability distribution $\mu \ll \nu$ with finite second moment. A probability distribution ν satisfies the Poincaré inequality (PI) with constant $C_{\text{PI}} > 0$ if for any smooth bounded function $u : \mathbb{R}^d \to \mathbb{R}$, it holds that

$$\mathrm{Var}_\nu(u) \le \frac{1}{C_{\text{PI}}} \mathbb{E}[\|\nabla u\|^2]. \tag{PI}$$

The (LSI) implies the (PI) with the same constant. In the following theorem, we assume x_0 is sampled from some initialization distribution μ_0.

Theorem 1 (Convergence of proximal sampler (Chen et al., 2022)) *Assuming the RGO is exact, we denote the corresponding iterates $y_t \sim \psi_t$ and $x_t \sim \mu_t$.*

1) If ν is log-concave (i.e. f is convex), $H_\nu(\mu_t) \le W_2^2(\mu_0,\nu)/(t\eta)$;

2) If ν satisfies C_{LSI}-LSI, $H_\nu(\mu_t) \le H_\nu(\mu_0)/(1+C_{\text{LSI}}\eta)^{2t}$;

3) If ν satisfies C_{PI}-PI, $\chi^2_\nu(\mu_t) \le \chi^2_\nu(\mu_0)/(1+C_{\text{PI}}\eta)^{2t}$.

Note that if ν is β-strongly-log-concave, then ν satisfies β-LSI. So the convergence of strongly-log-concave ν is also implicitly contained in Theorem 1.

2.2. Existing RGO implementations

Except for a few special cases with closed-form realizations (Lee et al., 2021; Mou et al., 2022), most of the existing implementations of RGO are based on rejection sampling. Denote the function $f_y^\eta := f(x) + \frac{1}{2\eta}\|x-y\|^2$. If $\pi^{X|Y}$ is strongly-log-concave, i.e., f_y^η is strongly-convex and smooth, one can naturally use a Gaussian with variance being the convexity of f_y^η and mean being the minimizer of f_y^η as the proposal distribution. In this case, with the step size $\eta = \Theta(1/(L_1 d))$, the expected number of iterations for rejection sampling is $\mathcal{O}(1)$ (Liang and Chen, 2022b; Chewi et al., 2022b). When the potential f is not smooth but L_0-Lipschitz, Liang and Chen (2022b) shows that a similar proposal gives $\mathcal{O}(1)$ complexity for rejection sampling when the step size is $\eta = \Theta(1/(L_0^2 d))$. In fact, it can be shown that, at least for quadratic potentials, the dependence $\eta = \Theta(1/d)$ is inevitable to ensure $\mathcal{O}(1)$ complexity for rejection sampling-based RGO. On the other hand, Gopi et al. (2022) proposes an approximate rejection sampling scheme for an inexact RGO for Lipschitz f which uses a larger step size $\eta = \tilde{\mathcal{O}}(1/L_0^2)$ to ensure $\mathcal{O}(1)$ complexity of RGO, rendering better dimension dependence of the proximal sampler. On a different route, one can apply any MCMC algorithm to implement RGO with dimension-free step size η, but the complexity of each RGO step is dimension dependent, thus, the overall complexity of this strategy is not necessarily better (Lee et al., 2021).

3. Improved dimension dependence of RGO for semi-smooth potential

The step size η of the proximal sampler is pivotal to the total complexity. As can be seen from Theorem 1 a larger step size points to a faster convergence rate. However, larger η also means higher complexity for each RGO step. In Liang and Chen (2022a), it is shown that, with $\eta = \Theta(1/(L_\alpha^{\frac{2}{\alpha+1}} d))$,

it is possible to use rejection sampling to realize RGO with $\mathcal{O}(1)$ complexity. In this section, we improve the step size to $\eta = \tilde{\mathcal{O}}(1/(L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}))$ while maintaining the same $\mathcal{O}(1)$ complexity by designing a new RGO implementation based on approximate rejection sampling.

Algorithm 2: Approximate rejection sampling implementation of RGO

Input: L_α-α-semi-smooth function $f(x)$, step size $\eta > 0$, current point y
Compute x_y such that $f'(x_y) + \frac{1}{\eta}(x_y - y) = 0$. Denote $g(x) = f(x) - \langle f'(x_y), x\rangle$.
repeat
  Sample x, z from the distribution $\phi(\cdot) \propto \exp(-\frac{1}{2\eta}\|\cdot - x_y\|_2^2)$
  $\rho = \exp(g(z) - g(x))$
  Sample u uniformly from $[0, 1]$.
until $u \le \frac{1}{2}\rho$;
Return x

Our proposed RGO implementation is in Algorithm 2. This is a variant of Gopi et al. (2022, Algorithm 2). Compared to Gopi et al. (2022), we eliminate the randomness in the inner loop of rejection sampling. More importantly, we modify the proposal distribution to ensure that the mean of the proposal Gaussian distribution ϕ is the same as the stationary point of the function f_y^η. This modification is justified by the following lemma. This modification is crucial to control the accuracy and complexity of our RGO algorithm via concentration inequality as will be seen in Theorem 4.

Lemma 2 *Sampling from $\pi^{X|Y}(x|y) \propto \exp(-f(x) - \frac{1}{2\eta}\|x - y\|^2)$ is equivalent to sampling from distribution $\propto \exp(-g(x) - \frac{1}{2\eta}\|x - x_y\|^2)$, where $g(x) = f(x) - \langle f'(x_y), x\rangle$ and x_y satisfies $f'(x_y) + \frac{1}{\eta}(x_y - y) = 0$.*

Thanks to Lemma 2, sampling from distribution $\propto \exp(-g(x) - \frac{1}{2\eta}\|x - x_y\|^2)$ is equivalent to the original RGO. Note that $g(x)$ shares the same semi-smooth constant as $f(x)$. Algorithm 2 requires the calculation of the stationary point x_y of $f(x) + \frac{1}{2\eta}\|x - y\|^2$. This could be challenging when f is non-convex. In that case, we can get an approximate stationary point instead. We can control the error introduced by approximate x_y and achieve the same complexity bound; the detailed discussions are postponed to §D. In the main paper, we assume the stationary point is solved exactly to make the analysis more comprehensible. To facilitate the analysis of Algorithm 2, we introduce another threshold $\bar{\rho} := \min(\rho, 2)$. As we will see, the acceptance probability in our RGO algorithm is directly related to $\bar{\rho}$.

Lemma 3 *Denote $\hat{\pi}^{X|Y}$ as the distribution of the output of Algorithm 2. Let ϕ be defined as in Algorithm 2. Define the random variables $\bar{\rho} := \min(\rho, 2)$, $V := \mathbb{E}[\rho|x]$, and $\overline{V} := \mathbb{E}[\bar{\rho}|x]$. Then*

$$\frac{\mathrm{d}\pi^{X|Y}}{\mathrm{d}x} = \frac{\mathrm{d}\phi}{\mathrm{d}x} \cdot \frac{\exp(-g(x))}{\mathbb{E}_{x\sim\phi}\exp(-g(x))} = \frac{\mathrm{d}\phi}{\mathrm{d}x} \cdot \frac{\mathbb{E}[\rho|x]}{\mathbb{E}[\rho]} = \frac{\mathrm{d}\phi}{\mathrm{d}x} \cdot \frac{V}{\mathbb{E}[V]},$$
$$\frac{\mathrm{d}\hat{\pi}^{X|Y}}{\mathrm{d}x} = \frac{\mathrm{d}\phi}{\mathrm{d}x} \cdot \frac{\mathbb{E}[\bar{\rho}|x]}{\mathbb{E}[\bar{\rho}]} = \frac{\mathrm{d}\phi}{\mathrm{d}x} \cdot \frac{\overline{V}}{\mathbb{E}[\overline{V}]}.$$

Moreover, the acceptance probability of rejection sampling is $\frac{1}{2}\mathbb{E}[\bar{\rho}] = \frac{1}{2}\mathbb{E}[\overline{V}]$.

Lemma 3 shows that the gap between ground truth $\pi^{X|Y}$ and our return $\hat{\pi}^{X|Y}$ is caused by the discrepancy between ρ and $\bar{\rho}$. At a high level, we need to control the probability that ρ and $\bar{\rho}$ are different by choosing a small enough step size. If η is sufficiently small, then the proposal distribution ϕ has a small variance and so that ρ is concentrated around 1 with high probability, which means ρ and $\bar{\rho}$ are equal with high probability. To give such a probabilistic bound, we establish a novel concentration inequality bound for a semi-smooth function over the Gaussian random variable.

Theorem 4 (Gaussian concentration inequality for semi-smooth functions) *Let $X \sim \mathcal{N}(m, \eta\mathbf{I})$ be a Gaussian random variable in $\mathbb{R}^d$, and let ℓ be an L_α-α-semi-smooth function. Assume $\ell'(m) = 0$. Then for any $r > 0,\ 0 \le \alpha \le 1$, one has*

$$\Pr(\ell(X) - \mathbb{E}(\ell(X)) \ge r) \le \left(1 - \frac{\epsilon}{d}\right)^{-d/2} \exp\left(-\frac{C\epsilon^{\frac{\alpha}{1+\alpha}} r^{\frac{2}{1+\alpha}}}{L_\alpha^{\frac{2}{1+\alpha}} d^{\frac{\alpha}{1+\alpha}} \eta}\right), \quad \forall \epsilon \in (0, d) \tag{2}$$

$$\textit{where} \qquad C = (1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}} \left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}} 2^{\frac{1-\alpha}{1+\alpha}}. \tag{3}$$

Sketch of proof for Theorem 4: By Maurey and Pisier argument (Pisier, 2006) and Young's inequality $\|G\|^{2\alpha} \le \alpha\|G\|^2/\omega + (1-\alpha)\omega^{\frac{\alpha}{1-\alpha}}$ for $\forall\, \omega > 0$, we have

$$\begin{aligned}\Pr(\ell(X) - \mathbb{E}(\ell(X)) \ge r) &\le \inf_{\lambda > 0} \frac{\mathbb{E}_G \exp(\frac{\pi^2}{8} L_\alpha^2 \eta \lambda^2 \|G\|_2^{2\alpha})}{\exp(\lambda r)} \\ &\le \frac{\mathbb{E}_G \exp\left(\frac{\pi^2}{8} L_\alpha^2 \eta \lambda^2 (\alpha\|G\|^2/\omega + (1-\alpha)\omega^{\frac{\alpha}{1-\alpha}})\right)}{\exp(\lambda r)}.\end{aligned}$$

Invoking the closed-form second moment of Gaussian distribution and choosing proper λ and ω, we can establish the result. The full proof is in §B.1.

In Theorem 4, ϵ is a tunable parameter, and we leave its choice to the user. It causes a trade-off between the coefficients $\left(1 - \frac{\epsilon}{d}\right)^{-d/2}$, which is in front of $\exp()$, and $\epsilon^{\frac{\alpha}{1+\alpha}}$, which is inside $\exp()$. We remark that the assumption $\ell'(m) = 0$ can be relaxed at the cost of an additional penalty coefficient (see Proposition 25). We also note that when r is in a small range, we can always get sub-Gaussian tail no matter what the α value is (see Proposition 22).

Remark 5 *One term in the coefficient (3) in Theorem 4 is not well-defined when $\alpha = 0$, but $C \to C_0 = 2/\pi^2$ as $\alpha \to 0$ and C is monotone w.r.t. α. Thus, for any $0 \le \alpha \le 1$, we have*

$$\Pr(\ell(X) - \mathbb{E}(\ell(X)) \ge r) \le \left(1 - \frac{\epsilon}{d}\right)^{-d/2} \exp\left(-\frac{2\epsilon^{\frac{\alpha}{1+\alpha}} r^{\frac{2}{1+\alpha}}}{\pi^2 L_\alpha^{\frac{2}{1+\alpha}} d^{\frac{\alpha}{1+\alpha}} \eta}\right), \quad \forall \epsilon \in (0, d).$$

Our concentration bound (2) can recover the known bounds in two extreme cases. When $\alpha = 0$, ℓ is Lipschitz, the RHS of (2) is $\propto \exp(-r^2/(L_0^2\eta))$. This recovers the sub-Gaussian tail for Lipschitz functions (Ledoux, 1999; Boucheron et al., 2013). When $\alpha = 1$, the RHS of (2) is $\propto \exp(-r/(L_1\sqrt{d}\eta))$, which recovers the sub-exponential tail in the Laurent-Massart bound (Laurent and Massart, 2000) for χ^2 distribution and Hanson-Wright inequality (Hanson and

Wright, 1971; Wright, 1973; Rudelson and Vershynin, 2013) for large enough r. A more detailed discussion is in §B.3.

With this concentration inequality, we are able to find a small enough η such that ρ and $\bar{\rho}$ are the same with high probability, and thus the difference between $\hat{\pi}^{X|Y}$ and $\pi^{X|Y}$ is small.

Theorem 6 (RGO complexity in total variation) *Assume f satisfies (1). For $\forall \zeta > 0$, if*

$$\eta \leq \left(49 L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}} (1 + \log(1 + 12/\zeta))\right)^{-1},$$

then Algorithm 2 returns a random point x that has ζ total variation distance to the distribution proportional to $\pi^{X|Y}(\cdot|y)$. Furthermore, if $0 < \zeta < 1$, then the algorithm access $\mathcal{O}(1)$ many $f(x)$ in expectation.

Theorem 7 (RGO complexity in Wasserstein distance) *Assume f satisfies (1). For $\forall \zeta > 0$, if*

$$\eta \leq \min\left(\left(49 L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}} (2 + \log(1 + 192(d^2 + 2d)/\zeta^4))\right)^{-1}, 1\right),$$

then Algorithm 2 returns a random point x that has ζ Wasserstein-2 distance to the distribution $\pi^{X|Y}(\cdot|y)$. Furthermore, if $0 < \zeta < 2\sqrt{2d}$, then the algorithm access $\mathcal{O}(1)$ many $f(x)$ in expectation.

Sketch of proof for Theorem 6 and Theorem 7: By definition and some elementary inequalities, the error $\|\pi^{X|Y} - \hat{\pi}^{X|Y}\|_{\mathrm{TV}}$ can be bounded by $2\mathbb{E}[\rho \mathbf{1}_{\rho \geq 2}] \leq 2\sum_{i=1}^{\infty} \exp((i+1)\log 2) \Pr((g(z) - g(x))/\log 2 \geq i)$, which is then bounded by applying our concentration inequality and choosing $\eta = \tilde{\mathcal{O}}(1/(L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}))$. Specifically, we apply our concentration inequality to $\ell(x) = g(x) = f(x) - \langle f'(x_y), x\rangle$. Wasserstein distance bound follows similarly. The full proofs are in §C.

Both step sizes have better dimension dependence than existing methods $\mathcal{O}(1/d)$ for non-convex semi-smooth potential (Liang and Chen, 2022a). In addition, Theorem 6 can recover Lemma 5.5 in Gopi et al. (2022). We further extend the results in Theorem 6, 7 to the χ^2-divergence in §E.1.

4. Improved complexity bounds of proximal sampling for semi-smooth potential

Our overall algorithm uses the RGO implementation in Algorithm 2 to implement line 4 in the proximal sampler (Algorithm 1). Combining the theoretical properties of our RGO implementation in §3, and the existing convergence results for the exact proximal sampler in Theorem 1, we establish superior sampling complexity bounds under various conditions.

Denote the distributions of the iterations y_t and x_t of the ideal proximal sampler by ψ_t and μ_t respectively. Our RGO implementation is not exact, rendering different distributions along the iterations of the proximal sampler, denoted by $y_t \sim \hat{\psi}_t$ and $x_t \sim \hat{\mu}_t$. To establish the final complexity bounds, we need to quantify the difference between $\hat{\psi}_t$ (or $\hat{\mu}_t$) and ψ_t (or μ_t) caused by the inexact RGO. We first present the following lemmas to control the accumulated error of the inexact RGO. Lemma 8 is also informally mentioned in Lee et al. (2021, §A). Lemma 9 is proved based on formulating the RGO as a backward diffusion, and then adopting a coupling argument, which is introduced in Chen et al. (2022, §A). The detailed proofs are in Appendix A.

Lemma 8 *Assume the output of Algorithm 2 follows $\hat{\pi}^{X|Y}(\cdot|y)$ that can achieve*

$$\left\|\hat{\pi}^{X|Y}(\cdot|y) - \pi^{X|Y}(\cdot|y)\right\|_{\mathrm{TV}} \leq \zeta$$

for $\forall\, y$, then $\|\hat{\mu}_T - \mu_T\|_{\mathrm{TV}} \leq \zeta T$.

Lemma 9 *Assume the output of Algorithm 2 follows $\hat{\pi}^{X|Y}(\cdot|y)$ that can achieve*

$$W_2\left(\hat{\pi}^{X|Y}(\cdot|y), \pi^{X|Y}(\cdot|y)\right) \leq \zeta$$

for $\forall\, y$, and suppose the target distribution $\nu \propto \exp(-f)$ is log-concave, then $W_2(\hat{\mu}_T, \mu_T) \leq \zeta T$.

We next establish the complexity bounds of our algorithm in three settings, with strongly convex, convex, and non-convex potentials. The idea of proof is simple. Assume we iterate T times in Algorithm 1. If the desired final error is δ, then we choose the RGO accuracy $\zeta = \Theta(\delta/T)$ by the above lemmas to ensure the accumulated error is small. Plugging the corresponding step size η by Theorem 6 or 7 into Theorem 1 gives the final results. The detailed proofs are in §C. Although our results in this section are with respect to TV / W_2, we extend them to the χ^2-divergence setting in §E.2.

4.1. Strongly convex and smooth potential

Proposition 10 *Suppose f is β-strongly convex and L_1-smooth. Let $\delta \in (0,1), \eta = \tilde{\mathcal{O}}\left(1/(L_1\sqrt{d})\right)$. Then Algorithm 1, with Algorithm 2 as RGO step and initialization $x_0 \sim \mu_0$, can find a random point x_T that has δ total variation distance to the distribution $\nu \propto \exp(-f(x))$ in*

$$T = \mathcal{O}\left(\frac{L_1\sqrt{d}}{\beta}\log\left(\frac{L_1\sqrt{d}}{\beta\delta}\right)\log\left(\frac{\sqrt{H_\nu(\mu_0)}}{\delta}\right)\right)$$

steps. And we can find x_T that has δ Wasserstein-2 distance to the distribution ν in

$$T = \mathcal{O}\left(\frac{L_1\sqrt{d}}{\beta}\log\left(\frac{L_1 d}{\beta\delta}\right)\log\left(\frac{1}{\delta}\sqrt{\frac{H_\nu(\mu_0)}{\beta}}\right)\right)$$

steps. Furthermore, each step accesses only $\mathcal{O}(1)$ many $f(x)$ queries in expectation.

In this most classical setting, we compare our results with others in the literature. Denote $\kappa := L_1/\beta$ as the condition number. By Lemma 20 and Chewi et al. (2022a, §A), we assume $H_\nu(\mu_0) = \mathcal{O}(d)$. Our total complexity becomes $\tilde{\mathcal{O}}(\kappa\sqrt{d})$ for both total variance and Wasserstein distance, which is better than $\tilde{\mathcal{O}}(\kappa d)$ in Chen et al. (2022, Corollary 7) also for the proximal sampler. Considering other sampling methods, our result $\tilde{\mathcal{O}}(\kappa\sqrt{d})$ surpasses most of the existing bounds, including randomized midpoint Unadjusted Langevin Monte Carlo (LMC) (He et al., 2020), ULMC (Cheng et al., 2018; Dalalyan and Riou-Durand, 2020; Ganesh and Talwar, 2020), MALA with a warm start (Chewi et al., 2021; Wu et al., 2021). In particular, Wu et al. (2021) shows $\tilde{\Omega}(\kappa\sqrt{d})$ is the lower bound for MALA to mix. Shen and Lee (2019) can achieve $\tilde{\mathcal{O}}\left(\kappa^{7/6}(\frac{2}{\delta}\sqrt{\frac{d}{\beta}})^{1/3} + \kappa(\frac{2}{\delta}\sqrt{\frac{d}{\beta}})^{2/3}\right)$ in terms of Wasserstein distance. Their bound is better in dimension dependence but depends polynomially on δ, and is therefore not a high-accuracy guarantee.

4.2. Convex and semi-smooth potential

Proposition 11 *Suppose f is convex and L_α-α-semi-smooth. Let $\delta \in (0,1), \eta = \tilde{\mathcal{O}}\left(1/(L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}})\right)$. Then we can find a random point x_T that has δ total variation distance to $\nu \propto \exp(-f(x))$ in*

$$T = \mathcal{O}\left(\frac{W_2^2(\mu_0,\nu) L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}}{\delta^2} \log\left(\frac{W_2^2(\mu_0,\nu) L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}}{\delta^3}\right)\right)$$

steps. Furthermore, each step accesses only $\mathcal{O}(1)$ many $f(x)$ queries in expectation.

