Sampling from Log-Concave Distributions with Infinity-Distance Guarantees

Oren Mangoubi
Worcester Polytechnic Institute

Nisheeth K. Vishnoi
Yale University

Abstract

For a d-dimensional log-concave distribution $\pi(\theta) \propto e^{-f(\theta)}$ constrained to a convex body K, the problem of outputting samples from a distribution ν which is ε-close in infinity-distance $\sup_{\theta \in K} |\log \frac{\nu(\theta)}{\pi(\theta)}|$ to π arises in differentially private optimization. While sampling within total-variation distance ε of π can be done by algorithms whose runtime depends polylogarithmically on $\frac{1}{\varepsilon}$, prior algorithms for sampling in ε infinity distance have runtime bounds that depend polynomially on $\frac{1}{\varepsilon}$. We bridge this gap by presenting an algorithm that outputs a point ε-close to π in infinity distance that requires at most $\mathrm{poly}(\log \frac{1}{\varepsilon}, d)$ calls to a membership oracle for K and evaluation oracle for f, when f is Lipschitz. Our approach departs from prior works that construct Markov chains on a $\frac{1}{\varepsilon^2}$-discretization of K to achieve a sample with ε infinity-distance error, and present a method to directly convert continuous samples from K with total-variation bounds to samples with infinity bounds. This approach also allows us to obtain an improvement on the dimension d in the running time for the problem of sampling from a log-concave distribution on polytopes K with infinity distance ε, by plugging in TV-distance running time bounds for the Dikin Walk Markov chain.

1 Introduction

The problem of sampling from a log-concave distribution is as follows: For a convex body $K \subseteq \mathbb{R}^d$ and a convex function $f : K \to \mathbb{R}$, output a sample θ from the distribution $\pi(\theta) \propto e^{-f(\theta)}$. This is a basic problem in computer science, statistics, and machine learning, with applications to optimization and integration [1, 29], Bayesian statistics [39], reinforcement learning [6], and differential privacy [32, 20, 2, 24]. Sampling exactly from π is known to be computationally hard for most interesting cases of K and f [16] and, hence, the goal is to output samples from a distribution ν that is at a small (specified) "distance" to π. For applications such as computing the integral of π, bounds in the total variation (TV) distance [1] or KL divergence (which implies a TV bound) are sufficient. In applications such as computing the expectation of a Lipschitz function with respect to π, Wasserstein distance may also be sufficient. In differentially private optimization [32, 20, 2, 19, 24], one requires bounds on the stronger infinity-distance – $\mathrm{d}_\infty(\nu, \pi) := \sup_{\theta \in K} |\log \frac{\nu(\theta)}{\pi(\theta)}|$ – to guarantee pure differential privacy, and TV, KL, or Wasserstein bounds are insufficient; see [14].

Pure differential privacy (DP) is the strongest notion of DP and has been extensively studied (see e.g. the survey [14]). It has advantages over weaker notions of differential privacy. E.g., when privacy of "groups" of individuals (rather than just single individuals) must be preserved, any mechanism which is (pure) ε-DP (with respect to single individuals), is also $k\varepsilon$-DP with respect to subsets of k individuals. Motivated by applications to differential privacy, we study the problem of designing efficient algorithms to output samples from a distribution ν which is ε-close in d_∞ to π.

Related works. Several lines of work have designed Markov chains that generate samples from distributions that are close to a given log-concave distribution. These results differ in both their as-

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

sumptions on the log-density and its support, as well as the distance used to measure closeness. One line of work includes bounds for sampling from a log-concave distribution on a compactly supported convex body within TV distance $O(\delta)$, including results with running time that is polylogarithmic in $\frac{1}{\delta}$ [1, 30, 29, 34] (as well as other results which give a running time bound that is polynomial in $\frac{1}{\delta}$ [18, 17, 5, 4]). In addition to assuming access to a value oracle for f, some Markov chains just need access to a membership oracle for K [1, 30, 29], while others assume that K is a given polytope: $\{\theta \in \mathbb{R}^d : A\theta \leq b\}$ [23, 33, 35, 34, 26]. They often also assume that K is contained in a ball of radius R and contains a ball of smaller radius r. Many of these results assume that the target log-concave distribution satisfies a "well-rounded" condition which says that the variance of the target distribution is $\Theta(d)$ [30, 29], or that it is in isotropic position (all eigenvalues of its covariance matrix are $\Theta(1)$) [25]; when applied to log-concave distributions that are not well-rounded or isotropic, these results require a "rounding" pre-processing procedure to find a linear transformation which makes the target distribution well-rounded or isotropic. Finally, it is often assumed that the function f is such that f is L-Lipschitz or β-smooth [34], including works handling the widely-studied special case when f is uniform on K where $L = \beta = 0$ (see e.g. [28, 23, 33, 35, 26, 8, 21]).

Another line of work gives sampling algorithms with bounds on the distance to the target density π in terms of Wasserstein distance [13, 11], KL divergence [40, 12], and Renyi divergence [36]. In contrast to works which assume access to an oracle for the value of f, many of these results instead assume access to an oracle for the gradient of f and require the log-density to be L-Lipschitz or β-smooth on all of $\mathbb{R}^d$ (or on, e.g., a cube containing K) for some $L, \beta > 0$. However, as noted earlier, bounds in the Wasserstein distance, KL divergence, and α-Renyi divergence (for $\alpha < \infty$) also do not imply bounds on the infinity distance, and the running time bounds provided by these works are polynomial in $\frac{1}{\varepsilon}$. (See also Appendix A for additional discussion and challenges.)

Among prior works that give algorithms with bounds on d_∞, [20] applies the grid walk Markov chain of [1] to sample from a uniform distribution on a convex body. [2] extends the approach of [20] to log-Lipschitz log-concave distributions. Unlike the TV-distance case where algorithms whose running time depends *logarithmically* on the error are known (e.g., [1, 29, 34]), the best available bounds for sampling within $O(\varepsilon)$ infinity-distance [20, 2] have runtime that is polynomial in $\frac{1}{\varepsilon}$ and a relatively large polynomial in d.

Our contributions. We present a new approach to output samples, which come with d_∞ bounds, from a log-concave and log-Lipschitz distribution constrained to a a convex body. Specifically, when $K := \{\theta : A\theta \leq b\}$ is a polytope (where one is given A and b) our main result (Theorem 2.1) guarantees samples from a distribution that is within $O(\varepsilon)$ error in d_∞ and whose runtime depends logarithmically on $\frac{1}{\varepsilon}$ compared to the polynomial dependence of [2]. Our approach departs from prior works that construct Markov chains on a $\frac{1}{\varepsilon^2}$-discretization of K to achieve a sample with ε infinity-distance error, and we present a method (Algorithm 1) to directly convert continuous samples from K with total-variation bounds to samples with infinity bounds (Theorem 2.2). This continuous-space approach also allows us to obtain an improvement on the dimension d in the running time when K is a polytope by plugging in TV-distance running time bounds for the Dikin Walk Markov chain of [34]. As immediate applications, we obtain faster algorithms for differentially private empirical risk minimization (Corollary 2.4) and low rank approximation (Corollary 2.5).

2 Results

Let $B(v, s) := \{z \in \mathbb{R}^d : \|z - v\|_2 \leq s\}$ and ω denote the matrix-multiplication constant.

Theorem 2.1 (Main result) *There exists an algorithm which, given $\varepsilon, L, r, R > 0$, $A \in \mathbb{R}^{m \times d}$, $b \in \mathbb{R}^m$ that define a polytope $K := \{\theta \in \mathbb{R}^d : A\theta \leq b\}$ contained in a ball of radius R, a point $a \in \mathbb{R}^d$ such that K contains a ball $B(0, r)$ of smaller radius r, and an oracle for the value of a convex function $f : K \to \mathbb{R}^d$, where f is L-Lipschitz, and defining π to be the distribution $\pi \propto e^{-f}$, outputs a point from a distribution ν such that $\mathrm{d}_\infty(\nu, \pi) < \varepsilon$. Moreover, with very high probability[1], this algorithm takes $O(T)$ function evaluations and $O(T \times md^{\omega-1})$ arithmetic operations, where $T = O((m^2d^3 + m^2dL^2R^2) \times [LR + d\log(\frac{Rd+LRd}{r\varepsilon})])$.*

[1]The number of steps is $O(\tau \times T)$, where $\mathbb{E}[\tau] \leq 3$, $\mathbb{P}(\tau \geq t) \leq \left(\frac{2}{3}\right)^t$ for $t \geq 0$, and $\tau \leq O(d\log(\frac{R}{r}) + LR)$ w.p. 1.

In comparison to the polynomial in $\frac{1}{\varepsilon}$ runtime bounds of [2], Theorem 2.1 guarantees a runtime that is *logarithmic* in $\frac{1}{\varepsilon}$, and also improves the dependence on the dimension d, in the setting where K is a polytope. Specifically, [2] show that the number of steps of the grid walk to sample from π with infinity-distance error at most ε is $O(\frac{1}{\varepsilon^2}(d^{10}+d^6L^4R^4)\times \mathrm{polylog}(\frac{1}{\varepsilon},\frac{1}{r},R,L,d))$ (Lemma 6.5 in the Arxiv version of [2]). When applying their algorithm to the setting where f is constrained to a polytope $K=\{x\in\mathbb{R}^d : Ax\leq b\}$, each step of their grid walk Markov chain requires computing a membership oracle for K and the value of the function f. The membership oracle can be computed in $O(md)$ arithmetic operations. Thus the bound on the number of arithmetic operations for each step of their grid walk is $O(md)$ (provided that each function evaluation takes at most $O(md)$ arithmetic operations). Thus the bound on the number of arithmetic operations to obtain a sample from π is $O(\frac{1}{\varepsilon^2}(md^{11}+md^7L^4R^4)\times \mathrm{polylog}(\frac{1}{\varepsilon},\frac{1}{r},R,L,d))$. Thus, Theorem 2.1 improves on this bound by a factor of roughly $\frac{1}{\varepsilon^2 m^3}d^{8-\omega}$. For example, when $m=O(d)$, as may be the case in differentially private applications, the improvement is $\frac{1}{\varepsilon^2}d^{5-\omega}$.

We note that the bounds of [2] also apply in the more general setting where K is a convex body with membership oracle. One can extend our bounds to achieve a runtime that is logarithmic in $\frac{1}{\varepsilon}$ (and polynomial in d,L,R) in the more general setting where K is a convex body with membership oracle; we omit the details (see Remark 2.3).

