High-Accuracy Log-Concave Sampling with Stochastic Queries

Fan Chen FANCHEN@MIT.EDU
Massachusetts Institute of Technology

Sinho Chewi SINHO.CHEWI@YALE.EDU
Yale University

Constantinos Daskalakis COSTIS@CSAIL.MIT.EDU
Massachusetts Institute of Technology

Alexander Rakhlin RAKHLIN@MIT.EDU
Massachusetts Institute of Technology

Editors: Steve Hanneke and Tor Lattimore

Abstract

We show that high-accuracy guarantees for log-concave sampling—that is, iteration and query complexities which scale as $\mathrm{poly}\log(1/\delta)$, where δ is the desired target accuracy—are achievable using stochastic gradients with sub-exponential tails. Notably, this exhibits a separation with the problem of convex optimization, where stochasticity (even additive Gaussian noise) in the gradient oracle incurs $\mathrm{poly}(1/\delta)$ queries. We also give an information-theoretic argument that light-tailed stochastic gradients are necessary for high accuracy: for example, in the bounded variance case, we show that the minimax-optimal query complexity scales as $\Theta(1/\delta)$. Our framework also provides similar high-accuracy guarantees under stochastic zeroth-order (value) queries, and an improved complexity result for sampling from finite-sum potentials.

1. Introduction

We study the problem of sampling from a log-concave density $\mu \propto e^{-f}$ given access to a stochastic gradient oracle for f. Our main result shows that if the stochastic gradients are unbiased and have light tails (e.g., sub-exponential), then it is possible to generate a δ-accurate sample in total variation distance in $\mathrm{polylog}(1/\delta)$ queries and time. We refer to such a guarantee as a *high-accuracy* guarantee.

Recent works take inspiration from the close connections between log-concave sampling and the better-understood field of convex optimization. From that standpoint, the phenomenon we highlight here could be surprising. Indeed, it is well-known that optimization in the presence of noisy gradients—even additive Gaussian noise—does not admit high-accuracy guarantees. Information-theoretic lower bounds (Agarwal et al., 2012; Raginsky and Rakhlin, 2011) establish that the optimal bounds are $1/\delta$ in the strongly convex case, and $1/\delta^2$ in the weakly convex case.

On the other hand, in the literature on Markov chain Monte Carlo (MCMC), there are remarkable examples of "exact MCMC" methods in which various components of the algorithm are replaced by unbiased estimates, yet the resulting Markov chain remains geometrically ergodic toward the original stationary distribution. For example, suppose that μ is the marginal distribution over a parameter θ,

but there is an additional latent variable z. In this case, the exact density can be difficult to compute, but unbiased estimators can be produced via importance sampling. Incorporating these estimators into Metropolis–Hastings algorithms leads to the class of pseudo-marginal MCMC methods (Andrieu and Roberts, 2009), some of which are exact.

When f is a finite sum of functions (e.g., negative log-likelihoods in a statistical context), there is a great need to develop samplers which make use of batched stochastic gradients, echoing the stochastic gradient revolution in machine learning. This led to the widespread use of stochastic gradient Langevin dynamics (SGLD) (Welling and Teh, 2011); see Nemeth and Fearnhead (2021) for a survey of recent developments. These methods are based on discretizations of diffusions and are therefore not exact, i.e., they do not admit high-accuracy guarantees. Other works propose minibatch variants of Metropolis–Hastings methods (Seita et al., 2018; Zhang et al., 2020; Wu et al., 2022b), leading to tailored algorithms but often without quantitative convergence guarantees. A notable recent advance is the work of Lee et al. (2021), which developed a high-accuracy sampler for the finite sum setting; subsequent work (Gopi et al., 2022, 2023) developed high-accuracy samplers with stochastic *value* (zeroth-order) queries. We revisit the finite-sum setting in Section 3.4.

Our interest lies in generalizing the above observations to the black-box setting, in which no particular structure for μ is assumed except log-concavity, the starting point of most non-asymptotic analyses (Chewi, 2026), as well as generic properties of the stochastic gradient oracle. In doing so, we aim to provide precise, non-asymptotic guarantees that incorporate recent state-of-the-art advances in sampling theory so that these guarantees are as sharp as possible.

1.1. Contributions

Our main contribution is the development of high-accuracy samplers in the presence of stochastic gradient noise, provided that the stochastic gradients have light tails (e.g., sub-exponential or sub-Gaussian). As a preview of our results, suppose that the target distribution is α-strongly log-concave and β-log-smooth. Then, state-of-the-art guarantees for high-accuracy sampling (Chewi et al., 2021; Wu et al., 2022a; Fan et al., 2023; Altschuler and Chewi, 2024; Chen et al., 2026) have established that it is possible to draw a δ-accurate sample in total variation distance from $\mu \propto e^{-f}$ using

$$\widetilde{O}\big(\kappa d^{1/2}\,\mathrm{polylog}(1/\delta)\big) \qquad \text{queries to exact oracles for } f,\ \nabla f\,,$$

where $\kappa := \beta/\alpha$ is the condition number of μ.

A consequence of our results is that it is in fact possible to draw a sample in

$$\widetilde{O}\big((\kappa d^{1/2} + \sigma^2/\alpha)\,\mathrm{polylog}(1/\delta)\big) \qquad \text{queries to a } \textit{stochastic} \text{ oracle for } \nabla f\,,$$

provided that the unbiased stochastic estimates $g(x)$ of the gradient $\nabla f(x)$ satisfy the sub-exponential tail bound $\mathbb{E}\exp(\|g(x) - \nabla f(x)\|/\sigma) \le 2$. We note that our main results are considerably more general, allowing for both log-concavity and log-smoothness to be relaxed, and covering noisy zeroth-order queries as well; see Section 3.3 for details.

This demonstrates a surprising *robustness to noise* for sampling, in that σ^2/α appears additively in the final bound and does not significantly deteriorate the dependence on the target accuracy δ. As discussed above, this is a stark departure from the corresponding results in optimization, in which stochasticity quickly degrades the rates to $\mathrm{poly}(1/\delta)$ regardless of the tail behavior.

We further remark that the work of Chatterji et al. (2022) established a lower bound of $\Omega(\sigma^2/\delta^2)$ in a certain regime, even under Gaussian additive noise. In Section 4.2, we explain why this does not contradict our results: their lower bound example requires the strong log-concavity parameter α to tend to zero with δ; in fact, $\alpha \lesssim \delta^2$. Hence, our results imply that the $\Omega(\delta^{-2})$-scaling is in fact a consequence of the target distribution being ill-conditioned. However, this raises the question of whether one can prove a lower bound which captures the dependence on δ, even when α remains bounded away from zero.

We resolve this question via a new lower bound that captures how the tail behavior of the stochastic gradients affects the complexity of sampling to high precision. In particular, when we only assume that the stochastic gradients have bounded variance, our lower bound reads $\Omega(1/\delta)$. This is actually attained by our upper bound algorithm in this setting, establishing that the optimal rate is $\Theta(1/\delta)$ under a bounded variance constraint. More generally, if we only assume that finitely many moments of the stochastic gradient are bounded, our lower bound shows that $\Omega(1/\delta^c)$ queries are necessary for some exponent $c > 0$.

Taken together, our results show that *light-tailed stochastic gradients are both necessary and sufficient for high-accuracy sampling*.

Finally, we apply our method to the finite-sum setting $f = m^{-1}\sum_{i=1}^m f_i$ and improve the complexity of high-accuracy sampling from $\widetilde{O}(m + \kappa\,(\sqrt{md} + d))$ (Lee et al., 2021) to $\widetilde{O}(m + \kappa\sqrt{md})$; see Section 3.4 for details.

2. Preliminaries

We first define the stochastic gradient/value oracle and its tail behavior.

Assumption 2.1 (Stochastic gradient oracle) *For any $x \in \mathbb{R}^d$, we can draw i.i.d. samples from a distribution $\mathsf{O}_{\mathsf{grad}}(x)$ such that under $g \sim \mathsf{O}_{\mathsf{grad}}(x)$, it holds that $\mathbb{E}[g] = \nabla f(x)$.*

We assume that there is a parameter $\mathfrak{m}_1 > 0$ such that $\mathbb{E}_{g\sim\mathsf{O}_{\mathsf{grad}}(x)}\|g - \nabla f(x)\| \le \mathfrak{m}_1$ for any $x \in \mathbb{R}^d$.

Assumption 2.2 (Stochastic value oracle) *For any $x \in \mathbb{R}^d$, we can draw i.i.d. samples from a distribution $\mathsf{O}_{\mathsf{eval}}(x)$ such that under $v \sim \mathsf{O}_{\mathsf{eval}}(x)$, it holds that $\mathbb{E}[v] = f(x)$.*

For any stochastic oracle O and integer $n \ge 1$, we define $O^{(n)}$ to be the *batch oracle* that, given input $x \in \mathbb{R}^d$, returns $y = \frac{1}{n}\sum_{i=1}^n y^i$ by generating i.i.d. samples $y^1, \dots, y^n \sim O(x)$.

Definition 2.3 (Oracle with ϵ-tail) *Suppose that $\epsilon = (\epsilon_n)_{n\ge1}$ is a sequence of functions. We say an oracle O is of ϵ-tail if for any $x \in \mathbb{R}^d$, $M > 0$, $n \ge 1$, it holds that under $g \sim O^{(n)}(x)$,*

$$\frac{1}{M}\,\mathbb{E}\big[\|g - \mathbb{E}[g]\|\,\mathbb{I}\{\|g - \mathbb{E}[g]\| > M\}\big] \le \epsilon_n(M; x)\,.$$

We also denote $\epsilon_n(M) \coloneqq \sup_{x\in\mathbb{R}^d}\epsilon_n(M; x)$.

Some cases of interest are as follows.

Example 2.4 (Sub-polynomial tail) *Suppose that for some parameter $\zeta > 0$ and $\sigma_{\mathsf{g}} > 0$, for any x, under $g \sim O(x)$, $\mathbb{E}\exp\big(\frac{\|g-\mathbb{E}[g]\|^\zeta}{\sigma_{\mathsf{g}}^\zeta}\big) \le 2$.*[1]

Then, we can choose $\epsilon_1(M) \le C_\zeta \exp(-c_\zeta(M/\sigma_{\mathsf{g}})^\zeta)$ for $M \ge \sigma_{\mathsf{g}}$. More generally, we can choose $\epsilon_n(M) \le C_\zeta \exp(-c_\zeta(\sqrt{n}M/\sigma_{\mathsf{g}})^{\bar\zeta})$ where $\bar\zeta := \min\{\zeta, 2\}$.

Example 2.5 (Polynomial tail) *Suppose that for some $k \ge 1$ and any x, $\mathbb{E}\|g - \mathbb{E}[g]\|^{2k} \le \sigma_{2k}^{2k}$. Then we can choose $\epsilon_n(M) \le \frac{(2k)!\sigma_{2k}^{2k}}{n^k M^{2k}}$.*

We carry out our analysis under the following Hölder continuity assumption for ∇f. It interpolates between the Lipschitz case ($s = 0$) and the smooth case ($s = 1$).

Assumption 2.6 (Hölder continuous gradient) *There exists $s \in [0, 1]$ and $\beta_s \ge 0$ such that $\|\nabla f(x) - \nabla f(y)\| \le \beta_s \|x - y\|^s$ for all $x, y \in \mathbb{R}^d$.*

For technical convenience, we also state our results using an approximate proximal oracle.

Assumption 2.7 (Approximate proximal oracle) *Given input x_0, the oracle $\mathsf{O}_{\mathsf{prox},\eta}(x_0)$ returns $\widehat{x}$ such that $\|\widehat{x} + \eta\nabla f(\widehat{x}) - x_0\| \le \eta\varepsilon_{\mathsf{prox}}$.*

Alternatively, if we assume that the guarantee in the assumption holds with high probability, then our results remain unchanged up to another error term in total variation distance. The following lemma shows that the approximate proximal oracle can be implemented using the stochastic gradient oracle.

Lemma 2.8 *Suppose that Assumption 2.6 holds with $s \in [0, 1]$ and denote $m_s = \beta_s^{1/(1+s)}$. Suppose that $\eta \le \frac{1}{2m_s}$ and we are given access to a stochastic gradient oracle with ϵ-tail.*

Then, as long as the input x_0 satisfies $\|\nabla f(x_0)\| \le G$, the approximate proximal oracle with $\varepsilon_{\mathsf{prox}} = 10(m_s + M)$ can be implemented with probability at least $1 - \epsilon_n(M)$ using $O(n\log(G/(m_s + M)))$ queries to the stochastic gradient oracle.

Notation. For any function $f : \mathbb{R}^d \to \mathbb{R}$ such that $Z_f := \int_{\mathbb{R}^d} e^{-f(x)}dx < +\infty$, we define μ_f to be the distribution over $\mathbb{R}^d$ with density $\mu_f(x) = \frac{1}{Z_f}e^{-f(x)}$.

For $B > 0$, we write $\mathsf{Clip}_B(\cdot) := \max\{-B, \min\{B, \cdot\}\}$ and $\tau_B(\cdot) := (|\cdot| - B)_+$. We use $\lesssim$ and $O(\cdot)$ to hide absolute constants, i.e., $f \lesssim g$ (and $f = O(g)$) if there is an absolute constant such that $f \le Cg$. The notation $\widetilde{O}(\cdot)$ hides logarithmic factors.

3. High-accuracy sampling with stochastic queries

We build up to our results in three steps. Our methods build upon first-order rejection sampling (FORS), a meta-algorithm recently developed in Chen et al. (2026) which simulates rejection sampling given unbiased estimators of the log-density ratio between the proposal and target. Therefore, we first review the FORS framework in Section 3.1. Then, in Section 3.2, we instantiate FORS for the problem of sampling from a Gaussian tilt distribution, thereby showing that the results of Chen et al. (2026) are robust to stochastic gradient noise. Finally, in Section 3.3, we combine the results of Section 3.2 with the proximal sampler algorithm (Lee et al., 2021; Chen et al., 2022) to establish our main results for log-concave sampling.

1. The case $\zeta = 2$ corresponds to a sub-Gaussian tail, and $\zeta = 1$ to a sub-exponential tail.

3.1. Background on first-order rejection sampling (FORS)

To motivate the FORS algorithm, we replicate the motivating example of Chen et al. (2026) here. Consider the simple problem of sampling from a density $p \propto e^{-f}$, where $f : [0,1] \to \mathbb{R}$, $f(0) = 0$, and $-1 \le f' \le 1$. In order to perform rejection sampling with the base measure $\mathsf{Unif}([0,1])$, we must generate randomness $b \sim \mathsf{Ber}(ce^{-f(x)})$ for any given $x \in [0,1]$. To do so, one typically assumes access to evaluations of f itself. The novelty of FORS lies in recognizing that this is unnecessary—it suffices to produce *unbiased estimators* of f.

A more general version of this idea is known as the "Bernoulli factory" problem (Keane and O'Brien, 1994; Nacu and Peres, 2005), and variants of this idea can be found in multiple domains (e.g., Wagner, 1988; Papaspiliopoulos, 2011). It can be stated as the following abstract task:

Task: Given i.i.d. random variables $W_1, W_2, W_3, \ldots$ in $[-1,1]$, generate a sample $b \sim \mathsf{Ber}(ce^{\mathbb{E}\,W_1})$.

