Structured Logconcave Sampling with a Restricted Gaussian Oracle

Yin Tat Lee YINTAT@UW.EDU
University of Washington and Microsoft Research

Ruoqi Shen SHENR3@CS.WASHINGTON.EDU
University of Washington

Kevin Tian KJTIAN@STANFORD.EDU
Stanford University

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We give algorithms for sampling several structured logconcave families to high accuracy.[1] We further develop a reduction framework, inspired by *proximal point methods* in convex optimization, which bootstraps samplers for regularized densities to generically improve dependences on problem conditioning κ from polynomial to linear. A key ingredient in our framework is the notion of a "restricted Gaussian oracle" (RGO) for $g : \mathbb{R}^d \to \mathbb{R}$, which is a sampler for distributions whose negative log-likelihood sums a quadratic (in a multiple of the identity) and g. By combining our reduction framework with our new samplers, we obtain the following bounds for sampling structured distributions to total variation distance ϵ.

- For composite densities $\exp(-f(x)-g(x))$, where f has condition number κ and convex (but possibly non-smooth) g admits an RGO, we obtain a mixing time of $O(\kappa d \log^3 \frac{\kappa d}{\epsilon})$, matching the state-of-the-art non-composite bound Lee et al. (2020). No composite samplers with better mixing than general-purpose logconcave samplers were previously known.
- For logconcave finite sums $\exp(-F(x))$, where $F(x) = \frac{1}{n}\sum_{i\in[n]} f_i(x)$ has condition number κ, we give a sampler querying $\widetilde{O}(n + \kappa \max(d, \sqrt{nd}))$ gradient oracles[2] to $\{f_i\}_{i\in[n]}$. No high-accuracy samplers with nontrivial gradient query complexity were previously known.
- For densities with condition number κ, we give an algorithm obtaining mixing time $O(\kappa d \log^2 \frac{\kappa d}{\epsilon})$, improving Lee et al. (2020) Lee et. al. by a logarithmic factor with a significantly simpler analysis. We also show a zeroth-order algorithm attains the same query complexity.

Keywords: Logconcave sampling, composite objectives, finite-sum objectives

1. Introduction

Since its study was pioneered by the celebrated randomized convex body volume approximation algorithm of Dyer, Frieze, and Kannan Dyer et al. (1991), designing samplers for logconcave distributions has been a very active area of research in theoretical computer science and statistics with

1. We say a sampler is "high-accuracy" if its mixing time has polylogarithmic dependence on the target accuracy ϵ.
2. For convenience of exposition, the $\widetilde{O}$ notation hides logarithmic factors in the dimension d, problem conditioning κ, desired accuracy ϵ, and summand count n (when applicable). A first-order (gradient) oracle for $f : \mathbb{R}^d \to \mathbb{R}$ returns $(f(x), \nabla f(x))$ on input x, and a zeroth-order (value) oracle only returns $f(x)$.

many connections to other fields. In a generic form, the problem can be stated as: sample from a distribution whose negative log-density is convex, under various access models to the distribution.

Developing efficient algorithms for sampling from *structured* logconcave densities is a topic that has received significant recent interest due to its widespread practical applications. There are many types of structure which densities commonplace in applications may possess that are exploitable for improved runtimes. Examples of such structure include derivative bounds ("well-conditioned densities") and various types of separability (e.g. "composite densities" corresponding to possibly non-smooth regularization or restrictions to a set, and "logconcave finite sums" corresponding to averages over multiple data points).[3] Building an algorithmic theory for sampling these latter two families, which are not well-understood in the literature, is a primary motivation of this work.

There are strong parallels between the types of structured logconcave families garnering recent attention and the classes of convex functions admitting efficient first-order optimization algorithms. Notably, gradient descent and its accelerated counterpart Nesterov (1983) are well-known to quickly optimize a well-conditioned function, and have become ubiquitous in both practice and theory. Similarly, methods have been developed for efficiently optimizing non-smooth but structured composite objectives Beck and Teboulle (2009) and well-conditioned finite sums Allen-Zhu (2017).

Logconcave sampling and convex optimization are intimately related primitives (Bertsimas and Vempala (2004); Abernethy and Hazan (2016)), so it is perhaps unsurprising that there are analogies between the types of structure algorithm designers may exploit. Nonetheless, our understanding of the complexity landscape for sampling is weak in comparison to counterparts in the field of optimization; few lower bounds are known for the complexity of sampling tasks, and obtaining stronger upper bounds is an extremely active research area (contrary to optimization, where matching bounds exist in many cases). Moreover (and perhaps relatedly), the toolkit for designing samplers is comparatively lacking; for many important primitives in optimization, it is unclear if there are sampling analogs, possibly impeding improvements. Our work broadly falls under the themes of (1) understanding which types of structured logconcave distributions admit efficient samplers, and (2) leveraging connections between optimization and sampling for algorithm design. We address these problems on two fronts, constituting our primary technical contributions.

1. We give a general reduction framework for bootstrapping samplers with mixing times with polynomial dependence on a conditioning measure κ to mixing times with linear dependence on κ. The framework is heavily motivated by a perspective on *proximal point methods* in structured convex optimization as instances of optimizing composite objectives, and leverages this connection via a surprisingly simple analysis (cf. Theorem 3).

2. We develop novel "base samplers" for composite logconcave distributions and logconcave finite sums (cf. Theorems 5, 8). The former is the first composite sampler with stronger guarantees than those known in the general logconcave setting. The latter constitutes the first high-accuracy finite sum sampler whose gradient query complexity improves upon the naïve strategy of querying full gradients of the negative log-density in each iteration.

Using our novel base samplers within our reduction framework, we obtain state-of-the-art samplers for all of the aforementioned structured families, i.e. well-conditioned, composite, and finite sum, as Corollaries 4, 7, and 9. We emphasize that even without our reduction technique, the

3. We make this terminology more precise in Section B.1, which contains various definitions used in this paper.

guarantees of our base samplers for composite and finite sum-structured densities are the first of their kind. However, by boosting their mixing via our reduction, we obtain guarantees for these structured distribution families which are essentially the best one can hope for without a significant improvement in the most commonly studied well-conditioned regime (cf. discussion in Section 2).

We formally state our results in Section 2, and situate them in the literature in Section 3. Section 4 is a technical overview of our approaches for developing our base samplers for composite and finite sum-structured densities (Sections 4.1 and 4.2), as well as our proximal reduction framework (Section 4.3). Finally, we refer the reader to Section A for a roadmap for the rest of the paper.

2. Our results

Before stating our results, we first require the notion of a restricted Gaussian oracle, whose definition is a key ingredient in giving our reduction framework as well as our later composite samplers.

Definition 1 (Restricted Gaussian oracle) *$\mathcal{O}(\lambda, v)$ is a* restricted Gaussian oracle *(RGO) for convex $g : \mathbb{R}^d \to \mathbb{R}$ if it returns*

$$\mathcal{O}(\lambda, v) \leftarrow \text{sample from the distribution with density} \propto \exp\left(-\frac{1}{2\lambda}\|x - v\|_2^2 - g(x)\right).$$

In other words, an RGO asks to sample from a multivariate Gaussian (with covariance a multiple of the identity), "restricted" by some convex function g. Intuitively, if we can reduce a sampling problem for the density $\propto \exp(-g)$ to calling an RGO a small number of times with a small value of λ, each RGO subproblem could be much easier to solve than the original problem. This can happen for a variety of reasons, e.g. if the regularized density is extremely well-conditioned, or because it inherits concentration properties of a Gaussian. This idea of reducing a sampling problem to multiple subproblems, each implementing an RGO, underlies the framework of Theorem 3. Because the idea of regularization by a large Gaussian component repeatedly appears in this paper, we make the following more specific definition for convenience, which lower bounds the size of the Gaussian.

Definition 2 (η-RGO) *We say $\mathcal{O}(\lambda, v)$ is an η-restricted Gaussian oracle (η-RGO) if it satisfies Definition 1 with the restriction that parameter λ is required to be always at most η in calls to $\mathcal{O}$.*

Variants of Definition 1 have implicitly appeared previously Cousins and Vempala (2018); Mou et al. (2019a), and efficient RGO implementation was a key subroutine in the fastest sampler for general logconcave distributions Cousins and Vempala (2018).[4] It extends a similar oracle used in composite optimization, to be discussed shortly. The explicit use of RGOs in a framework such as Theorem 3 is a technical innovation of our work, which we believe will find further use.

Proximal reduction framework. In Section C, we prove correctness of our proximal reduction framework, whose guarantees are stated in the following Theorem 3.

Theorem 3 *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-f_{\text{oracle}}(x))$ such that f_{oracle} is μ-strongly convex, and let $\epsilon \in (0, 1)$. Let $\eta > 0$, $T = \Theta(\frac{1}{\eta\mu} \log \frac{\log \beta}{\epsilon})$ for some $\beta \geq 1$, and $\mathcal{O}$ be a η-RGO for f_{oracle}. Algorithm 1, initialized at a β-warm start, runs in T iterations, each querying $\mathcal{O}$ a constant number of times, and obtains ϵ total variation distance to π.*

4. We note that the work Mou et al. (2019a) proposed a variant of Definition 1 with several additional requirements which we show are not necessary for applications such as composite sampling, discussed in more detail in Section 3.

In other words, if we can implement an η-RGO for a μ-strongly convex function f_{oracle} in time $\mathcal{T}_{\text{RGO}}$, we can sample from $\exp(-f_{\text{oracle}})$ in time $\widetilde{O}(\frac{1}{\eta\mu} \cdot \mathcal{T}_{\text{RGO}})$. To highlight the power of this reduction, suppose there was a sampler $\mathcal{A}$ for densities $\propto \exp(-f)$ with mixing time $\widetilde{O}(\kappa^{10}\sqrt{d})$, where $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth, μ-strongly convex, and has condition number $\kappa = \frac{L}{\mu}$ (cf. Section B.1 for definitions).[5] Choosing $\eta = \frac{1}{L}$ and $f_{\text{oracle}} \leftarrow f$ in Theorem 3 yields a sampler whose mixing time is $\widetilde{O}(\kappa \cdot \mathcal{T}_{\text{RGO}})$, where $\mathcal{T}_{\text{RGO}}$ is the cost of sampling from the density

$$\exp\left(-\frac{L}{2}\left\|x - v\right\|_2^2 - f(x)\right),$$

for some $v \in \mathbb{R}^d$. However, observe that this distribution has a negative log-density with constant condition number $\frac{L+L}{L+\mu} \leq 2$! By using $\mathcal{A}$ as our RGO, we have $\mathcal{T}_{\text{RGO}} = \widetilde{O}(\sqrt{d})$, and the overall mixing time is $\widetilde{O}(\kappa\sqrt{d})$. Leveraging Theorem 3 in applications, we obtain the following new results, improving mixing of various "base samplers" which we bootstrap as RGOs for regularized densities.

Well-conditioned densities. In Lee et al. (2020), it was shown that a variant of Metropolized HMC obtains a mixing time of $\widetilde{O}(\kappa d \log^3 \frac{\kappa d}{\epsilon})$ for sampling a density on $\mathbb{R}^d$ with condition number κ. The analysis of Lee et al. (2020) was delicate, and required reasoning about conditioning on a nonconvex set with desirable concentration properties. In Section D.1, we prove Corollary 4, improving Lee et al. (2020) by roughly a logarithmic factor with a significantly simpler analysis.

Corollary 4 *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp\left(-f(x)\right)$ such that f is L-smooth and μ-strongly convex, and let $\epsilon \in (0, 1)$, $\kappa = \frac{L}{\mu}$. Assume access to $x^* = \operatorname{argmin}_{x \in \mathbb{R}^d} f(x)$. Algorithm 1 with $\eta = \frac{1}{8Ld\log(\kappa)}$ using Algorithm 2 as a restricted Gaussian oracle for f uses $O(\kappa d \log \kappa \log \frac{\kappa d}{\epsilon})$ gradient queries in expectation, and obtains ϵ total variation distance to π.*

We include Corollary 4 as a warmup for our more complicated results, as a way to showcase the use of our reduction framework in a slightly different way than the one outlined earlier. In particular, in proving Corollary 4, we will choose a significantly smaller value of η, at which point a simple rejection sampling scheme implements each RGO with expected constant gradient queries.

We give another algorithm matching Corollary 4 with a deterministic query complexity bound as Corollary 22. The algorithm of Corollary 22 is interesting in that it is entirely a *zeroth-order* algorithm, and does not require access to a gradient oracle. To our knowledge, in the well-conditioned optimization setting, no zeroth-order query complexities better than roughly $\sqrt{\kappa}d$ are known, e.g. simulating accelerated gradient descent with a value oracle; thus, our sampling algorithm has a query bound off by only $\tilde{O}(\sqrt{\kappa})$ from the best-known optimization algorithm. We are hopeful this result may help in the search for query lower bounds for structured logconcave sampling.

Composite densities with a restricted Gaussian oracle. In Section E, we develop a sampler for densities on $\mathbb{R}^d$ proportional to $\exp(-f(x) - g(x))$, where f has condition number κ and g admits a restricted Gaussian oracle $\mathcal{O}$. We state its guarantees here.

Theorem 5 *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp\left(-f(x) - g(x)\right)$ such that f is L-smooth and μ-strongly convex, and let $\epsilon \in (0, 1)$. Let $\eta \leq \frac{1}{32L\kappa d\log(\kappa/\epsilon)}$ (where $\kappa = \frac{L}{\mu}$), $T =$*

5. No sampler with mixing time scaling as $\text{poly}(\kappa)\sqrt{d}$ is currently known.

$\Theta(\frac{1}{\eta\mu}\log(\frac{\kappa d}{\epsilon}))$*, and let* $\mathcal{O}$ *be a* η*-RGO for* g*. Assume access to* $x^* = \text{argmin}_{x\in\mathbb{R}^d}\{f(x) + g(x)\}$*. There is an algorithm which runs in* T *iterations in expectation, each querying a gradient oracle of* f *and* $\mathcal{O}$ *a constant number of times, and obtains* ϵ *total variation distance to* π*.*

The assumption that the composite component g admits an RGO can be thought of as a measure of "simplicity" of g. This mirrors the widespread use of a proximal oracle as a measure of simplicity in the context of composite optimization Beck and Teboulle (2009), which we now define.

Definition 6 (Proximal oracle) $\mathcal{O}(\lambda, v)$ *is a* proximal oracle *for convex* $g : \mathbb{R}^d \to \mathbb{R}$ *if it returns*

$$\mathcal{O}(\lambda, v) \leftarrow \text{argmin}_{x\in\mathbb{R}^d}\left\{\frac{1}{2\lambda}\|x - v\|_2^2 + g(x)\right\}.$$

Many regularizers g in defining composite optimization objectives, which are often used to enforce a quality such as sparsity or "simplicity" in a solution, admit efficient proximal oracles. In particular, if the proximal oracle subproblem admits a closed form solution (or otherwise is computable in $O(d)$ time), the regularized objective can be optimized at essentially no asymptotic loss. It is readily apparent that our RGO (Definition 1) is the extension of Definition 6 to the sampling setting. In Mou et al. (2019a), a variety of regularizations arising in practical applications including coordinate-separable g (such as restrictions to a coordinate-wise box, e.g. for a Bayesian inference task where we have side information on the ranges of parameters) and ℓ_1 or group Lasso regularized densities were shown to admit RGOs. Our composite sampling results achieve a similar "no loss" phenomenon for such regularizations, with respect to existing well-conditioned samplers.

By choosing the largest possible value of η in Theorem 5, we obtain an iteration bound of $\widetilde{O}(\kappa^2 d)$. In Section D.2, we prove Corollary 7, which improves Theorem 5 by roughly a κ factor.

Corollary 7 *Let* π *be a distribution on* $\mathbb{R}^d$ *with* $\frac{d\pi}{dx}(x) \propto \exp(-f(x) - g(x))$ *such that* f *is* L*-smooth and* μ*-strongly convex, and let* $\epsilon \in (0, 1)$*,* $\kappa = \frac{L}{\mu}$*. Assume access to* $x^* = \text{argmin}_{x\in\mathbb{R}^d}\{f(x)+g(x)\}$ *and let* $\mathcal{O}$ *be a RGO for* g*. There is an algorithm (Algorithm 1 using Theorem 5 as a RGO) which runs in* $O(\kappa d \log^3 \frac{\kappa d}{\epsilon})$ *iterations in expectation, each querying a gradient of* f *and* $\mathcal{O}$ *a constant number of times, and obtains* ϵ *total variation distance to* π*.*

To sketch the proof, choosing $\eta = \frac{1}{L}$ in Theorem 3 yields an algorithm running in $\widetilde{O}(\frac{1}{\eta\mu}) = \widetilde{O}(\kappa)$ iterations. In each iteration, the RGO asks to sample from the distribution with negative log-density $f(x) + g(x) + \frac{L}{2}\|x - v\|_2^2$ for some $v \in \mathbb{R}^d$, so we can call Theorem 5, where the "well-conditioned" portion $f(x) + \frac{L}{2}\|x - v\|_2^2$ has constant condition number. Thus, Theorem 5 runs in $\widetilde{O}(d)$ iterations to solve the subproblem, yielding the result. In fact, Corollary 7 nearly matches Corollary 4 in the case $g = 0$ uniformly. Surprisingly, this recovers the runtime of Lee et al. (2020) without appealing to gradient concentration bounds (Lee et al. (2020), Theorem 3.2).

Logconcave finite sums. In Section F, we initiate the study of mixing times for sampling logconcave finite sums with polylogarithmic dependence on accuracy. We give the following result.

Theorem 8 *Let* π *be a distribution on* $\mathbb{R}^d$ *with* $\frac{d\pi}{dx}(x) \propto \exp(-F(x))$*, where* $F(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ *is* μ*-strongly convex,* f_i *is* L*-smooth and convex* $\forall i \in [n]$*,* $\kappa = \frac{L}{\mu}$*, and* $\epsilon \in (0, 1)$*. Assume access to* $x^* = \text{argmin}_{x\in\mathbb{R}^d} F(x)$*. Algorithm 6 uses* $O\left(\kappa^2 d \log^4 \frac{n\kappa d}{\epsilon}\right)$ *value queries to summands* $\{f_i\}_{i\in[n]}$*, and obtains* ϵ *total variation distance to* π*.*

For a zeroth-order algorithm, Theorem 8 is a surprisingly strong baseline as it nearly matches the previous best bound for zeroth-order sampling when $n = 1$; however, when $\kappa \approx d$, the complexity is cubic. Using Theorem 8 in the framework of Theorem 3, we obtain the following improvement.

Corollary 9 (Improved first-order logconcave finite sum sampling) *In the setting of Theorem 8, Algorithm 1 using Algorithm 6 and SVRG Johnson and Zhang (2013) as a RGO for F uses*

$$O\left(n \log\left(\frac{n\kappa d}{\epsilon}\right) + \kappa\sqrt{nd}\log^{3.5}\left(\frac{n\kappa d}{\epsilon}\right) + \kappa d \log^{5}\left(\frac{n\kappa d}{\epsilon}\right)\right) = \widetilde{O}\left(n + \kappa \max\left(d, \sqrt{nd}\right)\right)$$

queries to first-order oracles for summands $\{f_i\}_{i\in[n]}$, and obtains ϵ total variation distance to π.

Corollary 9 has several surprising properties. First, its bound when $n = 1$ gives yet another way of nearly recovering the runtime of Lee et al. (2020) without gradient concentration. Second, up to a $\widetilde{O}(\max(1, \sqrt{\frac{n}{d}}))$ factor, it is the best runtime one could hope for without an improvement when $n = 1$. This is in the sense that $\widetilde{O}(\kappa d)$ is the best runtime for $n = 1$, and to our knowledge every efficient well-conditioned sampler requires minimizer access, i.e. $\widetilde{O}(n)$ gradient queries Woodworth and Srebro (2016). Interestingly, when $n = 1$, Algorithm 6 can be significantly simplified, and becomes the standard Metropolized random walk Dwivedi et al. (2018); this yields Corollary 22, an algorithm attaining the iteration complexity of Corollary 4 while only querying a value oracle for f.

3. Previous work

Logconcave sampling is a problem that, within the theoretical computer science field, has its origins in convex body sampling (a problem it generalizes). A long sequence of developments have made significant advances in the general model, where only logconcavity is assumed, and only value oracle access is given. In this prior work discussion, we focus on structured cases where all or part of the negative log-density has finite condition number, and refer the reader to Vempala (2005); Lovász and Vempala (2006); Cousins and Vempala (2015) regarding progress in the general case.

Well-conditioned densities. Significant recent efforts in the machine learning and statistics communities focused on obtaining provable guarantees for well-conditioned distributions, starting from pioneering work of Dalalyan (2017), and continued in e.g. Cheng et al. (2018); Dalalyan and Riou-Durand (2018); Chen and Vempala (2019); Chen et al. (2019); Dwivedi et al. (2018); Durmus and Moulines (2019); Durmus et al. (2019); Lee et al. (2018); Mou et al. (2019b); Shen and Lee (2019); Lee et al. (2020). In this setting, many methods based on discretizations of continuous-time first-order processes (such as the Langevin dynamics) have been proposed. Typically, error guarantees come in two forms: either in the 2-Wasserstein (W_2) distance, or in total variation (TV). The line Dwivedi et al. (2018); Chen et al. (2019); Lee et al. (2020) has brought the gradient complexity for obtaining ϵ TV distance to $\widetilde{O}(\kappa d)$ where d is the dimension, by exploiting gradient concentration properties. For progress in complexities depending polynomially on ϵ^{-1}, attaining W_2 guarantees (typically incomparable to TV bounds), we defer to Shen and Lee (2019), the state-of-the-art using $\widetilde{O}(\kappa^{\frac{7}{6}}\epsilon^{-\frac{1}{3}} + \kappa\epsilon^{-\frac{2}{3}})$ queries to obtain W_2 distance $\epsilon\sqrt{d\mu^{-1}}$ from the target.[6] We note incomparable guarantees can be obtained by assuming higher derivative bounds (e.g. a Lipschitz Hessian); our work uses only the minimal assumption of bounded second derivatives.

6. Here, $\sqrt{d\mu^{-1}}$ is the effective diameter; this accuracy measure allows for scale-invariant W_2 guarantees.

Composite densities. Recent works have studied sampling from densities of the form (1), or its specializations (e.g. restrictions to a convex set). Several Pereyra (2016); Brosse et al. (2017); Bernton (2018) are based on Moreau envelope or proximal regularization strategies, and demonstrate efficiency under more stringent assumptions on the structure of the composite function g, but under minimal assumptions obtain fairly large provable mixing times $\Omega(d^5)$. Proximal regularization algorithms have also been considered for non-composite sampling Wibisono (2019). Another discretization strategy based on projections was studied by Bubeck et al. (2018), but obtained mixing time $\Omega(d^7)$. Finally, improved algorithms for special constrained sampling problems have been proposed, such as simplex restrictions Hsieh et al. (2018).

Of particular relevance and inspiration to this work is Mou et al. (2019a). By generalizing and adapting Metropolized HMC algorithms of Dwivedi et al. (2018); Chen et al. (2019), adopting a Moreau envelope strategy, and using (a stronger version of) the RGO access model, Mou et al. (2019a) obtained a runtime which in the best case scales as $\tilde{O}\left(\kappa^2 d\right)$, similar to the guarantee of our Theorem 5. However, this result required a variety of additional assumptions, such as access to normalization factors of restricted Gaussians, Lipschitzness of g, a warm start, and various parameter tradeoffs. The general problem of sampling from (1) under minimal assumptions more efficiently than general-purpose logconcave algorithms is to the best of our knowledge unresolved (even under restricted Gaussian oracle access), a novel contribution of our mixing time bound. Our results also suggest that the RGO is a natural notion of tractability for the composite sampling problem.

Logconcave finite sums. Since Welling and Teh (2011) proposed the stochastic gradient Langevin dynamics, which at each step stochastically estimates the full gradient of the function, there has been a long line of work giving bounds for this method and other similar algorithms Dalalyan and Karagulyan (2019); Gao et al. (2018); Salim et al. (2019); Barkhagen et al. (2018); Nemeth and Fearnhead (2019). These convergence rates depend heavily on the variance of the stochastic estimates. Inspired by variance-reduced methods in convex optimization, samplers based on low-variance estimators have also been proposed Dubey et al. (2016); Durmus et al. (2016); Bierkens et al. (2019); Baker et al. (2019); Nagapetyan et al. (2017); Chen et al. (2017); Zou et al. (2018); Chatterji et al. (2018). Although our reduction-based approach is not designed specifically for solving problems of finite sum structure, our speedup can be viewed as due to a lower variance estimator implicitly defined through the oracle subproblems of Theorem 3 via repeated re-centering.

In Table 1, we list prior runtimes Zou et al. (2018); Chatterji et al. (2018) for sampling logconcave finite sums; note these results additionally require bounded higher derivatives (excepting the κ^4 dependence), obtain guarantees in Wasserstein distance, and depend polynomially on ϵ^{-1}. On the other hand, our reduction-based approach obtains total variation bounds with linear dependence on κ and polylogarithmic dependence on ϵ^{-1}. Our bound simultaneously matches the state-of-the-art bound when $n = 1$, a feature not shared by prior stochastic algorithms. To our knowledge, no previous nontrivial[7] bounds were previously known in the high-accuracy regime.