Assume $W_2^2(\mu_0,\nu) = \mathcal{O}(\mathcal{M}_2)$ where $\mathcal{M}_2$ is the second moment of ν. Then our bound becomes $\tilde{\mathcal{O}}(\mathcal{M}_2 L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}/\delta^2)$. When f is smooth and $\alpha = 1$, our result $\tilde{\mathcal{O}}(\mathcal{M}_2 L_1 \sqrt{d}/\delta^2)$ improves the bound $\tilde{\mathcal{O}}(\mathcal{M}_2 L_1 d/\delta^2)$ in Chen et al. (2022, Corollary 6), which is the state-of-art (in dimension) complexity for log-concave smooth sampling. When f is Lipschitz, our result $\tilde{\mathcal{O}}(\mathcal{M}_2 L_0^2/\delta^2)$ improves the bound $\tilde{\mathcal{O}}(\mathcal{M}_2 L_0^2 d/\delta^4)$ in Lehec (2021, Theorem 1). When f is semi-smooth, we also improve the bound $\tilde{\mathcal{O}}(\sqrt{\mathcal{M}_4} L_\alpha^{\frac{2}{\alpha+1}} d/\delta)$ in Liang and Chen (2022c, Theorem 4.2) in terms of dimension. Here $\mathcal{M}_4$ is the fourth central moment of ν.

Remark 12 *An alternative approach to sample from a log-concave distribution is to first construct a regularized strongly convex potential $\hat{f}(x) := f(x) + w\|x - x^*\|^2/2$ with x^* being an (approximated) minimizer of f. Following Liang and Chen (2022b), Algorithm 2 can be modified so that the proximal sampler can sample from $\exp(-\hat{f})$ with complexity $\tilde{\mathcal{O}}(L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}/w)$. With a proper choice of w, we arrive at an algorithm to sample from $\nu \propto \exp(-f)$ with complexity $\tilde{\mathcal{O}}(\sqrt{\mathcal{M}_4} L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}/\delta)$.*

4.3. Sampling from non-log-concave distributions satisfying isoperimetric inequalities

Since (LSI) implies that the distribution has a sub-Gaussian tail, we only present the result for smooth potentials ($\alpha = 1$) when LSI is satisfied.

Proposition 13 *Suppose $\nu \propto \exp(-f)$ satisfies C_{LSI}-LSI and f is L_1-smooth. Let $\delta \in (0,1), \eta = \tilde{\mathcal{O}}\left(1/(L_1\sqrt{d})\right)$. Then we can find a random point x_T that has δ total variation distance to ν in*

$$T = \mathcal{O}\left(\frac{L_1\sqrt{d}}{C_{\text{LSI}}} \log\left(\frac{L_1\sqrt{d}}{C_{\text{LSI}}\delta}\right) \log\left(\frac{\sqrt{H_\nu(\mu_0)}}{\delta}\right)\right)$$

steps. Furthermore, each step accesses only $\mathcal{O}(1)$ many $f(x)$ queries in expectation.

We can also define a "condition number" $\hat{\kappa} = L_1/C_{\text{LSI}}$, and assume $H_\nu(\mu_0) = \mathcal{O}(d)$. Then our result becomes $\tilde{\mathcal{O}}\left(\hat{\kappa}\sqrt{d}\right)$, whereas Chen et al. (2022, Corollary 7) and Liang and Chen (2022a, Theorem 3.1) give $\tilde{\mathcal{O}}(\hat{\kappa} d)$. Our bound is also better than the order $\tilde{\mathcal{O}}(\hat{\kappa}^2 d/\delta^2)$ in Chewi et al. (2022a); Erdogdu and Hosseinzadeh (2021, Theorem 7).

Proposition 14 *Suppose $\nu \propto \exp(-f)$ satisfies C_{PI}-PI and f is L_α-α-semi-smooth. Let $\delta \in (0,1), \eta = \tilde{\mathcal{O}}\left(1/(L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}})\right)$. Then we can find a random point x_T that has δ total variation distance to ν in*

$$T = \mathcal{O}\left(\frac{L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}}{C_{\text{PI}}} \log\left(\frac{L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}}{C_{\text{PI}}\delta}\right) \log\left(\frac{\chi^2_\nu(\mu_0)}{\delta^2}\right)\right)$$

steps. Furthermore, each step accesses only $\mathcal{O}(1)$ many $f(x)$ queries in expectation.

By Lemma 20 and Chewi et al. (2022a, §A), we assume $\chi^2_\nu(\mu_0) = \mathcal{O}(\exp(d))$. Then our result becomes $\tilde{\mathcal{O}}\left(\frac{L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{2\alpha+1}{\alpha+1}}}{C_{\text{PI}}}\right)$. This improves the result $\tilde{\mathcal{O}}\left(\frac{L_\alpha^{\frac{2}{\alpha+1}} d^2}{C_{\text{PI}}}\right)$ in Liang and Chen (2022a) and $\tilde{\mathcal{O}}\left(\frac{L_\alpha^{2/\alpha} d^{2+1/\alpha}}{C_{\text{PI}}^{1+1/\alpha}\delta^{2/\alpha}}\right)$ in Chewi et al. (2022a).

5. Proximal sampling for composite potentials

In this section, we consider the composite potential $f = \sum_{j=1}^n f_j$ that satisfies

$$\|f_j'(u) - f_j'(v)\| \le L_{\alpha_j}\|u-v\|^{\alpha_j}, \quad \forall\, u, v \in \mathbb{R}^d \ , \forall\, 1 \le j \le n \tag{4}$$

with $\alpha_j \in [0,1]$ for all j. When $n = 1$, this can recover the assumption (1). When $n = 2$, it can also recover the popular "smooth+non-smooth" function assumption. In this section we extend the analysis in the previous sections for semi-smooth f to composite f.

5.1. RGO complexity

Again, to simplify the argument, we assume that the stationary point of f_y^η can be computed exactly. Otherwise, we can adopt the same argument in §D to obtain the same complexity order.

Since f is composite, we naturally let $g = \sum_{j=1}^n g_j$, where $g_j(x) := f_j(x) - \langle f_j'(x_y), x\rangle$. Clearly, x_y is the stationary point of all g_j, namely, $g_j'(x_y) = 0$. It is easy to check that Lemma 2 and 3 still hold for the composite potential. Thus, the RGO step aims to sample from $\exp(-g(x) - 1/2\eta\|x-y\|^2)$ for a fixed y. Next, for the complexity analysis, the crux is also to develop a concentration inequality for composite function g. The proof is naturally based on the probabilistic uniform bound.

Corollary 15 (Gaussian concentration inequality for composite functions) *Let $X \sim \mathcal{N}(m, \eta\mathbf{I})$ be a Gaussian random variable in $\mathbb{R}^d$, and let ℓ be a composite function satisfying (4). Assume $\ell_j'(m) = 0$ for $\forall\, 1 \le j \le n$. Then for any $r > 0$, $0 \le \alpha_j \le 1$ and $\sum_{j=1}^n w_j = 1, w_j \ge 0$, one has*

$$\Pr(\ell(X) - \mathbb{E}(\ell(X)) \ge r) \le \left(1 - \frac{\epsilon}{d}\right)^{-\frac{d}{2}} \sum_{i=j}^n \exp\left(-\frac{C_j \epsilon^{\frac{\alpha_j}{1+\alpha_j}} (w_j r)^{\frac{2}{1+\alpha_j}}}{L_{\alpha_j}^{\frac{2}{1+\alpha_j}} d^{\frac{\alpha_j}{1+\alpha_j}} \eta}\right), \quad \forall \epsilon \in (0, d),$$

where $C_j = (1+\alpha_j)\left(\frac{1}{\alpha_j}\right)^{\frac{\alpha_j}{1+\alpha_j}} \left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha_j}} 2^{\frac{1-\alpha_j}{1+\alpha_j}}$.

With this concentration inequality, in view of the fact that $g_j'(x_y) = 0$, we obtain similar RGO complexity results. Our step size in the following theorems can recover the $n = 1$ case in §3.

Theorem 16 (RGO complexity in total variation) *Assume f satisfies (4). If the step size*

$$\eta \leq \left(49 \left(\sum_{j=1}^{n} L_{\alpha_j}^{\frac{1}{\alpha_j+1}} d^{\frac{\alpha_j}{2(\alpha_j+1)}} \right)^2 (1 + \log(1 + 12n/\zeta)) \right)^{-1},$$

then for any $\zeta > 0$, Algorithm 2 returns a random point x that has ζ total variation distance to the distribution proportional to $\pi^{X|Y}(\cdot|y)$. Furthermore, if $0 < \zeta < 1$, then the algorithm access $\mathcal{O}(1)$ many $f(x)$ in expectation.

Theorem 17 (RGO complexity in Wasserstein distance) *Assume f satisfies (4). If the step size*

$$\eta \leq \min \left(\left(49 \left(\sum_{j=1}^{n} L_{\alpha_j}^{\frac{1}{\alpha_j+1}} d^{\frac{\alpha_j}{2(\alpha_j+1)}} \right)^2 (2 + \log(1 + 192n(d^2 + 2d)/\zeta^4)) \right)^{-1}, \quad 1 \right),$$

then Algorithm 2 returns a random point x that has ζ Wasserstein-2 distance to the distribution $\pi^{X|Y}(\cdot|y)$. If $0 < \zeta < 2\sqrt{2d}$, then the algorithm access $\mathcal{O}(1)$ many $f(x)$ in expectation.

5.2. Complexity bounds of proximal sampling

Next we provide the total complexity under several conditions for composite potential. As in §4, the results in this section are obtained by plugging the RGO step size (Theorem 16, 17) into proximal sampler convergence results in Theorem 1. Due to the high similarity to §4, we omit the proofs for the results in this section. Throughout this section, we denote $M_{L,d} := \left(\sum_{j=1}^{n} L_{\alpha_j}^{\frac{1}{\alpha_j+1}} d^{\frac{\alpha_j}{2(\alpha_j+1)}} \right)^2$ and assume that $\eta = \tilde{\mathcal{O}}(1/M_{L,d})$.

Proposition 18 *Suppose f is β-strongly convex and satisfies (4) and $\delta \in (0, 1)$. Then Algorithm 1, with Algorithm 2 as RGO step, can find a random point x_T that has δ total variation distance to the distribution $\nu \propto \exp(-f(x))$ in*

$$T = \mathcal{O}\left(\frac{M_{L,d}}{\beta} \log\left(\frac{M_{L,d}}{\beta\delta} \right) \log\left(\frac{\sqrt{H_\nu(\mu_0)}}{\delta} \right) \right)$$

steps. And we can find x_T that has δ Wasserstein-2 distance to the distribution ν in

$$T = \mathcal{O}\left(\frac{M_{L,d}}{\beta} \log\left(\frac{M_{L,d}\sqrt{d}}{\beta\delta} \right) \log\left(\frac{1}{\delta} \sqrt{\frac{H_\nu(\mu_0)}{\beta}} \right) \right)$$

steps. Furthermore, each step accesses only $\mathcal{O}(1)$ many $f(x)$ queries in expectation.

We again assume $H_\nu(\mu_0) = \mathcal{O}(d)$ (Chewi et al., 2022a, §A). If $f = f_1 + f_2$, where f_1 is strongly convex and L_1-smooth, and f_2 is L_0-Lipschitz, then we have $\tilde{\mathcal{O}}((L_0 + L_1^{1/2} d^{1/4})^2/\beta)$, whereas Bernton (2018) gives $\tilde{\mathcal{O}}(L_0^2 d/(\beta\delta^4))$, Salim and Richtárik (2020) gives $\tilde{\mathcal{O}}((L_0^2 + L_1 d)/(\beta^2\delta^2))$, and Liang and Chen (2022c, Theorem 5.5) gives $\tilde{\mathcal{O}}(\max(L_0^2, L_1)\, d/\beta)$.

Proposition 19 *Suppose we can find a random point x_T that has δ total variation distance to ν in T steps. Let $\delta \in (0,1)$, and $\nu \propto \exp(-f)$, where f satisfies (4).*
1) If ν is log-concave, then $T = \mathcal{O}\left(\frac{M_{L,d} W_2^2(\mu_0,\nu)}{\delta^2} \log\left(\frac{M_{L,d} W_2^2(\mu_0,\nu)}{\delta^3}\right)\right)$;
2) If ν satisfies C_{LSI}-LSI, then $T = \mathcal{O}\left(\frac{M_{L,d}}{C_{\mathrm{LSI}}} \log\left(\frac{M_{L,d}}{C_{\mathrm{LSI}}\delta}\right) \log\left(\frac{\sqrt{H_\nu(\mu_0)}}{\delta}\right)\right)$;
3) If ν satisfies C_{PI}-PI, then $T = \mathcal{O}\left(\frac{M_{L,d}}{C_{\mathrm{PI}}} \log\left(\frac{M_{L,d}}{C_{\mathrm{PI}}\delta}\right) \log\left(\frac{\chi^2_\nu(\mu_0)}{\delta^2}\right)\right)$;
Furthermore, each step accesses only $\mathcal{O}(1)$ many $f(x)$ queries in expectation.

Consider log-concave ν and the potential f is a "smooth + semi-smooth" function. Our result becomes $\tilde{\mathcal{O}}(\mathcal{M}_2(L_1^{1/2} d^{1/4} + L_\alpha^{1/(\alpha+1)} d^{\alpha/(2(\alpha+1))})^2/\delta^2)$, which improves $\tilde{\mathcal{O}}(\sqrt{\mathcal{M}_4}\max(L_1, L_\alpha^{2/(\alpha+1)})d/\delta)$ in Liang and Chen (2022c, Theorem 5.4) in terms of dimension. Our result for non-log-concave ν also improves the bounds in Nguyen et al. (2021); Liang and Chen (2022a) (see Table 2).

6. Concluding remark

We propose and analyze a novel RGO realization of the proximal sampler. The core of our analysis is a new Gaussian concentration inequality for semi-smooth functions, which is itself of independent interest. With this concentration inequality, we significantly improve the dimension dependence of RGO. We then analyze the accumulated error caused by our inexact RGO, which is combined with proximal sampler convergence results to give the total complexity. Our complexity bounds are better than almost all existing results in all the classical settings (strongly log-concave, log-concave, LSI, PI) as well as more general settings with semi-smooth potentials or composite potentials.

We leave a few directions for future study: 1) How to generalize our RGO algorithm to settings where the target potential f is an empirical risk or population risk? This should be achievable by merging Algorithm 2 and Gopi et al. (2022, Algorithm 2). This will be useful for private optimization in differential privacy. 2) Is there any other RGO algorithm that has even better dimension dependence than Algorithm 2? 3) Our proof techniques make our concentration inequality only applicable to Gaussian distributions. Is there a similar concentration inequality as in Theorem 4 but for more general distributions satisfying LSI?

Acknowledgments

We thank Daogao Liu for useful discussions. We extend our gratitude to the anonymous reviewers for their invaluable feedback that enhanced this manuscript. Financial support from NSF under grants 1942523, 2008513, and 2206576 is greatly acknowledged.

References

Jason Altschuler and Kunal Talwar. Privacy of noisy stochastic gradient descent: More iterations without more privacy loss. *Advances in Neural Information Processing Systems*, 35:3788–3800, 2022.

Jason M. Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. *arXiv*, 2023.

Espen Bernton. Langevin Monte Carlo and JKO splitting. In *Conference On Learning Theory*, pages 1777–1798. PMLR, 2018.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. *arXiv preprint arXiv:2202.06386*, 2022.

Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. In *Conference on learning theory*, pages 300–323. PMLR, 2018.

Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the metropolis-adjusted langevin algorithm. In *Conference on Learning Theory*, pages 1260–1300. PMLR, 2021.

Sinho Chewi, Murat A Erdogdu, Mufan Li, Ruoqi Shen, and Shunshi Zhang. Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev. In *Conference on Learning Theory*, pages 1–2. PMLR, 2022a.

Sinho Chewi, Patrik R Gerber, Chen Lu, Thibaut Le Gouic, and Philippe Rigollet. The query complexity of sampling from strongly log-concave distributions in one dimension. In *Conference on Learning Theory*, pages 2041–2059. PMLR, 2022b.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3): 651–676, 2017.

Arnak S Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient bayesian computation by proximal markov chain monte carlo: when langevin meets moreau. *SIAM Journal on Imaging Sciences*, 11 (1):473–506, 2018.

Alain Durmus, Szymon Majewski, and Blazej Miasojedow. Analysis of langevin monte carlo via convex optimization. *The Journal of Machine Learning Research*, 20(1):2666–2711, 2019.

Murat A Erdogdu and Rasa Hosseinzadeh. On the convergence of langevin monte carlo: the interplay between tail growth and smoothness. In *Conference on Learning Theory*, pages 1776–1822. PMLR, 2021.

Murat A Erdogdu, Rasa Hosseinzadeh, and Shunshi Zhang. Convergence of langevin monte carlo in chi-squared and rényi divergence. In *International Conference on Artificial Intelligence and Statistics*, pages 8151–8175. PMLR, 2022.

JiaoJiao Fan, Qinsheng Zhang, Amir Hossein Taghvaei, and Yongxin Chen. Variational wasserstein gradient flow. In *International Conference on Machine Learning*, 2021.

Jiaojiao Fan, Bo Yuan, Jiaming Liang, and Yongxin Chen. Nesterov smoothing for sampling without smoothness. *arXiv preprint arXiv:2208.07459*, 2022.

Arun Ganesh and Kunal Talwar. Faster differentially private samplers via rényi divergence analysis of discretized langevin mcmc. *Advances in Neural Information Processing Systems*, 33:7222–7233, 2020.

Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. Private convex optimization via exponential mechanism. In *Conference on Learning Theory*, pages 1948–1989. PMLR, 2022.

David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

Ye He, Krishnakumar Balasubramanian, and Murat A Erdogdu. On the ergodicity, bias and asymptotic normality of randomized midpoint sampling method. *Advances in Neural Information Processing Systems*, 33:7366–7376, 2020.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Michel Ledoux. Concentration of measure and logarithmic sobolev inequalities. In *Seminaire de probabilites XXXIII*, pages 120–216. Springer, 1999.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted gaussian oracle. In *Conference on Learning Theory*, pages 2993–3050. PMLR, 2021.

Joseph Lehec. The langevin monte carlo algorithm in the non-smooth log-concave case. *arXiv preprint arXiv:2101.10695*, 2021.

Jiaming Liang and Yongxin Chen. A proximal algorithm for sampling from non-convex potentials. *ArXiv*, abs/2205.10188, 2022a.

Jiaming Liang and Yongxin Chen. A proximal algorithm for sampling from non-smooth potentials. In *2022 Winter Simulation Conference (WSC)*, pages 3229–3240. IEEE, 2022b.

Jiaming Liang and Yongxin Chen. A proximal algorithm for sampling. *arXiv preprint arXiv:2202.13975*, 2022c.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

Yuan Liu. The poincaré inequality and quadratic transportation-variance inequalities. *Electronic Journal of Probability*, 25:1–16, 2020.

Wenlong Mou, Nicolas Flammarion, Martin J Wainwright, and Peter L Bartlett. An efficient sampling algorithm for non-smooth composite potentials. *Journal of Machine Learning Research*, 23(233): 1–50, 2022.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Dao Nguyen, Xin Dang, and Yixin Chen. Unadjusted langevin algorithm for non-convex weakly smooth potentials. *arXiv preprint arXiv:2101.06369*, 2021.

Tomohiro Nishiyama and Igal Sason. On relations between the relative entropy and χ^2-divergence, generalizations and applications. *Entropy*, 22(5):563, 2020.

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

Gilles Pisier. Probabilistic methods in the geometry of banach spaces. In *Probability and Analysis: Lectures given at the 1st 1985 Session of the Centro Internazionale Matematico Estivo (CIME) held at Varenna (Como), Italy May 31–June 8, 1985*, pages 167–241. Springer, 2006.

Manuel Pulido and Peter Jan van Leeuwen. Sequential monte carlo with kernel embedded mappings: The mapping particle filter. *Journal of Computational Physics*, 396:400–415, 2019.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. 2013.

Adil Salim and Peter Richtárik. Primal dual interpretation of the proximal stochastic gradient langevin algorithm. *Advances in Neural Information Processing Systems*, 33:3786–3796, 2020.

Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. *Advances in Neural Information Processing Systems*, 32, 2019.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

Cédric Villani. Topics in optimal transportation. 2003.

Farrol Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric. *The Annals of Probability*, 1(6):1068–1070, 1973.

Keru Wu, Scott Schmidler, and Yuansi Chen. Minimax mixing time of the metropolis-adjusted langevin algorithm for log-concave sampling. *arXiv preprint arXiv:2109.13055*, 2021.