To solve this, write the Taylor series as

$$e^{\mathbb{E}\,W_1} = e^{-1} \cdot e^{\mathbb{E}[1+W_1]} = \sum_{j \ge 0} \frac{e^{-1}}{j!} \left(\mathbb{E}[1+W_1]\right)^j .$$

Suppose that $J \sim \mathsf{Poisson}(2)$ is independent of the i.i.d. sequence $W_1, W_2, W_3, \ldots$. Then we notice that

$$e^{\mathbb{E}\,W_1} = e\,\mathbb{E}\Big[\prod_{j=1}^{J} \big(\frac{1+W_j}{2}\big)\Big] ,$$

so we can set $b \sim \mathsf{Ber}\big(\prod_{j=1}^{J}\big(\frac{1+W_j}{2}\big)\big)$. Indeed, $\mathbb{P}(b=1) = \mathbb{E}\prod_{j=1}^{J}\big(\frac{1+W_j}{2}\big) = e^{-1+\mathbb{E}\,W_1}$.

In summary, to generate a sample $b \sim \mathsf{Ber}(ce^{-f(x)})$, it suffices to have access to (a random number of) unbiased estimates of $f(x)$. In Chen et al. (2026), this was leveraged to produce high-accuracy samplers which only use queries to the *derivative* f', via the representation $f(x) = \mathbb{E}_{y \sim \mathsf{Unif}([0,x])}[x f'(y)]$. In our work, our goal is to leverage this phenomenon in order to produce high-accuracy samplers that *tolerate stochasticity in the gradient oracle*.

We are now ready to state the general FORS meta-algorithm. Given a proposal distribution q, a tilt function w, and a tuneable parameter $B = \Theta(1)$, the goal of Algorithm 1 is to produce a sample from $\widehat{p}(x) \propto q(x)\, e^{w(x)}$ without having access to the *value* $w(x)$. Instead, for each $x \in \mathbb{R}^d$, we can generate i.i.d. samples $W_1, W_2, W_3 \ldots$ such that $\mathbb{E}[W_1 \mid x] = w(x)$. Let $\mathcal{W}_x$ denote the conditional distribution of W_1 given x.

Theorem 3.1 (FORS guarantee, Chen et al. (2026, Theorem 3.1)) *Algorithm 1 outputs a random point with density $\widehat{p}(x) \propto q(x)\, e^{\mathbb{E}[W_1 | x]}$. The number of sampled W_j's is bounded, with probability at least $1-\delta$, by $3Be^{2B}\log(2/\delta)$.*

Moreover, if Algorithm 1 is called T times, then with probability at least $1-\delta$, the total number of sampled W_j's is $O(Be^{2B}\,(T + \log(1/\delta)))$.

Algorithm 1 First-order rejection sampling (FORS)

Input: Parameter $B > 0$, proposal distribution q over $\mathbb{R}^d$, estimator distributions $(\mathcal{W}_x)_{x\in\mathbb{R}^d}$ supported on $[-B, B]$
for $i = 1, 2, 3, \ldots$ **do**
  Sample $x \sim q$.
  Sample $J \sim \mathsf{Poisson}(2B)$.
  Sample i.i.d. $W_1, \ldots, W_J \sim \mathcal{W}_x$.
  Output x with probability $\prod_{j=1}^{J} \frac{B+W_j}{2B}$.
end for

3.2. Sampling from Gaussian tilts with stochastic queries

The goal of this section is to sample from the following Gaussian tilt distribution:

$$\nu(x) \propto \exp\Big(-f(x) - \frac{1}{2\eta}\|x - x_0\|^2\Big)\,. \tag{1}$$

In Section 3.3, this will be used as a subroutine for the proximal sampler algorithm (Lee et al., 2021; Chen et al., 2022), leading to new guarantees for log-concave sampling.

Remark 3.2 (Diffusion models) *Leveraging the fact that the backward transition kernels along a diffusion model are also Gaussian tilts of the form (1), Chen et al. (2026) used FORS to provide the first* high-accuracy *sampling guarantees for diffusion models under minimal data assumptions. Similarly, the results we present below could also be applied to that setting to show that diffusion sampling can be made* robust to stochastic errors in the score evaluations. *For brevity, we do not pursue this application here.*

We now consider instantiating FORS for the Gaussian tilt distribution (1). Let $\gamma_{z,r}(x) := \gamma(x; z, r)$ be any *path function* such that $\gamma_{z,1}(x) = x$ and $\gamma_{z,0}(x) = \overline{\gamma}(z)$ is independent of x; here, $z \sim P$ is an external source of randomness. Then, by the fundamental theorem of calculus,

$$f(x) - \mathbb{E}_{z\sim P} f(\overline{\gamma}(z)) = \mathbb{E}_{r\sim\mathsf{Unif}([0,1]),\, z\sim P}\langle \dot{\gamma}_{z,r}(x), \nabla f(\gamma_{z,r}(x))\rangle\,.$$

If we choose the proposal $q = \mathsf{N}(\widehat{x}, \eta I)$, where $\widehat{x}$ is a fixed base point chosen so that $q \approx \nu$ (made precise in Theorem 3.3), then $q(x) \propto \exp\big(-\frac{1}{2\eta}\|x - \widehat{x}\|^2\big)$, and hence

$$\begin{aligned}\log \nu(x) - \log q(x) &= -f(x) - \frac{1}{2\eta}\|x - x_0\|^2 + \frac{1}{2\eta}\|x - \widehat{x}\|^2 + \text{const} \\ &= \frac{1}{\eta}\langle x_0 - \widehat{x}, x\rangle - f(x) + \text{const}\,.\end{aligned}$$

Thus, applying the path integral formula to $h(x) = \eta^{-1}\langle x_0 - \widehat{x}, x\rangle - f(x)$, we can express

$$\log \nu(x) - \log q(x) = \mathbb{E}_{r\sim\mathsf{Unif}([0,1]),\, z\sim P}\Big\langle \dot{\gamma}_{z,r}(x),\, \eta^{-1}(x_0 - \widehat{x}) - \nabla f(\gamma_{z,r}(x))\Big\rangle + \text{const}\,.$$

By the guarantee of Theorem 3.1, it suggests that we use the unbiased estimator $W_{r,z,x} := \langle \dot{\gamma}_{z,r}(x), u - \nabla f(\gamma_{z,r}(x))\rangle$, with $u := (x_0 - \widehat{x})/\eta$. Actually, since the W's in Algorithm 1 must lie

in $[-B, B]$, we truncate the estimator to lie in this range. Further, we replace the exact gradients by stochastic gradients, leading to

$$\widehat{W}_{r,z,g,x} \coloneqq \mathsf{Clip}_B(\langle \dot\gamma_{z,r}(x), u - g\rangle)\,, \qquad g \sim \mathsf{O}_{\mathsf{grad}}(\gamma_{z,r}(x))\,.$$

Below, we choose the base point of the proposal $\widehat{x}$, the path function $\gamma_{z,r}$, and the noise distribution P in order to optimize the dimension dependence of our result.

Theorem 3.3 (Sampling from Gaussian tilts) *Suppose that Assumption 2.6 holds, $\mathsf{O}_{\mathsf{grad}}(\cdot)$ has ϵ-tail, $n \geq 1$, and $B = \Theta(1)$.*

Instantiate Algorithm 1 as follows:

- *$q = \mathsf{N}(\widehat{x}, \eta I)$, where $\widehat{x}$ is drawn from $\mathsf{O}_{\mathsf{prox},\eta}(x_0)$. We write $u \coloneqq \frac{x_0 - \widehat{x}}{\eta}$.*
- *$\mathcal{W}_x$ is the law of $\mathsf{Clip}_B(W_{r,z,g,x})$, where*

$$W_{r,z,g,x} = \langle \dot\gamma_{z,r}(x), u - g\rangle\,, \quad r \sim \mathsf{Unif}([0,1])\,,\ z \sim \mathsf{N}(0, \eta I)\,,\ g \sim \mathsf{O}^{(n)}_{\mathsf{grad}}(\gamma_{z,r}(x))\,, \tag{2}$$

and

$$\gamma_{z,r,x_0}(x) = a_r x + (1 - a_r)\widehat{x} + b_r z\,, \qquad a_r = \sin(\pi r/2)\,,\ \ b_r = \cos(\pi r/2)\,, \tag{3}$$

so that $\dot\gamma_{z,r,x_0}(x) = a_r'(x - \widehat{x}) + b_r' z$.

Then, conditioned on $\|u - \nabla f(\widehat{x})\|^2 \leq \varepsilon^2_{\mathsf{prox}}$ and

$$\eta^{-1} \gg \left(\beta_s^2 d^s \log(1/\delta) + \frac{s\beta_s^2}{d^{1-s}} \log^2(1/\delta)\right)^{1/(1+s)} + (M^2 + \varepsilon^2_{\mathsf{prox}}) \log(1/\delta)\,,$$

the law $\widehat{\nu}$ of Algorithm 1 satisfies $D_{\mathsf{TV}}(\nu, \widehat{\nu}) \leq \delta + C\,\mathbb{E}_{x\sim\nu} \min\{\epsilon_n(M; x), 1\}$, where C is an absolute constant.

In our application to log-concave sampling, η will be interpreted as a step size, and hence the overall complexity of sampling will scale with η^{-1}, multiplied by the batch size n and other distribution-specific pre-factors. We pause to give several remarks to elucidate the dependencies in this result.

Remark 3.4 (Dimension dependence) *The first term requires taking an inverse step size $\eta^{-1} \gg \beta_0^2$ in the Lipschitz case ($s = 0$), and $\eta^{-1} \gg \beta_1 d^{1/2}$ in the smooth case ($s = 1$). This matches state-of-the-art results for high-accuracy sampling (Fan et al., 2023; Altschuler and Chewi, 2024; Chen et al., 2026), except that we allow for stochastic gradient queries.*

Remark 3.5 (Proximal tolerance) *Since the theorem already requires taking $\eta^{-1} \gg \beta_0^2 + M^2$ in the Lipschitz case, and $\eta^{-1} \gg M^2$ in the smooth case, then Lemma 2.8 (with $n = 1$) implies that implementing the approximate proximal oracle with the stochastic gradient oracle only incurs a* logarithmic overhead.

Remark 3.6 (Accuracy dependence) *To reach a final error of δ, we need to take η^{-1} at least of order $M^2 = \epsilon_n^{-1}(\delta)^2$. We elucidate this in two cases of particular interest.*

- *(Sub-Gaussian tails) Suppose that the stochastic gradients have sub-Gaussian tails, which corresponds to $\zeta = 2$ in Example 2.4. Then, we can take $n = 1$ and $M^2 \asymp \sigma_{\mathsf{g}}^2 \log(1/\delta)$, thus the final term requires $\eta^{-1} \gg \sigma_{\mathsf{g}}^2 \log^2(1/\delta)$. Hence, this leads to a* high-accuracy *guarantee.*

- *(Bounded variance) Suppose now that the stochastic gradients merely have variance bounded by σ^2. Since $\epsilon_1(M) \lesssim \sigma^2/M^2$, we can choose $M^2 \asymp \sigma^2/\delta$. Thus, the dependence on δ becomes $\eta^{-1} \gg \delta^{-1}\log(1/\delta)$. Although it suffices to take $n = 1$ here, to avoid error accumulation in the next section we will eventually have to apply batching ($n > 1$). After doing so, the iteration complexity remains $1/\delta$ (up to logarithmic factors).*

  In Section 4, we will show that the $1/\delta$ rate is in fact optimal *under the bounded variance assumption (Proposition 4.2). Thus, high-accuracy sampling requires light-tailed stochastic gradients.*

Parallel to Theorem 3.3, we show that it is also possible to sample from the Gaussian tilt distribution with only stochastic value queries, provided that the error of the stochastic value oracle is sufficiently small.

Theorem 3.7 *Suppose that Assumption 2.6 holds, and $\mathsf{O}_{\mathsf{eval}}(\cdot)$ has ϵ-tail. Suppose that $n \geq 1$ and $B = \Theta(1)$.*

Instantiate Algorithm 1 as follows:

- *$q = \mathsf{N}(\widehat{x}, \eta I)$, where $\widehat{x}$ is drawn from $\mathsf{O}_{\mathsf{prox},\eta}(x_0)$. We write $u = \frac{x_0 - \widehat{x}}{\eta}$.*
- *$\mathcal{W}_x$ is the law of $\mathsf{Clip}_B(W_{z,v,v',x})$, where*

$$W_{z,v,v',x} = v' - v - \langle u, x - z\rangle\,, \qquad z \sim q\,,\ v \sim \mathsf{O}_{\mathsf{eval}}^{(n)}(x)\,,\ v' \sim \mathsf{O}_{\mathsf{eval}}^{(n)}(z)\,. \tag{4}$$

Then, conditioned on $\|u - \nabla f(\widehat{x})\| \leq \varepsilon_{\mathsf{prox}}$, the law $\widehat{\nu}$ of Algorithm 1 satisfies $D_{\mathsf{TV}}(\nu, \widehat{\nu}) \leq \delta + C\,\mathbb{E}_{x\sim\nu} \min\{\epsilon_n(B/4; x), 1\}$, provided that

$$\eta^{-1} \gg \left(\beta_s^2 d^s \log(1/\delta) + \frac{s\beta_s^2}{d^{1-s}} \log^2(1/\delta)\right)^{1/(1+s)} + \varepsilon_{\mathsf{prox}}^2 \log(1/\delta).$$

We note that in general, implementing the proximal oracle with only noisy queries will incur additional computational cost. However, with the choice of $\widehat{x} = x_0$, it is trivially guaranteed that $\varepsilon_{\mathsf{prox}} = \|\nabla f(x_0)\|$. In this case, as we will see in Theorem 3.12 below, the term $\varepsilon_{\mathsf{prox}}^2$ in fact dominates the complexity.

3.3. Log-concave sampling

To apply our results to log-concave sampling (and beyond), we apply the results of the previous section to the proximal sampler algorithm (Lee et al., 2021; Chen et al., 2022). Given a target distribution $\mu \propto e^{-f}$, the proximal sampler aims to sample from the augmented distribution

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

It does so by applying Gibbs sampling to $\bar{\pi}$. Concretely, for $n = 0, 1, 2, \ldots$ and an initial point $X_0 \sim \mu_0$, repeat:

1. Sample $Y_n \sim \bar{\pi}^{Y|X=X_n} = \mathsf{N}(X_n, \eta I)$.
2. Sample $X_{n+1} \sim \bar{\pi}^{X|Y=Y_n}$.

The distribution $\bar{\pi}^{X|Y=y}$ is known as the *restricted Gaussian oracle* (RGO), and it is exactly the Gaussian tilt distribution (1) with $x_0 = y$. We therefore combine our result in Theorem 3.3 for implementing the RGO, together with existing results on the convergence of the proximal sampler itself, to deduce the following sampling corollaries. We begin by recalling the definitions of functional inequalities.

Definition 3.8 (Poincaré) *A distribution π satisfies a* Poincaré inequality (PI) *with constant C if for all compactly supported and smooth test functions $h : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathrm{Var}_{X\sim\pi}(h(X)) \le C\,\mathbb{E}_{X\sim\pi}[\|\nabla h(X)\|^2]\,.$$

We let $C_{\mathsf{PI}}(\pi)$ be the smallest constant C such that π satisfies PI with constant C.