Moreover, we also note that while there are several results which achieve $O(\delta)$ TV bounds in time logarithmic in $\frac{1}{\delta}$, TV bounds do not in general imply $O(\varepsilon)$ bounds on the KL or Renyi divergence, or on the infinity-distance, for any $\delta>0$.[2] On the other hand, an ε-infinity-distance bound does immediately imply a bound of ε on the KL divergence D_{KL}, and α-Renyi divergence D_α, since $D_{\mathrm{KL}}(\mu,\pi)\leq \mathrm{d}_\infty(\mu,\pi)$ and $D_\alpha(\mu,\pi)\leq \mathrm{d}_\infty(\mu,\pi)$ for any $\alpha>0$ and any pair of distributions μ,π. Thus, under the same assumptions on K and f, Theorem 2.1 implies a method of sampling from a Lipschitz concave log-density on K with ε KL and Renyi divergence error in a number of arithmetic operations that is *logarithmic* in $\frac{1}{\varepsilon}$, with the same bound on the number of arithmetic operations.

The polynomial dependence on $\frac{1}{\varepsilon}$ in [2] is due to the fact that they rely on a discrete-space Markov chain [1], on a grid with cells of width $w=O(\frac{\varepsilon}{L\sqrt{d}})$, to sample from π within $O(\varepsilon)$ infinity-distance. Since their Markov chain's runtime bound is polynomial in w^{-1}, they get a runtime bound for sampling within $O(\varepsilon)$ infinity-distance that is polynomial in $\frac{1}{\varepsilon}$. The proof of Theorem 2.1 bypasses the use of discrete grid-based Markov chains by introducing Algorithm 1 which transforms any sample within $\delta=O(\varepsilon e^{-d-nLR})$ TV distance of the distribution $\pi\propto e^{-f}$ on the *continuous* set K (as opposed to a discretization of K), into a sample within $O(\varepsilon)$ infinity-distance from π. This allows us to make use of a continuous-space Markov chain, whose step size is not restricted to a grid of width $O(\frac{\varepsilon}{L\sqrt{d}})$ and is instead independent of ε, to obtain a sample within $O(\varepsilon)$ infinity-distance from π in time that is logarithmic in $\frac{1}{\varepsilon}$.

Theorem 2.2 (Main technical contribution: Converting TV bounds to infinity-distance bounds) *There exists an algorithm (Algorithm 1) which, given $\varepsilon,r,R,L>0$, a membership oracle for a convex body K contained in a ball of radius R and containing a ball $B(0,r)$, and an oracle which outputs a point from a distribution μ which has TV distance $\delta\leq O(\varepsilon\times(\frac{R(d\log(R/r)+LR)^2}{\varepsilon r})^{-d}e^{-LR})$ from a distribution $\pi\propto e^{-f}$ where $f:K\to\mathbb{R}$ is an L-Lipschitz function (see Appendix B for the exact values of δ and related hyper-parameters), outputs a point $\hat{\theta}\in K$ such that the distribution ν of $\hat{\theta}$ satisfies $\mathrm{d}_\infty(\nu,\pi)\leq\varepsilon$. Moreover, with very high probability[3], this algorithm finishes in $O(1)$ calls to the sampling and membership oracles, plus $O(d)$ arithmetic operations.*

To the best of our knowledge Theorem 2.2 is the first result which for any $\varepsilon,L,r,R>0$, when provided as input a sample from a continuous-space distribution on a convex body K within some TV distance $\delta=\delta(\varepsilon,r,R,L)>0$ from a given L-log-Lipschitz distribution π on K, where K is contained in a ball of radius R and containing a ball of smaller radius r, outputs a sample with

[2] For instance, if $\pi(\theta)=1$ with support on $[0,1]$, for every $\delta>0$ there is a distribution ν where $\|\nu-\pi\|_{\mathrm{TV}}\leq 2\delta$ and yet $\mathrm{d}_\infty(\nu,\pi)\geq D_{\mathrm{KL}}(\nu,\pi)\geq\frac{1}{2}$. ($\nu(\theta)=e^{\frac{1}{\delta}}$ on $\theta\in[0,\delta e^{-\frac{1}{\delta}}]$, $\nu(\theta)=\frac{1-\delta}{1-\delta e^{-\frac{1}{\delta}}}$ on $(\delta e^{-\frac{1}{\delta}},1]$ and $\nu(\theta)=0$ otherwise)

[3] Algorithm 1 finishes in τ calls to the sampling and membership oracles, plus $O(\tau d)$ arithmetic operations, where $\mathbb{E}[\tau]\leq 3$ and $\mathbb{P}(\tau\geq t)\leq\left(\frac{2}{3}\right)^t$ for all $t\geq 0$ and $\tau\leq 5d\log(\frac{R}{r})+5LR+2$ w.p. 1.

distribution within infinity-distance $O(\varepsilon)$ from π. This is in contrast to previous works [2] (see also [20] which applies only to the special case where π is the uniform distribution on K) which instead require as input a sample with bounded TV distance from the restriction of π on a *discrete* grid on K, and then convert this discrete-space sample into a sample within infinity-distance $O(\varepsilon)$ from the continuous-space distribution $\pi : K \to \mathbb{R}$.

Algorithm 1: Interior point TV to infinity-distance converter

Input: $d \in \mathbb{N}$
Input: A membership oracle for a convex body $K \in \mathbb{R}^d$ and an $r > 0$ such that $B(0, r) \subseteq K$.
Input: A sampling oracle which outputs a point from a distribution $\mu : K \to \mathbb{R}$
Output: A point $\hat{\theta} \in K$.
Hyperparameters: $\Delta > 0$, $\tau_{\max} \in \mathbb{N}$ (set in Appendix B)
for $i = 1, \ldots, \tau_{\max}$ **do**
  Sample a point $\theta \sim \mu$
  Sample a point $\xi \sim \mathrm{Unif}(B(0, 1))$
  Set $Z \leftarrow \theta + \Delta r \xi$
  Set $\hat{\theta} \leftarrow \frac{1}{1-\Delta} Z$
  If $\hat{\theta} \in K$, output $\hat{\theta}$ with probability $\frac{1}{2}$ and halt. Otherwise, continue.
end
Sample a point $\hat{\theta} \sim \mathrm{Unif}(B(0, r))$
Output $\hat{\theta}$

Remark 2.3 (Extension to convex bodies with membership oracles) *We note that Theorem 2.1 can be extended to the general setting where K is an arbitrary convex body in a ball of radius R and containing a ball of smaller radius r, and we only have membership oracle access to K. Namely, one can plug in the results of [29] to our Theorem 2.2 to generate a sample from a L-Lipschitz concave log-density on an arbitrary convex body K in a number of operations that is (poly)-logarithmic in $\frac{1}{\varepsilon}, \frac{1}{r}$ and polynomial on d, L, R. We omit the details.*

Applications to differentially private optimization. Sampling from distributions with $O(\varepsilon)$ infinity-distance error has many applications to differential privacy. Here, the goal is to find a randomized mechanism $h : \mathcal{D}^n \to \mathcal{R}$ which, given a dataset $x \in \mathcal{D}^n$ consisting of n datapoints, outputs model parameters $\hat{\theta} \in \mathcal{R}$ in some parameter space $\mathcal{R}$, which minimize a given (negative) utility function $f(\theta, x)$, under the constraint that the output $\hat{\theta}$ preserves the pure ε-differential privacy of the data points x. A randomized mechanism $h : \mathcal{D}^n \to \mathcal{R}$ is said to be ε-differentially private if for any datasets $x, x' \in \mathcal{D}$ which differ by a single datapoint, and any $S \subseteq \mathcal{R}$, we have that $\mathbb{P}(h(x) \in S) \leq e^{\varepsilon} \mathbb{P}(h(x') \in S)$; see [14].

As one application of Theorem 2.1, we consider the problem of finding an (approximate) minimum $\hat{\theta}$ of an empirical risk function $f : K \times \mathcal{D}^n \to \mathbb{R}$ under the constraint that the output $\hat{\theta}$ is ε-differentially private, where $f(\theta, x) := \sum_{i=1}^n \ell_i(\theta, x_i)$. Following [2], we assume that the $\ell_i(\cdot, x)$ are L-Lipschitz for all $x \in \mathcal{D}^n$, $i \in \mathbb{N}$, for some given $L > 0$. In this setting [2] show that the minimum ERM utility bound under the constraint that $\hat{\theta}$ is pure ε-differentially private, $\mathbb{E}_{\hat{\theta}}[f(\hat{\theta}, x)] - \min_{\theta \in K} f(\theta, x) = \Theta(\frac{dLR}{\varepsilon})$, is achieved if one samples $\hat{\theta}$ from the exponential mechanism $\pi \propto e^{-\frac{\varepsilon}{2LR} f}$ with infinity-distance error at most $O(\varepsilon)$. Plugging Theorem 2.1 into the framework of the exponential mechanism, we obtain a pure ε-differentially private mechanism which achieves the minimum expected risk (Corollary 2.4, see Section C.1 for a proof).

Corollary 2.4 (Differentially private empirical risk minimization) *There exists an algorithm which, given $\varepsilon, L, r, R > 0$, $A \in \mathbb{R}^{m \times d}$, $b \in \mathbb{R}^m$ that define a polytope $K := \{\theta \in \mathbb{R}^d : A\theta \leq b\}$ contained in a ball of radius R and containing a ball $B(0, r)$ of smaller radius r, and a convex function $f(\theta, x) := \sum_{i=1}^n \ell_i(\theta, x_i)$, where each $\ell_i : K \to \mathbb{R}$ is L-Lipschitz, outputs a random point $\hat{\theta} \in K$ which is pure ε-differentially private and satisfies $\mathbb{E}_{\hat{\theta}}[f(\hat{\theta}, x)] - \min_{\theta \in K} f(\theta, x) \leq O(\frac{dLR}{\varepsilon})$. Moreover, this algorithm takes at most $T \times md^{\omega - 1}$ arithmetic operations plus T evaluations of the function f, where $T = O((m^2 d^3 + m^2 d n^2 \varepsilon^2) \times (\varepsilon n + d) \log^2(\frac{nRd}{r\varepsilon}))$.*

Corollary 2.4 improves on the previous bound [2] of $O((\frac{1}{\varepsilon^2}(m+n)d^{11} + \varepsilon^2 n^4(m+n)d^7) \times \mathrm{polylog}(\frac{nRd}{r\varepsilon})))$ arithmetic operations by a factor of roughly $\max(\frac{d^{8-\omega}}{\varepsilon^2 m^2}, \frac{1}{\varepsilon m^2} nd^5)$, in the setting where the ℓ_i are L-Lipschitz on a polytope K and each ℓ_i can be evaluated in $O(d)$ operations. See Appendix C.1 for a proof of this corollary.

As another application, we consider the problem of finding a low-rank approximation of a sample covariance matrix $\Sigma = \sum_{i=1}^n u_i u_i^\top$ where the datapoints $u_i \in \mathbb{R}^d$ satisfy $\|u_i\| \leq 1$, in a differentially private manner. Given any $k > 0$, the goal is to find a (random) rank-k projection matrix P which maximizes the average variance $\mathbb{E}_P[\langle \Sigma, P \rangle]$ of the matrix Σ (also reffered to as the utility of P), under the constraint that the mechanism which outputs the matrix P is ε-differentially private. This problem has many applications to statistics and machine learning, including differentially private principal component analysis (PCA) [7, 3, 15, 24].

