Proceedings of Machine Learning Research vol 195:1–41, 2023 36th Annual Conference on Learning Theory

Algorithmic Aspects of the Log-Laplace Transform and a Non-Euclidean Proximal Sampler

Sivakanth Gopi SIGOPI@MICROSOFT.COM
Microsoft Research

Yin Tat Lee YINTATLEE@MICROSOFT.COM
Microsoft Research

Daogao Liu DGLIU@UW.EDU
University of Washington

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

Kevin Tian TIANKEVIN@MICROSOFT.COM
Microsoft Research

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

The development of efficient sampling algorithms catering to non-Euclidean geometries has been a challenging endeavor, as discretization techniques which succeed in the Euclidean setting do not readily carry over to more general settings. We develop a non-Euclidean analog of the recent proximal sampler of Lee et al. (2021b), which naturally induces regularization by an object known as the log-Laplace transform (LLT) of a density. We prove new mathematical properties (with an algorithmic flavor) of the LLT, such as strong convexity-smoothness duality and an isoperimetric inequality, which are used to prove a mixing time on our proximal sampler matching Lee et al. (2021b) under a warm start. As our main application, we show our warm-started sampler improves the value oracle complexity of differentially private convex optimization in ℓ_p and Schatten-p norms for $p \in [1, 2]$ to match the Euclidean setting Gopi et al. (2022), while retaining state-of-the-art excess risk bounds Gopi et al. (2023). We find our investigation of the LLT to be a promising proof-of-concept of its utility as a tool for designing samplers, and outline directions for future exploration.
Keywords: sampling, private convex optimization, Non-Euclidean geometry, log-Laplace transform

1. Introduction

The development of samplers for continuous distributions, under weak oracle access to the corresponding densities, has seen a flurry of recent research activity. For applications in settings inspired by machine learning or computational statistics, this development has in large part built upon connections between sampling and continuous optimization. Inspired by perspectives on sampling as optimization in the space of measures Jordan et al. (1998) and starting with pioneering work of Dalalyan (2017b), a long sequence of results, e.g. Dalalyan (2017a); Cheng et al. (2018); Dwivedi et al. (2019); Durmus and Moulines (2019); Chen and Vempala (2019); Durmus et al. (2019); Shen and Lee (2019); Chen et al. (2020); Lee et al. (2020); Chewi et al. (2021), has used analysis techniques from convex optimization to bound the convergence rates of sampling algorithms for densities. We refer the reader to the survey Chewi (2023) for a more complete account, but note in almost all cases, the focus has been on sampling from densities satisfying regularity assumptions stated in the Euclidean (ℓ_2) norm, e.g. ℓ_2-bounded derivatives.

Continuous optimization under regularity assumptions stated for non-Euclidean geometries has played an important role in algorithm design. These geometries naturally arise when the optimization problem is over a structured constraint set, such as an ℓ_p ball or a polytope. In diverse applications such as learning from experts Arora et al. (2012), sparse recovery Candès et al. (2006), multi-armed bandits Bubeck and

Cesa-Bianchi (2012), matrix completion Agarwal et al. (2010), fair resource allocation Diakonikolas et al. (2020), and robust PCA Jambulapati et al. (2020), first-order mirror descent techniques for ℓ_p or Schatten-p geometries have been a remarkable success story. Beyond these applications, the theory of self-concordant barriers (and the Riemannian geometries induced by their Hessians) has been greatly influential to the theory of convex programming and interior point methods Nesterov and Todd (2002); Nemirovski (2004).[1]

Non-Euclidean samplers. A natural direction for building the theory of logconcave sampling (the analog of convex optimization) is thus to develop samplers handling non-Euclidean regularity assumptions and constraints. Unfortunately, progress in this direction has relatively lagged behind optimization counterparts, as discretization tools which work in the Euclidean case do not generalize. Briefly (with an extended discussion deferred to Section 1.3), most prior attempts at giving non-Euclidean samplers have focused on analyzing variants of the *mirrored Langevin dynamics*, building upon the ubiquitous mirror descent algorithm in optimization Nemirovski and Yudin (1983). The key idea of mirror descent is to choose a regularizer $\phi : \mathcal{X} \to \mathbb{R}$ over a constraint set $\mathcal{X}$, such that ϕ is strongly convex in an appropriate (possibly non-Euclidean) norm $\|\cdot\|_{\mathcal{X}}$. The regularizer ϕ is then used to define iterative methods for optimizing functions f with regularity in $\|\cdot\|_{\mathcal{X}}$.

The sampling analog of this non-Euclidean generalization is to extend the *Langevin dynamics*, a stochastic process inherently catered to the ℓ_2 geometry, to use Brownian motion reweighted by the Hessian of a regularizer ϕ. This process, the mirrored Langevin dynamics (MLD), was introduced recently by Zhang et al. (2020) (see also Hsieh et al. (2018) for an earlier incarnation). Several follow-up works attempted to bound convergence rates for discretizations of MLD, e.g. Ahn and Chewi (2021); Jiang (2021); Li et al. (2022). Unfortunately, many of these analyses imposed rather strong conditions on ϕ beyond strong convexity, e.g. a "modified self-concordance" assumption used in Zhang et al. (2020); Jiang (2021); Li et al. (2022) which (to our knowledge) is not known to be satisfied by standard regularizers. Even more problematically, these analyses (as well as an empirical evaluation by Jiang (2021)) suggest that without strong relative regularity assumptions between the target density and ϕ, naïve discretizations of MLD inherently do not converge to the target even in the limit. A notable exception is Ahn and Chewi (2021), which circumvented both issues (the modified self-concordance assumption and a biased limit) using a different MLD discretization; however, it is not clear that this discretization is feasible for standard choices of ϕ and $\mathcal{X}$.

An alternative to directly discretizing MLD is to use a filter to control bias, akin to the MALA or Metropolized HMC algorithms which are well-studied in the Euclidean case Besag (1994); Roberts and Tweedie (1996); Bou-Rabee and Hairer (2012); Dwivedi et al. (2019); Chen et al. (2020); Lee et al. (2020). However, here too generalizing existing analyses runs into obstacles: for example, typical analyses of MALA and Metropolized HMC rely on bounding the conductance of random walks via isoperimetric inequalities on the target distribution. Prior isoperimetry bounds appear to be tailored to the ℓ_2 geometry and properties of Gaussians (the basic strongly logconcave distribution in Euclidean settings). Potentially due to this difficulty, to our knowledge no general-purpose extension of MALA or its variants to non-Euclidean norms exists in the literature.[2]

Proximal samplers. In this paper, we overcome these difficulties by following a third strategy for the design of efficient samplers: a proximal approach recently proposed by Lee et al. (2021b). To sample from a density π on $\mathbb{R}^d$ proportional to $\exp(-f)$, the algorithm of Lee et al. (2021b) first extends the space to

1. Self-concordance requires that the second derivative of a function is stable to perturbations which are measured in the induced norm. For notation and definitions used throughout the paper, see Section 2.
2. We mention that in certain geometries induced by structured manifolds (discussed in part in Section 1.3), generalizations of MALA or Metropolized HMC have been previously proposed, e.g. Girolami and Calderhead (2011); Barp (2020). These works are motivated by related, but different, settings to the ones considered in this work (we mainly study norm regularity, akin to first-order convex optimization), and their focus is not on establishing non-asymptotic mixing time bounds.

$\mathbb{R}^d \times \mathbb{R}^d$, and defines a joint density $\hat{\pi}$ such that, for some parameter $\eta > 0$,

$$\mathrm{d}\hat{\pi}(z) \propto \exp\left(-f(x) - \frac{1}{2\eta}\|x-y\|_2^2\right)\mathrm{d}z \text{ where } z = (x,y) \in \mathbb{R}^d \times \mathbb{R}^d. \tag{1}$$

It is straightforward to see that for any η, the x-marginal of $\hat{\pi}$ is the original distribution π, and further Lee et al. (2021b) shows that alternating sampling from the conditional distributions of $\hat{\pi}$, i.e. $\hat{\pi}(x \mid y)$ or $\hat{\pi}(y \mid x)$, mixes rapidly. We give an extended discussion on recent activity on designing and harnessing proximal samplers building upon Lee et al. (2021b) in Section 1.3, but mention that instantiations of the framework have resulted in state-of-the-art runtimes for many structured density families Chen et al. (2022); Liang and Chen (2022); Gopi et al. (2022). Motivated by the success of proximal methods in the Euclidean setting, one goal of our work is to extend this technique to non-Euclidean geometries.

Our approach. Our main insight is that a generalization of the strategy in Lee et al. (2021b) induces a well-studied object in probability theory called the *log-Laplace transform* (LLT). Letting $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a convex function in the dual space $y \in \mathbb{R}^d$, our generalization of (1) defines the joint density

$$\begin{gathered}\mathrm{d}\hat{\pi}(z) \propto \exp\left(-f(x) + (\langle x, y\rangle - \varphi(y) - \psi(x))\right)\mathrm{d}z, \\ \text{where } \psi(x) := \log\left(\int \exp\left(\langle x, y\rangle - \varphi(y)\right)\mathrm{d}y\right).\end{gathered} \tag{2}$$

The function ψ is called the LLT of φ, and it has an interpretation as a normalizing constant for induced densities $\mathcal{D}_x^\varphi$ on the dual space proportional to $\exp(\langle x, \cdot\rangle - \varphi)$. Indeed, $\mathcal{D}_x^\varphi$ is defined exactly so the x-marginal of $\hat{\pi}$ is $\pi \propto \exp(-f)$. When $\eta = 1$ and φ, ψ are quadratics, this is exactly (1); we discuss the case of general η in Section 1.2. Moreover, the LLT is a well-studied mathematical object: it arises in probability theory as a *cumulant-generating function*, i.e. derivatives of the LLT yield cumulants of the induced distributions $\mathcal{D}_x^\varphi$, just as derivatives of the MGF yield moments.

The LLT famously appeared in Cramér's theorem on large deviations Cramér (1938), and its cumulant-generating properties have yielded fundamental concentration results in convex geometry Klartag (2006); Eldan and Klartag (2011); Klartag and Milman (2012). More recently, algorithmically-motivated properties of the LLT have been studied in settings such as optimization Bubeck and Eldan (2019), where it was used to define an optimal self-concordant barrier, as well as connections to localization schemes for sampling from discrete distributions Chen and Eldan (2022). We continue this investigation by demonstrating new mathematical properties of the LLT with an algorithmic flavor, and showcasing uses of the LLT as a tool for continuous logconcave sampling. Armed with a deeper understanding of the LLT, we overcome several of the aforementioned barriers to non-Euclidean sampler design and develop a generalized proximal sampler. We further apply it to obtain new complexity results for non-Euclidean differentially private convex optimization, building upon a connection discovered by Gopi et al. (2022, 2023). We are optimistic that the LLT will find additional uses in sampler design (potentially beyond proximal sampling, building upon the new properties we prove), and suggest avenues of future exploration to the community in Section 5.

1.1. Our results

In this section, we overview our results, which separate cleanly into three categories.

Algorithmic aspects of the LLT. It is well-known that the derivatives of the LLT at a point $x \in \mathbb{R}^d$ are *cumulants* of the induced density on $y \in \mathbb{R}^d$:

$$\mathrm{d}\mathcal{D}_x^\varphi(y) \propto \exp\left(\langle x, y\rangle - \varphi(y)\right)\mathrm{d}y.$$

For example, $\nabla\psi(x) = \mathbb{E}_{y\sim\mathcal{D}_x^\varphi}[y]$, and $\nabla^2\psi(x)$ is the covariance of $\mathcal{D}_x^\varphi$. Further, it was shown in Bubeck and Eldan (2019) that if ψ is the LLT of a convex function φ, then ψ is convex and self-concordant. Building upon these facts, in Section 3, we prove the following new properties of the LLT.

- *Strong convexity-smoothness duality.* Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. We prove that if $\varphi : \mathbb{R}^d \to \mathbb{R}$ is L-smooth in the dual norm $\|\cdot\|_*$, its LLT $\psi : \mathbb{R}^d \to \mathbb{R}$ is $\frac{1}{L}$-strongly convex in $\|\cdot\|$.[3] This fact parallels a similar, well-known form of strong convexity-smoothness duality for Fenchel conjugates Shalev-Shwartz (2007); Kakade et al. (2009). Our proof does not require φ to be convex. We further show that the converse holds as well: a $\frac{1}{L}$-strongly convex φ has a L-smooth LLT.

- *Isoperimetry in the Hessian norm.* We prove a one-dimensional isoperimetric inequality for densities of the form $\exp(-\phi)$, where $\phi : \mathbb{R} \to \mathbb{R}$ is self-concordant and convex. By appealing to (a strong variant of) the localization lemma of Lovász and Simonovits (1993), this proves that measures which are strongly logconcave with respect to convex and self-concordant $\phi : \mathbb{R}^d \to \mathbb{R}$ satisfy a similar isoperimetric inequality in the Riemannian geometry induced by $\nabla^2\phi$. Importantly, due to self-concordance of the LLT, this applies to strongly logconcave measures in an LLT.

- *Overlap of induced distributions $\mathcal{D}_x^\varphi$.* We provide a KL divergence bound on the distributions $\mathcal{D}_x^\varphi$ and $\mathcal{D}_{x'}^\varphi$ for x and x' which are close in the Riemannian distance induced by ψ. Combined with our isoperimetric inequality and a classical argument of Dyer et al. (1991), this proves a lower bound on the conductance of an alternating sampler for densities of the form (2).

These new properties of the LLT suggest that it may find uses in designing samplers under non-Euclidean geometries beyond those explored in Section 4 and Appendix A. For example, the LLT of a smooth function is strongly convex and self-concordant, which are the properties required by the MLD discretization scheme of Ahn and Chewi (2021). In optimization, regularizers ϕ for mirror descent typically only require strong convexity (and not self-concordance). However, controlling the evolution of the geometry induced by $\nabla^2\phi$ is critical for discretizing MLD schemes, so imposing self-concordance (as opposed to more non-standard regularity such as the modified self-concordance of Zhang et al. (2020); Jiang (2021); Li et al. (2022)) may be viewed as a minimal assumption. Problematically, standard strongly convex regularizers for mirror descent such as entropy or ℓ_p^2 *are not* self-concordant; LLTs are a way of bridging this gap. Moreover, our new isoperimetric inequality and conductance bounds suggest that LLTs may find use in Metropolized sampling schemes, paving the way for non-Euclidean generalizations of MALA and its variants.

Our new duality result is a generic way of taking a strongly convex regularizer and transform it, via the *Fenchel transform* and the *log-Laplace transform*, to another regularizer which is strongly convex in the same norm and self-concordant. The first transform yields smoothness in the dual Kakade et al. (2009), and the second undoes this change. We will later give an end-to-end application in improving the oracle complexity of private stochastic convex optimization in the ℓ_p geometry, using the LLT of the ℓ_q^2 regularizer.

Non-Euclidean proximal sampling. In Section 4, we build upon these aforementioned tools to analyze the mixing time of an alternating scheme for sampling densities π on convex, compact $\mathcal{X} \subset \mathbb{R}^d$ equipped with a norm $\|\cdot\|_{\mathcal{X}}$, where $\pi \propto \exp(-F(x) - \eta\mu\psi(x))\,\mathbf{1}_{\mathcal{X}}(x)$. Here, $F : \mathcal{X} \to \mathbb{R}$ is convex, $\eta, \mu > 0$ are tunable, and ψ is the LLT of η-smooth $\varphi : \mathbb{R}^d \to \mathbb{R}$ in $\|\cdot\|_{\mathcal{X}^*}$. We prove in Theorem 16 that alternately sampling from conditional distributions of the extended density on $z = (x, y) \in \mathcal{X} \times \mathbb{R}^d$ proportional to

$$\exp\left(-F(x) - \eta\mu\psi(x) + (\langle x, y\rangle - \varphi(y) - \psi(x))\right)\mathbf{1}_{\mathcal{X}}(x) \tag{3}$$

has stationary distribution π, and converges in $\approx \frac{1}{\eta\mu}$ iterations for a warm start. Our rate depends polylogarithmically on both the warmness β of the point it is initialized with, and the inverse of the total variation δ. The form of (3) is the same as (2), but we impose that f is $\eta\mu$-relatively strongly convex in ψ.

We first compare this result to the Euclidean proximal sampler of Lee et al. (2021b), who proved a similar result for alternating sampling densities of the form (1). The main result of Lee et al. (2021b) shows

3. The constant factor 1 here is optimal, as demonstrated by quadratics.

that if f is μ-strongly convex in the ℓ_2 norm, then alternating sampling from the marginals of (1) converges in $\approx \frac{1}{\eta\mu}$ iterations, also with polylogarithmic dependence on the target total variation error. Our result can be viewed as an extension of this result; instead of requiring μ-strong convexity in the ℓ_2 norm (which is equivalent to relative strong convexity with respect to the function $x \to \frac{1}{2}\|x\|_2^2$), we require μ-relative strong convexity in the function $\eta\psi$. In light of our duality result, $\eta\psi$ is 1-strongly convex in $\|\cdot\|_{\mathcal{X}}$, so it is the natural "unit" for measuring strong convexity.

We remark that the parameters η and μ play different roles: μ governs the strong logconcavity of the stationary distribution, and η controls the strong logconcavity of the x-conditional distribution of (3), which is tuned to govern the convergence rate of sampling from the conditional distribution. In particular, we further show that when F is G-Lipschitz in $\|\cdot\|_{\mathcal{X}}$, then as long as $\eta \lesssim G^{-2}$, the conditional sampling required by (3) can be performed in constant calls to a value oracle to F in expectation. This result holds even when F is a distribution over G-Lipschitz functions, and we only have sample access to this distribution. This extends a similar implementation of the marginal sampler required by Lee et al. (2021b) for log-Lipschitz densities in the ℓ_2 norm, given by Gopi et al. (2022). The remaining complexity of the marginal sampling depends on the structure of the chosen φ and $\mathcal{X}$, but is independent of F; we give a discussion of this aspect of our sampler in Appendix A.3 and Section 5.

One shortcoming of Theorem 25's rate is that it depends polylogarithmically on the warmness parameter. In contrast, the rate of Lee et al. (2021b) depends *doubly logarithmically* on the warmness, which is important because in many sampling applications, standard starting distributions yield warmness exponential in problem parameters, e.g. the dimension d. We refer to Section 1.1 of Lee et al. (2021a) on warmness assumptions under ℓ_2 geometry, which have created a $\approx \sqrt{d}$-sized gap on mixing time bounds for MALA, with and without a polynomially-bounded warm start Chewi et al. (2021); Lee et al. (2020). An interesting future direction is to close this gap in warmness assumptions for our sampler in Section 4, analogously to the result of Lee et al. (2021b). Notably, there has been an ongoing exploration of new proof techniques for the convergence of proximal samplers by the community Chen et al. (2022); Chen and Eldan (2022), and we are optimistic similar advancements can be made in non-Euclidean settings, discussed in Section 1.3.

Zeroth-order private convex optimization. As our main application, we use our sampler to give new algorithms for the problem of zeroth-order private convex optimization, where one is given access to a zeroth-order function value evaluation oracle. The zeroth-order access model is appealing in practical settings where gradients are expensive to compute, or the domain is not naturally differentiable. For example, in SCO problems where samples correspond to humans in a population, it may be straightforward to query the value of a sample (e.g. audit a human), but asking for gradients requires querying samples in a continuous manner, which may not be feasible.

Specifically, we achieve this in Appendix A by designing LLTs based on the smoothness of $\varphi_q(x) = \frac{p-1}{2}\|x\|_q^2$ in ℓ_q, where $\frac{1}{p} + \frac{1}{q} = 1$ and $p \in [1, 2]$, $q \geq 2$. We show that the additive range[4] of $\psi_{\eta,p}$,[5] the LLT of $\eta\varphi_q$ for $\eta \lesssim \frac{1}{d}$,[6] is bounded by $O(\frac{1}{(p-1)\eta})$ over the unit ℓ_p ball. This makes $\eta\psi_{\eta,p}$ competitive with the canonical choice of regularizer in ℓ_p norms for optimization, $r_p(x) := \frac{1}{2(p-1)}\|x\|_p^2$, which has the same additive range and strong convexity as $\eta\psi_{\eta,p}$ (up to constants). We further build efficient value oracles and samplers for induced densities for $\psi_{\eta,p}$ in Appendix A.3.

A critical difference between $\eta\psi$ and r_p, however, is that regularizing by a multiple of $\eta\psi$ admits efficient samplers via Section 4; to our knowledge no similar technique is known for r_p. This difference is important in the setting of *differentially private convex optimization*: see Problem 22 for a formal statement. Recently, Gopi et al. (2023) showed that to privately minimize population or empirical risk for a distribution over convex functions which are Lipschitz in a (possibly non-Euclidean) norm $\|\cdot\|_{\mathcal{X}}$, it suffices to sample from a

4. We use the additive range of a function r to mean $\max_x r(x) - \min_x r(x)$ evaluated over the domain of r.
5. We use slightly different notation than in Appendix A for convenience of exposition here.
6. This restriction is discussed further in Section 1.2, but does not bottleneck our privacy applications.

regularized density $\propto \exp(-k(F_{\text{erm}} + \mu r))$. Here, $F_{\text{erm}} = \frac{1}{n}\sum_{i\in[n]} f_i$ is the empirical risk over n samples $\{f_i\}_{i\in[n]}$, k, μ are tunable parameters, and r is a 1-strongly convex regularizer in $\|\cdot\|_{\mathcal{X}}$.