4. Technical overview

7. As mentioned previously, one can always compute the full ∇F in every iteration in a well-conditioned sampler.

Method	Gradient oracle complexity	
	$W_2 \le \epsilon$, $\mu = 1$	$W_2 \le \epsilon\sqrt{d\mu^{-1}}$
SAGA-LD Chatterji et al. (2018)	$n + \frac{\kappa^{1.5}\sqrt{d}+\kappa d+Md}{\epsilon} + \frac{\kappa d^{4/3}}{\epsilon^{2/3}}$	$n + \frac{\kappa^{1.5}+\kappa\sqrt{d}+M\sqrt{d}}{\epsilon} + \frac{\kappa d^{2/3}}{\epsilon^{2/3}}$
SVRG-LD Chatterji et al. (2018)	$n + \frac{\kappa^{1.5}\sqrt{d}+\kappa d+Md}{\epsilon} + \frac{\kappa d^{4/3}}{\epsilon^{2/3}}$	$n + \frac{\kappa^3}{\epsilon^2} + \frac{\kappa^{1.5}+M\sqrt{d}}{\epsilon}$
CV-ULD Chatterji et al. (2018)	$n + \frac{\kappa^4 d^{1.5}}{\epsilon^3}$	$n + \frac{\kappa^4}{\epsilon^3}$
SVRG-LD Zou et al. (2018)	$n + \frac{\kappa^{1.5}\sqrt{d}+Md}{\epsilon} + \frac{\kappa\sqrt{nd}}{\epsilon}$	$n + \frac{\kappa^{1.5}+M\sqrt{d}}{\epsilon} + \frac{\kappa\sqrt{n}}{\epsilon}$
State-of-the-art, $n = 1$ Shen and Lee (2019)	$\frac{\kappa^{7/6}d^{1/6}}{\epsilon^{1/3}} + \frac{\kappa d^{1/3}}{\epsilon^{2/3}}$	$\frac{\kappa^{7/6}}{\epsilon^{1/3}} + \frac{\kappa}{\epsilon^{2/3}}$

Method	Gradient oracle complexity (TV $\le \epsilon$)
Corollary 9	$n + \kappa d + \kappa\sqrt{nd}$
State-of-the-art, $n = 1$ Lee et al. (2020)	κd

Table 1: Complexity of sampling from $e^{-F(x)}$ where $F(x) = \frac{1}{n}\sum_{i\in[n]} f_i(x)$ on $\mathbb{R}^d$ is μ-strongly convex, each f_i is convex and L-smooth, and $\kappa = \frac{L}{\mu}$. When relevant, M is the Lipschitz constant of the Hessian $\nabla^2 F$, which our algorithm has no dependence on. Complexity is measured by the number of calls to f_i or ∇f_i for summands $\{f_i\}_{i\in[n]}$. We hide polylog$(\frac{n\kappa d}{\epsilon})$ factors for simplicity.

4.1. Composite logconcave sampling

We study the problem of approximately sampling from a distribution π on $\mathbb{R}^d$, with density

$$\frac{d\pi(x)}{dx} \propto \exp\left(-f(x) - g(x)\right). \tag{1}$$

Here, $f : \mathbb{R}^d \to \mathbb{R}$ is assumed to be "well-behaved" (i.e. has finite condition number), and $g : \mathbb{R}^d \to \mathbb{R}$ is a convex, but possibly non-smooth function. This problem generalizes the special case of sampling from $\exp(-f(x))$ for well-conditioned f, simply by letting g vanish. Even the specialization of (1) where g indicates a convex set (i.e. is 0 inside the set, and ∞ outside) is not well-understood; existing mixing time bounds for this restricted setting are large polynomials in d Brosse et al. (2017); Bubeck et al. (2018), and are typically weaker than guarantees in the general logconcave setting Lovász and Vempala (2006a,b). This is in contrast to the convex optimization setting, where first-order methods readily generalize to solve problem families such as $\min_{x\in\mathcal{X}} f(x)$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex set, as well as its generalization

$$\min_{x\in\mathbb{R}^d} f(x) + g(x), \text{ where } g : \mathbb{R}^d \to \mathbb{R} \text{ is convex and admits a proximal oracle.} \tag{2}$$

We defined proximal oracles in Definition 6; in short, they are prodecures which minimize the sum of a quadratic and g. Definition 6 is desirable as many natural non-smooth composite objectives arising in learning settings, such as the Lasso Tibshirani (1996) and elastic net Zou and Hastie (2005), admit efficient proximal oracles. The definition of a proximal oracle implies it can also handle arbitrary sums of linear functions and quadratics, as the resulting function can be rewritten as the sum of a constant and a single quadratic. The seminal work Beck and Teboulle (2009) extends fast gradient methods to solve (2) via proximal oracles, and has prompted many follow-up studies.

Motivated by the success of the proximal oracle framework in convex optimization, we study sampling from the family (1) through the lens of RGOs, a natural extension of Definition 6. The main result of Section E is a "base" algorithm efficiently sampling from (1), assuming access to an RGO for g. We now survey the main components of this algorithm.

Reduction to shared minimizers. We first observe that without loss of generality, f and g share a minimizer: by shifting f and g by linear terms, i.e. $\tilde{f}(x) := f(x) - \langle \nabla f(x^*), x \rangle$, $\tilde{g}(x) := g(x) + \langle \nabla f(x^*), x \rangle$, where x^* minimizes $f + g$, first-order optimality implies both $\tilde{f}$ and $\tilde{g}$ are minimized by x^*. Moreover, implementation of a first-order oracle for $\tilde{f}$ and an RGO for $\tilde{g}$ are immediate. This modification becomes crucial for our later developments, and we hope this simple observation, reminiscent of "variance reduction" techniques in stochastic optimization Johnson and Zhang (2013), is broadly applicable to improving algorithms for the sampling problem (1).

Beyond Moreau envelopes: expanding the space. A typical approach in convex optimization in handling non-smooth objectives g is to instead optimize its *Moreau envelope*, defined by

$$g^{\eta}(y) := \min_{x \in \mathbb{R}^d} \left\{ g(x) + \frac{1}{2\eta} \|x - y\|_2^2 \right\}. \tag{3}$$

Intuitively, the envelope g^{η} trades off function value with proximity to y; a standard exercise shows that g^{η} is smooth (has a Lipschitz gradient), with smoothness depending on η, and moreover that computing gradients of g^{η} reduces to calling a proximal oracle (Definition 6). It is natural to extend this idea to the composite sampling setting, e.g. via sampling from the density $\exp\left(-f(x) - g^{\eta}(x)\right)$. However, a variety of complications prevent such strategies from obtaining rates comparable to their smooth counterparts, including difficulty in bounding closeness of the resulting distribution, as well as biased drifts of the sampling process due to error in gradients.

Our approach departs from this smoothing strategy in a crucial way, inspired by Hamiltonian Monte Carlo (HMC) methods Kramers (1940); Neal (2011). HMC can be seen as a discretization of the ubiquitous Langevin dynamics, on an expanded space. In particular, discretizations of Langevin dynamics simulate the stochastic differential equation $\frac{dx_t}{dt} = -\nabla f(x_t) + \sqrt{2}\frac{dW_t}{dt}$, where W_t is Brownian motion. HMC methods instead simulate dynamics on an extended space $\mathbb{R}^d \times \mathbb{R}^d$, via an auxiliary "velocity" variable which accumulates gradient information. This is sometimes interpreted as a discretization of the underdamped Langevin dynamics Cheng et al. (2018). HMC often has desirable stability properties, and expanding the dimension via an auxiliary variable has been used in algorithms obtaining the fastest rates in the well-conditioned logconcave sampling regime Shen and Lee (2019); Lee et al. (2020). Inspired by this phenomenon, we consider the density on $\mathbb{R}^d \times \mathbb{R}^d$

$$\frac{d\hat{\pi}}{dz}(z) := \exp\left(-f(y) - g(x) - \frac{1}{2\eta} \|x - y\|_2^2 \right) \text{ where } z = (x, y). \tag{4}$$

Due to technical reasons, the family of distributions we use in our final algorithms are of slightly different form than (4), but this simplification is useful to build intuition. Note in particular that the form of (4) is directly inspired by (3), where rather than maximizing over x, we directly expand the space. The idea is that for small enough η and a set on x of large measure, smoothness of f will guarantee that the marginal of (4) on x will concentrate y near x, a fact we make rigorous. To sample from (1), we then show that a rejection filter applied to a sample x from the marginal of (4) will terminate in constant steps. Consequently, it suffices to develop a fast sampler for (4).

Alternating sampling with an oracle. The form of the distribution (4) suggests a natural strategy for sampling from it: starting from a current state (x_k, y_k), we iterate

1. Sample $y_{k+1} \sim \exp\left(-f(y) - \frac{1}{2\eta}\|x_k - y\|_2^2\right)$.
2. Sample $x_{k+1} \sim \exp\left(-g(x) - \frac{1}{2\eta}\|x - y_{k+1}\|_2^2\right)$, via a restricted Gaussian oracle.

When f and g share a minimizer, taking a first-order approximation in the first step, i.e. sampling $y_{k+1} \sim \exp(-f(x_k) - \langle \nabla f(x_k), y - x_k\rangle - \frac{1}{2\eta}\|y - x_k\|_2^2)$, can be shown to generalize the `Leapfrog` step of HMC updates. However, for η very small, we observe the first step itself reduces to the case of sampling from a distribution with constant condition number, performable in $\tilde{O}(d)$ gradient calls by e.g. Metropolized HMC Dwivedi et al. (2018); Chen et al. (2019); Lee et al. (2020). Moreover, it is not hard to see that this "alternating marginal" sampling strategy preserves the stationary distribution, so no filtering is necessary. Directly bounding the conductance of this walk, for small enough η, leads to an algorithm running in $\tilde{O}\left(\kappa^2 d^2\right)$ iterations, each calling an RGO once, and a gradient oracle for f roughly $\tilde{O}\left(d\right)$ times. This latter guarantee appeals to known bounds Chen et al. (2019); Lee et al. (2020) on the mixing time of Metropolized HMC for a well-conditioned distribution, a property satisfied by the y-marginal of (4) for small η.

Stability of Gaussians under bounded perturbations. To obtain our tightest runtime result, we use that η is chosen to be much smaller than L^{-1} to show structural results about distributions of the form (4), yielding tighter concentration for bounded perturbations of a Gaussian (i.e. the Gaussian has covariance $\frac{1}{\eta}\mathbf{I}$, and is restricted by L-smooth f for $\eta \ll L^{-1}$). To illustrate, let

$$\frac{d\mathcal{P}_x(y)}{dy} \propto \exp\left(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2\right)$$

and let its mean and mode be $\bar{y}_x$, y_x^*. It is standard that $\|\bar{y}_x - y_x^*\|_2 \le \sqrt{d\eta}$, by η^{-1}-strong logconcavity of $\mathcal{P}_x$. Informally, we show that for $\eta \ll L^{-1}$ and x not too far from the minimizer of f, we can improve this to $\|\bar{y}_x - y_x^*\|_2 = O(\sqrt{\eta})$; see Proposition 63 for a precise statement.

Using our structural results, we sharpen conductance bounds, improve the warmness of a starting distribution, and develop a simple rejection sampling scheme for sampling the y variable in expected constant gradient queries. Our proofs are continuous in flavor and based on gradually perturbing the Gaussian and solving a differential inequality; we believe they may of independent interest. These improvements lead to an algorithm running in $\tilde{O}\left(\kappa^2 d\right)$ iterations; ultimately, we use our reduction framework, stated in Theorem 3, to improve this dependence to $\tilde{O}\left(\kappa d\right)$.

4.2. Logconcave finite sums

We initiate the algorithmic study of the following task in the high-accuracy regime: sample $x \sim \pi$ within total variation distance ϵ, where $\frac{d\pi}{dx}(x) \propto \exp(-F(x))$ and

$$F(x) = \frac{1}{n}\sum_{i\in[n]} f_i(x), \tag{5}$$

all $f_i : \mathbb{R}^d \to \mathbb{R}$ are convex and L-smooth, and F is μ-strongly convex. We call such a distribution π a (well-conditioned) *logconcave finite sum*. In applications (where summands correspond to points in a dataset, e.g. in Bayesian linear and logistic regression tasks Dwivedi et al. (2018)), querying ∇F may be prohibitively expensive, so a natural goal is to obtain bounds on the number of required

queries to summands ∇f_i for $i \in [n]$. This motivation also underlies the development of stochastic gradient methods in optimization, a foundational tool in modern statistics and data processing. Naïvely, one can complete the task by using existing samplers for well-conditioned distributions and querying the full gradient ∇F in each iteration, resulting in a summand gradient query complexity of $\widetilde{O}(n\kappa d)$ Lee et al. (2020). Many recent works, inspired from recent developments in the complexity of optimizing a well-conditioned finite sum, have developed subsampled gradient methods for the sampling problem. However, to our knowledge, all such guarantees depend polynomially on the accuracy ϵ and are measured in the 2-Wasserstein distance; in the high-accuracy, total variation case no nontrivial query complexity is currently known.

We show in Section F that given access to the minimizer of F, a simple zeroth-order algorithm querying $\widetilde{O}(\kappa^2 d)$ values of summands $\{f_i\}_{i\in[n]}$ succeeds (i.e. it never requires a full value or gradient query of F). The algorithm is essentially the Metropolized random walk proposed in Dwivedi et al. (2018) for the $n = 1$ case with a cheap subsampled filter. Notably, because the random walk is conducted with respect to F, we cannot efficiently query the function value at any point; nonetheless, by sampling to compute a nearly-unbiased estimator of the rejection probability, we do not incur too much error. This random walk was shown in Chen et al. (2019) to mix in $\widetilde{O}(\kappa^2 d)$ iterations; we implement each step to sufficient accuracy using $\widetilde{O}(1)$ function evaluations.

It is natural to ask if first-order information can be used to improve this query complexity, through e.g. "variance reduction" techniques (e.g. Johnson and Zhang (2013)) developed for stochastic optimization. The idea behind variance reduction is to recenter gradient estimates in phases, occasionally computing full gradients to improve the estimate quality. One fundamental difficulty which arises from using variance reduction in high-accuracy sampling is that the resulting algorithms are not *stateless*. By this, we mean that the variance-reduced estimates depend on the algorithm history, and thus it is difficult to ascertain correctness of the stationary distribution. We take a different approach to achieve a linear query dependence on κ, which we now describe.

4.3. Proximal point reduction framework

To motivate Theorem 3, we first recast existing "proximal point" reduction-based optimization methods through the lens of composite optimization, and subsequently show that similar ideas underlying our composite sampler in Section 4.1 yield an analagous "proximal point reduction framework" for sampling. Chronologically, our composite sampler (originally announced in Shen et al. (2020)) predates our reduction framework, which was then inspired by the perspective given here. We hope these insights prove fruitful for further development of proximal approaches to sampling.

Proximal point methods as composite optimization. Proximal point methods are a powerful tool in optimization, developed by Rockafellar (1976); cf. Parikh and Boyd (2014) for a modern survey. The idea is that to minimize convex $F : \mathbb{R}^d \to \mathbb{R}$, it suffices to solve a sequence of subproblems

$$x_{k+1} \leftarrow \text{argmin}_{x\in\mathbb{R}^d} \left\{ F(x) + \frac{1}{2\lambda} \|x - x_k\|_2^2 \right\}. \tag{6}$$

Intuitively, by tuning the parameter $\lambda \geq 0$, we trade off how regularized the subproblems (6) are with how rapidly the overall method converges. Smaller values of λ result in larger regularization amounts which are amenable to algorithms for minimizing well-conditioned objectives.

For optimizing functions of the form (5) via stochastic gradient estimates to ϵ error, Johnson and Zhang (2013); Defazio et al. (2014); Schmidt et al. (2017) developed variance-reduced methods

obtaining a query complexity of $\widetilde{O}(n+\kappa)$. To match a known lower bound of $\widetilde{O}(n+\sqrt{n\kappa})$ due to Woodworth and Srebro (2016), two works Lin et al. (2015); Frostig et al. (2015) appropriately applied instances of accelerated proximal point methods Guler (1992) with careful analyses of how accurately subproblems (6) needed to be solved. These algorithms black-box called the $\widetilde{O}(n+\kappa)$ runtime as an oracle to solve the subproblems (6) for an appropriate choice of λ, obtaining an accelerated rate.[8] To shed some light on this acceleration procedure, we adopt an alternative view on proximal point methods.[9] Consider the following known composite optimization result.

Proposition 10 (Informal statement of Beck and Teboulle (2009)) *Let $f:\mathbb{R}^d\to\mathbb{R}$ be L-smooth and μ-strongly convex, and $g:\mathbb{R}^d\to\mathbb{R}$ admit a proximal oracle $\mathcal{O}(\lambda,v)$ (cf. Definition 6). There is an algorithm taking $\widetilde{O}(\sqrt{\kappa})$ iterations for $\kappa=\frac{L}{\mu}$ to find an ϵ-approximate minimizer to $f+g$, each querying ∇f and $\mathcal{O}$ a constant number of times. Further, $\lambda=\frac{1}{L}$ in all calls to $\mathcal{O}$.*

Ignoring subtleties of the error tolerance of $\mathcal{O}$, we show how to use an instance of Proposition 10 to recover the $\widetilde{O}(n+\sqrt{n\kappa})$ query complexity for optimizing (5). Let $f(x)=\frac{\mu}{2}\|x\|_2^2$, and $g=F-f$. For any $\Lambda\geq\mu$, f is both μ-strongly convex and Λ-smooth. Moreover, note that all calls to the proximal oracle $\mathcal{O}$ for g require solving subproblems of the form

$$\text{argmin}_{x\in\mathbb{R}^d}\left\{F(x)-\frac{\mu}{2}\|x\|_2^2+\frac{\Lambda}{2}\|x-v\|_2^2\right\}. \tag{7}$$

The upshot of choosing a smoothness bound $\Lambda\geq\mu$ is that the regularization amount in (7) increases, improving the conditioning of the subproblem, which is Λ-strongly convex and $L+\Lambda$-smooth. The algorithm of e.g. Johnson and Zhang (2013) solves each subproblem (7) in $\widetilde{O}(n+\frac{L+\Lambda}{\Lambda})$ gradient queries, leading to an overall query complexity (for Proposition 10) of $\widetilde{O}\left(\sqrt{\frac{\Lambda}{\mu}}\cdot\left(n+\frac{L}{\Lambda}\right)\right)$. Optimizing over $\Lambda\geq\mu$, i.e. taking $\Lambda=\max(\mu,\frac{L}{n})$, yields the desired bound of $\widetilde{O}(n+\sqrt{n\kappa})$.

Applications to sampling. In Sections E and F, we develop samplers for structured families with quadratic dependence on the conditioning κ. The proximal point approach suggests a strategy for accelerating these runtimes. Namely, if there is a framework which repeatedly calls a sampler for a regularized density (analogous to calls to (6)), one could trade off the regularization with the rate of the outer loop. Fortunately, in the spirit of interpreting proximal point methods as composite optimization, the composite sampler of Section E itself meets these reduction framework criteria.

We briefly recall properties of our composite sampler here. Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x)\propto\exp(-f_{\text{wc}}(x)-f_{\text{oracle}}(x))$,[10] where f_{wc} is well-conditioned (has finite condition number κ) and f_{oracle} admits an RGO, which solves subproblems of the form

$$\mathcal{O}(\eta,v)\sim\text{the density proportional to }\exp\left(-\frac{1}{2\eta}\|x-v\|_2^2-f_{\text{oracle}}(x)\right). \tag{8}$$

The algorithm of Section E only calls $\mathcal{O}$ with a fixed η; note the strong parallel between the RGO subproblem and the proximal oracle of Proposition 10. For a given value of $\eta\geq 0$, our composite

8. We note that an improved runtime without extraneous logarithmic factors was later obtained by Allen-Zhu (2017).
9. This perspective can also be found in the lecture notes Lee (2018).
10. To disambiguate, we sometimes also use the notation $f_{\text{wc}}+f_{\text{oracle}}$ rather than $f+g$ in defining instances of our reduction framework or composite samplers, when convenient in the context.

sampler runs in $\widetilde{O}(\frac{1}{\eta\mu})$ iterations, each requiring a call to $\mathcal{O}$. Smaller η improve the conditioning of the negative log-density of subproblem (8), but increase the overall iteration count, yielding a range of trade-offs. The algorithm of Section E has an upper bound requirement on η (cf. Theorem 5); in Section C, we observe that this may be lifted when $f_{\text{wc}} = 0$ uniformly, allowing for a full range of choices. Moreover, the analysis of the composite sampler becomes much simpler when $f_{\text{wc}} = 0$, as in Theorem 3. We give the framework as Algorithm 1, as well as a full (fairly short) convergence analysis. By trading off the regularization amount with the cost of implementing (8) via bootstrapping base samplers, we obtain a host of improved runtimes.

Beyond our specific applications, the framework we provide has strong implications as a generic reduction from mixing times scaling polynomially in κ to improved methods scaling linearly in κ. This is akin to the observation in Lin et al. (2015) that accelerated proximal point methods generically improve $\text{poly}(\kappa)$ dependences to $\sqrt{\kappa}$ dependences for optimization. We are optimistic this insight will find further implications in the logconcave sampling literature.

Acknowledgments

YL and RS are supported by NSF awards CCF-1749609, CCF-1740551, DMS-1839116, and DMS-2023166, a Microsoft Research Faculty Fellowship, a Sloan Research Fellowship, and a Packard Fellowship. KT is supported by NSF CAREER Award CCF-1844855 and a PayPal research gift.

References

Jacob D. Abernethy and Elad Hazan. Faster convex optimization: Simulated annealing with an efficient universal barrier. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 2520–2528, 2016.

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *J. Mach. Learn. Res.*, 18:221:1–221:51, 2017.

Jack Baker, Paul Fearnhead, Emily B Fox, and Christopher Nemeth. Control variates for stochastic gradient mcmc. *Statistics and Computing*, 29(3):599–615, 2019.

M Barkhagen, NH Chau, É Moulines, M Rásonyi, S Sabanis, and Y Zhang. On stochastic gradient langevin dynamics with dependent data streams in the logconcave case. *arXiv preprint arXiv:1812.02709*, 2018.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2(1):183–202, 2009.

Espen Bernton. Langevin monte carlo and JKO splitting. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, pages 1777–1798, 2018.

Dimitris Bertsimas and Santosh S. Vempala. Solving convex programs by random walks. *J. ACM*, 51(4):540–556, 2004.

Joris Bierkens, Paul Fearnhead, Gareth Roberts, et al. The zig-zag process and super-efficient sampling for bayesian analysis of big data. *The Annals of Statistics*, 47(3):1288–1320, 2019.

Nicolas Brosse, Alain Durmus, Eric Moulines, and Marcelo Pereyra. Sampling from a log-concave distribution with compact support with proximal langevin monte carlo. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 319–342, 2017.

Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected langevin monte carlo. *Discret. Comput. Geom.*, 59(4):757–783, 2018.

Niladri Chatterji, Nicolas Flammarion, Yian Ma, Peter Bartlett, and Michael Jordan. On the theory of variance reduction for stochastic gradient monte carlo. In *International Conference on Machine Learning*, pages 764–773, 2018.

Changyou Chen, Wenlin Wang, Yizhe Zhang, Qinliang Su, and Lawrence Carin. A convergence analysis for a class of practical variance-reduction stochastic gradient mcmc. *arXiv preprint arXiv:1709.01180*, 2017.

Yuansi Chen, Raaz Dwivedi, Martin J. Wainwright, and Bin Yu. Fast mixing of metropolized hamiltonian monte carlo: Benefits of multi-step gradients. *CoRR*, abs/1905.12247, 2019.

Zongchen Chen and Santosh S. Vempala. Optimal convergence rate of hamiltonian monte carlo for strongly logconcave distributions. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2019, September 20-22, 2019, Massachusetts Institute of Technology, Cambridge, MA, USA*, pages 64:1–64:12, 2019.

Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped langevin MCMC: A non-asymptotic analysis. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, pages 300–323, 2018.

Ben Cousins and Santosh S. Vempala. Gaussian cooling and $o^{*}(n^3)$ algorithms for volume and gaussian volume. *SIAM J. Comput.*, 47(3):1237–1273, 2018.

Benjamin Cousins and Santosh Vempala. Bypassing kls: Gaussian cooling and an oˆ*(n3) volume algorithm. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 539–548, 2015.

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

Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

Arnak S. Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic langevin diffusions. *CoRR*, abs/1807.09382, 2018.

Aaron Defazio, Francis R. Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1646–1654, 2014.

Kumar Avinava Dubey, Sashank J Reddi, Sinead A Williamson, Barnabas Poczos, Alexander J Smola, and Eric P Xing. Variance reduction in stochastic gradient langevin dynamics. In *Advances in neural information processing systems*, pages 1154–1162, 2016.