When privacy is not a concern, the solution P which maximizes the variance is just a projection matrix onto the subspace spanned by top-k eigenvectors of Σ, and the maximum variance satisfies $\langle \Sigma, P \rangle = \sum_{i=1}^k \lambda_i$, where $\lambda_1 \geq \cdots \geq \lambda_d > 0$ denote the eigenvalues of Σ. However, when privacy is a concern, there is a tradeoff between the desired privacy level ε and the utility $\mathbb{E}_P[\langle \Sigma, P \rangle]$, and the maximum utility $\mathbb{E}_P[\langle \Sigma, P \rangle]$ one can achieve decreases with the privacy parameter ε. The best current utility bound for an ε-differentially private low rank approximation algorithm was achieved in [24], who show that one can find a pure ε-differentially private random rank-k projection matrix P such that $\mathbb{E}_P[\langle \Sigma, P \rangle] \geq (1-\delta)\sum_{i=1}^k \lambda_i$ whenever $\sum_{i=1}^k \lambda_i \geq \Omega(\frac{dk}{\varepsilon\delta}\log\frac{1}{\delta})$ for any $\delta > 0$. To generate the matrix P, their algorithm generates a sample, with infinity-distance error $O(\varepsilon)$, from a Lipschitz concave log-density on a polytope, and transforms this sample into a projection matrix P. The sampling algorithm used in [24] has a bound of $\mathrm{poly}(\frac{1}{\varepsilon}, d, \lambda_1 - \lambda_d)$ arithmetic operations and they leave as an open problem whether this can be improved from a polynomial dependence on $\frac{1}{\varepsilon}$ to a logarithmic dependence on $\frac{1}{\varepsilon}$. Corollary 2.5 shows that a direct application of Theorem 2.1 resolves this problem. (See Section C.2 for a proof.)

Corollary 2.5 (Differentially private low rank approximation) *There exists an algorithm which, given a sample covariance matrix $\Sigma = \sum_{i=1}^n u_i u_i^\top$ for datapoints $u_i \in \mathbb{R}^d$ satisfying $\|u_i\| \leq 1$, its eigenvalues $\lambda_1 \geq \ldots \lambda_d > 0$, an integer k, and $\varepsilon, \delta > 0$, outputs a random rank-k symmetric projection matrix P such that P is ε-differentially private and satisfies the utility bound $\mathbb{E}_P[\langle \Sigma, P \rangle] \geq (1-\delta)\sum_{i=1}^k \lambda_i(\Sigma)$ whenever $\sum_{i=1}^k \lambda_i(\Sigma) \geq C\frac{dk}{\varepsilon\delta}\log\frac{1}{\delta}$ for some universal constant $C > 0$. Moreover the number of arithmetic operations is logarithmic in $\frac{1}{\varepsilon}$ and polynomial in d and $\lambda_1 - \lambda_d$.*

3 Proof Overviews

Given any ε, and a function $f : \mathbb{R}^d \to \mathbb{R}$, the goal is to sample from a distribution $\pi(\theta) \propto e^{-f(\theta)}$, constrained to a d-dimensional convex body K with infinity-distance error at most $O(\varepsilon)$ in a number of arithmetic operations that is logarithmic in $\frac{1}{\varepsilon}$. We assume that K is contained in a ball of some radius $R > 0$ and contains a ball of some radius $r > 0$, and f is L-Lipschitz. In addition we would also like our bounds to be polylogarithmic in $\frac{1}{r}$, and polynomial in d, L, R with a lower-order dependence on the dimension d than currently available bounds for sampling from Lipschitz concave log-densities on a polytope in infinity-distance [2]. We note that since whenever K is contained in a ball of radius R and contains a ball $B(0,r)$ of smaller radius r, we also have that $B(0,r) \subseteq K \subseteq B(0,2R)$, without loss of generality, we may assume that $B(0,r) \subseteq K \subseteq B(0,R)$ as this would only change the bounds provided in our main theorems by a constant factor.

The main ingredient in the proof of Theorem 2.1 is Theorem 2.2 that uses Algorithm 1 to transform a TV-bounded sample into a sample from π with error bounded in d_∞. Subsequently, we invoke Theorem 2.1 when K is given as a polytope $K := \{x \in \mathbb{R}^d : Ax \leq b\}$ and plug in the Dikin Walk Markov chain of [34] which generates independent samples from π with bounded TV error. We first present an overview of the proof of Theorem 2.2. (The full proof has been omitted to space restrictions and presented in Appendix B.) The proof of Theorem 2.1 is presented in Section 3.2.

3.1 Converting samples with TV bounds to infinity-distance bounds; proof of Theorem 2.2

Impossibility of obtaining log-dependence on infinity-distance via grid walk. One approach is to observe that if e^{-f} has support on a discrete space S with at most $|S|$ points, then any ν such that $\|\nu - \pi\|_{\mathrm{TV}} \leq \varepsilon$ also satisfies $\mathrm{d}_\infty(\nu, \pi) \leq 2|S| \frac{\max_{z \in S} e^{-f(z)}}{\min_{z \in S} e^{-f(z)}} \times \varepsilon$ for any $\varepsilon \leq \min_{z \in S} \pi(z)$. This suggests forming a grid G over K, then using a discrete Markov chain to generate a sample θ within $O(\varepsilon)$ TV distance of the discrete distribution $\pi_G \propto e^{-f}$ with support on the grid G, and then designing an algorithm which takes as input θ and outputs a point with bounded infinity-distance to the continuous distribution π. This approach was used in [20] in the special case when π is uniform on K, and then extended by [2] to log-Lipschitz log-concave distributions. In their approach, [2] first run a "grid-walk" Markov chain on a discrete grid in a cube containing K. They then apply the bound from [1] which says that the grid walk obtains a sample Z within TV distance $O(\delta)$ from the distribution $\propto e^{-f}$ (restricted to the grid) in time that is polylogarithmic in $\frac{1}{\delta}$ and quadratic in a^{-1}, where a is the distance between neighboring grid points. Since their grid has size $|S| = \Theta((\frac{R}{a})^d)$, a TV distance of $O(\delta)$ automatically implies an infinity-distance of $O(\delta c|S|)$, where c is the ratio of the maximum to the minimum probability mass satisfying $c \leq e^{LR}$ since f is L-Lipschitz on $K \subseteq B(0, R)$. Thus, by using the grid walk to sample within TV-distance $\delta = O(\frac{\varepsilon}{|S|c})$ from the discrete distribution π_G, they obtain a sample Z which also has infinity-distance $O(\varepsilon)$ from π_G. Finally, to obtain a sample from the distribution $\pi \propto e^{-f}$ on the continuous space K, they sample a point uniformly from the "grid cell" $[Z - a, Z + a]^d$ centered at Z. Since f is L-Lipschitz, the ratio $\frac{e^{-f(Z)}}{e^{-f(w)}}$ is bounded by $O(\varepsilon)$ for all w in the grid cell $[Z - a, Z + a]^d$ as long as $a = O(\frac{\varepsilon}{L\sqrt{d}})$, implying that the sample is an infinity-distance of $O(\varepsilon)$ from $\pi \propto e^{-f}$. However, since the running time bound of the grid walk is quadratic in a^{-1}, the grid coarseness $a = O(\frac{\varepsilon}{L\sqrt{d}})$ needed to achieve $O(\varepsilon)$ infinity-distance from π leads to a running time bound which is *quadratic* in $\frac{1}{\varepsilon}$.

To get around this problem, rather than relying on the use of a discrete-space Markov chain such as the grid walk to sample within $O(\varepsilon)$ infinity-distance from π, we introduce an algorithm (Algorithm 1) which transforms any sample within $\delta = O(\varepsilon e^{-d-LR})$ TV distance from the distribution $\pi \propto e^{-f}$ on the *continuous* space K (as opposed to a grid-based discretization of K), into a sample within $O(\varepsilon)$ infinity-distance from π. This allows us to make use of a continuous-space Markov chain, such as the Dikin walk of [34], whose step-size is not restricted by a grid of coarseness $w = O(\frac{\varepsilon}{L\sqrt{d}})$ and instead is independent of ε, in order to generate a sample within $O(\varepsilon)$ infinity-distance from π in runtime that is *logarithmic* in $\frac{1}{\varepsilon}$.

Converting continuous space TV-bounded samples to infinity-distance bounded samples. As discussed in Section 1, there are many Markov chain results which allow one to sample from a log-concave distribution on K with error bounded in weaker metrics such as total variation, Wasserstein, or KL divergence. However, when sampling from a continuous distribution, bounds in these metrics do not directly imply bounds in infinity-distance. And techniques used to prove bounds in weaker metrics do not easily extend to methods for bounding the infinity-distance; see Section A.1.

Convolving with continuous noise. As a first attempt, we consider the following simple algorithm: sample a point $\theta \sim \mu$ from a distribution μ with total variation error $\|\mu - \pi\|_{\mathrm{TV}} \leq O(\varepsilon)$. Since f is L-Lipschitz, for any $\Delta < \frac{\varepsilon}{L}$ and any ball $B(z, \Delta)$ in the Δ-interior of K (denoted by $\mathrm{int}_\Delta(K)$; see Definition B.1), we can obtain a sample from a distribution ν such that $\log(\frac{\nu(z)}{\mu(z)}) \leq \varepsilon$ for all $z \in \mathrm{int}_\Delta(K)$ by convolving μ with the uniform distribution on the ball $B(0, \Delta)$. Sampling from this distribution ν can be achieved by first sampling $\theta \sim \mu$ and then adding noise $\xi \sim \mathrm{Unif}(B(0, \Delta))$ to the sample θ.

Unfortunately, this simple algorithm does not allow us to guarantee that $\log(\frac{\nu(z)}{\mu(z)}) \leq \varepsilon$ at points $z \notin \mathrm{int}_\Delta(K)$ which are a distance less than Δ from the boundary of K. To see why, suppose that $K = [0, 1]^d$ is the unit cube, that f is constant on K, and consider a point $w = (1, \ldots, 1)$ at the corner of the cube K. In this case we could have that $\nu(z) \leq 2^{-d}\pi(z)$ for all z in some ball containing w, and hence $\mathrm{d}_\infty(\nu, \pi) = \sup |\log(\frac{\nu(z)}{\pi(z)})| \geq d\log(2)$, no matter how small we make Δ.

Stretching the convex body to handle points close to the boundary. To get around this problem, we would like to design an algorithm which samples from some distribution ν such that $|\log \frac{\nu(z)}{\pi(z)}| \leq \varepsilon$

for all $z \in K$, including at points z near the corners of K. Towards this end, we first consider the special case where K is itself contained in the Δ-interior of another convex body K', the function $f : K \to \mathbb{R}$ extends to an L-Lipschitz function on K' (also referred to here with slight abuse of notation as f) and where we are able to sample from the distribution $\propto e^{-f}$ on K' with $O(\varepsilon)$ total variation error. If we sample $\theta \sim e^{-f}$ on K' with total variation error $O(\delta)$ where $\delta \leq \varepsilon e^{-d\log(R)}$, add noise $\xi \sim \mathrm{Unif}(B(0,\Delta))$ to θ for $\Delta = \frac{\delta}{LR}$, and then reject $\theta + \xi$ only if it is not in K, we obtain a sample whose distribution is $O(\varepsilon)$ from the distribution $\propto e^{-f}$ on K in infinity-distance.