Definition 3.9 (Log-Sobolev) *A distribution π satisfies a* log-Sobolev inequality (LSI) *with constant C if for all compactly supported and smooth test functions $h : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathrm{Ent}_{X\sim\pi}(h^2(X)) \coloneqq \mathbb{E}_{X\sim\pi}\left[h^2(X)\log\frac{h^2(X)}{\mathbb{E}_{X\sim\pi}[h^2(X)]}\right] \le 2C\,\mathbb{E}_{X\sim\pi}[\|\nabla h(X)\|^2]\,.$$

We let $C_{\mathsf{LSI}}(\pi)$ be the smallest constant C such that π satisfies LSI with constant C.

It is well-known (e.g., Bakry et al., 2014) that if π is α-strongly log-concave (SLC), i.e., $-\log\pi$ is α-strongly convex, then it satisfies LSI with constant $1/\alpha$, and if π satisfies LSI with constant $1/\alpha$, then it satisfies PI with constant $1/\alpha$. These represent meaningful enlargements of the class of SLC measures which still allow for tractable sampling. For example, unlike SLC, LSI is robust to bounded perturbations of the log-density; and unlike LSI, PI allows for capturing measures without sub-Gaussian tails (e.g., the two-sided exponential). See Chewi (2026) for further background in the context of sampling. In addition, recent progress toward the KLS conjecture (Klartag, 2023) implies $C_{\mathsf{PI}}(\pi) \le O(\log d)\cdot\|\mathbb{E}_\pi[XX^\top]\|_{\mathrm{op}}$ under log-concavity of π.

We now present a suite of results by combining Theorem 3.3 (Theorem 3.7) with the guarantees of the proximal sampler (Chen et al., 2022). Let $\phi_M(\delta) \coloneqq \inf\{n \ge 1 : \epsilon_n(M) \le \delta/(10C)\}$. We note that this can be relaxed to the "in-distribution error":

$$\phi_{M,N}(\delta) \coloneqq \inf\left\{n \ge 1 : \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_{x\sim\mu_k}\min\{\epsilon_n(M;x),1\} \le \delta/(10C)\right\},$$

where μ_k is the distribution of the X_k in the exact proximal sampler.

Theorem 3.10 *Suppose that Assumption 2.6 holds for some $s \in [0,1]$, and that $\mathsf{O}_{\mathsf{grad}}(\cdot)$ has ϵ-tail. Suppose that we are given an initial distribution μ_0 such that $\log(1 + D_{\chi^2}(\mu_0 \parallel \mu)) \le \Delta$.*

Choose

$$\frac{1}{C\eta} = \left(\beta_s^2 d^s\log(N/\delta) + \beta_s^2\log^2(N/\delta)\right)^{1/(1+s)} + M^2\log(N/\delta) \tag{5}$$

for a sufficiently large universal constant $C > 0$. Let $\widehat{\mu}$ denote the law of the output of the proximal sampler initialized at μ_0, where in each step the RGO is implemented by Theorem 3.3. Then, the proximal sampler ensures $D_{\mathsf{TV}}(\widehat{\mu},\mu) \le \delta$ using at most $N\phi_M(\delta/N)\log A$ queries to $\mathsf{O}_{\mathsf{grad}}(\cdot)$ in the following situations.

1. *Suppose that μ satisfies a log-Sobolev inequality with constant $C_{\mathsf{LSI}}(\mu) < \infty$ and $s = 1$ (i.e., f is smooth). Then,*

$$N \lesssim C_{\mathsf{LSI}}(\mu) \left(\beta_1 d^{1/2} \log^{3/2} A + (\beta_1 + M^2) \log^2 A\right),$$

where $A := d + \Delta + \delta^{-1} + C_{\mathsf{LSI}}(\mu)\,(\beta_1 + M^2)$.

2. *Suppose that μ satisfies a Poincaré inequality with constant $C_{\mathsf{PI}}(\mu) < \infty$. Then,*

$$N \lesssim C_{\mathsf{PI}}(\mu) \left((\beta_s^2 d^s \log A + \beta_s^2 \log^2 A)^{1/(1+s)} + M^2 \log A\right)(\Delta + \log(1/\delta)),$$

where $A := d + \Delta + \delta^{-1} + C_{\mathsf{PI}}(\mu)\,(\beta_s^{2/(1+s)} + M^2)$.

3. *Suppose that μ is log-concave. Then,*

$$N \lesssim \left((\beta_s^2 d^s \log A + \beta_s^2 \log^2 A)^{1/(1+s)} + M^2 \log A\right) \cdot \frac{W_2^2(\mu_0, \mu)}{\delta^2},$$

where $A := d + \Delta + \delta^{-1} + (\beta_s^{2/(1+s)} + M^2)\, W_2^2(\mu_0, \mu)$.

Remark 3.11 (Dependence on δ) *As an illustration, we describe the implied query complexity in the following special cases.*

- *If the stochastic gradients have subexponential tails (Example 2.4 with $\zeta = 1$), then for $M \geq \sigma_{\mathsf{g}}$, we can take $\phi_M(\delta) \lesssim (\sigma_{\mathsf{g}}/M)\log(1/\delta)$. Therefore, we can take $M \asymp \sigma_{\mathsf{g}}$ in all of the results above, and the iteration complexity equals $\widetilde{O}(N)$. For example, in the smooth LSI case, the query complexity reads $\widetilde{O}(C_{\mathsf{LSI}}(\beta_1 d^{1/2} + \sigma_{\mathsf{g}}^2))$.*
- *On the other hand, in the bounded variance case (Example 2.5 with $k = 1$), we can take $\phi_M(\delta) \lesssim \sigma_2^2/(\delta M^2)$, and the total query complexity becomes $(N \vee N^2\sigma_2^2/(\delta M^2))\log A$. We then choose M to balance the terms. For example, in the smooth LSI case, the query complexity reads*

$$N \log A + C_{\mathsf{LSI}}\left[(\beta_1^2 d/M^2)\log^3 A + M^2 \log^4 A\right](\sigma_2^2 \log A)/\delta,$$

This leads to a total query complexity of $\widetilde{O}(\kappa d^{1/2}\,(1 + C_{\mathsf{LSI}}\sigma_2^2/\delta))$, where $\kappa = C_{\mathsf{LSI}}\beta_1$ is the condition number.

We emphasize that while sampling guarantees with stochastic gradients are well-studied (e.g., Dalalyan, 2017; Dalalyan and Karagulyan, 2019; Durmus et al., 2019; Balasubramanian et al., 2022; Huang et al., 2024; Lu et al., 2025), our contribution is to provide *high-accuracy* guarantees, provided that the stochastic gradients have light tails.

In the next section, we show that the assumption on the tails of the stochastic gradient is necessary.

We also provide a corresponding result for stochastic value queries.

Theorem 3.12 *Suppose that Assumption 2.6 holds and that $\mathsf{O}_{\mathsf{eval}}(\cdot)$ has ϵ-tail. Suppose that we are given an initial distribution μ_0 such that $\log(1 + D_{\chi^2}(\mu_0 \parallel \mu)) \leq \Delta$.*

Choose

$$\frac{1}{C\eta} = (\beta_s d^s)^{2/(1+s)} \left(1 + \frac{\Delta + \log(N/\delta)}{d}\right) \log(N/\delta) \tag{6}$$

for a sufficiently large universal constant $C > 0$. Let $\widehat{\mu}$ denote the law of the output of the proximal sampler initialized at μ_0, where in each step the RGO is implemented by Theorem 3.7. Then, the proximal sampler ensures $D_{\mathsf{TV}}(\widehat{\mu}, \mu) \le \delta$ using at most $N\phi_1(\delta/(4N))$ queries to $\mathsf{O}_{\mathsf{eval}}(\cdot)$ in the following situations.

1. *Suppose that μ satisfies a log-Sobolev inequality with constant $C_{\mathsf{LSI}}(\mu) < \infty$ and $s = 1$ (i.e., f is smooth). Then,*

$$N \lesssim C_{\mathsf{LSI}}(\mu)\, \beta_1\, (d + \Delta + \log A) \log^2 A\,,$$

   where $A \coloneqq d + \Delta + \delta^{-1} + C_{\mathsf{LSI}}(\mu)\, \beta_1$.

2. *Suppose that μ satisfies a Poincaré inequality with constant $C_{\mathsf{PI}}(\mu) < \infty$. Then,*

$$N \lesssim C_{\mathsf{PI}}(\mu)\, (\beta_s d^s)^{2/(1+s)} \left(1 + \frac{\Delta + \log A}{d}\right) (\Delta + \log(1/\delta)) \log A\,,$$

   where $A \coloneqq d + \Delta + \delta^{-1} + C_{\mathsf{PI}}(\mu)\, \beta_s^{2/(1+s)}$.

3. *Suppose that μ is log-concave. Then,*

$$N \lesssim (\beta_s d^s)^{2/(1+s)} \left(1 + \frac{\Delta + \log A}{d}\right) \log A \cdot \frac{W_2^2(\mu_0, \mu)}{\delta^2}\,,$$

   where $A \coloneqq d + \Delta + \delta^{-1} + \beta_s^{2/(1+s)} W_2^2(\mu_0, \mu)$.

Note that under noisy value queries of sub-Gaussian tail (Example 2.4), it holds that $\epsilon_n(1) \le e^{-n/\sigma_{\mathsf{g}}^2}$ for $n \gg \sigma_{\mathsf{g}}^2$, and hence $\phi_1(\delta) = O(\sigma_{\mathsf{g}}^2 \log(1/\delta) + 1)$. Thus, assuming that f is α-strongly convex and β-smooth, the query complexity (roughly) scales as $\widetilde{O}(\kappa d \cdot \max\{\sigma_{\mathsf{g}}^2, 1\})$, where $\kappa = \beta/\alpha$ is the condition number. By reduction to zeroth-order optimization, it is expected that in this setting, sublinear dependence on d cannot be achieved.

3.4. Application: finite-sum sampling

We now consider a slightly more abstract formulation of the finite-sum sampling problem. For simplicity, we focus on the smooth setting.

Assumption 3.13 *The function f takes the form $f(x) = \mathbb{E}_{w \sim P} F(x; w)$, and computing $\nabla F(x; w)$ requires unit cost. Furthermore, $\|\nabla F(x; w) - \nabla F(x'; w)\| \le \beta_1 \|x - x'\|$ for all w and $x, x' \in \mathbb{R}^d$.*

Theorem 3.14 *Suppose that Assumption 3.13 holds and that the initial distribution μ_0 satisfies $\log(1 + D_{\chi^2}(\mu_0 \,\|\, \mu)) \le \Delta$. Consider implementing the proximal sampler as follows.*

- *Initialize $X_0 \sim \mu_0$.*
- *For each $k \ge 0$, sample $Y_k \sim \mathsf{N}(X_k, \eta I)$.*
- *If $k \bmod K = 0$, query $\widehat{X}_{k+1} \sim \mathsf{O}_{\mathsf{prox},\eta}(Y_k)$ and compute $\nabla f(X_k)$. Otherwise, set $\widehat{X}_{k+1} \coloneqq \widehat{X}_{m(k)+1} + Y_k - Y_{m(k)}$, where $m(k) \coloneqq K\lfloor k/K \rfloor$.*
- *Let $O_{k+1}(x)$ denote the distribution of $\nabla F(x; w) - \nabla F(X_{m(k)}; w) + \nabla f(X_{m(k)})$ under $w \sim P$.*

- *Instantiate Theorem 3.3 with oracle O_{k+1} and center $\widehat{X}_{k+1}$ to generate X_{k+1}.*

Then, assuming that each call to the proximal oracle $\mathsf{O}_{\mathsf{prox},\eta}(\cdot)$ succeeds with probability at least $1-\delta$ and

$$\frac{1}{\beta_1\eta} \gg \sqrt{Kd} + K^{2/3}(d+\Delta+\log(K/\delta))^{1/3} + (\varepsilon_{\mathsf{prox}}^2/\beta_1+1)\log(K/\delta), \tag{7}$$

the distribution $\widehat{\mu}_N$ of our algorithm satisfies $D_{\mathsf{TV}}(\mu_N,\widehat{\mu}_N)\le N\delta$.

Note that when $P=\mathsf{Unif}([m])$, i.e., $f(x)=\frac{1}{m}\sum_{i=1}^m f_i(x)$, evaluating $\nabla f(x)$ has cost m, and the proximal oracle can be implemented via standard SVRG methods in $\widetilde{O}(m)$ time. Therefore, assuming $\Delta=\widetilde{O}(d)$, we can choose $K=m$ to obtain a query complexity of $\widetilde{O}(m+\kappa\,(\sqrt{md}+m^{2/3}d^{1/3}))$, where $\kappa := C_{\mathsf{LSI}}(\mu)\beta_1$. This improves upon the $\widetilde{O}(m+\kappa\,(\sqrt{md}+d))$ complexity achieved in Lee et al. (2021) in the regime $m\le d$. Moreover, by combining the two results, i.e., using our result for $m\le d$ and their result for $m>d$, it yields an overall bound of $\widetilde{O}(m+\kappa\sqrt{md})$.

4. Lower bound: light tails are necessary for high-accuracy sampling

4.1. A simple lower bound

We establish lower bounds for sampling with stochastic gradient queries under oracles with bounded ψ-moment.

Definition 4.1 *Let $\psi:[0,+\infty)\to[0,+\infty)$ be an increasing function such that $\psi(0)=0$. An oracle $\mathsf{O}_{\mathsf{grad}}(\cdot)$ is a ψ-oracle for f if for any $x\in\mathbb{R}^d$, under $g\sim\mathsf{O}_{\mathsf{grad}}(x)$, it holds that $\mathbb{E}[g]=\nabla f(x)$ and $\mathbb{E}\,\psi(\|g-\nabla f(x)\|)\le 1$.*

In the following, we present a simple information-theoretic argument based on the goal of sampling from a one-dimensional Gaussian $p_\theta=\mathsf{N}\big(\theta,\frac{1}{\alpha}I\big)$. Here $p_\theta(x)\propto\exp(-f_\theta(x))$ with $f_\theta(x)=\frac{\alpha}{2}(x-\theta)^2$ and $\nabla f_\theta(x)=\alpha(x-\theta)$. Since f_θ is α-strongly convex, the LSI holds with $C_{\mathsf{LSI}}(p_\theta)\le\frac{1}{\alpha}$.