Contents

Appendix A. Proof of technical lemmas

A.1. Proof of Lemma 2

Proof Since $f'(x_y) + \frac{1}{\eta}(x_y - y) = 0$, we have $x_y = y - \eta f'(x_y)$. This implies

$$\begin{aligned} g(x) + \frac{1}{2\eta}\|x - x_y\|^2 &= g(x) + \frac{1}{2\eta}\|x - y + \eta f'(x_y)\|^2 \\ &= g(x) + \frac{1}{2\eta}\|x - y\|^2 + \langle x, f'(x_y)\rangle - \langle y, f'(x_y)\rangle + \frac{\eta}{2}\|f'(x_y)\|^2 \\ &= g(x) + \frac{1}{2\eta}\|x - y\|^2 + \langle x, f'(x_y)\rangle + \text{constant} \\ &= f(x) + \frac{1}{2\eta}\|x - y\|^2 + \text{constant}. \end{aligned}$$

We use $f(x) = g(x) + \langle f'(x_y), x\rangle$ in the last equality. ■

A.2. Proof of Lemma 3

Proof Following the definition and Lemma 2, we have

$$\begin{aligned} \frac{\mathrm{d}\pi^{X|Y}}{\mathrm{d}x} &= \frac{\exp(-f(x) - \frac{1}{2\eta}\|x - y\|^2)}{\int \exp(-f(x) - \frac{1}{2\eta}\|x - y\|^2)\mathrm{d}x} = \frac{\exp(-g(x) - \frac{1}{2\eta}\|x - x_y\|^2)}{\int \exp(-g(x) - \frac{1}{2\eta}\|x - x_y\|^2)\mathrm{d}x} \\ &= \frac{\exp(-g(x))\exp(-\frac{1}{2\eta}\|x - x_y\|^2)/\int \exp(-\frac{1}{2\eta}\|x - x_y\|^2)\mathrm{d}x}{\int \exp(-g(x))(\exp(-\frac{1}{2\eta}\|x - x_y\|^2)/\int \exp(-\frac{1}{2\eta}\|x - x_y\|^2)\mathrm{d}x)\mathrm{d}x} \\ &= \frac{\exp(-\frac{1}{2\eta}\|x - x_y\|^2)}{\int \exp(-\frac{1}{2\eta}\|x - x_y\|^2)\mathrm{d}x} \cdot \frac{\exp(-g(x))}{\mathbb{E}[\exp(-g(x))]} = \frac{\mathrm{d}\phi}{\mathrm{d}x} \cdot \frac{\exp(-g(x))}{\mathbb{E}_{x\sim\phi}\exp(-g(x))}. \end{aligned}$$

Next, since

$$\begin{aligned} \mathbb{E}[\rho|x] = \mathbb{E}[\exp(g(z) - g(x))|x] = \mathbb{E}[\exp(g(z))]\exp(-g(x)), \\ \mathbb{E}[\rho] = \mathbb{E}[\exp(g(z))]\mathbb{E}[\exp(-g(x))], \end{aligned}$$

we get $\exp(-g(x))/\mathbb{E}\exp(-g(x)) = \mathbb{E}[\rho|x]/\mathbb{E}[\rho]$. Finally, since Algorithm 2 is rejection sampling,

$$\frac{\mathrm{d}\hat{\pi}^{X|Y}}{\mathrm{d}x} = \frac{\mathrm{d}\phi}{\mathrm{d}x} \cdot \frac{\Pr(u \le \frac{1}{2}\rho|x)}{\Pr(u \le \frac{1}{2}\rho)}.$$

By tower property and the fact that u follows the uniform distribution over $[0.1]$, the acceptance probability is

$$\Pr(u \le \frac{1}{2}\rho) = \mathbb{E}[\Pr(u \le \frac{1}{2}\rho|\rho)] = \frac{1}{2}\mathbb{E}[\bar{\rho}].$$

Similarly, $\Pr(u \le \frac{1}{2}\rho|x) = \frac{1}{2}\mathbb{E}[\bar{\rho}|x]$. These conclude that $\frac{\mathrm{d}\hat{\pi}^{X|Y}}{\mathrm{d}x} = \frac{\mathrm{d}\phi}{\mathrm{d}x} \cdot \frac{\mathbb{E}[\bar{\rho}|x]}{\mathbb{E}[\bar{\rho}]}$. ■

A.3. Proof of Lemma 8

Proof Indeed, by triangular inequality and Jensen inequality,

$$\begin{aligned}
&\|\hat{\mu}_t(x) - \mu_t(x)\|_{\mathrm{TV}} = \left\| \int \hat{\psi}_t(y)\hat{\pi}^{X|Y}(x|y)\mathrm{d}y - \int \psi_t(y)\pi^{X|Y}(x|y)\mathrm{d}y \right\|_{\mathrm{TV}} \\
\leq &\left\| \int \hat{\psi}_t(y) \left(\hat{\pi}^{X|Y}(x|y) - \pi^{X|Y}(x|y) \right) \mathrm{d}y \right\|_{\mathrm{TV}} + \left\| \int (\hat{\psi}_t(y) - \psi_t(y))\pi^{X|Y}(x|y)\mathrm{d}y \right\|_{\mathrm{TV}} \\
\leq &\int \hat{\psi}_t(y) \left\| \hat{\pi}^{X|Y}(x|y) - \pi^{X|Y}(x|y) \right\|_{\mathrm{TV}} \mathrm{d}y + \int \left(\left| \hat{\psi}_t(y) - \psi_t(y) \right| \int \pi^{X|Y}(x|y)\mathrm{d}x \right) \mathrm{d}y \\
\leq &\zeta + \left\| \hat{\psi}_t(y) - \psi_t(y) \right\|_{\mathrm{TV}}.
\end{aligned}$$

Moreover, denote G as the density of the distribution $\mathcal{N}(0, \eta\mathbf{I})$,

$$\left\| \hat{\psi}_t(y) - \psi_t(y) \right\|_{\mathrm{TV}} = \left\| \int (\hat{\mu}_{t-1}(x) - \mu_{t-1}(x))G(y-x)\mathrm{d}x \right\|_{\mathrm{TV}} \leq \|\hat{\mu}_{t-1} - \mu_{t-1}\|_{\mathrm{TV}}.$$

Thus, $\|\hat{\mu}_t - \mu_t\|_{\mathrm{TV}} \leq \zeta + \|\hat{\mu}_{t-1} - \mu_{t-1}\|_{\mathrm{TV}}$. Finally, because $\hat{\psi}_1 = \psi_1$, there is $\|\hat{\mu}_T - \mu_T\|_{\mathrm{TV}} \leq \zeta T$. ■

A.4. Proof of Lemma 9

Proof Denote the marginal distribution of Y in the t-th iteration of the idea proximal sampler and approximate proximal sampler by $\psi_t, \hat{\psi}_t$ respectively. It is well known the Gaussian convolution is contractive with respect to the Wasserstein-2 metric. Thus

$$W_2(\psi_t, \hat{\psi}_t) \leq W_2(\mu_{t-1}, \hat{\mu}_{t-1}). \tag{5}$$

Following the Doob's h-transform, the RGO $\pi^{X|Y}$ can be realized by simulating the backward diffusion (see (Chen et al., 2022, §A))

$$\mathrm{d}Z_s = -\nabla \log \pi_s(Z_s)\mathrm{d}t + \mathrm{d}B_s,$$

over the time interval $[0, \eta]$, where B_s is a standard Wiener process and $\pi_s = \nu * \mathcal{N}(0, s\mathbf{I})$. Let $Z_s, \hat{Z}_s$ be two copies of the above process initialized at $Z_\eta \sim \psi_t, \hat{Z}_\eta \sim \hat{\psi}_t$ respectively. We then use a coupling argument to show the RGO for log-concave distribution is also contractive with respect to W_2. In particular, let $Z_s, \hat{Z}_s$ be driven by a common Wiener process B_s and $Z_\eta, \hat{Z}_\eta$ be coupled in such a way that $\mathbb{E}\|Z_\eta - \hat{Z}_\eta\|^2 = W_2^2(\psi_t, \hat{\psi}_t)$, then

$$\frac{\mathrm{d}}{\mathrm{d}t}\|Z_s - \hat{Z}_s\|^2 = 2\langle -\nabla \log \pi_s(Z_s) + \nabla \log \pi_s(\hat{Z}_s), Z_s - \hat{Z}_s \rangle \geq 0.$$

The last inequality is due to the fact that ν and thus π_s is log-concave. It follows that

$$\begin{aligned}
&W_2^2 \left(\int \pi^{X|Y}(x|y)\psi_t(y)\mathrm{d}y, \int \pi^{X|Y}(x|y)\hat{\psi}_t(y)\mathrm{d}y \right) \\
\leq &\mathbb{E}\|Z_0 - \hat{Z}_0\|^2 \leq \mathbb{E}\|Z_\eta - \hat{Z}_\eta\|^2 = W_2^2(\psi_t, \hat{\psi}_t).
\end{aligned} \tag{6}$$

Moreover, by definition of Wasserstein distance, we have

$$W_2\left(\int \hat{\pi}^{X|Y}(x|y)\hat{\psi}_t(y)\mathrm{d}y, \int \pi^{X|Y}(x|y)\hat{\psi}_t(y)\mathrm{d}y\right) \le \sqrt{\int \hat{\psi}_t(y)W_2^2(\hat{\pi}^{X|Y}, \pi^{X|Y})\mathrm{d}y} \le \zeta. \quad (7)$$

By triangular inequality, we have

$$\begin{aligned} W_2(\hat{\mu}_t, \mu_t) &= W_2\left(\int \hat{\pi}^{X|Y}(x|y)\hat{\psi}_t(y)\mathrm{d}y, \int \pi^{X|Y}(x|y)\psi_t(y)\mathrm{d}y\right) \\ &\le W_2\left(\int \hat{\pi}^{X|Y}(x|y)\hat{\psi}_t(y)\mathrm{d}y, \int \pi^{X|Y}(x|y)\hat{\psi}_t(y)\mathrm{d}y\right) \\ &\quad + W_2\left(\int \pi^{X|Y}(x|y)\hat{\psi}_t(y)\mathrm{d}y, \int \pi^{X|Y}\psi_t(y)(x|y)\mathrm{d}y\right) \\ &\overset{(6),(7)}{\le} \zeta + W_2^2(\hat{\psi}_t, \psi_t) \overset{(5)}{\le} \zeta + W_2(\hat{\mu}_{t-1}, \mu_{t-1}). \end{aligned}$$

Finally, because $\hat{\psi}_1 = \psi_1$, there is $W_2(\hat{\mu}_T, \mu_T) \le \zeta T$. ■

A.5. Supportive lemmas

In this section, we list the lemmas pertaining to our analysis.

Lemma 20 *Under PI, Liu (2020) together with standard comparison inequalities implies that*

$$\max\left\{\frac{\|\mu-\nu\|_{\mathrm{TV}}^2}{2}, \log\left(1+\frac{C_{\mathrm{PI}}}{2}W_2^2(\mu,\nu)\right), H_\nu(\mu)\right\} \le R_{2,\nu}(\mu),$$

where $R_{2,\nu}(\mu) := \log\int(\mu^2/\nu)$ is Rényi divergence of order 2.

Lemma 21 (Proposition 7.10 in Villani (2003)) *Let μ and ν be two probability measures on $\mathbb{R}^d$. Then for any $m \in \mathbb{R}^d$,*

$$W_2^2(\mu,\nu) \le 2\int \|m-x\|_2^2 \mathrm{d}|\mu-\nu|(x) = 2\left\|\|m-\cdot\|_2^2(\mu-\nu)\right\|_{\mathrm{TV}}.$$

Appendix B. Gaussian concentration inequality

B.1. Proof of Theorem 4

Proof Without loss of generality, we assume $m = 0$. To see this, define $g(x) := \ell(x+m)$ which is also an L_α-α-semi-smooth function and satisfy $g'(0) = 0$, and notice that $\mathbb{E}(\ell(X)) = \mathbb{E}(g(Y))$ where $Y \sim \mathcal{N}(0, \eta\mathbf{I})$. It follows that

$$\begin{aligned} \Pr(\ell(X) - \mathbb{E}(\ell(X)) \ge r) &= (2\pi\eta)^{-d/2}\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2\eta}\|x-m\|_2^2\right)\mathbb{1}_{\ell(x)-\mathbb{E}(\ell(x))\ge r}\mathrm{d}x \\ &= (2\pi\eta)^{-d/2}\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2\eta}\|y\|_2^2\right)\mathbb{1}_{g(y)-\mathbb{E}(g(y))\ge r}\mathrm{d}y \\ &= \Pr(g(Y) - \mathbb{E}(g(Y)) \ge r). \end{aligned}$$

Hence in what follows, $m = 0$. The following proof is based on the elegant argument of Maurey and Pisier (Pisier, 2006, Theorem 2.1). Let G and H be two independent Gaussian variables following $\mathcal{N}(0, \eta\mathbf{I})$, then for any $\lambda > 0$, by independence and Jensen's inequality, one has

$$\begin{aligned}\mathbb{E}\exp(\lambda(\ell(G) - \ell(H))) &= \mathbb{E}\exp(\lambda(\ell(G) - \mathbb{E}\ell(G)))\mathbb{E}\exp(\lambda(\mathbb{E}\ell(H) - \ell(H))) \\ &\geq \mathbb{E}\exp(\lambda(\ell(G) - \mathbb{E}\ell(G)))\exp\mathbb{E}(\lambda(\mathbb{E}\ell(H) - \ell(H))) \\ &= \mathbb{E}\exp(\lambda(\ell(G) - \mathbb{E}\ell(G))).\end{aligned}$$

Let $G_\theta \coloneqq G\sin\theta + H\cos\theta$, $\theta \in [0, \pi/2]$. Using the fundamental theorem of calculus along θ, one obtains

$$\ell(G) - \ell(H) = \int_0^{\frac{\pi}{2}} \nabla\ell(G_\theta)\cdot(G\cos\theta - H\sin\theta)\mathrm{d}\theta.$$

It follows that

$$\begin{aligned}\mathbb{E}\exp(\lambda(\ell(G) - \ell(H))) &= \mathbb{E}\exp\left(\lambda\int_0^{\frac{\pi}{2}} \nabla\ell(G_\theta)\cdot(G\cos\theta - H\sin\theta)\mathrm{d}\theta\right) \\ &\leq \frac{2}{\pi}\mathbb{E}\int_0^{\frac{\pi}{2}} \exp\left(\frac{\pi}{2}\lambda\nabla\ell(G_\theta)\cdot(G\cos\theta - H\sin\theta)\right)\mathrm{d}\theta \\ &= \frac{2}{\pi}\int_0^{\frac{\pi}{2}} \mathbb{E}\exp\left(\frac{\pi}{2}\lambda\nabla\ell(G_\theta)\cdot(G\cos\theta - H\sin\theta)\right)\mathrm{d}\theta.\end{aligned}$$

Since G_θ and $G\cos\theta - H\sin\theta$ are two independent Gaussian variables, $\nabla\ell(G_\theta)\cdot(G\cos\theta - H\sin\theta)$ is equidistributed as $Z\|\nabla\ell(G_\theta)\|_2$ where $Z \sim \mathcal{N}(0, \eta)$ for any θ. This implies

$$\begin{aligned}\mathbb{E}\exp\left(\frac{\pi}{2}\lambda\nabla\ell(G_\theta)\cdot(G\cos\theta - H\sin\theta)\right) &= \mathbb{E}\exp\left(\frac{\pi}{2}\lambda Z\|\nabla\ell(G_\theta)\|_2\right) \\ &= \mathbb{E}_{G_\theta}\mathbb{E}_Z\left(\exp\left(\frac{\pi}{2}\lambda Z\|\nabla\ell(G_\theta)\|_2\right) \mid G_\theta\right) \\ &= \mathbb{E}_{G_\theta}\exp\left(\frac{\pi^2}{8}\eta\lambda^2\|\nabla\ell(G_\theta)\|_2^2\right).\end{aligned}$$

As the distribution of G_θ does not depend on θ, in the rest of the proof, we drop θ. Then,

$$\mathbb{E}\exp(\lambda(\ell(G) - \mathbb{E}\ell(G))) \leq \mathbb{E}\exp(\lambda(\ell(G) - \ell(H))) \leq \mathbb{E}_G\exp\left(\frac{\pi^2}{8}\eta\lambda^2\|\nabla\ell(G)\|_2^2\right). \quad (8)$$

Combining with Markov inequality yields

$$\begin{aligned}\Pr(\ell(X) - \mathbb{E}(\ell(X)) \geq r) &\leq \inf_{\lambda>0}\frac{\mathbb{E}\exp(\lambda(\ell(X) - \mathbb{E}(\ell(X))))}{\exp(\lambda r)} \\ &\leq \inf_{\lambda>0}\frac{\mathbb{E}_G\exp(\frac{\pi^2}{8}\eta\lambda^2\|\nabla\ell(G)\|_2^2)}{\exp(\lambda r)}.\end{aligned}$$

To study the properties of the optimization problem $\inf_{\lambda>0}\frac{\mathbb{E}_G\exp(\frac{\pi^2}{8}\eta\lambda^2\|\nabla\ell(G)\|_2^2)}{\exp(\lambda r)}$, we consider three cases based on different α.

1. $\alpha = 0$

   In this case, ℓ is an L_0-Lipschitz function, which means $\|\nabla \ell(G)\|_2^2 \leq L_0^2$. Hence, our result coincides with the classical Gaussian concentration inequality for Lipschitz functions. Taking $\lambda = \frac{4r}{\pi^2 L_0^2 \eta}$ yields

$$\Pr(\ell(X) - \mathbb{E}(\ell(X)) \geq r) \leq \exp(-\frac{2}{\pi^2}\frac{r^2}{L_0^2\eta}). \tag{9}$$

2. $\alpha = 1$

   In this case, one could simplify the optimization function in an explicit way. Assume $-\frac{1}{2\eta} + \frac{\pi^2}{8}L_1^2\eta\lambda^2 < 0$ (the condition that ensures $\mathbb{E}_G$ is finite), then

$$\begin{aligned}\frac{\mathbb{E}_G \exp(\frac{\pi^2}{8}\eta\lambda^2\|\nabla\ell(G)\|_2^2)}{\exp(\lambda r)} &\leq \frac{\mathbb{E}_G \exp(\frac{\pi^2}{8}L_1^2\eta\lambda^2\|G\|_2^2)}{\exp(\lambda r)} \\ &= \left(\frac{1}{1-\frac{\pi^2}{4}L_1^2\eta^2\lambda^2}\right)^{d/2}\exp(-\lambda r).\end{aligned}$$

   Let $\lambda = \frac{k}{L_1^2 d\eta^2}$ where k is a positive parameter. Here k is chosen such that the condition $-\frac{1}{2\eta} + \frac{\pi^2}{8}L_1^2\eta\lambda^2 < 0$ holds. Then,

$$\Pr(\ell(X) - \mathbb{E}(\ell(X)) \geq r) \leq (1 - \frac{\pi^2 k^2}{4L_1^2 d^2\eta^2})^{-\frac{d}{2}}\exp(-\frac{kr}{L_1^2 d\eta^2}).$$

   In the last step, let $\frac{\pi^2 k^2}{4L_1^2 d\eta^2} = \epsilon$, leading to

$$\Pr(\ell(X) - \mathbb{E}(\ell(X)) \geq r) \leq (1-\frac{\epsilon}{d})^{-\frac{d}{2}}\exp(-\sqrt{\frac{4\epsilon}{\pi^2}}\frac{r}{L_1 d^{1/2}\eta}),\ \ \forall \epsilon \in (0,d). \tag{10}$$

   Note that given the value of ϵ, $(1-\frac{\epsilon}{d})^{-\frac{d}{2}}$ is bounded for d.

3. $0 < \alpha < 1$

   By Young's inequality, for any $\omega > 0$, one obtains $\|G\|_2^{2\alpha} \leq \alpha\|G\|_2^2/\omega + (1-\alpha)\omega^{\frac{\alpha}{1-\alpha}}$. Hence, with the assumption $1 - \frac{\pi^2}{4}L_\alpha^2\eta^2\lambda^2\frac{\alpha}{\omega} > 0$,

$$\begin{aligned}\frac{\mathbb{E}_G \exp(\frac{\pi^2}{8}\eta\lambda^2\|\nabla\ell(G)\|_2^2)}{\exp(\lambda r)} &\leq \frac{\mathbb{E}_G \exp(\frac{\pi^2}{8}L_\alpha^2\eta\lambda^2\|G\|_2^{2\alpha})}{\exp(\lambda r)} \\ &\leq \frac{\mathbb{E}_G \exp\left(\frac{\pi^2}{8}L_\alpha^2\eta\lambda^2(\alpha\|G\|^2/\omega + (1-\alpha)\omega^{\frac{\alpha}{1-\alpha}})\right)}{\exp(\lambda r)} \\ &= (1-\frac{\pi^2}{4}L_\alpha^2\eta^2\lambda^2\frac{\alpha}{\omega})^{-\frac{d}{2}}\exp(\frac{\pi^2}{8}L_\alpha^2\eta\lambda^2(1-\alpha)\omega^{\frac{\alpha}{1-\alpha}})\exp(-\lambda r).\end{aligned} \tag{11}$$

Denote $F(\lambda,\omega) \coloneqq (1-\frac{\pi^2}{4}L_\alpha^2\eta^2\lambda^2\frac{\alpha}{\omega})^{-\frac{d}{2}}\exp(\frac{\pi^2}{8}L_\alpha^2\eta\lambda^2(1-\alpha)\omega^{\frac{\alpha}{1-\alpha}})\exp(-\lambda r)$. Since the exact optimal solution of $\inf_{\lambda,\omega}F(\lambda,\omega)$ is not a closed-form expression, one can seek for the suboptimal values instead.