However, we would still need to define and sample from such a convex body K', and to make sure that K' is not too large when compared to K; otherwise the samples from the distribution $\propto e^{-f}$ on K' may be rejected with high probability. Moreover, another issue we need to deal with is that f may not even be defined outside of K.

To get around these two problems, in Algorithm 1, we begin by taking as input a point $\theta \sim \mu$ sampled from some distribution μ supported on K where $\|\mu - \pi\|_{\mathrm{TV}} \leq \delta$ for some $\delta \leq \varepsilon e^{-d\log(R)}$, and add noise $\xi \sim \mathrm{unif}(B(0,\Delta r))$ in order to sample from a distribution $\hat{\mu}$ which satisfies $\left|\log \frac{\hat{\mu}(z)}{\pi(z)}\right| \leq \varepsilon$ for all $z \in \mathrm{int}_{\Delta r}(K)$. Here r is the radius of the small ball contained in K; the choice of radius Δr for the noise distribution is because we will show in the following paragraphs that to sample from the distribution π on K with infinity-distance error ε it is sufficient sample a point in the Δr-interior of K and to then apply a "stretching" operation to K.

We still need a method of sampling within $O(\varepsilon)$ infinity-distance error of π on all of K, including in the region $K \backslash \mathrm{int}_{\Delta r}(K)$ near the boundary of K. Towards this end, after Algorithm 1 generates a point $Z = \theta + \xi$ from the above-mentioned distribution $\hat{\mu}$, it then multiplies Z by $\frac{1}{1-\Delta}$ and returns the resulting point $\hat{\theta} := \frac{1}{1-\Delta} Z$ if it is K, in other words, if $Z \in (1-\Delta)K$. If we can show that $(1-\Delta)K \subseteq \mathrm{int}_{\Delta r}(K)$, then this would imply that $\left|\log \frac{\hat{\mu}(z)}{\pi(z)}\right| \leq \varepsilon$ for all $z \in (1-\Delta)K$, and hence that the distribution of $\hat{\nu}$ of the returned point $\hat{\theta}$ satisfies

$$\left|\log \frac{\hat{\nu}(z)}{\pi((1-\Delta)z)}\right| \leq \left|\log \frac{\hat{\mu}((1-\Delta)z)}{\pi((1-\Delta)z)}\right| + \log \frac{1}{(1-\Delta)^d} \leq O(\varepsilon)$$

for all $z \in K$. Since f is L-Lipschitz we have $\left|\log \frac{\pi(\theta)}{\pi((1-\Delta)\theta)}\right| = O(\varepsilon)$ for all $\theta \in K$, and hence we would then have that the distribution $\hat{\nu}$ of the point $\hat{\theta}$ returned by Algorithm 1 satisfies

$$\left|\log \frac{\hat{\nu}(z)}{\pi(z)}\right| = \left|\log \frac{\hat{\nu}(z)}{\pi((1-\Delta)z)}\right| + O(\varepsilon) \leq O(\varepsilon) \qquad \forall z \in K. \tag{1}$$

However, for (1) to hold, we still need to show that $(1-\Delta)K \subseteq \mathrm{int}_{\Delta r}(K)$ (proved in Lemma B.1). In other words, we would like to show that for any point $Z \in (1-\Delta)K$, there is a ball $B(Z, \Delta r)$ centered at Z of radius Δr contained in K. To show this fact, it is sufficient to consider the convex hull $\mathcal{C}$ of $B(0,r) \cup \left\{\frac{1}{1-\Delta} Z\right\} \subseteq K$, and show that it contains the ball $B(Z, \Delta r)$. Towards this end, we make the following geometric construction: we let p be a point such that the line $p\hat{\theta}$ is tangent to $B(0,r)$, and q the point on $p\hat{\theta}$ which minimizes the distance $\|q - Z\|_2$ (see Figure 1 in the Appendix). Since $\angle 0p\hat{\theta}$ and $\angle Zq\hat{\theta}$ are both right angles, we have that the triangles $0p\hat{\theta}$ and $Zq\hat{\theta}$ are similar triangles, and hence that $\frac{\|Z-q\|_2}{r} = \frac{\|Z-\hat{\theta}\|_2}{\|\hat{\theta}-0\|_2}$. In other words,

$$\|Z - q\|_2 = \frac{\|Z-\hat{\theta}\|_2}{\|\hat{\theta}-0\|_2} \times r = \frac{\|Z - \frac{1}{1-\Delta} Z\|_2}{\|\frac{1}{1-\Delta} Z\|_2} \times r = \Delta \times r,$$

implying a ball of radius Δr centered at Z is contained in $\mathcal{C} \subseteq K$, and hence $(1-\Delta)K \subseteq \mathrm{int}_{\Delta r}(K)$.

Bounding the infinity distance error. To complete the bound on the infinity-distance of the distribution $\hat{\nu}$ of the point returned by Algorithm 1 to the target distribution π, we must show both a lower bound (Lemma B.5) and an upper bound (Lemma B.6) on the ratio $\frac{\hat{\nu}(\theta)}{\pi(\theta)}$ at every point $\theta \in K$. Both the upper and lower bounds are necessary to bound the infinity-distance $\mathrm{d}_\infty(\hat{\nu}(\theta), \pi(\theta)) = \sup_{\theta \in K} |\log \frac{\hat{\nu}(\theta)}{\pi(\theta)}|$.

Both Lemmas B.5 and B.6 require the input point to have TV error $\delta < \varepsilon(\frac{R}{\Delta r})^{-d} e^{-LR}$. The term $(\frac{R}{\Delta r})^{-d}$ is a lower bound on the ratio of the volume of K to the volume of the smoothing ball

$B(0, \Delta r)$; this bound holds since K is contained in a ball of radius R. The term e^{-LR} is a lower bound on the ratio $\frac{\min_{w \in K} \pi(w)}{\max_{w \in K} \pi(w)}$ of the minimum value of the density π to the maximum value of π at any two points in K; this bound holds since f is L-Lipschitz.

The above choice of δ ensures that in any ball $B(z, \Delta r)$ with center z in the Δr-interior of K, the distribution μ of the input point, which satisfies $\|\mu - \pi\|_{\mathrm{TV}} \leq \delta$, will have between $e^{-\varepsilon}$ and e^{ε} times the probability mass which the target distribution π has inside the ball $B(z, \Delta r)$. Thus, when the distribution μ is smoothed by adding noise uniformly distributed on a ball of radius Δr, the smoothed distribution $\tilde{\nu}(\theta)$ is within $e^{-\varepsilon}$ and e^{ε} times the target probability density $\pi(\theta)$ at any point θ in the Δr-interior of K, allowing us to bound the infinity distance error of the smoothed distribution $\tilde{\nu}$ at any point θ in the Δr-interior of K. We then apply this fact, together with Lemma B.1 which says that for any point $\theta \in K$ the point $(1-\Delta)\theta$ is in the Δr-interior of K, to bound the distribution ν of the output point (after the stretching operation) as follows,

$$\nu(\theta) \geq (1-\Delta)^d \tilde{\nu}((1-\Delta)\theta) \overset{\text{Lemma } B.1}{\geq} (1-\Delta)^d \pi((1-\Delta)\theta) \times e^{-\varepsilon} \geq \pi(\theta) e^{-\frac{\varepsilon}{2}}. \tag{2}$$

Here our choice of hyperparameter $\Delta \leq \frac{\varepsilon}{\max(d, LR)}$ ensures that $(1-\Delta)^d = \Omega(1)$ and, since f is L-Lipschitz, that $\pi((1-\Delta)\theta) \geq e^{-\varepsilon}\pi(\theta)$. This proves the lower bound (Lemma B.5). The proof of the upper bound (Lemma B.6) follows in a similar way as equation (2) but with the inequalities going in the opposite direction.

Bounding the number of iterations and concluding the proof of Theorem 2.2. We still need to deal with the problem that the point $\hat{\theta}$ may not be accepted. If this occurs, roughly speaking, we repeat the above procedure until a point $\hat{\theta}$ with distribution $\hat{\nu}$ is accepted. To bound the number of iterations, we show that $\hat{\theta}$ is in K with high probability. Towards this end, we first use the facts that f is L-Lipschitz and $K \subseteq B(0, R)$, to show that the probability a point sampled from $\pi \propto e^{-f}$ lies inside $(1-\Delta)K$ is at least $(1-\Delta)^d e^{-L\Delta R} \geq \frac{9}{10}$ (Lemma B.2). Lemma B.2 says that if you stretch the polytope by a factor of $\frac{1}{1-\Delta}$, then most of the volume of the stretched polytope $(\frac{1}{1-\Delta})K$ remains inside the original polytope K. The term $(1-\Delta)^d$ is just the ratio of the volume of $(1-\Delta)K$ to the volume of K. And, since f is L-Lipschitz, the term $e^{-L\Delta R}$ bounds the ratio $\frac{\pi(\theta)}{\pi(\frac{1}{1-\Delta}\theta)}$ of the target density at any point $\theta \in K$ to the value of π at the point $\frac{1}{1-\Delta}\theta$ to which the stretching operation transports θ, whenever $\frac{1}{1-\Delta}\theta \in K$. The choice of hyperparameter $\Delta \leq \frac{\varepsilon}{\max(d, LR)}$ ensures that the acceptance probability $(1-\Delta)^d e^{-L\Delta R}$ guaranteed by Lemma B.2 is at least $\frac{9}{10}$.

Since the convex body $(1-\Delta)K$ contains the ball $B(0, \frac{r}{2})$, applying Lemma B.1 a second time (this time to the convex body $(1-\Delta)K$) we get that $(1-3\Delta)K \subseteq \mathrm{int}_{\Delta r}((1-\Delta)K)$. Thus, by Lemma B.2 we have that θ lies inside $\mathrm{int}_{\Delta r}((1-\Delta)K)$ with probability at least $\frac{9}{10} - \delta \geq \frac{8}{10}$ (as θ is sampled from π with TV error $\leq \delta$). Therefore, since $\xi \sim B(0, \Delta r)$, we must also have that the probability that the point $\hat{\theta} = \frac{1}{1-\Delta}(\theta + \xi)$ is in K (and is therefore not rejected) is greater than $\frac{8}{10}$ [4]. This implies that the number of iterations until our algorithm returns a point $\hat{\theta}$ is less than $k > 0$ with probability at least $1 - 2^{-k}$, and the expected number of iterations is at most 2 (proved in Corollary B.4).