Proposition 4.2 (Lower bound) *Fix any increasing function $\psi:[0,+\infty)\to[0,+\infty)$ such that $\psi(0)=0$. Suppose that $T\ge 1$ and $\delta\in(0,1]$, and there is an algorithm $\mathtt{Alg}$ such that for any $\theta\in\{0,\delta/\sqrt{\alpha}\}$, given any ψ-oracle O for f_θ, return a sample $x\sim\mathtt{Alg}(O)$ using T queries to O and $D_{\mathsf{TV}}(p_\theta,\mathtt{Alg}(O))\le\frac{\delta}{10}$. Then, it holds that*

$$T\ge\frac{1}{10\sqrt{\alpha}}F_\psi(\sqrt{\alpha}\delta)\,,\qquad F_\psi(\theta):=\sup\{u\ge\theta:(1-\psi(\theta))\cdot u\ge\theta\cdot\psi(u)\}\,.$$

Remark 4.3 *As demonstration, we describe the implied query complexity lower bound in the following special cases.*

- *Consider the case where the only assumption on $\mathsf{O}_{\mathsf{grad}}(\cdot)$ is that the variance is bounded by σ^2. Then we can take $\psi(m)=\frac{m^2}{\sigma^2}$ and $F_\psi(\theta)=\frac{\sigma^2}{\theta}-\theta$ for any $\theta\le\frac{\sigma}{2}$, i.e., $\Omega(\sigma^2/(\alpha\delta))$ queries are necessary for stochastic gradients with only bounded second moment. Thus, in this case, our upper bounds are optimal (Remark 3.11), at least with respect to the dependence on δ.*
- *More generally, for $\psi(m)=(m/\sigma)^s$ with $s>1$, we have $F_\psi(\delta)\asymp\frac{\sigma^{s/(s-1)}}{\delta^{1/(s-1)}}$, i.e., if the only assumption on $\mathsf{O}_{\mathsf{grad}}(\cdot)$ is a bounded s-th moment, Proposition 4.2 yields a lower bound of*

$\Omega\big(\frac{(\sigma/\sqrt{\alpha})^{s/(s-1)}}{\delta^{1/(s-1)}}\big)$ *queries. In particular, taking* $s \to 1$ *implies that it is intractable to sample with stochastic gradients with only bounded first moment.*

- *For* $\mathsf{O}_{\mathsf{grad}}(\cdot)$ *with sub-exponential tail, we take* $\psi(m) = e^{(m/\sigma)^\zeta} - 1$ *with* $\zeta > 0$*. Then* $F_\psi(\delta) \geq \Omega(\sigma \log^{1/\zeta}(\sigma/\delta))$*, and Proposition 4.2 yields a lower bound of* $\Omega\big(\frac{\sigma}{\sqrt{\alpha}} \log^{1/\zeta} \frac{\sigma}{\sqrt{\alpha}\delta}\big)$*. On the other hand, in this case, the argument of Chatterji et al. (2022) yields an alternate lower bound of* $\Omega(\frac{\sigma^2}{\alpha})$ *in this case.*

Remark 4.4 (Dimensional dependence) *For light-tailed (e.g., sub-exponential) stochastic gradients, our upper bound scales as* $\widetilde{O}(\kappa d^{1/2} + \sigma^2/\alpha)$*, whereas our lower bound does not scale with* d*. The* $\kappa d^{1/2}$ *term is therefore not captured by the lower bound; since it is independent of the variance proxy* σ*, it reflects the baseline cost of sampling even with an exact oracle. It may be possible to reduce this term, as the best-known lower bound for exact-oracle sampling scales only as* $\min\{\sqrt{\kappa}, d\}$ *(Chewi et al., 2023). However, closing this gap remains a long-standing open question.*

4.2. Revisiting the lower bound of Chatterji et al. (2022)

Here, we discuss the lower bound of Chatterji et al. (2022), which also applies to sampling with stochastic gradients, in order to avoid potential misunderstandings.

Their main lower bound shows that there is a strongly log-concave and log-smooth distribution, with condition number $\kappa = O(1)$, such that it requires $\Omega(\sigma^2/\delta^2)$ queries to reach δ error in TV distance. Here, σ^2 is the variance of the stochastic gradients.[2] This appears to contradict our upper bound, which only requires $O(1/\delta)$ queries in the bounded variance case. Moreover, inspection of their lower bound reveals that it holds when the stochastic gradient oracle is produced by adding Gaussian noise; in particular, the stochastic gradients have sub-Gaussian tails. In such a setting, we have produced algorithms whose complexity scales as $O(\mathrm{polylog}(1/\delta))$.

To resolve this apparent contradiction, we remark that Chatterji et al. (2022, Theorem 4.1) requires taking the strong log-concavity parameter $\alpha \lesssim \delta^2$. Our upper bounds, which generally incur a dependence of σ^2/α, therefore match their $\Omega(\sigma^2/\delta^2)$ lower bound for their hard examples up to logarithmic terms. However, their lower bounds do not address the question of what the best dependence on δ is, provided that α is bounded away from zero. This is the reason why we proved Proposition 4.2.

We leave it as an open question to prove a more general lower bound which captures the dependence, not just on δ, but on other problem parameters such as κ and d. In the case of exact oracle access, proving lower bounds for sampling remains notoriously challenging, with existing results providing sharp characterizations only for Gaussians or in low dimension (Chewi et al., 2022, 2023).

5. Conclusion

In this work, we have shown that high-accuracy guarantees—$\mathrm{polylog}(1/\delta)$ rates—are achievable for sampling, provided that the stochastic gradients have light tails. Moreover, via an information-theoretic argument, we have shown that light tails are necessary for such a result. In fact, as a by-product of our analysis, we identified that the optimal dependence is $\Theta(1/\delta)$ if the stochastic

2. A variance bound of σ^2 in our convention corresponds to a variance bound of $\sigma^2 d$ in theirs.

gradients are only assumed to have a bounded variance. We then improved the state-of-the-art for high-accuracy sampling from finite-sum potentials.

Several open questions remain, of which we list two: (1) Can the lower bound be extended to capture dependence on other problem parameters, such as the dimension d? (2) What is the optimal complexity in the finite-sum setting?

Acknowledgments

We thank Sam Power for bringing to our attention useful references. We acknowledge support from AFOSR through award FA9550-25-1-0375, Simons Foundation and the NSF through awards DMS-2031883 and PHY-2019786, and DARPA AIQ award. CD is supported by a Simons Investigator Award, a Simons Collaboration on Algorithmic Fairness, ONR MURI grant N00014-25-1-2116, and ONR grant N00014-25-1-2296.

References

Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Trans. Inform. Theory*, 58(5):3235–3249, 2012.

Jason M. Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. *J. ACM*, 71(3), 6 2024.

Christophe Andrieu and Gareth O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37(2):697–725, 2009.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham, 2014.

Krishna Balasubramanian, Sinho Chewi, Murat A. Erdogdu, Adil Salim, and Matthew S. Zhang. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for Langevin Monte Carlo. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2896–2923. PMLR, 7 2022.

Niladri S. Chatterji, Peter L. Bartlett, and Philip M. Long. Oracle lower bounds for stochastic gradient sampling algorithms. *Bernoulli*, 28(2):1074–1092, 2022.

Fan Chen, Sinho Chewi, Constantinos Daskalakis, and Alexander Rakhlin. High-accuracy sampling for diffusion models and log-concave distributions. *arXiv preprint arXiv:2602.01338*, 2026.

Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2984–3014. PMLR, 7 2022.

Sinho Chewi. *Log-concave sampling*. Forthcoming, 2026. Available online at https://chewisinho.github.io/.

Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1260–1300. PMLR, 8 2021.

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 Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2041–2059. PMLR, 7 2022.

Sinho Chewi, Jaume De Dios Pont, Jerry Li, Chen Lu, and Shyam Narayanan. Query lower bounds for log-concave sampling. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 2139–2148, 2023.

Arnak S. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 678–689. PMLR, 7 2017.

Arnak S. Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Process. Appl.*, 129(12):5278–5311, 2019.

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

Jiaojiao Fan, Bo Yuan, and Yongxin Chen. Improved dimension dependence of a proximal algorithm for sampling. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 1473–1521. PMLR, 7 2023.

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

Sivakanth Gopi, Yin Tat Lee, Daogao Liu, Ruoqi Shen, and Kevin Tian. Algorithmic aspects of the log-Laplace transform and a non-Euclidean proximal sampler. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2399–2439. PMLR, 2023.

Xunpeng Huang, Difan Zou, Hanze Dong, Yian Ma, and Tong Zhang. Faster sampling via stochastic gradient proximal sampler. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20559–20596. PMLR, 7 2024.

MS Keane and George L. O'Brien. A Bernoulli factory. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 4(2):213–219, 1994.

Bo'az Klartag. Logarithmic bounds for isoperimetry and slices of convex sets. *arXiv preprint arXiv:2303.14938*, 2023.

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

Jianfeng Lu, Xuda Ye, and Zhennan Zhou. Mean square error analysis of stochastic gradient and variance-reduced sampling algorithms. *arXiv preprint 2511.04413*, 2025.

¸Serban Nacu and Yuval Peres. Fast simulation of new coins from old. *Ann. Appl. Probab.*, 15(1A): 93–115, 2005.

Christopher Nemeth and Paul Fearnhead. Stochastic gradient Markov chain Monte Carlo. *J. Amer. Statist. Assoc.*, 116(533):433–450, 2021.

Omiros Papaspiliopoulos. Monte Carlo probabilistic inference for diffusion processes: a methodological framework. In *Bayesian time series models*, pages 82–103. Cambridge Univ. Press, Cambridge, 2011.

Maxim Raginsky and Alexander Rakhlin. Information-based complexity, feedback and dynamics in convex programming. *IEEE Transactions on Information Theory*, 57(10):7036–7056, 2011.

Daniel Seita, Xinlei Pan, Haoyu Chen, and John Canny. An efficient minibatch acceptance test for Metropolis–Hastings. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5359–5363. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

Wolfgang Wagner. Monte Carlo evaluation of functionals of solutions of stochastic differential equations. Variance reduction and numerical examples. *Stochastic Anal. Appl.*, 6(4):447–468, 1988.

Max Welling and Yee-Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, ICML'11, pages 681–688, Madison, WI, USA, 2011. Omnipress.

Keru Wu, Scott Schmidler, and Yuansi Chen. Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling. *Journal of Machine Learning Research*, 23(270): 1–63, 2022a.

Tung-Yu Wu, Y. X. Rachel Wang, and Wing H. Wong. Mini-batch Metropolis–Hastings with reversible SGLD proposal. *J. Amer. Statist. Assoc.*, 117(537):386–394, 2022b.

Ruqi Zhang, A. Feder Cooper, and Christopher M. De Sa. Asymptotically optimal exact minibatch Metropolis–Hastings. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19500–19510. Curran Associates, Inc., 2020.

Contents

Appendix A. Implementation of the approximate proximal oracle

In this section, we show how to implement the approximate proximal oracle (Assumption 2.7) using a stochastic gradient oracle.

Proof of Lemma 2.8. Our goal is to approximately compute

$$\widehat{x} \approx \arg\min_{x\in\mathbb{R}^d}\big(f(x) + \frac{1}{2\eta}\|x - x_0\|^2\big)\,.$$

Let $x^\star$ be an optimal solution to the minimization problem. We denote $m_s \coloneqq \beta_s^{1/(1+s)}$ and assume that $\eta \le \frac{1}{2m_s}$.

We consider the following linearized update rule: Let $X_0 = x_0$, and

$$X_{k+1} = \frac{X_k - \eta g_k + x_0}{2}\,, \quad g_k \sim \mathsf{O}^{(n)}_{\mathsf{grad}}(X_k)\,, \qquad \forall k \ge 0\,.$$

Note that $x^\star + \eta \nabla f(x^\star) = x_0$. Therefore,

$$\begin{aligned} 2\|X_{k+1} - x^\star\| &= \|X_k - \eta g_k - x^\star + \eta \nabla f(x^\star)\| \\ &\le \|X_k - x^\star - \eta(\nabla f(X_k) - \nabla f(x^\star))\| + \eta \|g_k - \nabla f(X_k)\| \,. \end{aligned}$$

In the following we denote

$$\Delta_k \coloneqq \|X_k - x^\star\| \,, \qquad E_k \coloneqq \|g_k - \nabla f(X_k)\| \,.$$

Note that by Assumption 2.6,

$$\begin{aligned} \|X_k - x^\star - \eta(\nabla f(X_k) - \nabla f(x^\star))\| &\le \|X_k - x^\star\| + \eta \|\nabla f(X_k) - \nabla f(x^\star)\| \\ &\le \Delta_k + \eta \beta_s \Delta_k^s \,. \end{aligned}$$

When $s \in (0,1)$, we can use AM–GM inequality to derive $\eta \beta_s \Delta^s \le \frac{1}{2}\Delta + \eta m_s$. It is also straightforward to verify this inequality holds for $s \in \{0,1\}$.

Then, it holds that

$$\Delta_{k+1} \le \frac{\frac{3}{2}\Delta_k + \eta(m_s + E_k)}{2} \,, \qquad \forall k \ge 0 \,.$$

Applying this inequality recursively, we get

$$\Delta_k \le \Big(\frac{3}{4}\Big)^k \Delta_0 + 2\eta m_s + \frac{\eta}{2} \sum_{i=1}^{k} \Big(\frac{3}{4}\Big)^{i-1} E_{k-i} \,.$$

Note that $\mathbb{P}(Y \ge 2y) \le \frac{1}{y}\,\mathbb{E}(Y - y)_+$ for $y > 0$, and hence

$$\begin{aligned} \mathbb{P}\Big(\Delta_k \ge 2\Big(\frac{3}{4}\Big)^k \Delta_0 + 4\eta(m_s + M)\Big) &\le \frac{1}{8M} \sum_{i=1}^{k} \Big(\frac{3}{4}\Big)^{i-1} \mathbb{E}(E_{k-i} - M)_+ \\ &\le \frac{1}{2}\epsilon_n(M) \,. \end{aligned}$$

Note that

$$\begin{aligned} \Delta_0 = \|x_0 - x^\star\| = \eta \|\nabla f(x^\star)\| &\le \eta \|\nabla f(x_0)\| + \eta \|\nabla f(x_0) - \nabla f(x^\star)\| \\ &\le \eta G + \eta \beta_s \|x_0 - x^\star\|^s \le \eta G + \eta m_s + \frac{1}{2}\Delta_0, \end{aligned}$$

and hence $\Delta_0 \le 2\eta G + 2\eta m_s$. In particular, when $k \ge 10 \log(4G/(M + m_s))$, we know

$$\mathbb{P}(\Delta_k \ge 5\eta(m_s + M)) \le \epsilon_n(M) \,.$$

Finally, note that

$$\begin{aligned} \|X_k + \eta \nabla f(X_k) - x_0\| &\le \|X_k - x^\star\| + \eta \|\nabla f(X_k) - \nabla f(x^\star)\| \\ &\le \Delta_k + \eta \beta_s \Delta_k^s \le \frac{3}{2}\Delta_k + \eta m_s, \end{aligned}$$

and hence we know

$$\mathbb{P}(\|X_k + \eta \nabla f(X_k) - x_0\| \ge 10\eta(m_s + M)) \le \epsilon_n(M) \,.$$

□

Appendix B. Technical tools

Lemma B.1 *Let $\lambda > 0$, $B > 0$. Then*

$$\mathbb{E}_{Z\sim\mathsf{N}(0,\sigma^2)}[e^{\lambda(|Z|-B)_+} - 1] \le 2e^{\frac{1}{2}\lambda^2\sigma^2}\,.$$

Further, when $B \ge 2\max\{\lambda\sigma^2, \sigma\}$, we can bound

$$\mathbb{E}_{Z\sim\mathsf{N}(0,\sigma^2)}[e^{\lambda(|Z|-B)_+} - 1] \le e^{-\frac{B^2}{8\sigma^2}}.$$

Proof. By rescaling, we may assume $\sigma = 1$. Then, we can upper bound

$$\begin{aligned}\mathbb{E}_{Z\sim\mathsf{N}(0,\sigma^2)}[e^{\lambda(|Z|-B)_+} - 1] &\le \sqrt{\frac{2}{\pi}}\int_B^\infty e^{\lambda(z-B)-\frac{1}{2}z^2}\,dz \\ &= \sqrt{\frac{2}{\pi}}\int_B^\infty e^{\frac{1}{2}\lambda^2-\lambda B-\frac{1}{2}(z-\lambda)^2}\,dz = 2e^{\frac{1}{2}\lambda^2-\lambda B}\Phi(B-\lambda)\,,\end{aligned}$$

where $\Phi(w) = \frac{1}{\sqrt{2\pi}}\int_w^\infty e^{-\frac{1}{2}z^2}\,dz$. The first inequality then follows from $\Phi(B-\lambda) \le 1$. Further, using the inequality $\Phi(w) \le \frac{1}{\sqrt{2\pi}w}e^{-\frac{1}{2}w^2}$ for $w > 0$, we can bound

$$\begin{aligned}\mathbb{E}_{Z\sim\mathsf{N}(0,\sigma^2)}[e^{\lambda(|Z|-B)_+} - 1] &\le 2e^{\frac{1}{2}\lambda^2-\lambda B}\Phi(B-\lambda) \\ &\le e^{\frac{1}{2}\lambda^2-\lambda B-\frac{1}{8}B^2} \le e^{-\frac{1}{8}B^2}\,.\end{aligned}$$

□

The following lemma is standard (Chen et al., 2026).