To this end, let

$$\hat{\lambda} = \frac{k\, r^{\frac{1-\alpha}{1+\alpha}}}{L_\alpha^2 d^\alpha \eta^{\alpha+1}},\ \ \hat{\omega} = c\, r^{\frac{2(1-\alpha)}{1+\alpha}}\eta^{1-\alpha}d^{1-\alpha} \quad \text{for some } k>0,\ \ c>0$$

It follows that

$$F(\hat{\lambda},\hat{\omega}) = (1-\frac{\pi^2\alpha k^2}{4cL_\alpha^2 d^{\alpha+1}\eta^{\alpha+1}})^{-d/2}\exp\left(-(k-\frac{\pi^2}{8}k^2(1-\alpha)c^{\frac{\alpha}{1-\alpha}})\frac{r^{\frac{2}{\alpha+1}}}{L_\alpha^2 d^\alpha\eta^{\alpha+1}}\right).$$

Similarly, let $\frac{\pi^2\alpha k^2}{4cL_\alpha^2 d^\alpha\eta^{\alpha+1}} = \epsilon \in (0,d)$, and plugging $k = \sqrt{\frac{4c\epsilon\eta^{\alpha+1}d^\alpha L_\alpha^2}{\pi^2\alpha}}$ into $F(\hat{\lambda},\hat{\omega})$ yields

$$F(\hat{\lambda},\hat{\omega}) = (1-\frac{\epsilon}{d})^{-d/2}\exp\left(-D\frac{r^{\frac{2}{\alpha+1}}}{L_\alpha^2 d^\alpha\eta^{\alpha+1}}\right)$$

with

$$D = \sqrt{\frac{4\epsilon L_\alpha^2 d^\alpha\eta^{\alpha+1}}{\pi^2\alpha}}c^{1/2} - \frac{(1-\alpha)\epsilon L_\alpha^2 d^\alpha\eta^{\alpha+1}}{2\alpha}c^{\frac{1}{1-\alpha}}.$$

Directly optimizing D as a function of c gives

$$\max D = (1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}}2^{\frac{1-\alpha}{1+\alpha}}\epsilon^{\frac{\alpha}{\alpha+1}}L_\alpha^{\frac{2\alpha}{\alpha+1}}d^{\frac{\alpha^2}{\alpha+1}}\eta^\alpha.$$

This implies

$$\Pr(f(X)-\mathbb{E}(f(X))\geq r) \leq (1-\frac{\epsilon}{d})^{-d/2}\exp\left(-\frac{C\epsilon^{\frac{\alpha}{\alpha+1}}r^{\frac{2}{\alpha+1}}}{L_\alpha^{\frac{2}{\alpha+1}}d^{\frac{\alpha}{\alpha+1}}\eta}\right),\ \forall\epsilon\in(0,d), \tag{12}$$

with C being $(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}}2^{\frac{1-\alpha}{1+\alpha}}$.

It is noteworthy that (12) can recover the results for both smooth functions and Lipschitz functions. More precisely, plugging $\alpha=1$ into (12) leads to (10). Moreover, if $\alpha=0$, by choosing $\epsilon\to 0$, (12) coincides with (9).

■

B.2. Proof of Corollary 15

Proof By Theorem 4 and a probabilistic uniform bound, we have for any $\sum_{j=1}^n w_j = 1,\ w_j\geq 0$

$$\Pr(\ell(X)-\mathbb{E}(\ell(X))\geq r) \leq \left(1-\frac{\epsilon}{d}\right)^{-\frac{d}{2}}\sum_{j=1}^n\exp\left(-\frac{C_j\epsilon^{\frac{\alpha_j}{1+\alpha_j}}(w_j r)^{\frac{2}{1+\alpha_j}}}{L_{\alpha_j}^{\frac{2}{1+\alpha_j}}d^{\frac{\alpha_j}{1+\alpha_j}}\eta}\right),\ \ \forall\epsilon\in(0,d).$$

■

B.3. Relations to existing concentration inequalities

We now compare Theorem 4 with other existing concentration inequalities. In this section, we assume $X \sim \mathcal{N}(0, \eta\mathbf{I})$.

Lipschitz function $\alpha = 0$. Taking $\epsilon \to 0$, our concentration can be simplified to

$$\Pr(\ell(X) - \mathbb{E}(\ell(X)) \geq r) \leq \exp\left(-\frac{2r^2}{\pi^2 L_0^2 \eta}\right).$$

Obviously, it has a sub-optimal coefficient because based on the entropy argument (Boucheron et al., 2013), one can obtain

$$\Pr(\ell(X) - \mathbb{E}(\ell(X)) \geq r) \leq \exp\left(-\frac{r^2}{2L_0^2 \eta}\right).$$

Smooth function $\alpha = 1$. Taking $\epsilon = 0.5$, our concentration bound is simplified to

$$\Pr(\ell(X) - \mathbb{E}(\ell(X)) \geq r) \leq 1.5 \exp\left(-\frac{0.45r}{L_1\sqrt{d}\eta}\right). \tag{13}$$

Suppose $Y = \sum_{i=1}^d Y_i^2$, $Y_i \sim \mathcal{N}(0, \mathbf{I})$, then $Y \sim \chi^2(d)$ follows the chi-squared distribution, and $\mathbb{E}[Y] = d$. Thus, taking $\ell(X) = \|X\|^2$, the **Laurent-Massart bound** (Laurent and Massart, 2000) can be equivalently written as

$$\Pr(\|X\|^2 - \mathbb{E}(\|X\|^2) \geq 2\sqrt{\mathbb{E}(\ell(X))r} + 2r) \leq \exp(-r).$$

This further implies, if $X \sim \mathcal{N}(0, \eta\mathbf{I})$, it holds that

$$\Pr(\|X\|^2 - \mathbb{E}(\|X\|^2) \geq r) \leq \exp\left(-\frac{d + r/\eta - \sqrt{d(d + 2r/\eta)}}{2}\right), \tag{14}$$

Take $L_1 = 2, \eta = 1$, we compare (13) and (14) in terms of d. Since $(d+r-\sqrt{d(d+2r)})\sqrt{d}/r \to 0$ for a fixed r, our bound is even tighter on dimension when d is large enough.

On the other hand, we compare with **Hanson-Wright inequality** (Rudelson and Vershynin, 2013): Let A be an $d \times d$ matrix,

$$\Pr(X^\top AX - \mathbb{E}(X^\top AX) > r) \leq \exp\left(-c\min\left(\frac{4r^2}{\eta^2\|A\|_{\text{HS}}^2}, \frac{2r}{\eta\|A\|}\right)\right). \tag{15}$$

Take $\ell(X) = X^\top AX$, and suppose A is positive semi-definite, then $L_1 = \lambda_{max}(2A) = 2\|A\|$, thus our bound becomes

$$\Pr(X^\top AX - \mathbb{E}(X^\top AX) \geq r) \leq 1.5 \exp\left(-\frac{0.225r}{\|A\|\sqrt{d}\eta}\right).$$

When r is large enough, $2r/\eta$ would dominate, and Hanson-Wright is tighter than our bound in terms of d. When r is in a small range, $r^2/d\eta^2$ dominates, and our bound is tighter on dimension.

B.4. Low range concentration inequality has sub-Gaussian tail

In this part, we show that no matter what $0 \leq \alpha \leq 1$ is, we can always get a sub-Gaussian concentration when r is in a low range.

Proposition 22 *Let $X \sim \mathcal{N}(m, \eta \mathbf{I})$ be a Gaussian random variable in $\mathbb{R}^d$, and let ℓ be an L_α-α-semi-smooth function. Assume $\ell'(m) = 0$. Then for any $0 \leq \alpha \leq 1$,*

$$0 < r \leq \frac{\pi L_\alpha d^{\frac{1+\alpha}{2}} \eta^{\frac{1+\alpha}{2}}}{\sqrt{\alpha 2^\alpha}}, \tag{16}$$

we have

$$\Pr(\ell(X) - \mathbb{E}(\ell(X)) \geq r) \leq \exp\left(-\frac{r^2}{\pi^2 L_\alpha^2 d^\alpha \eta^{1+\alpha}}\right). \tag{17}$$

Proof It suffices to prove the case $\eta = 1$ because we can define $\hat{\ell}(X) = \ell(\sqrt{\eta} X)$ and correspondingly $\hat{L}_\alpha = L_\alpha \eta^{(1+\alpha)/2}$. Similar to the proof of Theorem 4, we use Maurey and Pisier argument and apply Young's inequality, and obtain

$$\Pr(\ell(X) - \mathbb{E}(\ell(X)) \geq r) \leq \inf_{\lambda > 0,\ \omega > 0} F(\lambda, \omega) \tag{18}$$

$$\text{where } F(\lambda, \omega) := \left(1 - \frac{\pi^2}{4} L_\alpha^2 \lambda^2 \frac{\alpha}{\omega}\right)^{-\frac{d}{2}} \exp\left(\frac{\pi^2}{8} L_\alpha^2 \lambda^2 (1-\alpha) \omega^{\frac{\alpha}{1-\alpha}}\right) \exp(-\lambda r).$$

We define $H = \frac{\pi^2}{8} \lambda^2 L_\alpha^2$. Firstly, we assume that (will prove in the end)

$$\frac{2H\alpha}{\hat{\omega}} \in [0, 0.5], \tag{19}$$

which suffices to conclude

$$\left(1 - \frac{\pi^2}{4} L_\alpha^2 \lambda^2 \frac{\alpha}{\hat{\omega}}\right)^{-\frac{d}{2}} = \left(1 - \frac{2H\alpha}{\hat{\omega}}\right)^{-\frac{d}{2}} \leq \exp\left(\frac{2H\alpha d}{\hat{\omega}}\right).$$

Plugging this inequality, and the value $\hat{\omega} = (2d)^{1-\alpha}$ into $F(\lambda, \omega)$, we obtain

$$\begin{aligned} F(\lambda, \hat{\omega}) \leq & \exp\left(\frac{2H\alpha d}{(2d)^{1-\alpha}} + H(1-\alpha)(2d)^\alpha - \lambda r\right) \\ \leq & \exp(2Hd^\alpha - \lambda r) = \exp\left(\frac{\pi^2}{4} \lambda^2 L_\alpha^2 d^\alpha - \lambda r\right). \end{aligned}$$

We then take $\hat{\lambda} = \frac{2r}{\pi^2 L_\alpha^2 d^\alpha}$, and obtain $F(\hat{\lambda}, \hat{\omega}) = \exp\left(-\frac{r^2}{\pi^2 L_\alpha^2 d^\alpha}\right)$. This together with (18) gives (17). Finally, to make (19) hold, we need to insure

$$\frac{2H\alpha}{\hat{\omega}} = \frac{\pi^2 \lambda^2 L_\alpha^2 \alpha}{4\hat{\omega}} = \frac{r^2 \alpha}{\pi^2 L_\alpha^2 2^{1-\alpha} d^{1+\alpha}} \leq 0.5,$$

which is equivalent to the constraint (16). ■

Proposition 22 can recover the Hanson-Wright inequality (15) in the low range. If we take $A = \mathbf{I}$, then (15) becomes

$$\Pr(\|X\|^2 - \mathbb{E}(\|X\|^2) \geq r) \leq \exp\left(-c\min\left(\frac{4r^2}{d\eta^2}, \frac{2r}{\eta}\right)\right).$$

This implies that, when $r \leq \Theta(d\eta)$,

$$\Pr(\|X\|^2 - \mathbb{E}(\|X\|^2) \geq r) \leq \exp\left(-\frac{cr^2}{d\eta^2}\right).$$

It is easy to check that Proposition 22 shows exactly the same result when $\alpha = 1$.

Appendix C. Proof of main results

C.1. Proof of Theorem 6

Proof Recall that the return of Algorithm 2 follows the distribution $\hat{\pi}^{X|Y}$, and the target distribution of RGO is $\pi^{X|Y} \propto \exp(-g(x) - \frac{1}{2\eta}\|x - x_y\|^2)$. According to Lemma 3, we have

$$\begin{aligned}\|\pi^{X|Y} - \hat{\pi}^{X|Y}\|_{\mathrm{TV}} &= \mathbb{E}\left|\frac{V}{\mathbb{E}V} - \frac{\overline{V}}{\mathbb{E}\overline{V}}\right| \leq \mathbb{E}\left|\frac{V}{\mathbb{E}V} - \frac{\overline{V}}{\mathbb{E}V}\right| + \mathbb{E}\left|\frac{\overline{V}}{\mathbb{E}V} - \frac{\overline{V}}{\mathbb{E}\overline{V}}\right| \\ &\leq \frac{\mathbb{E}|V - \overline{V}|}{|\mathbb{E}V|} + \frac{\mathbb{E}[\overline{V}]|\mathbb{E}[V - \overline{V}]|}{|\mathbb{E}V||\mathbb{E}\overline{V}|} \leq \frac{2\mathbb{E}|V - \overline{V}|}{|\mathbb{E}V|}.\end{aligned}$$

Then by Cauchy-Schwartz inequality on the Hilbert space $L^2(\phi)$, we have

$$\begin{aligned}&|\mathbb{E}V| = \mathbb{E}V = \mathbb{E}_{x\sim\phi}\exp(-g(x))\mathbb{E}_{x\sim\phi}\exp(g(x)) = \|\exp(-g(x)/2)\|^2_{L^2(\phi)}\|\exp(g(x)/2)\|^2_{L^2(\phi)} \\ \geq&\langle\exp(-g(x)/2), \exp(g(x)/2)\rangle^2 = [\mathbb{E}_{x\sim\phi}(\exp(-g(x)/2)\exp(g(x)/2))]^2 = 1. \end{aligned} \tag{20}$$

Since $\rho = \exp(g(z) - g(x))$ is always non-negative, there is

$$\mathbb{E}|V - \overline{V}| = \mathbb{E}|\mathbb{E}[\rho|x] - \mathbb{E}[\bar{\rho}|x]| = \mathbb{E}[(\rho - 2)\mathbf{1}_{\rho\geq 2}] \leq \mathbb{E}[\rho\mathbf{1}_{\rho\geq 2}].$$

Denote $\Delta = g(z) - g(x)$ and $\bar{\Delta} = \Delta/\log 2$, we have

$$\begin{aligned}\mathbb{E}[\rho\mathbf{1}_{\rho\geq 2}] &= \mathbb{E}[\exp(\Delta)\mathbf{1}_{\rho\geq 2}] = \mathbb{E}[\exp(\Delta)\mathbf{1}_{\Delta\geq\log 2}] = \mathbb{E}[\exp(\bar{\Delta}\log 2)\mathbf{1}_{\bar{\Delta}\geq 1}] \\ &\leq \sum_{i=1}^{\infty}\exp((i+1)\log 2)\Pr(\bar{\Delta}\geq i).\end{aligned}$$

Note that $g(x) = f(x) - \langle f'(x_y), x\rangle$ satisfies that $g'(x_y) = 0$, and $\mathbb{E}[x] = x_y$. Moreover, $g(x)$ is also L_α-α-semi-smooth because $g'(x_1) - g'(x_2) = f'(x_1) - f'(x_2)$ for any x_1, x_2. We also have the inequality $(1 - 0.5/a)^{-a/2} \leq 1.5$ when $a \geq 1$. We plug $\epsilon = 0.5$ into Theorem 4, and obtain that, for $\forall r > 0$,

$$\Pr[g(x) - \mathbb{E}g(x) \geq r] \leq \frac{3}{2}\exp\left(-\frac{(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}}2^{\frac{1-2\alpha}{1+\alpha}}r^{\frac{2}{1+\alpha}}}{L_\alpha^{\frac{2}{\alpha+1}}d^{\frac{\alpha}{\alpha+1}}\eta}\right). \tag{21}$$

This further implies that

$$\begin{aligned}\Pr[g(z)-g(x)\geq r] &= \Pr[g(z)-\mathbb{E}g(z)+\mathbb{E}g(x)-g(x)\geq r]\\ &\leq \Pr\left[g(z)-\mathbb{E}g(z)\geq \frac{r}{2}\right]+\Pr\left[\mathbb{E}g(x)-g(x)\geq \frac{r}{2}\right]\\ &= \Pr\left[g(z)-\mathbb{E}g(z)\geq \frac{r}{2}\right]+\Pr\left[-g(x)-\mathbb{E}[-g(x)]\geq \frac{r}{2}\right]\\ &\leq 3\exp\left(-\frac{(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}}2^{\frac{1-2\alpha}{1+\alpha}}\left(\frac{r}{2}\right)^{\frac{2}{1+\alpha}}}{L_\alpha^{\frac{2}{\alpha+1}}d^{\frac{\alpha}{\alpha+1}}\eta}\right).\end{aligned}$$

Thus $\Pr[\bar{\Delta}\geq i]\leq 3\exp\left(-\frac{(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}}2^{\frac{1-2\alpha}{1+\alpha}}\left(\frac{i\log 2}{2}\right)^{\frac{2}{1+\alpha}}}{L_\alpha^{\frac{2}{\alpha+1}}d^{\frac{\alpha}{\alpha+1}}\eta}\right)$. Denote

$$C_\eta := \frac{(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}}2^{\frac{-1-2\alpha}{1+\alpha}}(\log 2)^{\frac{2}{1+\alpha}}}{L_\alpha^{\frac{2}{\alpha+1}}d^{\frac{\alpha}{\alpha+1}}\eta}-1.$$

Since $\log 2<1$ and $i^{\frac{2}{1+\alpha}}\geq i$ for any $i\geq 1$, $0\leq\alpha\leq 1$, we have that

$$\sum_{i=1}^{\infty}\exp((i+1)\log 2)\Pr(\bar{\Delta}\geq i)\leq 6\sum_{i=1}^{\infty}\exp\left(i-\frac{(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}}2^{\frac{1-2\alpha}{1+\alpha}}\left(\frac{i\log 2}{2}\right)^{\frac{2}{1+\alpha}}}{L_\alpha^{\frac{2}{\alpha+1}}d^{\frac{\alpha}{\alpha+1}}\eta}\right)$$

$$\leq 6\sum_{i=1}^{\infty}\exp\left(-C_\eta i\right)=\frac{6}{\exp(C_\eta)-1}\leq\frac{\zeta}{2} \tag{22}$$

by the choice

$$\eta\leq\frac{(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}}2^{\frac{-1-2\alpha}{1+\alpha}}(\log 2)^{\frac{2}{1+\alpha}}}{L_\alpha^{\frac{2}{\alpha+1}}d^{\frac{\alpha}{\alpha+1}}(1+\log(1+12/\zeta))}.$$

Since

$$0.024<(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}}2^{\frac{-1-2\alpha}{1+\alpha}}(\log 2)^{\frac{2}{1+\alpha}}<0.16 \tag{23}$$

when $0\leq\alpha\leq 1$, the bound

$$\eta\leq\frac{1}{49L_\alpha^{\frac{2}{\alpha+1}}d^{\frac{\alpha}{\alpha+1}}(1+\log(1+12/\zeta))}$$

suffices. When $\alpha=0$, it can recover the bound in Gopi et al. (2022).