Since each iteration requires one random sample θ from the distribution μ, and one call to a membership oracle for K (to determine if $\frac{1}{1-\Delta}Z \in K$), the number of sampling oracle and membership oracle calls required by Algorithm 1 is $O(1)$ with very high probability. Therefore, with high probability, Algorithm 1 returns a point $\hat{\theta}$ from a distribution with infinity-distance at most ε from π after $O(1)$ calls to the sampling and membership oracles.

Since Algorithm 1 succeeds with probability $1 - 2^{-k}$ after k iterations, after $\tau_{\max} = 5d\log(\frac{R}{r}) + 5LR + \varepsilon$ iterations Algorithm 1 will have succeeded with probability roughly $1 - \varepsilon(\frac{R}{r})^{-5d}e^{-5LR}$. In the very unlikely event that Algorithm 1 still has not succeeded after $\tau_{\max}$ iterations, Algorithm 1 simply outputs a point sampled from the uniform distribution on the ball $B(0, r)$ of radius r contained in K. The probability mass of the target distribution π inside this ball is at least as large

[4] In Algorithm 1 we reject $\hat{\theta}$ with a slightly higher probability to ensure that, in differential privacy applications, in addition to the privacy of the point returned by the algorithm, the runtime is also ε-differentially private.

as $(\frac{R}{r})^{-d}e^{-LR}$; thus, since f is L-Lipschitz, we show in Corollary B.4 that outputing a sample from the uniform distribution on this ball with probability $\varepsilon(\frac{R}{r})^{-5d}e^{-5LR}$ does not change the ∞-distance error of the sample returned by the algorithm by more than ε.

3.2 Completing the proof of Theorem 2.1

Proof: By Theorem 1, the output of Algorithm 1 has infinity-distance error bounded by ε as long as the input samples have TV distance error bounded by $\delta \leq O\left(\varepsilon \times \left(\frac{R(d\log(R/r)+LR)^2}{\varepsilon r}\right)^{-d} e^{-LR}\right)$, and, with high probability, Algorithm 1 requires $O(1)$ such independent samples. To generate a sample from π with TV error $O(\delta)$ when $K = \{\theta \in \mathbb{R}^d : A\theta \leq b\}$ is a polytope defined by m inequalities, we use the Dikin Walk Markov chain of [34]. This Markov chain requires an initial point from some distribution μ_0 which is w-warm with respect to the stationary distribution π, that is, $\sup_{z\in K} \frac{\mu_0(z)}{\pi(z)} \leq w$. To obtain a warm start, we let μ_0 be the uniform distribution on the ball with radius r contained in K and sample from μ_0. Since f is L-Lipschitz, and K is contained in a ball of radius R, μ_0 is w-warm with

$$w \leq \frac{1}{\mathrm{Vol}(B(0,r))} \times \left(\frac{\max_{z\in K}\pi(\theta)}{\min_{z\in K}\pi(\theta)} \times \mathrm{Vol}(B(0,R))\right) \leq \left(\frac{R}{r}\right)^d \times e^{RL}.$$

From [34], we have that from this w-warm start the Dikin Walk Markov chain requires at most $O((m^2d^4 + m^2d^2L^2R^2)\log(\frac{w}{\delta}))$ steps to generate a sample with TV distance at most δ from π, where each step makes one function evaluation and $O(md^{\omega-1})$ arithmetic operations. Plugging in the above values of δ, w the number of Markov chain steps is $T = O((m^2d^3 + m^2dL^2R^2) \times [LR + d\log(\frac{Rd+LRd}{r\varepsilon})])$ to generate each independent sample with the required TV error $O(\delta)$. Since the number of independent samples required as input for Algorithm 1 is $O(1)$ w.h.p., the number of arithmetic operations for Algorithm 1 to output a point with at most ε infinity-distance error is $O(T \times md^{\omega-1})$.

Finally, we note that in the more general setting where K is a convex body with membership oracle (but not necessarily) a polytope, we can instead use, for instance, the hit-and-run Markov chain of [29] to generate samples from π with TV error $O(\delta)$ in a number of membership and function evaluation oracle calls that is polynomial in d and poly-logarithmic in $\frac{1}{\delta}, R, r$. We can then plug this sample into our Algorithm 1 to obtain a sample from π with infinity-distance error $O(\varepsilon)$ in a number of oracle calls that is (poly)-logarithmic in $\frac{1}{\varepsilon}, \frac{1}{r}$ and polynomial on d, L, R. (see Remark 2.3). ■

4 Conclusion, Limitations, and Future Work

To the best of our knowledge, this is the first work that presents an algorithm for sampling from log-concave distributions on convex bodies that comes with infinity-distance bounds and whose running time depends logarithmically on $1/\varepsilon$. Towards this, the main technical contribution is Algorithm 1 (and Theorem 2.2) which achieves this improved dependence on ε by taking as input continuous samples from a convex body with TV bounds and converting them to samples with infinity-distance bounds.

On the other hand, our bounds are polynomial in LR, yet there are algorithms for sampling from logconcave distributions $\pi \propto e^{-f}$ on a convex body in the total variation distance that are poly-logarithmic in R and do not assume f to be Lipschitz [29]. Thus, the main open problem that remains is whether one can also obtain running time bounds for sampling in the infinity-distance which are poly-logarithmic in R and do not require f to be Lipschitz.

Our main result also has direct applications to differentially private optimization (Corollaries 2.4 and 2.5). Differential privacy is a notion which has been embraced in many technologies in societal contexts where privacy of individuals is a concern. Hence, we see our work to have a potential of positive societal impact and do not foresee any potential negative societal impacts.

Acknowledgments and Disclosure of Funding

This research was supported in part by NSF CCF-2104528, CCF-1908347, and CCF-2112665 awards.

References

[1] David Applegate and Ravi Kannan. Sampling and integration of near log-concave functions. In *Proceedings of the twenty-third annual ACM symposium on Theory of computing*, pages 156–163, 1991.

[2] 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.

[3] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the sulq framework. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 128–138, 2005.

[4] 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.

[5] Sebastien Bubeck, Ronen Eldan, and Joseph Lehec. Finite-time analysis of projected Langevin Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 1243–1251. Citeseer, 2015.

[6] Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. *Advances in neural information processing systems*, 24:2249–2257, 2011.

[7] Kamalika Chaudhuri, Anand Sarwate, and Kaushik Sinha. Near-optimal differentially private principal components. *Advances in neural information processing systems*, 25:989–997, 2012.

[8] Yuansi Chen, Raaz Dwivedi, Martin J Wainwright, and Bin Yu. Vaidya walk: A sampling algorithm based on the volumetric barrier. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1220–1227. IEEE, 2017.

[9] Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Algorithmic Learning Theory*, pages 186–211. PMLR, 2018.

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

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

[12] Alain Durmus, Szymon Majewski, and Blazej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *J. Mach. Learn. Res.*, 20:73–1, 2019.

[13] Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Annals of Applied Probability*, 27(3):1551–1587, 2017.

[14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[15] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 11–20, 2014.

[16] M. E. Dyer and A. M. Frieze. On the complexity of computing the volume of a polyhedron. *SIAM Journal on Computing*, 17(5):967–974, 1988.

[17] Alan Frieze and Ravi Kannan. Log-Sobolev inequalities and sampling from log-concave distributions. *The Annals of Applied Probability*, 9(1):14–26, 1999.

[18] Alan Frieze, Ravi Kannan, and Nick Polson. Sampling from log-concave distributions. *The Annals of Applied Probability*, pages 812–837, 1994.

[19] Arun Ganesh and Kunal Talwar. Faster differentially private samplers via Rényi divergence analysis of discretized Langevin MCMC. *Advances in Neural Information Processing Systems*, 33, 2020.

[20] Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 705–714, 2010.

[21] 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.

[22] 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.

[23] Ravi Kannan and Hariharan Narayanan. Random walks on polytopes and an affine interior point method for linear programming. *Mathematics of Operations Research*, 37(1):1–20, 2012.

[24] Jonathan Leake, Colin S McSwiggen, and Nisheeth K Vishnoi. A polynomial-time algorithm and applications for matrix sampling from Harish-Chandra–Itzykson-Zuber densities. In *ACM symposium on theory of computing STOC*, 2021.

[25] Yin Tat Lee and Santosh Vempala. Eldan's stochastic localization and the KLS hyperplane conjecture: An improved lower bound for expansion. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 998–1007. IEEE, 2017.

[26] 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.

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

[28] László Lovász and Santosh Vempala. Hit-and-run is fast and fun. *preprint, Microsoft Research*, 2003.

[29] László Lovász and Santosh Vempala. Fast algorithms for logconcave functions: Sampling, rounding, integration and optimization. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 57–68. IEEE, 2006.

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

[31] Oren Mangoubi and Aaron Smith. Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions: Continuous dynamics. *The Annals of Applied Probability*, in press.

[32] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

[33] Hariharan Narayanan. Randomized interior point methods for sampling and optimization. *The Annals of Applied Probability*, 26(1):597–641, 2016.

[34] Hariharan Narayanan and Alexander Rakhlin. Efficient sampling from time-varying log-concave distributions. *The Journal of Machine Learning Research*, 18(1):4017–4045, 2017.

[35] Sushant Sachdeva and Nisheeth K Vishnoi. The mixing time of the Dikin walk in a polytope—a simple proof. *Operations Research Letters*, 44(5):630–634, 2016.

[36] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, 2019.

[37] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[38] Nisheeth K. Vishnoi. An introduction to hamiltonian monte carlo method for sampling. *CoRR*, abs/2108.12107, 2021.

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

[40] Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR, 2018.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [N/A]
- Did you include the license to the code and datasets? [N/A]
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] (See Section 4)
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] (See Section 3.2, Appendix B and C)
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Contents

1 Introduction **1**

2 Results **2**

3 Proof Overviews **5**

3.1 Converting samples with TV bounds to infinity-distance bounds; proof of Theorem 2.2 . . . 6

3.2 Completing the proof of Theorem 2.1 . . . 9

4 Conclusion, Limitations, and Future Work **9**

A Challenges obtaining infinity-distance bounds from continuous-space Markov chains **15**

A.1 Challenges in obtaining infinity-distance bounds via spectral gap methods . . . 15

A.2 Challenges in extending coupling-based analysis from Wasserstein (and TV) bounds to infinity-distance bounds . . . 15

A.3 Challenges in extending methods based on gradient flows in space of distributions from KL bounds to infinity-distance bounds . . . 16

B Proofs of Theorem 2.2 and Theorem 2.1 **17**

B.1 Bounding the number of iterations and completing the proof of Theorem 2.2 . . . 17

B.2 Stretching the polytope to avoid samples near the boundary . . . 18

B.3 Bounding the acceptance probability . . . 19

B.4 Bounding the infinity-distance error . . . 21

B.5 Proof of Theorem 2.1 . . . 23

C Proofs of applications to differentially private optimization **24**

C.1 Proof of Corollary 2.4 . . . 24

C.2 Proof of Corollary 2.5 . . . 26

D Performance of Algorithm 1 on simple test functions **26**

A Challenges obtaining infinity-distance bounds from continuous-space Markov chains

A.1 Challenges in obtaining infinity-distance bounds via spectral gap methods

Many TV bounds for Markov chains have been obtained by applying isoperimetric inequalities for log-concave distributions π, to bound the spectral gap $\gamma := 1 - \lambda_2(\mathcal{K})$ of the Markov transition kernel operator $\mathcal{K}$ (see e.g. [27, 28]). If the initial distribution μ_0 is such that $\sup_{\theta \in K} \frac{\mu_0(\theta)}{\pi(\theta)}$ is bounded by some number w (e.g., by initializing the Markov chain at a uniform random point in a ball contained in the interior of the polytope), a bound on the spectral gap of $\mathcal{K}$ implies $O(\delta)$ bounds on the total variation error in a number of steps that is logarithmic in $\frac{w}{\delta}$.