Lemma B.2 *Suppose that $\eta > 0$ and $0 \le \lambda \le \frac{d^{1-s}}{4s\eta^s}$. Then, it holds that*

$$\mathbb{E}_{W\sim\mathsf{N}(0,\eta I)}\exp\bigl(\lambda\|W\|^{2s}\bigr) \le \exp(2(\eta d)^s\lambda)\,.$$

For two probability measures μ, ν, and $\ell > 1$, we write $D_\ell(\mu\,\|\,\nu) \coloneqq \mathbb{E}_\mu(\frac{\mu}{\nu})^{\ell-1} - 1$.

Lemma B.3 *For any $\ell > 1$, it holds that*

$$\max\{D_\ell(\mu_f\,\|\,\mu_g), D_\ell(\mu_g\,\|\,\mu_f)\} \le \mathbb{E}_{x\sim\mu_f}[e^{2\ell\,|f(x)-g(x)|} - 1]\,. \tag{8}$$

Furthermore, it holds that

$$D_{\mathsf{TV}}(\mu_f, \mu_g) \le \mathbb{E}_{x\sim\mu_f}[e^{2\,|f(x)-g(x)|} - 1]\,. \tag{9}$$

Proof. By definition, we can write

$$Z_g = \int_{\mathbb{R}^d} e^{-f(x)+(f(x)-g(x))}\,dx = Z_f\cdot\mathbb{E}_{x\sim\mu_f}e^{f(x)-g(x)},$$

and hence

$$\frac{\mu_f(x)}{\mu_g(x)} = \exp(g(x) - f(x))\frac{Z_g}{Z_f} = e^{g(x)-f(x)}\,\mathbb{E}_{\mu_f}[e^{f-g}]\,.$$

Therefore, we have

$$\begin{aligned}1 + D_\ell(\mu_f \,\|\, \mu_g) = \mathbb{E}_{\mu_f}\big(\frac{\mu_f}{\mu_g}\big)^{\ell-1} &= \Big(\mathbb{E}_{\mu_f}[e^{f-g}]\Big)^{\ell-1} \cdot \mathbb{E}_{\mu_f}[e^{(\ell-1)(g-f)}] \\ &\le \big(\mathbb{E}_{\mu_f}\, e^{(\ell-1)|f-g|}\big)^2 \le \mathbb{E}_{\mu_f}[e^{2\ell|f-g|}]\,.\end{aligned}$$

Similarly,

$$\begin{aligned}1 + D_\ell(\mu_g \,\|\, \mu_f) = \mathbb{E}_{\mu_f}\big(\frac{\mu_g}{\mu_f}\big)^{\ell} &= \Big(\mathbb{E}_{\mu_f}[e^{f-g}]\Big)^{-\ell} \cdot \mathbb{E}_{\mu_f}[e^{\ell(f-g)}] \le \mathbb{E}_{\mu_f}[e^{-\ell(f-g)}] \cdot \mathbb{E}_{\mu_f}[e^{\ell(f-g)}] \\ &\le \big(\mathbb{E}_{\mu_f}\, e^{\ell|f-g|}\big)^2 \le \mathbb{E}_{\mu_f}[e^{2\ell|f-g|}]\,.\end{aligned}$$

Combining both inequalities completes the proof of the first inequality.

To prove the second inequality, we note that

$$\begin{aligned}2D_{\mathsf{TV}}(\mu_f, \mu_g) = \mathbb{E}_{\mu_f}\Big|\frac{\mu_g}{\mu_f} - 1\Big| &= \mathbb{E}_{x\sim\mu_f}\Big|\frac{e^{f(x)-g(x)}}{\mathbb{E}_{\mu_f}[e^{f-g}]} - 1\Big| \\ &\le \frac{1}{\mathbb{E}_{\mu_f}[e^{f-g}]}\left(\mathbb{E}_{\mu_f}|e^{f-g} - 1| + |\mathbb{E}_{\mu_f}[e^{f-g}] - 1|\right) \\ &\le \frac{2}{\mathbb{E}_{\mu_f}[e^{f-g}]}\,\mathbb{E}_{\mu_f}|e^{f-g} - 1|\,.\end{aligned}$$

Note that $|e^w - 1| \le e^{|w|} - 1$ and $\frac{1}{\mathbb{E}_{\mu_f}[e^{f-g}]} \le \mathbb{E}_{\mu_f}[e^{g-f}] \le \mathbb{E}_{\mu_f}[e^{|f-g|}]$, so we can deduce

$$D_{\mathsf{TV}}(\mu_f, \mu_g) \le \mathbb{E}_{\mu_f}[e^{|f-g|}]\left(\mathbb{E}_{\mu_f}[e^{|f-g|}] - 1\right) \le \mathbb{E}_{\mu_f}[e^{2|f-g|} - 1]\,.$$

□

Lemma B.4 *For any $f : \mathcal{X} \to [0, 1]$, it holds that*

$$\mathbb{E}_p[f] - M\,\mathbb{E}_q[f] \le \inf_{\lambda>1} M^{-(\lambda-1)}\left(1 + D_\lambda(p \,\|\, q)\right).$$

Proof. By definition,

$$\mathbb{E}_p[f] - M\,\mathbb{E}_q[f] = \mathbb{E}_q\Big[\Big(\frac{dp}{dq} - M\Big)\,f\Big] \le \mathbb{E}_q\Big(\frac{dp}{dq} - M\Big)_+\,.$$

Note that for any random variable $Y \ge 0$, we have

$$\mathbb{E}(Y - M)_+ = \mathbb{E}[\mathbb{I}\{Y > M\}\,(Y - M)_+] \le \mathbb{P}(Y > M)^{1-\frac{1}{\lambda}}\Big(\mathbb{E}[(Y - M)_+^\lambda]\Big)^{\frac{1}{\lambda}} \le \frac{\mathbb{E}[Y^\lambda]}{M^{\lambda-1}}\,.$$

Combining these inequalities and taking infimum over $\lambda > 1$ completes the proof. □

Lemma B.5 (Sub-additivity for TV distance) *Suppose that $X_1 \to \cdots \to X_T$ is a Markov chain. Given a family of transition kernels $\rho = (\rho_t : \mathcal{X} \to \Delta(\mathcal{X}))_{t\in[T]}$, we let $\mathbb{P}_\rho$ be the law of $X_1, \dots, X_T$ under $X_1 \sim \rho_1$, $X_t \sim \rho_t(\cdot \mid X_{t-1})$. Then, for any families of transition kernels ρ, ρ', it holds that*

$$D_{\mathsf{TV}}\big(\mathbb{P}_\rho, \mathbb{P}_{\rho'}\big) \le \sum_{t=1}^{T} \mathbb{E}_{X_{t-1}\sim\mathbb{P}_\rho}\big[D_{\mathsf{TV}}\big(\rho_t(\cdot \mid X_{t-1}), \rho'_t(\cdot \mid X_{t-1})\big)\big]\,,$$

where we regard $X_0 = \perp$ and $\rho_1(\cdot \mid \perp) = \rho_1$.

Lemma B.6 *Suppose that Y is a random variable such that $\mathbb{E}[Y] = 0$. Then for any $B > 0$, $X \in \mathbb{R}$,*

$$|X - \mathbb{E}\,\mathsf{Clip}_B(X+Y)| \le \tau_{B/2}(X) + \min\{2B, \mathbb{E}\,\tau_{B/2}(Y)\}.$$

Proof. First, note that $|\mathsf{Clip}_B(X) - \mathsf{Clip}_B(X+Y)| \le 2B$, we know

$$\begin{aligned}|X - \mathbb{E}\,\mathsf{Clip}_B(X+Y)| &\le |X - \mathsf{Clip}_B(X)| + |\mathbb{E}[\mathsf{Clip}_B(X) - \mathsf{Clip}_B(X+Y)]| \\ &\le \tau_B(X) + 2B.\end{aligned}$$

On the other hand, we know $X = \mathbb{E}[X+Y]$, and hence

$$\begin{aligned}|X - \mathbb{E}\,\mathsf{Clip}_B(X+Y)| &\le \mathbb{E}|X+Y - \mathsf{Clip}_B(X+Y)| \\ &= \mathbb{E}\,\tau_B(X+Y) \le \tau_{B/2}(X) + \mathbb{E}\,\tau_{B/2}(Y).\end{aligned}$$

Combining both inequalities completes the proof. □

Appendix C. Proofs from Section 3

C.1. Proof of Theorem 3.3

Without loss of generality we only consider the case $n = 1$. We denote by P_x the joint distribution of (r, z, g) under (2).

By Theorem 3.1, the output of Algorithm 1 with the specified choices samples from $\widehat{\nu}$, such that

$$\log\widehat{\nu}(x) - \log q(x) = \text{const} + \mathbb{E}_{(r,z,g)\sim P_x}\,\mathsf{Clip}_B(W_{r,z,g,x})\,.$$

We denote $\overline{W}_{r,z,x} \coloneqq \langle\dot\gamma_{z,r}(x), u - \nabla f(\gamma_{z,r}(x))\rangle$, and we know

$$W_{r,z,g,x} - \overline{W}_{r,z,x} = \langle\dot\gamma_{z,r}(x), \nabla f(\gamma_{z,r}(x)) - g\rangle$$

has mean zero under $g \sim \mathsf{O}_{\mathsf{grad}}(\gamma_{z,r}(x))$. On the other hand, we know

$$\begin{aligned}\mathbb{E}_{r,z}\,\overline{W}_{r,z,x} = \mathbb{E}_{r,z}\langle\dot\gamma_{z,r}(x), u - \nabla f(\gamma_{z,r}(x))\rangle &= -f(x) + \langle u, x\rangle + \text{const} \\ &= \log\nu(x) - \log q(x) + \text{const}\,.\end{aligned}$$

Then, using Lemma B.6, we have

$$\begin{aligned}&|\log\nu(x) - \log\widehat{\nu}(x) - \text{const}| \\ &\le \mathbb{E}_{r,z}\Big|\overline{W}_{r,z,x} - \mathbb{E}_{g\sim\mathsf{O}_{\mathsf{grad}}(\gamma_{z,r}(x))}\,\mathsf{Clip}_B(W_{r,z,g,x})\Big| \\ &\le \mathbb{E}_{r,z}\,\tau_{B/2}(\overline{W}_{r,z,x}) + \mathbb{E}_{r,z}\min\Big\{2B, \mathbb{E}_{g\sim\mathsf{O}_{\mathsf{grad}}(\gamma_{z,r}(x))}\,\tau_{B/2}(W_{r,z,g,x} - \overline{W}_{r,z,x})\Big\} =: V(x).\end{aligned}$$

Then, using Lemma B.3,

$$D_{\mathsf{TV}}(\nu, \widehat{\nu}) \le \mathbb{E}_{x\sim\widehat{\nu}}[e^{2V(x)} - 1] \le e^{2B}\, \mathbb{E}_{x\sim q}[e^{2V(x)} - 1]. \tag{10}$$

where the second inequality uses $\frac{d\widehat{\nu}}{dq}(x) \le e^{2B}$ for $x \in \mathbb{R}^d$. Using the definition of V, $2e^{u+v} \le e^{2u} + e^{2v}$ and the convexity of $w \mapsto e^w$, we also know

$$\begin{aligned} 2e^{2V(x)} \le\ & \mathbb{E}_{r,z} \exp\big(2\tau_{B/2}(\langle \dot\gamma_{z,r}(x), u - \nabla f(\gamma_{z,r}(x))\rangle)\big) \\ & + \mathbb{E}_{r,z} \exp\Big(2\min\Big\{2B, \mathbb{E}_{g\sim \mathsf{O}_{\mathsf{grad}}(\gamma_{z,r}(x))}\, \tau_{B/2}(\langle \dot\gamma_{z,r}(x), \nabla f(\gamma_{z,r}(x)) - g\rangle)\Big\}\Big). \end{aligned}$$

Now, note that for any fixed $r \in [0,1]$ under $x \sim q = \mathsf{N}(\widehat{x}, \eta I)$ and $z \sim P = \mathsf{N}(0, \eta I)$, $[\gamma_{z,r}(x); \dot\gamma_{z,r}(x)]$ are jointly distributed as

$$[\gamma_{z,r}(x); \dot\gamma_{z,r}(x)] \sim \mathsf{N}\left(\begin{bmatrix}\widehat{x}\\ 0\end{bmatrix}, \begin{bmatrix}\eta I & \\ & (\pi/2)^2\eta I\end{bmatrix}\right). \tag{11}$$

This implies that

$$\begin{aligned} & 2(1 + e^{-2B} D_{\mathsf{TV}}(\nu, \widehat{\nu})) \le 2\, \mathbb{E}_{x\sim q}\, e^{2V(x)} \\ \le\ & \mathbb{E}_r\, \mathbb{E}_{x\sim q,\, z\sim\mathsf{N}(0,\eta I)} \exp\big(\tau_{B/2}(\langle \dot\gamma_{z,r}(x), u - \nabla f(\gamma_{z,r}(x))\rangle)\big) \\ & + \mathbb{E}_r\, \mathbb{E}_{x\sim q,\, z\sim\mathsf{N}(0,\eta I)} \exp\Big(\min\Big\{2B, \mathbb{E}_{g\sim \mathsf{O}_{\mathsf{grad}}(\gamma_{z,r}(x))}\, \tau_{B/2}(\langle \dot\gamma_{z,r}(x), \nabla f(\gamma_{z,r}(x)) - g\rangle)\Big\}\Big) \\ =\ & \mathbb{E}_{x\sim q,\, Z_1\sim\mathsf{N}(0,(\pi/2)^2\eta I)} \exp\big(\tau_{B/2}(\langle Z_1, u - \nabla f(Z)\rangle)\big) \\ & + \mathbb{E}_{x\sim q,\, Z_1\sim\mathsf{N}(0,(\pi/2)^2\eta I)} \exp\Big(\min\Big\{2B, \mathbb{E}_{g\sim \mathsf{O}_{\mathsf{grad}}(x)}\, \tau_{B/2}(\langle Z_1, \nabla f(x) - g\rangle)\Big\}\Big). \end{aligned}$$