Alain Durmus and Éric Moulines. High-dimensional bayesian inference via the unadjusted langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Alain Durmus, Umut Simsekli, Eric Moulines, Roland Badeau, and Gaël Richard. Stochastic gradient richardson-romberg markov chain monte carlo. In *Advances in Neural Information Processing Systems*, pages 2047–2055, 2016.

Alain Durmus, Szymon Majewski, and Blazej Miasojedow. Analysis of langevin monte carlo via convex optimization. *J. Mach. Learn. Res.*, 20:73:1–73:46, 2019.

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

Martin E. Dyer, Alan M. Frieze, and Ravi Kannan. A random polynomial time algorithm for approximating the volume of convex bodies. *J. ACM*, 38(1):1–17, 1991.

Roy Frostig, Rong Ge, Sham M. Kakade, and Aaron Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *Proceedings of the*

32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, pages 2540–2548, 2015.

Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu. Global convergence of stochastic gradient hamiltonian monte carlo for non-convex stochastic optimization: Non-asymptotic performance bounds and momentum-based acceleration. *arXiv preprint arXiv:1809.04618*, 2018.

Sharad Goel, Ravi Montenegro, and Prasad Tetali. Mixing time bounds via the spectral profile. *Electronic Journal of Probability*, 11:1–26, 2006.

Osman Guler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.

Gilles Hargé. A convex/log-concave correlation inequality for gaussian measure and an application to abstract wiener spaces. *Probability theory and related fields*, 130(3):415–440, 2004.

Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored langevin dynamics. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 2883–2892, 2018.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 315–323, 2013.

Ravi Kannan, László Lovász, and Ravi Montenegro. Blocking conductance and mixing in random walks. *Combinatorics, Probability & Computing*, 15(4):541–570, 2006.

Hendrik Anthony Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940.

Yin Tat Lee. Lecture 8: Stochastic methods and applications. Class notes, UW CSE 599: Interplay between Convex Optimization and Geometry, 2018.

Yin Tat Lee, Zhao Song, and Santosh S. Vempala. Algorithmic theory of odes and sampling from well-conditioned logconcave densities. *CoRR*, abs/1812.06243, 2018.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Logsmooth gradient concentration and tighter runtimes for metropolized hamiltonian monte carlo. In *Conference on Learning Theory, COLT 2020*, 2020.

David Asher Levin, Yuval Peres, and Elizabeth Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2009.

Hongzhou Lin, Julien Mairal, and Zaïd Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3384–3392, 2015.

László Lovász and Ravi Kannan. Faster mixing via average conductance. In *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing, May 1-4, 1999, Atlanta, Georgia, USA*, pages 282–287, 1999.

László Lovász and Santosh Vempala. Simulated annealing in convex bodies and an o*(n4) volume algorithm. *Journal of Computer and System Sciences*, 72(2):392–417, 2006.

László Lovász and Santosh S. Vempala. Hit-and-run from a corner. *SIAM J. Comput.*, 35(4):985–1005, 2006a.

László Lovász and Santosh S. Vempala. Fast algorithms for logconcave functions: Sampling, rounding, integration and optimization. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings*, pages 57–68, 2006b.

Wenlong Mou, Nicolas Flammarion, Martin J. Wainwright, and Peter L. Bartlett. An efficient sampling algorithm for non-smooth composite potentials. *CoRR*, abs/1910.00551, 2019a.

Wenlong Mou, Yi-An Ma, Martin J. Wainwright, Peter L. Bartlett, and Michael I. Jordan. High-order langevin diffusion yields an accelerated MCMC algorithm. *CoRR*, abs/1908.10859, 2019b.

Tigran Nagapetyan, Andrew B Duncan, Leonard Hasenclever, Sebastian J Vollmer, Lukasz Szpruch, and Konstantinos Zygalakis. The true cost of stochastic gradient langevin dynamics. *arXiv preprint arXiv:1706.02692*, 2017.

Radford M Neal. Mcmc using hamiltonian dynamics. *Handbook of Markov chain Monte Carlo*, 2 (11):2, 2011.

Christopher Nemeth and Paul Fearnhead. Stochastic gradient markov chain monte carlo. *arXiv preprint arXiv:1907.06986*, 2019.

Yurii Nesterov. A method for solving a convex programming problem with convergence rate $o(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983.

Neal Parikh and Stephen P. Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239, 2014.

Marcelo Pereyra. Proximal markov chain monte carlo algorithms. *Stat. Comput.*, 26(4):745–760, 2016.

R Tyrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

Adil Salim, Dmitry Koralev, and Peter Richtárik. Stochastic proximal langevin algorithm: Potential splitting and nonasymptotic rates. In *Advances in Neural Information Processing Systems*, pages 6653–6664, 2019.

Mark Schmidt, Nicolas Le Roux, and Francis R. Bach. Minimizing finite sums with the stochastic average gradient. *Math. Program.*, 162(1-2):83–112, 2017.

Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems*, pages 2100–2111, 2019.

Ruoqi Shen, Kevin Tian, and Yin Tat Lee. Composite logconcave sampling with a restricted gaussian oracle. *CoRR*, abs/2006.05976, 2020.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1):267–288, 1996.

Santosh Vempala. Geometric random walks: A survey. *MSRI Combinatorial and Computational Geometry*, 52:573–612, 2005.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

Andre Wibisono. Proximal langevin algorithm: Rapid convergence under isoperimetry. *CoRR*, abs/1911.01469, 2019.

Blake E. Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3639–3647, 2016.

Difan Zou, Pan Xu, and Quanquan Gu. Subsampled stochastic variance-reduced gradient langevin dynamics. In *International Conference on Uncertainty in Artificial Intelligence*, 2018.

Hui Zou and Trevor Hastie. Regularization and variable section via the elastic net. *Journal of the Royal Statistical Society, Series B (Methodological)*, 67(2):301–320, 2005.

Appendix A. Roadmap

The appendix contains detailed proofs omitted from the main body. It can naturally be divided into four (mostly self-contained) parts, which can be read in any desired order (after notations and preliminaries are discussed in Section B), but we give a suggested order of reading here.

Proximal reduction framework. In Section C we give our framework for bootstrapping samplers, and prove its correctness (Theorem 3). This framework is a generic reduction from sampling a desired density to calling a sampler for a number of recentered regularized densities. It has generic use in improving conditioning dependence (and trade off other parameters). Assuming the "base samplers" of Theorems 5 and 8, in Section D we apply our reduction to obtain our strongest guarantees for the structured densities considered in this paper: Corollaries 4, 7, and 9.

Base composite sampler. In Section E, we give the base composite sampler developed in this paper (proving Theorem 5). We defer technical parts of its analysis to Sections H and I.

Base finite sum sampler. In Section F, we give a self-contained exposition of the base finite sum sampler developed in this paper (proving Theorem 8).

Technical components. In Section G, we discuss tolerance to various forms of approximation error, as well as the assumption of access to the mode of the distribution. In Section J, we prove several structural tools which we repeatedly use throughout the paper.

Appendix B. Preliminaries

B.1. Notation

General notation. For $d \in \mathbb{N}$, $[d]$ refers to the set of naturals $1 \le i \le d$; $\|\cdot\|_2$ is the Euclidean norm on $\mathbb{R}^d$ when d is clear from context. $\mathcal{N}(\mu, \mathbf{\Sigma})$ is the multivariate Gaussian of specified mean and variance, $\mathbf{I}$ is the identity of appropriate dimension when clear from context, and $\preceq$ is the Loewner order on symmetric matrices.

Functions. We say twice-differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and μ-strongly convex if $\mu\mathbf{I} \preceq \nabla^2 f(x) \preceq L\mathbf{I}$ for all $x \in \mathbb{R}^d$; it is well-known that L-smoothness implies that f has an L-Lipschitz gradient, and that for any $x, y \in \mathbb{R}^d$,

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2 \le f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2 .$$

If f is L-smooth and μ-strongly convex, we say it has a condition number $\kappa := \frac{L}{\mu}$. We call a zeroth-order oracle, or "value oracle", an oracle which returns $f(x)$ on any input point $x \in \mathbb{R}^d$; similarly, a first-order oracle, or "gradient oracle", returns both the value and gradient $(f(x), \nabla f(x))$.

Distributions. We call distribution π on $\mathbb{R}^d$ logconcave if $\frac{d\pi}{dx}(x) = \exp(-f(x))$ for convex f; π is μ-strongly logconcave if f is μ-strongly convex. For $A \subseteq \mathbb{R}^d$, A^c is its complement, and we let $\pi(A) := \int_{x \in A} d\pi(x)$. We say distribution ρ is β-warm with respect to π if $\frac{d\pi}{d\rho}(x) \le \beta$ everywhere, and define the total variation $\|\pi - \rho\|_{\mathrm{TV}} := \sup_{A \subseteq \mathbb{R}^d} \pi(A) - \rho(A)$. We will frequently use the fact that $\|\pi - \rho\|_{\mathrm{TV}}$ is also the probability that $x \sim \pi$ and $x' \sim \rho$ are unequal under the best coupling of

(π, ρ); this allows us to "locally share randomness" when comparing two random walk procedures. We define the expectation $\mathbb{E}_\pi$ and variance Var_π with respect to distribution π in the standard way,

$$\mathbb{E}_\pi[h(x)] := \int h(x) d\pi(x), \ \mathrm{Var}_\pi[h(x)] := \mathbb{E}_\pi\left[(h(x))^2\right] - (\mathbb{E}_\pi[h(x)])^2.$$

Structured distributions. This work considers two types of distributions with additional structure, which we call *composite logconcave densities* and *logconcave finite sums*. A composite logconcave density has the form $\exp(-f(x) - g(x))$, where both f and g are convex. In this context throughout, f will either be uniformly 0 or have a finite condition number (be "well-conditioned"), and g will represent a "simple" but possibly non-smooth function, as measured by admitting an RGO (cf. Definition 1). We will sometimes refer to the components as f and g as f_{wc} and f_{oracle} respectively, to disambiguate when the functions f and g are already defined in context. In our reduction-based approaches, we have additional structure on the parameter λ which an RGO is called with (cf. Definition 2). Specifically, in our instances typically $\lambda^{-1} \gg L$ (or some other "niceness" parameter associated with the negative log-density); this can be seen as heavily regularizing the negative log-density, and often makes the implementation simpler.

Finally, a logconcave finite sum has density of the form $\exp(-F(x))$ where $F(x) = \frac{1}{n}\sum_{i\in[n]} f_i(x)$. When treating such densities, we make the assumption that each constituent summand f_i is L-smooth and convex, and the overall function F is μ-strongly convex. We measure complexity of algorithms for logconcave finite sums by gradient queries to summands, i.e. $\nabla f_i(x)$ for some $i \in [n]$.

Optimization. Throughout this work, we are somewhat liberal with assuming access to minimizers to various functions (namely, the negative log-densities of target distributions). We give a more thorough discussion of this assumption in Section G, but note here that for all function families we consider (well-conditioned, composite, and finite sum), efficient first-order methods exist for obtaining high accuracy minimizers, and this optimization query complexity is never the leading-order term in any of our algorithms assuming polynomially bounded initial error.

B.2. Technical facts

We will repeatedly use the following results.

Fact 1 (Gaussian integral) *For any $\lambda \geq 0$ and $v \in \mathbb{R}^d$,*

$$\int \exp\left(-\frac{1}{2\lambda}\|x - v\|_2^2\right) dx = (2\pi\lambda)^{\frac{d}{2}}.$$

Fact 2 (Dwivedi et al. (2018), Lemma 1) *Let π be a μ-strongly logconcave distribution, and let x^* minimize its negative log-density. Then, for $x \sim \pi$ and any $\delta \in [0, 1]$, with probability at least $1 - \delta$,*

$$\|x - x^*\|_2 \leq \sqrt{\frac{d}{\mu}}\left(2 + 2\max\left(\sqrt[4]{\frac{\log(1/\delta)}{d}}, \sqrt{\frac{\log(1/\delta)}{d}}\right)\right).$$

Fact 3 (Hargé (2004), Theorem 1.1) *Let π be a μ-strongly logconcave density. Let $d\gamma_\mu(x)$ be the Gaussian density with covariance matrix $\mu^{-1}\mathbf{I}$. For any convex function h,*

$$\mathbb{E}_\pi[h(x - \mathbb{E}_\pi[x])] \leq \mathbb{E}_{\gamma_\mu}[h(x - \mathbb{E}_{\gamma_\mu}[x])].$$

Fact 4 (Durmus and Moulines (2019), Theorem 1) *Let π be a μ-strongly logconcave distribution, and let x^* minimize its negative log-density. Then, $\mathbb{E}_\pi[\|x - x^*\|_2^2] \le \frac{d}{\mu}$.*

Appendix C. Proximal reduction framework

The reduction framework of Theorem 3 can be thought of as a specialization of a more general composite sampler which we develop in Section E, whose guarantees are reproduced here.

Theorem 5 *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-f(x) - g(x))$ such that f is L-smooth and μ-strongly convex, and let $\epsilon \in (0, 1)$. Let $\eta \le \frac{1}{32L\kappa d \log(\kappa/\epsilon)}$ (where $\kappa = \frac{L}{\mu}$), $T = \Theta(\frac{1}{\eta\mu}\log(\frac{\kappa d}{\epsilon}))$, and let $\mathcal{O}$ be a η-RGO for g. Assume access to $x^* = \operatorname{argmin}_{x\in\mathbb{R}^d}\{f(x) + g(x)\}$. There is an algorithm which runs in T iterations in expectation, each querying a gradient oracle of f and $\mathcal{O}$ a constant number of times, and obtains ϵ total variation distance to π.*

Our main observation, elaborated on more formally for specific applications in Section D, is that a variety of structured logconcave densities have negative log-densities f_{oracle}, where we can implement an efficient restricted Gaussian oracle for f_{oracle} via calling an existing sampling method. Crucially, in these instantiations we use the fact that the distributions which $\mathcal{O}$ is required to sampled from are heavily regularized (restricted by a quadratic with large leading coefficient) to obtain fast samplers. We further note that the upper bound requirement on η in Theorem 5 can be lifted when the "well-conditioned" component is uniformly 0. By setting $f = 0$ and $g = f_{\text{oracle}}$ in Theorem 5, and refining the analysis for this special case to tolerate arbitrary η, we obtain Theorem 3, which follows straightforwardly from the analysis in Section E. We provide a self-contained proof here for convenience, as this particular structure (the composite setting where f_{wc} is uniformly zero and f_{oracle} is strongly convex) admits significant simplifications from the more general case.

Theorem 3 *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-f_{\text{oracle}}(x))$ such that f_{oracle} is μ-strongly convex, and let $\epsilon \in (0, 1)$. Let $\eta > 0$, $T = \Theta(\frac{1}{\eta\mu}\log\frac{\log\beta}{\epsilon})$ for some $\beta \ge 1$, and $\mathcal{O}$ be a η-RGO for f_{oracle}. Algorithm 1, initialized at a β-warm start, runs in T iterations, each querying $\mathcal{O}$ a constant number of times, and obtains ϵ total variation distance to π.*

For simplicity of notation, we replace f_{oracle} in the statement of Theorem 3 with g throughout just this section. Let π be a density on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-g(x))$ where g is μ-strongly convex (but possibly non-smooth), and let $\mathcal{O}$ be a restricted Gaussian oracle for g. Consider the joint distribution $\hat{\pi}$ supported on an expanded space $z = (x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ with density, for some $\eta > 0$,

$$\frac{d\hat{\pi}}{dz}(z) \propto \exp\left(-g(x) - \frac{1}{2\eta}\|x - y\|_2^2\right).$$

Note that the x-marginal of $\hat{\pi}$ is precisely π, so it suffices to sample from the x-marginal. We consider a simple alternating Markov chain for sampling from $\hat{\pi}$, described in the following Algorithm 1.

By observing that the distributions π_x and π_y in the above method are precisely the marginal distributions of $\hat{\pi}$ with one variable fixed, it is straightforward to see that $\hat{\pi}$ is a stationary distribution of the process. We make this formal in the following lemma.

Algorithm 1 `AlternateSample`(g, η, T)

1: **Input:** μ-strongly convex $g : \mathbb{R}^d \to \mathbb{R}$, $\eta > 0$, $T \in \mathbb{N}$, x_0 drawn from a β-warm distribution for π
2: **for** $k \in [T]$ **do**
3: Sample $y_k \sim \pi_{x_{k-1}}$, where for all $x \in \mathbb{R}^d$, $\frac{d\pi_x}{dy}(y) \propto \exp\left(-\frac{1}{2\eta}\|x-y\|_2^2\right)$.
4: Sample $x_k \sim \pi_{y_k}$, where for all $y \in \mathbb{R}^d$, $\frac{d\pi_y}{dx}(x) \propto \exp\left(-g(x) - \frac{1}{2\eta}\|x-y\|_2^2\right)$.
5: **end for**
6: **return** x_T

Lemma 11 (Alternating marginal sampling) *Let $\hat{\pi}$ be a density on two blocks (x, y). Sample $(x, y) \sim \hat{\pi}$, and then sample $\tilde{x} \sim \hat{\pi}(\cdot, y)$, $\tilde{y} \sim \hat{\pi}(\tilde{x}, \cdot)$. Then, the distribution of $(\tilde{x}, \tilde{y})$ is $\hat{\pi}$.*

Proof The density of the resulting distribution at $(\tilde{x}, y)$ is proportional to the product of the (marginal) density at y and the conditional distribution of $\tilde{x} \mid y$, which by definition is $\hat{\pi}$. Therefore, $(\tilde{x}, y)$ is distributed as $\hat{\pi}$, and the argument for $\tilde{y}$ follows symmetrically. ■

Next, let $\mathcal{P}_x$ be the density of y_k after one step of the above procedure starting from $x_{k-1} = x$, and let $\mathcal{T}_x$ be the resulting density of x_k. We first make the following simplifying observation.

Observation 1 *For any two points x, $x' \in \mathbb{R}^d$, $\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\mathrm{TV}} \le \|\mathcal{P}_x - \mathcal{P}_{x'}\|_{\mathrm{TV}}$.*

Proof This follows by the coupling characterization of total variation (see e.g. Chapter 5 of Levin et al. (2009)). Per the optimal coupling of $y \sim \mathcal{P}_x$ and $y' \sim \mathcal{P}_{x'}$, whenever the total variation sets $y = y'$ in Line 2 of `AlternateSample`, we can couple the resulting distributions in Line 3 as well. ■

In order to prove Theorem 3, we follow the well-established average conductance framework studied and generalized by a long line of works (e.g. Lovász and Kannan (1999); Kannan et al. (2006); Goel et al. (2006)). We state a mixing time bound due to Chen et al. (2019) here which is convenient for our purposes. In Section H, we will use a more fine-grained variant of Proposition 12, which is broken up as Propositions 44 and 45.

Proposition 12 (Lemma 1, Lemma 2, Chen et al. (2019)) *Let a random walk with a μ-strongly logconcave stationary distribution π on $x \in \mathbb{R}^d$ have transition distributions $\{\mathcal{T}_x\}_{x \in \mathbb{R}^d}$. For some $\epsilon \in [0, 1]$, let convex set $\Omega \subseteq \mathbb{R}^d$ have $\pi(\Omega) \ge 1 - \frac{\epsilon^2}{2\beta^2}$. Let π_{start} be a β-warm start for π, and let the algorithm be initialized at $x_0 \sim \pi_{\text{start}}$. Suppose for any $x, x' \in \Omega$ with $\|x - x'\|_2 \le \Delta$,*

$$\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\mathrm{TV}} \le \frac{7}{8}. \tag{9}$$

Then, the random walk mixes to total variation distance within ϵ of π in $O(\frac{1}{\Delta^2 \mu} \log \frac{\log \beta}{\epsilon})$ iterations.

In the following, we will set $\Omega = \mathbb{R}^d$. By combining Proposition 12 with our earlier Observation 1, it suffices to bound the total variation distance between $\mathcal{P}_x$ and $\mathcal{P}_{x'}$ for nearby points x, x', which is done in the following.

Lemma 13 *Let $x, x' \in \mathbb{R}^d$ with $\|x - x'\|_2 \le \sqrt{\eta}$. Then $\|\mathcal{P}_x - \mathcal{P}_{x'}\|_{\mathrm{TV}} \le \frac{1}{2}$.*

Proof Note that $\mathcal{P}_x$ as defined previously is simply the distribution $\mathcal{N}(x, \eta\mathbf{I})$; similarly, $\mathcal{P}_{x'}$ is $\mathcal{N}(x', \eta\mathbf{I})$. It is a standard calculation that the KL divergence between these two distributions (multivariate Gaussians with equal covariances) is

$$\frac{\|x - x'\|_2^2}{2\eta} \le \frac{1}{2}.$$

Here we used the assumed bound on $\|x - x'\|_2$. The result follows from Pinsker's inequality. ■

Finally, we put these pieces together to obtain a proof of Theorem 3.

Proof [Proof of Theorem 3] Let the transition distributions $\{\mathcal{T}_x\}_{x\in\mathbb{R}^d}$ be defined by one step of Lines 2 and 3 of Algorithm 1. By combining Observation 1 with Lemma 13, the requirement (9) of Proposition 12 is satisfied for $\Delta = \sqrt{\eta}$. The conclusion follows immediately upon overloading $\hat{\pi}$ to mean its x-marginal, which we noted earlier is the same distribution as the desired π. Finally, we note that all calls to $\mathcal{O}$ are in updating the x variable, and indeed have $\lambda = \eta$. ■

We note that Theorem 3 is robust to a small amount of error tolerance in the sampler $\mathcal{O}$. Specifically, if $\mathcal{O}$ has tolerance $\frac{\epsilon}{2T}$, then by calling Theorem 3 with desired accuracy $\frac{\epsilon}{2}$ and adjusting constants appropriately, the cumulative error incurred by all calls to $\mathcal{O}$ is within the total requisite bound (formally, this can be shown via the coupling characterization of total variation). We defer a more formal elaboration on this inexactness argument to Section G and the proof of Proposition 26.

Appendix D. Tighter runtimes for structured densities

In this section, we use applications of Theorem 3 to obtain simple analyses of novel state-of-the-art high-accuracy runtimes for the well-conditioned densities studied in Dwivedi et al. (2018); Chen et al. (2019); Lee et al. (2020), as well as the composite and finite sum densities studied in this work. We will assume the conclusions of Theorems 5 and 8 respectively in deriving the results of Sections D.2 and D.3.

D.1. Well-conditioned logconcave sampling: proof of Corollary 4

In this section, let π be a distribution on $\mathbb{R}^d$ with density proportional to $\exp(-f(x))$, where f is L-smooth and μ-strongly convex (and $\kappa = \frac{L}{\mu}$) and has pre-computed minimizer x^*. We will instantiate Theorem 3 with $f_{\mathrm{oracle}}(x) = f(x)$, and choose $\eta = \frac{1}{8Ld\log(\kappa)}$. Applying Lemma 18 (given in the following section) with no composite term implies sampling from $\mathcal{N}(x^*, \frac{1}{L}\mathbf{I})$ yields a $\beta = \kappa^{\frac{d}{2}}$-warm start for π. We now require an η-RGO $\mathcal{O}$ for $f_{\mathrm{oracle}} = f$ to use in Theorem 3.

Our implementation of $\mathcal{O}$ is a rejection sampling scheme. We use the following helpful guarantee.

Lemma 14 (Rejection sampling) *Let π, $\hat{\pi}$ be distributions on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto p(x)$, $\frac{d\hat{\pi}}{dx}(x) \propto \hat{p}(x)$. Suppose for some $C \ge 1$ and all $x \in \mathbb{R}^d$, $\frac{p(x)}{\hat{p}(x)} \le C$. The following is termed "rejection sampling": repeat independent runs of the following procedure until a point is outputted.*

1. *Draw $x \sim \hat{\pi}$.*

2. *With probability* $\frac{p(x)}{C\hat{p}(x)}$*, output* x*.*

Rejection sampling terminates in $\frac{C\int \hat{p}(x)dx}{\int p(x)dx}$ *runs in expectation, and the output distribution is* π*.*

Proof The second claim follows from Bayes' rule which implies the conditional density of the output point is proportional to $\hat{p}(x) \cdot \frac{p(x)}{C\hat{p}(x)} \propto p(x)$, so the distribution is π. To see the first claim, the probability any sample outputs is

$$\int_x \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}(x) = \frac{1}{C}\int_x \frac{\int_x p(x)dx}{\int_x \hat{p}(x)dx} d\pi(x) = \frac{\int_x p(x)dx}{C\int_x \hat{p}(x)dx}.$$

The conclusion follows by independence and linearity of expectation. ■

We further state a concentration bound shown first in Lee et al. (2020) regarding the norm of the gradient of a point drawn from a logsmooth distribution.