Unfortunately, bounding the spectral gap does not in general allow one to obtain bounds on the infinity-distance. While one can bound the χ^2-divergence (which implies a bound on the TV distance), by using the fact (first shown in [27]) that $\|\mathcal{K}^t u_0\|_2 \leq (1-\gamma)^t \|u_0\|_2$, where $u_0 := \mu_0 - \mathrm{proj}_\pi(\mu_0)$, a bound on the spectral gap does not imply a bound in the infinity-distance error. The difficulty in bounding the infinity-distance error arises because, if the space S is continuous, even though $\|\mathcal{K}^t u_0\|_2 \leq (1-\gamma)^t \|u_0\|_2$, there may still be some $c > 0$ for which $\|\mathcal{K}^t u_0\|_\infty \geq c$ for all $t > 0$. For instance this is the case when π is the uniform distribution on a polytope K and $\mathcal{K}$ is the transition Kernel of the Dikin walk, since for every time t, there is always a ball $B_t \subseteq K$ sufficiently close to the boundary of K such that the Dikin walk has probability zero of entering B_t after t steps (for the Gaussian Dikin walk, the probability of entering B_t is very low, but still nonzero). Thus, $\mathcal{K}^t \mu_0(z) = 0$ for all points $z \in B_t$, and, since $\pi(z) = \frac{1}{\mathrm{vol}(K)}$ at every point $z \in K$, we must have $\|\mathcal{K}^t \mu_0 - \pi\|_\infty = \|\mathcal{K}^t u_0\|_\infty \geq \frac{1}{\mathrm{vol}(K)}$ for all $t \geq 0$.

On the other hand, in the special case when the space S is discrete and has a finite number of elements $|S|$, the (ℓ^2-normalized) eigenvectors v have bounded infinity norm, $\|v\|_\infty \leq 1$, and, hence $\|\mathcal{K}^t \mu_0 - \mathrm{proj}_\pi(\mu_0)\|_\infty \leq (1-\gamma)^t |S|$. Thus, if S is discrete with finitely many elements, bounding the spectral gap implies one can sample from π with infinity-distance error $O(\varepsilon)$ in a number of steps that is logarithmic in $\frac{1}{\varepsilon} \times \frac{|S|}{\min_{i \in [S]} \pi[i]}$ if $\min_{i \in [S]} \pi[i] > 0$ (where, with slight abuse of notation, we denote by π the probability mass function of the discrete Markov chain's stationary distribution).

Aside from bounding the spectral gap, many works instead make use of probabilistic coupling methods to bound the distance of a continuous-space Markov chain to the target distribution in, e.g., the Wasserstein distance metric (see for instance [13, 10]). And other works instead achieve bounds in the KL divergence metric for, e.g., the Langevin dynamics Markov chain by analyzing the (continuous-time) Langevin diffusion as a gradient flow of the KL divergence functional in the space of probability distributions (see for instance [40, 9, 12]). In Appendix A.2 and A.3 we discuss challenges which arise if one seeks to extend either of these methods to obtain $O(\varepsilon)$ bounds in the infinity-distance metric, in a number of Markov chain steps that is polylogarithmic in $\frac{1}{\varepsilon}$.

A.2 Challenges in extending coupling-based analysis from Wasserstein (and TV) bounds to infinity-distance bounds

As an alternative to bounding the spectral gap of a Markov chain, one can oftentimes instead make use of a probabilistic coupling method to bound the error of the Markov chain. Here, one considers two Markov chains, one Markov chain started at an arbitrary initial point which is the initial point provided to the Markov chain sampling algorithm, and another "imaginary" Markov chain (oftentimes just a copy of the algorithm's Markov chain) which is started at a random point distributed according to the target distribution π and for which π is a stationary distribution. The goal is to find a joint distribution– also called a probabilistic coupling– for the steps of the two Markov chains such that the distance between the two chains becomes very small in some metric of interest after the Markov chains take multiple steps. If one can find such a coupling, then one can bound the distance of the algorithm's Markov chain to the target distribution π in the relevant metric. If a coupling is found such that the distance between the two Markov chains contracts in the Euclidean distance, then this implies bounds in the Wasserstein metrics. A contraction of the expected (squared) Euclidean distance implies bounds in the 1- (or 2)-Wasserstein metric (see e.g. [13, 10]). However, for any $k \in \mathbb{N}$, a bound on the k-Wasserstein distance

$W_k(\nu, \pi) := \sup_{\rho \sim \Pi(\mu,\pi)} (\mathbb{E}_{(X,Y)\sim\rho}[\|X - Y\|^k])^{\frac{1}{k}} < \varepsilon$, where $\Pi(\mu, \pi)$ denotes the set of all possible couplings of the distributions μ and π, does not imply a bound on the infinity-distance $\sup_{\theta \in S} \left| \log \frac{\nu(\theta)}{\pi(\theta)} \right|$. For instance, consider $\pi \propto \text{unif}[0, 1]$, and ν the uniform distribution on the grid $\{1, 2, \ldots, n\}$ for $n < \frac{1}{4\varepsilon}$. And consider the coupling $\Pi(\pi, \nu)$ of π and ν, which, for every $i \in \{1, 2, \ldots, n\}$, transports all the probability mass of π in the interval $(\frac{i}{n}, \frac{i+1}{n}]$ to a point mass at $\frac{i+1}{n}$. This coupling does not transport any of the probability mass a distance of more that $\frac{1}{2}\varepsilon$, and hence $W_k(\pi, \nu) < \varepsilon$ for any $k \in \mathbb{N} \cup \{\infty\}$. On the other hand, since ν has atomic point-masses while π is a continuous distribution, we have $\mathrm{d}_\infty(\pi, \mu) = \sup_{\theta \in S} \left| \log \frac{\nu(\theta)}{\pi(\theta)} \right| = \infty$.

In some cases, a contraction can be shown to occur with probability 1, yielding a sample with bounds in the ∞-Wasserstein metric. For instance, this is the case for "idealized" versions of the Hamiltonian Monte Carlo Markov chain whose steps are determined by continuous trajectories determined by the Hamiltonian mechanics [31, 38]. For this idealized version of the Hamiltonian Monte Carlo Markov chain, in the special case where the target log-density is strongly convex and smooth on all of $\mathbb{R}^d$, one can oftentimes show that the two chains contract to within a Euclidean distance of $O(\varepsilon)$, and generate a sample $\hat{\theta}$ from π with an error of $O(\varepsilon)$ in the ∞-Wasserstein metric, after a number of Markov chain steps that is logarithmic in $\frac{1}{\varepsilon}$. From such $\hat{\theta}$, if f is 1-Lipschitz on $\mathbb{R}^d$, one can obtain a sample from π with $O(\varepsilon d)$ infinity-distance error by adding a uniform random vector on a ball of radius roughly εd. One can extend this approach to the problem of sampling from, e.g., smooth and Lipschitz convex log-densities supported on a polytope, by extending f to a L-Lipschitz log-density on all of $\mathbb{R}^d$, and adding a strongly convex regularizer. However, to implement such a Markov chain as an algorithm, the trajectories must be discretized, and the number of discretization steps to bring the two Markov chains within $O(\varepsilon)$ Euclidean distance is polynomial in $\frac{1}{\varepsilon}$ (a polynomial dependence on ε also occurs when one only seeks an $O(\varepsilon)$ bound on the 1- or 2- Wasserstein error for many Markov chain algorithms via contractive couplings, including "Unadjusted" Langevin dynamics Markov chains [13, 10]). Thus, we need a different approach if we wish to achieve $O(\varepsilon)$ infinity-divergence bounds in runtime logarithmic in $\frac{1}{\varepsilon}$.

Another approach would be to design a coupling such that the two Markov chains contract within some distance $O(\frac{1}{d})$, and then to propose to add a uniform random vector on a ball of radius roughly $\Theta(1)$ to each Markov chain and accept this proposed step according to the Metropolis acceptance rule for π. If f is 1-Lipschitz on $\mathbb{R}^d$ (or on the constraint set K), the acceptance probability will be at least $\frac{1}{2}$ and one can show that the total variation distance of two Markov chains is at least $\frac{1}{2}$ after some number T steps where T is polynomial in $\frac{1}{\varepsilon}$. Repeating this coupling every T steps, one can show that the total variation distance of the two chains decreases by a factor of at least $\frac{1}{2}$ every T steps, allowing one to generate a sample from π with total variation error $O(\varepsilon)$ in a number of steps that is logarithmic in $\frac{1}{\varepsilon}$. In other words, there exists a coupling of the two chains such that after a number of steps that is logarithmic in $\frac{1}{\varepsilon}$, the algorithm's Markov chain is equal to the chain with distribution π with probability $1 - O(\varepsilon)$. However, we may still have that, with probability roughly ε, the algorithm's Markov chain is concentrated in a region of space of volume $O(\varepsilon)$ where the total probability mass of π is much smaller than ε, which would mean that the infinity-distance to π would be $\Omega(1)$ even though the total variation distance to π is $O(\varepsilon)$.

A.3 Challenges in extending methods based on gradient flows in space of distributions from KL bounds to infinity-distance bounds

In addition to Markov chain bounds achieved via probabilistic coupling methods, bounds for certain Markov chains can also be achieved by analyzing, e.g., the Langevin diffusion process as a gradient flow of the KL divergence functional (also called the relative entropy functional) in the space of probability distributions under the 2-Wasserstein metric. Using this approach, one can show that the Langevin diffusion process with stationary distribution π converges to within KL divergence distance ε of π in (continuous) time t that is logarithmic in $\frac{1}{\varepsilon}$, if, for instance, π is strongly log-concave and smooth (see e.g. [37, 22]). One can then discretize the Langevin diffusion process using a discrete-time Markov chain algorithm (such as the Langevin dynamics Markov chain), and bound the distance between the distribution of the algorithm's Markov chain and the diffusion process in the KL or Renyi divergence metrics (see for instance [40, 9, 12] for Langevin dynamics Markov chains). However, these bounds are polynomial in $\frac{1}{\varepsilon}$ rather than logarithmic in $\frac{1}{\varepsilon}$. The polyno-

mial dependence on ε is due to the fact that the discretization step size for the Langevin dynamics algorithms to approximate the Langevin diffusion process with error $O(\varepsilon)$ is polynomial in $\frac{1}{\varepsilon}$.