Therefore, by Lemma C.1, there is a constant c_1 such that as long as $\frac{1}{\eta} \ge c_1 M^2(\log(1/\delta) + \lambda)$, it holds that for any $x \in \mathbb{R}^d$ such that $\epsilon(M; x) \le C := \frac{e^{4B}-1}{8}$,

$$\mathbb{E}_{Z_1\sim\mathsf{N}(0,(\pi/2)^2\eta I)} \exp\big(2\, \mathbb{E}_{g\sim \mathsf{O}_{\mathsf{grad}}(Y)}\, \tau_{B/2}(|\langle Z_1, g - \nabla f(x)\rangle|)\big) - 1 \le \delta + 8\epsilon(M)\,.$$

This immediately implies that

$$\begin{aligned} & \mathbb{E}_{x\sim q,\, Z_1\sim\mathsf{N}(0,(\pi/2)^2\eta I)} \exp\Big(2\min\Big\{2B, \mathbb{E}_{g\sim \mathsf{O}_{\mathsf{grad}}(x)}\, \tau_{B/2}(\langle Z_1, \nabla f(x) - g\rangle)\Big\}\Big) - 1 \\ & \quad \le \delta + 8\, \mathbb{E}_{x\sim q} \min\{\epsilon(M; x), C\}. \end{aligned}$$

By Lemma C.2 and Corollary C.3, there is a constant c_2 such that as long as

$$\frac{1}{\eta^{1+s}} \ge c_2\beta_s^2 \left(d^s \log(1/\delta) + \frac{s}{d^{1-s}} \log^2(1/\delta)\right) + c_2 \left(\varepsilon_{\mathsf{prox}}^2 \log(1/\delta)\right)^{1+s},$$

it holds that

$$\mathbb{E}_{x\sim q,\, Z_1\sim\mathsf{N}(0,(\pi/2)^2\eta I)} \exp\big(2\tau_{B/2}(\langle Z_1, u - \nabla f(Z)\rangle)\big) - 1 \le \delta,$$

and $\mathbb{E}_q[f] \le e\, \mathbb{E}_\nu[f] + \delta$ for any bounded function $f : \mathbb{R}^d \to [0,1]$. This immediately implies that

$$\mathbb{E}_{x\sim q} \min\{\epsilon(M; x), C\} \le e\, \mathbb{E}_{x\sim\nu} \min\{\epsilon(M; x), C\} + \delta.$$

Combining the inequalities above and rescale $\delta \leftarrow \frac{\delta}{3}$ completes the proof. □

Lemma C.1 *Suppose that $C > 1$ is a constant, $\sigma M \le \frac{B}{2}$ and $\lambda \le \min\{\frac{B}{2C^2\sigma^2M^2}, \frac{1}{2CM\sigma}\}$. Then, as long as $\epsilon(M;x) \le C$, it holds that*

$$\mathbb{E}_{Z\sim\mathsf{N}(0,\sigma^2 I)} \exp\Big(\lambda\, \mathbb{E}_{g\sim\mathsf{O}_{\mathsf{grad}}(x)}\, \tau_B(|\langle Z, \nabla f(x) - g\rangle|)\Big) - 1 \le e^{-\frac{B^2}{8\sigma^2M^2}} + 8\epsilon(M;x)\,.$$

Proof. We denote $P = \mathsf{N}\big(0,\sigma^2 I\big)$ and let Q be the distribution of $v = g - \nabla f(x)$. Without loss of generality, we assume the support of Q does not contain 0, and define $\mathfrak{m} = \mathbb{E}_{v\sim Q}\|v\|$. Note that we assume $\epsilon(M;x) \le 1$, i.e.,

$$(C-1)M \ge \mathbb{E}_{v\sim Q}[\|v\|\mathbb{I}\{\|v\| > M\}] = \mathfrak{m} - \mathbb{E}_{v\sim Q}[\|v\|\mathbb{I}\{\|v\| \le M\}] \ge \mathfrak{m} - M,$$

i.e., $\mathfrak{m} \le CM$. Define $\alpha_v = \frac{\mathfrak{m}+M}{\|v\|+M}$. Then we know $\mathbb{E}_{v\sim Q}[1/\alpha_v] = 1$, and hence we can consider the distribution $\overline{Q}$ over $\mathbb{R}^d$ such that $\frac{d\overline{Q}}{dQ}(g) = 1/\alpha_v$. We can rewrite

$$\begin{aligned} I &\coloneqq \mathbb{E}_{Z\sim P} \exp\big(\lambda\, \mathbb{E}_{v\sim Q}\, \tau_B(|\langle Z, v\rangle|)\big) - 1 \\ &= \mathbb{E}_{Z\sim P} \exp\big(\lambda\, \mathbb{E}_{v\sim \overline{Q}}[\alpha_v \tau_B(|\langle Z, v\rangle|)]\big) - 1 \\ &\le \mathbb{E}_{Z\sim P, v\sim\overline{Q}} \exp\big(\lambda\alpha_v \tau_B(|\langle Z, v\rangle|)\big) - 1\,. \end{aligned}$$

Then, using Lemma B.1, we can obtain the following upper bounds:

(1) When $B \ge 2\max\{\|v\|\sigma, \lambda\alpha_v^2\|v\|^2\sigma^2\}$, i.e., when $\|v\| \le \frac{B}{2\sigma}$ *and* $\mathfrak{m} + M \le \frac{\sqrt{B/(2\lambda)}}{\sigma}$, it holds that

$$\mathbb{E}_{Z\sim P} \exp(\lambda\alpha_v \tau_B(|\langle Z, v\rangle|)) - 1 \le e^{-\frac{B^2}{8\sigma^2\|v\|^2}}\,.$$

(2) For any $g \ne 0$, it holds that

$$\mathbb{E}_{Z\sim P} \exp(\lambda\alpha_v \tau_B(|\langle Z, v\rangle|)) - 1 \le 2e^{\frac{1}{2}\lambda^2\alpha_v^2\|v\|^2\sigma^2} \le 2e^{\frac{1}{2}\lambda^2(\mathfrak{m}+M)^2\sigma^2} \le 4\,,$$

where we use the condition $\lambda \le \frac{1}{2CM\sigma}$ and the fact that $\mathfrak{m} \le CM$.

Now, we note that we assume $M \le \frac{B}{2\sigma}$ *and* $CM \le \frac{\sqrt{B/(2\lambda)}}{\sigma}$. Then we can upper bound

$$\begin{aligned} I &\le \mathbb{E}_{v\sim\overline{Q}}[\mathbb{I}\{\|v\| \le M\}(\mathbb{E}_{Z\sim P} \exp(\lambda\alpha_v \tau_B(|\langle Z, v\rangle|)) - 1)] \\ &\quad + \mathbb{E}_{v\sim\overline{Q}}[\mathbb{I}\{\|v\| > M\}(\mathbb{E}_{Z\sim P} \exp(\lambda\alpha_v \tau_B(|\langle Z, v\rangle|)) - 1)] \\ &\le e^{-\frac{B^2}{8\sigma^2M^2}} + 4\mathbb{P}_{v\sim\overline{Q}}(\|v\| \ge M)\,. \end{aligned}$$

Finally, using $\frac{d\overline{Q}}{dQ}(g) = \frac{\|v\|+M}{\mathfrak{m}+M}$, we have

$$\mathbb{P}_{v\sim\overline{Q}}(\|v\| \ge M) \le \frac{2}{M}\, \mathbb{E}_{v\sim Q}[\|v\|\mathbb{I}\{\|v\| \ge M\}]\,.$$

Combining these inequalities gives the desired upper bound. □

Lemma C.2 *Suppose that $Z_0 \sim \mathsf{N}(0, \eta I)$, $Z_1 \sim \mathsf{N}\big(0, \sigma^2 I\big)$, and $Y = \widehat{x} + Z_0$. Suppose Assumption 2.6 holds and $\|\nabla f(\widehat{x}) - u\| \le \varepsilon_{\mathsf{prox}}$. Then it holds that*

$$\mathbb{E}\exp(\lambda \tau_B(|\langle Z_1, \nabla f(Y) - u\rangle|)) - 1 \le 2\exp\Big(-\frac{B}{12}\min\Big\{\sqrt{\frac{d^{1-s}}{s\eta^s\sigma^2\beta_s^2}}, \frac{B}{\sigma^2\varepsilon_{\mathsf{prox}}^2 + \sigma^2\beta_s^2(\eta d)^s}\Big\}\Big),$$
$$\text{for } 0 \le \lambda \le \frac{1}{6}\min\Big\{\sqrt{\frac{d^{1-s}}{s\eta^s\sigma^2\beta_s^2}}, \frac{B}{\sigma^2\varepsilon_{\mathsf{prox}}^2 + \sigma^2\beta_s^2(\eta d)^s}\Big\}.$$

Proof. We write

$$\begin{aligned} M_\lambda &:= \mathbb{E}\exp(\lambda\tau_B(|\langle Z_1, \nabla f(Y) - u\rangle|)) - 1 \\ &\le \mathbb{E}\exp(\lambda(|\langle Z_1, \nabla f(Y) - u\rangle| - B)) \\ &\le 2e^{-\lambda B}\,\mathbb{E}\exp\Big(\frac{1}{2}\lambda^2\sigma^2\|\nabla f(Y) - u\|^2\Big) \\ &\le 2e^{-\lambda B + \lambda^2\sigma^2\|\nabla f(\widehat{x}) - u\|^2}\,\mathbb{E}\exp\big(\lambda^2\sigma^2\|\nabla f(Y) - \nabla f(\widehat{x})\|^2\big). \end{aligned}$$

Using Assumption 2.6 and Lemma B.2, it holds that as long as $\lambda^2\sigma^2\beta_s^2 \le \frac{d^{1-s}}{4s\eta^s}$,

$$\mathbb{E}\exp\big(\lambda^2\sigma^2\|\nabla f(Y) - \nabla f(\widehat{x})\|^2\big) \le \mathbb{E}\exp\big(\lambda^2\sigma^2\beta_s^2\|Y - \widehat{x}\|^{2s}\big) \le \exp\big(2\lambda^2\sigma^2\beta_s^2(\eta d)^s\big).$$

Therefore, we have shown

$$M_\lambda \le 2\exp\big(\lambda^2\sigma^2\varepsilon_{\mathsf{prox}}^2 + 2\lambda^2\sigma^2\beta_s^2(\eta d)^s - \lambda B\big), \qquad \text{for } \lambda \le \sqrt{\frac{d^{1-s}}{4s\eta^s\sigma^2\beta_s^2}}.$$

Note that $\lambda \mapsto M_\lambda$ is an increasing function, and hence we can choose

$$\lambda_\star = \min\Big\{\sqrt{\frac{d^{1-s}}{4s\eta^s\sigma^2\beta_s^2}}, \frac{B}{2\sigma^2\varepsilon_{\mathsf{prox}}^2 + 4\sigma^2\beta_s^2(\eta d)^s}\Big\},$$

so that for any $\lambda \le \lambda_\star$, it holds that

$$M_\lambda \le M_{\lambda_\star} \le 2\exp\Big(-\frac{B}{12}\min\Big\{\sqrt{\frac{d^{1-s}}{s\eta^s\sigma^2\beta_s^2}}, \frac{B}{\sigma^2\varepsilon_{\mathsf{prox}}^2 + \sigma^2\beta_s^2(\eta d)^s}\Big\}\Big).$$

□

Corollary C.3 *There is an absolute constant $c > 0$ such that the following holds. Suppose $\|u - \nabla f(\widehat{x})\| \le \varepsilon_{\mathsf{prox}}$.*

For any $\delta \in (0, \frac{1}{2}]$, as long as $\frac{1}{\eta} \ge c\varepsilon_{\mathsf{prox}}^2\log(1/\delta)$ and $\frac{1}{\eta^{1+s}} \ge c\beta_s^2 d^s\log(1/\delta)$, for any $f : \mathbb{R}^d \to [0,1]$ it holds that

$$\mathbb{E}_q[f] \le e\,\mathbb{E}_\nu[f] + \delta.$$

Proof. Recall that (in the proof of Theorem 3.3) we denote $\overline{W}_{r,z,x} \coloneqq \langle \dot{\gamma}_{z,r}(x), u - \nabla f(\gamma_{z,r}(x)) \rangle$ and

$$\log \nu(x) - \log q(x) + \text{const} = \mathbb{E}_{r,z}\, \overline{W}_{r,z,x}.$$

By Lemma B.3, it holds that for any $\ell \geq 1$,

$$\begin{aligned}
\max\{D_\ell(\nu \,\|\, q), D_\ell(q \,\|\, \nu)\} + 1 &\leq \mathbb{E}_{x\sim q}\, \mathbb{E}_{r,z}\, e^{2\ell|\overline{W}_{r,z,x}|} \\
&= \mathbb{E}_r\, \mathbb{E}_{x\sim q,\, z\sim \mathsf{N}(0,\eta I)} \exp(2\ell|\langle \dot{\gamma}_{z,r}(x), u - \nabla f(\gamma_{z,r}(x))\rangle|) \\
&= \mathbb{E}_{x'\sim q,\, z'\sim \mathsf{N}(0,(\pi/2)^2\eta I)} \exp\big(2\ell\big|\langle z', u - \nabla f(x')\rangle\big|\big),
\end{aligned}$$

where the last line uses Eq. (11). Then, from our proof of Lemma C.2, we know that as long as $50\ell^2\eta\beta_s^2 \leq \frac{d^{1-s}}{4s\eta^s}$,

$$\begin{aligned}
\max\{D_\ell(\nu \,\|\, q), D_\ell(q \,\|\, \nu)\} + 1 &\leq \mathbb{E}_{x'\sim q,\, z'\sim \mathsf{N}(0,(\pi/2)^2\eta I)} \exp\big(2\ell\big|\langle z', u - \nabla f(x')\rangle\big|\big) \\
&\leq 2\exp\big(100\ell^2\eta\|\nabla f(\widehat{x}) - u\|^2 + 100\ell^2\eta\beta_s^2(\eta d)^s\big)\,.
\end{aligned}$$

Finally, by Lemma B.4, it holds that for any $f : \mathbb{R}^d \to [0,1]$,

$$\mathbb{E}_q[f] - e\,\mathbb{E}_\nu[f] \leq \inf_{\ell\geq 1} e^{1-\ell}(1 + D_\ell(q \,\|\, \nu)).$$

Then, we set $\ell_\star = \frac{1}{200\eta(\varepsilon_{\mathsf{prox}}^2 + \beta_s^2(\eta d)^s)}$. As long as $\ell_\star \geq 1$, we have $\mathbb{E}_q[f] - e\,\mathbb{E}_\nu[f] \leq e^{1-\frac{1}{2}\ell_\star}$. This is the desired result. □

Corollary C.4 *There is an absolute constant $c > 0$ such that the following holds. Suppose $\|u - \nabla f(\widehat{x})\| \leq \varepsilon_{\mathsf{prox}}$. As long as $\frac{1}{\eta} \geq c\varepsilon_{\mathsf{prox}}^2$ and $\frac{1}{\eta^{1+s}} \geq c\beta_s^2 d^s$, it holds that $1 + D_{\chi^2}(\nu \,\|\, q) \leq e^{c(\eta\varepsilon_{\mathsf{prox}}^2 + \eta^{1+s}\beta_s^2 d^s)}$.*

Proof. It is straightforward to verify that for $V(x) \coloneqq \mathbb{E}_{r,z}\, \overline{W}_{r,z,x}$, it holds that

$$1 + D_{\chi^2}(\nu \,\|\, q) = \frac{\mathbb{E}_{x\sim q}\, e^{2V(x)}}{(\mathbb{E}_{x\sim q}\, e^{V(x)})^2}.$$

Note that $\mathbb{E}_{x\sim q}[V(x)] = 0$ by Eq. (11). Hence,

$$\begin{aligned}
1 + D_{\chi^2}(\nu \,\|\, q) &= \mathbb{E}_{x\sim q}\, e^{2V(x)} \leq \mathbb{E}_{x\sim q}\, \mathbb{E}_{r,z}\, e^{2\overline{W}_{r,z,x}} \\
&= \mathbb{E}_r\, \mathbb{E}_{x\sim q,\, z\sim \mathsf{N}(0,\eta I)} \exp(2\langle \dot{\gamma}_{z,r}(x), u - \nabla f(\gamma_{z,r}(x))\rangle) \\
&= \mathbb{E}_{x'\sim q,\, z'\sim \mathsf{N}(0,(\pi/2)^2\eta I)} \exp\big(2\langle z', u - \nabla f(x')\rangle\big) \\
&= \mathbb{E}_{x'\sim q} \exp\Big(\frac{\pi^2}{2}\eta\,\|u - \nabla f(x')\|^2\Big)\,.
\end{aligned}$$

The remaining proof is the same. □

C.2. Proof of Theorem 3.7

Without loss of generality we only consider the case $n = 1$. We denote by P_x the joint distribution of (z, v, v') under (4).