Our results show a demonstrable algorithmic advantage of using $\eta\psi_{\eta,p}$ in ℓ_p geometries, as opposed to r_p. In Theorem 25, we give algorithms for private convex optimization matching the state-of-the-art excess risk bounds in Gopi et al. (2023) (who used r_p as their regularizer). Under a warm start, our new algorithms improve the zeroth-order oracle complexities under ℓ_p regularity in dimension d by $\text{poly}(d)$ factors compared to Gopi et al. (2023), i.e. the number of queries to $\{f_i\}_{i\in[n]}$ used. We show these new complexities extend straightforwardly to improve private convex optimization over matrix spaces satisfying Schatten-p norm regularity. Our results match (up to logarithmic factors) the value oracle complexities in the ℓ_2 setting obtained by Gopi et al. (2022), for all ℓ_p norms where $p \in [1, 2]$. In Appendix D, we extend lower bounds for stochastic optimization from Duchi et al. (2015); Gopi et al. (2022) to the ℓ_p setting to show the value oracle complexities of Theorems 16 and 25 are near-optimal, given a polynomially warm start.

1.2. Our techniques

Analogously to Section 1.1, in this section we split our discussion of our techniques into three parts.

Algorithmic aspects of the LLT. We first discuss our strong convexity-smoothness duality result. From a convex geometry perspective, smoothness of φ (with LLT ψ) ensures that the induced distributions $\propto \exp(\langle x, \cdot\rangle - \varphi)$ are heavy-tailed (because their log-densities cannot grow quickly), which means their variances are "large." We also know that $\nabla^2\psi$ is the covariance matrix of the induced distribution which means that $\nabla^2\psi$ should be lower-bounded. Interestingly, we formalize this intuition by using a perspective inspired by differential privacy: we show that small shifts of the induced distributions are difficult to distinguish, by smoothness of φ. If the variance was small, the shifts would be easy to distinguish, proving the result. Our converse proof is simpler, and uses the Brascamp-Lieb inequality Brascamp and Lieb (1976).

To prove our isoperimetric inequality, we draw inspiration from a similar bound shown in Lemma 35 of Lee and Vempala (2018), but for a family of convex functions ϕ satisfying a strange condition that ϕ'' was convex (which fortunately includes the $\log$ barrier function). Noticing that $-\log$ is self-concordant, we extend the Lee and Vempala (2018) result to hold for all self-concordant functions. Further we show by a direct calculation that the KL divergence between the induced distributions of two nearby points x and x' is essentially the LLT ψ at one of the points, up to a linear term. This lets us use stability of the Hessian of self-concordance functions to demonstrate stability of nearby induced distributions, a key ingredient in proving conductance bounds by the machinery of Dyer et al. (1991).

Non-Euclidean proximal sampling. Given the results of Section 3, establishing our main proximal sampling result Theorem 16 is fairly routine. Our algorithm consists of an "outer loop" which is stated and analyzed in Section 4, and an "inner loop" for sampling from the x-conditional distribution of (3) (Appendix C). Our outer loop analysis is directly based on the mixing time-to-conductance reduction of Lovász and Simonovits (1993) and the technique of Dyer et al. (1991) to lower bound conductance, using facts from Section 3. Our inner loop handling functions F in (3) which are Lipschitz (or distributions over Lipschitz functions) is a small modification of a similar result in Gopi et al. (2022). The only LLT property needed there is strong convexity: this implies a rejection sampler terminates quickly via concentration of Lipschitz functions under strongly logconcave distributions (in any norm) Ledoux (1999); Bobkov and Ledoux (2000).

We note there is a design decision on how to define "scaling up the LLT by $\frac{1}{\eta}$," unlike in the case of (1) where using $\mathcal{N}(x, \eta^{-1}\mathbf{I}_d)$ is natural. Given r, a 1-strongly convex function in $\|\cdot\|_{\mathcal{X}}$, and letting r^* be its (smooth) Fenchel conjugate, two ways of defining a scaled up induced distribution are to choose densities

$$\propto \exp\left(\langle x, y\rangle - \eta r^*(y) - \psi(x)\right), \tag{4}$$

or

$$\propto \exp\left(\frac{1}{\eta}\left(\langle x, y\rangle - r^*(y) - \psi(x)\right)\right). \tag{5}$$

The choice (4) results in ψ which is $\Omega(\eta^{-1})$-strongly convex, suitable for our applications. However, the second results in $\eta^{-1}\psi$ which is also $\Omega(\eta^{-1})$-strongly convex. Interestingly, when $r = r^* = \frac{1}{2}\|\cdot\|_2^2$, (1) agrees with (5) but not (4). Unfortunately, the ψ from (5) is not self-concordant: its Hessian scales with η^{-1} and its third derivative with η^{-2}. Our choice to use (4) has further implications, elaborated on next.

Zeroth-order private convex optimization. The frameworks of Gopi et al. (2022, 2023) show that to use our proximal sampler for ℓ_p private convex optimization, it suffices to design an LLT with small additive range. Perhaps surprisingly, we exploit the *non-scale invariance* of LLT for this task: the LLT of $\eta\varphi$ does not behave like η^{-1} times the LLT of φ.[7] To see why this helps, consider the case when $\varphi = \frac{1}{2}\|\cdot\|_\infty^2$. Although one would hope $\psi(x)$ has additive range comparable to $\frac{1}{2}\|x\|_1^2$, the Fenchel conjugate of $\frac{1}{2}\|x\|_\infty^2$, it is not hard to show that $\psi(e_1) - \psi(0) = \Omega(\sqrt{d})$; we give a proof in Appendix E. This shows the additive range of ψ on the ℓ_1 ball is larger than $\frac{1}{2}\|\cdot\|_1^2$ by $\text{poly}(d)$ factors.

We show the non-scale invariance of (4) actually helps improve additive ranges. Letting ψ_η denote the LLT of $\eta\|x\|_q^2$, we show the additive range of $\eta\psi_\eta$ (a ≈ 1-strongly convex function) is $\approx \max(\eta, 1, \sqrt{d\eta})$. For sufficiently small η, this implies $\eta\psi_\eta$ is much smaller than ψ; graciously, our applications require $\eta \lesssim \frac{1}{d^2}$. We find it potentially useful to explore how generic this non-scale invariance of the LLT is.

1.3. Prior work

Non-Euclidean sampling. A recurring issue that arises in bounding the convergence rate of non-Euclidean samplers is that naïve discretizations can result in significant error. As a result, most prior works either require strong assumptions or oracles for accurate discretization or adopt more sophisticated discretization methods that are difficult to analyze. For example, earlier in the introduction this was discussed for discretizations of MLD Zhang et al. (2020); Jiang (2021); Ahn and Chewi (2021); Li et al. (2022). Part of the intrinsic difficulty of bounding discretized MLD lies in third-order error terms emerging from non-Euclidean geometries, which are hard to control under standard assumptions.

Under structured settings different than, but related to, those in this paper, an interesting alternative sampling strategy is discretizing Riemannian Langevin or Hamiltonian dynamics. For example, Gatmiry and Vempala (2022) studied the Riemmanian Langevin dynamics assuming access to an oracle to sample from Brownian motion on a manifold, whose complexity heavily depends on the manifold. Further, the convergence rate of Riemannian Hamiltonian Monte Carlo (RHMC) in polytopes was studied in Lee and Vempala (2018), and a discretized version was analyzed in Kook et al. (2022); the results apply to a limited family of distributions, and the convergence rate is fairly large. For RHMC to converge to the correct target distribution, sophisticated discretization methods such as Implicit Midpoint Method are necessary. Though efficient in practice, these methods are challenging to analyze theoretically.

Proximal sampling. A long line of works has used proximal methods in sampling (inspired by optimization). Several considered proximal Langevin algorithms Pereyra (2016); Brosse et al. (2017); Bernton (2018); Wibisono (2019), which combine proximal methods and discretizations of Langevin dynamics. Further, Mou et al. (2022) proposed a sampler based on a proximal sampling oracle. However, these algorithms required either stringent assumptions or a large mixing time. Recently, Lee et al. (2021b) proposed a proximal sampler overcoming many assumptions and efficiency issues in prior methods. Several works have focused on generalizing Lee et al. (2021b) and applying it in different settings: Chen et al. (2022) proved convergence results using weaker assumptions than strong logconcavity. The framework has been used

7. On the other hand, the Fenchel conjugate of $\eta\varphi$ is η times the Fenchel conjugate of $\varphi(\frac{\cdot}{\eta})$.

to obtain state-of-the-art samplers for various structured families, e.g. smooth, composite, and finite-sum densities Lee et al. (2021b) as well as non-smooth densities Gopi et al. (2022); Liang and Chen (2022).

Log-Laplace transform. The LLT is a powerful tool that emerges frequently in probability theory and convex geometry. Notably, Bubeck and Eldan (2019); Chewi (2021) showed that the Legendre-Fenchel dual of LLT of the uniform measure on a convex body in $\mathbb{R}^n$ is an n-self-concordant barrier, giving the first universal barrier for convex bodies with optimal self-concordance parameter. In Chen and Eldan (2022), the LLT serves as one of the key ingredients of entropy conservation in localization schemes for sampling. In addition, the LLT shows up in the solution to the entropic optimal transport problem, where a KL divergence is added to regularize the optimal transport objective Chewi and Pooladian (2022).

Private convex optimization. Differentially private convex optimization is one of the most extensively studied problems in the privacy literature and captures an increasing number of critical applications in various domains, including machine learning, statistics, and data analysis. There is a rich body of works on this topic Chaudhuri and Monteleoni (2008); Chaudhuri et al. (2011); Kifer et al. (2012); Bassily et al. (2014); Wang et al. (2017); Bassily et al. (2019); Feldman et al. (2020), which have mainly focused on the Euclidean geometry, e.g. assuming the ℓ_2 diameter of the domain and ℓ_2 norms of gradients are bounded. Motivated by applications not captured by these assumptions, there has been growing interest in studying differentially private convex optimization in non-Euclidean geometries, as seen in Talwar et al. (2015); Asi et al. (2021); Bassily et al. (2021); Han et al. (2022); Gopi et al. (2023). Of particular relevance, Gopi et al. (2023) develops an exponential mechanism based method attaining state-of-the-art excess risk bounds for ℓ_p and Schatten-p norms, which are matched by our algorithms in Appendix A.

2. Preliminaries

General notation. In Section 1 only, $\widetilde{O}$, $\approx$, and $\lesssim$ hide logarithmic factors in problem parameters for expositional convenience. For $n \in \mathbb{N}$, $[n]$ refers to the naturals $1 \le i \le n$. We use $\mathcal{X}$ to denote a compact convex subset of $\mathbb{R}^d$. For all $p \ge 1$ including $p = \infty$, we let $\|\cdot\|_p$ applied to a vector argument denote the ℓ_p norm. We denote matrices in boldface and when $\|\cdot\|_p$ is applied to a matrix argument it denotes the corresponding Schatten-p norm (ℓ_p norm of the singular values). For any $\mathcal{X} \subset \mathbb{R}^d$ we let its indicator function (i.e. the function which is 1 on $\mathcal{X}$ and 0 otherwise) be denoted $\mathbf{1}_{\mathcal{X}}$. We will be concerned with optimizing functions $f : \mathcal{X} \to \mathbb{R}$, and $\|\cdot\|_{\mathcal{X}}$ refers to a norm on $\mathcal{X}$. We let $\mathcal{X}^*$ be the dual space to $\mathcal{X}$, and equip it with the dual norm $\|y\|_{\mathcal{X}^*} := \sup_{\|x\|_{\mathcal{X}}=1} x^\top y$. We let $\mathcal{N}(\mu, \mathbf{\Sigma})$ be the Gaussian density of given mean and covariance. For a positive definite matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, we denote the induced norm by $\|v\|_{\mathbf{M}} := \sqrt{v^\top \mathbf{M} v}$. When making asymptotic statements we will typically assume the dimension d is at least a sufficently large constant, else we can pad and affect statements by at most constant factors.

Optimization. We say f is G-Lipschitz in $\|\cdot\|_{\mathcal{X}}$ if for all $x, x' \in \mathcal{X}$, $|f(x) - f(x')| \le G\|x - x'\|_{\mathcal{X}}$. If f is differentiable, we say it is L-smooth in $\|\cdot\|_{\mathcal{X}}$ if for all $x, x' \in \mathcal{X}$, $\|\nabla f(x) - \nabla f(x')\|_{\mathcal{X}^*} \le L\|x - x'\|_{\mathcal{X}}$; this implies $f(x') \le f(x) + \langle \nabla f(x), x' - x\rangle + \frac{L}{2}\|x - x'\|_{\mathcal{X}}^2$. We say f is m-strongly convex in $\|\cdot\|_{\mathcal{X}}$ if for all $x, x' \in \mathcal{X}$, $t \in [0,1]$, $f(tx + (1-t)y) \le tf(x) + (1-t)f(y) - \frac{mt(1-t)}{2}\|x - x'\|_{\mathcal{X}}^2$. We say f is m-relatively strongly convex in ϕ if $f - m\phi$ is convex. For k-times differentiable f, $\nabla^k f(x)[v_1, v_2, \ldots, v_k]$ denotes the corresponding k^{th} order directional derivative at f. If f is twice-differentiable and m-strongly convex in $\|\cdot\|_{\mathcal{X}}$, $\nabla^2 f(x)[v, v] \ge m\|v\|_{\mathcal{X}}^2$ for all $x \in \mathcal{X}$, $v \in \mathbb{R}^d$. We say convex $\phi : \mathbb{R}^d \to \mathbb{R}$ is self-concordant if it satisfies for all $x, h \in \mathbb{R}^d$, $\left|\nabla^3\phi(x)[h,h,h]\right| \le 2\left(\nabla^2\phi(x)[h,h]\right)^{\frac{3}{2}}$. A key implication of self-concordance is Hessian stability under small distances: see Lemma 2.

Probability. For a density π supported on $\mathcal{X}$, we let $\pi(S) := \Pr_{x\sim\pi}[x \in S]$. For two densities μ, π, their total variation distance is $\|\mu - \pi\|_{\text{TV}} := \frac{1}{2}\int |\mu(x) - \pi(x)| \text{d}x$ and (when the Radon-Nikodym derivative exists) their KL divergence is $D_{\mathrm{KL}}(\mu\|\pi) := \int \mu(x)\log\frac{\mu(x)}{\pi(x)}\mathrm{d}x$. For $1 < \alpha < \infty$, we also define the α-Rényi divergence between densities μ, π by $D_\alpha(\mu\|\pi) := \frac{1}{\alpha-1}\log(\int(\frac{\mu(x)}{\pi(x)})^\alpha \pi(x)\mathrm{d}x)$. We say density π is logconcave (respectively, m-strongly logconcave in $\|\cdot\|_{\mathcal{X}}$) if $-\log\pi$ is convex (respectively, m-strongly convex in $\|\cdot\|_{\mathcal{X}}$). We say π is m-relatively strongly logconcave in ϕ if $-\log\pi$ is m-relatively strongly convex in ϕ. We say a density π_0 is β-warm with respect to a density π if for all x, $\frac{\mathrm{d}\pi_0(x)}{\mathrm{d}\pi(x)} \le \beta$.

Log-Laplace transform. We define the log-Laplace transform (LLT) of $\varphi : \mathbb{R}^d \to \mathbb{R}$ by $\psi(x) := \log\left(\int \exp\left(\langle x, y\rangle - \varphi(y)\right)\mathrm{d}y\right)$. When φ, ψ are clear from context, we define the density

$$\mathcal{D}_x^\varphi(y) = \exp\left(\langle x, y\rangle - \varphi(y) - \psi(x)\right). \tag{6}$$

Note that the normalization constant is exactly given by $\exp(-\psi(x))$ and hence $\mathcal{D}_x^\varphi$ is indeed a valid density. We use $\propto$ to indicate proportionality, e.g. if μ is a density and we write $\mu \propto \exp(-f)$, we mean $\mu(x) = \frac{\exp(-f)}{Z}$ where $Z := \int \exp(-f(x))\mathrm{d}x$ and the integration is over the support of μ.

Riemannian geometry. In Sections 3 and 4 we will use geometry induced by the Hessian of a self-concordant, convex function $\phi : \mathbb{R}^d \to \mathbb{R}$. We summarize the important points here, and defer a extended treatment to Nesterov and Todd (2002). When ϕ is clear from context, we denote $\|h\|_x := \|h\|_{\nabla^2\phi(x)}$. Throughout this discussion let $M \subseteq \mathbb{R}^d$ be a Riemannian manifold equipped with the local metric $\|\cdot\|_x$. The induced Riemannian distance of a curve $c : [0,1] \to M$ is defined as $L_\phi(c) := \int_0^1 \left\|\frac{\mathrm{d}}{\mathrm{d}t}c(t)\right\|_{c(t)}\mathrm{d}t$, where $\frac{\mathrm{d}}{\mathrm{d}t}c(t)$ is the velocity element of the curve in the tangent space at $c(t)$. For $x, y \in M$, $d_\phi(x,y)$ is the infimum of the length $L_\phi(c)$ over all curves with $c(0) = x$ and $c(1) = y$. We use the following properties of the Riemannian geometry over $M = \mathbb{R}^d$ induced by self-concordant, convex functions.

Lemma 1 (Nesterov and Todd (2002), Lemma 3.1) *Suppose $\phi : \mathbb{R}^d \to \mathbb{R}$ is convex and self-concordant. For $x, y \in \mathbb{R}^d$, if $d_\phi(x,y) \le \delta - \delta^2 < 1$ for some $\delta \in (0,1)$, then $\|y - x\|_x \le \delta$.*

Lemma 2 (Nemirovski (2004), Section 2.2.1) *Suppose $\phi : \mathbb{R}^d \to \mathbb{R}$ is convex and self-concordant. For any $h, x \in \mathbb{R}^d$ such that $\|h\|_x < 1$, $(1 - \|h\|_x)^2\nabla^2\phi(x) \preceq \nabla^2\phi(x+h) \preceq (1 - \|h\|_x)^{-2}\nabla^2\phi(x)$.*

3. Properties of the LLT

We collect facts about the log-Laplace transform used to develop our sampling scheme in Section 4, deferring most proofs to Appendix B due to space constraints. We use the first three derivatives of ψ.

Lemma 3 (LLT derivatives) *For any $x, h \in \mathbb{R}^d$, we have $\nabla\psi(x) = \mu(\mathcal{D}_x^\varphi) := \mathbb{E}_{y\sim\mathcal{D}_x^\varphi}[y]$, $\nabla^2\psi(x) = \mathrm{Cov}(\mathcal{D}_x^\varphi) := \mathbb{E}_{y\sim\mathcal{D}_x^\varphi}\left[(y - \mu(\mathcal{D}_x^\varphi))(y - \mu(\mathcal{D}_x^\varphi))^\top\right]$, $\nabla^3\psi(x)[h,h,h] = \mathbb{E}_{y\sim\mathcal{D}_x^\varphi}\left[\langle y - \mu(\mathcal{D}_x^\varphi), h\rangle^3\right]$.*

By a fact on one-dimensional distributions in Bubeck and Eldan (2019), this implies the following.

Lemma 4 (Self-concordance) *If ψ is the LLT of a convex function, it is self-concordant.*

Next, we prove that a form of strong convexity-smoothness duality holds with respect to φ and ψ, analogous to the type of duality satisfied by Fenchel conjugates Kakade et al. (2009).