Finally, by Lemma 3, the acceptance probability is $\frac{1}{2}\mathbb{E}[\overline{V}]$. Since $\zeta<1$, we have

$$\mathbb{E}[\overline{V}]\geq\mathbb{E}[V]-\mathbb{E}|V-\overline{V}|\geq 1-\frac{\zeta}{2}\geq\frac{1}{2}. \tag{24}$$

Thus, the expected number of the iterations for the rejection sampling step is $\mathcal{O}(1)$. ■

C.2. Proof of Theorem 7

Proof It suffices to prove the claim: *If the step size*

$$\eta \leq \min\left(\frac{1}{49L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}(2+\log(1+\frac{192(d^2+2d)}{\zeta^2}))}, 1\right), \tag{25}$$

*then Algorithm 2 returns a random point x that has ζ **squared** Wasserstein-2 distance to the distribution $\pi^{X|Y}$. Furthermore, if $0 < \zeta < 8d$, then the algorithm access $\mathcal{O}(1)$ many $f(x)$ in expectation.*

Recall that the return of Algorithm 2 follows the distribution $\hat{\pi}^{X|Y}$, and the target distribution of RGO is $\pi^{X|Y} \propto \exp(-g(x) - \frac{1}{2\eta}\|x - x_y\|^2)$. By Lemma 21 (Villani (2003, Proposition 7.10)) and triangular inequality,

$$\begin{aligned}
W_2^2(\pi^{X|Y}, \hat{\pi}^{X|Y}) &\leq 2\left\|\|\cdot - x_y\|_2^2(\pi^{X|Y} - \hat{\pi}^{X|Y})\right\|_{\text{TV}} = 2\mathbb{E}\left(\|x - x_y\|_2^2\left|\frac{V}{\mathbb{E}V} - \frac{\overline{V}}{\mathbb{E}\overline{V}}\right|\right) \\
&\leq 2\mathbb{E}\|x - x_y\|_2^2\left(\left|\frac{V}{\mathbb{E}V} - \frac{\overline{V}}{\mathbb{E}V}\right| + \left|\frac{\overline{V}}{\mathbb{E}V} - \frac{\overline{V}}{\mathbb{E}\overline{V}}\right|\right) \\
&\leq \frac{2\mathbb{E}[\|x - x_y\|^2|V - \overline{V}|]}{\mathbb{E}V} + \frac{2\mathbb{E}[\|x - x_y\|^2\overline{V}|\mathbb{E}[V - \overline{V}]|]}{\mathbb{E}V\mathbb{E}\overline{V}} \\
&\overset{(20)}{\leq} 2\mathbb{E}[\|x - x_y\|^2|V - \overline{V}|] + \frac{2\mathbb{E}|V - \overline{V}| \cdot \mathbb{E}[\|x - x_y\|^2\overline{V}]}{\mathbb{E}\overline{V}}.
\end{aligned} \tag{26}$$

Firstly, by Cauchy-Schwartz inequality,

$$\mathbb{E}[\|x - x_y\|^2|V - \overline{V}|] \leq \left(\mathbb{E}\|x - x_y\|^4\mathbb{E}|V - \overline{V}|^2\right)^{\frac{1}{2}}. \tag{27}$$

Denote $G \sim \mathcal{N}(0, \mathbf{I})$. Since $x \sim \mathcal{N}(x_y, \eta\mathbf{I})$ and the constraint $\eta \leq 1$, we have

$$\begin{aligned}
\mathbb{E}\|x - x_y\|^4 &= \eta^2\mathbb{E}\|G\|^4 = \eta^2\left(\sum_i \mathbb{E}[G_i^4] + \sum_{i\neq j}[\mathbb{E}\, G_i^2\,][\mathbb{E}\, G_j^2]\right) \\
&= \eta^2(3d + d^2 - d) = \eta^2(d^2 + 2d) \leq d^2 + 2d.
\end{aligned} \tag{28}$$

On the other hand, following the same logic of bounding $\mathbb{E}[V - \overline{V}]$ in Theorem 6, we have

$$\begin{aligned}
\mathbb{E}|V - \overline{V}|^2 &= \mathbb{E}|\mathbb{E}[\rho - \bar{\rho}|x]|^2 \leq \mathbb{E}\mathbb{E}[|\rho - \bar{\rho}|^2|x] \leq \mathbb{E}|\rho - \bar{\rho}|^2 \leq \mathbb{E}[\rho^2\mathbf{1}_{\rho\geq 2}] \\
&\leq \sum_{i=1}^{\infty}\exp(2(i+1)\log 2)\Pr(\bar{\Delta} \geq i).
\end{aligned} \tag{29}$$

Denote

$$C_\eta := \frac{(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}} 2^{\frac{-1-2\alpha}{1+\alpha}}(\log 2)^{\frac{2}{1+\alpha}}}{L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}\eta} - 2.$$

Since the bound (23), we can verify that the assumption (25) satisfies

$$\eta \leq \frac{(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}}2^{\frac{-1-2\alpha}{1+\alpha}}(\log 2)^{\frac{2}{1+\alpha}}}{L_\alpha^{\frac{2}{\alpha+1}}d^{\frac{\alpha}{\alpha+1}}(2+\log(1+\frac{192(d^2+2d)}{\zeta^2}))}, \tag{30}$$

By the concentration inequality (21), following the same proof of bounding (22), we have

$$\begin{aligned}
&\sum_{i=1}^{\infty}\exp(2(i+1)\log 2)\Pr(\bar{\Delta}\geq i)\\
\overset{(21)}{\leq}&12\sum_{i=1}^{\infty}\exp\left(2i-\frac{(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}}2^{\frac{1-2\alpha}{1+\alpha}}\left(\frac{i\log 2}{2}\right)^{\frac{2}{1+\alpha}}}{L_\alpha^{\frac{2}{\alpha+1}}d^{\frac{\alpha}{\alpha+1}}\eta}\right)\\
\leq&12\sum_{i=1}^{\infty}\exp\left(-C_\eta i\right)=\frac{12}{\exp(C_\eta)-1}\overset{(30)}{\leq}\frac{\zeta^2}{16(d^2+2d)}.
\end{aligned}$$

This, together with the inequalities (27) (28), gives that

$$\mathbb{E}[\|x-x_y\|^2|V-\overline{V}|]\leq\frac{\zeta}{4}. \tag{31}$$

Now we bound the second term in (26). By the assumption (25), we have

$$\eta\leq\frac{1}{49L_\alpha^{\frac{2}{\alpha+1}}d^{\frac{\alpha}{\alpha+1}}(1+\log(1+96d/\zeta))}.$$

Thus, following the same steps in bounding (22) and (24), we obtain

$$\mathbb{E}|V-\overline{V}|\leq\frac{\zeta}{16d}\ \text{ and }\ \mathbb{E}[\overline{V}]\geq\frac{1}{2}. \tag{32}$$

Next, since $0\leq\overline{V}:=\mathbb{E}[\bar{\rho}|x]\leq 2$, there is

$$\mathbb{E}[\|x-x_y\|^2\overline{V}]\leq 2\mathbb{E}[\|x-x_y\|^2]=2\eta d. \tag{33}$$

By the choice $\eta\leq 1$ and the bounds (32) and (33), we have

$$\frac{\mathbb{E}|V-\overline{V}|\cdot\mathbb{E}[\|x-x_y\|^2\overline{V}]}{\mathbb{E}\overline{V}}\leq\frac{\zeta}{4}.$$

This, together with (26) and (31), gives that $W_2^2(\pi^{X|Y},\hat{\pi}^{X|Y})\leq\zeta$.

Finally, because of the choice $\zeta<8d$, and the bound (32), we have $\mathbb{E}|V-\overline{V}|\leq\frac{1}{2}$. Thus, the acceptance probability $\frac{1}{2}\mathbb{E}[\overline{V}]\geq\frac{1}{2}(\mathbb{E}[V]-\mathbb{E}|V-\overline{V}|)\geq\frac{1}{4}$. So the expectation of iterations in rejection sampling follows as $\mathcal{O}(1)$. ■

C.3. Proof of Proposition 10

Proof **(For TV distance)** According to Theorem 6, if choosing $\eta \leq \frac{1}{49L_1\sqrt{d}(1+\log(1+24T/\delta))}$, we can guarantee $\left\|\hat{\pi}^{X|Y}(\cdot|y) - \pi^{X|Y}(\cdot|y)\right\|_{\text{TV}} \leq \delta/(2T)$ for $\forall\ y$. Then by Lemma 8, we have $\|\hat{\mu}_T - \mu_T\|_{\text{TV}} \leq \frac{\delta}{2}$.

By Pinsker's inequality, there is $\|\mu_T - \nu\|_{\text{TV}} \leq \sqrt{2H_\nu(\mu_T)}$. Then by Theorem 1 (Chen et al., 2022, Theorem 3), Algorithm 1 returns a random point x_T that satisfies

$$H_\nu(\mu_T) \leq \frac{H_\nu(\mu_0)}{(1+\eta\beta)^{2T}} \leq \frac{\delta^2}{8}$$

in $T \geq \log\left(\frac{2}{\delta}\sqrt{2H_\nu(\mu_0)}\right)/\log(1+\beta\eta)$ steps. Thus $\|\mu_T - \nu\| \leq \delta/2$.

Putting two together, we have

$$\|\hat{\mu}_T - \nu\|_{\text{TV}} \leq \|\hat{\mu}_T - \mu_T\|_{\text{TV}} + \|\mu_T - \nu\|_{\text{TV}} \leq \delta.$$

Since $\eta\beta = \mathcal{O}(1)$, we have $\log(1+\eta\beta) = \mathcal{O}(\eta\beta)$. Thus, plugging in the value of η, we need

$$T = \mathcal{O}\left(\frac{L_1\sqrt{d}}{\beta}\log\left(\frac{L_1\sqrt{d}}{\beta\delta}\right)\log\left(\frac{\sqrt{H_\nu(\mu_0)}}{\delta}\right)\right)$$

steps. Each step accesses only $\mathcal{O}(1)$ many $f(x)$ in expectation because of Theorem 6.

(For Wasserstein distance) Since ν is β-strongly-log-concave, it satisfies Talagrand inequality

$$W_2(\mu_T, \nu) \leq \sqrt{\frac{2}{\beta}H_\nu(\mu_T)}.$$

Next, Theorem 1 (Chen et al., 2022, Theorem 3) guarantees that if $T \geq \log\left(\frac{2}{\delta}\sqrt{\frac{2H_\nu(\mu_0)}{\beta}}\right)/\log(1+\beta\eta)$ then

$$H_\nu(\mu_T) \leq \frac{H_\nu(\mu_0)}{(1+\beta\eta)^{2T}} \leq \frac{\beta\delta^2}{8}.$$

Thus, $W_2(\mu_T, \nu) \leq \frac{\delta}{2}$. On the other hand, according to Theorem 7, if choosing

$$\eta \leq \min\left(\frac{1}{49L_1 d^{\frac{1}{2}}(2+\log(1+3100(d^2+2d)T^4/\delta^4))}, 1\right),$$

we can guarantee $W_2(\hat{\pi}^{X|Y}(\cdot|y), \pi^{X|Y}(\cdot|y)) \leq \delta/2T$ for $\forall\ y$. Then Lemma 9 guarantees that $W_2(\mu_T, \hat{\mu}_T) \leq \frac{\delta}{2}$. Putting two pieces together, we get

$$W_2(\hat{\mu}_T, \nu) \leq W_2(\hat{\mu}_T, \mu_T) + W_2(\mu_T, \nu) \leq \delta.$$

Since $\beta\eta = \mathcal{O}(1)$, we have $\log(1+\beta\eta) = \mathcal{O}(\beta\eta)$. So, plugging in the value of η, we only need $T = \mathcal{O}\left(\frac{1}{\beta\eta}\log\left(\frac{2}{\delta}\sqrt{\frac{2H_\nu(\mu_0)}{\beta}}\right)\right) = \mathcal{O}\left(\frac{L_1\sqrt{d}}{\beta}\log\left(\frac{L_1 d}{\beta\delta}\right)\log\left(\frac{1}{\delta}\sqrt{\frac{H_\nu(\mu_0)}{\beta}}\right)\right)$ steps. Each step accesses only $\mathcal{O}(1)$ many $f(x)$ in expectation because of Theorem 7. ■

C.4. Proof of Proposition 11

Proof According to Theorem 6, if choosing $\eta \leq \frac{1}{49L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}} (1+\log(1+24T/\delta))}$, we can guarantee $\left\|\hat{\pi}^{X|Y}(\cdot|y) - \pi^{X|Y}(\cdot|y)\right\|_{\text{TV}} \leq \delta/(2T)$ for $\forall\, y$. Then by Lemma 8, we have $\|\hat{\mu}_T - \mu_T\|_{\text{TV}} \leq \frac{\delta}{2}$.

By Pinsker's inequality, there is $\|\mu_T - \nu\|_{\text{TV}} \leq \sqrt{2H_\nu(\mu_T)}$. Since $f(x)$ is convex and L_1-smooth, by Theorem 1 (Chen et al., 2022, Theorem 2), Algorithm 1 returns a random point x_T that satisfies

$$H_\nu(\mu_T) \leq \frac{W_2^2(\mu_0, \nu)}{T\eta} \leq \frac{\delta^2}{8}$$

in $T = \mathcal{O}\left(W_2^2(\mu_0, \nu)/(\delta^2\eta)\right)$ steps. Thus $\|\mu_T - \nu\| \leq \delta/2$.

Putting two together, we have

$$\|\hat{\mu}_T - \nu\|_{\text{TV}} \leq \|\hat{\mu}_T - \mu_T\|_{\text{TV}} + \|\mu_T - \nu\|_{\text{TV}} \leq \delta.$$

Thus, plugging in the value of η, we need

$$T = \mathcal{O}\left(\frac{W_2^2(\mu_0, \nu) L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}}{\delta^2} \log\left(\frac{W_2^2(\mu_0, \nu) L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}}{\delta^3}\right)\right)$$

steps. Each step accesses only $\mathcal{O}(1)$ many $f(x)$ in expectation because of Theorem 6. ■

C.5. Proof of Proposition 13

Proof According to Theorem 6, if choosing $\eta \leq \frac{1}{49L_1\sqrt{d}(1+\log(1+24T/\delta))}$, we can guarantee $\left\|\hat{\pi}^{X|Y}(\cdot|y) - \pi^{X|Y}(\cdot|y)\right\|_{\text{TV}} \leq \delta/(2T)$ for $\forall\, y$. Then by Lemma 8, we have $\|\hat{\mu}_T - \mu_T\|_{\text{TV}} \leq \frac{\delta}{2}$.

By Pinsker's inequality, there is $\|\mu_T - \nu\|_{\text{TV}} \leq \sqrt{2H_\nu(\mu_T)}$. Then by Theorem 1 (Chen et al., 2022, Theorem 3), Algorithm 1 returns a random point x_T that satisfies

$$H_\nu(\mu_T) \leq \frac{H_\nu(\mu_0)}{(1 + \eta C_{\text{LSI}})^{2T}} \leq \frac{\delta^2}{8}$$

in $T \geq \log\left(\frac{2}{\delta}\sqrt{2H_\nu(\mu_0)}\right)/\log(1 + C_{\text{LSI}}\,\eta)$ steps. Thus $\|\mu_T - \nu\| \leq \delta/2$.

Putting two together, we have

$$\|\hat{\mu}_T - \nu\|_{\text{TV}} \leq \|\hat{\mu}_T - \mu_T\|_{\text{TV}} + \|\mu_T - \nu\|_{\text{TV}} \leq \delta.$$

Since $\eta C_{\text{LSI}} = \mathcal{O}(1)$, we have $\log(1 + \eta C_{\text{LSI}}) = \mathcal{O}(\eta C_{\text{LSI}})$. Thus, plugging in the value of η, we need

$$T = \mathcal{O}\left(\frac{L_1\sqrt{d}}{C_{\text{LSI}}} \log\left(\frac{L_1\sqrt{d}}{C_{\text{LSI}}\delta}\right) \log\left(\frac{\sqrt{H_\nu(\mu_0)}}{\delta}\right)\right)$$

steps. Each step accesses only $\mathcal{O}(1)$ many $f(x)$ in expectation because of Theorem 6. ■

C.6. Proof of Proposition 14

Proof According to Theorem 6, if choosing $\eta \leq \frac{1}{49L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}} (1+\log(1+24T/\delta))}$, we can guarantee $\left\|\hat{\pi}^{X|Y}(\cdot|y) - \pi^{X|Y}(\cdot|y)\right\|_{\text{TV}} \leq \delta/(2T)$ for $\forall\, y$. Then by Lemma 8, we have $\|\hat{\mu}_T - \mu_T\|_{\text{TV}} \leq \frac{\delta}{2}$.

Together with Pinsker's inequality, Nishiyama and Sason (2020) implies that $\|\mu_T - \nu\|_{\text{TV}} \leq \sqrt{2\log((1+\chi^2_\nu(\mu_T)))}$. Then by Chen et al. (2022, Theorem 4), Algorithm 1 returns a random point x_T that satisfies

$$\chi^2_\nu(\mu_T) \leq \frac{\chi^2_\nu(\mu_0)}{(1+C_{\text{PI}}\eta)^{2T}} \leq \exp(\delta^2/8) - 1$$

in $T \geq \frac{1}{2}\log\left(\frac{\chi^2_\nu(\mu_0)}{\exp(\delta^2/8)-1}\right)/\log(1+C_{\text{PI}}\eta) = \mathcal{O}\left(\log\left(\frac{\chi^2_\nu(\mu_0)}{\delta^2}\right)/\log(1+C_{\text{PI}}\eta)\right)$ steps. Thus $\|\mu_T - \nu\| \leq \delta/2$.

Putting two together, we have

$$\|\hat{\mu}_T - \nu\|_{\text{TV}} \leq \|\hat{\mu}_T - \mu_T\|_{\text{TV}} + \|\mu_T - \nu\|_{\text{TV}} \leq \delta.$$

Since $\eta C_{\text{PI}} = \mathcal{O}(1)$, we have $\log(1+\eta C_{\text{PI}}) = \mathcal{O}(\eta C_{\text{PI}})$. Thus, plugging in the value of η, we need

$$T = \mathcal{O}\left(\frac{L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}}{C_{\text{PI}}} \log\left(\frac{L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}}{C_{\text{PI}}\delta}\right) \log\left(\frac{\chi^2_\nu(\mu_0)}{\delta^2}\right)\right)$$

steps. Each step accesses only $\mathcal{O}(1)$ many $f(x)$ in expectation because of Theorem 6. ■

C.7. Proof of Theorem 16

Proof Recall that the return of Algorithm 2 follows the distribution $\hat{\pi}^{X|Y}$, and the target distribution of RGO is $\pi^{X|Y} \propto \exp(-g(x) - \frac{1}{2\eta}\|x - x_y\|^2)$. Following the same proof of Theorem 6, we have $\|\pi^{X|Y} - \hat{\pi}^{X|Y}\|_{\text{TV}} \leq \frac{2\mathbb{E}|V-\overline{V}|}{|\mathbb{E}V|}$, $|\mathbb{E}V| \geq 1$, and $\mathbb{E}[V - \overline{V}] \leq \mathbb{E}[\rho \mathbf{1}_{\rho\geq 2}]$.