Proposition 15 (Logsmooth gradient concentration, Corollary 3.3, Lee et al. (2020)) *Let* π *be a distribution on* $\mathbb{R}^d$ *with* $\frac{d\pi}{dx}(x) \propto \exp(-f(x))$ *where* f *is convex and* L*-smooth. With probability at least* $1 - \kappa^{-d}$*,*

$$\|\nabla f(x)\|_2 \le 3\sqrt{L}d\log\kappa \text{ for } x \sim \pi. \tag{10}$$

By the requirements of Theorem 3, the restricted Gaussian oracle $\mathcal{O}$ only must be able to draw samples from densities of the form, for some $y \in \mathbb{R}^d$,

$$\exp\left(-f_{\text{oracle}}(x) - \frac{1}{2\eta}\|x-y\|_2^2\right) = \exp\left(-f(x) - 4Ld\log\kappa\,\|x-y\|_2^2\right). \tag{11}$$

We will use the following Algorithm 2 to implement $\mathcal{O}$.

Algorithm 2 `XSample`(f, y, η)

Input: L-smooth, μ-strongly convex $f : \mathbb{R}^d \to \mathbb{R}$, $y \in \mathbb{R}^d$, $\eta > 0$
if $\|\nabla f(y)\|_2 \le 3\sqrt{L}d\log\kappa$ **then**
  while true do
    Draw $x \sim \mathcal{N}(y - \nabla f(y), \eta\mathbf{I})$
    $\tau \sim \text{Unif}[0,1]$
    if $\tau \le \exp(f(y) + \langle\nabla f(y), x-y\rangle - f(x))$ **then**
      return x
    end if
  end while
end if
Use Chen et al. (2019) to sample x from (11) to total variation distance $\frac{\epsilon}{\Theta(\kappa d^2\log^3(\frac{\kappa d}{\epsilon}))}$ using $O(d\log\frac{\kappa d}{\epsilon})$ queries to ∇f (Theorem 1, Chen et al. (2019), where (11) has constant condition number)
return x

Lemma 16 *Let $\eta = \frac{1}{8Ld\log(\kappa)}$, and suppose y satisfies the bound in (10), i.e. $\|\nabla f(y)\|_2 \le 3\sqrt{L}d\log\kappa$. Then, Line 3 of Algorithm 2 runs an expected 2 times, and Algorithm 2 samples exactly from (11), whenever the condition of Line 1 is met.*

Proof Note that when the assumption of Line 1 is met, Algorithm 2 is an instantiation of rejection sampling (Lemma 14) with

$$p(x) = \exp\left(-f(x) - \frac{1}{2\eta}\|x-y\|_2^2\right),$$
$$\hat{p}(x) = \exp\left(-f(y) - \langle\nabla f(y), x-y\rangle - \frac{1}{2\eta}\|x-y\|_2^2\right).$$

By convexity, we may take $C = 1$. Next, by applying Fact 1 twice and L-smoothness of f_{oracle},

$$\begin{aligned}\int_x p(x)dx &\ge \int_x \exp\left(-f(y) - \langle\nabla f(y), x-y\rangle - \frac{1+\eta L}{2\eta}\|x-y\|_2^2\right)dx\\ &= \exp\left(-f(y) + \frac{\eta}{2(1+\eta L)}\|\nabla f(y)\|_2^2\right)\int_x \exp\left(-\frac{1+\eta L}{2\eta}\left\|x-y+\frac{\eta}{1+\eta L}\nabla f(y)\right\|_2^2\right)dx\\ &= \exp\left(-f(y) + \frac{\eta}{2(1+\eta L)}\|\nabla f(y)\|_2^2\right)\left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}},\\ \int_x \hat{p}(x)dx &= \exp\left(-f(y) + \frac{\eta}{2}\|\nabla f(y)\|_2^2\right)(2\pi\eta)^{\frac{d}{2}},\end{aligned}$$

which implies the desired bound (recalling Lemma 14 and our assumed bound on $\|\nabla f(y)\|_2$)

$$\begin{aligned}\frac{\int\hat{p}(x)dx}{\int p(x)dx} &\le \exp\left(\left(\frac{\eta}{2} - \frac{\eta}{2(1+\eta L)}\right)\|\nabla f(y)\|_2^2\right)(1+\eta L)^{\frac{d}{2}}\\ &\le 1.5\exp\left(\frac{\eta^2 L}{2(1+\eta L)}\|\nabla f(y)\|_2^2\right) \le 2.\end{aligned}$$

■

We are now equipped to prove our main result concerning well-conditioned densities.

Corollary 17 *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-f(x))$ such that f is L-smooth and μ-strongly convex, and let $\epsilon \in (0,1)$, $\kappa = \frac{L}{\mu}$. Assume access to $x^* = \operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$. Algorithm 1 with $\eta = \frac{1}{8Ld\log(\kappa)}$ using Algorithm 2 as a restricted Gaussian oracle for f uses $O(\kappa d\log\kappa\log\frac{\kappa d}{\epsilon})$ gradient queries in expectation, and obtains ϵ total variation distance to π.*

Proof By applying Theorem 3 with the chosen η and $\beta = \kappa^{\frac{d}{2}}$ using the warm start from Lemma 18, and noting that the cumulative error due to all calls to Line 10 cannot amount to more than $\frac{\epsilon}{2}$ total variation error throughout the algorithm, it suffices to show that Algorithm 2 uses $O(1)$ gradient queries each iteration in expectation. This happens whenever the condition in Line 1 is met via Lemma 16, so we must show Line 10 is executed with probability $O((d\log\frac{\kappa d}{\epsilon})^{-1})$.

To show this, note that combining Proposition 15 with the warmness of the start x_0 in Algorithm 2, this event occurs with probability at most $\kappa^{-\frac{d}{2}}$ in the first iteration.[11] Since warmness is monotonically decreasing[12] throughout using an exact oracle in Algorithm 1, and the total error accumulated due to Line 10 throughout the algorithm is $O((d \log \frac{\kappa d}{\epsilon})^{-1})$, we have the desired conclusion. ■

We show a bound nearly-matching Corollary 4 using only value access to f, and with a deterministic iteration complexity (rather than an expected one), as Corollary 22 in Section D.3.

D.2. Composite logconcave sampling: proof of Corollary 7

In this section, let π be a distribution on $\mathbb{R}^d$ with density proportional to $\exp(-f(x) - g(x))$, where f is L-smooth and μ-strongly convex (and $\kappa = \frac{L}{\mu}$), and g is convex and admits a restricted Gaussian oracle $\mathcal{O}$. Without loss of generality, we assume that f and g share a minimizer x^* which we have pre-computed; if this is not the case, we can redefine $f(x) \leftarrow f(x) - \langle \nabla f(x^*), x \rangle$ and $g(x) \leftarrow g(x) + \langle \nabla f(x^*), x \rangle$; see Section E.1 for this reduction.

We will instantiate Theorem 3 with $f_{\text{oracle}} = f + g$, which is a μ-strongly convex function. In order to apply Theorem 3, we give a warm start and an η-restricted Gaussian oracle for f_{oracle}.

Lemma 18 *Define the distribution π_{start} on $\mathbb{R}^d$ with density*

$$\frac{d\pi_{\text{start}}}{dx}(x) \propto \exp\left(-f(x^*) - \frac{L}{2}\|x - x^*\|_2^2 - g(x)\right).$$

Then, π_{start} is a β-warm distribution for π with $\beta = \kappa^{\frac{d}{2}}$.

Proof We wish to show that for all $x \in \mathbb{R}^d$,

$$\frac{d\pi_{\text{start}}}{d\pi}(x) = \frac{\exp\left(-f(x^*) - \frac{L}{2}\|x - x^*\|_2^2 - g(x)\right)}{\exp\left(-f(x) - g(x)\right)} \cdot \frac{\int \exp(-f(x) - g(x))dx}{\int \exp\left(-f(x^*) - \frac{L}{2}\|x - x^*\|_2^2 - g(x)\right)dx} \leq \beta.$$

First, by smoothness and the assumption that x^* minimizes f (e.g. $\nabla f(x^*) = 0$), $-f(x^*) - \frac{L}{2}\|x - x^*\|_2^2 \leq -f(x)$, so it suffices to bound the ratio of the normalization constants. Note that

$$\frac{\int \exp(-f(x) - g(x))dx}{\int \exp\left(-f(x^*) - \frac{L}{2}\|x - x^*\|_2^2 - g(x)\right)dx} \leq \frac{\int \exp(-f(x^*) - \frac{\mu}{2}\|x - x^*\|_2^2 - g(x))dx}{\int \exp\left(-f(x^*) - \frac{L}{2}\|x - x^*\|_2^2 - g(x)\right)dx}.$$

Finally, applying Proposition 61, where we recall $\frac{\mu}{2}\|x - x^*\|_2^2 + g(x)$ has minimizer x^* and is μ-strongly convex, we have the desired

$$\frac{d\pi_{\text{start}}}{d\pi}(x) \leq \frac{\int \exp(-f(x^*) - \frac{\mu}{2}\|x - x^*\|_2^2 - g(x))dx}{\int \exp\left(-f(x^*) - \frac{L}{2}\|x - x^*\|_2^2 - g(x)\right)dx} \leq \kappa^{\frac{d}{2}}.$$

11. Formally, Line 2 of Algorithm 1 has $y_1 \sim \mathcal{N}(x_0, \eta\mathbf{I})$, but by smoothness $\|\nabla f(y_1)\|_2 \leq \|\nabla f(x_0)\|_2 + L\|x - y\|_2$ and $L\|x - y\|_2 \leq \widetilde{O}(L\sqrt{\eta})$ with high probability, adding a negligible constant to the bound of Proposition 15.
12. This is a standard fact in the literature, and can be seen as follows: each transition step in the chain is a convex combination of warm point masses, preserving warmness.

■

Our main result of this section follows directly from combining Theorem 3 with Lemma 18, and using Theorem 5 as the required oracle $\mathcal{O}$, stated more precisely in the following.

Corollary 19 *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-f(x) - g(x))$ such that f is L-smooth and μ-strongly convex, and let $\epsilon \in (0,1)$, $\kappa = \frac{L}{\mu}$. Assume access to $x^* = \operatorname{argmin}_{x\in\mathbb{R}^d}\{f(x)+g(x)\}$ and let $\mathcal{O}$ be a RGO for g. There is an algorithm (Algorithm 1 using Theorem 5 as a RGO) which runs in $O(\kappa d \log^3 \frac{\kappa d}{\epsilon})$ iterations in expectation, each querying a gradient of f and $\mathcal{O}$ a constant number of times, and obtains ϵ total variation distance to π.*

Proof As discussed at the beginning of this section, assume without loss that f and g both are minimized by x^*. We apply the algorithm of Theorem 3 with $\eta = \frac{1}{L}$ to the μ-strongly convex function $f + g$ and the warm start from Lemma 18 with $\beta = \kappa^{\frac{d}{2}}$, which requires one call to $\mathcal{O}$ to implement. Thus, the iteration count parameter in Theorem 3 is $T = O(\kappa \log \frac{\kappa d}{\epsilon})$.

Recall that we chose $\eta = \frac{1}{L}$. To bound the total complexity of this algorithm, it suffices to give an η-RGO $\mathcal{O}^+$ for sampling from distributions with densities of the form, for some $y \in \mathbb{R}^d$,

$$\exp\left(-f(x) - g(x) - \frac{1}{2\eta}\|x-y\|_2^2\right) = \exp\left(-f(x) - g(x) - \frac{L}{2}\|x-y\|_2^2\right)$$

to total variation distance $\frac{\epsilon}{\Theta(T)}$ (see discussion at the end of Section C). To this end, we apply Theorem 5 with the well-conditioned component $f(x) + \frac{L}{2}\|x-y\|_2^2$, the composite component $g(x)$, and the largest possible choice of η. Note that we indeed have access to a restricted Gaussian oracle for g (namely, $\mathcal{O}$), and this choice of well-conditioned component is $2L$-smooth and L-strongly convex, so its condition number is a constant. Thus, Theorem 5 requires $O(d \log^2 \frac{\kappa d}{\epsilon})$ calls to $\mathcal{O}$ and gradients of f to implement the desired $\mathcal{O}^+$ on any query y (where we note $\frac{\epsilon}{\Theta(T)} = \frac{1}{\text{poly}(\kappa,d,\epsilon^{-1})}$). Combining these complexity bounds yields the desired conclusion. ■

D.3. Sampling logconcave finite sums: proof of Corollary 9

In this section, let π be a distribution on $\mathbb{R}^d$ with density proportional to $\exp(-F(x))$, where $F(x) = \frac{1}{n}\sum_{i\in[n]} f_i(x)$ is μ-strongly convex, and for all $i \in [n]$, f_i is L-smooth (and $\kappa = \frac{L}{\mu}$). We will instantiate Theorem 3 with $f_{\text{oracle}}(x) = F(x)$, and Theorem 8 as an η-RGO for some choice of η.

More precisely, Theorem 8 shows that given access to the minimizer x^*, only zeroth-order access to the summands of F is necessary to obtain the iteration bound. In order to obtain the minimizer to high accuracy however, variance reduced stochastic gradient methods (e.g. Johnson and Zhang (2013)) require $\Omega(n+\kappa)$ gradient queries, which amounts to $\Omega((n+\kappa)d)$ function evaluations. We state a convenient corollary of Theorem 8 which removes the requirement of accessing x^*, via an optimization pre-processing step using the method of Johnson and Zhang (2013) (see further discussion in Section G). This is useful to us in proving Theorem 9 because in the sampling tasks required by the RGO, the minimizer changes (and thus must be recomputed every time).

Corollary 20 (First-order logconcave finite sum sampling) *In the setting of Theorem 8, using Johnson and Zhang (2013) to precompute the minimizer x^* and running Algorithm 6 uses $O(n \log \frac{\kappa d}{\epsilon} +$*

$\kappa^2 d \log^4 \frac{n\kappa d}{\epsilon}$) *first-order oracle queries to summands* $\{f_i\}_{i\in[n]}$ *and obtains* ϵ *total variation distance to* π.

We now apply the reduction framework developed in Section B to our Algorithm 6 to obtain an improved query complexity for sampling from logconcave finite sums.

Corollary 21 (Improved first-order logconcave finite sum sampling) *In the setting of Theorem 8, Algorithm 1 using Algorithm 6 and SVRG Johnson and Zhang (2013) as a RGO for* F *uses*

$$O\left(n \log\left(\frac{n\kappa d}{\epsilon}\right) + \kappa\sqrt{nd}\log^{3.5}\left(\frac{n\kappa d}{\epsilon}\right) + \kappa d \log^5\left(\frac{n\kappa d}{\epsilon}\right)\right) = \widetilde{O}\left(n + \kappa \max\left(d, \sqrt{nd}\right)\right)$$

queries to first-order oracles for summands $\{f_i\}_{i\in[n]}$, *and obtains* ϵ *total variation distance to* π.

Proof We use the warm start $x_0 \sim \mathcal{N}(x^*, \frac{1}{L}\mathbf{I})$ which obtains warmness $\beta = \kappa^{\frac{d}{2}}$ via Lemma 18. We then apply Theorem 3 with $f_{\text{oracle}} = F(x)$, using Algorithm 6 as the required η-RGO $\mathcal{O}$ for sampling from distributions with densities of the form

$$\exp\left(-F(x) - \frac{1}{\eta}\|x - y\|_2^2\right)$$

for some $y \in \mathbb{R}^d$, to total variation $\frac{\epsilon}{\Theta(T)}$ (see Section C) for T the iteration bound of Algorithm 1. We apply Theorem 8 to the function $\widetilde{F}(x) = F(x) + \frac{1}{\eta}\|x - y\|_2^2$; we can express this in finite sum form by adding $\frac{1}{\eta}\|x - y\|_2^2$ to every constituent function, and the effect on gradient oracles is $\frac{1}{\eta}(x - y)$. Note $\widetilde{F}$ has condition number $O(1 + \eta L)$. For a given η, the overall complexity is

$$\frac{\log\frac{\kappa d}{\epsilon}}{\eta\mu}\left(n\log\left(\frac{n\kappa d}{\epsilon}\right) + d\log^4\left(\frac{n\kappa d}{\epsilon}\right) + (\eta L)^2 d\log^4\left(\frac{n\kappa d}{\epsilon}\right)\right)$$

Here, the inner loop complexity uses Corollary 20 to also find the minimizer (for warm starts), and the outer loop complexity is by Theorem 3. The result follows by optimizing over η, namely picking $\eta = \max(\frac{1}{L}, \sqrt{\frac{n}{L^2 d \log^3(n\kappa d/\epsilon)}})$, and that Algorithm 1 always must have at least one iteration. ■

Note the only place that Corollary 9 used gradient evaluations was in determining minimizers of subproblems, via the first step of Corollary 20. Consider now the $n = 1$ case. By running e.g. accelerated gradient descent for smooth and strongly convex functions, it is well-known Nesterov (1983) that we can obtain a minimizer in $\widetilde{O}(\sqrt{\kappa})$ iterations, each querying a gradient oracle, where κ is the condition number. By smoothness, we can approximate every coordinate of the gradient to arbitrary precision using 2 function evaluations, so this is a $\widetilde{O}(\sqrt{\kappa}d)$ value oracle complexity.

Finally, for every optimization subproblem in Corollary 9 where $\eta = (L \cdot \text{polylog}\frac{\kappa d}{\epsilon})^{-1}$, the condition number is a constant, which amounts to a $\widetilde{O}(d)$ value oracle complexity for computing a minimizer. This is never the dominant term compared to Theorem 8, yielding the following conclusion.

Corollary 22 *In the setting of Corollary 4, Algorithm 1 using Algorithm 6 as a restricted Gaussian oracle uses* $O(\kappa d \log^2 \frac{\kappa d}{\epsilon})$ *value queries and obtains* ϵ *total variation distance to* π.

We note that the polylogarithmic factor is significantly improved when compared to Corollary 9 by removing the random sampling steps in Algorithm 6. A precise complexity bound of the resulting Metropolized random walk, a zeroth-order algorithm mixing in $O(\kappa^2 d \log \frac{\kappa d}{\epsilon})$ for a logconcave distribution with condition number κ, is given as Theorem 2 of Chen et al. (2019).

Finally, in the case $n \geq 1$, we also exhibit an improved query complexity in terms of an entirely zeroth-order sampling algorithm which interpolates with Corollary 22 (up to logarithmic factors). By trading off the $\widetilde{O}(nd+\kappa d)$ zeroth-order complexity of minimizing a finite sum function Johnson and Zhang (2013), and the $\widetilde{O}(\kappa^2 d)$ zeroth-order complexity of sampling, we can run Theorem 3 for the optimal choice of $\eta = \widetilde{O}(\frac{\sqrt{n}}{L})$. The overall zeroth-order complexity can be seen to be $\widetilde{O}(nd + \sqrt{n}\kappa d)$.

Appendix E. Composite logconcave sampling with a restricted Gaussian oracle

In this section, we provide our "base sampler" for composite logconcave densities as Algorithm 3, and give its guarantees by proving Theorem 5. Throughout, fix distribution π with density

$$\frac{d\pi}{dx}(x) \propto \exp\left(-f(x) - g(x)\right), \text{where } f : \mathbb{R}^d \to \mathbb{R} \text{ is } L\text{-smooth, } \mu\text{-strongly convex,} \\ \text{and } g : \mathbb{R}^d \to \mathbb{R} \text{ admits a restricted Gaussian oracle } \mathcal{O}. \tag{12}$$

We will define $\kappa := \frac{L}{\mu}$, and assume that we have precomputed $x^* := \text{argmin}_{x \in \mathbb{R}^d} \{f(x) + g(x)\}$. Our algorithm proceeds in stages following the outline in Section 4.1.

1. `Composite-Sample` is reduced to `Composite-Sample-Shared-Min`, which takes as input a distribution with negative log-density $f + g$, where f and g share a minimizer; this reduction is given in Section E.1, and the remainder of the section handles the shared-minimizer case.

2. The algorithm `Composite-Sample-Shared-Min` is a rejection sampling scheme built on top of sampling from a joint distribution $\hat{\pi}$ on $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ whose x-marginal approximates π. We give this reduction in Section E.2.

3. The bulk of our analysis is for `Sample-Joint-Dist`, an alternating marginal sampling algorithm for sampling from $\hat{\pi}$. To implement marginal sampling, it alternates calls to $\mathcal{O}$ and a rejection sampling algorithm `YSample`. We prove its correctness in Section E.3.

We put these pieces together in Section E.4 to prove Theorem 5. We remark that for simplicity, we will give the algorithms corresponding to the largest value of step size η in the theorem statement; it is straightforward to modify the bounds to tolerate smaller values of η, which will cause the mixing time to become correspondingly larger (in particular, the value of K in Algorithm 5).

E.1. Reduction from `Composite-Sample` to `Composite-Sample-Shared-Min`

Correctness of `Composite-Sample` is via the following properties.

Proposition 23 *Let $\tilde{f}$ and $\tilde{g}$ be defined as in `Composite-Sample`.*

1. *The density $\propto \exp(-f(x) - g(x))$ is the same as the density $\propto \exp(-\tilde{f}(x) - \tilde{g}(x))$.*

Algorithm 3 `Composite-Sample`(π, x^*, ϵ)

Input: Distribution π of form (12), x^* minimizing negative log-density of π, $\epsilon \in [0, 1]$.
Output: Sample x from a distribution π' with $\|\pi' - \pi\|_{\mathrm{TV}} \le \epsilon$.
$\tilde{f}(x) \leftarrow f(x) - \langle \nabla f(x^*), x \rangle$, $\tilde{g}(x) \leftarrow g(x) + \langle \nabla f(x^*), x \rangle$
return `Composite-Sample-Shared-Min`$(\pi, \tilde{f}, \tilde{g}, x^*, \epsilon)$

Algorithm 4 `Composite-Sample-Shared-Min`$(\pi, f, g, x^*, \epsilon)$

Input: Distribution π of form (12), where f and g are both minimized by x^*, $\epsilon \in [0, 1]$.
Output: Sample x from a distribution π' with $\|\pi' - \pi\|_{\mathrm{TV}} \le \epsilon$.
while true do
  Define the set

$$\Omega := \left\{ x \mid \|x - x^*\|_2 \le 4\sqrt{\frac{d \log(288\kappa/\epsilon)}{\mu}} \right\} \tag{13}$$

  $x \leftarrow$ `Sample-Joint-Dist`$(f, g, x^*, \mathcal{O}, \frac{\epsilon}{18})$
  if $x \in \Omega$ **then**
    $\tau \sim \mathrm{Unif}[0, 1]$
    $y \leftarrow$ `YSample`(f, x, η)
    $\alpha \leftarrow \exp\left(f(y) - \langle \nabla f(x), y - x \rangle - \frac{L}{2}\|y - x\|_2^2 + g(x) + \frac{\eta L^2}{2}\|x - x^*\|_2^2 \right)$
    $\hat{\theta} \leftarrow \exp\left(-f(x) - g(x) + \frac{\eta}{2(1+\eta L)}\|\nabla f(x)\|_2^2 \right)(1 + \eta L)^{\frac{d}{2}}\alpha$
    if $\tau \le \frac{\hat{\theta}}{4}$ **then**
      return x
    end if
  end if
end while

2. *Assuming first-order (function and gradient evaluation) access to f, and restricted Gaussian oracle access to g, we can implement the same accesses to $\tilde{f}$, $\tilde{g}$ with constant overhead.*
3. *$\tilde{f}$ and $\tilde{g}$ are both minimized by x^*.*

Proof For f and g with properties as in (12), with x^* minimizing $f + g$, define the functions

$$\tilde{f}(x) := f(x) - \langle \nabla f(x^*), x \rangle, \ \tilde{g}(x) := g(x) + \langle \nabla f(x^*), x \rangle,$$

and observe that $\tilde{f} + \tilde{g} = f + g$ everywhere. This proves the first claim. Further, implementation of a first-order oracle for $\tilde{f}$ and a restricted Gaussian oracle for $\tilde{g}$ are immediate assuming a first-order oracle for f and a restricted Gaussian oracle for g, showing the second claim; any quadratic shifted by a linear term is the sum of a quadratic and a constant. We now show $\tilde{f}$ and $\tilde{g}$ have the same minimizer. By strong convexity, $\tilde{f}$ has a unique minimizer; first-order optimality shows that

$$\nabla \tilde{f}(x^*) = \nabla f(x^*) - \nabla f(x^*) = 0,$$

so this unique minimizer is x^*. Moreover, optimality of x^* for $f + g$ implies that for all $x \in \mathbb{R}^d$,

$$\langle \partial g(x^*) + \nabla f(x^*), x^* - x \rangle \le 0.$$

Algorithm 5 `Sample-Joint-Dist`$(f, g, x^*, \eta, \mathcal{O}, \delta)$

Input: f, g of form (12) both minimized by x^*, $\delta \in [0,1]$, $\eta > 0$, $\mathcal{O}$ restricted Gaussian oracle for g.

Output: Sample x from a distribution $\hat{\pi}'$ with $\|\hat{\pi}' - \hat{\pi}\|_{\mathrm{TV}} \le \delta$, where we overload $\hat{\pi}$ to mean the marginal of (14) on the x variable.