Lemma 5 (Strong convexity-smoothness duality) *If $\varphi : \mathbb{R}^d \to \mathbb{R}$ is L-smooth with respect to $\|\cdot\|_*$, then $\psi : \mathbb{R}^d \to \mathbb{R}$ is $\frac{1}{L}$-strongly convex with respect to $\|\cdot\|$.*

Proof By definition of strong convexity it suffices to prove for any $x, v \in \mathbb{R}^d$, $v^\top \nabla^2 \psi(x) v \geq \frac{1}{L} \|v\|^2$. Without loss of generality, by scale invariance we can assume $\|v\| = 1$. Let $Y = \langle y, v\rangle$, where $y \sim \mathcal{D}_x^\varphi$. By Lemma 3, $\nabla^2\psi(x) = \mathrm{Cov}(\mathcal{D}_x^\varphi)$, so it suffices to prove that

$$\mathrm{Var}(Y) = \mathbb{E}_{y\sim\mathcal{D}_x^\varphi}\left[\langle y - \mu(\mathcal{D}_x^\varphi), v\rangle^2\right] \geq \frac{1}{L}.$$

Letting $\mathbf{M} := \mathbb{E}_{y\sim\mathcal{D}_x^\varphi}\nabla^2\varphi(y)$, we first observe

$$\frac{L}{2} v^\top \mathbf{M}^{-1} v = \max_{u\in\mathbb{R}^d} \langle u, v\rangle - \frac{1}{2L} u^\top \mathbf{M} u \geq \max_{u\in\mathbb{R}^d} \langle u, v\rangle - \frac{1}{2}\|u\|_*^2 = \frac{1}{2}\|v\|^2.$$

In the only inequality, we used that $u^\top \mathbf{M} u = \mathbb{E}_{y\sim\mathcal{D}_x^\varphi} u^\top \nabla^2\varphi(y) u \leq L\|u\|_*^2$ by smoothness of φ, and the last equality follows by optimizing over $\|u\|_*$. This shows $v^\top \mathbf{M}^{-1} v \geq \frac{1}{L}$. The Cramér-Rao inequality (see Lemma 2, Chewi and Pooladian (2022)) then implies

$$\mathrm{Var}(Y) \geq v^\top \mathbf{M}^{-1} v \geq \frac{1}{L},$$

since the Hessian of $-\log \mathcal{D}_x^\varphi$ at any $x \in \mathbb{R}^d$ is $\nabla^2\varphi$. ■

Lemma 6 (Smoothness-strong convexity duality) *If $\varphi : \mathbb{R}^d \to \mathbb{R}$ is $\frac{1}{L}$-strongly convex with respect to $\|\cdot\|_*$, then $\psi : \mathbb{R}^d \to \mathbb{R}$ is L-smooth with respect to $\|\cdot\|$.*

Proof Let $v, x \in \mathbb{R}^d$ and assume $\|v\| = 1$. As in Lemma 5, defining $Y = \langle y, v\rangle$ for $y \sim \mathcal{D}_x^\varphi$, we have $v^\top \nabla^2\psi(x) v = \mathrm{Var}(Y)$, and want to show $\mathrm{Var}(Y) \leq L$. First note that for any $y \in \mathbb{R}^d$,

$$\frac{1}{2L} v^\top \left(\nabla^2\varphi(y)\right)^{-1} v = \max_{u\in\mathbb{R}^d} \langle u, v\rangle - \frac{L}{2} u^\top \nabla^2\varphi(y) u \leq \max_{u\in\mathbb{R}^d} \langle u, v\rangle - \frac{1}{2}\|u\|_*^2 = \frac{1}{2}\|v\|^2.$$

The first inequality used strong convexity of φ and the last equality optimizes over $\|u\|_*$. This shows $v^\top(\nabla^2\varphi(y))^{-1} v \leq L$ for all y. The Brascamp-Lieb inequality Brascamp and Lieb (1976) then implies $\mathrm{Var}(Y) \leq \mathbb{E}_{y\sim\mathcal{D}_x^\varphi} v^\top \left(\nabla^2\varphi(y)\right)^{-1} v \leq L$, since the Hessian of $-\log \mathcal{D}_x^\varphi$ at any $x \in \mathbb{R}^d$ is $\nabla^2\varphi$. ■

We next state a new 1-d isoperimetric inequality for self-concordant functions, proven in Appendix B.

Lemma 7 *Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is convex and self-concordant. For any $x \in \mathbb{R}$,*

$$\frac{\exp(-\phi(x))}{\sqrt{\phi''(x)}} \geq \frac{1}{12}\min\left\{\int_{-\infty}^{x} \exp(-\phi(t))dt, \int_{x}^{\infty} \exp(-\phi(t))dt\right\}.$$

By a localization argument deferred to Appendix B, we have the following result in high dimensions.

Lemma 8 (Self-concordant isoperimetry) *Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be convex and self-concordant, and let $f : \mathbb{R}^d \to \mathbb{R}$ be m-relatively strongly convex in ϕ. Given any partition S_1, S_2, S_3 of $\mathbb{R}^d$,*

$$\frac{\int_{S_3} \exp\left(-f(x)\right) \mathrm{d}x}{\min\left\{\int_{S_1} \exp(-f(x))\mathrm{d}x, \int_{S_2} \exp(-f(x))\mathrm{d}x\right\}} = \Omega\left(\sqrt{m}\, d_\phi(S_1, S_2)\right),$$

where $d_\phi(S_1, S_2) = \min_{x\in S_1, y\in S_2} d_\phi(x, y)$.

Finally, we provide a bound on the total variation distance of "nearby" induced distributions.

Lemma 9 (TV of induced distributions) *For $x, x' \in \mathbb{R}^d$ with $d_\psi(x, x') \leq \frac{1}{4}$, $\left\|\mathcal{D}_x^\varphi - \mathcal{D}_{x'}^\varphi\right\|_{\mathrm{TV}} \leq \frac{1}{2}$.*

4. Proximal LLT sampler

In this section, we study a sampling problem in the following setting, assumed throughout.

Problem 10 *For $D, G, \eta > 0$, let $\mathcal{X} \subset \mathbb{R}^d$ be compact and convex, with diameter in a norm $\|\cdot\|_{\mathcal{X}}$ at most D. Let $F : \mathcal{X} \to \mathbb{R}$ have the stochastic form $F(x) := \mathbb{E}_{i \sim \mathcal{I}}[f_i(x)]$, for a distribution $\mathcal{I}$ over (a possibly infinite) family of indices i, such that each $f_i : \mathcal{X} \to \mathbb{R}$ is convex and G-Lipschitz in $\|\cdot\|_{\mathcal{X}}$. Finally, let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be convex and η-smooth in the dual norm $\|\cdot\|_{\mathcal{X}^*}$. Given $\mu > 0$, and letting $\psi : \mathbb{R}^d \to \mathbb{R}$ be the LLT of φ, the goal is to sample from the density π satisfying*

$$\mathrm{d}\pi(x) \propto \exp\left(-F(x) - \eta\mu\psi(x)\right) \mathbf{1}_{\mathcal{X}}(x)\mathrm{d}x. \tag{7}$$

Note that by Lemma 5, $\eta\mu\psi$ is μ-strongly convex in $\|\cdot\|_{\mathcal{X}}$. Letting $z = (x, y)$ denote a variable on $\mathcal{X} \times \mathbb{R}^d$, it is convenient for us to define the extended density on the joint space of z:

$$\mathrm{d}\hat{\pi}(z) \propto \exp\left(-F(x) - \eta\mu\psi(x) + \left(\langle x, y\rangle - \psi(x) - \varphi(y)\right)\right) \mathbf{1}_{\mathcal{X}}(x)\mathrm{d}z. \tag{8}$$

Our sampling framework for (7) generalizes an approach pioneered by Lee et al. (2021b), and is stated in the following Algorithm 1. The algorithm simply alternately samples from each marginal of (8). Before stating it, we define the following notation for conditional densities throughout the section:

$$\begin{aligned} &\mathrm{d}\pi_x(y) = \exp\left(\langle x, y\rangle - \psi(x) - \varphi(y)\right)\mathrm{d}y \text{ for all } x \in \mathcal{X}, \\ &\mathrm{d}\pi_y(x) \propto \exp\left(-F(x) - (1 + \eta\mu)\psi(x) + \langle x, y\rangle\right) \mathbf{1}_{\mathcal{X}}(x)\mathrm{d}x \text{ for all } y \in \mathbb{R}^d. \end{aligned} \tag{9}$$

In particular, we observe that $\mathrm{d}\pi_x(y) = \mathrm{d}\hat{\pi}(\cdot \mid x)$ and $\mathrm{d}\pi_y(x) = \mathrm{d}\hat{\pi}(\cdot \mid y)$.

Algorithm 1: `AlternateSample`$(\mathcal{X}, F, \varphi, T, \mu, x_0)$

1 **Input:** $\mathcal{X}, F, \varphi$ in the setting of Problem 10, $T \in \mathbb{N}$, $\mu > 0$, $x_0 \in \mathcal{X}$.
2 **for** $k \in [T]$ **do**
3 Sample $y_k \sim \pi_{x_{k-1}}$.
4 Sample $x_k \sim \pi_{y_k}$.
5 **end**
6 **Return:** x_T

Correctness of Algorithm 1 builds upon the following basic facts.

Lemma 11 *The total x-marginal of $\hat{\pi}$ in (8) is π in (7). Furthermore, the stationary distribution of Algorithm 1 is $\hat{\pi}$, and the induced Markov chains in Algorithm 1 restricted to either $\{x_k\}_{0 \le k \le T}$ (a Markov chain on $\mathcal{X}$) or $\{y_k\}_{k \in [T]}$ (a Markov chain on $\mathbb{R}^d$) are both reversible.*

Proof The first conclusion is a direct calculation, and the remainder is Lemma 1 in Lee et al. (2021b). ■

In Appendix C, we develop Algorithm 2, a rejection sampler for implementing Line 4 of Algorithm 1, based on Gopi et al. (2022). We defer a bound on its performance in Proposition 38 in Appendix C. We now give our analysis of Algorithm 1 via a conductance argument, using tools from Section 3.

Definition 12 *For a reversible Markov chain with stationary distribution π supported on $\mathcal{X}$ and transition distributions $\{\mathcal{T}_x\}_{x \in \mathcal{X}}$, we define the conductance of the Markov chain by $\Phi := \inf_{S \subset \mathcal{X}} \frac{\int_S \mathcal{T}_x(\mathcal{X} \setminus S)\mathrm{d}\pi(x)}{\min\{\pi(S), \pi(\mathcal{X} \setminus S)\}}$.*

We further recall a standard way of lower bounding conductance via isoperimetry.

Lemma 13 (Lee and Vempala (2018), Lemma 13) *In the setting of Definition 12, let $d : \mathcal{X} \times \mathcal{X}$ be a metric on $\mathcal{X}$. Suppose for $x, x' \in \mathcal{X}$ with $d(x, x') \leq \Delta$, $\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\text{TV}} \leq \frac{1}{2}$. Also, suppose for any partition S_1, S_2, S_3 of $\mathbb{R}^d$, π satisfies $\pi(S_3) \geq C_{\text{iso}}(\min_{x \in S_1, y \in S_2} d(x, y)) \min\{\pi(S_1), \pi(S_2)\}$. Then $\Phi = \Omega(\Delta C_{\text{iso}})$.*

A classical result of Lovász and Simonovits (1993) upper bounds mixing time via conductance.

Lemma 14 (Lovász and Simonovits (1993), Corollary 1.5) *In the setting of Definition 12, let π_t be the distribution after t steps of the Markov chain. If π_0 is β-warm with respect to π, $\|\pi_t - \pi\|_{\text{TV}} \leq \sqrt{\beta}(1 - \frac{\Phi^2}{2})^t$.*

Proposition 15 *Assume the input x_0 to Algorithm 1 is drawn from a β-warm distribution with respect to π, $\eta\mu \leq 1$, and $T = \Omega(\frac{1}{\eta\mu} \log \frac{\beta}{\delta})$ for a sufficiently large constant. Then the output of Algorithm 1 has total variation distance to π bounded by δ.*

Proof Following the optimal coupling characterization of total variation, whenever the optimal coupling of $y \sim \mathcal{D}_x^\varphi$ and $y' \sim \mathcal{D}_{x'}^\varphi$ sets $y = y'$ in Line 3 of Algorithm 1, we can couple the resulting distributions in Line 4 as well. This shows that $\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\text{TV}} \leq \|\mathcal{D}_x^\varphi - \mathcal{D}_{x'}^\varphi\|_{\text{TV}}$. By Lemma 4, since φ is convex, ψ is a self-concordant function. Lemma 9 then implies if $d_\psi(x, x') \leq \frac{1}{4}$, $\|\mathcal{T}_x - \mathcal{T}_{x'}\|_{\text{TV}} \leq \left\|\mathcal{D}_x^\varphi - \mathcal{D}_{x'}^\varphi\right\|_{\text{TV}} \leq \frac{1}{2}$.

By Lemma 8, since $F + \eta\mu\psi$ is $\eta\mu$-relatively strongly convex in ψ, π satisfies the isoperimetric inequality such that for any partition S_1, S_2, S_3 of $\mathbb{R}^d$, $\pi(S_3) = \Omega(\sqrt{\eta\mu}) \left(\min_{x \in S_1, y \in S_2} d_\psi(x, y)\right) \min\{\pi(S_1), \pi(S_2)\}$. By Lemma 13, we can then lower bound the conductance by $\Phi = \Omega(\sqrt{\eta\mu})$. Choosing a sufficiently large constant in T, we conclude by Lemma 14 the desired $\|\pi_T - \pi\|_{\text{TV}} \leq \sqrt{\beta} \exp(-\frac{T\Phi^2}{2}) \leq \delta$. ■

Theorem 16 *In the setting of Problem 10, let $\eta\mu \leq 1$ and assume x_0 has a β-warm distribution with respect to π defined in (7). Further for sufficiently large constants suppose $\frac{1}{\eta} = \Omega(G^2 \log \frac{\log \beta}{\delta\eta\mu})$ and $T = \Theta\left(\frac{1}{\eta\mu} \log \frac{\beta}{\delta}\right)$. Algorithm 1 using Algorithm 2 with error parameter $\frac{\delta}{2T}$ to implement Line 4 returns a point with δ total variation distance to π, querying $O(T)$ random f_i in expectation.*

Proof Proposition 15 guarantees that if each call to Line 4 of Algorithm 1 is implemented exactly, we obtain $\frac{\delta}{2}$ total variation to π. Further, the total variation error accumulated over T calls to Algorithm 2 is less than $\frac{\delta}{2}$ by a union bound on Proposition 38. Combining these bounds results in the desired total variation guarantee, and the complexity bound follows from Proposition 38. ■

We note that given sample access to $\exp(-\eta\mu\psi(x))\mathbf{1}_{x \in \mathcal{X}}$, a distribution which only depends on the choice of φ and $\mathcal{X}$ (and not the function F), we obtain $\beta \leq \exp(GD)$ in Theorem 16.

Lemma 17 *The density $\mathrm{d}\nu(x) \propto \exp(-\eta\mu\psi(x))\mathbf{1}_{\mathcal{X}}(x)\mathrm{d}x$ is $\exp(GD)$-warm for π defined in (7).*

5. Conclusion

We believe our work is a significant step towards developing the theory of LLTs and paving the way for their use in designing sampling algorithms. There are a number of important questions left open by our work, which we find interesting and potentially fruitful for the community to explore.

Stronger mixing time bounds. Perhaps the most immediate open question regarding our alternating sampling framework in Section 4 is to obtain a better understanding of its mixing time. As discussed in Section 1.1, Theorem 16's mixing time scales linearly in $\log \beta$, which as demonstrated by Lemma 17 (and related other settings, e.g. MALA Chewi et al. (2021); Lee et al. (2021a)) can result in additional polynomial overhead in problem parameters: for what φ, ψ is this avoidable? Notably, it is avoided for the Euclidean proximal sampler Lee et al. (2021b) by working directly with KL divergence (as opposed to the larger χ^2 distance typically used by proofs using conductance bounds). Different proofs of this $\log \log \beta$ dependency for the Euclidean proximal sampler were then subsequently obtained by Chen et al. (2022); Chen and Eldan (2022). We also mention that $\log \log \beta$ dependences may sometimes follow via average conductance techniques (e.g. Lovász and Kannan (1999)), which may apply to our Markov chain.

Samplers for explicit distributions. Our results Theorem 16 and 25 mainly focused on bounding the query complexity to the function F, or samples f_i from the distribution defining it. The total computational complexity of a practical implementation of Algorithm 1 also includes the cost of sampling from the distribution families π_x (9), γ_y (24) (the latter is used in our rejection sampler subroutine, Algorithm 2). In Appendix A.3, we give a linear-time sampler for π_x and a polynomial-time sampler for γ_y under the ℓ_p geometry, but it is interesting to obtain faster samplers for particular choices of $(\varphi, \mathcal{X})$.

LLT beyond proximal sampling. More generally, we believe it is worthwhile to obtain a better understanding of specific choices of (φ, ψ), e.g. the examples in Appendix A.1, from an algorithmic perspective. LLTs satisfy appealing properties such as self-concordance, strong convexity, and isoperimetry making them well-suited for frameworks beyond Algorithm 1, such as discretized MLD Ahn and Chewi (2021) and Metropolized sampling methods discussed in Section 1. Bounding the complexity of their use in these applications necessitates an improved understanding of specific LLTs.

LLT as a dual object. Finally, a tantalizing open question in the theory of well-conditioned sampling (even in the ℓ_2 setting) is whether acceleration is achievable, i.e. mixing scaling with the square root of the condition number (which is possible in optimization Nesterov (1983)). The duality of Fenchel conjugates appears to play a key role in acceleration Wang and Abernethy (2018); Cohen et al. (2021), so a better understanding of duality may be helpful in the corresponding endeavor for sampling. The LLT is a natural candidate for a dual object, as it arises via joint densities on an extended space (2), and satisfies strong convexity-smoothness duality. Can we demystify this relationship, and use it to obtain faster samplers?

Acknowledgments

We thank Sam Power for helpful comments, suggesting several references, noticing that an improved strong convexity-smoothness duality result follows from combining our smoothness-strong convexity duality result and a tool in Chewi and Pooladian (2022), and letting us include his elegant observation.

References

Alekh Agarwal, Sahand N. Negahban, and Martin J. Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010*, pages 37–45. Curran Associates, Inc., 2010.

Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-langevin algorithm. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, pages 28405–28418, 2021.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory Comput.*, 8(1):121–164, 2012.

Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in ℓ_1 geometry. *arXiv preprint arXiv:2103.01516*, 2021.

Keith Ball, Eric A. Carlen, and Elliott H. Lieb. Sharp uniform convexity and smoothness estimates for trace norms. *Inventiones mathematicae*, 115(1):463–482, 1994.

Alessandro Andrea Barp. *The bracket geometry of statistics*. PhD thesis, Imperial College London, 2020.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pages 464–473. IEEE, 2014.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

Raef Bassily, Cristóbal Guzmán, and Anupama Nandi. Non-euclidean differentially private stochastic convex optimization. *arXiv preprint arXiv:2103.01278*, 2021.

Espen Bernton. Langevin monte carlo and jko splitting. In *Conference on learning theory*, pages 1777–1798. PMLR, 2018.

Julian Besag. Comments on "representations of knowledge in complex systems" by u. grenander and mi miller. *Journal of the Royal Statistical Society, Series B*, 56:591–592, 1994.

Sergey G Bobkov and Michel Ledoux. From brunn-minkowski to brascamp-lieb and to logarithmic sobolev inequalities. *GAFA, Geometric and Functional Analysis*, 10:1028–1052, 2000.

Nawaf Bou-Rabee and Martin Hairer. Nonasymptotic mixing of the mala algorithm. *IMA Journal of Numerical Analysis*, 33(1):80–110, 2012.

Herm Jan Brascamp and Elliott H Lieb. On extensions of the brunn-minkowski and prékopa-leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22(4):366–389, 1976.

Nicolas Brosse, Alain Durmus, Éric Moulines, and Marcelo Pereyra. Sampling from a log-concave distribution with compact support with proximal langevin monte carlo. In *Conference on learning theory*, pages 319–342. PMLR, 2017.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends Mach. Learn.*, 5(1):1–122, 2012.

Sébastien Bubeck and Ronen Eldan. The entropic barrier: Exponential families, log-concave geometry, and self-concordance. *Math. Oper. Res.*, 44(1):264–276, 2019.

Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, 2006.

Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. *Advances in neural information processing systems*, 21, 2008.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2984–3014. PMLR, 2022.

Yuansi Chen. An almost constant lower bound of the isoperimetric coefficient in the kls conjecture. *Geometric and Functional Analysis*, 31(1):34–61, 2021.

Yuansi Chen and Ronen Eldan. Localization schemes: A framework for proving mixing bounds for markov chains (extended abstract). In *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022*, pages 110–122. IEEE, 2022.

Yuansi Chen, Raaz Dwivedi, Martin J. Wainwright, and Bin Yu. Fast mixing of metropolized hamiltonian monte carlo: Benefits of multi-step gradients. *J. Mach. Learn. Res.*, 21:92:1–92:72, 2020.

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*, volume 145 of *LIPIcs*, pages 64:1–64:12. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 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*, volume 75 of *Proceedings of Machine Learning Research*, pages 300–323. PMLR, 2018. URL http://proceedings.mlr.press/v75/cheng18a.html.

Sinho Chewi. The entropic barrier is n-self-concordant. *CoRR*, abs/2112.10947, 2021.

Sinho Chewi. *Log-Concave Sampling*. 2023.

Sinho Chewi and Aram-Alexandre Pooladian. An entropic generalization of Caffarelli's contraction theorem via covariance inequalities. *arXiv e-prints*, art. arXiv:2203.04954, 2022.

Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the metropolis-adjusted langevin algorithm. In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 1260–1300. PMLR, 2021.