Denote $\Delta = g(z) - g(x)$ and $\bar{\Delta} = \Delta/\log 2$, we have

$$\begin{aligned}\mathbb{E}[\rho\mathbf{1}_{\rho\geq 2}] &= \mathbb{E}[\exp(\Delta)\mathbf{1}_{\rho\geq 2}] = \mathbb{E}[\exp(\Delta)\mathbf{1}_{\Delta\geq\log 2}] = \mathbb{E}[\exp(\bar{\Delta}\log 2)\mathbf{1}_{\bar{\Delta}\geq 1}] \\ &\leq \sum_{i=1}^{\infty} \exp((i+1)\log 2)\Pr(\bar{\Delta} \geq i).\end{aligned}$$

Note that $g(x) = f(x) - \langle f'(x_y), x\rangle$ satisfies that $g'(x_y) = \sum_{i=j}^n (f_j'(x_y) - f_j'(x_y)) = 0$, and $\mathbb{E}[x] = x_y$. Moreover, $g(x)$ also satisfies (4) because $g_j'(x_1) - g_j'(x_2) = f_j'(x_1) - f_j'(x_2)$ for any x_1, x_2. We also have the inequality $(1 - 0.5/a)^{-a/2} \leq 1.5$ when $a \geq 1$. We plug $\epsilon = 0.5$ in Corollary 15, and obtain that, for $\forall r > 0$,

$$\Pr[g(x) - \mathbb{E}g(x) \geq r] \leq \frac{3}{2}\sum_{j=1}^n \exp\left(-\frac{(1+\alpha_j)\left(\frac{1}{\alpha_j}\right)^{\frac{\alpha_j}{1+\alpha_j}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha_j}} 2^{\frac{1-2\alpha_j}{1+\alpha_j}} (w_j r)^{\frac{2}{1+\alpha_j}}}{L_{\alpha_j}^{\frac{2}{\alpha_j+1}} d^{\frac{\alpha_j}{\alpha_j+1}} \eta}\right).$$

This further implies that

$$\begin{aligned}\Pr[g(z)-g(x)\geq r] &= \Pr[g(z)-\mathbb{E}g(z)+\mathbb{E}g(x)-g(x)\geq r]\\ &\leq \Pr\left[g(z)-\mathbb{E}g(z)\geq\frac{r}{2}\right]+\Pr\left[\mathbb{E}g(x)-g(x)\geq\frac{r}{2}\right]\\ &= \Pr\left[g(z)-\mathbb{E}g(z)\geq\frac{r}{2}\right]+\Pr\left[-g(x)-\mathbb{E}[-g(x)]\geq\frac{r}{2}\right]\\ &\leq 3\sum_{j=1}^{n}\exp\left(-\frac{(1+\alpha_j)\left(\frac{1}{\alpha_j}\right)^{\frac{\alpha_j}{1+\alpha_j}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha_j}}2^{\frac{1-2\alpha_j}{1+\alpha_j}}\left(\frac{w_j r}{2}\right)^{\frac{2}{1+\alpha_j}}}{L_{\alpha_j}^{\frac{2}{\alpha_j+1}}d^{\frac{\alpha_j}{\alpha_j+1}}\eta}\right).\end{aligned}$$

Thus $\Pr[\bar{\Delta}\geq i]\leq 3\sum_{j=1}^{n}\exp\left(-\frac{(1+\alpha_j)\left(\frac{1}{\alpha_j}\right)^{\frac{\alpha_j}{1+\alpha_j}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha_j}}2^{\frac{1-2\alpha_j}{1+\alpha_j}}\left(\frac{w_j i\log 2}{2}\right)^{\frac{2}{1+\alpha_j}}}{L_{\alpha_j}^{\frac{2}{\alpha_j+1}}d^{\frac{\alpha_j}{\alpha_j+1}}\eta}\right)$. We choose

$$w_j=\frac{L_{\alpha_j}^{\frac{1}{\alpha_j+1}}d^{\frac{\alpha_j}{2(\alpha_j+1)}}}{\sum_{j=1}^{n}L_{\alpha_j}^{\frac{1}{\alpha_j+1}}d^{\frac{\alpha_j}{2(\alpha_j+1)}}}. \tag{34}$$

Since $i^{\frac{2}{1+\alpha}}\geq i$ for any $i\geq 1$, $0\leq\alpha\leq 1$, and the bound (23), we have that

$$\begin{aligned}&\sum_{i=1}^{\infty}\exp((i+1)\log 2)\Pr(\bar{\Delta}\geq i)\\ \leq&6\sum_{i=1}^{\infty}\sum_{j=1}^{n}\exp\left(i-\frac{(1+\alpha_j)\left(\frac{1}{\alpha_j}\right)^{\frac{\alpha_j}{1+\alpha_j}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha_j}}2^{\frac{1-2\alpha_j}{1+\alpha_j}}\left(\frac{w_j i\log 2}{2}\right)^{\frac{2}{1+\alpha_j}}}{L_{\alpha_j}^{\frac{2}{\alpha_j+1}}d^{\frac{\alpha_j}{\alpha_j+1}}\eta}\right)\\ \leq&6\sum_{i=1}^{\infty}\sum_{j=1}^{n}\exp\left(i-\frac{w_j^2 i}{49L_{\alpha_j}^{\frac{2}{\alpha_j+1}}d^{\frac{\alpha_j}{\alpha_j+1}}\eta}\right)\overset{(34)}{=}6n\sum_{i=1}^{\infty}\exp\left(i-\frac{i}{49\left(\sum_{j=1}^{n}L_{\alpha_j}^{\frac{1}{\alpha_j+1}}d^{\frac{\alpha_j}{2(\alpha_j+1)}}\right)^2\eta}\right)\\ =&6n/\left(\exp\left(\frac{1}{49\left(\sum_{j=1}^{n}L_{\alpha_j}^{\frac{1}{\alpha_j+1}}d^{\frac{\alpha_j}{2(\alpha_j+1)}}\right)^2\eta}-1\right)-1\right)\leq\frac{\zeta}{2}\end{aligned} \tag{35}$$

by the choice

$$\eta\leq\frac{1}{49\left(\sum_{j=1}^{n}L_{\alpha_j}^{\frac{1}{\alpha_j+1}}d^{\frac{\alpha_j}{2(\alpha_j+1)}}\right)^2(1+\log(1+12n/\zeta))}.$$

Finally, by Lemma 3, the acceptance probability is $\frac{1}{2}\mathbb{E}[\overline{V}]$. Since $\zeta<1$, we have

$$\mathbb{E}[\overline{V}]\geq\mathbb{E}[V]-\mathbb{E}|V-\overline{V}|\geq 1-\frac{\zeta}{2}\geq\frac{1}{2}.$$

Thus, the expected number of the iterations for the rejection sampling step is $\mathcal{O}(1)$. ■

C.8. Proof of Theorem 17

Proof It suffices to prove the claim: *If the step size*

$$\eta \le \min\left(\frac{1}{49\left(\sum_{j=1}^n L_{\alpha_j}^{\frac{1}{\alpha_j+1}} d^{\frac{\alpha_j}{2(\alpha_j+1)}}\right)^2 (2+\log(1+192n(d^2+2d)/\zeta^2))}, 1\right), \tag{36}$$

then Algorithm 2 returns a random point x that has ζ ***squared*** *Wasserstein-2 distance to the distribution $\pi^{X|Y}$. Furthermore, if $0 < \zeta < 8d$, then the algorithm access $\mathcal{O}(1)$ many $f(x)$ in expectation.*

Recall that the return of Algorithm 2 follows the distribution $\hat{\pi}^{X|Y}$, and the target distribution of RGO is $\pi^{X|Y} \propto \exp(-g(x) - \frac{1}{2\eta}\|x - x_y\|^2)$. Following the same proof of Theorem 7, we have

$$W_2^2(\pi^{X|Y}, \hat{\pi}^{X|Y}) \le 2\left(\mathbb{E}\|x-x_y\|^4 \mathbb{E}|V-\overline{V}|^2\right)^{\frac{1}{2}} + \frac{2\mathbb{E}|V-\overline{V}| \cdot \mathbb{E}[\|x-x_y\|^2\overline{V}]}{\mathbb{E}\overline{V}}, \tag{37}$$

$$\mathbb{E}\|x-x_y\|^4 \le d^2+2d, \tag{38}$$

$$\mathbb{E}|V-\overline{V}|^2 \le \sum_{i=1}^{\infty} \exp(2(i+1)\log 2)\Pr(\bar{\Delta} \ge i), \tag{39}$$

$$\mathbb{E}[\|x-x_y\|^2\overline{V}] \le 2\eta d. \tag{40}$$

Applying the same concentration inequality and weight (34) in the proof of Theorem 16, we have

$$\begin{aligned}
&\sum_{i=1}^{\infty} \exp(2(i+1)\log 2)\Pr(\bar{\Delta} \ge i) \\
\le& 12n/\left(\exp\left(\frac{1}{49\left(\sum_{j=1}^n L_{\alpha_j}^{\frac{1}{\alpha_j+1}} d^{\frac{\alpha_j}{2(\alpha_j+1)}}\right)^2 \eta} - 2\right) - 1\right) \overset{(36)}{\le} \frac{\zeta^2}{16(d^2+2d)}.
\end{aligned}$$

This, together with the inequalities (38), (39), gives that

$$2\left(\mathbb{E}\|x-x_y\|^4 \mathbb{E}|V-\overline{V}|^2\right)^{\frac{1}{2}} \le \frac{\zeta}{4}.$$

Now we bound the second term in (37). By the assumption (36), we have

$$\eta \le \frac{1}{49\left(\sum_{j=1}^n L_{\alpha_j}^{\frac{1}{\alpha_j+1}} d^{\frac{\alpha_j}{2(\alpha_j+1)}}\right)^2 (1+\log(1+96nd/\zeta))}.$$

Thus, follow the same steps in bounding (35) and (24), we obtain

$$\mathbb{E}|V - \overline{V}| \leq \frac{\zeta}{16d} \text{ and } \mathbb{E}[\overline{V}] \geq \frac{1}{2}.$$

By the choice $\eta \leq 1$, the above bounds, and the inequalities (37), (40), we reach the result $W_2^2(\pi^{X|Y}, \hat{\pi}^{X|Y}) \leq \zeta$.

The expectation of iterations in rejection sampling also follows the proof of Theorem 7. ■

Appendix D. RGO with approximate proximal optimization error

In Algorithm 2 of Section 3, we assume that we can solve the step 2 exactly, which is to solve the stationary point of $f_y^\eta(x) = f(x) + \frac{1}{2\eta}\|x - y\|^2$. However, it could be challenging to solve this optimization problem exactly when f is non-convex. In this section, we tackle this issue and present the complexity analysis when we can only obtain an approximate stationary point of it.

D.1. Algorithm

Algorithm 3: Rejection sampling implementation of RGO with proximal optimization error

1 **Input**: L_α-α-semi-smooth function $f(x)$, step size $\eta > 0$, current point y
2 Compute an approximation solution x_y satisfying (42) by using Algorithm 3 in Liang and Chen (2022a). Denote $g(x) = f(x) - \langle f'(x_y)), x\rangle,\ w = y - \eta f'(x_y)$.
3 **repeat**
4 Sample x, z from the distribution $\phi(\cdot) \propto \exp(-\frac{1}{2\eta}\|\cdot - w\|_2^2)$
5 $\rho = \exp(g(z) - g(x))$
6 Sample u uniformly from $[0, 1]$.
7 **until** $u \leq \frac{1}{2}\rho$;
8 **Return** x

We will use Algorithm 3 as the RGO algorithm, and Lemma 3 still holds for this algorithm. We resort to Nesterov's accelerated gradient descent method to compute w that is an s-approximation to x_y, i.e., $\|x_y - w\| \leq s$, see (Liang and Chen, 2022a, Algorithm 3). And we have the following lemma to guarantee a similar equivalence as in Lemma 2.

Lemma 23 *For any x_y, sampling from $\pi^{X|Y}(x|y) \propto \exp(-f(x) - \frac{1}{2\eta}\|x - y\|^2)$ is equivalent to sampling from distribution $\propto \exp(-g(x) - \frac{1}{2\eta}\|x - w\|^2)$, where $g(x) = f(x) - \langle f'(x_y), x\rangle$ and $w = y - \eta f'(x_y)$. If we further assume x_y is an approximate stationary point to $f(x) + \frac{1}{2\eta}\|x - y\|^2$, i.e.,*

$$\left\| f'(x_y) + \frac{1}{\eta}(x_y - y) \right\| \leq \frac{s}{\eta}, \tag{41}$$

then we have $\|x_y - w\| \leq s$.

Proof We first show the potentials of two distributions are the same up to a constant.

$$\begin{aligned} g(x) + \frac{1}{2\eta}\|x - w\|^2 &= g(x) + \frac{1}{2\eta}\|x - y + \eta f'(x_y)\|^2 \\ &= g(x) + \frac{1}{2\eta}\|x - y\|^2 + \langle x, f'(x_y)\rangle - \langle y, f'(x_y)\rangle + \frac{\eta}{2}\|f'(x_y)\|^2 \\ &= g(x) + \frac{1}{2\eta}\|x - y\|^2 + \langle x, f'(x_y)\rangle + \text{constant} \\ &= f(x) + \frac{1}{2\eta}\|x - y\|^2 + \text{constant}. \end{aligned}$$

We use $f(x) = g(x) + \langle f'(x_y), x\rangle$ in the last equality. By the definition of $g(x)$, we have that $g'(x_y) = 0$, thus

$$\|x_y - w\| = \|x_y - y + \eta f'(x_y)\| = \eta \left\| f'(x_y) + \frac{1}{\eta}(x_y - y) \right\| \overset{(41)}{\leq} s.$$

■

D.2. Complexity analysis

We first investigate the complexity of the optimization algorithm to reach a small enough tolerance. Later, it turns out that the tolerance in (42) will be enough.

Lemma 24 *Assume $\eta = 1/(98L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}(1 + \log(1 + 12/\zeta)))$. Let $x_y \in \mathbb{R}^d$ be an approximate stationary point of f_y^η, i.e.,*

$$\left\| f'(x_y) + \frac{1}{\eta}(x_y - y) \right\| \leq \frac{d^{\frac{1}{2(1+\alpha)}}}{7L_\alpha^{\frac{1}{1+\alpha}}\eta}. \tag{42}$$

Then, the iteration complexity to find x_y by Liang and Chen (2022a, Algorithm 3) is $\tilde{\mathcal{O}}(1)$.

Proof According to Liang and Chen (2022a, Lemma A.2), $f_y^\eta := f(x) + \frac{1}{2\eta}\|x - y\|^2$ satisfies that

$$\frac{\beta}{2}\|u - v\|^2 - \theta \leq f_y^\eta(u) - f_y^\eta(v) - \left\langle (f_y^\eta)'(v), u - v \right\rangle \leq \frac{L}{2}\|u - v\|^2 + \theta, \quad \forall u, v \in \mathbb{R}^d,$$

with

$$M = \frac{L_\alpha^{\frac{2}{1+\alpha}}}{(1+\alpha)^{\frac{1-\alpha}{1+\alpha}}}, \ \ \beta = \frac{1}{\eta} - M, \ \ L = \frac{1}{\eta} + M, \ \ \theta = \frac{1-\alpha}{2}.$$

Denote the upper bound in (42) as $\rho := \frac{d^{\frac{1}{2(1+\alpha)}}}{7L_\alpha^{\frac{1}{1+\alpha}}\eta}$. Since $(1+\alpha)^{\frac{1-\alpha}{1+\alpha}} \geq 1$, by simple calculation, we can verify that

$$2\sqrt{2}(\beta + L)\theta/\sqrt{\beta} = \frac{2\sqrt{2}(1-\alpha)}{\eta\sqrt{\frac{1}{\eta} - M}} \leq \frac{2\sqrt{2}(1-\alpha)}{L_\alpha^{\frac{1}{\alpha+1}}\eta\sqrt{98d^{\frac{\alpha}{\alpha+1}}(1 + \log(1 + 12/\zeta)) - 1}} \leq \rho.$$

Then by Liang and Chen (2022a, Lemma B.4), the number of iterations to obtain (42) is at most

$$\frac{2\sqrt{L}+\sqrt{\beta}}{2\sqrt{\beta}}\log\left(\frac{(\beta+L)^2\|x^{(0)}-x^*\|^2}{\rho^2}\frac{2\sqrt{L}+\sqrt{\beta}}{2\sqrt{\beta}}+1\right)=\tilde{\mathcal{O}}(1),$$

where $x^{(0)}$ is the initialization point in the optimization algorithm and x^* is a ground truth stationary point of f_y^η. ■

Considering the error in solving the stationary point of f_y^η, we will need another concentration bound for the sampling complexity proof. Compared with Theorem 4, this concentration inequality only has an additional coefficient $\exp\left(\frac{s^2}{2\eta(d/\epsilon-1)}\right)$ with all other terms intact.

Proposition 25 *Let $X\sim\mathcal{N}(m_0,\eta\mathbf{I})$ be a Gaussian random variable in $\mathbb{R}^d$, and let ℓ be an L_α-α-semi-smooth function. Assume $\ell'(m_1)=0$, and $\|m_0-m_1\|\le s$. Then for any $r>0,\ 0\le\alpha\le 1$, one has $\forall\epsilon\in(0,d)$,*

$$\Pr(\ell(X)-\mathbb{E}(\ell(X))\ge r)\le\exp\left(\frac{s^2}{2\eta(d/\epsilon-1)}\right)\left(1-\frac{\epsilon}{d}\right)^{-d/2}\exp\left(-\frac{C\epsilon^{\frac{\alpha}{1+\alpha}}r^{\frac{2}{1+\alpha}}}{L_\alpha^{\frac{2}{1+\alpha}}d^{\frac{\alpha}{1+\alpha}}\eta}\right),$$

where

$$C=(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}}2^{\frac{1-\alpha}{1+\alpha}}.$$

Proof With the same procedure in the proof of Theorem 4, one can obtain

$$\Pr(\ell(X)-\mathbb{E}(\ell(X))\ge r)\le\inf_{\lambda>0}\frac{\mathbb{E}_Z\exp(\frac{\pi^2}{8}\eta\lambda^2\|\nabla\ell(Z)\|_2^2)}{\exp(\lambda r)}.$$

where $Z\sim\mathcal{N}(0,\eta\mathbf{I})$ and $\nabla\ell(m_1-m_0)=0$. In what follows, we denote m_1-m_0 by ι. Again, we proceed by considering three cases.

1. $\alpha=0$

   Notice that $\|\nabla\ell(Z)\|_2^2\le L_0^2$. By taking $\lambda=\frac{4r}{\pi^2\eta L_0^2}$, one could get

   $$\Pr(\ell(X)-\mathbb{E}(\ell(X))\ge r)\le\exp(-\frac{2r^2}{\pi^2\eta L_0^2}).$$

   Note that for the case that $\alpha=0$, the result is the same as the one in Theorem 4.

2. $\alpha=1$

   Assume $-\frac{1}{2\eta}+\frac{\pi^2}{8}L_1^2\eta\lambda^2<0$, then

   $$\begin{aligned}\frac{\mathbb{E}_Z\exp(\frac{\pi^2}{8}\lambda^2\eta\|\nabla\ell(Z)\|_2^2)}{\exp(\lambda r)}&\le\frac{\mathbb{E}_Z\exp(\frac{\pi^2}{8}L_1^2\eta\lambda^2\|Z-\iota\|_2^2)}{\exp(\lambda r)}\\&=\left(\frac{1}{1-\frac{\pi^2}{4}L_1^2\eta^2\lambda^2}\right)^{d/2}\exp\left(\frac{1/2\eta}{\frac{4}{\pi^2L_1^2\eta^2\lambda^2}-1}\|\iota\|_2^2\right)\exp(-\lambda r).\end{aligned}$$

Let $\lambda = \frac{k}{L_1^2 d\eta^2}$ and denote $\frac{\pi^2 k^2}{4\eta^2 L_1^2 d}$ by ϵ. Then,

$$\Pr(f(X) - \mathbb{E}(f(X)) \geq r) \leq (1 - \frac{\epsilon}{d})^{-\frac{d}{2}} \exp\left(\frac{1/2\eta}{d/\epsilon - 1}\|\iota\|_2^2\right) \exp(-\sqrt{\frac{4\epsilon}{\pi^2}}\frac{r}{\eta L_1 d^{1/2}}),\ \forall \epsilon \in (0, d)$$

Note that given the value of ϵ, $(1 - \frac{\epsilon}{d})^{-\frac{d}{2}}$ is bounded for d. One can also observe that compared with the result in Theorem 4, the only difference is the additional term $\exp\left(\frac{1/2\eta}{d/\epsilon - 1}\|\iota\|_2^2\right)$.

3. $0 < \alpha < 1$

By Young's inequality, for any $\omega > 0$, one obtains $\|Z - \iota\|^{2\alpha}\omega \leq \alpha\|Z - \iota\|^2 + (1 - \alpha)\omega^{\frac{1}{1-\alpha}}$. Hence, with the assumption $1 - \frac{\pi^2}{4}L_\alpha^2\eta^2\lambda^2\frac{\alpha}{\omega} > 0$,

$$\begin{aligned}
&\frac{\mathbb{E}_Z \exp(\frac{\pi^2}{8}\eta\lambda^2\|\nabla\ell(Z)\|_2^2)}{\exp(\lambda r)} \leq \frac{\mathbb{E}_Z \exp(\frac{\pi^2}{8}L_\alpha^2\eta\lambda^2\|Z - \iota\|_2^{2\alpha})}{\exp(\lambda r)} \\
\leq &(1 - \frac{\pi^2}{4}\eta^2\lambda^2 L_\alpha^2\frac{\alpha}{\omega})^{-\frac{d}{2}} \exp\left(\frac{1/2\eta\|\iota\|_2^2}{1/\left(\frac{\pi^2}{4}\eta^2\lambda^2 L_\alpha^2\frac{\alpha}{\omega}\right) - 1}\right) \exp(\frac{\pi^2}{8}\eta\lambda^2 L_\alpha^2(1 - \alpha)\omega^{\frac{\alpha}{1-\alpha}}) \exp(-\lambda r).
\end{aligned} \tag{43}$$

One can notice that compared with the case $\alpha \in (0, 1)$ in the proof of Theorem 4, the only additional term is $\exp\left(\frac{1/2\eta}{1/\left(\frac{\pi^2}{4}\eta^2\lambda^2 L_\alpha^2\frac{\alpha}{\omega}\right) - 1}\|\iota\|_2^2\right)$. With the same suboptimal $\hat{\lambda}$ and $\hat{\omega}$ as in the proof of Theorem 4, we have $\forall \epsilon \in (0, d)$,

$$\Pr(\ell(X) - \mathbb{E}(\ell(X)) \geq r) \leq \left(1 - \frac{\epsilon}{d}\right)^{-d/2} \exp\left(\frac{1/2\eta}{d/\epsilon - 1}\|\iota\|_2^2\right) \exp\left(-\frac{C\epsilon^{\frac{\alpha}{\alpha+1}} r^{\frac{2}{\alpha+1}}}{L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}\eta}\right),$$

with C being $(1 + \alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}} 2^{\frac{1-\alpha}{1+\alpha}}$.