$\eta \leftarrow \frac{1}{32L\kappa d \log(16\kappa/\delta)}$

Let $\hat{\pi}$ be the density with

$$\frac{d\hat{\pi}}{dx}(z) \propto \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\|y - x\|_2^2 - \frac{\eta L^2}{2}\|x - x^*\|_2^2\right) \tag{14}$$

Call $\mathcal{O}$ to sample $x_0 \sim \pi_{\text{start}}$, for

$$\frac{d\pi_{\text{start}}(x)}{dx} \propto \exp\left(-\frac{L + \eta L^2}{2}\|x - x^*\|_2^2 - g(x)\right) \tag{15}$$

$K \leftarrow \frac{2^{26}\cdot 100}{\eta\mu}\log\left(\frac{d\log(16\kappa)}{4\delta}\right)$ (see Remark 25)

for $k \in [K]$ **do**

  Call `YSample`$\left(f, x_{k-1}, \eta, \frac{\delta}{2Kd\log(\frac{d\kappa}{\delta})}\right)$ to sample $y_k \sim \pi_{x_{k-1}}$ (Algorithm 7), for

$$\frac{d\pi_x}{dy}(y) \propto \exp\left(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2\right) \tag{16}$$

  Call $\mathcal{O}$ to sample $x_k \sim \pi_{y_k}$, for

$$\frac{d\pi_y}{dx}(x) \propto \exp\left(-g(x) - \frac{1}{2\eta}\|y - x\|_2^2 - \frac{\eta L^2}{2}\|x - x^*\|_2^2\right) \tag{17}$$

end for

return x_K

Here, ∂g is a subgradient. This shows first-order optimality of x^* for $\tilde{g}$ also, so x^* minimizes $\tilde{g}$. ■

E.2. Reduction from `Composite-Sample-Shared-Min` to `Sample-Joint-Dist`

`Composite-Sample-Shared-Min` is a rejection sampling scheme, which accepts samples from subroutine `Sample-Joint-Dist` in the high-probability region Ω defined in (13). We give a general analysis for approximate rejection sampling in Section H.1.1, and Section H.1.2 bounds relationships between distributions π and $\hat{\pi}$, defined in (12) and (14) respectively (i.e. relative densities and normalization constant ratios). Combining these pieces proves the following main claim.

Proposition 24 *Let* $\eta = \frac{1}{32L\kappa d \log(288\kappa/\epsilon)}$, *and assume* `Sample-Joint-Dist`$(f, g, x^*, \mathcal{O}, \delta)$ *samples within* δ *total variation of the* x*-marginal on* (14). `Composite-Sample-Shared-Min` *outputs a sample within total variation* ϵ *of* (12) *in an expected* $O(1)$ *calls to* `Sample-Joint-Dist`.

E.3. Implementing `Sample-Joint-Dist`

`Sample-Joint-Dist` alternates between sampling marginals in the joint distribution $\hat{\pi}$, as seen by definitions (16), (17). We showed that marginal sampling attains the correct stationary distribution as Lemma 11. We bound the conductance of the induced walk on iterates $\{x_k\}$ by combining an isoperimetry bound with a total variation guarantee between transitions of nearby points in Section H.2.1. Finally, we give a simple rejection sampling scheme `YSample` as Algorithm 7 for implementing the step (16). Since the y-marginal of $\hat{\pi}$ is a bounded perturbation of a Gaussian (intuitively, f is L-smooth and $\eta^{-1} \gg L$), we show in a high probability region that rejecting from the sum of a first-order approximation to f and the Gaussian succeeds in 2 iterations.

Remark 25 *For simplicity of presentation, we were conservative in bounding constants throughout; in practice, we found that the constant in Line 4 is orders of magnitude too large (a constant* < 10 *sufficed), which can be found as Section 4 of Shen et al. (2020). Several constants were inherited from prior analyses, which we do not rederive to save on redundancy.*

We now give a complete guarantee on the complexity of `Sample-Joint-Dist`.

Proposition 26 `Sample-Joint-Dist` *outputs a point with distribution within* δ *total variation distance from the* x*-marginal of* $\hat{\pi}$. *The expected number of gradient queries per iteration is constant.*

E.4. Putting it all together: proof of Theorem 5

We show Theorem 5 follows from the guarantees of Propositions 23, 24, and 26. Formally, Theorem 5 is stated for an arbitrary value of η which is upper bounded by the value in Line 1 of Algorithm 5; however, it is straightforward to see that all our proofs go through for any smaller value. By observing the value of K in `Sample-Joint-Dist`, we see that the number of total iterations in each call to `Sample-Joint-Dist` $O\left(\frac{1}{\eta\mu}\log(\frac{\kappa d}{\epsilon})\right) = O\left(\kappa^2 d \log^2\left(\frac{\kappa d}{\delta}\right)\right)$. Proposition 26 also shows that every iteration, we require an expected constant number of gradient queries and calls to $\mathcal{O}$, the restricted Gaussian oracle for g, and that the resulting distribution has δ total variation from the desired marginal of $\hat{\pi}$. Next, Proposition 24 implies that the number of calls to `Sample-Joint-Dist` in a run of `Composite-Sample-Shared-Min` is bounded by a constant, the choice of δ is $\Theta(\epsilon)$, and the resulting point has total variation ϵ from the original distribution π. Finally, Proposition 23 shows sampling from a general distribution of the form (1) is reducible to one call of `Composite-Sample-Shared-Min`, and the requisite oracles are implementable.

Appendix F. Logconcave finite sums

In this section, we provide our "base sampler" for logconcave finite sums as Algorithm 6, and give its guarantees by proving Theorem 8. Throughout, fix distribution π with density

$$\frac{d\pi}{dx}(x) \propto \exp(-F(x)), \text{ where } F(x) = \frac{1}{n}\sum_{i\in[n]} f_i(x) \text{ is } \mu\text{-strongly convex,}$$
$$\text{and for all } i \in [n],\ f_i \text{ is } L\text{-smooth.}$$

We will define $\kappa := \frac{L}{\mu}$, and assume that we have precomputed $x^* := \text{argmin}_{x\in\mathbb{R}^d}\{F(x)\}$.

Algorithm 6 `FiniteSum-MRW`(F, h, x_0, p, K)

Input: $F(x) = \frac{1}{n}\sum_{i\in[n]} f_i(x)$, step size $h > 0$, initial x_0, $p \in [0,1]$, iteration count $K \in \mathbb{N}$
for $0 \le k < K$ **do**
  Draw $\xi_k \sim \mathcal{N}(0, \mathbf{I})$
  $y_{k+1} \leftarrow x_k + \sqrt{2h}\xi_k$
  Draw $S_k \subseteq [n]$ by including each $i \in S_k$ independently with probability p
  For each $i \in [n]$,

$$\gamma_k^{(i)} \leftarrow \begin{cases} \frac{1}{p}\left(\sqrt{\exp\left(-\frac{1}{n}f_i(y_{k+1}) + \frac{1}{n}f_i(x_k)\right)} - 1\right) + 1 & i \in S_k \\ 1 & i \notin S_k \end{cases}$$

  $\gamma_k \leftarrow \prod_{i=1}^n \gamma_k^{(i)}$, $\tau \sim \text{Unif}[0,1]$
  if $\tau \le \frac{3}{4}\gamma_k$ and $|S_k| \le 2pn$ **then**
    $x_{k+1} \leftarrow y_{k+1}$
  else
    $x_{k+1} \leftarrow x_k$
  end if
end for
return x_K.

Algorithm 6 is the zeroth-order Metropolized random walk of Dwivedi et al. (2018) with an efficient, but biased, filter step; the goal of our analysis is to show this bias does not incur significant error.

F.1. Approximate Metropolis-Hastings

We first recall the following well-known fact underlying Metropolis-Hastings (MH) filters.

Proposition 27 *Consider a random walk on $\mathbb{R}^d$ with proposal distributions $\{\mathcal{P}_x\}_{x\in\mathbb{R}^d}$ and acceptance probabilities $\{\alpha(x,x')\}_{x,x'\in\mathbb{R}^d}$ conducted as follows: at a current point x,*

1. *Draw a point $x' \sim \mathcal{P}_x$.*
2. *Move the random walk to x' with probability $\alpha(x,x')$, else stay at x.*

Suppose $\mathcal{P}_x(x') = \mathcal{P}_{x'}(x)$ for all pairs $x, x' \in \mathbb{R}^d$, and further $\frac{d\pi}{dx}(x)\alpha(x,x') = \frac{d\pi}{dx}(x')\alpha(x',x)$. Then, π is a stationary distribution for the random walk.

Proof This follows because the walk satisfies detailed balance (reversibility) with respect to π. ■

We propose an algorithm that applies a variant of the Metropolis-Hastings filter to a Gaussian random walk. Specifically, we define the following algorithm, which we call `Inefficient-MRW`.

Definition 28 (`Inefficient-MRW`) *Consider the following random walk for some step size $h > 0$: for each iteration k at a current point $x_k \in \mathbb{R}^d$,*

1. *Set $y_{k+1} \leftarrow x_k + \sqrt{2h}\xi$, where $\xi \sim \mathcal{N}(0, \mathbf{I})$.*
2. *$x_{k+1} \leftarrow y_{k+1}$ with probability $\alpha(x_k, y_{k+1})$ (otherwise, $x_{k+1} \leftarrow x_k$), where*

$$\alpha(x,y) = \begin{cases} 1 & \sqrt{\frac{\exp(-F(y))}{\exp(-F(x))}} > \frac{4}{3}, \\ \frac{3}{4}\sqrt{\frac{\exp(-F(y))}{\exp(-F(x))}} & \frac{3}{4} \le \sqrt{\frac{\exp(-F(y))}{\exp(-F(x))}} \le \frac{4}{3}, \\ \frac{\exp(-F(y))}{\exp(-F(x))} & \sqrt{\frac{\exp(-F(y))}{\exp(-F(x))}} < \frac{3}{4}. \end{cases} \tag{18}$$

Lemma 29 *Distribution π with $\frac{d\pi}{dx}(x) \propto \exp(-F(x))$ is stationary for `Inefficient-MRW`.*

Proof Without loss of generality, assume that π has been normalized so that $\frac{d\pi}{dx}(x) = \exp(-F(x))$. We apply Proposition 27, dropping subscripts in the following. It is clear that $\mathcal{P}_x(y) = \mathcal{P}_y(x)$ for any x, y, so it suffices to check the second condition. When $\frac{3}{4} \le \sqrt{\frac{\exp(-F(y))}{\exp(-F(x))}} \le \frac{4}{3}$, this follows from

$$\frac{d\pi}{dx}(x)\alpha(x,x') = \frac{3}{4}\sqrt{\exp(-F(x) - F(y))} = \frac{d\pi}{dx}(x')\alpha(x',x).$$

The other case is similar (as it is a standard Metropolis-Hastings filter). ■

In Algorithm 6, we implement an approximate version of the modified MH filter in Definition 28, where we always assume the pair x, y are in the second case of (18). In Lemma 30, we show that if a certain boundedness condition holds, then Algorithm 6 approximates `Inefficient-MRW` well. We then show that the output distributions of `Inefficient-MRW` and our Algorithm 6 have small total variation distance in Lemma 31.

Lemma 30 *Suppose that in an iteration $0 \le k < K$ of Algorithm 6, the following three conditions hold for some parameters R_x, C_ξ, $C_x \in \mathbb{R}_{\ge 0}$:*

1. *$\|x_k - x^*\|_2 \le R_x$.*
2. *$\|\xi_k\|_2 \le C_\xi\sqrt{d}$.*
3. *For all $i \in [n]$, $|\nabla f_i(x_k)^\top \xi_k| \le C_x \|\nabla f_i(x_k)\|_2$.*

Then, for any

$$h \le \frac{1}{98C_x^2L^2R_x^2 + 7LC_\xi^2 d}, \tag{19}$$

$\frac{3}{4} \le \sqrt{\frac{\exp(-F(y_{k+1}))}{\exp(-F(x_k))}} \le \frac{4}{3}$. Moreover, we have $\mathbb{E}[\gamma_k] = \sqrt{\frac{\exp(-F(y_{k+1}))}{\exp(-F(x_k))}}$, and when $|S_k| \le 2pn$, $\gamma_k \le \frac{4}{3}$.

Proof We first show $\mathbb{E}[\gamma_k] = \sqrt{\frac{\exp(-F(y_{k+1}))}{\exp(-F(x_k))}}$. Since each $i \in S_k$ is generated independently,

$$\begin{aligned}\mathbb{E}[\gamma_k] &= \prod_{i\in[n]} \mathbb{E}\left[\gamma_k^{(i)}\right] \\ &= \prod_{i\in[n]} \left[(1-p) + p\left(\frac{1}{p}\left(\sqrt{\exp\left(-\frac{1}{n}f_i(y_{k+1}) + \frac{1}{n}f_i(x_k)\right)} - 1\right) + 1\right)\right] \\ &= \prod_{i\in[n]} \sqrt{\exp\left(-\frac{1}{n}f_i(y_{k+1}) + \frac{1}{n}f_i(x_k)\right)} = \sqrt{\frac{\exp(-F(y_{k+1}))}{\exp(-F(x_k))}}.\end{aligned}$$

Next, for any $i \in [n]$, we lower and upper bound $-f_i(y_{k+1}) + f_i(x_k)$. First,

$$\begin{aligned}-f_i(y_{k+1}) + f_i(x_k) &\le \nabla f_i(x_k)^\top (x_k - y_{k+1}) \\ &\le \sqrt{2h} C_x \|\nabla f_i(x_k)\|_2 \le \sqrt{2h} C_x L R_x.\end{aligned}$$

The first inequality followed from convexity of f_i, the second from $y_{k+1} - x_k = \sqrt{2h}\xi_k$ and our assumed bound, and the third from smoothness and $\nabla f(x^*) = 0$. To show a lower bound,

$$\begin{aligned}f_i(y_{k+1}) - f_i(x_k) &\le \nabla f_i(x_k)^\top (y_{k+1} - x_k) + \frac{L}{2}\|y_{k+1} - x_k\|_2^2 \\ &\le \sqrt{2h} C_x L R_x + h L C_\xi^2 d.\end{aligned}$$

The first inequality was smoothness. Repeating this argument for each $i \in [n]$ and averaging,

$$-\sqrt{2h} C_x L R_x - h L C_\xi^2 d \le -F(y_{k+1}) + F(x_k) \le \sqrt{2h} C_x L R_x. \tag{20}$$

Then, when $h \le \frac{1}{98 C_x^2 L^2 R_x^2 + 7 L C_\xi^2 d}$,

$$\frac{3}{4} \le \sqrt{\frac{\exp(-F(y_{k+1}))}{\exp(-F(x_k))}} \le \frac{4}{3}, \text{ and for all } i \in [n],\ -f_i(y_{k+1}) + f_i(x_k) \le \frac{1}{4}.$$

Thus, we can bound each $\gamma_k^{(i)}$:

$$\gamma_k^{(i)} \le \frac{1}{p}\left(\exp\left(\frac{1}{8n}\right) - 1\right) + 1 \le 1 + \frac{1}{7pn}.$$

Finally, when $|S_k| \le 2pn$, $\gamma_k \le (1 + \frac{1}{7pn})^{2pn} \le \frac{4}{3}$ as desired. ■

Lemma 31 *Draw $x_0 \sim \mathcal{N}(x^*, \frac{1}{L}\mathbf{I})$. Let $\hat{\pi}_K$ be the output distribution of the algorithm of Definition 28 for K steps starting from x_0, and let π_K be the output distribution of Algorithm 6 starting from x_0. For any $\delta \in [0,1]$, let $p = \frac{5\log\frac{12K}{\delta}}{n}$ in Algorithm 6. There exist*

$$C_\xi = O\left(1 + \sqrt{\frac{\log\frac{K}{\delta}}{d}}\right),\ C_x = O\left(\sqrt{\log\frac{nK}{\delta}}\right),\ \text{and } R_x = O\left(\sqrt{\frac{d\log\frac{\kappa K}{\delta}}{\mu}}\right),$$

so that when $h \le \frac{1}{98 C_x^2 L^2 R_x^2 + 7 L C_\xi^2 d}$, we have $\|\pi_K - \hat{\pi}_K\|_{\mathrm{TV}} \le \delta$.

Proof By the coupling definition of total variation, it suffices to upper bound the probability that the algorithms' trajectories, sharing all randomness in proposing points y_{k+1}, differ. This can happen for two reasons: either we used an incorrect filtering step (i.e. the pair (x_k, y_{k+1}) did not lie in the second case of (18)), or we incorrectly rejected in Line 7 of Algorithm 6 because $|S_k| \geq 2pn$. We bound the error due to either happening over any iteration by δ, yielding the conclusion.

Incorrect filtering. Consider some iteration k. Lemma 30 shows that as long as its three conditions hold in iteration k, we are in the second case of (18), so it suffices to show all conditions hold. By Fact 2 and as ξ_k is independent of all $\{\nabla f_i(x_k)\}_{i\in[n]}$, with probability at least $1 - \frac{\delta}{2K}$, both of the conditions $\|\xi_k\|_2 \leq C_\xi\sqrt{d}$ and[13] $|\nabla f_i(x_k)^\top \xi_k| \leq C_x \|\nabla f_i(x_k)\|_2$ for all $i \in [n]$ hold for some

$$C_\xi = O\left(1 + \sqrt{\frac{\log \frac{K}{\delta}}{d}}\right), \; C_x = O\left(\sqrt{\log \frac{nK}{\delta}}\right).$$

Next, by Lemma 18, x_0 is drawn from a $\kappa^{\frac{d}{2}}$ warm start for π. By Fact 2, we have $\|x_0 - x^*\|_2 \leq R_x$ for x_0 drawn from π with probability at least $1 - \frac{\delta}{4K} \cdot \kappa^{-\frac{d}{2}}$, for some

$$R_x = O\left(\sqrt{\frac{d \log \frac{\kappa K}{\delta}}{\mu}}\right).$$

Since warmness of the exact algorithm of Definition 28 is monotonic, as long as the trajectories have not differed up to iteration k, $\|x_k - x^*\|_2 \leq R_x$ also holds with probability $\geq 1 - \frac{\delta}{4K}$. Inductively, the total variation error caused by incorrect filtering over K steps is at most $\frac{3\delta}{4}$.

Error due to large $|S_k|$. Supposing all the conditions of Lemma 30 are satisfied in iteration k, we show that with high probability, `Inefficient-MRW` and Algorithm 6 make the same accept or reject decision. By Lemma 30, `Inefficient-MRW` (18) accepts with probability $\alpha'_k = \frac{3}{4}\sqrt{\frac{\exp(-F(y_{k+1}))}{\exp(-F(x_k))}}$. On the other hand, Algorithm 6 accepts with probability

$$\alpha_k = \frac{3}{4}\mathbb{E}\left[\gamma_k \mid |S_k| \leq 2pn\right] \cdot \Pr[|S_k| \leq 2pn].$$

The total variation between the output distributions is $|\alpha_k - \alpha'_k|$. Further, since by Lemma 30,

$$\begin{aligned}
\alpha'_k &= \frac{3}{4}\mathbb{E}\left[\gamma_k\right] \\
&= \frac{3}{4}\left(\mathbb{E}\left[\gamma_k \mid |S_k| \leq 2pn\right] \cdot \Pr[|S_k| \leq 2pn] + \mathbb{E}\left[\gamma_k \mid |S_k| > 2pn\right] \cdot \Pr[|S_k| > 2pn]\right) \\
&= \alpha_k + \frac{3}{4}\mathbb{E}\left[\gamma_k \mid |S_k| > 2pn\right] \cdot \Pr[|S_k| > 2pn],
\end{aligned}$$

13. We recall that the distribution of $v^\top \xi$ for $\xi \sim \mathcal{N}(0, \mathbf{I})$ is the one-dimensional $\mathcal{N}(0, \|v\|_2^2)$.

it suffices to upper bound this latter quantity. First, by Lemma 32, when $p = \frac{5 \log \frac{12K}{\delta}}{n}$, we have $\Pr[|S_k| > 2pn] \le \frac{\delta}{12K}$. Finally, since each $i \in S_k$ is generated independently,

$$\mathbb{E}\left[\gamma_k \mid |S_k| > 2pn\right] \le \max_{S' : |S'| = 2pn} \mathbb{E}\left[\prod_{i \in [n]} \gamma_k^{(i)} \mid S' \subseteq S_k\right]$$
$$\le 2\mathbb{E}\left[\prod_{i \in [n] \setminus S'} \gamma_k^{(i)}\right] = 2\sqrt{\prod_{[n] \setminus S'} \exp\left(-\frac{1}{n} f_i(y_{k+1}) + \frac{1}{n} f_i(x_k)\right)} \le 4.$$

Here, we used Lemma 30 applied to the set S', and the upper bound (20) we derived earlier. Combining these calculations shows that the total variation distance incurred in any iteration k due to $|S_k|$ being too large is at most $\frac{\delta}{4K}$, so the overall contribution over K steps is at most $\frac{\delta}{4}$. ■

We used the following helper lemma in our analysis.

Lemma 32 *Let $S \subseteq [n]$ be formed by independently including each $i \in [n]$ with probability p. Then,*

$$\Pr\left[|S| > 2pn\right] \le \exp\left(-\frac{3pn}{14}\right).$$

Proof For $i \in [n]$, let $\mathbf{1}_{i \in S}$ be the indicator random variable of the event $i \in S$, so $\mathbb{E}\left[\mathbf{1}_{i \in S}\right] = p$ and

$$\mathrm{Var}\left[\mathbf{1}_{i \in S} - p\right] = p(1-p)^2 + (1-p)p^2 \le 2p.$$

By Bernstein's inequality,

$$\Pr\left[\sum_{i \in [n]} \mathbf{1}_{i \in S} \ge np + r\right] \le \exp\left(-\frac{\frac{1}{2}r^2}{2np + \frac{1}{3}r}\right).$$

In particular, when $r = pn$, we have the desired conclusion. ■

F.2. Conductance analysis

We next bound the mixing time of `Inefficient-MRW` via average conductance (Proposition 12). Consider an iteration of `Inefficient-MRW` from $x_k = x$. Let $\mathcal{P}_x$ be the density of y_{k+1}, and let $\mathcal{T}_x$ be the density of x_{k+1} after filtering. Define a convex set $\Omega \subseteq \mathbb{R}^d$ parameterized by $R_\Omega \in \mathbb{R}_{\ge 0}$:

$$\Omega = \{x \in \mathbb{R}^d : \|x - x^*\|_2 \le R_\Omega\}.$$

We show that for two close points $x, x' \subseteq \Omega$, the total variation between $\mathcal{T}_x$ and $\mathcal{T}_{x'}$ is small.

Lemma 33 *For some $h = O(\frac{1}{L^2 R_\Omega^2 + Ld})$ and $x, x' \subseteq \Omega$ with $\|x - x'\|_2 \le \frac{1}{8}\sqrt{h}$, $\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\mathrm{TV}} \le \frac{7}{8}$.*

Proof By the triangle inequality of total variation distance,

$$\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\mathrm{TV}} \le \|\mathcal{T}_x - \mathcal{P}_x\|_{\mathrm{TV}} + \|\mathcal{P}_x - \mathcal{P}_{x'}\|_{\mathrm{TV}} + \|\mathcal{T}_{x'} - \mathcal{P}_{x'}\|_{\mathrm{TV}}\,.$$

First, by Pinsker's inequality and the KL divergence between Gaussian distributions,

$$\|\mathcal{P}_x - \mathcal{P}_{x'}\|_{\mathrm{TV}} \le \sqrt{2\mathrm{KL}(\mathcal{P}_x||\mathcal{P}_{x'})} = \frac{\|x - x'\|_2}{\sqrt{2h}}.$$

When $\|x - x'\|_2 \le \frac{1}{8}\sqrt{h}$, $\|\mathcal{P}_x - \mathcal{P}_{x'}\|_{\mathrm{TV}} \le \frac{1}{8}$. Next, we bound $\|\mathcal{T}_x - \mathcal{P}_x\|_{\mathrm{TV}}$: by a standard calculation (e.g. Lemma D.1 of Lee et al. (2020)), we have

$$\|\mathcal{T}_x - \mathcal{P}_x\|_{\mathrm{TV}} = 1 - \frac{3}{4}\mathbb{E}_{\xi_{k+1}}\left[\sqrt{\frac{\exp\left(-F(y_{k+1})\right)}{\exp\left(-F(x_k)\right)}}\right].$$

We show that $\|\mathcal{T}_x - \mathcal{P}_x\|_{\mathrm{TV}} \le \frac{3}{8}$. It suffices to show that $\mathbb{E}_{\xi_{k+1}}\left[\sqrt{\exp\left(-F(y_{k+1}) + F(x_k)\right)}\right] \ge \frac{5}{6}$.