Michael B. Cohen, Aaron Sidford, and Kevin Tian. Relative lipschitzness in extragradient methods and a direct recipe for acceleration. In *12th Innovations in Theoretical Computer Science Conference, ITCS 2021*, volume 185 of *LIPIcs*, pages 62:1–62:18, 2021.

Harald Cramér. Sur un nouveau théorème-limite de la théorie des probabilités. *Act. Sci. et Ind.*, 736, 1938.

Arnak S. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pages 678–689, 2017a.

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

Etienne De Klerk and Monique Laurent. Comparison of lasserre's measure-based bounds for polynomial optimization to bounds obtained by simulated annealing. *Mathematics of Operations Research*, 43(4): 1317–1325, 2018.

Jelena Diakonikolas, Maryam Fazel, and Lorenzo Orecchia. Fair packing and covering on a relative scale. *SIAM J. Optim.*, 30(4):3284–3314, 2020.

John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

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

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. *J. Mach. Learn. Res.*, 20:183:1–183:42, 2019.

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.

Ronen Eldan and Bo"az Klartag. Approximately gaussian marginals and the hyperplane conjecture. *Contemporary Math.*, 545:44–68, 2011.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

Khashayar Gatmiry and Santosh S Vempala. Convergence of the riemannian langevin algorithm. *arXiv preprint arXiv:2204.10818*, 2022.

Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society, Series B*, 73(2):123–214, 2011.

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

Sivakanth Gopi, Yin Tat Lee, Daogao Liu, Ruoqi Shen, and Kevin Tian. Private convex optimization in general norms. In *Proceedings of the 2023 ACM-SIAM Symposium on Discrete Algorithms, SODA 2023*. SIAM, 2023.

Yuxuan Han, Zhicong Liang, Zhipeng Liang, Yang Wang, Yuan Yao, and Jiheng Zhang. Private streaming sco in ℓ_p geometry with applications in high dimensional online decision making. In *International Conference on Machine Learning*, pages 8249–8279. PMLR, 2022.

Harish-Chandra. Differential operators on a semisimple lie groups. *American Journal of Mathematics*, 79: 87–120, 1957.

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*, pages 2883–2892, 2018.

C. Itzykson and J.-. Zuber. The planar approximation. ii. *Journal of Mathematical Physics*, 21:411–421, 1980.

Arun Jambulapati, Jerry Li, and Kevin Tian. Robust sub-gaussian principal component analysis and width-independent schatten packing. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020.

Arun Jambulapati, Yin Tat Lee, and Santosh S. Vempala. A slightly improved bound for the KLS constant. *CoRR*, abs/2208.11644, 2022.

He Jia, Aditi Laddha, Yin Tat Lee, and Santosh S. Vempala. Reducing isotropy and volume to KLS: an $o^*(n^3\psi^2)$ volume algorithm. In *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 961–974. ACM, 2021.

Qijia Jiang. Mirror langevin monte carlo: the case under isoperimetry. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, pages 715–725, 2021.

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

Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Applications of strong convexity–strong smoothness duality to learning with matrices. *arXiv e-prints*, abs/0910.0610, 2009.

Ravi Kannan, László Lovász, and Miklós Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete & Computational Geometry*, 13(3):541–559, 1995.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1. JMLR Workshop and Conference Proceedings, 2012.

Bo"az Klartag. On convex perturbations with a bounded isotropic constant. *Geometric and Functional Analysis*, 16(6):1274–1290, 2006.

Bo"az Klartag and Joseph Lehec. Bourgain's slicing problem and kls isoperimetry up to polylog. *CoRR*, abs/2203.15551, 2022.

Bo"az Klartag and Emanuel Milman. Centroid bodies and the logarithmic laplace transform: a unified approach. *Journal of Functional Analysis*, 262(1):10–34, 2012.

Yunbum Kook, Yin Tat Lee, Ruoqi Shen, and Santosh S Vempala. Condition-number-independent convergence rate of riemannian hamiltonian monte carlo with numerical integrators. *arXiv preprint arXiv:2210.07219*, 2022.

Jonathan Leake, Colin S. McSwiggen, and Nisheeth K. Vishnoi. Sampling matrices from harish-chandra-itzykson-zuber densities with applications to quantum inference and differential privacy. In *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1384–1397. ACM, 2021.

Michel Ledoux. *Concentration of measure and logarithmic Sobolev inequalities*. Seminaire de probabilities XXXIII, 1999.

Yin Tat Lee and Santosh S Vempala. Convergence rate of riemannian hamiltonian monte carlo and faster polytope volume computation. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1115–1121, 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*, volume 125 of *Proceedings of Machine Learning Research*, pages 2565–2597. PMLR, 2020.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Lower bounds on metropolized sampling methods for well-conditioned distributions. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, pages 18812–18824, 2021a.

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

Ruilin Li, Molei Tao, Santosh S. Vempala, and Andre Wibisono. The mirror langevin algorithm converges with vanishing bias. In *International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pages 718–742. PMLR, 2022.

Jiaming Liang and Yongxin Chen. A proximal algorithm for sampling from non-smooth potentials. In *2022 Winter Simulation Conference (WSC)*, pages 3229–3240, 2022. doi: 10.1109/WSC57314.2022.10015293.

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 Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random structures & algorithms*, 4(4):359–412, 1993.

László Lovász and Santosh S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Struct. Algorithms*, 30(3):307–358, 2007.

Francesco Mainardi. Lévy stable distributions in the theory of probability. *Lecture Notes on Mathematical Physics*, 2007.

Colin McSwiggen. The harish-chandra integral: An introduction with examples. *L'Enseignement Mathématique*, 67(3):229–299, 2021.

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

A. Nemirovski and D.B̃. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

Arkadi Nemirovski. Interior point polynomial time methods in convex programming. *Lecture notes*, 42(16):3215–3224, 2004.

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

Yurii E Nesterov and Michael J Todd. On the riemannian geometry defined by self-concordant barriers and interior-point methods. *Foundations of Computational Mathematics*, 2(4):333–361, 2002.

Marcelo Pereyra. Proximal markov chain monte carlo algorithms. *Statistics and Computing*, 26:745–760, 2016.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Shai Shalev-Shwartz. Online learning: Theory, algorithms, and applications. PhD thesis, Hebrew University, 2007.

Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, pages 2098–2109, 2019.

Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Nearly-optimal private lasso. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pages 3025–3033, 2015.

Terence Tao. The harish-chandra-itzykson-zuber integral formula. https://terrytao.wordpress.com/2013/02/08/the-harish-chandra-itzykson-zuber-integral-formula/, 2013. Accessed: 2023-02-05.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30, 2017.

Jun-Kun Wang and Jacob D. Abernethy. Acceleration through optimistic no-regret dynamics. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, pages 3828–3838, 2018.

Andre Wibisono. Proximal langevin algorithm: Rapid convergence under isoperimetry. *arXiv preprint arXiv:1911.01469*, 2019.

Kelvin Shuangjian Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. Wasserstein control of mirror langevin monte carlo. In *Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3814–3841. PMLR, 2020.

Appendix A. Applications

In this section, we discuss applications of the sampling scheme we develop in Section 4. We begin by specializing our machinery to ℓ_p and Schatten-p norms in Appendix A.1. We then give new algorithms with improved zeroth-order query complexity for private convex optimization in Appendix A.2. Finally, in Appendix A.3 we discuss computational issues regarding the specific LLT we introduce.

A.1. LLT for ℓ_p and Schatten-p norms

Throughout this section we fix some $p \in [1, 2]$, and define the dual value $q \geq 2$ such that $\frac{1}{q} + \frac{1}{p} = 1$. It is well-known that the ℓ_q norm and ℓ_p norm are dual, as are the corresponding Schatten norms. In light of Lemma 5, to obtain a sampler catering to the ℓ_p geometry for example, it suffices to take the LLT of a smooth function in ℓ_q. We provide the latter by recalling the following fact.

Lemma 18 *Let $p \in [1, 2]$, $q \geq 2$ satisfy $\frac{1}{p} + \frac{1}{q}$. If $\|\cdot\|_q$ is a vector ℓ_q norm, $\frac{1}{2}\|\cdot\|_q^2$ is $\frac{1}{p-1}$-smooth in the ℓ_q norm, and if $\|\cdot\|_q$ is a matrix Schatten-q norm, $\frac{1}{2}\|\cdot\|_q^2$ is $\frac{1}{p-1}$-smooth in the Schatten-q norm.*

Proof This follows (for example) from three well-known facts: 1) that $\frac{1}{2}\|\cdot\|_q^2$ and $\frac{1}{2}\|\cdot\|_p^2$ are conjugate functions in both the vector and matrix cases, 2) that the conjugate of a m-strongly convex function in a norm is $\frac{1}{m}$-smooth in the dual norm Kakade et al. (2009), and 3) that $\frac{1}{2}\|\cdot\|_p^2$ is $(p-1)$-strongly convex in $\|\cdot\|_p$ in both the vector and matrix cases Ball et al. (1994). ∎

ℓ_p norms. Next, for any $a > 0$, when the context is clearly about vector spaces, we define

$$\psi_{p,a}(x) := \log\left(\int \exp\left(\langle x, y\rangle - a\|y\|_q^2\right) \mathrm{d}y\right). \tag{10}$$

Note that as the LLT of a $\frac{2a}{p-1}$-smooth function in ℓ_q, $\psi_{p,a}$ is $\Omega(\frac{p-1}{a})$-strongly convex in ℓ_p by Lemma 5. In applications we fix a value of $\eta > 0$, set $a = \Theta((p-1)\eta)$, and use $\eta\psi_{p,a}$ as our strongly convex regularizer in ℓ_p. We next provide a bound on the range of $\psi_{p,a}$.

Lemma 19 *Let $a > 0$ and let $d \in \mathbb{N}$ be at least a sufficiently large constant. The additive range of $\psi_{p,a}$ over $\{x \in \mathbb{R}^d \mid \|x\|_p \leq 1\}$ is*

$$O\left(1 + \frac{1}{a} + \sqrt{\frac{d}{a}\log\left(a + \frac{d}{a}\right)}\right).$$

In particular, for $a \leq \frac{1}{d\log d}$, the additive range is $O(\frac{1}{a})$.

Proof Throughout the proof denote for simplicity $\psi := \psi_{p,a}$ and let

$$\mathcal{D}_x^{\varphi}(y) \propto \exp\left(\langle x, y\rangle - a\|y\|_q^2\right)$$

be the associated density. By the characterization of $\nabla\psi$ in Lemma 3 and the fact that the associated density $\mathcal{D}_x^{\varphi}$ is symmetric in y for $x = 0$, we have $\nabla\psi(0) = 0$ and hence it suffices to bound $\psi(x) - \psi(0)$ for $\|x\|_q \leq 1$. We simplify this expression as

$$\begin{aligned}\psi(x) - \psi(0) &= \log\left(\int \exp\left(\langle x, y\rangle - a\|y\|_q^2\right) \mathrm{d}y\right) - \log\left(\int \exp\left(-a\|y\|_q^2\right) \mathrm{d}y\right) \\ &= \log\left(\int \exp\left(\langle x, y\rangle\right) \frac{\exp\left(-a\|y\|_q^2\right)}{\int \exp\left(-a\|y\|_q^2\right) \mathrm{d}y} \mathrm{d}y\right) = \log\left(\mathbb{E}_{y\sim\mathcal{D}_0^{\varphi}}\left[\exp\left(\langle x, y\rangle\right)\right]\right).\end{aligned} \tag{11}$$

Next, let π be the probability density on $\mathbb{R}_{\geq 0}$ such that

$$\mathrm{d}\pi(r) \propto r^{d-1} \exp\left(-ar^2\right) \mathrm{d}r.$$

We note $\mathrm{d}\pi(r)$ is the density of the scalar quantity $r = \|y\|_q$ for $y \sim \mathcal{D}_0^{\varphi}$. This can be seen by taking a derivative of the volume of the ℓ_p ball of radius r, which scales as r^d, so the surface area of the ball scales as r^{d-1}. By Hölder's inequality, $\langle x, y\rangle \leq \|y\|_q$ for all y, since $\|x\|_p \leq 1$. We then continue (11) and bound $\psi(x) - \psi(0) \leq \log(\mathbb{E}_{r\sim\pi} \exp(r))$, and the conclusion follows from Lemma 20. ■

Lemma 20 *For any $a > 0$ and $d \in \mathbb{N}$ at least a sufficiently large constant,*

$$\log\left(\frac{\int_0^\infty \exp\left((d-1)\log r + r - ar^2\right) \mathrm{d}r}{\int_0^\infty \exp\left((d-1)\log r - ar^2\right) \mathrm{d}r}\right) \leq 8 + \frac{8}{a} + \sqrt{\frac{8d}{a}\log\left(a + \frac{d}{a}\right)}.$$

Proof Throughout this proof let

$$Z := \int_0^\infty \exp\left((d-1)\log r - \alpha r^2\right) \mathrm{d}r = \frac{\Gamma(\frac{d}{2})}{2a^{\frac{d}{2}}},\ \tau := 7 + \frac{8}{a} + \sqrt{\frac{8d}{a}\log\left(a + \frac{d}{a}\right)}.$$

Next we split the numerator of the left-hand side into two integrals:

$$I_1 := \int_0^\tau \exp\left((d-1)\log r + r - ar^2\right) \mathrm{d}r,$$
$$I_2 := \int_\tau^\infty \exp\left((d-1)\log r + r - ar^2\right) \mathrm{d}r.$$

It is immediate that $I_1 \leq \exp(\tau)Z$. Further, we recognize that for $r \geq \tau$,

$$\max\left(r, (d-1)\log r\right) \leq \frac{ar^2}{4}.$$

The first piece in the maximum is clear from $\tau \geq \frac{4}{a}$. The second follows since $\frac{r^2}{\log r}$ is an increasing function for $r \geq 7$, and either $\frac{4d}{a} \leq 10$ in which case we use $\frac{7^2}{\log 7} \geq 10$, or we let $C := \frac{4d}{a}$ and use

$$\frac{r^2}{\log r} \geq C \text{ for } r \geq \sqrt{2C\log\frac{C}{4}},\ C \geq 10.$$

Hence we may bound

$$I_2 \leq \int_\tau^\infty \exp\left(-\frac{ar^2}{2}\right) = \sqrt{\frac{2\pi}{a}} \Pr_{t\sim\mathcal{N}(0,a^{-1})}[t \geq \tau] \leq \frac{2}{a\tau}\exp\left(-\frac{a\tau^2}{2}\right).$$

Above, we used Mill's inequality

$$\Pr_{t\sim\mathcal{N}(0,\sigma^2)}[t \geq \tau] \leq \sqrt{\frac{2}{\pi}}\frac{\sigma}{\tau}\exp\left(-\frac{\tau^2}{2\sigma^2}\right).$$

Further for our τ, our upper bound on I_1 is larger than our upper bound on I_2. To see this,

$$\begin{aligned}\tau\left(1 + \frac{a\tau}{2}\right) + \frac{d}{3}\log d \geq \frac{d}{2}\log a &\implies \exp\left(\tau\left(1 + \frac{a\tau}{2}\right)\right)\Gamma\left(\frac{d}{2}\right) \geq a^{\frac{d}{2}} \\ &\implies \frac{\exp(\tau)\Gamma(\frac{d}{2})}{2a^{\frac{d}{2}}} \geq \frac{4}{a\tau}\exp\left(-\frac{a\tau^2}{2}\right).\end{aligned}$$

The first inequality is because $a\tau^2 \geq d \log a$. The first implication then follows by exponentiating and using $\log \Gamma(\frac{d}{2}) \geq \frac{d}{3} \log d$ for sufficiently large d, and the second implication follows by rearranging and using $a\tau \geq 4$. Finally the conclusion follows from

$$\log \left(\frac{\int_0^\infty \exp\left((d-1)\log r + r - ar^2\right) \mathrm{d}r}{\int_0^\infty \exp\left((d-1)\log r - ar^2\right) \mathrm{d}r} \right) \leq \log \left(\frac{2\exp(\tau) Z}{Z} \right) \leq \tau + 1.$$

■

Schatten-p norms. When the context is clearly about matrix spaces, we analogously define

$$\psi_{p,a}(\mathbf{X}) := \log \left(\int \exp\left(\langle \mathbf{X}, \mathbf{Y} \rangle - a \left\| \mathbf{Y} \right\|_q^2 \right) \mathrm{d}y \right).$$

The proof of Lemma 19 implies the following analogous range bound in this setting.

Corollary 21 *Let $a > 0$ and let $d_1, d_2 \in \mathbb{N}$ be at least sufficiently large constants. The additive range of $\psi_{p,a}$ over $\{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2} \mid \left\| \mathbf{X} \right\|_p \leq 1\}$ is*

$$O\left(1 + \frac{1}{a} + \sqrt{\frac{d_1 d_2}{a} \log\left(a + \frac{d_1 d_2}{a} \right)} \right).$$

In particular, for $a \leq \frac{1}{d_1 d_2 \log(d_1 d_2)}$, the additive range is $O(\frac{1}{a})$.

A.2. Zeroth-order private convex optimization

In this section, we consider a pair of closely-related problems in private convex optimization. Let $\mathcal{S}$ be a domain, and let $n \in \mathbb{N}$. We say that a mechanism (randomized algorithm) $\mathcal{M} : \mathcal{S}^n \to \Omega$ satisfies (ϵ, δ)-differential privacy (DP) if for any event $S \subseteq \Omega$ where Ω is the output space, and any two datasets $\mathcal{D}, \mathcal{D}' \in \mathcal{S}^n$ which differ in exactly one element,

$$\Pr[\mathcal{M}(\mathcal{D}) \in S] \leq \exp(\epsilon) \Pr[\mathcal{M}(\mathcal{D}') \in S] + \delta.$$

We next define the private optimization problems we study.

Problem 22 (DP-ERM and DP-SCO) *Let $n \in \mathbb{N}$, $\epsilon, \delta \in (0,1)$, $D, G \geq 0$, and let $\mathcal{X} \subset \mathbb{R}^d$ be compact and convex with diameter in a norm $\|\cdot\|_{\mathcal{X}}$ at most D. Let $\mathcal{P}$ be a distribution over a set $\mathcal{S}$ such that for any $s \in \mathcal{S}$, there is a $f(\cdot; s) : \mathcal{X} \to \mathbb{R}$ which is convex and G-Lipschitz in $\|\cdot\|_{\mathcal{X}}$. Let $\mathcal{D} := \{s_i\}_{i \in [n]}$ consist of n independent draws from $\mathcal{P}$, and let $f_i := f(\cdot; s_i)$ for all $i \in [n]$.*

In the differentially private empirical risk minimization (DP-ERM) *problem, we receive $\mathcal{D}$ and wish to design a mechanism $\mathcal{M}$ which satisfies (ϵ, δ)-DP and approximately minimizes*

$$F_{\text{erm}}(x) := \frac{1}{n} \sum_{i \in [n]} f_i(x).$$

In the differentially private stochastic convex optimization (DP-SCO) *problem, we receive $\mathcal{D}$ and wish to design a mechanism $\mathcal{M}$ which satisfies (ϵ, δ)-DP and approximately minimizes*

$$F_{\text{sco}}(x) := \mathbb{E}_{s \sim \mathcal{P}} \left[f(x; s) \right].$$

The following powerful general-purpose result was proven in Gopi et al. (2023) reducing the DP-ERM and DP-SCO problems to logconcave sampling problems catered to the $\|\cdot\|_{\mathcal{X}}$ geometry. We slightly improve the parameter settings used by Theorem 4 of Gopi et al. (2023) for DP-SCO by noting that a smaller value of k also suffices (due to the larger error bound), as observed by Gopi et al. (2022).

Proposition 23 (Theorem 3, Theorem 4, Gopi et al. (2023), Theorem 6.9, Gopi et al. (2022)) *In the setting of Problem 22, let $k \geq 0$, and let $r : \mathcal{X} \to \mathbb{R}$ be 1-strongly convex with respect to $\|\cdot\|_{\mathcal{X}}$, with additive range at most Θ. Let ν be the density on $\mathcal{X}$ satisfying $\mathrm{d}\nu(x) \propto \exp(-k(F_{\text{erm}}(x) + \mu r(x)))\mathbf{1}_{\mathcal{X}}(x)\mathrm{d}x$. Then the algorithm which returns a sample from ν for*

$$k = \frac{\sqrt{d}n\epsilon}{G\sqrt{2\Theta \log \frac{1}{2\delta}}},\ \mu = \frac{2G^2 k \log \frac{1}{2\delta}}{n^2\epsilon^2},$$

satisfies (ϵ, δ)-DP, and guarantees

$$\mathbb{E}_{x\sim\nu}\left[F_{\text{erm}}(x)\right] - \min_{x\in\mathcal{X}} F_{\text{erm}}(x) \leq O\left(G\sqrt{\Theta} \cdot \frac{\sqrt{d\log\frac{1}{\delta}}}{n\epsilon}\right).$$

Further, the algorithm which returns a sample from ν for

$$k = \frac{1}{G\sqrt{\Theta}} \cdot \sqrt{\left(\frac{d\log\frac{1}{2\delta}}{\epsilon^2 n^2} + \frac{1}{n}\right) \cdot \min\left(\frac{\epsilon^2 n^2}{\log\frac{1}{2\delta}}, nd\right)},\ \mu = G^2 k \cdot \max\left(\frac{\log\frac{1}{2\delta}}{n^2\epsilon^2}, \frac{1}{nd}\right)$$

satisfies (ϵ, δ)-DP, and guarantees

$$\mathbb{E}_{x\sim\nu}\left[F_{\text{sco}}(x)\right] - \min_{x\in\mathcal{X}} F_{\text{sco}}(x) \leq O\left(G\sqrt{\Theta} \cdot \left(\frac{\sqrt{d\log\frac{1}{\delta}}}{n\epsilon} + \frac{1}{\sqrt{n}}\right)\right).$$

Armed with Proposition 23 and the sampler in Theorem 16, we give our main results on Problem 22.