By Theorem 3.1, the output of Algorithm 1 with the specified choices samples from $\widehat{\nu}$, such that

$$\log \widehat{\nu}(x) - \log q(x) = \text{const} + \mathbb{E}_{(z,v,v')\sim P_x} \mathsf{Clip}_B(W_{z,v,v',x})\,.$$

On the other hand, we know

$$\begin{aligned}\mathbb{E}_{(z,v,v')\sim P_x}[W_{z,v,v',x}] &= -f(x) + \langle u, x\rangle + \text{const}\\ &= \log \nu(x) - \log q(x) + \text{const}\,.\end{aligned}$$

Then, using Lemma B.6, $|\log \nu(x) - \log \widehat{\nu}(x) - \text{const}| \le V(x)$, where

$$\begin{aligned}V(x) &\coloneqq \mathbb{E}_{(z,v,v')\sim P_x} \tau_B(W_{z,v,v',x})\\ &\le \mathbb{E}_{z\sim q} \min\big\{2B, \mathbb{E}_{v\sim \mathsf{O}_{\mathsf{eval}}(x),\, v'\sim \mathsf{O}_{\mathsf{eval}}(z)} \tau_{B/2}(v - f(x) + f(z) - v')\big\}\\ &\quad + \mathbb{E}_{z\sim q} \tau_{B/2}(f(z) - f(x) - \langle u, z - x\rangle)\\ &\le \mathbb{E}_{z\sim q} \min\{2B, \epsilon(B/4; x) + \epsilon(B/4; z)\} + \mathbb{E}_{z\sim q} \tau_{B/2}(f(z) - f(x) - \langle u, z - x\rangle)\,.\end{aligned}$$

In the following, we denote $\Delta_{x,z} \coloneqq f(x) - f(z) - \langle u, x - z\rangle$. Then, using Lemma B.3,

$$\begin{aligned}D_{\mathsf{TV}}(\nu, \widehat{\nu}) &\le \mathbb{E}_{x\sim\widehat{\nu}}\big[\exp\big(2\,\mathbb{E}_{z\sim q} \tau_{B/2}(\Delta_{x,z}) + 2\min\{2B, \epsilon(B/4; x) + \epsilon(B/4; z)\}\big) - 1\big]\\ &\le e^{2B}\, \mathbb{E}_{x,z\sim q}\big[\exp\big(2\tau_{B/2}(\Delta_{x,z}) + 2\min\{2B, \epsilon(B/4; x) + \epsilon(B/4; z)\}\big) - 1\big]\end{aligned} \tag{12}$$

where we use $\frac{d\widehat{\nu}}{dq}(x) \le e^{2B}$ for $x \in \mathbb{R}^d$ and the convexity of $w \mapsto e^w$. In the following, it remains to prove the following lemma (the rest of the proof then follows from the argument of Theorem 3.3). □

Lemma C.5 *Let $\Delta_{x,z} \coloneqq f(x) - f(z) - \langle u, x - z\rangle$ and $q = \mathsf{N}(\widehat{x}, \eta I)$. Suppose Assumption 2.6 holds and $\|\nabla f(\widehat{x}) - u\| \le \varepsilon_{\mathsf{prox}}$. Then it holds that*

$$\mathbb{E}_{x,z\sim q}\, e^{\lambda\tau_B(\Delta_{x,z})} - 1 \le 2\exp\Big(-\frac{B}{32}\min\Big\{\sqrt{\frac{d^{1-s}}{s\eta^s\eta\beta_s^2}},\ \frac{B}{\eta\varepsilon_{\mathsf{prox}}^2 + \eta\beta_s^2(\eta d)^s}\Big\}\Big)\,,$$

$$\text{for } 0 \le \lambda \le \frac{1}{16}\min\Big\{\sqrt{\frac{d^{1-s}}{s\eta^s\eta\beta_s^2}},\ \frac{B}{\eta\varepsilon_{\mathsf{prox}}^2 + \eta\beta_s^2(\eta d)^s}\Big\}\,.$$

Proof. We can express

$$\Delta_{x,z} = f(x) - f(z) - \langle u, x - z\rangle = \int_0^1 \langle \dot\gamma_{z,r}(x), \nabla f(\gamma_{z,r}(x)) - u\rangle\, dr,$$

where $\gamma_{z,r}(x)$ and $\dot\gamma_{z,r}(x)$ are defined in Eq. (3). Then, using the convexity of $w \to \tau_B(w)$, we can upper bound

$$\mathbb{E}_{x,z\sim q}\, e^{\lambda\tau_B(\Delta_{x,z})} \le \mathbb{E}_{r\sim\mathsf{Unif}([0,1])}\, \mathbb{E}_{x,z\sim q}\, e^{\lambda\tau_B(\langle\dot\gamma_{z,r}(x), \nabla f(\gamma_{z,r}(x)) - u\rangle)}.$$

Now, following the proof of Theorem 3.3, we know that for any fixed $r \in [0,1]$, under $x, z \sim q$, the vector $\gamma_{z,r}(x), \dot{\gamma}_{z,r}(x)$ are jointly distributed as

$$[\gamma_{z,r}(x); \dot{\gamma}_{z,r}(x)] \sim \mathsf{N}\left(\begin{bmatrix}\widehat{x}\\ 0\end{bmatrix}, \begin{bmatrix}\eta I & \\ & (\pi/2)^2 \eta I\end{bmatrix}\right).$$

Therefore, we have shown that

$$\mathbb{E}_{x,z\sim q}\, e^{\lambda \tau_B(\Delta_{x,z})} \le \mathbb{E}_{Z_0\sim\mathsf{N}(0,\eta I), Z_1\sim\mathsf{N}(0,(\pi/2)^2\eta I)} \exp(\lambda \tau_B(|\langle Z_1, \nabla f(\widehat{x}+Z_0) - u\rangle|)).$$

Applying Lemma C.2 completes the proof. □

C.3. Proof of Theorem 3.10

Let ρ be the transition kernel on $\mathbb{R}^d$ induced by the proximal sampler, i.e., $X' \sim \rho(\cdot \mid X)$ is generated by $Y' \sim \mathsf{N}(X, \eta I)$ and $X' \sim \bar{\pi}^{X|Y=Y'}$. Then, ρ induces a Markov chain $X_0, X_1, \cdots$ by $X_0 \sim \mu_0, X_{n+1} \sim \rho(\cdot \mid X_n)$ for $n \ge 0$. For $n \ge 0$, let μ_n be the law of X_n.

Similarly, we let $\widehat{\rho}$ be the transition with $\bar{\pi}^{X|Y=Y'}$ implemented via Theorem 3.3, and the induced Markov chain $X_0, X_1, \cdots$ is given by $X_0 \sim \mu_0, X_{n+1} \sim \widehat{\rho}(\cdot \mid X_n)$ for $n \ge 0$. For $n \ge 0$, let $\widehat{\mu}_n$ be the law of X_n.

Then, by Lemma B.5 and data-processing inequality, it holds that

$$D_{\mathsf{TV}}(\widehat{\mu}_N, \mu_N) \le D_{\mathsf{TV}}(\mathbb{P}_{\widehat{\rho}}, \mathbb{P}_{\rho}) \le \sum_{n=0}^{N-1} \mathbb{E}_{X_n\sim\mu_n} D_{\mathsf{TV}}(\widehat{\rho}(\cdot \mid X_n), \rho(\cdot \mid X_n)). \tag{13}$$

In all cases, we use the following error analysis. By Lemma C.6 and the fact that $D_{\chi^2}(\mu_n \parallel \mu) \le D_{\chi^2}(\mu_0 \parallel \mu)$, it holds that with $G = O(\beta_s^{1/(1+s)}(d + \Delta + \log(N/\delta)))$,

$$\max_{n\in[N]} \mathbb{P}_{X\sim\mu_n}(\|\nabla f(X_n)\| \ge G_\delta) \le \frac{\delta}{10N}.$$

Note that as long as $\|\nabla f(X_n)\| \le G$, we can implement the proximal oracle at X_n with $\varepsilon_{\mathsf{prox}} = 10(\beta_s^{1/(1+s)} + M))$ and success probability at least $1 - \epsilon_n(M)$ by Lemma 2.8, using

$$O(n \log(G/\beta_s^{1/(1+s)})) = O(n \log A) \quad \text{queries to } \mathsf{O}_{\mathsf{grad}}(\cdot).$$

Therefore, by the choice of η (5), we can implement the RGO via Theorem 3.3 so that

$$D_{\mathsf{TV}}(\widehat{\rho}(\cdot \mid X_n), \rho(\cdot \mid X_n)) \le \frac{\delta}{10N} + 5\epsilon_n(M)$$

as long as $\|\nabla f(X_n)\| \le G$, using $O(n \log A)$ queries. Then, by a conditioning argument, we see that

$$\begin{aligned}\mathbb{E}_{X_n\sim\mu_n} D_{\mathsf{TV}}(\widehat{\rho}(\cdot \mid X_n), \rho(\cdot \mid X_n)) &\le \mathbb{P}_{\mu_n}(\|\nabla f(X_n)\| \ge G) + \frac{\delta}{10N} + 5\epsilon_n(M)\\ &\le \frac{\delta}{5N} + 5\epsilon_n(M).\end{aligned}$$

Therefore, taking summation over $k = 0, 1, \cdots, N-1$ gives

$$D_{\mathsf{TV}}(\widehat{\mu}_N, \mu_N) \le \frac{\delta}{5} + 5N\epsilon_n(M). \tag{14}$$

Finally, we can set $n = \phi_M(\frac{\delta}{10N})$ so that $\epsilon_n(M) \le \frac{\delta}{10N}$. Note that this implies

$$D_{\mathsf{TV}}(\widehat{\mu}_N, \mu) \le \frac{3}{4}\,\delta + D_{\mathsf{TV}}(\mu_N, \mu)\,.$$

Now, we apply results from Chen et al. (2022) to bound N such that $D_{\mathsf{TV}}(\mu_N, \mu) \le \frac{\delta}{4}$:

- **LSI case.** Here, $N \asymp \frac{1}{\alpha\eta} \log \frac{D_{\mathsf{KL}}(\mu_0 \,\|\, \mu)}{\delta^2}$.
- **PI case.** Here, $N \asymp \frac{1}{\alpha\eta} \log \frac{D_{\chi^2}(\mu_0 \,\|\, \mu)}{\delta^2}$.
- **LC case.** Here, $N \asymp \frac{W_2^2(\mu_0,\mu)}{\eta\delta^2}$.

Plugging in the choice of η in (5) gives the desired results. □

Lemma C.6 *Suppose that Assumption 2.6 holds and ν is a distribution such that*

$$\log(1 + D_{\chi^2}(\nu \parallel \mu_f)) \le \Delta\,.$$

Then it holds that for $\delta \in (0,1)$,

$$\mathbb{P}_{X\sim\nu}\Big(\|\nabla f(X)\|^2 \ge \frac{64\beta_s^{2/(1+s)}}{d^{(1-s)/(1+s)}}\,(\Delta + d + \log(1/\delta))\Big) \le \delta.$$

Proof. We follow the proof of Chewi (2026, Lemma 6.2.7). For any vector $v \in \mathbb{R}^d$, we bound

$$f(x+v) - f(x) - \langle v, \nabla f(x)\rangle = \int_0^1 \langle v, \nabla f(x+rv) - \nabla f(x)\rangle\, dr \le \beta_s \|v\|^{1+s}\,, \qquad \forall x \in \mathbb{R}^d\,.$$

Then, we can bound

$$\int_{\mathbb{R}^d} e^{-f(x+v)}\,dx \ge \int_{\mathbb{R}^d} e^{-f(x)-\langle v,\nabla f(x)\rangle-\beta_s\|v\|^{1+s}}\,dx\,.$$

Re-organizing gives

$$\mathbb{E}_{X\sim\mu_f}[e^{\langle v,\nabla f(X)\rangle}] \le e^{\beta_s\|v\|^{1+s}}\,, \qquad \forall v \in \mathbb{R}^d\,.$$

For any $m \ge 0$ such that $\beta_s \le \frac{d^{(1-s)/2}}{2(1+s)m^{(1+s)/2}}$, we can take expectation over $v \sim \mathsf{N}(0, mI)$, and then Lemma B.2 gives

$$\mathbb{E}_{X\sim\mu_f} \exp\Big(\frac{m}{2}\|\nabla f(X)\|^2\Big) \le \exp\big(2\beta_s(md)^{(1+s)/2}\big)\,.$$

Therefore, we choose $m > 0$ such that $m^{1+s} = \frac{d^{1-s}}{16\beta_s^2}$, and then

$$\begin{aligned}\mathbb{E}_{X\sim\nu} \exp\Big(\frac{m}{4}\|\nabla f(X)\|^2\Big) &\le \sqrt{(1 + D_{\chi^2}(\nu \parallel \mu_f))\, \mathbb{E}_{X\sim\mu_f} \exp\Big(\frac{m}{2}\|\nabla f(X)\|^2\Big)} \\ &\le \exp\Big(\frac{1}{2}\Delta + \beta_s (md)^{(1+s)/2}\Big) \le \exp\Big(\frac{1}{2}(\Delta + d)\Big)\,.\end{aligned}$$

Applying Markov's inequality gives

$$\mathbb{P}_{X\sim\nu}(\|\nabla f(X)\| \ge G) \le e^{-mG^2/4}\, \mathbb{E}_{X\sim\nu} \exp\Big(\frac{m}{4}\|\nabla f(X)\|^2\Big) \le \delta\,,$$

as long as $G^2 \ge \frac{4}{m}(\Delta + d + \log(1/\delta))$. This is the desired result. □

C.4. Proof of Theorem 3.12

The proof is very similar to the proof of Theorem 3.10.