Since $\frac{15}{16}\sqrt{\exp\left(-\frac{1}{16}\right)} \ge \frac{5}{6}$, it suffices to show that with probability at least $\frac{15}{16}$ over the randomness of ξ_{k+1}, $-F(y_{k+1}) + F(x_k) \ge -\frac{1}{16}$. As $\xi_{k+1} \sim \mathcal{N}(0, \mathbb{I}_d)$, by applying Fact 2 twice,

$$\begin{aligned} \Pr\left[\|\xi_{k+1}\|_2^2 > 36d\right] &\le \exp(-4) \le \frac{1}{32}, \\ \Pr\left[\left|\nabla F(x_k)^\top \xi_{k+1}\right|^2 \ge 36\,\|\nabla F(x_k)\|_2^2\right] &\le \frac{1}{32}. \end{aligned} \tag{21}$$

We upper bound the term $F(y_{k+1}) - F(x_k)$ by smoothness and Cauchy-Schwarz:

$$\begin{aligned} F(y_{k+1}) - F(x_k) &\le \nabla F(x_k)^\top (y_{k+1} - x_k) + \frac{L}{2}\|y_{k+1} - x_k\|_2^2 \\ &\le \sqrt{2h}\left|\nabla F(x_k)^\top \xi_{k+1}\right| + hL\|\xi_{k+1}\|_2^2. \end{aligned}$$

Then, since $\|\nabla F(x_k)\| \le LR_\Omega$ when $x \in \Omega$, it is enough to choose $h = O(\frac{1}{L^2R_\Omega^2 + Ld})$ so that

$$-F(y_{k+1}) + F(x_k) \ge -\frac{1}{16},$$

as long as the events of (21) hold, which occurs with probability at least $\frac{15}{16}$. Similarly, we can show that $\|\mathcal{T}_{x'} - \mathcal{P}_{x'}\|_{\mathrm{TV}} \le \frac{3}{8}$. Combining the three bounds, we have the desired conclusion. ■

Theorem 8 *Let π be a distribution on $\mathbb{R}^d$ with $\frac{d\pi}{dx}(x) \propto \exp(-F(x))$, where $F(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ is μ-strongly convex, f_i is L-smooth and convex $\forall i \in [n]$, $\kappa = \frac{L}{\mu}$, and $\epsilon \in (0,1)$. Assume access to $x^* = \mathrm{argmin}_{x\in\mathbb{R}^d} F(x)$. Algorithm 6 uses $O\left(\kappa^2 d \log^4 \frac{n\kappa d}{\epsilon}\right)$ value queries to summands $\{f_i\}_{i\in[n]}$, and obtains ϵ total variation distance to π.*

Proof First, by Lemma 18, $\mathcal{N}(x^*, \frac{1}{L}\mathbf{I})$ yields a $\beta = \kappa^{\frac{d}{2}}$-warm start for π. For this value of β, by Fact 2 it suffices to choose

$$R_\Omega = \Theta\left(\sqrt{\frac{d \log \frac{\kappa}{\epsilon}}{\mu}}\right)$$

for $\pi(\Omega) \geq 1 - \frac{\epsilon^2}{2\beta^2}$. Letting $\delta = \frac{\epsilon}{2}$, we will choose the step size h and iteration count K so that

$$\frac{1}{h} = \Theta\left(L\kappa d \log^2 \frac{n\kappa d}{\epsilon}\right), \; K = \Theta\left(\kappa^2 d \log^3 \frac{n\kappa d}{\epsilon}\right)$$

have constants compatible with Lemma 31. Note that this choice of h is also sufficiently small to apply Lemma 33 for our choice of R_Ω. By applying Proposition 12 to the algorithm of Definition 28, and using the bound from Lemma 33, in K iterations `Inefficient-MRW` will mix to total variation distance δ to π. Furthermore, applying Lemma 31, we conclude that Algorithm 6 has total variation distance at most $2\delta = \epsilon$ from π.

It remains to bound the oracle complexity of Algorithm 6. Note in every iteration, we never compute more than $4pn$ values of $\{f_i\}_{i \in [n]}$, since we always reject if $|S_k| \geq 2pn$, and we only compute values for indices in S_k. For the value of p in Lemma 31, this amounts to $O(\log \frac{n\kappa d}{\epsilon})$ value queries. ■

Appendix G. Discussion of inexactness tolerance

We briefly discuss the tolerance of our algorithm to approximation error in two places: computation of minimizers, and implementation of RGOs in the methods of Sections C and E.

Inexact minimization. For all function classes considered in this work, there exist efficient optimization methods converging to a minimizer with logarithmic dependence on the target accuracy.

Specifically, for negative log-densities with condition number κ, accelerated gradient descent Nesterov (1983) converges at a rate $O(\sqrt{\kappa})$ with logarithmic dependence on initial error and target accuracy (we implicitly assumed in stating our runtimes that one can attain initial error polynomial in problem parameters for negative log-densities; otherwise, there is additional logarithmic overhead in the quality of the initial point to optimization procedures). For composite functions $f_{\text{wc}} + f_{\text{oracle}}$ where f_{wc} has condition number κ, the FISTA method of Beck and Teboulle (2009) converges at the same rate with each iteration querying ∇f_{wc} and a proximal oracle for f_{oracle} once; typically, access to a proximal oracle is a weaker assumption than access to a restricted Gaussian oracle, so this is not restrictive. Finally, for minimizing finite sums with condition number κ, the algorithm of Allen-Zhu (2017) obtains a convergence rate linearly dependent on $n + \sqrt{n\kappa} \leq n + \kappa$; alternatively, Johnson and Zhang (2013) has a dependence on $n + \kappa$. In all our final runtimes, these optimization rates do not constitute the bottleneck for oracle complexities.

The only additional difficulty our algorithms may present is if the function requiring minimization, say of the form $f_{\text{oracle}}(x) + \frac{1}{2\eta}\|x - y\|_2^2$ for some $y \in \mathbb{R}^d$ where we have computed the minimizer x^* to f_{oracle}, has $\|y - x^*\|_2^2$ very large (so the initial function error is bad). However, in all our settings y is drawn from a distribution with sub-Gaussian tails, so $\|y - x^*\|_2^2$ decays exponentially (whereas the complexity of first-order methods increases only logarithmically), negligibly affecting the expected oracle query complexity for our methods.

Finally, by solving the relevant optimization problems to high accuracy as a subroutine in each of our methods, and adjusting various distance bounds to the minimizer by constants (e.g. by expanding the radius in the definition of the sets Ω in Algorithm 4 and Section F.2), this accomodates tolerance to inexact minimization and only affects all bounds throughout the paper by constants. The only other place that x^* is used in our algorithms is in initializing warm starts; tolerance to inexactness in our warmness calculations follows essentially identically to Section 3.2.1 of Dwivedi et al. (2018).

Inexact oracle implementation. Our algorithms based on restricted Gaussian oracle access are tolerant to total variation error inverse polynomial in problem parameters for the restricted Gaussian oracle for g. We discussed this at the end of Section C, in the case of RGO use for our reduction framework. To see this in the case of the composite sampler in Section E, we pessimistically handled the case where the sampler `YSample` for a quadratic restriction of f resulted in total variation error in the proof of Proposition 26, assuming that the error was incurred in every iteration. By accounting for similar amounts of error in calls to $\mathcal{O}$ (on the order of $\frac{\epsilon}{T}$, where T is the number of times an RGO was used), the bounds in our algorithm are only affected by constants.

Appendix H. Deferred proofs from Section E

H.1. Deferred proofs from Section E.2

H.1.1. APPROXIMATE REJECTION SAMPLING

We first define the rejection sampling framework we will use, and prove various properties.

Definition 34 (Approximate rejection sampling) *Let π be a distribution, with $\frac{d\pi}{dx}(x) \propto p(x)$. Suppose set Ω has $\pi(\Omega) = 1 - \epsilon'$, and distribution $\hat{\pi}$ with $\frac{d\hat{\pi}}{dx}(x) \propto \hat{p}(x)$ has for some $C \geq 1$,*

$$\frac{p(x)}{\hat{p}(x)} \leq C \text{ for all } x \in \Omega, \text{ and } \frac{\int \hat{p}(x)dx}{\int p(x)dx} \leq 1.$$

Suppose there is an algorithm $\mathcal{A}$ which draws samples from a distribution $\hat{\pi}'$, such that $\|\hat{\pi}' - \hat{\pi}\|_{\text{TV}} \leq 1 - \delta$. We call the following scheme approximate rejection sampling: repeat independent runs of the following procedure until a point is outputted.

1. *Draw x via $\mathcal{A}$ until $x \in \Omega$.*
2. *With probability $\frac{p(x)}{C\hat{p}(x)}$, output x.*

Lemma 35 *Consider an approximate rejection sampling scheme with relevant parameters defined as in Definition 34, with $2\delta \leq \frac{1-\epsilon'}{C}$. The algorithm terminates in at most*

$$\frac{1}{\frac{1-\epsilon'}{C} - 2\delta} \tag{22}$$

calls to $\mathcal{A}$ in expectation, and outputs a point from a distribution π' with $\|\pi' - \pi\|_{\text{TV}} \leq \epsilon' + \frac{2\delta C}{1-\epsilon'}$.

Proof Define for notational simplicity normalization constants $Z := \int p(x)dx$ and $\hat{Z} := \int \hat{p}(x)dx$. First, we bound the probability any particular call to $\mathcal{A}$ returns in the scheme:

$$\begin{aligned}\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x) &\geq \int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}(x) - \left|\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} (d\hat{\pi}'(x) - d\hat{\pi}(x))\right| \\ &= \int_{x\in\Omega} \frac{Z}{C\hat{Z}} d\pi(x) - \left|\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} (d\hat{\pi}'(x) - d\hat{\pi}(x))\right| \\ &\geq \frac{1-\epsilon'}{C} - \int_{x\in\Omega} |d\hat{\pi}'(x) - d\hat{\pi}(x)| \geq \frac{1-\epsilon'}{C} - 2\delta.\end{aligned} \tag{23}$$

The second line followed by the definitions of Z and $\hat{Z}$, and the third followed by triangle inequality, the assumed lower bound on $Z/\hat{Z}$, and the total variation distance between $\hat{\pi}'$ and $\hat{\pi}$. By linearity of expectation and independence, this proves the first claim.

Next, we claim the output distribution is close in total variation distance to the conditional distribution of π restricted to Ω. The derivation of (23) implies

$$\begin{aligned}&\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}(x) \geq \frac{1-\epsilon'}{C}, \ \left|\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} (d\hat{\pi}'(x) - d\hat{\pi}(x))\right| \leq 2\delta, \\ &\implies 1 - \frac{2\delta C}{1-\epsilon'} \leq \frac{\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x)}{\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}(x)} \leq 1 + \frac{2\delta C}{1-\epsilon'}.\end{aligned} \tag{24}$$

Thus, the total variation of the true output distribution from π restricted to Ω is

$$\begin{aligned}&\frac{1}{2}\int_{x\in\Omega} \left|\frac{d\pi(x)}{1-\epsilon'} - \frac{\frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x)}{\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x)}\right| \\ \leq &\frac{1}{2}\int_{x\in\Omega} \left|\frac{d\pi(x)}{1-\epsilon'} - \frac{\frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x)}{\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}(x)}\right| + \frac{1}{2}\int_{x\in\Omega} \left|\frac{\frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x)}{\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}(x)} - \frac{\frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x)}{\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x)}\right| \\ \leq &\frac{1}{2}\int_{x\in\Omega} \left|\frac{d\pi(x)}{1-\epsilon'} - \frac{\frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x)}{\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}(x)}\right| + \frac{\delta C}{1-\epsilon'} = \frac{1}{2}\int_{x\in\Omega} \frac{d\pi(x)}{1-\epsilon'} \left|1 - \frac{d\hat{\pi}'}{d\hat{\pi}}(x)\right| + \frac{\delta C}{1-\epsilon'}.\end{aligned}$$

The first inequality was triangle inequality, and we bounded the second term by (24). To obtain the final equality, we used

$$\begin{aligned}&\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}(x) = \int_{x\in\Omega} \frac{Z}{C\hat{Z}} d\pi(x) = \frac{(1-\epsilon')Z}{C\hat{Z}} \\ \implies &\frac{\frac{p(x)}{C\hat{p}(x)} d\hat{\pi}'(x)}{\int_{x\in\Omega} \frac{p(x)}{C\hat{p}(x)} d\hat{\pi}(x)} = \frac{p(x)}{Z} \cdot \frac{\hat{Z}}{\hat{p}(x)} \cdot \frac{1}{1-\epsilon'} \cdot d\hat{\pi}'(x) = \frac{d\pi(x)}{1-\epsilon'} \cdot \frac{d\hat{\pi}'}{d\hat{\pi}}(x).\end{aligned}$$

We now bound this final term. Observe that the given conditions imply that $\frac{d\pi}{d\hat{\pi}}(x)$ is bounded by C everywhere in Ω. Thus, expanding we have

$$\frac{1}{2}\int_{x\in\Omega} \frac{d\pi(x)}{1-\epsilon'} \left|1 - \frac{d\hat{\pi}'}{d\hat{\pi}}(x)\right| \leq \frac{C}{2(1-\epsilon')} \int_{x\in\Omega} |d\hat{\pi}(x) - d\hat{\pi}'(x)| \leq \frac{\delta C}{1-\epsilon'}.$$

Finally, combining these guarantees, and the fact that restricting π to Ω loses ϵ' in total variation distance, yields the desired conclusion by triangle inequality. ■

Corollary 36 *Let $\hat{\theta}(x)$ be an unbiased estimator for $\frac{p(x)}{\hat{p}(x)}$, and suppose $\hat{\theta}(x) \le C$ with probability 1 for all $x \in \Omega$. Then, implementing the procedure of Definition 34 with acceptance probability $\frac{\hat{\theta}(x)}{C}$ has the same runtime bound and total variation guarantee as given by Lemma 35.*

Proof It suffices to take expectations over the randomness of $\hat{\theta}$ everywhere in the proof of Lemma 35. ■

H.1.2. DISTRIBUTION RATIO BOUNDS

We next show two bounds relating the densities of distributions π and $\hat{\pi}$. We first define the normalization constants of (12), (14) for shorthand, and then tightly bound their ratio.

Definition 37 (Normalization constants) *We denote normalization constants of π and $\hat{\pi}$ by*

$$Z_\pi := \int_x \exp\left(-f(x) - g(x)\right) dx,$$
$$Z_{\hat{\pi}} := \int_{x,y} \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\left\|y - x\right\|_2^2 - \frac{\eta L^2}{2}\left\|x - x^*\right\|_2^2\right) dxdy.$$

Lemma 38 (Normalization constant bounds) *Let Z_π and $Z_{\hat{\pi}}$ be as in Definition 37. Then,*

$$\left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}} \left(1 + \frac{\eta L^2}{\mu}\right)^{-\frac{d}{2}} \le \frac{Z_{\hat{\pi}}}{Z_\pi} \le (2\pi\eta)^{\frac{d}{2}}.$$

Proof For each x, by convexity we have

$$\begin{aligned}
&\int_y \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\left\|y - x\right\|_2^2 - \frac{\eta L^2}{2}\left\|x - x^*\right\|_2^2\right) dy \\
&\le \exp\left(-g(x) - \frac{\eta L^2}{2}\left\|x - x^*\right\|_2^2\right) \int_y \exp\left(-f(x) - \langle \nabla f(x), y - x\rangle - \frac{1}{2\eta}\left\|y - x\right\|_2^2\right) dy \\
&= \exp\left(-f(x) - g(x) - \frac{\eta L^2}{2}\left\|x - x^*\right\|_2^2\right) \int_y \exp\left(\frac{\eta}{2}\left\|\nabla f(x)\right\|_2^2 - \frac{1}{2\eta}\left\|y - x + \eta\nabla f(x)\right\|_2^2\right) dy \\
&= (2\pi\eta)^{\frac{d}{2}} \exp\left(-f(x) - g(x)\right) \exp\left(\frac{\eta}{2}\left\|\nabla f(x)\right\|_2^2 - \frac{\eta L^2}{2}\left\|x - x^*\right\|_2^2\right) \\
&\le (2\pi\eta)^{\frac{d}{2}} \exp\left(-f(x) - g(x)\right).
\end{aligned} \tag{25}$$

Integrating both sides over x yields the upper bound on $\frac{Z_{\hat{\pi}}}{Z_\pi}$. Next, for the lower bound we have a similar derivation. For each x, by smoothness

$$\begin{aligned}
&\int_y \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\|y-x\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2\right) dy \\
&\geq \exp\left(-f(x) - g(x) - \frac{\eta L^2}{2}\|x-x^*\|_2^2\right) \int_y \exp\left(\langle \nabla f(x), x-y\rangle - \frac{1+\eta L}{2\eta}\|y-x\|_2^2\right) dy \\
&= \exp\left(-f(x) - g(x) - \frac{\eta L^2}{2}\|x-x^*\|^2 + \frac{\eta}{2(1+\eta L)}\|\nabla f(x)\|^2\right)\left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}} \\
&\geq \exp\left(-f(x) - g(x) - \frac{\eta L^2}{2}\|x-x^*\|_2^2\right)\left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}}.
\end{aligned}$$

Integrating both sides over x yields

$$\frac{Z_{\hat{\pi}}}{Z_\pi} \geq \left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}} \frac{\int_x \exp\left(-f(x) - g(x) - \frac{\eta L^2}{2}\|x-x^*\|_2^2\right) dx}{\int_x \exp\left(-f(x) - g(x)\right) dx} \geq \left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}} \left(1 + \frac{\eta L^2}{\mu}\right)^{-\frac{d}{2}}.$$

The last inequality followed from Proposition 61, where we used $f+g$ is μ-strongly convex. ■

Lemma 39 (Relative density bounds) *Let $\eta = \frac{1}{32L\kappa d\log(288\kappa/\epsilon)}$. For all $x \in \Omega$, as defined in (13), $\frac{d\pi}{d\hat{\pi}}(x) \leq 2$. Here, $\frac{d\hat{\pi}}{dx}(x)$ denotes the marginal density of $\hat{\pi}$. Moreover, for all $x \in \mathbb{R}^d$, $\frac{d\pi}{d\hat{\pi}}(x) \geq \frac{1}{2}$.*

Proof We first show the upper bound. By Lemma 38,

$$\begin{aligned}
\frac{d\pi}{d\hat{\pi}}(x) &= \frac{\exp\left(-f(x) - g(x)\right)}{\int_y \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\|y-x\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2\right) dy} \cdot \frac{Z_{\hat{\pi}}}{Z_\pi} \\
&\leq \frac{\exp\left(-f(x) - g(x)\right)}{\int_y \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\|y-x\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2\right) dy} \cdot (2\pi\eta)^{\frac{d}{2}}.
\end{aligned} \tag{26}$$

We now bound the first term, for $x \in \Omega$. By smoothness, we have

$$\frac{\exp\left(-f(y) - g(x)\right)}{\exp\left(-f(x) - g(x)\right)} \geq \exp\left(\langle \nabla f(x), x-y\rangle - \frac{L}{2}\|y-x\|_2^2\right),$$

so applying this for each y,

$$\begin{aligned}
&\frac{\int_y \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\|y-x\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2\right) dy}{\exp\left(-f(x) - g(x)\right)} \\
&\geq \exp\left(-\frac{\eta L^2}{2}\|x-x^*\|_2^2\right) \int_y \exp\left(\langle \nabla f(x), x-y\rangle - \frac{1+\eta L}{2\eta}\|y-x\|_2^2\right) dy \\
&= \exp\left(-\frac{\eta L^2}{2}\|x-x^*\|_2^2 + \frac{\eta}{2(1+\eta L)}\|\nabla f(x)\|_2^2\right) \int_y \exp\left(-\frac{1+\eta L}{2\eta}\left\|x - y - \frac{\eta}{1+\eta L}\nabla f(x)\right\|_2^2\right) dy \\
&\geq \exp\left(-\frac{\eta L^2}{2} \cdot \frac{16d\log(288\kappa/\epsilon)}{\mu}\right)\left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}} \geq \frac{3}{4}\left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}}.
\end{aligned}$$

In the last line, we used that $x \in \Omega$ implies $\|x - x^*\|_2^2 \leq \frac{16d\log(288\kappa/\epsilon)}{\mu}$, and the definition of η. Combining this bound with (26), we have the desired

$$\frac{d\pi}{d\hat{\pi}}(x) \leq \frac{4}{3}(1+\eta L)^{\frac{d}{2}} \leq 2.$$

Next, we consider the lower bound. By combining (25) with Lemma 38, we have the desired

$$\begin{aligned}\frac{d\pi}{d\hat{\pi}}(x) &= \frac{\exp\left(-f(x)-g(x)\right)}{\int_y \exp\left(-f(y)-g(x)-\frac{1}{2\eta}\|y-x\|_2^2-\frac{\eta L^2}{2}\|x-x^*\|_2^2\right)dy}\cdot\frac{Z_{\hat{\pi}}}{Z_\pi}\\ &\geq (2\pi\eta)^{-\frac{d}{2}}\cdot\left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}}\left(1+\frac{\eta L^2}{\mu}\right)^{-\frac{d}{2}} = \left(\frac{1}{1+\eta L}\right)^{\frac{d}{2}}(1+\eta L\kappa)^{-\frac{d}{2}} \geq \frac{1}{2}.\end{aligned}$$

■

H.1.3. CORRECTNESS OF COMPOSITE-SAMPLE-SHARED-MIN

Proposition 40 *Let $\eta = \frac{1}{32L\kappa d\log(288\kappa/\epsilon)}$, and assume* `Sample-Joint-Dist`$(f, g, x^*, \mathcal{O}, \delta)$ *samples within δ total variation of the x-marginal on (14).* `Composite-Sample-Shared-Min` *outputs a sample within total variation ϵ of (12) in an expected $O(1)$ calls to* `Sample-Joint-Dist`*.*

Proof We remark that $\eta = \frac{1}{32L\kappa d\log(288\kappa/\epsilon)}$ is precisely the choice of η in `Sample-Joint-Dist` where $\delta = \epsilon/18$, as in `Composite-Sample-Shared-Min`. First, we may apply Fact 2 to conclude that the measure of set Ω with respect to the μ-strongly logconcave density π is at least $1-\epsilon/3$. The conclusion of correctness will follow from an appeal to Corollary 36, with parameters

$$C = 4,\ \epsilon' = \frac{\epsilon}{3},\ \delta = \frac{\epsilon}{18}.$$

Note that indeed we have $\epsilon' + \frac{2\delta C}{1-\epsilon'}$ is bounded by ϵ, as $1-\epsilon' \geq \frac{2}{3}$. Moreover, the expected number of calls (22) is clearly bounded by a constant as well.

We now show that these parameters satisfy the requirements of Corollary 36. Define the functions

$$\begin{aligned}p(x) &:= \exp(-f(x)-g(x)),\\ \hat{p}(x) &:= (2\pi\eta)^{-\frac{d}{2}}\int_y \exp\left(-f(y)-g(x)-\frac{1}{2\eta}\|y-x\|_2^2-\frac{\eta L^2}{2}\|x-x^*\|_2^2\right)dy,\end{aligned}$$

and observe that clearly the densities of π and $\hat{\pi}$ are respectively proportional to p and $\hat{p}$. Moreover, define $Z = \int p(x)dx$ and $\hat{Z} = \int \hat{p}(x)dx$. By comparing these definitions with Lemma 38, we have $Z = Z_\pi$ and $\hat{Z} = (2\pi\eta)^{-\frac{d}{2}}Z_{\hat{\pi}}$, so by the upper bound in Lemma 38, $\hat{Z}/Z \leq 1$. Next, we claim that the following procedure produces an unbiased estimator for $\frac{p(x)}{\hat{p}(x)}$.