For instance, if one wishes to sample from $\pi \sim e^{-f(\theta)}$ with support on the interval $[-2, 2]$, one can use the Langevin dynamics Markov chain, which is a (first-order) "Euler" discretization of the Langevin diffusion on $\mathbb{R}$ with updates $\hat{\theta}_{i+1}$ at each step $i+1$ given as follows: $\hat{\theta}_{i+1} = \hat{\theta}_i - \eta \nabla f(\hat{\theta}_i) + \sqrt{2\eta}\xi$, where $\xi \sim N(0, I_d)$. To (approximately) sample from $\pi \sim e^{-f(\theta)}$ with support on $[-2, 2]$, one could then output only those steps of the Markov chain which fall inside the constraint interval $[-2, 2]$. If $f : \mathbb{R}^1 \to \mathbb{R}$, $f(\theta) = \frac{1}{2}\theta^2$, then the Langevin diffusion is $\mathrm{d}\theta = -\theta \mathrm{d}t + \sqrt{2}\mathrm{d}W_t$, and the discretization is $\hat{\theta}_{i+1} = \hat{\theta}_i - \eta\hat{\theta}_i + 2\sqrt{\eta}\xi$. The solution to the (continuous-time) Langevin diffusion at time t is a Gaussian random variable $\theta_t \sim \mathcal{N}(\theta_0 e^{-t}, 1 - e^{-2t})$, and its stationary distribution is the target distribution $\propto e^{-\frac{1}{2}\theta^2}$. On the other hand, the stationary distribution of the discrete-time Markov chain with step size parameter η is $N(0, \frac{1}{1-\frac{1}{2}\eta})$. Therefore, to have the Langevin Markov chain approximate the Langevin diffusion to within infinity-distance error $O(\varepsilon)$ (even just in a compact constraint interval such as $K = [-2, 2]$), we would need to have a step size $\eta = \mathrm{poly}(\varepsilon)$, and the number of Markov chain steps required to sample within infinity-distance error $O(\varepsilon)$ would be *polynomial* in $\frac{1}{\varepsilon}$.

B Proofs of Theorem 2.2 and Theorem 2.1

In this section, we first prove our main technical result, Theorem 2.2 (Appendix B.1). The Lemmas we use to prove Theorem 2.2 are proved in Appendix B.2, B.3, and B.4. Finally, we plug in TV bounds for the Dikin Walk Markov chain [34] to Theorem 2.2 to complete the proof of Theorem 2.1 (Appendix B.5).

In the following, we define the random variable τ to be the number of iterations of the "for" loop started by Algorithm 1 if Algorithm 1 halts while it is running the for loop. Otherwise, we set $\tau = \tau_{\max} + 1$.

Setting the parameters. In the following, we assume that

1. $\tau_{\max} \geq 5d\log(\frac{R}{r}) + 5LR + \varepsilon$,
2. $\Delta \leq \frac{\varepsilon}{512\tau_{\max}\max(d, LR)}$,
3. and $\delta \leq \frac{1}{64}\varepsilon \times (\frac{R}{\Delta r})^{-d} e^{-LR}$.

B.1 Bounding the number of iterations and completing the proof of Theorem 2.2

Proof: [of Theorem 2.2]

Correctness. By Lemmas B.5 and B.6, we have that

1. The distribution ν of the output $\hat{\theta}$ of Algorithm 1 satisfies $\mathrm{d}_\infty(\nu, \pi) \leq \varepsilon$.
2. Moreover, the distribution $\hat{\nu}$ of $\hat{\theta}$ conditional on $\tau \leq t$ satisfies $\mathrm{d}_\infty(\hat{\nu}, \pi) \leq \varepsilon$ for any $t < \tau_{\max}$.

Bounding the number of operations. Each iteration of Algorithm 1 requires one call to the sampling oracle (Line 3) for μ, and one call to the membership oracle for K (Line 8). In addition, each line of Algorithm 1 requires no more than $O(d)$ arithmetic operations. Thus, each iteration of Algorithm 1 can be computed in one call to the sampling oracle for μ, one call to the membership oracle for K, plus $O(d)$ arithmetic operations. The number of iterations τ is random, and can be bounded as follows:

By Corollary B.4 we have that $\mathbb{E}[\tau] \leq 3$ and $\mathbb{P}(\tau \geq t) \leq \left(\frac{2}{3}\right)^t$. By Corollary B.4 we also have that

$$\left(\frac{1}{2}\right)^t \leq \mathbb{P}(\tau \geq t) \leq \left(\frac{1}{2} + \frac{\varepsilon}{8\tau_{\max}}\right)^t \qquad \forall t \leq \tau_{\max},$$

and, hence, that

$$\begin{aligned}\left(\frac{1}{2}\right)^t &\leq \mathbb{P}(\tau \geq t) \leq \left(\frac{1}{2}\right)^t \times \left(1 + \frac{\varepsilon}{4\tau_{\max}}\right)^t \qquad \forall t \leq \tau_{\max} \\ &\leq \left(\frac{1}{2}\right)^t \times e^{\frac{\varepsilon}{4}} \qquad \forall t \leq \tau_{\max}.\end{aligned}$$

Thus, we have that, for all $t \leq \tau_{\max}$,

$$\mathbb{P}(\tau = t) = \mathbb{P}(\tau \geq t) - \mathbb{P}(\tau \geq t+1) \leq \left(\frac{1}{2}\right)^t e^{\frac{\varepsilon}{4}} - \left(\frac{1}{2}\right)^{t+1} \tag{3}$$

and that

$$\mathbb{P}(\tau = t) = \mathbb{P}(\tau \geq t) - \mathbb{P}(\tau \geq t+1) \geq \left(\frac{1}{2}\right)^t - \left(\frac{1}{2}\right)^{t+1} e^{\frac{\varepsilon}{4}}. \tag{4}$$

Thus, we have that

$$\left(\frac{1}{2}\right)^t \left(1 - \frac{1}{2}e^{\frac{\varepsilon}{4}}\right) \overset{\text{Eq. (4)}}{\leq} \mathbb{P}(\tau = t) \overset{\text{Eq. (3)}}{\leq} \left(\frac{1}{2}\right)^t \left(e^{\frac{\varepsilon}{4}} - \frac{1}{2}\right) \qquad \forall t \leq \tau_{\max},$$

and, hence, that

$$\left(\frac{1}{2}\right)^t e^{-\frac{\varepsilon}{2}} \leq \mathbb{P}(\tau = t) \leq \left(\frac{1}{2}\right)^t e^{\frac{\varepsilon}{2}} \qquad \forall t \leq \tau_{\max}, \tag{5}$$

since $\varepsilon \leq 1$.

■

B.2 Stretching the polytope to avoid samples near the boundary

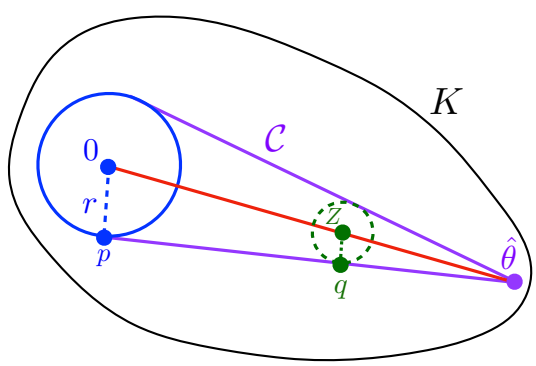

Figure 1: The construction used in the proof of Lemma B.1.

Definition B.1 (Interior) *For any $\Delta \geq 0$ and any $S \subseteq \mathbb{R}^d$, we define the Δ-interior of S, $\mathrm{int}_\Delta(S)$, as*

$$\mathrm{int}_\Delta(S) = \{z \in S : B(z, \Delta) \in S\}.$$

Lemma B.1 *Let $Z \in \mathbb{R}^d$ and $0 \leq \Delta \leq \frac{1}{2}$. Then if $\frac{1}{1-\Delta} Z \in K$, we also have that $Z \in \mathrm{int}_{\Delta r}(K)$.*

Proof: Let $\mathcal{C}$ be the convex hull of $B(0, r) \cup \left\{\frac{1}{1-\Delta} Z\right\}$. Since $B(0, r) \subseteq K$ and $\frac{1}{1-\Delta} Z \in K$, we have that the convex hull $\mathcal{C} \subseteq K$.

Let $h := \max\{s > 0 : B(Z,\ s) \in \mathcal{C}\}$. Defining the point $\hat{\theta} := \frac{1}{1-\Delta} Z$, the point p to be a point such that the line $p\hat{\theta}$ is tangent to $B(0, r)$, and the point q to be a point on $p\hat{\theta}$ which minimizes the distance $\|q - Z\|_2$ (see the diagram in Figure 1). Then we have that the triangles $0p\hat{\theta}$ and $Zq\hat{\theta}$ are similar triangles, since $\angle 0p\hat{\theta}$ and $\angle Zq\hat{\theta}$ are both right angles. Thus, $\frac{\|Z-q\|_2}{r} = \frac{\|Z-\hat{\theta}\|_2}{\|\hat{\theta}-0\|_2}$.

Therefore,

$$\begin{aligned} h &= \|Z - q\|_2 \\ &= \frac{\|Z - \hat{\theta}\|_2}{\|\hat{\theta} - 0\|_2} \times r \\ &= \frac{\|Z - \frac{1}{1-\Delta} Z\|_2}{\|\frac{1}{1-\Delta} Z\|_2} \times r \\ &= \frac{\|\frac{\Delta}{1-\Delta} Z\|_2}{\|\frac{1}{1-\Delta} Z\|_2} \times r \\ &= \Delta \times r \end{aligned}$$

Thus, $Z \in \mathrm{int}_{\Delta r}(\mathcal{C}) \subseteq \mathrm{int}_{\Delta r}(K)$.

■

B.3 Bounding the acceptance probability

The following lemma allows us to bound the expected number of oracle calls in Algorithm 1.