Assumption 24 *Fix $p \in [1, 2]$ and $k, a, \eta, \mu > 0$. In the setting of Problem 22, assume there is an algorithm $\mathcal{A}$ which returns a point drawn from a β-warm start to the density ν satisfying*

$$\mathrm{d}\nu(x) \propto \exp\left(-k\left(F_{\text{erm}}(x) + \eta\mu\psi_{p,a}(x)\right)\right)\mathbf{1}_{\mathcal{X}}(x)\mathrm{d}x.$$

Theorem 25 *Let $p \in [1, 2]$, $\epsilon, \delta \in (0, 1)$. In the setting of Problem 22 where $\|\cdot\|_{\mathcal{X}}$ is the ℓ_p norm on $\mathbb{R}^d$, there is an (ϵ, δ)-differentially private algorithm $\mathcal{M}_{\text{erm}}$ which produces $x \in \mathcal{X}$ such that*

$$\mathbb{E}_{\mathcal{M}_{\text{erm}}}\left[F_{\text{erm}}(x)\right] - \min_{x\in\mathcal{X}} F_{\text{erm}}(x) = O\left(\frac{GD}{\sqrt{p-1}} \cdot \frac{\sqrt{d\log\frac{1}{\delta}}}{n\epsilon}\right) \text{ for } p \in (1, 2],$$

$$\mathbb{E}_{\mathcal{M}_{\text{erm}}}\left[F_{\text{erm}}(x)\right] - \min_{x\in\mathcal{X}} F_{\text{erm}}(x) = O\left(GD\sqrt{\log d} \cdot \frac{\sqrt{d\log\frac{1}{\delta}}}{n\epsilon}\right) \text{ for } p = 1.$$

Further, there is an (ϵ, δ)-differentially private algorithm $\mathcal{M}_{\text{sco}}$ which produces $x \in \mathcal{X}$ such that

$$\mathbb{E}_{\mathcal{M}_{\text{sco}}}\left[F_{\text{sco}}(x)\right] - \min_{x \in \mathcal{X}} F_{\text{sco}}(x) = O\left(\frac{GD}{\sqrt{p-1}} \cdot \left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d \log \frac{1}{\delta}}}{n\epsilon}\right)\right) \text{ for } p \in (1, 2],$$

$$\mathbb{E}_{\mathcal{M}_{\text{sco}}}\left[F_{\text{sco}}(x)\right] - \min_{x \in \mathcal{X}} F_{\text{sco}}(x) = O\left(GD\sqrt{\log d} \cdot \left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d \log \frac{1}{\delta}}}{n\epsilon}\right)\right) \text{ for } p = 1.$$

Both $\mathcal{M}_{\text{erm}}$ and $\mathcal{M}_{\text{sco}}$ call $\mathcal{A}$ in Assumption 24, appropriately parameterized, once. $\mathcal{M}_{\text{erm}}$ uses

$$O\left(\left(1 + \frac{n^2\epsilon^2}{\log \frac{1}{\delta}}\right) \log\left(\frac{(1+n\epsilon)\log \beta}{\delta}\right) \log \frac{\beta}{\delta}\right).$$

additional value queries to some $f(\cdot; s_i)$, and $\mathcal{M}_{\text{sco}}$ uses

$$O\left(\min\left(nd,\ 1 + \frac{n^2\epsilon^2}{\log \frac{1}{\delta}}\right) \log\left(\frac{(1+n\epsilon)\log \beta}{\delta}\right) \log \frac{\beta}{\delta}\right)$$

additional value queries to some $f(\cdot; s_i)$.

Proof First, we slightly simplify the setting of Problem 22. We may first assume that $D = 1$, i.e. $\mathcal{X}$ has diameter at most 1 in $\|\cdot\|_{\mathcal{X}}$. If the diameter is bounded by some $D \neq 1$, we can rescale the domain $\mathcal{X} \leftarrow \frac{1}{D}\mathcal{X}$, and remap to the modified functions $f(x; s) \leftarrow f(Dx; s)$ over this modified domain for all $s \in \mathcal{S}$. It is clear the Lipschitz constant rescales as $G \leftarrow GD$ as a result. Next, we assume $(n\epsilon)^2 \geq d\Theta \log \frac{1}{\delta}$ where $\Theta = \min(\frac{1}{p-1}, \log d)$. In the other case, in light of the diameter bound on $\mathcal{X}$ and the Lipschitz assumption, returning a random point in $\mathcal{X}$ attains the error bound claimed. Finally, assume $p \in (1, 2]$, as otherwise we set $p \leftarrow 1 + \frac{1}{\log d}$, which only affects bounds by constant factors, since $\|\cdot\|_p$ is affected by $O(1)$ multplicatively everywhere under this change.

Under these simplifications, we choose the parameters k and μ according to Proposition 22 for each problem. Assume for now that Θ for the regularizer r we choose is bounded by a universal constant times $\frac{1}{p-1}$. Then the Lipschitz constant of kF_{erm} in either case of Proposition 22 is

$$kG = \Omega\left(\min\left(\frac{\sqrt{(p-1)d}n\epsilon}{\sqrt{\log \frac{1}{\delta}}}, d\sqrt{n}\right)\right) = \Omega(d),$$

as implied by our earlier simplification. We hence may choose $\mathcal{I}$ to be uniform over $[n]$, and

$$\eta = O\left(\frac{1}{k^2G^2 \log \frac{(1+n\epsilon)\log \beta}{\delta}}\right)$$

for a sufficiently small constant to use Theorem 16. Under this setting we certainly have $\eta = O(\frac{1}{d^2})$, so letting $r := 12\eta\psi_{p,a}$ for $a := \frac{\eta(p-1)}{2}$ shows that r is 12η times the LLT of an η-smooth function in ℓ_q. By Lemma 5, r is indeed 1-strongly convex in ℓ_p, and Lemma 19 bounds its range by $\Theta = O(\frac{1}{p-1})$ satisfying our earlier assumption, where we use $a = O(\frac{1}{d^2})$. The runtime finally follows by applying our choices of k, μ in Proposition 23, with our choice of η, in Theorem 16, where we ensure that $\eta \cdot k\mu \leq 1$ by choosing a smaller η if this is not the case (so Theorem 16 applies). Finally, to account for total variation error in our sampler, it suffices to adjust the failure probability δ by a constant and take a union bound over the privacy definition and the failure of Theorem 16. ■

By combining the proof strategy of Theorem 25 with Corollary 21 instead of Lemma 19, we immediately obtain the following corollary in the case of Schatten norms.

Corollary 26 *Let $p \in [1, 2]$, $\epsilon, \delta \in (0, 1)$. In the setting of Problem 22 where $\|\cdot\|_{\mathcal{X}}$ is the Schatten-p norm on $\mathbb{R}^{d_1 \times d_2}$, there is an (ϵ, δ)-differentially private algorithm $\mathcal{M}_{\text{erm}}$ which produces $\mathbf{X} \in \mathcal{X}$ such that*

$$\mathbb{E}_{\mathcal{M}_{\text{erm}}}\left[F_{\text{erm}}(\mathbf{X})\right] - \min_{\mathbf{X}\in\mathcal{X}} F_{\text{erm}}(\mathbf{X}) = O\left(\frac{GD}{\sqrt{p-1}} \cdot \frac{\sqrt{d_1 d_2 \log\frac{1}{\delta}}}{n\epsilon}\right) \text{ for } p \in (1, 2],$$

$$\mathbb{E}_{\mathcal{M}_{\text{erm}}}\left[F_{\text{erm}}(\mathbf{X})\right] - \min_{\mathbf{X}\in\mathcal{X}} F_{\text{erm}}(\mathbf{X}) = O\left(GD\sqrt{\log(d_1 d_2)} \cdot \frac{\sqrt{d_1 d_2 \log\frac{1}{\delta}}}{n\epsilon}\right) \text{ for } p = 1.$$

Further, there is an (ϵ, δ)-differentially private algorithm $\mathcal{M}_{\text{sco}}$ which produces $\mathbf{X} \in \mathcal{X}$ such that

$$\mathbb{E}_{\mathcal{M}_{\text{sco}}}\left[F_{\text{sco}}(\mathbf{X})\right] - \min_{\mathbf{X}\in\mathcal{X}} F_{\text{sco}}(\mathbf{X}) = O\left(\frac{GD}{\sqrt{p-1}} \cdot \left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d_1 d_2 \log\frac{1}{\delta}}}{n\epsilon}\right)\right) \text{ for } p \in (1, 2],$$

$$\mathbb{E}_{\mathcal{M}_{\text{sco}}}\left[F_{\text{sco}}(\mathbf{X})\right] - \min_{\mathbf{X}\in\mathcal{X}} F_{\text{sco}}(\mathbf{X}) = O\left(GD\sqrt{\log(d_1 d_2)} \cdot \left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d_1 d_2 \log\frac{1}{\delta}}}{n\epsilon}\right)\right) \text{ for } p = 1.$$

Both $\mathcal{M}_{\text{erm}}$ and $\mathcal{M}_{\text{sco}}$ call $\mathcal{A}$ in Assumption 24, appropriately parameterized, once. $\mathcal{M}_{\text{erm}}$ uses

$$O\left(\left(1 + \frac{n^2\epsilon^2}{\log\frac{1}{\delta}}\right) \log\left(\frac{(1+n\epsilon)\log\beta}{\delta}\right) \log\frac{\beta}{\delta}\right).$$

additional value queries to some $f(\cdot; s_i)$, and $\mathcal{M}_{\text{sco}}$ uses

$$O\left(\min\left(nd_1 d_2,\ 1 + \frac{n^2\epsilon^2}{\log\frac{1}{\delta}}\right) \log\left(\frac{(1+n\epsilon)\log\beta}{\delta}\right) \log\frac{\beta}{\delta}\right)$$

additional value queries to some $f(\cdot; s_i)$.

A.3. Oracle access for $\psi_{p,a}$

In Theorem 25 and Corollary 26, we only bounded the value oracle complexity of our sampling algorithms. The remainder of the steps in Algorithm 1 and its subroutine Algorithm 2 require samples from densities of the form $\mathrm{d}\pi_x$ (for some $x \in \mathcal{X}$) or $\mathrm{d}\gamma_y$ (for some $y \in \mathbb{R}^d$), defined in (9) and (24) respectively and reproduced here for convenience:

$$\begin{aligned} \mathrm{d}\pi_x(y) &= \exp\left(\langle x, y\rangle - \psi(x) - \varphi(y)\right) \mathrm{d}y, \\ \mathrm{d}\gamma_y(x) &\propto \exp\left(-\eta\mu\psi(x) - (\psi(x) - \langle x, y\rangle)\right) \mathbf{1}_{\mathcal{X}}(x)\mathrm{d}x. \end{aligned} \quad (12)$$

These densities are independent of the function F in Problem 10 and hence do not require additional value oracle queries in the setting of Problem 10. In general, the complexity of these steps depends on the complexity of the functions φ and ψ, and the set $\mathcal{X}$. We now discuss strategies for sampling from π_x and γ_y in specific settings described by Appendix A.1, which we first briefly summarize.

(1) We describe a method based on the inverse Laplace transform for sampling from π_x and evaluating $\psi_{p,a}$ with complexity linear in the dimension d in the vector setting.

(2) Under efficient value oracle access to $\psi_{p,a}$ and membership oracle access to $\mathcal{X}$, general-purpose results Lovász and Vempala (2007); Jia et al. (2021); Jambulapati et al. (2022) imply polynomial-time samplers for γ_y.

(3) We discuss generalizations of these methods to the matrix setting, and naïve sampling methods. We draw a loose connection to the HCIZ integral from harmonic analysis, and suggest how it may potentially help in the structured sampling task for LLTs in Schatten norms.

ℓ_p setting. We first discuss the case when $\mathcal{X} \subset \mathbb{R}^d$ is a set on vectors equipped with the ℓ_p norm for some $p \in [1, 2]$, and we let $q \geq 2$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. We follow the notation (10).

In order to sample from the density π_x, we use an *inverse Laplace transform* decomposition. For a parameter $c \in [0, 1)$, we define the density μ_c supported on $\mathbb{R}_{\geq 0}$, such that for all $t \geq 0$,

$$\exp(-t^c) = \int_0^\infty \exp(-\lambda t)\, \mu_c(\lambda) \mathrm{d}\lambda. \tag{13}$$

Intuitively, the density $\mu_c(\lambda)$ and the corresponding decomposition (inverse Laplace transform) (13) aims to express the more heavy-tailed function $\exp(-t^c)$ as a distribution over the lighter-tailed functions $\exp(-\lambda t)$. The inverse Laplace transform densities μ_c are well-studied in the probability theory literature, and correspond to *stable count distributions* parameterized by c. For example, it is well-known that $\mu_{\frac{1}{2}}$ is the Lévy distribution

$$\mathrm{d}\mu_{\frac{1}{2}}(\lambda) = \frac{1}{2\sqrt{\pi}\lambda^{\frac{3}{2}}} \exp\left(-\frac{1}{4\lambda}\right) \mathrm{d}\lambda.$$

We refer the reader to references e.g. Mainardi (2007) on properties of the densities μ_c, and for now assume we can access and sample from these one-dimensional distributions in closed form for simplicity. Given this decomposition, we can then write

$$\begin{aligned}
\exp(\psi_{p,a}(x)) &= \int \exp\left(\langle x, y\rangle - a\left\|y\right\|_q^2\right) \mathrm{d}y \\
&= \int_0^\infty \left(\int \exp\left(\langle x, y\rangle - \lambda a^{\frac{q}{2}} \left\|y\right\|_q^q\right) \mathrm{d}y\right) \mu_{\frac{2}{q}}(\lambda)\mathrm{d}\lambda \\
&= \int_0^\infty \prod_{i\in[d]} \left(\int_{-\infty}^\infty \exp\left(x_i y_i - \lambda a^{\frac{q}{2}} y_i^q\right) \mathrm{d}y_i\right) \mu_{\frac{2}{q}}(\lambda)\mathrm{d}\lambda.
\end{aligned} \tag{14}$$

The decomposition (14) reduces the problem of sampling from π_x to d one-dimensional problems. To sample $\propto \exp(\langle x, y\rangle - a\left\|y\right\|_q^2)$, we can first sample λ from the density μ_c for $c = \frac{2}{q}$, and then sample each coordinate y_i proportionally to $\exp(x_i y_i - \lambda a^{\frac{q}{2}} y_i^q)$ conditioned on the sampled λ.

This decomposition also gives us an efficient value oracle for $\psi_{p,a}$, by evaluating (14) as a one-dimensional integral over λ, where the integrand may be evaluated as a product of d one-dimensional integrals. Under membership oracle access to $\mathcal{X}$, the problem of sampling from γ_y then falls under a generic logconcave sampling setup studied in a long line of work building upon Dyer et al. (1991). The state-of-the-art general-purpose logconcave sampler, which combines the algorithms of Lovász and Vempala (2007); Jia et al. (2021) with the isoperimetric bound in Jambulapati et al. (2022) (improving recent breakthroughs by Chen (2021); Klartag and Lehec (2022)), requires roughly d^3 value oracle calls to $\psi_{p,a}$ and membership oracle calls to $\mathcal{X}$.

In principle, for structured sets $\mathcal{X}$ (such as ℓ_p balls), the particular explicit structure of $\psi_{p,a}$ and $\mathcal{X}$ may be exploited to design more efficient samplers for the densities γ_y, analogously to our custom linear-time sampler for π_x. However, it should be noted that the sampling problem for γ_y appears to be quite a bit more challenging than the problem for π_x. We leave the investigation of explicit sampler design for γ_y as an interesting open problem for future work.

Schatten-p setting. The situation is somewhat less straightforward in the matrix case. Here, the key computational problem in replicating the strategy suggested by (14) is evaluating the integral

$$\int \exp\left(\langle \mathbf{X}, \mathbf{Y}\rangle - C\left\|\mathbf{Y}\right\|_q^q\right) \mathrm{d}\mathbf{Y}, \tag{15}$$

where the integral is over $\mathbf{Y} \in \mathbb{R}^{d_1 \times d_2}$, and $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, $C > 0$ are fixed. The difficulty is $\langle \mathbf{X}, \mathbf{Y}\rangle$ decomposes coordinatewise, whereas $\left\|\mathbf{Y}\right\|_q^q$ decomposes spectrally.[8] At least superficially, this is similar to the challenge faced when evaluating the Harish-Chandra-Itzykson-Zuber (HCIZ) formula

$$\int \exp\left(\mathrm{Tr}\left(\mathbf{AUBU}^\dagger\right)\right) \mathrm{d}\mathbf{U}, \tag{16}$$

where the integral is over the Haar measure on (complex) unitary matrices $\mathbf{U}$, and $\mathbf{A}$, $\mathbf{B}$ are Hermitian. By dropping the $-C\left\|\mathbf{Y}\right\|_q^q$ term in (15) and only integrating over unitary conjugations of a fixed matrix $\mathbf{Y}$, we arrive at a generalization of (16). The difficulty in evaluating (16) is also a sort of tension between the eigenspaces of $\mathbf{A}$ and $\mathbf{B}$. Nonetheless, (16) has a (polynomial-time computable) exact formula, which was famously discovered independently by Harish-Chandra (1957); Itzykson and Zuber (1980). Furthermore, Leake et al. (2021) recently obtained a polynomial-time sampler for the density induced by (16); while a sampler for (15) would follow from logconcavity and general-purpose results, it would be far from cheap, so ways of exploiting structure are fruitful to explore.

As a proof-of-concept, evaluating the integral (15) in (polynomial-time computable) closed form is a minimal requirement for implementing the $\mathbf{X}$-oracles in (12) used by our algorithm. Even this problem appears challenging, but (as summarized cleanly by Tao (2013); McSwiggen (2021)) a plethora of techniques exist for proving the HCIZ formula, some based on tools from stochastic processes. We pose the efficient computability of the integral (15) as another explicit open question.