■

Following the same procedures in Section 3, with the concentration inequality at our disposal, we can probabilistically bound the difference of ρ and $\bar{\rho}$, which determines the discrepancy between $\hat{\pi}^{X|Y}$ and $\pi^{X|Y}$.

Theorem 26 (RGO complexity in total variation) *If the step size*

$$\eta \leq \frac{1}{98 L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}(1 + \log(1 + 12/\zeta))},$$

then for any $\zeta > 0$, Algorithm 3 returns a random point x that has ζ total variation distance to the distribution proportional to $\pi^{X|Y}(\cdot|y)$. Furthermore, if $0 < \zeta < 1$, then the algorithm access $\mathcal{O}(1)$ many $f(x)$ in expectation.

Proof Following the same proof as Theorem 6, we obtain

$$\|\pi^{X|Y} - \hat{\pi}^{X|Y}\|_{\text{TV}} \leq 2\sum_{i=1}^{\infty} \exp((i+1)\log 2)\Pr(\bar{\Delta} \geq i).$$

Note that $g(x) = f(x) - \langle \nabla f(x_y), x\rangle$ satisfies that $\nabla g(x_y) = 0$, and $\mathbb{E}[x] = w$. Moreover, $g(x)$ is also L_α-α-semi-smooth because $\nabla g(x_1) - \nabla g(x_2) = \nabla f(x_1) - \nabla f(x_2)$ for any x_1, x_2. We also have the inequality $(1 - 0.5/a)^{-a/2} \leq 1.5$ and $0 < 1/(2a-1) \leq 1/a$ when $a \geq 1$. Since (42) in Lemma 24 is satisfied, further by Lemma 23, we have

$$\|x_y - w\| \leq s := \frac{d^{\frac{1}{2(1+\alpha)}}}{7L_\alpha^{\frac{1}{1+\alpha}}}. \tag{44}$$

Thus, we plug in $\epsilon = 0.5$ in Proposition 25, and obtain that, for $\forall r > 0$,

$$\Pr[g(x) - \mathbb{E}g(x) \geq r] \leq \frac{3}{2}\exp\left(\frac{s^2}{2d\eta} - \frac{(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}} 2^{\frac{1-2\alpha}{1+\alpha}} r^{\frac{2}{1+\alpha}}}{L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}\eta}\right). \tag{45}$$

This further implies that

$$\begin{aligned}
\Pr[g(z) - g(x) \geq r] &= \Pr[g(z) - \mathbb{E}g(z) + \mathbb{E}g(x) - g(x) \geq r] \\
&\leq \Pr\left[g(z) - \mathbb{E}g(z) \geq \frac{r}{2}\right] + \Pr\left[\mathbb{E}g(x) - g(x) \geq \frac{r}{2}\right] \\
&= \Pr\left[g(z) - \mathbb{E}g(z) \geq \frac{r}{2}\right] + \Pr\left[-g(x) - \mathbb{E}[-g(x)] \geq \frac{r}{2}\right] \\
&\leq 3\exp\left(-\left(\frac{(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}} 2^{\frac{1-2\alpha}{1+\alpha}}\left(\frac{r}{2}\right)^{\frac{2}{1+\alpha}}}{L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}} - \frac{s^2}{2d}\right)\frac{1}{\eta}\right).
\end{aligned}$$

Thus $\Pr[\bar{\Delta} \geq i] \leq 3\exp\left(-\left(\frac{(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}} 2^{\frac{1-2\alpha}{1+\alpha}}\left(\frac{i\log 2}{2}\right)^{\frac{2}{1+\alpha}}}{L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}} - \frac{s^2}{2d}\right)\frac{1}{\eta}\right)$. Denote

$$C_\eta := \left(\frac{(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}} 2^{\frac{-1-2\alpha}{1+\alpha}}(\log 2)^{\frac{2}{1+\alpha}}}{L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}} - \frac{s^2}{2d}\right)\frac{1}{\eta} - 1.$$

Since $\log 2 < 1$ and $i^{\frac{2}{1+\alpha}} \geq i$ for any $i \geq 1$, $0 \leq \alpha \leq 1$, we have that

$$\begin{aligned}
&\sum_{i=1}^{\infty} \exp((i+1)\log 2)\Pr(\bar{\Delta} \geq i) \\
\leq& 6\sum_{i=1}^{\infty}\exp\left(i - \left(\frac{(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}} 2^{\frac{1-2\alpha}{1+\alpha}}\left(\frac{i\log 2}{2}\right)^{\frac{2}{1+\alpha}}}{L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}} - \frac{s^2}{2d}\right)\frac{1}{\eta}\right) \\
\leq& 6\sum_{i=1}^{\infty}\exp\left(-C_\eta i\right) = \frac{6}{\exp(C_\eta) - 1} \leq \frac{\zeta}{2}
\end{aligned} \tag{46}$$

by the choice

$$\eta \leq \left(\frac{(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}} 2^{\frac{-1-2\alpha}{1+\alpha}} (\log 2)^{\frac{2}{1+\alpha}}}{L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}} - \frac{s^2}{2d} \right) \frac{1}{(1+\log(1+12/\zeta))}.$$

Since the bound (23) and the value of s in (44), the bound

$$\eta \leq \frac{1}{98 L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}} (1+\log(1+12/\zeta))}$$

suffices. This finishes the proof.

Finally, by Lemma 3, the acceptance probability is $\frac{1}{2}\mathbb{E}[\overline{V}]$. Since $\zeta < 1$, we have

$$\mathbb{E}[\overline{V}] \geq \mathbb{E}[V] - \mathbb{E}|V - \overline{V}| \geq 1 - \frac{\zeta}{2} \geq \frac{1}{2}.$$

Thus, the expected number of iterations for the rejection sampling step is $\mathcal{O}(1)$. ■

Theorem 27 (RGO complexity in Wasserstein distance) *If the step size*

$$\eta \leq \min\left(\frac{1}{98 L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}} (2+\log(1+192(d^2+2d)/\zeta^4))}, 1 \right),$$

then Algorithm 3 returns a random point x that has ζ Wasserstein-2 distance to the distribution $\pi^{X|Y}(\cdot|y)$. Furthermore, if $0 < \zeta < 2\sqrt{2d}$, then the algorithm access $\mathcal{O}(1)$ many $f(x)$ in expectation.

Proof It suffices to prove the claim: *If the step size*

$$\eta \leq \min\left(\frac{1}{98 L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}} (2+\log(1+192(d^2+2d)/\zeta^2))}, 1 \right), \tag{47}$$

then Algorithm 2 returns a random point x that has ζ ***squared*** *Wasserstein-2 distance to the distribution $\pi^{X|Y}$. Furthermore, if $0 < \zeta < 8d$, then the algorithm access $\mathcal{O}(1)$ many $f(x)$ in expectation.*

By Lemma 21 (Villani (2003, Proposition 7.10)) and triangular inequality,

$$\begin{aligned}
W_2^2(\pi^{X|Y}, \hat{\pi}^{X|Y}) &\leq 2\left\| \|\cdot - w\|_2^2 (\pi^{X|Y} - \hat{\pi}^{X|Y}) \right\|_{\mathrm{TV}} = 2\mathbb{E}\left(\|x-w\|_2^2 \left| \frac{V}{\mathbb{E}V} - \frac{\overline{V}}{\mathbb{E}\overline{V}} \right| \right) \\
&\leq 2\mathbb{E}\|x-w\|_2^2 \left(\left| \frac{V}{\mathbb{E}V} - \frac{\overline{V}}{\mathbb{E}V} \right| + \left| \frac{\overline{V}}{\mathbb{E}V} - \frac{\overline{V}}{\mathbb{E}\overline{V}} \right| \right) \\
&\leq \frac{2\mathbb{E}[\|x-w\|^2 |V-\overline{V}|]}{\mathbb{E}V} + \frac{2\mathbb{E}[\|x-w\|^2 \overline{V}|\mathbb{E}[V-\overline{V}]|]}{\mathbb{E}V\mathbb{E}\overline{V}} \\
&\overset{(20)}{\leq} 2\mathbb{E}[\|x-w\|^2 |V-\overline{V}|] + \frac{2\mathbb{E}|V-\overline{V}| \cdot \mathbb{E}[\|x-w\|^2 \overline{V}]}{\mathbb{E}\overline{V}}.
\end{aligned}$$

Firstly, by Cauchy-Schwartz inequality,

$$\mathbb{E}[\|x-w\|^2|V-\overline{V}|] \le \left(\mathbb{E}\|x-w\|^4\mathbb{E}|V-\overline{V}|^2\right)^{\frac{1}{2}}.$$

Similar to (28) and (33), we have

$$\mathbb{E}\|x-w\|^4 \le d^2+2d \text{ and } \mathbb{E}[\|x-w\|^2\overline{V}] \le 2\eta d. \tag{48}$$

On the other hand, following the same logic of bounding $\mathbb{E}[V-\overline{V}]$ in Theorem 6,

$$\mathbb{E}|V-\overline{V}|^2 \le \sum_{i=1}^{\infty} \exp(2(i+1)\log 2)\Pr(\bar{\Delta}\ge i). \tag{49}$$

Note that the s value in (44) still holds. By the concentration (45), and the bound (23), we have

$$\begin{aligned}
&\sum_{i=1}^{\infty} \exp(2(i+1)\log 2)\Pr(\bar{\Delta}\ge i)\\
\overset{(45)}{\le}& 12\sum_{i=1}^{\infty}\exp\left(2i-\left(\frac{(1+\alpha)\left(\frac{1}{\alpha}\right)^{\frac{\alpha}{1+\alpha}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\alpha}}2^{\frac{1-2\alpha}{1+\alpha}}\left(\frac{i\log 2}{2}\right)^{\frac{2}{1+\alpha}}}{L_\alpha^{\frac{2}{\alpha+1}}d^{\frac{\alpha}{\alpha+1}}}-\frac{s^2}{2d}\right)\frac{1}{\eta}\right)\\
\le& 12\sum_{i=1}^{\infty}\exp\left(2i-\left(\frac{i}{49L_\alpha^{\frac{2}{\alpha+1}}d^{\frac{\alpha}{\alpha+1}}}-\frac{s^2}{2d}\right)\frac{1}{\eta}\right) \le 12\exp\left(2i-\left(\frac{1}{49L_\alpha^{\frac{2}{\alpha+1}}d^{\frac{\alpha}{\alpha+1}}}-\frac{s^2}{2d}\right)\frac{i}{\eta}\right)\\
=&\frac{12}{\exp\left(\left(\frac{1}{49L_\alpha^{\frac{2}{\alpha+1}}d^{\frac{\alpha}{\alpha+1}}}-\frac{s^2}{2d}\right)/\eta-2\right)-1} \overset{(44)}{=} \frac{12}{\exp\left(\frac{1}{98L_\alpha^{\frac{2}{\alpha+1}}d^{\frac{\alpha}{\alpha+1}}\eta}-2\right)-1} \overset{(47)}{\le} \frac{\zeta^2}{16(d^2+2d)}.
\end{aligned}$$

This gives $\mathbb{E}[\|x-w\|^2|V-\overline{V}|] \le \frac{\zeta}{4}$. Next, following the same steps to bound (46) and (24), we obtain

$$\mathbb{E}|V-\overline{V}| \le \frac{\zeta}{16d} \text{ and } \mathbb{E}[\overline{V}] \ge \frac{1}{2}. \tag{50}$$

The above three inequalities, together with (48), gives that $W_2^2(\pi^{X|Y},\hat{\pi}^{X|Y}) \le \zeta$.

Finally, because of the choice $\zeta < 8d$, and the bound (50), we have $\mathbb{E}|V-\overline{V}| \le \frac{1}{2}$. Thus, the acceptance probability $\frac{1}{2}\mathbb{E}[\overline{V}] \ge \frac{1}{2}(\mathbb{E}[V]-\mathbb{E}|V-\overline{V}|) \ge \frac{1}{4}$. So the expectation of iterations in rejection sampling follows as $\mathcal{O}(1)$. ■

These two theorems show that the step size will have the same order as in Section 3. Thus all the results in Section 4 also follow if the RGO is realized by Algorithm 3.

Appendix E. Extension to the convergence in χ^2-divergence

We now extend our results in §3, §4, §D to the strong notion of χ^2-divergence. We do not modify the the concentration inequalities nor the RGO algorithm, but only the proof of RGO step size. The new RGO results in χ^2 with both accurate and inaccurate optimization step are shown in §E.1. Then in §E.2, we combine our RGO result with the techniques in Altschuler and Chewi (2023) to get the final convergence results in χ^2.

E.1. RGO for semi-smooth potential

We first consider we can solve optimization step 2 accurately in Algorithm 2. This theorem is more general than Theorem 6 and 7, which are presented under the weaker metric TV or W_2.

Theorem 28 (RGO complexity in χ^2-divergence with accurate optimization step 2) *Assume f satisfies (1). For $\forall \zeta > 0$, if*

$$\eta \leq \left(49 L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}} (2 + \log(1 + 522/\zeta))\right)^{-1},$$

then Algorithm 2 returns a random point x that has ζ χ^2-divergence to the distribution proportional to $\pi^{X|Y}(\cdot|y)$. Furthermore, if $0 < \zeta < 0.5$, then the algorithm access $\mathcal{O}(1)$ many $f(x)$ in expectation.

Proof Recall that the return of Algorithm 2 follows the distribution $\hat{\pi}^{X|Y}$, and the target distribution of RGO is $\pi^{X|Y} \propto \exp(-g(x) - \frac{1}{2\eta}\|x - x_y\|^2)$. According to Lemma 3, we have

$$\begin{aligned}\chi^2_{\pi^{X|Y}}(\hat{\pi}^{X|Y}) &= \frac{\mathbb{E}[\overline{V}^2 V^{-1}]\mathbb{E}[V]}{(\mathbb{E}\overline{V})^2} - 1 = \frac{\mathbb{E}[(V-\overline{V})^2 V^{-1}]\mathbb{E}[V] - (\mathbb{E}[V-\overline{V}])^2}{(\mathbb{E}\overline{V})^2} \\ &\leq \frac{\mathbb{E}[(V-\overline{V})^2 V^{-1}]\mathbb{E}[V]}{(\mathbb{E}\overline{V})^2}\end{aligned}$$

We first bound the two expectations in the numerator. Denote $\Delta = g(z) - g(x)$. Following (8) in the proof of concentration inequality in Theorem 4, we have

$$\begin{aligned}\mathbb{E}[V] = \mathbb{E}_{x,z}[\rho] = \mathbb{E}_{x,z}[\exp(\Delta)] &\leq \mathbb{E}_x[\exp(\pi^2\eta\|\nabla g(x)\|^2/8)] \\ &\overset{(11)}{\leq} \left(1 - \frac{\pi^2 L_\alpha^2 \eta^2 \alpha}{4w}\right)^{-\frac{d}{2}} \exp\left(\frac{\pi^2}{8} L_\alpha^2 \eta (1-\alpha) w^{\frac{\alpha}{1-\alpha}}\right)\end{aligned}$$

for $\forall w > 0$. By choosing $w = (1-\alpha)^{-\frac{1-\alpha}{\alpha}}(\eta d)^{1-\alpha}$, the above term becomes

$$\left(1 - \frac{\pi^2 L_\alpha^2 \eta^{1+\alpha}\alpha(1-\alpha)^{\frac{1-\alpha}{\alpha}}}{4d^{1-\alpha}}\right)^{-\frac{d}{2}} \exp\left(\frac{\pi^2}{8} L_\alpha^2 \eta^{1+\alpha} d^\alpha\right).$$

We further choose $\eta \leq \frac{1}{2L_\alpha^{\frac{2}{1+\alpha}} d^{\frac{\alpha}{1+\alpha}}} \leq \left(\frac{8\log(1.2)}{\pi^2 L_\alpha^2 d^\alpha}\right)^{\frac{1}{1+\alpha}}$, so the above term is smaller than

$$1.2\left(1 - \frac{2\alpha(1-\alpha)^{\frac{1-\alpha}{\alpha}}\log(1.2)}{d}\right)^{-\frac{d}{2}} \leq 1.2\exp(2\alpha(1-\alpha)^{\frac{1-\alpha}{\alpha}}\log(1.2)) \leq 1.2 \times 1.5 = 2.$$

We use the inequality $(1/(1-x))^{1/x} \leq \exp(2)$ if $0 \leq x \leq 0.5$ above. So we obtain $\mathbb{E}[V] \leq 2$.

When $V = \mathbb{E}[\rho|x]$ is small, ρ is likely to be small and thus $V - \overline{V} = \mathbb{E}[(\rho-2)\mathbf{1}_{\rho>2}]$ is also small. We bound the term $\mathbb{E}[(V-\overline{V})^2 V^{-1}]$ by splitting it into

$$\mathbb{E}\frac{(V-\overline{V})^2}{V} \leq \mathbb{E}\frac{(V-\overline{V})^2}{V}\mathbf{1}_{V\leq\frac{1}{e}} + \mathbb{E}\frac{(V-\overline{V})^2}{V}\mathbf{1}_{V>\frac{1}{e}} \leq \mathbb{E}\frac{(V-\overline{V})^2}{V}\mathbf{1}_{V\leq\frac{1}{e}} + e\mathbb{E}(V-\overline{V})^2.$$

Following the same logic of (29) in the proof of Theorem 7, we can get that if

$$\eta \leq \frac{1}{49L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}(2+\log(1+\frac{522}{\zeta}))},$$

then $e\mathbb{E}(V-\overline{V})^2 \leq \zeta/(16)$. We then notice that when $V \leq 1/e$, by the definition of V and Jensen inequality, it holds that $g(x) - \mathbb{E}[g] \geq 1$. To bound the term $\mathbb{E}\frac{(V-\overline{V})^2}{V}\mathbf{1}_{V\leq\frac{1}{e}}$, we write

$$\mathbb{E}\frac{(V-\overline{V})^2}{V}\mathbf{1}_{V\leq\frac{1}{e}} \leq \frac{1}{e}\mathbb{E}\left(\frac{V-\overline{V}}{V}\right)^2\mathbf{1}_{V\leq\frac{1}{e}} \leq \frac{1}{e}\mathbb{E}\left(\frac{V-\overline{V}}{V}\right)^2\mathbf{1}_{g(x)-\mathbb{E}[g]\geq 1}.$$

Moreover, since $g(x) - \mathbb{E}[g]$ is always no less than 1, with $c = 0.2$, we have

$$\Pr(\Delta > r|x) \leq 1.5\exp\left(-c\frac{(g(x)-\mathbb{E}[g]+r)^{\frac{2}{\alpha+1}}}{L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}\eta}\right) \leq 1.5\exp\left(-c\frac{g(x)-\mathbb{E}[g]+r}{L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}\eta}\right) \quad (51)$$

for $\forall r > 0$.

Denote $\bar{\Delta} = \Delta/\log 2 = (g(z)-g(x))/\log 2$, and $S = L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}\eta$. We can write

$$\begin{aligned}
\frac{V-\overline{V}}{V} =& \frac{\mathbb{E}[(\rho-2)\mathbf{1}_{\rho>2}|x]}{\mathbb{E}[\exp(\Delta)|x]} \leq \frac{\mathbb{E}[\exp(\Delta)\mathbf{1}_{\rho>2}|x]}{\mathbb{E}[\exp(\Delta)|x]} = \frac{1}{\mathbb{E}[\exp(\Delta)\mathbf{1}_{\rho\leq 2}|x]/\mathbb{E}[\exp(\Delta)\mathbf{1}_{\rho>2}|x]+1} \\
\leq& \frac{\mathbb{E}[\exp(\Delta)\mathbf{1}_{\rho>2}|x]}{\mathbb{E}[\exp(\Delta)\mathbf{1}_{\rho\leq 2}|x]} \leq \frac{2\sum_{i=1}^\infty \exp(i\log 2)\Pr(i\leq\bar{\Delta}\leq i+1|x)}{\exp(\frac{-1+\mathbb{E}[g]-g(x)}{\log 2})\Pr(\frac{-1+\mathbb{E}[g]-g(x)}{\log 2}\leq\bar{\Delta}\leq 1|x)} \\
\leq& \frac{2\sum_{i=1}^\infty \exp(i\log 2)\Pr(\bar{\Delta}\geq i|x)}{\exp(\frac{-1+\mathbb{E}[g]-g(x)}{\log 2})[1-\Pr(\Delta>\log 2|x)-\Pr(g(z)-\mathbb{E}[g]<-1)]} \\
\overset{(51)}{\leq}& \frac{3\sum_{i=1}^\infty \exp(i\log 2)\exp(-c\frac{g(x)-\mathbb{E}[g]+i\log 2}{S})\exp(\frac{g(x)-\mathbb{E}[g]}{\log 2})}{\exp(-\frac{1}{\log 2})(1-1.5[\exp(-c\frac{g(x)-\mathbb{E}[g]+\log 2}{S})+\exp(-\frac{c}{S})])} \\
\leq& \frac{15\sum_{i=1}^\infty \exp(i(\log 2-\frac{c\log 2}{S}))\exp(-(\frac{c}{S}-\frac{1}{\log 2})(g(x)-\mathbb{E}[g]))}{1-1.5[\exp(-c\frac{g(x)-\mathbb{E}[g]+\log 2}{S})+\exp(-\frac{c}{S})]} \\
=& \frac{15\exp(-(\frac{c}{S}-\frac{1}{\log 2})(g(x)-\mathbb{E}[g]))/(\exp(\frac{c\log 2}{S}-\log 2)-1)}{1-1.5[\exp(-c\frac{g(x)-\mathbb{E}[g]+\log 2}{S})+\exp(-\frac{c}{S})]} \\
\leq& \frac{15\exp(-(\frac{c}{S}-\frac{1}{\log 2}))/(\exp(\frac{c\log 2}{S}-\log 2)-1)}{1-1.5[\exp(-c\frac{1+\log 2}{S})+\exp(-\frac{c}{S})]},
\end{aligned} \quad (52)$$

where we use the condition $g(x) - \mathbb{E}[g] \geq 1$ in the last inequality. By choosing

$$\eta \leq (8L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}\max\{1.5+\log(1+7/\sqrt{\zeta}),\log(30)\}))^{-1},$$

under the condition $g(x) - \mathbb{E}[g] \geq 1$, we guarantee that $1-1.5[\exp(-c\frac{1+\log 2}{S})+\exp(-\frac{c}{S}) \geq 0.8$ and $15\exp(-(\frac{c}{S}-\frac{1}{\log 2}))/(\exp(\frac{c\log 2}{S}-\log 2)-1) \lesssim \zeta$, thus $(\frac{V-\overline{V}}{V})^2 \leq \frac{e\zeta}{16}$. So we obtain that $\mathbb{E}\frac{(V-\overline{V})^2}{V}\mathbf{1}_{V\leq\frac{1}{e}} \leq \frac{\zeta}{16}$, and thus $\mathbb{E}\frac{(V-\overline{V})^2}{V} \leq \frac{\zeta}{8}$.