By Lemma C.6 and the fact that $D_{\chi^2}(\mu_n \parallel \mu) \le D_{\chi^2}(\mu_0 \parallel \mu)$, it holds that with $G > 0$ chosen as

$$G^2 = \frac{64\beta_s^{2/(1+s)}}{d^{(1-s)/(1+s)}} \left(\Delta + d + \log(10N/\delta)\right),$$

it holds that

$$\max_{n\in[N]} \mathbb{P}_{X\sim\mu_n}(\|\nabla f(X_n)\| \ge G_\delta) \le \frac{\delta}{10N}\,.$$

Note that as long as $\|\nabla f(X_n)\| \le G$, we can implement the proximal oracle at X_n with $\varepsilon_{\mathsf{prox}} = G$ by trivially returning $x = X_n$. Therefore, by the choice of η (6), we can implement the RGO via Theorem 3.7 so that $D_{\mathsf{TV}}(\widehat{\rho}(\cdot \mid X_n), \rho(\cdot \mid X_n)) \le \frac{\delta}{10N} + 10\epsilon_n(M)$ as long as $\|\nabla f(X_n)\| \le G$, using $O(n)$ queries. The rest of the proof is concluded as before. □

C.5. Proof of Theorem 3.14

Denote

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

We let $\mathbb{P}_\star(\cdot)$ be the probability law of $(X_0, Y_0), \dots, (X_N, Y_N)$ induced by the proximal sampler, and $\mathbb{E}_\star[\cdot]$ be the corresponding expectation.

In the following, we choose $M > 0$ as

$$M^2 = 4\beta_1^2 CK\eta\,(d + \log(K/\delta)) + 4\beta_1^2 CK^2\eta^2\beta_1\,(d + \Delta + \log(K/\delta))\,,$$

where the constant $C > 0$ is from Lemma C.7, and $A = M + \varepsilon_{\mathsf{prox}}$.

For each $i \ge 0$, we consider each time step in the epoch $\mathcal{K}_i := [iK, (i+1)K)$. By definition of the oracle O_{k+1}, as long as $\|x - X_{iK}\| \le \frac{M}{2\beta_1}$, it holds that $\|g - \nabla f(x)\| \le M$ deterministically under the oracle $O_{k+1}(x)$. Therefore, by Theorem 3.3, as long as

$$\frac{1}{\eta} \gg \beta_1\sqrt{d\log(1/\delta)} + (A^2 + M^2 + \beta_1)\log(1/\delta)\,, \tag{15}$$

and $\|Y_k - \widehat{X}_{k+1} - \eta\nabla f(\widehat{X}_{k+1})\| \le \eta A$, it holds that our algorithm generates a sample X_{k+1} following a distribution $\widehat{\nu}_{k+1}(\cdot \mid Y_k, X_{iK})$ satisfying

$$D_{\mathsf{TV}}(\nu(\cdot \mid Y_k), \widehat{\nu}_{k+1}(\cdot \mid Y_k, X_{iK})) \lesssim \delta + \mathbb{P}_{X_{k+1}\sim\nu(\cdot|Y_k)}\Big(\|X_k - X_{iK}\| \ge \frac{M}{2\beta_1}\Big)\,.$$

Note that Eq. (15) can indeed be ensured by Eq. (7). Then, taking expectation over $(Y_k, X_{iK}) \sim \mathbb{P}_\star$, we have

$$\begin{aligned}\mathbb{E}_\star\, D_{\mathsf{TV}}(\nu(\cdot \mid Y_k), \widehat{\nu}_{k+1}(\cdot \mid Y_k, X_{iK})) \lesssim \delta &+ \mathbb{P}_\star\Big(\|X_k - X_{iK}\| \ge \frac{M}{2\beta_1}\Big)\\ &+ \mathbb{P}_\star\big(\|Y_k - \widehat{X}_{k+1} - \eta\nabla f(\widehat{X}_{k+1})\| \ge \eta A\big)\,.\end{aligned}$$

By definition, $\widehat{X}_{k+1} = \widehat{X}_{iK} + Y_k - Y_{iK}$, and hence

$$\begin{aligned}\|Y_k - \widehat{X}_{k+1} - \eta\nabla f(\widehat{X}_{k+1})\| &= \|Y_{iK} - \widehat{X}_{iK} - \eta\nabla f(\widehat{X}_{k+1})\|\\ &\le \|Y_{iK} - \widehat{X}_{iK} - \eta\nabla f(\widehat{X}_{iK})\| + \beta_1\eta\,\|\widehat{X}_{iK} - \widehat{X}_{k+1}\|\\ &= \|Y_{iK} - \widehat{X}_{iK} - \eta\nabla f(\widehat{X}_{iK})\| + \beta_1\eta\,\|Y_k - Y_{iK}\|\,.\end{aligned}$$

Therefore, we can bound

$$\begin{aligned}&\mathbb{P}_\star\big(\|Y_k - \widehat{X}_{k+1} - \eta\nabla f(\widehat{X}_{k+1})\| \ge \eta A\big)\\ &\le \mathbb{P}_\star\big(\|Y_{iK} - \widehat{X}_{iK} - \eta\nabla f(\widehat{X}_{iK})\| \ge \eta\varepsilon_{\mathsf{prox}}\big) + \mathbb{P}_\star\Big(\|Y_k - Y_{iK}\| \ge \frac{M}{\beta_1}\Big)\,.\end{aligned}$$

By our definition of the proximal oracle $\mathsf{O}_{\mathsf{prox},\eta}(\cdot)$, the first term of the RHS is bounded by δ. Combining the inequalities and taking summation over k and apply Lemma B.5, we know

$$\begin{aligned}D_{\mathsf{TV}}(\mu_N, \widehat{\mu}_N) &\le \sum_{i=0}^{\lfloor N/K\rfloor}\sum_{k\in\mathcal{K}_i\cap[N]} \mathbb{E}_\star\, D_{\mathsf{TV}}(\nu(\cdot \mid Y_k), \widehat{\nu}_{k+1}(\cdot \mid Y_k, X_{iK}))\\ &\lesssim N\delta + \sum_{i=0}^{\lfloor N/K\rfloor}\sum_{k\in\mathcal{K}_i\cap[N]}\Big[\mathbb{P}_\star\Big(\|X_k - X_{iK}\| \ge \frac{M}{2\beta_1}\Big) + \mathbb{P}_\star\Big(\|Y_k - Y_{iK}\| \ge \frac{M}{\beta_1}\Big)\Big]\\ &\lesssim N\delta\,,\end{aligned}$$

where the last inequality follows from Lemma C.7. □

Lemma C.7 *Suppose that Assumption 2.6 holds and ν is a distribution such that*

$$\log(1 + D_{\chi^2}(\nu \parallel \mu_f)) \le \Delta\,.$$

Consider the Markov chain $X_0 \to Y_0 \to \cdots \to X_K \to Y_K$ generated by the proximal sampler. Then as long as $\eta \le \frac{1}{C\beta_1\sqrt{dK}}$, it holds that for $\delta \in (0,1)$,

$$\mathbb{P}\Big(\max_{k\in[K]}\|Y_k - Y_0\| \ge R\Big) \le \delta\,, \qquad \mathbb{P}\Big(\max_{k\in[K]}\|X_k - Y_0\| \ge R\Big) \le \delta\,,$$

where $R > 0$ is defined as ($C > 0$ is an absolute constant):

$$R^2 \coloneqq CK\eta(d + \log(K/\delta)) + CK^2\eta^2\beta_1(d + \Delta + \log(K/\delta))\,.$$

Proof. Denote $\overline{\nu}(\cdot \mid y) = \mathsf{N}\big(\mathsf{prox}_{\eta f}(y), \eta I\big)$. We consider the following distributions of Markov chain $X_0 \to Y_0 \to \cdots \to X_K \to Y_K$:

(1) P is the distribution of the exact proximal sampler, i.e., $X_0 \sim \nu$, and for each $k \in [K]$, $Y_k \mid X_k \sim \mathsf{N}(X_k, \eta I)$ and $X_{k+1} \mid Y_k \sim \nu(\cdot \mid Y_k)$.

(2) Q is the following distribution: $X_0 \sim \nu$, and for each $k \in [K]$, $Y_k \mid X_k \sim \mathsf{N}(X_k, \eta I)$ and $X_{k+1} \sim \overline{\nu}(\cdot \mid Y_k)$.

By Corollary C.4, there is a constant $c_1 > 0$ such that as long as $\frac{1}{\eta} \geq c_1 \beta_1 \sqrt{d}$, it holds that

$$1 + D_{\chi^2}(\nu(\cdot \mid y) \parallel \overline{\nu}(\cdot \mid y)) \leq \exp\big(c\eta^2 \beta_1^2 d\big)\,, \qquad \forall y \in \mathbb{R}^d\,.$$

Therefore, it is straightforward to verify that

$$1 + D_{\chi^2}(P \parallel Q) \leq (1 + \max_{y \in \mathbb{R}^d} D_{\chi^2}(\nu(\cdot \mid y) \parallel \overline{\nu}(\cdot \mid y)))^K \leq \exp\big(c\eta^2 \beta_1^2 dK\big) \leq O(1)\,.$$

In the following, we denote $\overline{X}_k = \mathsf{prox}_{\eta f}(Y_{k-1})$, $Z_k = X_k - \overline{X}_{k-1}$, and $Z_k' = Y_k - X_k$. Then, we can express $Y_{k-1} = \overline{X}_k + \eta \nabla f(\overline{X}_k)$, and hence $Y_k = Y_{k-1} + Z_k + Z_k' - \eta \nabla f(\overline{X}_k)$. Apply this recursively, we get

$$Y_k - Y_0 = \sum_{i=1}^k (Z_i + Z_i') - \eta \sum_{i=1}^k \nabla f(\overline{X}_i)\,.$$

Therefore, we can bound

$$\|Y_k - Y_0\| \leq \Big\|\sum_{i=1}^k (Z_i + Z_i')\Big\| + \eta \sum_{i=1}^k \|\nabla f(X_i)\| + \eta\beta_1 \sum_{i=1}^k \|Z_i\|\,.$$

Note that under Q, $\sum_{i=1}^k (Z_i + Z_i') \sim \mathsf{N}(0, 2k\eta I)$ and hence for any $k \in [K]$,

$$Q\bigg(\Big\|\sum_{i=1}^k (Z_i + Z_i')\Big\| \geq \sqrt{2k\eta}\,(\sqrt{d} + 2\sqrt{\log(1/\delta)})\bigg) \leq \delta\,.$$

In addition, $Q\big(\|Z_i\| \geq \sqrt{\eta}\,(\sqrt{d} + 2\sqrt{\log(1/\delta)})\big) \leq \delta$. Therefore, using the union bound, we get

$$Q\bigg(\max_{k \in [K]} \|Y_k - Y_0\| \geq (\sqrt{2K\eta} + K\beta_1\eta\sqrt{\eta})\,(\sqrt{d} + 2\sqrt{\log(2K/\delta)}) + \eta \sum_{i=1}^K \|\nabla f(X_i)\|\bigg) \leq \delta\,.$$

Note that $\eta \leq \frac{1}{\beta_1\sqrt{dK}}$. Then, applying the change-of-measure argument, we get

$$P\bigg(\max_{k \in [K]} \|Y_k - Y_0\| \geq C_1\sqrt{K\eta}\,(\sqrt{d} + \sqrt{\log(K/\delta)}) + \eta \sum_{i=1}^K \|\nabla f(X_i)\|\bigg) \leq \delta\,.$$

Finally, by Lemma C.6, we can show $P(\|\nabla f(X_i)\| \geq C_2\sqrt{\beta_1(\Delta + d + \log(K/\delta))}) \leq \frac{\delta}{K}$. Taking the union bound completes the proof of the first inequality. The second inequality follows similar by noting

$$X_k - X_0 = Z_0' + \sum_{i=1}^{k-1} (Z_i + Z_i') + Z_k - \eta \sum_{i=1}^k \nabla f(\overline{X}_i)\,.$$

□

Appendix D. Proofs from Section 4

D.1. Proof of Proposition 4.2

Fix any $p \in [0,1]$ such that $p\psi(\sqrt{\alpha}\delta/p) + \psi(\sqrt{\alpha}\delta) \le 1$. We denote $M = \frac{\sqrt{\alpha}\delta}{p}$.

We denote $\theta = \delta/\sqrt{\alpha}$. Consider ψ-oracle O_0 and O_θ: for any $x \in \mathbb{R}$, $O_0(x)$ returns αx with probability 1, and $O_\theta(x)$ returns $\alpha x - M$ with probability p and returns αx otherwise. Then, O_0 is a ψ-oracle for f_0, and O_θ is a ψ-oracle for f_θ, because

$$\mathbb{E}_{g\sim O_\theta(x)}\, \psi(|g - f_\theta'(x)|) = p\psi(M - \alpha\theta) + (1-p)\psi(\alpha\theta) \le 1\,.$$

Note that

$$D_{\mathsf{TV}}(O_0(x), O_\theta(x)) = p\,,$$

and hence by the sub-additivity of the TV distance (Lemma B.5), we have

$$D_{\mathsf{TV}}(\mathtt{Alg}(O_0), \mathtt{Alg}(O_\theta)) \le Tp\,.$$

On the other hand, by our assumption, it holds that

$$D_{\mathsf{TV}}(p_0, \mathtt{Alg}(O_0)) \le \frac{\delta}{10}\,, \qquad D_{\mathsf{TV}}(p_\theta, \mathtt{Alg}(O_\theta)) \le \frac{\delta}{10}\,.$$

An elementary calculation also yields

$$D_{\mathsf{TV}}(p_0, p_\theta) = D_{\mathsf{TV}}\big(\mathsf{N}\big(0, \alpha^{-1}\big), \mathsf{N}\big(\theta, \alpha^{-1}\big)\big) = D_{\mathsf{TV}}(\mathsf{N}(0,1), \mathsf{N}(\delta, 1)) \ge \frac{\delta}{3}\,.$$

Therefore, by triangle inequality,

$$\begin{aligned}\frac{\delta}{3} \le D_{\mathsf{TV}}(p_0, p_\theta) &\le D_{\mathsf{TV}}(p_0, \mathtt{Alg}(O_0)) + D_{\mathsf{TV}}(\mathtt{Alg}(O_0), \mathtt{Alg}(O_\theta)) + D_{\mathsf{TV}}(p_\theta, \mathtt{Alg}(O_\theta)) \\ &\le \frac{\delta}{5} + Tp\,,\end{aligned}$$

and this implies $T \ge \frac{\delta}{10p}$. Taking infimum over $p \in (0,1]$ such that $p\psi(\sqrt{\alpha}\delta/p) + \psi(\sqrt{\alpha}\delta) \le 1$ completes the proof. □