Appendix B. Deferred proofs from Section 3

Lemma 27 (LLT derivatives) *For any $x, h \in \mathbb{R}^d$, we have $\nabla\psi(x) = \mu(\mathcal{D}_x^\varphi) := \mathbb{E}_{y\sim\mathcal{D}_x^\varphi}[y]$, $\nabla^2\psi(x) = \mathrm{Cov}(\mathcal{D}_x^\varphi) := \mathbb{E}_{y\sim\mathcal{D}_x^\varphi}\left[(y-\mu(\mathcal{D}_x^\varphi))(y-\mu(\mathcal{D}_x^\varphi))^\top\right]$, $\nabla^3\psi(x)[h,h,h] = \mathbb{E}_{y\sim\mathcal{D}_x^\varphi}\left[\langle y-\mu(\mathcal{D}_x^\varphi), h\rangle^3\right]$.*

Proof For any $x \in \mathbb{R}^d$, a straightforward calculation shows that

$$\nabla\psi(x) = \nabla\left(\log\int\exp\left(\langle x,y\rangle - \varphi(y)\right)\mathrm{d}y\right) = \frac{\int\exp\left(\langle x,y\rangle - \varphi(y)\right)y\mathrm{d}y}{\int\exp\left(\langle x,y\rangle - \varphi(y)\right)\mathrm{d}y} = \mu(\mathcal{D}_x^\varphi).$$

Further,

$$\begin{aligned}\nabla^2\psi(x) &= \nabla\left(\frac{\int\exp\left(\langle x,y\rangle - \varphi(y)\right)y\mathrm{d}y}{\int\exp\left(\langle x,y\rangle - \varphi(y)\right)\mathrm{d}y}\right)\\ &= \frac{\int\exp\left(\langle x,y\rangle - \varphi(y)\right)yy^\top\mathrm{d}y}{\int\exp\left(\langle x,y\rangle - \varphi(y)\right)\mathrm{d}y} - \frac{\left(\int\exp\left(\langle x,y\rangle - \varphi(y)\right)y\mathrm{d}y\right)\left(\int\exp\left(\langle x,y\rangle - \varphi(y)\right)y\mathrm{d}y\right)^\top}{\left(\int\exp\left(\langle x,y\rangle - \varphi(y)\right)\mathrm{d}y\right)^2}.\end{aligned}$$

8. Note that because $\left\|\cdot\right\|_q$ is unitarily invariant, we may assume $\mathbf{X}$ is diagonal.

Finally,

$$\begin{aligned}\nabla^3\psi(x)[h,h,h] &= h^\top \nabla \left(\frac{\int \exp\left(\langle x,y\rangle - \varphi(y)\right)\left(y^\top h\right)^2 \mathrm{d}y}{\int \exp\left(\langle x,y\rangle - \varphi(y)\right)\mathrm{d}y} - \frac{\left(\int \exp\left(\langle x,y\rangle - \varphi(y)\right) y^\top h \mathrm{d}y\right)^2}{\left(\int \exp\left(\langle x,y\rangle - \varphi(y)\right)\mathrm{d}y\right)^2} \right) \\ &= \frac{\int \exp\left(\langle x,y\rangle - \varphi(y)\right)\left(y^\top h\right)^3 \mathrm{d}y}{\int \exp\left(\langle x,y\rangle - \varphi(y)\right)\mathrm{d}y} + 2\left(\frac{\int \exp\left(\langle x,y\rangle - \varphi(y)\right) y^\top h \mathrm{d}y}{\int \exp\left(\langle x,y\rangle - \varphi(y)\right)\mathrm{d}y}\right)^3 \\ &\quad - \frac{3\int \exp\left(\langle x,y\rangle - \varphi(y)\right)\left(y^\top h\right)^2 \mathrm{d}y \int \exp\left(\langle x,y\rangle - \varphi(y)\right) y^\top h \mathrm{d}y}{\left(\int \exp\left(\langle x,y\rangle - \varphi(y)\right)\mathrm{d}y\right)^2}.\end{aligned}$$

■

Lemma 28 (Self-concordance) *If ψ is the LLT of a convex function, it is self-concordant.*

Proof By the definition of self-concordance and Lemma 3, it suffices to show for any $h \in \mathbb{R}^d$,

$$\mathbb{E}_{y\sim\mathcal{D}_x^\varphi}\left[\langle y-\mu(\mathcal{D}_x^\varphi), h\rangle\right]^3 \le 2\left(\mathbb{E}_{y\sim\mathcal{D}_x^\varphi}\left[\langle y-\mu(\mathcal{D}_x^\varphi), h\rangle^2\right]\right)^{\frac{3}{2}}. \tag{17}$$

We then note that the random variable $\langle y-\mu(\mathcal{D}_x^\varphi), h\rangle$ for $y\sim\mathcal{D}_x^\varphi$ follows a logconcave distribution because affine transformations preserve logconcavity. Finally Lemma 2 of Bubeck and Eldan (2019) implies (17) holds. ■

Lemma 29 *Suppose $F : \mathbb{R}^d \to \mathbb{R}$ is L-smooth in $\|\cdot\|$. Let π be the distribution on $\mathcal{X}$ given by $\pi(x) \propto \exp(-F(x))$, and for any $u\in\mathbb{R}^d$, let $\pi^{(u)}$ be the distribution given by $\pi^{(u)}(x)\propto\exp(-F(x+u))$. Then, for any $\alpha\ge 1$,*

$$D_\alpha(\pi^{(u)}\|\pi) \le \frac{\alpha L}{2}\|u\|^2.$$

Proof Recall that L-smoothness of F implies that for any $x,y\in\mathbb{R}^d$,

$$F(x) \le F(y) + \langle \nabla F(y), y-x\rangle + \frac{L}{2}\|y-x\|^2. \tag{18}$$

For $\alpha > 1$, applying (18) on $(x, x+u)$ and $(x+\alpha u, x+u)$ and taking a weighted combination gives

$$F(x+\alpha u) \le \alpha F(x+u) + (1-\alpha)F(x) + \frac{\alpha(\alpha-1)L}{2}\|u\|^2.$$

Let $Z = \int \exp(-F(x))\mathrm{d}x$ which is the normalizing constant for all $\pi^{(u)}$. Then,

$$\begin{aligned} D_\alpha(\pi^{(u)}\|\pi) &= \frac{1}{\alpha-1}\log\int\frac{\pi^{(u)}(x)^\alpha}{\pi(x)^{\alpha-1}}\mathrm{d}x \\ &= \frac{1}{\alpha-1}\log\int\frac{1}{Z}\exp\left(-\alpha F(x+u) - (1-\alpha)F(x)\right)\mathrm{d}x \\ &\le \frac{1}{\alpha-1}\log\int\frac{1}{Z}\exp\left(-F(x+\alpha u) + \frac{\alpha(\alpha-1)L}{2}\|u\|^2\right)\mathrm{d}x \\ &= \frac{1}{\alpha-1}\log\exp\left(\frac{\alpha(\alpha-1)L}{2}\|u\|^2\right) = \frac{\alpha L}{2}\|u\|^2. \end{aligned}$$

The case $\alpha = 1$ follows by taking a limit as $\alpha\to 1$. ■

Lemma 30 *Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is convex and self-concordant. For any $x \in \mathbb{R}$,*

$$\frac{\exp(-\phi(x))}{\sqrt{\phi''(x)}} \geq \frac{1}{12} \min \left\{ \int_{-\infty}^{x} \exp(-\phi(t))dt, \int_{x}^{\infty} \exp(-\phi(t))dt \right\}.$$

Proof Assume $\phi'(x) \geq 0$ (the other case will follow analogously by bounding the integral on $(-\infty, x]$). Define $r := x + \frac{1}{4\sqrt{\phi''(x)}}$. By self-concordance (Lemma 2), for all $t \in [x, r]$,

$$\frac{1}{2}\phi''(x) \leq \phi''(t) \leq 2\phi''(x).$$

Hence, we have for all $t \in [x, r]$, since $\phi'(x) \geq 0$,

$$\phi(t) = \phi(x) + \phi'(x)(t - x) + \int_{x}^{t} (t - s)\phi''(s)\mathrm{d}s \geq \phi(x) + \frac{1}{4}(t - x)^2\phi''(x). \tag{19}$$

We use (19) to bound the integral on $[x, r]$:

$$\begin{aligned} \int_{x}^{r} \exp(-\phi(t))\mathrm{d}t &\leq \exp(-\phi(x)) \int_{x}^{r} \exp\left(-\frac{1}{4}(t - x)^2\phi''(x)\right) \mathrm{d}t \\ &\leq \exp(-\phi(x)) \int_{-\infty}^{\infty} \exp\left(-\frac{1}{4}(t - x)^2\phi''(x)\right) \mathrm{d}t = 2\sqrt{\pi} \cdot \frac{\exp(-\phi(x))}{\sqrt{\phi''(x)}}. \end{aligned} \tag{20}$$

Next, to bound the integral on $[r, \infty)$, we first observe

$$\phi'(r) \geq \phi'(x) + \int_{x}^{r} \phi''(r)\mathrm{d}t \geq \frac{1}{2} \int_{x}^{r} \phi''(x)\mathrm{d}t \geq \frac{1}{8}\sqrt{\phi''(x)}.$$

Hence, by convexity from r,

$$\begin{aligned} \int_{r}^{\infty} \exp(-\phi(t))\mathrm{d}t &\leq \int_{r}^{\infty} \exp\left(-\phi(r) - \phi'(r)(t - r)\right) \mathrm{d}t \\ &\leq \exp(-\phi(x)) \int_{r}^{\infty} \exp\left(-\frac{1}{8}\sqrt{\phi''(x)}(t - r)\right) \mathrm{d}t = 8 \cdot \frac{\exp(-\phi(x))}{\sqrt{\phi''(x)}}. \end{aligned} \tag{21}$$

We used $\phi(r) \geq \phi(x)$ by convexity and $\phi'(x) \geq 0$. Combining (20) and (21) yields the claim. ∎

Lemma 31 (Modification of the localization lemma, Kannan et al. (1995), Theorem 2.7) *Let f_1, f_2, f_3, f_4 be four nonnegative functions on $\mathbb{R}^d$ such that f_1 and f_2 are upper semicontinuous and f_3 and f_4 are lower semicontinuous, let $c_1, c_2 > 0$, and let $\phi : \mathbb{R}^d \to \mathbb{R}$ be convex. Then, the following are equivalent:*

- *For every density $\pi : \mathbb{R}^d \to \mathbb{R}$ which is 1-relatively strongly logconcave in ϕ,*

$$\left(\int f_1(x)\pi(x)\mathrm{d}x\right)^{c_1} \left(\int f_2(x)\pi(x)\mathrm{d}x\right)^{c_2} \leq \left(\int f_3(x)\pi(x)\mathrm{d}x\right)^{c_1} \left(\int f_4(x)\pi(x)\mathrm{d}x\right)^{c_2}.$$

- *For every $a, b \in \mathbb{R}^d$ and $\gamma \in \mathbb{R}$,*

$$\begin{aligned} &\left(\int_0^1 f_1((1 - t)a + tb)e^{\gamma t - \phi((1-t)a+tb)}\mathrm{d}t\right)^{c_1} \left(\int_0^1 f_2((1 - t)a + tb)e^{\gamma t - \phi((1-t)a+tb)}\mathrm{d}t\right)^{c_2} \\ \leq &\left(\int_0^1 f_3((1 - t)a + tb)e^{\gamma t - \phi((1-t)a+tb)}\mathrm{d}t\right)^{c_1} \left(\int_0^1 f_4((1 - t)a + tb)e^{\gamma t - \phi((1-t)a+tb)}\mathrm{d}t\right)^{c_2}. \end{aligned}$$

Proof The proof follows identically to the case where $\phi = 0$, which was proven in Lovász and Simonovits (1993); Kannan et al. (1995) via a bisection argument (see Lemma 2.5, Lovász and Simonovits (1993)). The only fact the bisection argument relies on is that restricting logconcave densities to subsets of $\mathbb{R}^d$ preserves logconcavity, which remains true for densities which are relatively strongly logconcave with respect to a given convex function. For a more formal treatment of this generalized bisection argument, see Lemma 1 of Gopi et al. (2023). Finally the change on the continuity assumptions on the $\{f_i\}_{i\in[4]}$ follows by Remark 2.3 of Kannan et al. (1995). ■

Lemma 32 (Self-concordant isoperimetry) *Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be convex and self-concordant, and let $f : \mathbb{R}^d \to \mathbb{R}$ be m-relatively strongly convex in ϕ. Given any partition S_1, S_2, S_3 of $\mathbb{R}^d$,*

$$\frac{\int_{S_3} \exp\left(-f(x)\right) \mathrm{d}x}{\min\left\{\int_{S_1} \exp(-f(x))\mathrm{d}x, \int_{S_2} \exp(-f(x))\mathrm{d}x\right\}} = \Omega\left(\sqrt{m} d_\phi(S_1, S_2)\right),$$

where $d_\phi(S_1, S_2) = \min_{x\in S_1, y\in S_2} d_\phi(x, y)$.

Proof We assume $m = 1$ by rescaling $\phi \leftarrow m\phi$ which results in $d_\phi(S_1, S_2) \leftarrow \sqrt{m} d_\phi(S_1, S_2)$. We first show that without loss of generality, we can assume

$$\max_{i\in\{1,2\}} \frac{\int_{S_i} \exp(-f(x))\mathrm{d}x}{\int \exp(-f(x))\mathrm{d}x} = \Omega(1). \tag{22}$$

To see this, let $S_1^\star$, $S_2^\star$ and $S_3^\star$ be the partition that achieves the minimum of

$$\beta(S_1, S_2, S_3) = \frac{\int_{S_3} \exp\left(-f(x)\right) \mathrm{d}x}{d_\phi(S_1, S_2) \min\left\{\int_{S_1} \exp(-f(x))\mathrm{d}x, \int_{S_2} \exp(-f(x))\mathrm{d}x\right\}}.$$

Let $\delta = d_\phi(S_1^\star, S_2^\star)$. For any $z \in S_3^\star$, let $x \in S_1^\star$ minimize $d_\phi(x, z)$ and let $y \in S_2^\star$ minimize $d_\phi(y, z)$. By the triangle inequality we have

$$d_\phi(x, z) + d_\phi(y, z) \geq \delta$$

and hence $\max(d_\phi(x, z), d_\phi(y, z)) \geq \frac{\delta}{2}$. Consequently we can partition $S_3^\star$ into S_3' and S_3'' such that $d_\phi(S_1^\star, S_3') \geq \frac{\delta}{2}$ and $d_\phi(S_2^\star, S_3'') \geq \frac{\delta}{2}$ by placing each z into an appropriate set. Moreover, we can assume without loss of generality that

$$\frac{\int_{S_3'} \exp\left(-f(x)\right) \mathrm{d}x}{\frac{\delta}{2} \min\left\{\int_{S_1^\star} \exp(-f(x))\mathrm{d}x, \int_{S_2^\star} \exp(-f(x))\mathrm{d}x\right\}} \leq \beta.$$

as otherwise the above is true for S_3''. Thus, $\beta(S_1^\star \cup S_3'', S_2^\star, S_3') \leq \beta(S_1^\star, S_2^\star, S_3^\star)$, proving (22) (else we may halve the measure of S_3). Given (22), it suffices to show that there is a constant C with

$$\begin{aligned} C d_\phi(S_1, S_2) \int \exp(-f(x))\mathbf{1}_{S_1}(x)\mathrm{d}x \int \exp(-f(x))\mathbf{1}_{S_2}(x)\mathrm{d}x \\ \leq \int \exp(-f(x))\mathrm{d}x \int \exp(-f(x))\mathbf{1}_{S_3}(x)\mathrm{d}x. \end{aligned}$$

Using the localization lemma (Lemma 31), letting $f_i = \mathbf{1}_{S_i}$ for $i \in [3]$ and $f_4 = (Cd_\phi(S_1, S_2))^{-1}$,[9] it suffices to prove for every $a, b \in \mathbb{R}^d$ and $\gamma \in \mathbb{R}$,

$$\begin{aligned} &Cd_\phi(S_1, S_2) \int_0^1 \exp\left(\gamma t - \phi((1-t)a + tb)\right) \mathbf{1}_{S_1}((1-t)a + tb)\mathrm{d}t \\ &\cdot \int_0^1 \exp\left(\gamma t - \phi((1-t)a + tb)\right) \mathbf{1}_{S_2}((1-t)a + tb)\mathrm{d}t \\ &\le \int_0^1 \exp\left(\gamma t - \phi((1-t)a + tb)\right) \mathrm{d}t \int_0^1 \exp\left(\gamma t - \phi((1-t)a + tb)\right) \mathbf{1}_{S_3}((1-t)a + tb)\mathrm{d}t. \end{aligned}$$

Redefine $\phi(t) \leftarrow \phi((1-t)a + tb) - \gamma t$ for $t \in \mathbb{R}$, which is a one-dimensional self-concordant function, and redefine $S_i \leftarrow \{t \mid (1-t)a + tb \in S_i\}$ for $i \in [3]$, such that each S_i is a union of intervals. It is straightforward to check that the distance $d_\phi(S_1, S_2)$ only increases under this transformation, because it can only take fewer paths, and each path has the same length (the change in $\sqrt{\phi''}$ is negated by the change in distance traveled by the path).

So, it suffices to consider the special one-dimensional case with $\gamma = 0$, where $d_\phi(x, y) = \int_x^y \sqrt{\phi''(t)}\mathrm{d}t$. We next note that it suffices to consider the case when S_3 is a single interval, i.e. for any $a \le a' \le b' \le b$, we have $S_1 = [a, a']$, $S_2 = [b', b]$, $S_3 = [a', b']$, and wish to show for some constant C

$$\frac{\int_{a'}^{b'} \exp(-\phi(t))\mathrm{d}t}{\int_{a'}^{b'} \sqrt{\phi''(t)}dt} \ge C \frac{\int_a^{a'} \exp(-\phi(t))\mathrm{d}t \int_{b'}^{b} \exp(-\phi(t))\mathrm{d}t}{\int_a^b \exp(-\phi(t))\mathrm{d}t}. \tag{23}$$

When S_3 has multiple intervals, by Theorem 2.6 in Lovász and Simonovits (1993), we show (23) for each interval in S_3 and its adjacent segments in S_1 and S_2, and sum over all such inequalities. By Lemma 7, when ϕ is convex and self-concordant, we have for any $x \in [a, b]$,

$$\frac{\exp(-\phi(x))}{\sqrt{\phi''(x)}} \ge \frac{1}{12} \min\left(\int_a^x \exp(-\phi(t))\mathrm{d}t, \int_x^b \exp(-\phi(t))\mathrm{d}t\right)$$

which combined with $\frac{\int_{a'}^{b'} \exp(-\phi(t))\mathrm{d}t}{\int_{a'}^{b'} \sqrt{\phi''(t)}\mathrm{d}t} \ge \min_{x \in [a', b']} \frac{\exp(-\phi(x))}{\sqrt{\phi''(x)}}$ shows (23). ■

Lemma 33 (TV of induced distributions) *For $x, x' \in \mathbb{R}^d$ with $d_\psi(x, x') \le \frac{1}{4}$, $\left\|\mathcal{D}_x^\varphi - \mathcal{D}_{x'}^\varphi\right\|_{\mathrm{TV}} \le \frac{1}{2}$.*

Proof Let $h = x' - x$ and note that the KL divergence between $\mathcal{D}_x^\varphi$ and $\mathcal{D}_{x'}^\varphi$ may be rewritten as

$$\begin{aligned} D_{\mathrm{KL}}\left(\mathcal{D}_x^\varphi \| \mathcal{D}_{x'}^\varphi\right) &= \mathbb{E}_{y \sim \mathcal{D}_x^\varphi}\left[\log \frac{\mathrm{d}\mathcal{D}_x^\varphi}{\mathrm{d}\mathcal{D}_{x'}^\varphi}(y)\right] = \mathbb{E}_{y \sim \mathcal{D}_x^\varphi}\left[\psi(x') - \psi(x) - \langle h, y\rangle\right] \\ &= \psi(x') - \psi(x) - \langle h, \nabla\psi(x)\rangle. \end{aligned}$$

In the last equation, we used Lemma 3. We recognize that the KL divergence is the Bregman divergence (first-order Taylor approximation) in ψ, and hence letting $x_t = x + th$ for $t \in [0, 1]$ such that $x_0 = x$ and $x_1 = x'$, we continue bounding

$$\begin{aligned} D_{\mathrm{KL}}\left(\mathcal{D}_x^\varphi \| \mathcal{D}_{x'}^\varphi\right) &= \int_0^1 (1-t)\nabla^2\psi(x_t)[h, h]\mathrm{d}t \\ &\le \int_0^1 4(1-t)\nabla^2\psi(x)[h, h]\mathrm{d}t \le \frac{1}{2}. \end{aligned}$$

9. Without loss of generality we can assume S_1 and S_2 are closed (implying S_3 is open) by taking their closures. This implies f_1, f_2 are upper semicontinuous and f_3, f_4 are lower semicontinuous.

The first inequality used that when $d_\psi(x, x') \le \frac{1}{4}$, Lemma 1 shows $\|x_t - x\|_x \le \|x' - x\|_x \le \frac{1}{2}$, so Lemma 2 gives $\nabla^2\psi(x_t) \preceq 4\nabla^2\psi(x)$; the second used $\|h\|_x \le \frac{1}{2}$. Finally by Pinsker's inequality,

$$\left\|\mathcal{D}_x^\varphi - \mathcal{D}_{x'}^\varphi\right\|_{\text{TV}} \le \sqrt{\frac{1}{2} D_{\text{KL}}(\mathcal{D}_x^\varphi \| \mathcal{D}_{x'}^\varphi)} \le \frac{1}{2}.$$

■

Appendix C. Deferred proofs from Section 4

Throughout this section, we assume the setting in Problem 10, and fix some $y \in \mathbb{R}^d$. We provide a sampler for the marginal density π_y (following notation (9)), and denote the component of the density independent of F by γ_y, i.e.

$$\mathrm{d}\gamma_y(x) \propto \exp\left(-\eta\mu\psi(x) - (\psi(x) - \langle x, y\rangle)\right) \mathbf{1}_{\mathcal{X}}(x)\mathrm{d}x. \tag{24}$$

By Lemma 5, γ_y (and hence π_y) is $\frac{1}{\eta}$-strongly logconcave in $\|\cdot\|_{\mathcal{X}}$. Our rejection sampler leverages this fact and the stochastic nature of F to build a rejection sampling scheme similarly to Gopi et al. (2022). For completeness, we state our Algorithm 2 below, and provide the details of its analysis here.