To bound $\mathbb{E}\overline{V}$, we notice that if $\chi^2_{\pi^{X|Y}}(\hat{\pi}^{X|Y}) \le 0.5$, then $\|\hat{\pi}^{X|Y} - \pi^{X|Y}\|_{\text{TV}} \le 1$, thus by (24), we get $\mathbb{E}\overline{V} \ge \frac{1}{2}$. Combining it with the bounds $\mathbb{E}[V] \le 2$ and $\mathbb{E}\frac{(V-\overline{V})^2}{V} \le \frac{\zeta}{8}$, we get $\chi^2_{\pi^{X|Y}}(\hat{\pi}^{X|Y}) \le \zeta$.

The expected number of the iterations for the rejection sampling step is $\mathcal{O}(1)$ because the acceptance probability is $\frac{1}{2}\mathbb{E}[\overline{V}]$ (Lemma 3). ■

We then consider the case where we can only solve the optimization step 2 inaccurately. The corresponding algorithm and analysis have been discussed in §D. The following theorem extends the results of Theorem 26 and 27 in §D.

Theorem 29 (RGO complexity in χ^2-divergence) *Assume f satisfies (1). For $\forall \zeta > 0$, if*

$$\eta \le \left(98 L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}} (2 + \log(1 + 522/\zeta))\right)^{-1},$$

then Algorithm 3 returns a random point x that has ζ χ^2-divergence to the distribution proportional to $\pi^{X|Y}(\cdot|y)$. Furthermore, if $0 < \zeta < 0.5$, then the algorithm access $\mathcal{O}(1)$ many $f(x)$ in expectation.

Proof Following the proof with accurate optimization, we get

$$\chi^2_{\pi^{X|Y}}(\hat{\pi}^{X|Y}) \le \frac{\mathbb{E}[(V - \overline{V})^2 V^{-1}]\mathbb{E}[V]}{(\mathbb{E}\overline{V})^2}$$

When bounding $\mathbb{E}V$, without loss of generality, we assume w is 0, i.e., the mean of x and z is 0, and $g'(x_y - w) = 0$. Denote $\Delta = g(z) - g(x)$, and $\iota = x_y - w$. Following (8) in the proof of concentration inequality in Theorem 4, we have

$$\begin{aligned}
\mathbb{E}[V] &= \mathbb{E}_{x,z}[\rho] = \mathbb{E}_{x,z}[\exp(\Delta)] \le \mathbb{E}_x[\exp(\pi^2\eta\|\nabla g(x)\|^2/8)] \\
&\overset{(43)}{\le} \left(1 - \frac{\pi^2 L_\alpha^2 \eta^2 \alpha}{4w}\right)^{-\frac{d}{2}} \exp\left(\frac{\pi^2}{8} L_\alpha^2 \eta (1-\alpha) w^{\frac{\alpha}{1-\alpha}}\right) \exp\left(\frac{\|\iota\|_2^2/(2\eta)}{\left(\frac{\pi^2}{4}\eta^2 L_\alpha^2 \frac{\alpha}{\omega}\right)^{-1} - 1}\right)
\end{aligned}$$

for $\forall w > 0$. The only additional term compared to proof of Theorem 28 is the right-most term above. By choosing $w = (1-\alpha)^{-\frac{1-\alpha}{\alpha}}(\eta d)^{1-\alpha}$, and denote $b = (1-\alpha)^{-\frac{1-\alpha}{\alpha}}$, the above equation becomes

$$\left(1 - \frac{\pi^2 L_\alpha^2 \eta^{1+\alpha} \alpha (1-\alpha)^{\frac{1-\alpha}{\alpha}}}{4 d^{1-\alpha}}\right)^{-\frac{d}{2}} \exp\left(\frac{\pi^2}{8} L_\alpha^2 \eta^{1+\alpha} d^\alpha\right) \exp\left(\frac{\|\iota\|_2^2}{\frac{8b^2 d^{1-\alpha}}{\pi^2 \alpha \eta^\alpha L_\alpha^2} - 2\eta}\right)$$

Recall that (44) still holds, i.e. $\|\iota\|^2 \le s := \frac{d^{\frac{1}{2(1+\alpha)}}}{7 L_\alpha^{\frac{1}{1+\alpha}}}$, since we use the same Algorithm 3. We further choose $\eta \le \frac{1}{2 L_\alpha^{\frac{2}{1+\alpha}} d^{\frac{\alpha}{1+\alpha}}} \le \left(\frac{8\log(1.2)}{\pi^2 L_\alpha^2 d^\alpha}\right)^{\frac{1}{1+\alpha}}$, so the above equation is bounded by

$$2\exp\left(\frac{\alpha\pi^2\|\iota\|^2\eta^\alpha L_\alpha^2}{4b^2 d^{1-\alpha}}\right) \le 3.$$

So we obtain $\mathbb{E}[V] \leq 3$. Next, we bound the term $\mathbb{E}[(V-\overline{V})^2 V^{-1}]$:

$$\mathbb{E}\frac{(V-\overline{V})^2}{V} \leq \mathbb{E}\frac{(V-\overline{V})^2}{V}\mathbf{1}_{V\leq\frac{1}{e}} + \mathbb{E}\frac{(V-\overline{V})^2}{V}\mathbf{1}_{V>\frac{1}{e}} \leq \mathbb{E}\frac{(V-\overline{V})^2}{V}\mathbf{1}_{V\leq\frac{1}{e}} + e\mathbb{E}(V-\overline{V})^2.$$

Following the same logic of (49) in the proof of Theorem 27, we can get that if

$$\eta \leq \frac{1}{98L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}(2+\log(1+\frac{522}{\zeta}))},$$

then $e\mathbb{E}(V-\overline{V})^2 \leq \zeta/(24)$.

We then notice that since $V \leq 1/e$, by the definition of V and Jensen inequality, it holds that $g(x) - \mathbb{E}[g] \geq 1$. The term $\mathbb{E}\frac{(V-\overline{V})^2}{V}\mathbf{1}_{V\leq\frac{1}{e}}$ satisfies $\mathbb{E}\frac{(V-\overline{V})^2}{V}\mathbf{1}_{V\leq\frac{1}{e}} \leq \frac{1}{e}\mathbb{E}\left(\frac{V-\overline{V}}{V}\right)^2\mathbf{1}_{g(x)-\mathbb{E}[g]\geq 1}$. Moreover, since $g(x)-\mathbb{E}[g]$ is always no less than 1, we can apply the inequality (45). With $c = 0.2$ and $\|x_y - w\| \leq s$, we have

$$\Pr(\Delta > r|x) \leq 1.5\exp\left(\frac{s^2}{2d\eta} - c\cdot\frac{g(x)-\mathbb{E}[g]+r}{L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}\eta}\right) \tag{53}$$

for $\forall r > 0$.

Denote $\bar{\Delta} = \Delta/\log 2 = (g(z)-g(x))/\log 2$, and $S = L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}\eta$. We can write

$$\begin{aligned}
\frac{V-\overline{V}}{V} &\overset{(52)}{\leq} \frac{2\sum_{i=1}^\infty \exp(i\log 2)\Pr(\bar{\Delta}\geq i|x)}{\exp(\frac{-1+\mathbb{E}[g]-g(x)}{\log 2})[1-\Pr(\Delta>\log 2|x)-\Pr(g(z)-\mathbb{E}[g]<-1)]} \\
&\overset{(53)}{\leq} \frac{3\sum_{i=1}^\infty \exp(i\log 2)\exp(\frac{s^2}{2d\eta} - c\frac{g(x)-\mathbb{E}[g]+i\log 2}{S})\exp(\frac{g(x)-\mathbb{E}[g]}{\log 2})}{\exp(-\frac{1}{\log 2})(1-1.5[\exp(\frac{s^2}{2d\eta}-c\frac{g(x)-\mathbb{E}[g]+\log 2}{S})+\exp(\frac{s^2}{2d\eta}-\frac{c}{S})])} \\
&\leq \frac{15\sum_{i=1}^\infty \exp(i(\log 2 - \frac{c\log 2}{S}))\exp(\frac{s^2}{2d\eta} - (\frac{c}{S}-\frac{1}{\log 2})(g(x)-\mathbb{E}[g]))}{1-1.5[\exp(\frac{s^2}{2d\eta}-c\frac{g(x)-\mathbb{E}[g]+\log 2}{S})+\exp(\frac{s^2}{2d\eta}-\frac{c}{S})]} \\
&= \frac{15\exp(\frac{s^2}{2d\eta} - (\frac{c}{S}-\frac{1}{\log 2})(g(x)-\mathbb{E}[g]))/(\exp(\frac{c\log 2}{S}-\log 2)-1)}{1-1.5[\exp(\frac{s^2}{2d\eta}-c\frac{g(x)-\mathbb{E}[g]+\log 2}{S})+\exp(\frac{s^2}{2d\eta}-\frac{c}{S})]} \\
&\leq \frac{15\exp(\frac{s^2}{2d\eta} - (\frac{c}{S}-\frac{1}{\log 2}))/(\exp(\frac{c\log 2}{S}-\log 2)-1)}{1-1.5[\exp(\frac{s^2}{2d\eta}-c\frac{1+\log 2}{S})+\exp(\frac{s^2}{2d\eta}-\frac{c}{S})]},
\end{aligned}$$

where we use the condition $g(x) - \mathbb{E}[g] \geq 1$ in the last inequality. By by the choice s in (44) and

$$\eta \leq (16L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}\max\{1.5+\log(1+7/\sqrt{\zeta}), \log(30)\})^{-1},$$

under the condition $g(x) - \mathbb{E}[g] \geq 1$, we guarantee that $1 - 1.5[\exp(-c\frac{1+\log 2}{S}) + \exp(-\frac{c}{S}) \geq 0.8$ and $15\exp(-(\frac{c}{S}-\frac{1}{\log 2}))/(\exp(\frac{c\log 2}{S}-\log 2)-1) \lesssim \zeta$. So we obtain that $\mathbb{E}\frac{(V-\overline{V})^2}{V}\mathbf{1}_{V\leq\frac{1}{e}} \leq \frac{\zeta}{24}$, and thus $\mathbb{E}\frac{(V-\overline{V})^2}{V} \leq \frac{\zeta}{12}$.

To bound $\mathbb{E}\overline{V}$, we notice that if $\chi^2_{\pi^{X|Y}}(\hat{\pi}^{X|Y}) \le 0.5$, then $\|\hat{\pi}^{X|Y} - \pi^{X|Y}\|_{\mathrm{TV}} \le 1$, thus by (24), we get $\mathbb{E}\overline{V} \ge \frac{1}{2}$. Combining it with the bounds $\mathbb{E}[V] \le 3$ and $\mathbb{E}\frac{(V-\overline{V})^2}{V} \le \frac{\zeta}{12}$, we get $\chi^2_{\pi^{X|Y}}(\hat{\pi}^{X|Y}) \le \zeta$.

The expected number of the iterations for the rejection sampling step is $\mathcal{O}(1)$ because the acceptance probability is $\frac{1}{2}\mathbb{E}[\overline{V}]$ (Lemma 3). ■

E.2. Complexity bounds of proximal sampling for semi-smooth potentials

The following Proposition 30, 31 extend the results in Proposition 10, 13, 14 to χ^2-divergence notion. Moreover, they only assume we have inaccurate optimization step 2. Our proofs in this section use the same techniques from Altschuler and Chewi (2023, Theorem 5.1-5.4).

Proposition 30 (Convergence in χ^2 for well-conditioned targets) *Suppose f is β-strongly convex and L_1-smooth. Let $\delta \in (0,1), \eta = \tilde{\mathcal{O}}\left(1/(L_1\sqrt{d})\right)$. Then Algorithm 1, with Algorithm 3 as RGO step and initialization $x_0 \sim \mu_0$, can find a random point $x_T \sim \hat{\mu}_T$ such that $\chi^2_\nu(\hat{\mu}_T) \le \delta$ in*

$$T = \mathcal{O}\left(\frac{L_1\sqrt{d}}{\beta}\log\left(\frac{L_1\sqrt{d}}{\beta\delta}\right)\log\left(\frac{\sqrt{R_{2,\nu}(\mu_0)}}{\delta}\right)\right)$$

steps. Furthermore, each step accesses only $\mathcal{O}(1)$ many $f(x)$ queries in expectation.

Proof Recall that the distributions of the iterations y_t and x_t of the ideal proximal sampler are ψ_t and μ_t respectively (see the discussion at the beginning of §4). Denote the Y-marginal of π^{XY} as π^Y. By analyzing the simultaneous heat flow, Chen et al. (2022, §A.4) shows that the forwards step of the proximal algorithm is a contraction in Rényi divergence, i.e.

$$R_{2,\pi^Y}(\psi_t) \le \frac{R_{2,\nu}(\mu_t)}{(1+\eta\beta)^{1/2}}. \tag{54}$$

We assume that for $\forall y \in \mathbb{R}^d$, we have

$$\chi^2_{\pi^{X|Y}(\cdot|y)}(\hat{\pi}^{X|Y}(\cdot|y)) \le \zeta$$

using $O(1)$ many $f(x)$ queries in expectation. Then according to the data-processing inequality for Rényi divergence (Altschuler and Chewi, 2023, Lemma 2.3), strong composition rule for Rényi differential privacy (Altschuler and Talwar, 2022, Lemma 2.9), we have

$$\begin{aligned} R_{2,\nu}(\hat{\mu}_{t+1}) &\le R_{2,\pi^{XY}}(\mathrm{law}(\hat{X}_{t+1}, \hat{Y}_t)) \\ &\le R_{2,\pi^Y}(\hat{\psi}_t) + \sup_{y_t\in\mathbb{R}^d} R_{2,\pi^{X|Y}(\cdot|y_t)}(\hat{\pi}^{X|Y}(\cdot|y_t)) \\ &\le R_{2,\pi^Y}(\hat{\psi}_t) + \log(1+\zeta) \end{aligned} \tag{55}$$

Combining (54) and (55) gives

$$R_{2,\nu}(\hat{\mu}_{t+1}) \le \frac{R_{2,\nu}(\hat{\mu}_t)}{(1+\eta\beta)^{1/2}} + \log(1+\zeta).$$

Iterating this bound for T times gives

$$R_{2,\nu}(\hat{\mu}_T) \le \frac{R_{2,\nu}(\mu_0)}{(1+\eta\beta)^{T/2}} + \log(1+\zeta)\sum_{t=0}^{T-1}\frac{1}{(1+\eta\beta)^{t/2}} \le \frac{R_{2,\nu}(\mu_0)}{(1+\eta\beta)^{T/2}} + \frac{\log(1+\zeta)}{1-\frac{1}{\sqrt{1+\eta\beta}}}$$

This error is at most δ if

$$T = \mathcal{O}\left(\frac{1}{\log(1+\eta\beta)}\log\left(\frac{R_{2,\nu}(\mu_0)}{\delta}\right)\right), \quad \zeta = \Theta(\delta\eta\beta).$$

According to Theorem 29, plugging in the necessary step size $\eta = \mathcal{O}(\frac{1}{L_1\sqrt{d}\log(L_1\sqrt{d}/(\delta\beta))})$ finishes the proof. ■

Proposition 31 (Convergence in χ^2 for non-log-concave targets) *We use Algorithm 1 with Algorithm 3 as RGO step. Let $\delta \in (0,1)$, and $\nu \propto \exp(-f)$.*
1) If ν satisfies C_{LSI}-LSI and L_1-smooth, then we need
$T = \mathcal{O}\left(\frac{L_1\sqrt{d}}{C_{\mathrm{LSI}}}\log\left(\frac{L_1\sqrt{d}}{C_{\mathrm{LSI}}\delta}\right)\log\left(\frac{R_{2,\nu}(\mu_0)}{\delta}\right)\right)$ *to have $\chi^2_\nu(\hat{\mu}_T) \le \delta$.*
2) If ν satisfies C_{PI}-PI and f is L_α-α-semi-smooth, then we need
$T = \mathcal{O}\left(\frac{L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}}{C_{\mathrm{PI}}}\log\left(\frac{L_\alpha^{\frac{2}{\alpha+1}} d^{\frac{\alpha}{\alpha+1}}}{C_{\mathrm{PI}}\delta}\right)\log\left(\frac{\chi^2_\nu(\mu_0)}{\delta}\right)\right)$ *to have $\chi^2_\nu(\hat{\mu}_T) \le \delta$.*
Furthermore, each step accesses only $\mathcal{O}(1)$ many $f(x)$ queries in expectation.

Proof 1) The proof is the same as of Proposition 30 since (54) also holds under LSI:

$$R_{2,\pi^Y}(\psi_t) \le \frac{R_{2,\nu}(\mu_t)}{(1+\eta C_{\mathrm{LSI}})^{1/2}}.$$

2) Chen et al. (2022, §A.4) shows

$$\chi^2_{\pi^Y}(\psi_t) \le \frac{\chi^2_\nu(\mu_t)}{(1+\eta C_{\mathrm{PI}})}. \tag{56}$$

In the same time, (55) still holds as it does not use the LSI assumption. Recall that $R_{2,\nu} = \log(1+\chi^2_\nu)$. Combining (56) and (55) gives

$$\chi^2_\nu(\hat{\mu}_{t+1}) \le \left(\frac{1+\zeta}{1+\eta C_{\mathrm{PI}}}\right)\chi^2_\nu(\hat{\mu}_t) + \zeta.$$

Iterating this bound for T times gives

$$\chi^2_\nu(\hat{\mu}_T) \le \left(\frac{1+\zeta}{1+\eta\beta}\right)^T \chi^2_\nu(\mu_0) + \zeta\sum_{t=0}^{T-1}\left(\frac{1+\zeta}{1+\eta C_{\mathrm{PI}}}\right)^t \le \left(\frac{1+\zeta}{1+\eta\beta}\right)^T \chi^2_\nu(\mu_0) + \frac{\zeta}{1-\frac{1+\zeta}{1+\eta C_{\mathrm{PI}}}}$$

This error is at most δ if we choose

$$T = \mathcal{O}\left(\frac{1}{\log((1+\eta C_{\mathrm{PI}})/(1+\zeta))}\log\left(\frac{\chi^2_\nu(\mu_0)}{\delta}\right)\right), \quad \zeta \le \frac{\delta\eta C_{\mathrm{PI}}}{2(1+\eta C_{\mathrm{PI}}+\delta/2)}.$$

Plugging in the choice of η in Theorem 29 finishes the proof.

■

Remark 32 (Convergence stability) *In the convergence analysis, an interesting question is whether we can allow the number of iterations to go to infinity. Lemma 8, 9 do not guarantee stability since the RGO accumulation error linearly depends on the iterations T. However, the proof of Lemma 9 can be easily strengthened when π is strongly-log-concave to bound accumulation error. Moreover, with the techniques in Theorem 5.1-5.4 of Altschuler and Chewi (2023), our Proposition 30, 31 now ensure stable convergence.*