1. Sample $y \sim \pi_x$, where $\frac{d\pi_x(y)}{dy} \propto \exp\left(-f(y)-\frac{1}{2\eta}\|y-x\|_2^2\right)$
2. $\alpha \leftarrow \exp\left(f(y)-\langle\nabla f(x), y-x\rangle-\frac{L}{2}\|y-x\|_2^2+g(x)+\frac{\eta L^2}{2}\|x-x^*\|_2^2\right)$

3. Output $\hat{\theta}(x) \leftarrow \exp\left(-f(x) - g(x) + \frac{\eta}{2(1+\eta L)} \|\nabla f(x)\|_2^2\right)(1+\eta L)^{\frac{d}{2}}\alpha$

To prove correctness of this estimator $\hat{\theta}$, define for simplicity

$$Z_x := \int_y \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\|y-x\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2\right) dy.$$

We compute, using $\frac{d\pi_x(y)}{dy} = \frac{\exp(-f(y)-g(x)-\frac{1}{2\eta}\|y-x\|_2^2-\frac{\eta L^2}{2}\|x-x^*\|_2^2)}{Z_x}$, that

$$\begin{aligned}
\mathbb{E}_{\pi_x}[\alpha] &= \int_y \exp\left(f(y) - \langle \nabla f(x), y-x\rangle - \frac{L}{2}\|y-x\|_2^2 + g(x) + \frac{\eta L^2}{2}\|x-x^*\|_2^2\right) d\pi_x(y) \\
&= \frac{1}{Z_x}\int_y \exp\left(-\langle \nabla f(x), y-x\rangle - \frac{L}{2}\|y-x\|_2^2 - \frac{1}{2\eta}\|y-x\|_2^2\right) dy \\
&= \frac{1}{Z_x}\exp\left(-\frac{\eta}{2(1+\eta L)}\|\nabla f(x)\|_2^2\right)\left(\frac{2\pi\eta}{1+\eta L}\right)^{\frac{d}{2}}.
\end{aligned}$$

This implies that the output quantity

$$\hat{\theta}(x) = \exp\left(-f(x) - g(x) + \frac{\eta}{2(1+\eta L)}\|\nabla f(x)\|_2^2\right)(1+\eta L)^{\frac{d}{2}}\alpha$$

is unbiased for $\frac{p(x)}{\hat{p}(x)} = \exp(-f(x)-g(x))Z_x^{-1}(2\pi\eta)^{\frac{d}{2}}$. Finally, note that for any y used in the definition of $\hat{\theta}(x)$, by using $f(y) - f(x) - \langle \nabla f(x), y-x\rangle - \frac{L}{2}\|y-x\|_2^2 \le 0$ via smoothness, we have

$$\begin{aligned}
\hat{\theta}(x) &= \exp\left(-f(x) - g(x) + \frac{\eta}{2(1+\eta L)}\|\nabla f(x)\|_2^2\right)(1+\eta L)^{\frac{d}{2}}\alpha \\
&\le (1+\eta L)^{\frac{d}{2}}\exp\left(\frac{\eta}{2(1+\eta L)}\|\nabla f(x)\|_2^2 + \frac{\eta L^2}{2}\|x-x^*\|_2^2\right) \\
&\le (1+\eta L)^{\frac{d}{2}}\exp\left(\eta L^2\|x-x^*\|_2^2\right) \le 4.
\end{aligned}$$

Here, we used the definition of η and $L^2\|x-x^*\|_2^2 \le 16 L\kappa d\log(288\kappa/\epsilon)$ by the definition of Ω. ■

H.2. Deferred proofs from Section E.3

Throughout this section, for error tolerance $\delta \in [0,1]$ which parameterizes `Sample-Joint-Dist`, we denote for shorthand a high-probability region Ω_δ and its radius R_δ by

$$\Omega_\delta := \{x \mid \|x-x^*\|_2 \le R_\delta\}, \text{ for } R_\delta := 4\sqrt{\frac{d\log(16\kappa/\delta)}{\mu}}. \tag{27}$$

The following density ratio bounds hold within this region, by simply modifying Lemma 39.

Corollary 41 *Let $\eta = \frac{1}{32L\kappa d\log(16\kappa/\delta)}$, and let $\hat{\pi}$ be parameterized by this choice of η in (14). For all $x \in \Omega_\delta$, as defined in (27), $\frac{d\pi}{d\hat{\pi}}(x) \le 2$. Moreover, for all $x \in \mathbb{R}^d$, $\frac{d\pi}{d\hat{\pi}}(x) \ge \frac{1}{2}$.*

The following claim follows immediately from applying Fact 2.

Lemma 42 *With probability at least $1 - \frac{\delta^2}{8(1+\kappa)^d}$, $x \sim \hat{\pi}$ lies in Ω_δ.*

Finally, when clear from context, we overload $\hat{\pi}$ as a distribution on $x \in \mathbb{R}^d$ to be the x component marginal of the distribution (14), i.e. with density

$$\frac{d\hat{\pi}}{dx}(x) \propto \int_y \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\|y-x\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2\right) dy.$$

We first note that $\hat{\pi}$ is stationary for `Sample-Joint-Dist`; this follows immediately from Lemma 11. In Section H.2.1, we bound the *conductance* of the walk. We then use this bound in Section H.2.2 to bound the mixing time and overall complexity of `Sample-Joint-Dist`.

H.2.1. CONDUCTANCE OF SAMPLE-JOINT-DIST

We bound the conductance of this random walk, as a process on the iterates $\{x_k\}$, to show the final point has distribution close to the marginal of $\hat{\pi}$ on x. To do so, we break Proposition 12 into two pieces, which we will use in a more white-box manner to prove our conductance bound.

Definition 43 (Restricted conductance) *Let a random walk with stationary distribution $\hat{\pi}$ on $x \in \mathbb{R}^d$ have transition densities $\mathcal{T}_x$, and let $\Omega \subseteq \mathbb{R}^d$. The Ω-restricted conductance, for $v \in (0, \frac{1}{2}\hat{\pi}(\Omega))$, is*

$$\Phi_\Omega(v) = \inf_{\hat{\pi}(S\cap\Omega)\in(0,v]} \frac{\mathcal{T}_S(S^c)}{\hat{\pi}(S\cap\Omega)}, \text{ where } \mathcal{T}_S(S^c) := \int_{x\in S}\int_{x'\in S^c} \mathcal{T}_x(x')d\hat{\pi}(x)dx'.$$

Proposition 44 (Lemma 1, Chen et al. (2019)) *Let π_{start} be a β-warm start for $\hat{\pi}$, and let $x_0 \sim \pi_{\text{start}}$. For some $\delta > 0$, let $\Omega \subseteq \mathbb{R}^d$ have $\hat{\pi}(\Omega) \ge 1 - \frac{\delta^2}{2\beta^2}$. Suppose that a random walk with stationary distribution $\hat{\pi}$ satisfies the Ω-restricted conductance bound*

$$\Phi_\Omega(v) \ge \sqrt{B\log\left(\frac{1}{v}\right)}, \text{ for all } v \in \left[\frac{4}{\beta}, \frac{1}{2}\right].$$

Let x_K be the result of K steps of this random walk, starting from x_0. Then, for

$$K \ge \frac{64}{B}\log\left(\frac{\log\beta}{2\delta}\right),$$

the resulting distribution of x_K has total variation at most $\frac{\delta}{2}$ from $\hat{\pi}$.

We state a well-known strategy for lower bounding conductance, via showing the stationary distribution has good *isoperimetry* and that transition distributions of nearby points have large overlap.

Proposition 45 (Lemma 2, Chen et al. (2019)) *Let a random walk with stationary distribution $\hat{\pi}$ on $x \in \mathbb{R}^d$ have transition distribution densities $\mathcal{T}_x$, and let $\Omega \subseteq \mathbb{R}^d$, and let $\hat{\pi}_\Omega$ be the conditional distribution of $\hat{\pi}$ on Ω. Suppose for any $x, x' \in \Omega$ with $\|x - x'\|_2 \le \Delta$,*

$$\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\mathrm{TV}} \le \frac{1}{2}.$$

Also, suppose $\hat{\pi}_\Omega$ satisfies, for any partition S_1, S_2, S_3 of Ω, where $d(S_1, S_2)$ is the minimum Euclidean distance between points in S_1, S_2, the log-isoperimetric inequality

$$\hat{\pi}_\Omega(S_3) \ge \frac{1}{2\psi} d(S_1, S_2) \cdot \min\left(\hat{\pi}_\Omega(S_1), \hat{\pi}_\Omega(S_2)\right) \cdot \sqrt{\log\left(1 + \frac{1}{\min\left(\hat{\pi}_\Omega(S_1), \hat{\pi}_\Omega(S_2)\right)}\right)}. \tag{28}$$

Then, we have the bound for all $v \in (0, \frac{1}{2}]$

$$\Phi_\Omega(v) \ge \frac{\Delta}{128\psi} \sqrt{\log\left(\frac{1}{v}\right)}.$$

To utilize Propositions 44 and 45, we prove the following bounds in Appendices I.1, I.2, and I.3.

Lemma 46 (Warm start) *For $\eta \le \frac{1}{L\kappa d}$, π_{start} defined in (15) is a $2(1+\kappa)^{\frac{d}{2}}$-warm start for $\hat{\pi}$.*

Lemma 47 (Transitions of nearby points) *Suppose $\eta L \le 1$, $\eta L^2 R_\delta^2 \le \frac{1}{2}$, and $400 d^2 \eta \le R_\delta^2$. For a point x, let $\mathcal{T}_x$ be the density of x_k after sampling according to Lines 6 and 7 of Algorithm 5 from $x_{k-1} = x$. For $x, x' \in \Omega_\delta$ with $\|x - x'\|_2 \le \frac{\sqrt{\eta}}{10}$, for Ω_δ defined in (27), we have $\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\mathrm{TV}} \le \frac{1}{2}$.*

Lemma 48 (Isoperimetry) *Density $\hat{\pi}$ and set Ω_δ defined in (14), (27) satisfy (28) with $\psi = 8\mu^{-\frac{1}{2}}$.*

We note that the parameters of Algorithm 5 and the set Ω_δ in (27) satisfy all assumptions of Lemmas 46, 47, and 48. By combining these results in the context of Proposition 45, we see that the random walk satisfies the bound for all $v \in (0, \frac{1}{2}]$:

$$\Phi_{\Omega_\delta}(v) \ge \sqrt{\frac{\eta\mu}{2^{20} \cdot 100} \cdot \log\left(\frac{1}{v}\right)}.$$

Plugging this conductance lower bound, the high-probability guarantee of Ω_δ by Lemma 42, and the warm start bound of Lemma 46 into Proposition 44, we have the following conclusion.

Corollary 49 (Mixing time of ideal `Sample-Joint-Dist`) *Assume that calls to `YSample` are exact in the implementation of `Sample-Joint-Dist`. Then, for any error parameter δ, and*

$$K := \frac{2^{26} \cdot 100}{\eta\mu} \log\left(\frac{d \log(16\kappa)}{4\delta}\right),$$

the distribution of x_K has total variation at most $\frac{\delta}{2}$ from $\hat{\pi}$.

H.2.2. COMPLEXITY OF SAMPLE-JOINT-DIST

We first state a guarantee on the subroutine `YSample`, which we prove in Section I.4.

Lemma 50 (`YSample` guarantee) *For $\delta \in [0,1]$, define R_δ as in (27), and let $\eta = \frac{1}{32L\kappa d\log(16\kappa/\delta)}$. For any x with $\|x - x^*\|_2 \le \sqrt{\kappa d\log(16\kappa/\delta)}\cdot R_\delta$, Algorithm 7 (`YSample`) draws an exact sample y from the density proportional to $\exp\left(-f(y) - \frac{1}{2\eta}\|y-x\|_2^2\right)$ in an expected 2 iterations.*

We also state a result due to Chen et al. (2019), which bounds the mixing time of 1-step Metropolized HMC for well-conditioned distributions; this handles the case when $\|x - x^*\|_2$ is large in Algorithm 7.

Proposition 51 (Theorem 1, Chen et al. (2019)) *Let π be a distribution on $\mathbb{R}^d$ whose negative log-density is convex and has condition number bounded by a constant. Then, Metropolized HMC from an explicit starting distribution mixes to total variation δ to the distribution π in $O(d\log(\frac{d}{\delta}))$ iterations.*

Proposition 52 *`Sample-Joint-Dist` outputs a point with distribution within δ total variation distance from the x-marginal of $\hat\pi$. The expected number of gradient queries per iteration is constant.*

Proof Under an exact `YSample`, Corollary 49 shows the output distribution of `Sample-Joint-Dist` has total variation at most $\frac{\delta}{2}$ from $\hat\pi$. Next, the resulting distribution of the subroutine `YSample` is never larger than $\delta/(2Kd\log(\frac{d\kappa}{\delta}))$ in total variation distance away from an exact sampler. By running for K steps, and using the coupling characterization of total variation, it follows that this can only incur additional error $\delta/(2d\log(\frac{d\kappa}{\delta}))$, proving correctness (in fact, the distribution is always at most $O((d\log(d\kappa/\delta))^{-1})$ away in total variation from an exact `YSample`).

Next, we prove the guarantee on the expected gradient evaluations per iteration. Lemma 50 shows whenever the current iterate x_k has $\|x - x^*\|_2 \le \sqrt{\kappa d\log(16\kappa/\delta)}\cdot R_\delta$, the expected number of gradient evaluations is constant, and moreover Proposition 51 shows that the number of gradient evaluations is never larger than $O(d\log(\frac{d\kappa}{\delta}))$, where we use that the condition number of the log-density in (16) is bounded by a constant. Therefore, it suffices to show in every iteration $0 \le k \le K$, the probability $\|x_k - x^*\|_2 > \sqrt{\kappa d\log(16\kappa/\delta)}\cdot R_\delta$ is $O((d\log(d\kappa/\delta))^{-1})$. By the warmness assumption in Lemma 46, and the concentration bound in Fact 2, the probability x_0 does not satisfy this bound is negligible (inverse exponential in $\kappa d^2\log(\kappa/\delta)$). Since warmness is monotonically decreasing with an exact sampler,[14] and the accumulated error due to inexactness of `YSample` is at most $O((d\log(d\kappa/\delta))^{-1})$ through the whole algorithm, this holds for all iterations. ■

Appendix I. Mixing time ingredients

We now prove facts which are used in the mixing time analysis of `Sample-Joint-Dist`. Throughout this section, as in the specification of `Sample-Joint-Dist`, f and g are functions with properties as in (12), and share a minimizer x^*.

14. This fact is well-known in the literature, and a simple proof is that if a distribution is warm, then taking one step of the Markov chain induces a convex combination of warm point masses, and is thus also warm.

I.1. Warm start

We show that we obtain a warm start for the distribution $\hat{\pi}$ in algorithm `Sample-Joint-Dist` via one call to the restricted Gaussian oracle for g, by proving Lemma 46.

Lemma 53 (Warm start) *For $\eta \le \frac{1}{L\kappa d}$, π_{start} defined in (15) is a $2(1+\kappa)^{\frac{d}{2}}$-warm start for $\hat{\pi}$.*

Proof By the definitions of $\hat{\pi}$ and π_{start} in (14), (15), we wish to bound everywhere the quantity

$$\frac{d\pi_{\text{start}}}{d\hat{\pi}}(x) = \frac{Z_{\hat{\pi}}}{Z_{\text{start}}} \cdot \frac{\exp\left(-\frac{L}{2}\|x-x^*\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2 - g(x)\right)}{\int_y \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\|y-x\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2\right) dy}. \tag{29}$$

Here, $Z_{\hat{\pi}}$ is as in Definition 37, and we let Z_{start} denote the normalization constant of π_{start}, i.e.

$$Z_{\text{start}} := \int_x \exp\left(-\frac{L}{2}\|x-x^*\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2 - g(x)\right) dx.$$

Regarding the first term of (29), the earlier derivation (25) showed

$$\int_y \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\|y-x\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2\right) dy \le (2\pi\eta)^{\frac{d}{2}} \exp\left(-f(x) - g(x)\right).$$

Then, integrating, we can bound the ratio of the normalization constants

$$\begin{aligned}
\frac{Z_{\hat{\pi}}}{Z_{\pi_{\text{start}}}} &\le \frac{\int_x (2\pi\eta)^{\frac{d}{2}} \exp\left(-f(x) - g(x)\right) dx}{\int_x \exp\left(-\frac{L}{2}\|x-x^*\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2 - g(x)\right) dx} \\
&\le \frac{\int_x (2\pi\eta)^{\frac{d}{2}} \exp\left(-f(x^*) - \frac{\mu}{2}\|x-x^*\|_2^2 - g(x)\right) dx}{\int_x \exp\left(-\frac{L}{2}\|x-x^*\|_2^2 - \frac{\mu}{2}\|x-x^*\|_2^2 - g(x)\right) dx} \\
&\le (2\pi\eta)^{\frac{d}{2}} \exp\left(-f(x^*)\right) \left(1 + \frac{L}{\mu}\right)^{\frac{d}{2}}.
\end{aligned} \tag{30}$$

The second inequality followed from f is μ-strongly convex and $\eta L^2 \le \mu$ by assumption. The last inequality followed from Proposition 61, where we used $\frac{\mu}{2}\|x-x^*\|_2^2 + g(x)$ is μ-strongly convex. Next, to bound the second term of (29), notice first that

$$\frac{\exp\left(-\frac{L}{2}\|x-x^*\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2 - g(x)\right)}{\int_y \exp\left(-f(y) - g(x) - \frac{1}{2\eta}\|y-x\|_2^2 - \frac{\eta L^2}{2}\|x-x^*\|_2^2\right) dy} = \frac{\exp\left(-\frac{L}{2}\|x-x^*\|_2^2\right)}{\int_y \exp\left(-f(y) - \frac{1}{2\eta}\|y-x\|_2^2\right) dy}.$$

It thus suffices to lower bound $\exp\left(\frac{L}{2}\|x-x^*\|_2^2\right)\int_y \exp\left(-f(y)-\frac{1}{2\eta}\|y-x\|_2^2\right)dy$. We have

$$\begin{aligned}
&\exp\left(\frac{L}{2}\|x-x^*\|_2^2\right)\int_y \exp\left(-f(y)-\frac{1}{2\eta}\|y-x\|_2^2\right)dy \\
&\geq \exp\left(-f(x)+\frac{L}{2}\|x-x^*\|_2^2\right)\int_y \exp\left(-\langle\nabla f(x), y-x\rangle-\left(\frac{1}{2\eta}+\frac{L}{2}\right)\|y-x\|_2^2\right)dy \\
&= \exp\left(-f(x)+\frac{L}{2}\|x-x^*\|_2^2\right)\left(\frac{2\pi\eta}{1+L\eta}\right)^{\frac{d}{2}}\exp\left(\frac{\eta}{2(1+L\eta)}\|\nabla f(x)\|_2^2\right) \\
&\geq \exp(-f(x^*))\left(\frac{2\pi\eta}{1+L\eta}\right)^{\frac{d}{2}}
\end{aligned} \tag{31}$$

The first and third steps followed from L-smoothness of f, and the second applied the Gaussian integral (Fact 1). Combining the bounds in (30) and (31), (29) becomes

$$\frac{d\pi_{\text{start}}}{d\hat{\pi}}(x) \leq \left(1+\frac{L}{\mu}\right)^{\frac{d}{2}}(1+L\eta)^{\frac{d}{2}} \leq 2(1+\kappa)^{\frac{d}{2}},$$

where $x \in \mathbb{R}^d$ was arbitrary, which completes the proof. ■

I.2. Transitions of nearby points

Here, we prove Lemma 47. Throughout this section, $\mathcal{T}_x$ is the density of x_k, according to the steps in Lines 6 and 7 of `Sample-Joint-Dist` (Algorithm 5) starting at $x_{k-1} = x$. We also define $\mathcal{P}_x$ to be the density of y_k, by just the step in Line 6. We first make a simplifying observation: by Observation 1, for any two points x, x', we have

$$\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\text{TV}} \leq \|\mathcal{P}_x - \mathcal{P}_{x'}\|_{\text{TV}}.$$

Thus, it suffices to understand $\|\mathcal{P}_x - \mathcal{P}_{x'}\|_{\text{TV}}$ for nearby $x, x' \in \Omega_\delta$. Our proof of Lemma 47 combines two pieces: (1) bounding the ratio of normalization constants Z_x, $Z_{x'}$ of $\mathcal{P}_x$ and $\mathcal{P}_{x'}$ for nearby x, x' in Lemma 56 and (2) the structural result Proposition 63. To bound the normalization constant ratio, we state two helper lemmas. Lemma 54 characterizes facts about the minimizer of

$$f(y)+\frac{1}{2\eta}\|y-x\|_2^2. \tag{32}$$

Lemma 54 *Let f be convex with minimizer x^*, and y_x minimize (32) for a given x. Then,*

1. $\|y_x - y_{x'}\|_2 \leq \|x-x'\|_2$.
2. *For any x,* $\|y_x - x^*\|_2 \leq \|x-x^*\|_2$.
3. *For any x with $\|x-x^*\|_2 \leq R$,* $\|x-y_x\|_2 \leq \eta LR$.

Proof By optimality conditions in the definition of y_x,

$$\eta \nabla f(y_x) = x - y_x.$$

Fix two points x, x', and let $x_t := (1-t)x + tx'$. Letting $\mathbf{J}_x(y_x)$ be the Jacobian matrix of y_x,

$$\begin{aligned}\frac{d}{dt}\eta\nabla f(y_{x_t}) = \frac{d}{dt}(x_t - y_{x_t}) &\implies \eta\nabla^2 f(y_{x_t})\mathbf{J}_x(y_{x_t})(x'-x) = (\mathbf{I} - \mathbf{J}_x(y_{x_t}))(x'-x) \\ &\implies \mathbf{J}_x(y_{x_t})(x'-x) = (\mathbf{I} + \eta\nabla^2 f(y_{x_t}))^{-1}(x'-x).\end{aligned}$$

We can then compute

$$y_{x'} - y_x = \int_0^1 \frac{d}{dt} y_{x_t} dt = \int_0^1 \mathbf{J}_x(y_{x_t})(x'-x)dt = \int_0^1 (\mathbf{I} + \eta\nabla^2 f(y_{x_t}))^{-1}(x'-x)dt.$$

By triangle inequality and convexity of f, the first claim follows:

$$\|y_{x'} - y_x\|_2 \le \int_0^1 \left\|(\mathbf{I} + \eta\nabla^2 f(y_{x_t}))^{-1}\right\|_2 \left\|x' - x\right\|_2 dt \le \left\|x' - x\right\|_2.$$

The second claim follows from the first by $y_{x^*} = x^*$. The third claim follows from the second via

$$\|x - y_x\|_2 = \eta \|\nabla f(y_x)\|_2 \le \eta L \|y_x - x^*\|_2 \le \eta L R.$$

■

Next, Lemma 55 states well-known bounds on the integral of a well-conditioned function h.

Lemma 55 *Let h be a L_h-smooth, μ_h-strongly convex function and let y_h^* be its minimizer. Then*

$$\left(2\pi L_h^{-1}\right)^{\frac{d}{2}} \exp\left(-h(y_h^*)\right) \le \int_y \exp\left(-h(y)\right) \le \left(2\pi\mu_h^{-1}\right)^{\frac{d}{2}} \exp\left(-h(y_h^*)\right).$$

Proof By smoothness and strong convexity,

$$\exp\left(-h(y_h^*) - \frac{L_h}{2}\|y - y_h^*\|_2^2\right) \le \exp(-h(y)) \le \exp\left(-h(y_h^*) - \frac{\mu_h}{2}\|y - y_h^*\|_2^2\right).$$

The result follows by Gaussian integrals, i.e. Fact 1. ■

We now define the normalization constants of $\mathcal{P}_x$ and $\mathcal{P}_{x'}$:

$$\begin{aligned}Z_x &= \int_y \exp\left(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2\right) dy, \\ Z_{x'} &= \int_y \exp\left(-f(y) - \frac{1}{2\eta}\|y - x'\|_2^2\right) dy.\end{aligned} \tag{33}$$

We apply Lemma 54 and Lemma 55 to bound the ratio of Z_x and $Z_{x'}$.

Lemma 56 *Let f be μ-strongly convex and L-smooth. Let $x, x' \in \Omega_\delta$, for Ω_δ defined in (27), and let $\|x - x'\|_2 \le \Delta$. Then, the normalization constants Z_x and $Z_{x'}$ in (33) satisfy*

$$\frac{Z_x}{Z_{x'}} \le 1.05 \exp\left(3LR\Delta + \frac{L\Delta^2}{2}\right).$$

Proof First, applying Lemma 55 to Z_x and $Z_{x'}$ yields that the ratio is bounded by

$$\begin{aligned}\frac{Z_x}{Z_{x'}} &\le \frac{\exp\left(-f(y_x) - \frac{1}{2\eta}\|y_x - x\|_2^2\right)\left(2\pi\left(\mu + \frac{1}{\eta}\right)^{-1}\right)^{\frac{d}{2}}}{\exp\left(-f(y_{x'}) - \frac{1}{2\eta}\|y_{x'} - x\|_2^2\right)\left(2\pi\left(L + \frac{1}{\eta}\right)^{-1}\right)^{\frac{d}{2}}} \\ &\le 1.05 \exp\left(f(y_{x'}) - f(y_x) + \frac{1}{2\eta}\left(\|y_{x'} - x'\|_2^2 - \|y_x - x\|_2^2\right)\right).\end{aligned}$$

Here, we used the bound for $\eta^{-1} \ge 32Ld$ that

$$\left(\frac{L + \frac{1}{\eta}}{\mu + \frac{1}{\eta}}\right)^{d/2} \le 1.05.$$

Regarding the remaining term, recall x, x' both belong to Ω_δ, and $\|x - x'\|_2 \le \Delta$. We have

$$\begin{aligned}&f(y_{x'}) - f(y_x) + \frac{1}{2\eta}\left(\|y_{x'} - x'\|_2^2 - \|y_x - x\|_2^2\right) \\ \le\ &\langle \nabla f(y_x), y_{x'} - y_x\rangle + \frac{L}{2}\|y_{x'} - y_x\|_2^2 + \frac{1}{2\eta}\langle y_{x'} - x' + y_x - x, y_{x'} - y_x + x - x'\rangle \\ \le\ &LR\Delta + \frac{L\Delta^2}{2} + \frac{1}{2\eta}\left(\|y_x - x\|_2 + \|y_{x'} - x'\|_2\right)\left(\|y_{x'} - y_x\|_2 + \|x' - x\|_2\right) \\ \le\ &LR\Delta + \frac{L\Delta^2}{2} + \frac{2\eta LR}{2\eta}\left(\|y_{x'} - y_x\|_2 + \|x' - x\|_2\right) \le 3LR\Delta + \frac{L\Delta^2}{2}.\end{aligned}$$

The first inequality was smoothness and expanding the difference of quadratics. The second was by $\|\nabla f(y_x)\|_2 \le L\|y_x - x^*\|_2 \le LR$ and $\|y_{x'} - y_x\|_2 \le \Delta$, where we used the first and second parts of Lemma 54; we also applied Cauchy-Schwarz and triangle inequality. The third used the third part of Lemma 54. Finally, the last inequality was by the first part of Lemma 54 and $\|x' - x\|_2 \le \Delta$. ■

We now are ready to prove Lemma 47.