Algorithm 2: $\texttt{InnerLoop}(y, \delta, \mathcal{X}, F, \varphi, \mu)$

Input: $\delta \in (0, \frac{1}{2})$, $y \in \mathbb{R}^d$, $\mathcal{X}, F, \varphi$ in the setting of Problem 10 for $\frac{1}{\eta} \ge 10^4 G^2 \log\frac{1}{\delta}$
Output: Sample within total variation distance δ of

$$\mathrm{d}\pi_y(x) \propto \exp\left(-F(x) - \eta\mu\psi(x) - (\psi(x) - \langle x, y\rangle)\right) \mathbf{1}_{x\in\mathcal{X}}\mathrm{d}x.$$

$u \leftarrow 1, \rho \leftarrow 1$
while $u > \frac{1}{2}\rho$ **do**
    Sample $x_1, x_2 \sim \gamma_y$ defined in (24) independently
    $\rho \leftarrow 1$, $u \sim_{\text{unif.}} [0, 1]$
    Draw $a \in \mathbb{N}$ such that for all $b \in \mathbb{N}$, $\Pr[a \ge b] = \frac{1}{b!}$
    for $b \in [a]$ **do**
        Draw $j_{i,b} \sim \mathcal{I}$ for $i \in [b]$
        $\rho \leftarrow \rho + \prod_{i\in[b]}(f_{j_{i,b}}(x_2) - f_{j_{i,b}}(x_1))$
    end
end
Return: x_1

In order to analyze Algorithm 2, we first state a general result about concentration of Lipschitz functions with respect to a strongly logconcave measure, in general norms. The following is a direct adaptation of standard results on log-Sobolev inequalities contained in Ledoux (1999); Bobkov and Ledoux (2000).

Lemma 34 (Ledoux (1999), Section 2.3 and Bobkov and Ledoux (2000), Proposition 3.1) *Let $X \sim \pi$ for density $\pi : \mathcal{X} \to \mathbb{R}$ which is μ-strongly logconcave in $\|\cdot\|_{\mathcal{X}}$, and let $\ell : \mathcal{X} \to \mathbb{R}$ be G-Lipschitz in $\|\cdot\|_{\mathcal{X}}$. For all $t \ge 0$,*

$$\Pr_{x\sim\pi}\left[\ell(x) \ge \mathbb{E}_\pi[\ell] + t\right] \le \exp\left(-\frac{\mu t^2}{2G^2}\right).$$

In the remainder of the section, let $\tilde{\pi}_y$ be the distribution of the output of Algorithm 2 and recall the target stationary distribution is π_y. When ρ is clear from context, we define $\bar{\rho} := \mathrm{med}(0, \rho, 2)$ to be the truncation of ρ to $[0, 2]$. We also denote the index set drawn on Line 9 by

$$\mathcal{J} := \{j_{i,b}\}_{b \in [a], i \in [b]},$$

when a is clear from context. We first provide the following characterization of $\|\pi_y - \tilde{\pi}_y\|_{\mathrm{TV}}$.

Lemma 35 *Define r_x to be the random variable $\mathbb{E}[\rho \mid x_1 = x]$ (where the expectation is over x_2, a, and the random indices $\mathcal{J}$, and similarly let $\bar{r}_x := \mathbb{E}[\bar{\rho} \mid x_1 = x]$. Then,*

$$\|\pi_y - \tilde{\pi}_y\|_{\mathrm{TV}} \le \mathbb{E}_{x \sim \gamma_y} |r_x - \bar{r}_x|.$$

Proof First, by definition of π_y, we have

$$\pi_y(x) = \frac{\exp(-F(x))\gamma_y(x)}{\int \exp(-F(w))\gamma_y(w)\mathrm{d}w} = \gamma_y(x) \cdot \frac{\exp(-F(x))}{\mathbb{E}_{w \sim \gamma_y} \exp(-F(w))}. \tag{25}$$

Moreover, by definition of the algorithm,

$$\tilde{\pi}_y(x) = \frac{\gamma_y(x) \Pr[u \le \frac{1}{2}\rho \mid x_1 = x]}{\Pr[u \le \frac{1}{2}\rho]} = \frac{\gamma_y(x)\mathbb{E}[\bar{\rho} \mid x_1 = x]}{\mathbb{E}[\bar{\rho}]} \tag{26}$$

where all probabilities and expectations are x_2, a, and $\mathcal{J}$. Furthermore, note that for fixed $b \in [a]$,

$$\mathbb{E}_{\mathcal{J}}\left[\prod_{i \in [b]} (f_{j_{i,b}}(x_2) - f_{j_{i,b}}(x_1))\right] = \left(\mathbb{E}_{j \sim \mathcal{I}}\left[f_j(x_2) - f_j(x_1)\right]\right)^b = (F(x_2) - F(x_1))^b.$$

Hence, taking expectations over a, we have for any fixed x_1, x_2,

$$\begin{aligned}
\mathbb{E}\left[\rho \mid x_1, x_2\right] &= \sum_{b \ge 0} \Pr[a \ge b](F(x_2) - F(x_1))^b \\
&= \sum_{b \ge 0} \frac{1}{b!}(F(x_2) - F(x_1))^b = \exp\left(F(x_2) - F(x_1)\right).
\end{aligned} \tag{27}$$

Next, by combining (25) and (26), we have

$$\begin{aligned}
\|\pi - \tilde{\pi}\|_{\mathrm{TV}} &= \frac{1}{2} \int \left| \frac{\exp(-F(x))}{\mathbb{E}_{w \sim \gamma_y} \exp(-F(w))} - \frac{\mathbb{E}[\bar{\rho} \mid x_1 = x]}{\mathbb{E}[\bar{\rho}]} \right| \gamma_y(x)\mathrm{d}x \\
&= \frac{1}{2}\mathbb{E}_{x \sim \gamma_y}\left[\left| \frac{\exp(-F(x))}{\mathbb{E}_{w \sim \gamma_y} \exp(-F(w))} - \frac{\mathbb{E}[\bar{\rho} \mid x_1 = x]}{\mathbb{E}[\bar{\rho}]} \right|\right].
\end{aligned}$$

By taking expectations over x_2 in (27), and recalling the definitions of $r_x, \bar{r}_x$, we obtain $r_x = \mathbb{E}[\rho \mid x_1 = x] = \exp(-F(x))\mathbb{E}_{x_2 \sim \gamma_y} \exp(F(x_2))$. We thus have

$$\|\pi - \tilde{\pi}\|_{\mathrm{TV}} = \frac{1}{2}\mathbb{E}_{x \sim \gamma_y}\left[\left| \frac{r_x}{\mathbb{E}_{w \sim \gamma_y} r_w} - \frac{\bar{r}_x}{\mathbb{E}_{w \sim \gamma_y} \bar{r}_w} \right|\right].$$

Next, we lower bound $\mathbb{E}_{w \sim \gamma_y} r_w$ as follows. By taking expectations over (27) and using independence of x_1 and x_2, we have that for the random variable $Z = \exp(-F(x))$ where $x \sim \gamma_y$, we have

$$\mathbb{E}_{w \sim \gamma_y} r_w = (\mathbb{E}Z) \cdot \left(\mathbb{E}Z^{-1}\right) \ge 1, \tag{28}$$

where we used Jensen's inequality which implies the last inequality for any nonnegative random variable Z. Finally, combining the above two displays, we derive the desired bound as follows:

$$\begin{aligned}\frac{1}{2}\mathbb{E}_{x\sim\gamma_y}\left[\left|\frac{r_x}{\mathbb{E}_{w\sim\gamma_y}r_w}-\frac{\bar{r}_x}{\mathbb{E}_{w\sim\gamma_y}\bar{r}_w}\right|\right] &\le \frac{1}{2}\mathbb{E}_{x\sim\gamma_y}\left[\left|\frac{r_x}{\mathbb{E}_{w\sim\gamma_y}r_w}-\frac{\bar{r}_x}{\mathbb{E}_{w\sim\gamma_y}r_w}\right|\right] \\ &+\frac{1}{2}\mathbb{E}_{x\sim\gamma_y}\left[\left|\frac{\bar{r}_x}{\mathbb{E}_{w\sim\gamma_y}r_w}-\frac{\bar{r}_x}{\mathbb{E}_{w\sim\gamma_y}\bar{r}_w}\right|\right] \\ &\le \frac{1}{2}\mathbb{E}_{x\sim\gamma_y}\left[|r_x-\bar{r}_x|\right]+\frac{\mathbb{E}_{x\sim\gamma_y}[|\bar{r}_x|]}{2}\cdot\left|\frac{1}{\mathbb{E}_{w\sim\gamma_y}\bar{r}_w}-\frac{1}{\mathbb{E}_{w\sim\gamma_y}r_w}\right| \\ &= \frac{1}{2}\mathbb{E}_{x\sim\gamma_y}\left[|r_x-\bar{r}_x|\right]+\frac{1}{2}\left|1-\frac{\mathbb{E}_{x\sim\gamma_y}\bar{r}_x}{\mathbb{E}_{x\sim\gamma_y}r_x}\right| \\ &\le \frac{1}{2}\mathbb{E}_{x\sim\gamma_y}\left[|r_x-\bar{r}_x|\right]+\frac{1}{2|\mathbb{E}_{x\sim\gamma_y}r_x|}\cdot\mathbb{E}_{x\sim\gamma_y}\left[|r_x-\bar{r}_x|\right] \\ &\le \mathbb{E}_{x\sim\gamma_y}\left[|r_x-\bar{r}_x|\right].\end{aligned}$$

In the second and last inequalities, we use the bound (28). The third line follows since $\bar{r}_x$ is always nonnegative by definition, and the third inequality used convexity of $|\cdot|$. ■

Lemma 35 shows it remains to bound $\mathbb{E}_{x\sim\gamma_y}|r_x-\bar{r}_x|$. Fixing x_1 and x_2, we know ρ and $\bar{\rho}$ as random variables of a and $\mathcal{J}$ are equal, except for the effect of truncating ρ to $[0,2]$. Hence,

$$\mathbb{E}_{x\sim\gamma_y}|r_x-\bar{r}_x|\le\mathbb{E}[|\rho|\mathbf{1}_{\rho\notin[0,2]}]. \tag{29}$$

In the remainder of the section, define

$$H:=\left\lceil 10\log\frac{1}{\delta}\right\rceil. \tag{30}$$

We then let

$$\begin{aligned}\lambda &:= \sum_{b>H}\mathbf{1}_{a\ge b}\prod_{i\in[b]}(f_{j_{i,b}}(x_2)-f_{j_{i,b}}(x_1)),\\ \sigma &:= \sum_{b=0}^{H}\mathbf{1}_{a\ge b}\prod_{i\in[b]}(f_{j_{i,b}}(x_2)-f_{j_{i,b}}(x_1)),\end{aligned} \tag{31}$$

be random variables depending on the choices of $x_1,x_2,a,\mathcal{J}$, where λ captures the effect of the "large" b, and σ captures the effect of the "small" b (where the $b=0$ term is 1 by convention). Since $\rho=\sigma+\lambda$, in light of (29) it suffices to bound $\mathbb{E}[|\sigma|\mathbf{1}_{\rho\notin[0,2]}]+\mathbb{E}[|\lambda|\mathbf{1}_{\rho\notin[0,2]}]$, as

$$\mathbb{E}_{x\sim\gamma_y}|r_x-\bar{r}_x|\le\mathbb{E}[|\rho|\mathbf{1}_{\rho\notin[0,2]}]\le\mathbb{E}[|\sigma|\mathbf{1}_{\rho\notin[0,2]}]+\mathbb{E}[|\lambda|\mathbf{1}_{\rho\notin[0,2]}]. \tag{32}$$

We next provide bounds on λ and σ, using small modifications to Gopi et al. (2022).

Lemma 36 *For λ defined in (31),*

$$\mathbb{E}\left[|\lambda|\mathbf{1}_{\rho\notin[0,2]}\right]\le\frac{\delta}{4}.$$

Proof Clearly, it suffices to show $\mathbb{E}|\lambda|\le\frac{\delta}{4}$. Define random variables,

$$\Delta_i:=|f_i(x_2)-f_i(x_1)|,\ \Delta:=\mathbb{E}_{i\sim\mathcal{I}}\Delta_i,$$

whose randomness comes from $x_1, x_2 \sim \gamma_y$. By definition,

$$\mathbb{E}|\lambda| = \sum_{b>H} \frac{1}{b!} \mathbb{E}_{x_1,x_2\sim\gamma}[\Delta]^B.$$

Define $\Phi(t) := \sum_{b>H} \frac{t^b}{b!}$. For $H = \lceil 10 \log \frac{1}{\delta} \rceil$, it is straightforward to check $\Phi(t) \le \frac{\delta}{16}$ for any $|t| \le 1$, and for all nonnegative t, $\Phi(t) \le \exp(t)$. Hence, letting p_Δ be the density of Δ,

$$\begin{aligned} \mathbb{E}|\lambda| &\le \frac{\delta}{16} + \mathbb{E}[\mathbf{1}_{\Delta>1} e^{\Delta}] \le \frac{\delta}{16} + \int_1^\infty \exp\left(\lceil \Delta \rceil\right) p_\Delta(\Delta) \mathrm{d}\Delta \\ &\le \frac{\delta}{16} + \sum_{k\ge 1} \exp(k+1) \Pr_{x_1,x_2\sim\gamma}[\Delta \ge k]. \end{aligned} \tag{33}$$

It now suffices to bound on $\Pr[\Delta \ge k]$. Define a function $h_{x_1,x_2}(k) := \Pr_{i\sim\mathcal{I}}[|f_i(x_1) - f_i(x_2)| \ge k]$. Since each f_i is G-Lipschitz, and γ_y is $\frac{1}{12\eta}$-strongly logconcave in by Lemma 5, by Lemma 34:

$$\mathbb{E}_{x_1,x_2}[h_{x_1,x_2}(k)] = \Pr_{x_1,x_2,i\sim\mathcal{I}}[|f_i(x_1) - f_i(x_2)| \ge k] \le 4 \exp\left(-\frac{k^2}{96\eta G^2}\right),$$

and so by Markov's inequality we have

$$\Pr_{x_1,x_2}[h_{x_1,x_2}(k) \ge e^{-t}] \le 4 \exp\left(t - \frac{k^2}{96\eta G^2}\right). \tag{34}$$

For fixed x_1, x_2, as each f_i is G-Lipschitz in $\|\cdot\|_{\mathcal{X}}$, $|f_i(x_1) - f_i(x_2)| \le G \|x_1 - x_2\|_{\mathcal{X}}$, and hence

$$\mathbb{E}_{i\sim\mathcal{I}}[|f_i(x_1) - f_i(x_2)|] \le \min_{k\ge 0} k + h_{x_1,x_2}(k) \cdot G \|x_1 - x_2\|_{\mathcal{X}}.$$

This then shows that if for some k, $h_{x_1,x_2}(k) \le \exp(-\frac{k^2}{192\eta G^2})$,

$$\mathbb{E}_{i\sim\mathcal{I}}[|f_i(x_1) - f_i(x_2)|] \le k + \exp\left(-\frac{k^2}{192\eta G^2}\right) \cdot G \|x_1 - x_2\|_{\mathcal{X}},$$

which implies via (34) that

$$\begin{aligned} &\Pr_{x_1,x_2}\left[\Delta \ge k + \exp\left(-\frac{k^2}{192\eta G^2}\right) \cdot G \|x_1 - x_2\|_{\mathcal{X}}\right] \\ &\le \Pr_{x_1,x_2}\left[h_{x_1,x_2}(k) \ge \exp\left(-\frac{k^2}{192\eta G^2}\right)\right] \le 4 \exp\left(-\frac{k^2}{192\eta G^2}\right). \end{aligned} \tag{35}$$

Further, since $\|x_1 - \mathbb{E}x_1\|_{\mathcal{X}}$ is a 1-Lipschitz function in x_1 with a nonnegative mean, by Lemma 34,

$$\Pr\left[\|x_1 - x_2\|_{\mathcal{X}} \ge k\right] \le 2 \Pr\left[\|x_1 - \mathbb{E}x_1\|_{\mathcal{X}} \ge k\right] \le 2 \exp\left(-\frac{k^2}{96\eta G^2}\right). \tag{36}$$

Combining (35) and (36),

$$\begin{aligned} \Pr_{x_1,x_2}[\Delta \ge 2k] &= \Pr_{x_1,x_2}\left[\Delta \ge 2k \wedge \|x_1 - x_2\|_{\mathcal{X}} \ge \frac{k}{G}\right] + \Pr_{x_1,x_2}\left[\Delta \ge 2k \wedge \|x_1 - x_2\|_{\mathcal{X}} \le \frac{k}{G}\right] \\ &\le 2 \exp\left(-\frac{k^2}{96\eta G^2}\right) + \Pr_{x_1,x_2}\left[\Delta \ge k + \exp\left(-\frac{k^2}{192\eta G^2}\right) G \|x_1 - x_2\|_{\mathcal{X}}\right] \\ &\le 6 \exp\left(-\frac{k^2}{192\eta G^2}\right). \end{aligned} \tag{37}$$

Plugging (37) into (33), and using $\eta^{-1} \geq 10^4 G^2 \log \frac{1}{\delta}$, we have the desired

$$\mathbb{E}(|\lambda| \mathbf{1}_{\rho \notin [0,2]}) \leq \frac{\delta}{16} + \sum_{k=1}^{\infty} 6 \exp\left(k - \frac{k^2}{768 \eta G^2}\right) \leq \frac{\delta}{4}.$$

■

Lemma 37 *For σ defined in (31),*

$$\mathbb{E}\left[|\sigma| \mathbf{1}_{\rho \notin [0,2]}\right] \leq \frac{\delta}{4}.$$

Proof We begin by bounding, analogously to (33),

$$\mathbb{E}[|\sigma| \mathbf{1}_{\rho \notin [0,2]}] \leq 2^H \Pr[\rho \notin [0,2]] + \sum_{k \geq 1} \Pr\left[|\sigma| > 2^{kH}\right] 2^{(k+1)H}. \tag{38}$$

Recall when $a \leq H$, $|\mathcal{J}| \leq \frac{1}{2} H^2$. By a union bound over Lemma 34,

$$\Pr_{x_1, x_2}\left[|f_i(x_1) - f_i(x_2)| \geq \frac{2^k}{3} \ \forall i \in \mathcal{J}\right] \leq H^2 \exp\left(-\frac{4^k}{864 \eta G^2}\right).$$

If for each $i \in \mathcal{J}$, $|f_i(x_1) - f_i(x_2)| \leq \frac{2^k}{3}$, we have for $k \geq 1$

$$|\sigma| = \sum_{b=0}^{H} \mathbf{1}_{a \geq b} \prod_{i \in [b]} (f_{j_{i,b}}(x_2) - f_{j_{i,b}}(x_1)) \leq 1 + \sum_{b=1}^{H} \left(\frac{2^k}{3}\right)^b \leq 2^{kH},$$

which implies that $\Pr[|\sigma| \geq 2^{kH}] \leq H^2 \exp(-\frac{4^k}{864 \eta G^2})$ and hence using our choice of $\eta \leq \frac{1}{500 G^2 H}$,

$$\begin{aligned} \sum_{k=1}^{\infty} 2^{(k+1)H} \Pr\left[|\sigma| > 2^{kH}\right] &\leq \sum_{k=1}^{\infty} 2^{(k+1)H} H^2 \exp\left(-\frac{4^k}{864 \eta G^2}\right) \\ &\leq \sum_{k=1}^{\infty} 2^{4kH} \exp(-2 \cdot 4^k H) \leq \sum_{k=1}^{\infty} 2^{-kH} \leq \frac{\delta}{8}. \end{aligned} \tag{39}$$

It remains to bound $\Pr[\rho \notin [0,2]]$. Recall $\Pr[a > H] \leq \frac{1}{H!}$ so since $a \leq H \implies \sigma = \rho$, $\Pr[\rho \notin [0,2]] \leq \frac{1}{H!} + \Pr[\sigma \notin [0,2]]$. Next, by a union bound over Lemma 34 and $\frac{1}{2} H^2$ indices in $\mathcal{J}$,

$$\Pr_{x_1, x_2}\left[|f_i(x_1) - f_i(x_2)| \geq \frac{1}{2} \ \forall i \in \mathcal{I}\right] \leq 2H^2 \exp\left(-\frac{1}{384 \eta G^2}\right).$$

Under the event that $|f_i(x_1) - f_i(x_2)| \leq \frac{1}{2}$ for all $i \in \mathcal{I}$, $0 \leq \sigma \leq 2$ by definition. Hence we know $\Pr[\sigma \notin [0,2]] \leq 2H^2 \exp(-\frac{1}{384 \eta G^2})$ and by our setting that $H > 10 \log \frac{1}{\delta}$, we have

$$\Pr[\rho \notin [0,2]] \cdot 2^H \leq 2^H \left(2H^2 \exp\left(-\frac{1}{384 \eta G^2}\right) + \frac{1}{H!}\right) \leq \frac{\delta}{8}. \tag{40}$$

Combining (38), (39) and (40) completes the proof. ■

Putting together these pieces, we finally obtain the following guarantee on Algorithm 2.