Lemma 57 (Transitions of nearby points) *Suppose $\eta L \le 1$, $\eta L^2 R_\delta^2 \le \frac{1}{2}$, and $400d^2\eta \le R_\delta^2$. For a point x, let $\mathcal{T}_x$ be the density of x_k after sampling according to Lines 6 and 7 of Algorithm 5 from $x_{k-1} = x$. For $x, x' \in \Omega_\delta$ with $\|x - x'\|_2 \le \frac{\sqrt{\eta}}{10}$, for Ω_δ defined in (27), we have $\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\mathrm{TV}} \le \frac{1}{2}$.*

Proof First, by Observation 1, it suffices to show $\|\mathcal{P}_x - \mathcal{P}_{x'}\|_{\mathrm{TV}} \le \frac{1}{2}$. Pinsker's inequality states

$$\|\mathcal{P}_x - \mathcal{P}_{x'}\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2} d_{\mathrm{KL}}\left(\mathcal{P}_x, \mathcal{P}_{x'}\right)},$$

where d_{KL} is KL-divergence, so it is enough to show $d_{\mathrm{KL}}\left(\mathcal{P}_x, \mathcal{P}_{x'}\right) \le \frac{1}{2}$. Notice that

$$d_{\mathrm{KL}}\left(\mathcal{P}_x, \mathcal{P}_{x'}\right) = \log\left(\frac{Z_{x'}}{Z_x}\right) + \int_y \mathcal{P}_x(y) \log\left(\frac{\exp\left(-f(y) - \frac{1}{2\eta}\left\|y-x\right\|_2^2\right)}{\exp\left(-f(y) - \frac{1}{2\eta}\left\|y-x'\right\|_2^2\right)}\right) dy.$$

By Lemma 56, the first term satisfies, for $\Delta := \frac{\sqrt{\eta}}{10}$,

$$\log\left(\frac{Z_{x'}}{Z_x}\right) \le 3LR\Delta + \frac{L\Delta^2}{2} + \log(1.05).$$

To bound the second term, we have

$$\begin{aligned}
\int_y \mathcal{P}_x(y) \log\left(\frac{\exp\left(-f(y) - \frac{1}{2\eta}\left\|y-x\right\|_2^2\right)}{\exp\left(-f(y) - \frac{1}{2\eta}\left\|y-x'\right\|_2^2\right)}\right) dy &= \frac{1}{2\eta}\int_y \mathcal{P}_x(y)\left(\left\|y-x'\right\|_2^2 - \left\|y-x\right\|_2^2\right) dy \\
&= \frac{1}{2\eta}\int_y \mathcal{P}_x(y)\left\langle x-x', 2\left(y-x\right) + \left(x-x'\right)\right\rangle dy \\
&\le \frac{\Delta^2}{2\eta} + \frac{\Delta}{\eta}\left\|\int_y y\mathcal{P}_x(y)dy - x\right\|_2.
\end{aligned}$$

Here, the second line was by expanding and the third line was by $\left\|x-x'\right\|_2 \le \Delta$ and Cauchy-Schwarz. By Proposition 63, $\left\|\int_y y\mathcal{P}_x(y)dy - x\right\|_2 \le 2\eta LR$, where by assumption the parameters satisfy the conditions of Proposition 63. Then, combining the two bounds, we have

$$d_{\mathrm{KL}}\left(\mathcal{P}_x, \mathcal{P}_{x'}\right) \le 3LR\Delta + \frac{L\Delta^2}{2} + \frac{\Delta^2}{2\eta} + 2LR\Delta + \log(1.05) = 5LR\Delta + \frac{L\Delta^2}{2} + \frac{\Delta^2}{2\eta} + \log(1.05).$$

When $\Delta = \frac{\sqrt{\eta}}{10}$, $\eta L \le 1$, and $\eta L^2 R^2 \le \frac{1}{2}$, we have the desired

$$d_{\mathrm{KL}}\left(\mathcal{P}_x, \mathcal{P}_{x'}\right) \le \frac{\sqrt{\eta}LR}{2} + \frac{L\eta}{200} + \frac{1}{200} + \log(1.05) \le \frac{1}{2}.$$

■

I.3. Isoperimetry

In this section, we prove Lemma 48, which asks to show that $\hat{\pi}_{\Omega_\delta}$ satisfies a log-isoperimetric inequality (28). Here, we define $\hat{\pi}_{\Omega_\delta}$ to be the conditional distribution of the $\hat{\pi}$ x-marginal on set Ω_δ. We recall this means that for any partition S_1, S_2, S_3 of Ω_δ,

$$\hat{\pi}_{\Omega_\delta}(S_3) \ge \frac{1}{2\psi} d(S_1, S_2) \cdot \min\left(\hat{\pi}_{\Omega_\delta}(S_1), \hat{\pi}_{\Omega_\delta}(S_2)\right) \cdot \sqrt{\log\left(1 + \frac{1}{\min\left(\hat{\pi}_{\Omega_\delta}(S_1), \hat{\pi}_{\Omega_\delta}(S_2)\right)}\right)}.$$

The following fact was shown in Chen et al. (2019).

Lemma 58 (Chen et al. (2019), Lemma 11) *Any μ-strongly logconcave distribution π satisfies the log-isoperimetric inequality (28) with $\psi = \mu^{-\frac{1}{2}}$.*

Observe that π_{Ω_δ}, the restriction of π to the convex set Ω_δ, is μ-strongly logconcave by the definition of π (12), so it satisfies a log-isoperimetric inequality. We now combine this fact with the relative density bounds Lemma 39 to prove Lemma 48.

Lemma 59 (Isoperimetry) *Density $\hat{\pi}$ and set Ω_δ defined in (14), (27) satisfy (28) with $\psi = 8\mu^{-\frac{1}{2}}$.*

Proof Fix some partition S_1, S_2, S_3 of Ω_δ, and without loss of generality let $\hat{\pi}_{\Omega_\delta}(S_1) \le \hat{\pi}_{\Omega_\delta}(S_2)$. First, by applying Corollary 41, which shows $\frac{d\pi}{d\hat{\pi}}(x) \in [\frac{1}{2}, 2]$ everywhere in Ω_δ, we have the bounds

$$\frac{1}{2}\pi_{\Omega_\delta}(S_1) \le \hat{\pi}_{\Omega_\delta}(S_1) \le 2\pi_{\Omega_\delta}(S_1), \ \frac{1}{2}\pi_{\Omega_\delta}(S_2) \le \hat{\pi}_{\Omega_\delta}(S_2) \le 2\pi_{\Omega_\delta}(S_2), \text{ and } \hat{\pi}_{\Omega_\delta}(S_3) \ge \frac{1}{2}\pi_{\Omega_\delta}(S_3).$$

Therefore, we have the sequence of conclusions

$$\begin{aligned}
\hat{\pi}_{\Omega_\delta}(S_3) &\ge \frac{1}{2}\pi_{\Omega_\delta}(S_3) \\
&\ge \frac{d(S_1,S_2)\sqrt{\mu}}{4} \cdot \min\left(\pi_{\Omega_\delta}(S_1), \pi_{\Omega_\delta}(S_2)\right) \cdot \sqrt{\log\left(1 + \frac{1}{\min\left(\pi_{\Omega_\delta}(S_1), \pi_{\Omega_\delta}(S_2)\right)}\right)} \\
&\ge \frac{d(S_1,S_2)\sqrt{\mu}}{8} \cdot \hat{\pi}_{\Omega_\delta}(S_1) \cdot \sqrt{\log\left(1 + \frac{1}{2\hat{\pi}_{\Omega_\delta}(S_1)}\right)} \\
&\ge \frac{d(S_1,S_2)\sqrt{\mu}}{16} \cdot \hat{\pi}_{\Omega_\delta}(S_1) \cdot \sqrt{\log\left(1 + \frac{1}{\hat{\pi}_{\Omega_\delta}(S_1)}\right)}.
\end{aligned}$$

Here, the second line was by applying Lemma 58 to the μ-strongly logconcave distribution π_{Ω_δ}, and the final line used $\sqrt{\log(1+\alpha)} \le 2\sqrt{\log(1+\frac{\alpha}{2})}$ for all $\alpha > 0$. ■

I.4. Correctness of `YSample`

In this section, we show how we can sample y efficiently in the alternating scheme of the algorithm `Sample-Joint-Dist`, within an extremely high probability region. Specifically, for any x with $\|x - x^*\|_2 \le \sqrt{\kappa d \log(16\kappa/\delta)} \cdot R_\delta$, where R_δ is defined in (27), we give a method for implementing

$$\text{draw } y \propto \exp\left(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2\right) dy.$$

The algorithm is Algorithm 7, which is a simple rejection sampling scheme.

We recall that we gave guarantees on rejection sampling procedures in Lemma 14 (an "exact" version of Lemma 35 and Corollary 36). We now prove Lemma 50 via a direct application of Lemma 14.

Lemma 60 (`YSample` guarantee) *For $\delta \in [0,1]$, define R_δ as in (27), and let $\eta = \frac{1}{32L\kappa d\log(16\kappa/\delta)}$. For any x with $\|x - x^*\|_2 \le \sqrt{\kappa d \log(16\kappa/\delta)} \cdot R_\delta$, Algorithm 7 (`YSample`) draws an exact sample y from the density proportional to $\exp\left(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2\right)$ in an expected 2 iterations.*

Algorithm 7 `YSample`(f, x, η, δ)

Input: L-smooth, μ-strongly convex $f : \mathbb{R}^d \to \mathbb{R}$ with minimizer x^*, $\eta > 0$, $\delta \in [0, 1]$, $x \in \mathbb{R}^d$.
Output: If $\|x - x^*\|_2 \le \sqrt{\kappa d \log(16\kappa/\delta)} \cdot R_\delta$, return exact sample from distribution with density $\propto \exp(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2)$ (see (27) for definition of R_δ). Otherwise, return sample within δ TV from distribution with density $\propto \exp(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2)$.
if $\|x - x^*\|_2 \le \sqrt{\kappa d \log(16\kappa/\delta)} \cdot R_\delta$ **then**
  while true do
    Draw $y \sim \mathcal{N}(x - \eta\nabla f(x), \eta\mathbf{I})$
    $\tau \sim \text{Unif}[0, 1]$
    if $\tau \le \exp(f(x) + \langle \nabla f(x), y - x\rangle - f(y))$ **then**
      return y
    end if
  end while
end if
return Sample x within TV δ from density $\propto \exp(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2)$ using Chen et al. (2019)

Proof For $\|x - x^*\|_2 \le \sqrt{\kappa d \log(16\kappa/\delta)} \cdot R_\delta$, `YSample` is a rejection sampling scheme with

$$p(y) = \exp\left(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2\right), \ \hat{p}(y) = \exp\left(-f(x) - \langle \nabla f(x), y - x\rangle - \frac{1}{2\eta}\|y - x\|_2^2\right).$$

It is clear that $p(y) \le \hat{p}(y)$ everywhere by convexity of f, so we may choose $C = 1$. To bound the expected number of iterations and obtain the desired conclusion, Lemma 14 requires a bound on

$$\frac{\int_y \exp\left(-f(x) - \langle \nabla f(x), y - x\rangle - \frac{1}{2\eta}\|y - x\|_2^2\right) dy}{\int_y \exp\left(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2\right) dy}, \tag{34}$$

the ratio of the normalization constants of $\hat{p}$ and p. First, by Fact 1,

$$\int_y \exp\left(-f(x) - \langle \nabla f(x), y - x\rangle - \frac{1}{2\eta}\|y - x\|_2^2\right) dy = \exp\left(-f(x) + \frac{\eta}{2}\|\nabla f(x)\|_2^2\right)(2\pi\eta)^{\frac{d}{2}}.$$

Next, by smoothness and Fact 1 once more,

$$\begin{aligned}\int_y \exp\left(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2\right) dy &\ge \int_y \exp\left(-f(x) - \langle \nabla f(x), y - x\rangle - \frac{1 + \eta L}{2\eta}\|y - x\|_2^2\right) dy \\ &= \exp\left(-f(x) + \frac{\eta}{2(1 + \eta L)}\|\nabla f(x)\|_2^2\right)\left(\frac{2\pi\eta}{1 + \eta L}\right)^{\frac{d}{2}}.\end{aligned}$$

Taking a ratio, the quantity in (34) is bounded above by

$$\begin{aligned}\exp\left(\left(\frac{\eta}{2} - \frac{\eta}{2(1 + \eta L)}\right)\|\nabla f(x)\|_2^2\right)(1 + \eta L)^{\frac{d}{2}} &\le 1.5\exp\left(\frac{\eta^2 L}{2(1 + \eta L)}\|\nabla f(x)\|_2^2\right) \\ &\le 1.5\exp\left(\frac{\eta^2 L^3}{2} \cdot \left(\frac{16\kappa d^2 \log^2(16\kappa/\delta)}{\mu}\right)\right) \le 2.\end{aligned}$$

The first inequality was $(1+\eta L)^{\frac{d}{2}} \le 1.5$, the second used smoothness and the assumed bound on $\|x-x^*\|_2$, and the third again used our choice of η. ■

Appendix J. Structural results

Here, we prove two structural results about distributions whose negative log-densities are small perturbations of a quadratic, which obtain tighter concentration guarantees compared to naive bounds on strongly logconcave distributions. They are used in obtaining our bounds in Section I (and for the warm start bounds in Section D), but we hope both the statements and proof techniques are of independent interest to the community. Our first structural result is a bound on normalization constant ratios, used throughout the paper.

Proposition 61 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex with minimizer x^*, and let $\lambda > 0$. Then,*

$$\frac{\int \exp(-f(x))dx}{\int \exp\left(-f(x) - \frac{1}{2\lambda}\|x-x^*\|_2^2\right)dx} \le \left(1+\frac{1}{\mu\lambda}\right)^{\frac{d}{2}}.$$

Proof Define the function

$$R(\alpha) := \frac{\int \exp\left(-f(x) - \frac{1}{2\lambda\alpha}\|x-x^*\|_2^2\right)dx}{\int \exp\left(-f(x) - \frac{1}{2\lambda}\|x-x^*\|_2^2\right)dx}.$$

Let $d\pi_\alpha(x)$ be the density proportional to $\exp\left(-f(x) - \frac{1}{2\lambda\alpha}\|x-x^*\|_2^2\right)dx$. We compute

$$\begin{aligned}
\frac{d}{d\alpha}R(\alpha) &= \int \frac{\exp\left(-f(x) - \frac{1}{2\lambda\alpha}\|x-x^*\|_2^2\right)}{\int \exp\left(-f(x) - \frac{1}{2\lambda}\|x-x^*\|_2^2\right)dx}\frac{1}{2\lambda\alpha^2}\|x-x^*\|_2^2\,dx \\
&= \frac{R(\alpha)}{2\lambda\alpha^2}\int \frac{\exp\left(-f(x) - \frac{1}{2\lambda\alpha}\|x-x^*\|_2^2\right)\|x-x^*\|_2^2}{\int \exp\left(-f(x) - \frac{1}{2\lambda\alpha}\|x-x^*\|_2^2\right)dx}dx \\
&= \frac{R(\alpha)}{2\lambda\alpha^2}\int \|x-x^*\|_2^2\,d\pi_\alpha(x) \le \frac{R(\alpha)}{2\alpha}\cdot\frac{d}{\mu\lambda\alpha+1}.
\end{aligned}$$

Here, the last inequality was by Fact 4, using the fact that the function $f(x) + \frac{1}{2\lambda\alpha}\|x-x^*\|_2^2$ is $\mu + \frac{1}{\lambda\alpha}$-strongly convex. Moreover, note that $R(1) = 1$, and

$$\frac{d}{d\alpha}\log\left(\frac{\alpha}{\mu\lambda\alpha+1}\right) = \frac{1}{\alpha} - \frac{\mu\lambda}{\mu\lambda\alpha+1} = \frac{1}{\mu\lambda\alpha^2+\alpha}.$$

Solving the differential inequality

$$\frac{d}{d\alpha}\log(R(\alpha)) = \frac{dR(\alpha)}{d\alpha}\cdot\frac{1}{R(\alpha)} \le \frac{d}{2}\cdot\frac{1}{\mu\lambda\alpha^2+\alpha},$$

we obtain the bound for any $\alpha \geq 1$ (since $\log(R(1)) = 0$)

$$\log(R(\alpha)) \leq \frac{d}{2} \log\left(\frac{\mu\lambda\alpha + \alpha}{\mu\lambda\alpha + 1}\right) \implies R(\alpha) \leq \left(\frac{\mu\lambda\alpha + \alpha}{\mu\lambda\alpha + 1}\right)^{\frac{d}{2}} \leq \left(1 + \frac{1}{\mu\lambda}\right)^{\frac{d}{2}}.$$

Taking a limit $\alpha \to \infty$ yields the conclusion. ■

Our second structural result uses a similar proof technique to show that the mean of a bounded perturbation f of a Gaussian is not far from its mode, as long as the gradient of the mode is small. We remark that one may directly apply strong logconcavity, i.e. a variant of Fact 4, to obtain a weaker bound by roughly a $\sqrt{d}$ factor, which would result in a loss of $\Omega(d)$ in the guarantees of Theorem 5. This tighter analysis is crucial in our improved mixing time result.

Before stating the bound, we apply Fact 3 to the convex functions $h(x) = (\theta^\top x)^2$ and $h(x) = \|x\|_2^4$ to obtain the following conclusions which will be used in the proof of Proposition 63.

Corollary 62 *Let π be a μ-strongly logconcave density. Then,*

1. $\mathbb{E}_\pi[(\theta^\top(x - \mathbb{E}_\pi[x]))^2] \leq \mu^{-1}$, *for all unit vectors θ.*
2. $\mathbb{E}_\pi[\|x - \mathbb{E}_\pi[x]\|_2^4] \leq 3d^2\mu^{-2}$.

Proposition 63 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and convex with minimizer x^*, let $x \in \mathbb{R}^d$ with $\|x - x^*\|_2 \leq R$, and let $d\pi_\eta(y)$ be the density proportional to $\exp\left(-f(y) - \frac{1}{2\eta}\|y - x\|_2^2\right) dy$. Suppose that $\eta \leq \min\left(\frac{1}{2L^2R^2}, \frac{R^2}{400d^2}\right)$. Then,*

$$\left\|\mathbb{E}_{\pi_\eta}[y] - x\right\|_2 \leq 2\eta LR.$$

Proof Define a family of distributions π^α for $\alpha \in [0, 1]$, with

$$d\pi^\alpha(y) \propto \exp\left(-\alpha\left(f(y) - f(x) - \langle\nabla f(x), y - x\rangle\right) - f(x) - \langle\nabla f(x), y - x\rangle - \frac{1}{2\eta}\|y - x\|_2^2\right) dy.$$

In particular, $\pi^1 = \pi_\eta$, and π^0 is a Gaussian with mean $x - \eta\nabla f(x)$. We define $\bar{y}_\alpha := \mathbb{E}_{\pi_\alpha}[y]$, and

$$y_\alpha^* := \text{argmin}_y \left\{\alpha\left(f(y) - f(x) - \langle\nabla f(x), y - x\rangle\right) + f(x) + \langle\nabla f(x), y - x\rangle + \frac{1}{2\eta}\|y - x\|_2^2\right\}.$$

Define the function $D(\alpha) := \|\bar{y}_\alpha - x\|_2$, such that we wish to bound $D(1)$. First, by smoothness

$$D(0) = \|\mathbb{E}_{\pi_0}[y] - x\|_2 = \|\eta\nabla f(x)\|_2 \leq \eta LR.$$

Next, we observe

$$\frac{d}{d\alpha}D(\alpha) = \left\langle\frac{\bar{y}_\alpha - x}{\|\bar{y}_\alpha - x\|_2}, \frac{d\bar{y}_\alpha}{d\alpha}\right\rangle \leq \left\|\frac{d\bar{y}_\alpha}{d\alpha}\right\|_2.$$

In order to bound $\left\|\frac{d\bar{y}_\alpha}{d\alpha}\right\|_2$, fix a unit vector θ. We have

$$\begin{aligned}\left\langle \frac{d\bar{y}_\alpha}{d\alpha}, \theta \right\rangle &= \frac{d}{d\alpha}\left\langle \int (y-x)d\pi^\alpha(y), \theta \right\rangle \\ &= \int \langle y-x,\theta\rangle \left(f(x) + \langle \nabla f(x), y-x\rangle - f(y)\right) d\pi^\alpha(y) \\ &\le \sqrt{\int (\langle y-x,\theta\rangle)^2 d\pi^\alpha(y)} \sqrt{\int (f(x) + \langle \nabla f(x), y-x\rangle - f(y))^2 d\pi^\alpha(y)} \\ &\le \sqrt{\int (\langle y-x,\theta\rangle)^2 d\pi^\alpha(y)} \sqrt{\int \frac{L^2}{4}\|y-x\|_2^4 \, d\pi^\alpha(y)}.\end{aligned} \tag{35}$$

The third line was Cauchy-Schwarz and the last line used smoothness and convexity, i.e.

$$-\frac{L}{2}\|y-x\|_2^2 \le f(x) + \langle \nabla f(x), y-x\rangle - f(y) \le 0.$$

We now bound these terms. First,

$$\begin{aligned}\int (\langle y-x,\theta\rangle)^2 d\pi^\alpha(y) &\le 2\int (\langle y-\bar{y}_\alpha,\theta\rangle)^2 d\pi^\alpha(y) + 2\int (\langle \bar{y}_\alpha - x,\theta\rangle)^2 d\pi^\alpha(y) \\ &\le 2\eta + 2\|\bar{y}_\alpha - x\|_2^2 = 2\eta + 2D(\alpha)^2.\end{aligned} \tag{36}$$

Here, we applied the first part of Corollary 62, as π^α is η^{-1}-strongly logconcave, and the definition of $D(\alpha)$. Next, using for any $a, b \in \mathbb{R}^d$, $\|a+b\|_2^4 \le (\|a\|_2 + \|b\|_2)^4 \le 16\|a\|_2^4 + 16\|b\|_2^4$, we have

$$\begin{aligned}\int \frac{L^2}{4}\|y-x\|_2^4 \, d\pi^\alpha(y) &\le \int 4L^2\|y-\bar{y}_\alpha\|_2^4 \, d\pi^\alpha(y) + \int 4L^2\|x-\bar{y}_\alpha\|_2^4 \, d\pi^\alpha(y) \\ &\le 12L^2d^2\eta^2 + 4L^2D(\alpha)^4.\end{aligned} \tag{37}$$

Here, we used the second part of Corollary 62. Maximizing (35) over θ, and applying (36), (37),

$$\begin{aligned}\frac{d}{d\alpha}D(\alpha) \le \left\|\frac{d\bar{y}_\alpha}{d\alpha}\right\|_2 &\le \sqrt{8L^2(\eta + D(\alpha)^2)(3d^2\eta^2 + D(\alpha)^4)} \\ &\le 4L(\sqrt{\eta} + D(\alpha)) \cdot \max(2\eta d, D(\alpha)^2).\end{aligned} \tag{38}$$

Assume for contradiction that $D(1) > 2\eta LR$, violating the conclusion of the proposition. By continuity of D, there must have been some $\bar{\alpha} \in (0,1)$ where $D(\bar{\alpha}) = 2\eta LR$, and for all $0 \le \alpha < \bar{\alpha}$, $D(\alpha) < 2\eta LR$. By the mean value theorem, there then exists $0 \le \hat{\alpha} \le \bar{\alpha}$ such that

$$\frac{dD(\hat{\alpha})}{d\alpha} = \frac{D(\bar{\alpha}) - D(0)}{\bar{\alpha}} > \eta LR.$$

On the other hand, by our assumption that $2\eta L^2R^2 \le 1$, for any $d \ge 1$ it follows that

$$2\eta d \ge 4\eta^2L^2R^2 > D(\hat{\alpha})^2, \;\; \sqrt{2\eta} \ge 2\eta LR > D(\hat{\alpha}).$$

Then, plugging these bounds into (38) and using $\sqrt{\eta} + D(\hat{\alpha}) \le \frac{5}{2}\sqrt{\eta}$ as $\sqrt{2} \le \frac{3}{2}$,

$$\frac{d}{d\alpha}D(\hat{\alpha}) \le 4L \cdot \frac{5}{2}\sqrt{\eta} \cdot 2\eta d = 20\sqrt{\eta}\frac{d}{R} \cdot \eta LR \le \eta LR.$$

We used $\eta \le \frac{R^2}{400d^2}$ in the last inequality. This is a contradiction, implying $D(1) \le 2\eta LR$. ■