Proposition 38 *If $\eta^{-1} = \Omega(G^2 \log \frac{1}{\delta})$ for an appropriate constant, Algorithm 2 obtains total variation distance to π_y at most δ. In expectation, Algorithm 2 queries $O(1)$ random f_i and $O(1)$ samples from γ_y.*

Proof The total variation distance bound comes from combining Lemma 35, (32), Lemma 36, and Lemma 37. Further, the end probability of each "while" loop is $\Pr[u \le \frac{1}{2}\rho] = \mathbb{E}[\bar{\rho}] = \mathbb{E}_{x\sim\gamma}\bar{r}_x \ge \mathbb{E}_{x\sim\gamma_y} r_x - \mathbb{E}_{x\sim\gamma_y}|\bar{r}_x - r_x|$. We proved in (28) that $\mathbb{E}_{x\sim\gamma_y} r_x \ge 1$, and combining (32), Lemma 36 and Lemma 37, shows $\mathbb{E}_{x\sim\gamma_y}|\bar{r}_x - r_x| \le \delta \le \frac{1}{2}$. Hence the expected number of loops is ≤ 2, and each loop draws two samples from γ_y, and $O(1)$ many f_i in expectation since $\mathbb{E}a^2 = O(1)$. ∎

Finally, we prove the following (exponentially) warm start result from the main body.

Lemma 39 *The density $\mathrm{d}\nu(x) \propto \exp(-\eta\mu\psi(x))\mathbf{1}_{\mathcal{X}}(x)\mathrm{d}x$ is $\exp(GD)$-warm for π defined in (7).*

Proof Note that for all $x, w \in \mathcal{X}$, $|F(x) - F(w)| \le GD$. Further recall $\pi \propto \exp(-F)\nu$. We conclude by

$$\frac{\exp\left(-F(x)\right)\nu(x)}{\int_{\mathcal{X}} \exp(-F(w))\nu(w)\mathrm{d}w} \cdot \frac{\int_{\mathcal{X}} \nu(w)\mathrm{d}w}{\nu(x)} = \frac{\int_{\mathcal{X}} \nu(w)\mathrm{d}w}{\int_{\mathcal{X}} \exp(F(x) - F(w))\nu(w)\mathrm{d}w} \le \exp\left(GD\right).$$

∎

Appendix D. Information-theoretic lower bound

In this section, we show that prior information-theoretic lower bounds from Duchi et al. (2015) and Gopi et al. (2022) can be straightforwardly extended to the settings studied by this paper to show that the value oracle complexities used by our algorithms in Sections 3 and A are near-optimal. We first recall some notation from prior work and summarize previous results we will leverage.

Setup. We consider the setting of stochastic optimization where there is a distribution over distributions $\{\mathcal{P}_v\}_v$ indexed by v. An index v is randomly selected, and we consider algorithms interacting with $\mathcal{P}_v$ in one of two different ways. Letting $k \in \mathbb{N}$ and $\mathcal{X} \subset \mathbb{R}^d$, Duchi et al. (2015) defined a family of algorithms $\mathbb{A}_k$ such that $\mathcal{A} \in \mathbb{A}_k$ can (adaptively) query a sequence of k values $f(x; s)$ where $x \in \mathcal{X}$ and s is a fresh random sample from $\mathcal{P}_v$. The follow-up work Gopi et al. (2022) defined another family of algorithms $\mathbb{B}_k$ which takes as input a dataset $\mathcal{D} = \{s_i\}_{i\in[n]}$ and can (adaptively) query a sequence of k values $f(x; s)$ where $x \in \mathcal{X}$ and $s \in \mathcal{D}$. These algorithm families model the SCO and ERM problems stated in Problem 22, without the privacy requirement. In a slight abuse of notation, we denote the output of an algorithm $\mathcal{A} \in \mathbb{A}_k \cup \mathbb{B}_k$ in a SCO or ERM problem corresponding to a distribution $\mathcal{P}$ by $\mathcal{A}(\mathcal{P})$, where $\mathcal{A} \in \mathbb{B}_k$ also depends on the dataset received.

Both Duchi et al. (2015); Gopi et al. (2022) let v be drawn uniformly at random from $\mathcal{V} := \{-1, 1\}^d$ and let

$$\mathcal{P}_v := \mathcal{N}\left(\kappa v, \sigma^2 \mathbf{I}_d\right), \; f(x; s) := \langle s, x\rangle$$

for parameters κ, σ to be chosen. We fix this notation throughout this section. For any algorithm $\mathcal{A} \in \mathbb{A}_k \cup \mathbb{B}_k$ corresponding to a set $\mathcal{X}$ and a distribution $\mathcal{P}$, we define the optimality gap

$$\epsilon_k(\mathcal{A}, \mathcal{X}, \mathcal{P}) := \mathbb{E}\left[\mathbb{E}_{s\sim\mathcal{P}} f(\mathcal{A}(\mathcal{P}); s)\right] - \min_{x\in\mathcal{X}} \mathbb{E}_{s\sim\mathcal{P}} f(x; s),$$

where the first outer expectation is over any randomness in $\mathcal{A}$, as well as in the samples used. We also define the minimax risk over a family of distributions P,

$$\epsilon_k^\star(\mathbb{A}_k \cup \mathbb{B}_k, P, \mathcal{X}) := \inf_{\mathcal{A}\in\mathbb{A}_k\cup\mathbb{B}_k} \sup_{\mathcal{P}\in P} \epsilon_k\left(\mathcal{A}, \mathcal{P}, \mathcal{X}\right).$$

For $p \in [1,2]$, we let $P_{G,p}$ denote the family of distributions $\mathcal{P}$ over vectors s such that

$$\mathbb{E}_{s\sim\mathcal{P}} \|s\|_q^2 \le G^2, \text{ where } \frac{1}{p} + \frac{1}{q} = 1.$$

Our lower bounds in this section will be on $\epsilon_k^\star(\mathbb{A}_k \cup \mathbb{B}_k, P_{G,p}, \mathcal{X})$, where $\mathcal{X}$ is a scaled ℓ_p ball. The family $P_{G,p}$ induces random linear functions $\langle s, \cdot\rangle$ with gradient s, and hence $\mathcal{P} \in P_{G,p}$ implies that the induced function $\mathbb{E}_{s\sim\mathcal{P}} \langle s, \cdot\rangle$ has a bounded-variance gradient oracle in the ℓ_p norm via queries to $\mathcal{P}$. We use the following facts from prior work in our proofs.

Lemma 40 (Section 5.1, Duchi et al. (2015)) *Let $\mathcal{X}$ be the ℓ_p ball of diameter D for $p \in [1,2]$. For any $v \in \mathcal{V}$ and $x \in \mathcal{X}$, letting $x_v^\star := \min_{x\in\mathcal{X}} \mathbb{E}_{s\sim\mathcal{P}_v} f(x;s)$, and letting $\mathbf{1}(\text{sign}(a) = \text{sign}(b))$ be the 0-1 function which is 1 if and only if the signs of a and b agree,*

$$\mathbb{E}_{s\sim\mathcal{P}_v}\left[f(x;s)\right] - \mathbb{E}_{s\sim\mathcal{P}_v}\left[f(x_v^\star;s)\right] \ge \frac{(1-\frac{1}{p})\kappa D}{2d^{\frac{1}{p}}} \sum_{j\in[d]} \mathbf{1}\left(\text{sign}(x_j) = \text{sign}(v_j)\right).$$

Lemma 40 shows that it suffices to lower bound the expected Hamming distance between the signs of an estimate x and a randomly sampled $-v$. Such a lower bound was given in Duchi et al. (2015); Gopi et al. (2022) for estimates returned by $\mathcal{A} \in \mathbb{A}_k \cup \mathbb{B}_k$ via information-theoretic arguments.

Lemma 41 (Section 5.1, Duchi et al. (2015), Lemma 7.4, Gopi et al. (2022)) *Let $\mathcal{X}$ be the ℓ_p ball of diameter D, and let $\mathcal{A} \in \mathbb{A}_k \cup \mathbb{B}_k$ be parameterized by $\mathcal{X}$ and $\mathcal{P}_v$. Then*

$$\mathbb{E}_{v\sim_{\text{unif.}}\mathcal{V}}\left[\sum_{j\in[d]} \mathbf{1}(\text{sign}(\mathcal{A}(\mathcal{P}_v)_j) = \text{sign}(v_j))\right] \ge \frac{d}{2}\left(1 - \frac{\kappa\sqrt{k}}{\sigma\sqrt{d}}\right).$$

To lower bound the oracle query complexity of our sampler we use the following standard result.

Lemma 42 (De Klerk and Laurent (2018), Corollary 1) *Let $\mathcal{X} \subset \mathbb{R}^d$ be compact and convex, $f : \mathcal{X} \to \mathbb{R}$ be convex, $\tau > 0$, and π be the density over $\mathcal{X}$ proportional to $\exp(-\tau f)$. Then,*

$$\mathbb{E}_{x\sim\pi}[f(x)] - \min_{x\in\mathcal{X}} f(x) \le \frac{d}{\tau}.$$

Lower bounds. We now state three lower bounds generalizing results from Duchi et al. (2015); Gopi et al. (2022). Our results follow straightforwardly from Lemmas 40, 41, and 42 with appropriate parameters.

Proposition 43 (Minimax risk lower bound, $P_{G,p}$) *Let $G, D > 0$, and let $p \in [1,2]$, $q \ge 2$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Let $\mathcal{X}$ be the ℓ_p ball of diameter D. Then,*

$$\epsilon_k^\star\left(\mathbb{A}_k \cup \mathbb{B}_k, P_{G,p}, \mathcal{X}\right) = \Omega\left(GD \max\left(1 - \frac{1}{p}, \frac{1}{\log d}\right) \min\left(1, \sqrt{\frac{d}{k\log d}}\right)\right).$$

Proof Throughout the proof, let $\kappa = \frac{\sigma\sqrt{d}}{2\sqrt{k}}$, and let

$$\sigma = \frac{Gd^{-\frac{1}{q}}}{\sqrt{\frac{d}{k} + 4\log d}}. \quad (41)$$

By well-known bounds on the expected maximum of d standard Gaussians, we have

$$\begin{aligned}\mathbb{E}_{s\sim\mathcal{P}_v}\left[\|s\|_q^2\right] &\le 2\kappa^2\|v\|_q^2 + 2\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2\mathbf{I}_d)}\left[\|u\|_q^2\right]\\ &\le 2\kappa^2 d^{\frac{2}{q}} + 2d^{\frac{2}{q}}\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2\mathbf{I}_d)}\left[\|u\|_\infty^2\right]\\ &\le \sigma^2 d^{\frac{2}{q}}\left(\frac{d}{k} + 4\log d\right) \le G^2.\end{aligned}$$

Hence, $\mathcal{P}_v \in P_{G,p}$ for all $v \in \mathcal{V}$, so it suffices to lower bound $\epsilon_k(\mathcal{A}, \mathcal{P}_v, \mathcal{X})$. Combining Lemmas 40 and 41 with our choices of parameters,

$$\epsilon_k(\mathcal{A}, \mathcal{P}_v, \mathcal{X}) \ge \frac{(1-\frac{1}{p})\kappa D d^{1-\frac{1}{p}}}{8} = \Omega\left(GD\left(1-\frac{1}{p}\right)\min\left(1, \sqrt{\frac{d}{k\log d}}\right)\right).$$

The conclusion then follows because for $p \le 1 + \frac{1}{\log d}$, choosing a larger value of p only affects problem parameters by constant factors by norm conversions. ■

We give a slight extension of Proposition 43 for the family $\overline{P}_{G,p}$ of distributions over linear functions $\langle s, \cdot\rangle$, where s is required to satisfy $\|s\|_q \le G$ with probability 1, by simply truncating a draw from $\mathcal{P}_v$. This family is compatible with the setting in Problem 22.

Corollary 44 (Minimax risk lower bound, $\overline{P}_{G,p}$) *In the setting of Proposition 43,*

$$\epsilon_k^\star\left(\mathbb{A}_k \cup \mathbb{B}_k, \overline{P}_{G,p}, \mathcal{X}\right) = \Omega\left(GD\max\left(1-\frac{1}{p}, \frac{1}{\log d}\right)\min\left(1, \sqrt{\frac{d}{k\log(dk)}}\right)\right).$$

Proof We define a distribution $\overline{\mathcal{P}}_v$ as follows: first $s \sim \mathcal{P}_v$, and then if $\|s\|_q \ge G$, we set $s \leftarrow 0$. By adjusting the logarithmic term in (41) to be $O(\log(dk))$, with probability at most $\text{poly}((dk)^{-1})$, all k draws from $\mathcal{P}_v$ and $\overline{\mathcal{P}}_v$ used are identical by a union bound. Further, due to problem constraints the function error is always at most GD. So, the risk is affected by at most $GD \cdot \text{poly}((dk)^{-1})$. ■

Corollary 44 shows that when β in Assumption 24 is polynomially bounded, the value oracle complexities used by Theorem 25 for both DP-SCO and DP-ERM are optimal up to logarithmic factors for the expected excess risk bounds they produce, even without the requirement of privacy. Finally, we show that the value oracle complexity of our sampler in Theorem 16 is also near-optimal.

Corollary 45 *In the setting of Proposition 43, let $r : \mathcal{X} \to \mathbb{R}$ be 1-strongly convex in $\|\cdot\|_p$ with additive range $O(D^2\min(\log d, \frac{1}{p-1}))$. Let $\mathcal{I}$ be a distribution over i such that all $f_i : \mathcal{X} \to \mathbb{R}$ are G-Lipschitz in $\|\cdot\|_p$, and let $F := \mathbb{E}_{i\sim\mathcal{I}} f_i$. No algorithm using $o(\frac{G^2}{\mu}\log^{-4} d)$ value oracle queries to some f_i samples within total variation*

$$o\left(\min\left(\frac{1}{\log d}, \sqrt{\frac{d}{k\log^3(dk)}}\right)\right)$$

of the density proportional to $\exp(-F - \mu r(x))\mathbf{1}_{\mathcal{X}}(x)$.

Proof Assume for contradiction that $\mathcal{A}$ is an algorithm satisfying the stated criterion using $k = o(\frac{G^2}{\mu}\log^{-4} d)$ value oracle queries, and let F be minimized by $x^\star \in \mathcal{X}$. We choose

$$\mu = \frac{d}{D^2\min(\log d, \frac{1}{p-1})}.$$

Lemma 42 then shows that the sampled x satisfies

$$\mathbb{E}_{x\sim\mathcal{A}}\left[F(x)\right] - F(x^\star) \le \mu\left(r(x^\star) - r(x)\right) + d + GD \cdot o\left(\min\left(\frac{1}{\log d}, \sqrt{\frac{d}{k\log^3(dk)}}\right)\right)$$
$$= O(d) + o\left(\frac{GD}{\log d}\min\left(1, \sqrt{\frac{d}{k\log(dk)}}\right)\right).$$

For the given values of k and μ, this contradicts Corollary 44. ■

Corollary 45 implies that for samplers with value query complexity depending polylogarithmically on the total variation distance, $\frac{G^2}{\mu}$ queries are required (up to polylogarithmic factors). This applies to the setting of our sampler in Theorem 16; we also note that the LLT-based regularizers we use in our ℓ_p applications (Appendix A.2) satisfy the additive range bound in Corollary 45.

Appendix E. Lower bound on the range of $\psi_{1,1}$

In this section, we provide a lower bound on the range of $\psi_{1,1}$ (10) which grows with the dimension d, demonstrating non-scale invariance of our family of LLTs. Recall that $\psi_{1,1}(x)$ is defined by

$$\psi_{1,1}(x) := \log\left(\int \exp\left(\langle x, y\rangle - \|y\|_\infty^2\right) \mathrm{d}y\right).$$

Lemma 46 *The additive range of $\psi_{1,1}$ over $\{x \in \mathbb{R}^d \mid \|x\|_1 \le 1\}$ is $\Omega(\sqrt{d})$.*

Proof Throughout the proof denote for simplicity $\psi := \psi_{1,1}$ and let

$$\mathcal{D}_x^\varphi(y) \propto \exp\left(\langle x, y\rangle - \|y\|_\infty^2\right).$$

Then, following (11), we can write $\psi(x) - \psi(0)$ as

$$\psi(x) - \psi(0) = \log\left[\mathbb{E}_{y\sim\mathcal{D}_0^\varphi}\exp(\langle x, y\rangle)\right],$$

where $\mathcal{D}_0^\varphi \propto \exp(-\|y\|_\infty^2)$. Let π be the probability density on $\mathbb{R}_{\ge 0}$ such that

$$\mathrm{d}\pi(r) \propto r^{d-1}\exp(-r^2)\mathrm{d}r.$$

Here, $\mathrm{d}\pi(r)$ is the density of the scalar quantity $r = \|y\|_\infty$ for $y \sim \mathcal{D}_0^\varphi$. Note that the distribution of y conditioned on $\|y\|_\infty = r$ is uniform over the surface of the ℓ_∞ ball, where one random coordinate is set to $\pm r$, and the remaining coordinates are uniform on a $d-1$ dimensional hypercube with side length r. We denote this distribution as $\mathcal{P}_r$, and write

$$\mathbb{E}_{y\sim\mathcal{D}_0^\varphi}\exp(\langle x, y\rangle) = \mathbb{E}_{r\sim\pi}\left[\mathbb{E}_{y\sim\mathcal{P}_r}\exp(\langle x, y\rangle)\right]$$
$$= \mathbb{E}_{r\sim\pi}\left[\frac{1}{d}\sum_{i^\star\in[d]}\frac{1}{2}\sum_{y_{i^\star}\in\{-r,r\}}\exp(x_{i^\star}y_{i^\star})\prod_{i\neq i^\star}\int_{-r}^{r}\frac{1}{2r}\exp(x_iy_i)\mathrm{d}y_i\right].$$

Let $x = e_1$ and $g_{i^\star}^{(r)} = \exp(x_{i^\star}r)\prod_{i\neq i^\star}\int_{-r}^{r}\frac{1}{2r}\exp(x_iy_i)\mathrm{d}y_i$. Then,

$$\mathbb{E}_{y\sim\mathcal{D}_0^\varphi}\exp(\langle x, y\rangle) \ge \frac{1}{2d}\sum_{i^\star\in[d]}\mathbb{E}_{r\sim\pi(r)}g_{i^\star}^{(r)}$$

since this drops terms where $y_{i^\star} = -r$. When $i^\star = 1$, we have $g_{i^\star}^{(r)} = \exp(r)$. When $i^\star \neq 1$, we have

$$g_{i^\star}^{(r)} = \int_{-r}^{r} \frac{1}{2r} \exp(y_1) \mathrm{d}y_1 = \frac{1}{2r} \left(\exp(r) - \exp(-r) \right).$$

Now, consider $r_1 = \sqrt{\frac{d-1}{2}}$. For any $r \leq r_1$, $\frac{\mathrm{d}}{\mathrm{d}r}[(d-1)\log r - r^2] = \frac{d-1}{r} - 2r \geq 0$. Thus, we have

$$I := \int_0^{\frac{1}{2}r_1} \exp((d-1)\log r - r^2)\mathrm{d}r \leq \int_{\frac{1}{2}r_1}^{r_1} \exp((d-1)\log r - r^2)\mathrm{d}r. \tag{42}$$

Letting $Z := \int_0^\infty \exp((d-1)\log r - r^2)\mathrm{d}r$, (42) shows that

$$\int_{\frac{1}{2}r_1}^{\infty} \exp((d-1)\log r - r^2)\mathrm{d}r = Z - I \geq Z - \frac{1}{2}Z = \frac{1}{2}Z.$$

Then, for all $i^\star \in [d]$,

$$\begin{aligned}
\mathbb{E}_{r\sim\pi} g_{i^\star} &= \frac{\int_0^\infty \exp((d-1)\log r - r^2) g_{i^\star}^{(r)} \mathrm{d}r}{Z} \\
&\geq \frac{\int_{\frac{1}{2}r_1}^\infty \exp((d-1)\log r - r^2) g_{i^\star}^{(r)} \mathrm{d}r}{Z} \\
&\geq \frac{2\int_{\frac{1}{2}r_1}^\infty \exp((d-1)\log r - r^2) g_{i^\star}^{(r)} \mathrm{d}r}{\int_{\frac{1}{2}r_1}^\infty \exp((d-1)\log r - r^2)\mathrm{d}r} \\
&\geq 2 \min_{r \geq r_1} \exp(r - \log(4r)) = 2\exp(r_1 - \log(4r_1)).
\end{aligned}$$

The fourth step follows from $g_{i^\star}^{(r)} \geq \frac{1}{4r}\exp(r)$ for $r \geq r_1$. The last step follows from $r - \log 4r$ increases on $r \geq r_1$. Combining with $\mathbb{E}_{y\sim\mathcal{P}_0} \exp(\langle x, y\rangle) \geq \frac{1}{2d}\sum_{i^\star\in[d]} \mathbb{E}_{r\sim\pi(r)} g_{i^\star}$,

$$\psi(x) - \psi(0) = \log \mathbb{E}_{y\sim\mathcal{P}_0} \exp(\langle x, y\rangle) \geq \log\left(\frac{d-1}{d}\exp(r_1 - \log(4r_1))\right) = \Omega(\sqrt{d}).$$

■