Lemma B.2 *For any $0 \leq \Delta \leq \frac{1}{4}$, we have that*

$$\mathbb{P}_{Z \sim \pi}((Z \in (1-\Delta)K) \geq (1-\Delta)^d e^{-2L\Delta R}.$$

Hence, since by Lemma B.1 $(1-\Delta)K \subseteq \mathrm{int}_{\Delta r}(K)$, we also have that

$$\mathbb{P}_{Z \sim \pi}(Z \in \mathrm{int}_{\Delta r}(K)) \geq (1-\Delta)^d e^{-2L\Delta R}.$$

Proof:

Let $c = (\int_K e^{-f(\theta)} \mathrm{d}\theta)^{-1}$ be the normalizing constant of π, that is, $\pi(\theta) = c e^{-f(\theta)}$ for $\theta \in K$. And let $\tilde{\pi}$ be the distribution

$$\tilde{\pi}(\theta) = \begin{cases} \tilde{c} e^{-f\left(\frac{1}{1-\Delta}\theta\right)} & \text{if } \theta \in (1-\Delta)K \\ 0 & \text{otherwise} \end{cases}$$

where $\tilde{c} = (\int_{(1-\Delta)K} e^{-f\left(\frac{1}{1-\Delta}\theta\right)} \mathrm{d}\theta)^{-1}$ is the normalizing constant of $\tilde{\pi}$. In other words, if $Z \sim \pi$ then we have $(1-\Delta)Z \sim \tilde{\pi}$. Then

$$\tilde{c} = \left(\frac{1}{1-\Delta}\right)^d c. \tag{6}$$

Thus,

$$\frac{\tilde{\pi}(\theta)}{\pi(\theta)} = \frac{\tilde{c}}{c} e^{f(\theta) - f\left(\frac{1}{1-\Delta}\theta\right)} = \left(\frac{1}{1-\Delta}\right)^d e^{f(\theta) - f\left(\frac{1}{1-\Delta}\theta\right)} \qquad \forall \theta \in (1-\Delta)K. \tag{7}$$

Since $K \subseteq B(0, R)$, we have that

$$\left\| \frac{1}{1-\Delta}\theta - \theta \right\|_2 = \frac{\Delta}{1-\Delta} \|\theta\|_2 \leq 2\Delta \|\theta\|_2 \leq 2\Delta R, \tag{8}$$

for all $\theta \in K$. Therefore, since f is L-Lipschitz, Equations (6) and (7) imply that

$$\left(\frac{1}{1-\Delta}\right)^d e^{-2L\Delta R} \leq \frac{\tilde{\pi}(\theta)}{\pi(\theta)} \leq \left(\frac{1}{1-\Delta}\right)^d e^{2L\Delta R} \qquad \forall \theta \in (1-\Delta)K. \tag{9}$$

Therefore,

$$
\begin{aligned}
\mathbb{P}_{Z\sim\pi}(Z \in (1-\Delta)K) &= \int_{\theta\in(1-\Delta)K} \pi(\theta)\mathrm{d}\theta \\
&\overset{\text{Eq. 9}}{\geq} \int_{\theta\in(1-\Delta)K} (1-\Delta)^d e^{-2L\Delta R}\tilde{\pi}(\theta)\mathrm{d}\theta \\
&= (1-\Delta)^d e^{-2L\Delta R} \int_{\theta\in(1-\Delta)K} \tilde{\pi}(\theta)\mathrm{d}\theta \\
&= (1-\Delta)^d e^{-2L\Delta R}.
\end{aligned}
$$

■

Corollary B.3 *For any* $0 \leq \Delta \leq \frac{1}{2}$, *we have that*

$$
\mathbb{P}_{Z\sim\pi}(Z \in \mathrm{int}_{\Delta r}((1-\Delta)K)) \geq [(1-\Delta)^d e^{-2L\Delta R}]^2.
$$

Proof: First, we apply Lemma B.2 to the convex body K and the distribution π to show that

$$
\mathbb{P}_{Z\sim\pi}((Z \in (1-\Delta)K) \geq (1-\Delta)^d e^{-2L\Delta R}. \tag{10}
$$

Next, we let $\pi^\dagger(\theta) \propto e^{-f(\theta)}\mathbb{1}\{\theta \in (1-\Delta)K\}$, and we apply Lemma B.2 again, but this time to the convex body $(1-\Delta)K$ (which, like K, is contained in $B(0,R)$) and the distribution $\pi^\dagger$ (which, like π, has L-Lipschitz log-density) to show that

$$
\mathbb{P}_{W\sim\pi^\dagger}(W \in \mathrm{int}_{\Delta r}((1-\Delta)K)) \geq (1-\Delta)^d e^{-2L\Delta R}. \tag{11}
$$

Thus,

$$
\begin{aligned}
&\mathbb{P}_{Z\sim\pi}(Z \in \mathrm{int}_{\Delta r}((1-\Delta)K)) \\
&= \mathbb{P}_{Z\sim\pi}\left(Z \in \mathrm{int}_{\Delta r}((1-\Delta)K) \middle| Z \in (1-\Delta)K\right) \times \mathbb{P}_{Z\sim\pi}((Z \in (1-\Delta)K) \\
&= \mathbb{P}_{W\sim\pi^\dagger}(W \in \mathrm{int}_{\Delta r}((1-\Delta)K)) \times \mathbb{P}_{Z\sim\hat{\pi}}((Z \in (1-\Delta)K) \\
&\overset{\text{Eq. (10),(11)}}{\geq} [(1-\Delta)^d e^{-2L\Delta R}]^2.
\end{aligned}
$$

■

Corollary B.4 *Algorithm 1 finishes in* τ *calls to the sampling oracle, where*

$$
\left(\frac{1}{2}\right)^t \leq \mathbb{P}(\tau \geq t) \leq \left(\frac{1}{2} + \frac{\varepsilon}{8\tau_{\max}}\right)^t \leq \left(\frac{2}{3}\right)^t \qquad \forall t \in [\tau_{\max}],
$$

and, hence,

$$
\mathbb{E}[\tau] \leq 3.
$$

Proof: By Corollary B.3, we have that, conditional on Algorithm 1 reaching some iteration $i \in \mathbb{N}$, the probability that Algorithm i will reject $\hat{\theta}$ at step i is

$$
\begin{aligned}
\mathbb{P}(\hat{\theta} \text{ rejected at step } i | \text{step } i \text{ reached}) &\leq \frac{1}{2} + \frac{1}{2}\left(1 - \mathbb{P}_{Z\sim\pi}(Z \in \mathrm{int}_{\Delta r}((1-\Delta)K))\right) \\
&\overset{\text{Corr. } B.3}{\leq} \frac{1}{2} + \frac{1}{2}\left(1 - [(1-\Delta)^d e^{-2L\Delta R}]^2\right) \\
&\leq \left(\frac{1}{2} + \frac{\varepsilon}{8\tau_{\max}}\right)^t \\
&\leq \frac{2}{3},
\end{aligned} \tag{12}
$$

where the second-to-last inequality holds because $\Delta \leq \frac{\varepsilon}{512\tau_{\max}\max(d,LR)}$. The last inequality holds because $\tau_{\max} \geq \frac{\varepsilon}{3}$. Hence,

$$\begin{aligned}\mathbb{P}(\tau \geq t) \quad &\leq \Pi_{i=1}^{t}\mathbb{P}(\hat{\theta} \text{ rejected at step } i | \text{step } i \text{ reached}) \\ &\overset{\text{Eq. (12)}}{\leq} \left(\tfrac{2}{3}\right)^t \qquad\qquad \forall t \geq 0, \end{aligned} \tag{13}$$

and

$$\begin{aligned}\mathbb{E}[\tau] &\leq \sum_{t=1}^{\infty}\mathbb{P}(\tau \leq t) \\ &= \sum_{t=1}^{\infty}\mathbb{P}(\tau \geq t) \\ &\overset{\text{Eq. 13}}{\leq} \sum_{t=1}^{\infty}\left(\frac{2}{3}\right)^t \\ &= \frac{1}{1-\frac{2}{3}} \\ &= 3. \end{aligned} \tag{14}$$

■

B.4 Bounding the infinity-distance error

The next lemma allows us to provide a lower bound on the density ν of the output $\hat{\theta}$. Let $\tilde{\nu}$ be the distribution of the random variable $Z = y + \Delta r\xi$, where $y \sim \mu$, and $\xi \sim \text{unif}(B(0,1))$. Let ν^* be the distribution of $\left(\frac{1}{1-\Delta}\right) Z$ conditional on the event that $\left(\frac{1}{1-\Delta}\right) Z \in K$.

Lemma B.5 *For every $\theta \in K$ we have*

$$\nu^*(\theta) \geq e^{-\frac{\varepsilon}{2}}\pi(\theta),$$

and

$$\nu(\theta) \geq e^{-\varepsilon}\pi(\theta).$$

Proof: For all $\theta \in \text{int}_{\Delta r}(K)$, we have

$$\begin{aligned}\tilde{\nu}(\theta) &= \frac{1}{\text{Vol}(B(0,\Delta r))}\int_{w\in B(0,\Delta r)}\mu(\theta+w)\mathrm{d}w \\ &\geq \frac{1}{\text{Vol}(B(0,\Delta r))}\left[\int_{w\in B(0,\Delta r)}\pi(\theta+w)\mathrm{d}w - \delta\right] \\ &\geq \pi(\theta)e^{-L\Delta r} - \frac{\delta}{\text{Vol}(B(0,\Delta r))} \\ &= \pi(\theta)e^{-L\Delta r} - \frac{\delta}{\text{Vol}(B(0,\Delta r))} \times \pi(\theta) \times \frac{1}{\pi(\theta)} \\ &\geq \pi(\theta)e^{-L\Delta r} - \frac{\delta}{\text{Vol}(B(0,\Delta r))} \times \pi(\theta) \times \left(\frac{\max_{w\in K}\pi(w)}{\min_{w\in K}\pi(w)} \times \text{Vol}(B(0,R))\right) \\ &\geq \pi(\theta)e^{-L\Delta r} - \left(\frac{R}{\Delta r}\right)^d \times \delta \times \pi(\theta) \times e^{LR} \\ &= \pi(\theta) \times \left[e^{-L\Delta r} - \left(\frac{R}{\Delta r}\right)^d \times \delta \times e^{LR}\right] \\ &\geq \pi(\theta) \times e^{-\frac{\varepsilon}{8}}. \end{aligned} \tag{15}$$

Where the last inequality holds since $\Delta \leq \frac{\varepsilon}{16Lr}$ and $\delta \leq (e^{-\frac{\varepsilon}{16}} - e^{-\frac{\varepsilon}{8}}) \times \left(\frac{R}{\Delta r}\right)^{-d} e^{-LR}$. Thus, with probability at least $1 - \hat{\delta}$, we have that the conditional distribution $\tilde{\nu}$ satisfies:

$$\tilde{\nu}(\theta) \geq \pi(\theta) \times e^{-\frac{\varepsilon}{8}} \qquad \forall \theta \in \mathrm{int}_{\Delta r}(K). \tag{16}$$

Recall that $\tilde{\nu}$ is the distribution of the random variable $Z = y + \Delta r \xi$, where $y \sim \mu$. And recall that ν^* is the distribution of $(\frac{1}{1-\Delta})Z$ conditional on the event that $(\frac{1}{1-\Delta})Z \in K$. Then Inequality (16) implies that $\forall \theta \in K$

$$\begin{aligned}
\nu^*(\theta) &\geq (1-\Delta)^d \tilde{\nu}((1-\Delta)\theta) \\
&\overset{\text{Eq. (16), Lemma } B.1}{\geq} (1-\Delta)^d \pi((1-\Delta)\theta) \times e^{\frac{-\varepsilon}{8}} \\
&\geq \pi((1-\Delta)\theta) e^{-\frac{\varepsilon}{4}} \\
&\geq \pi(\theta) e^{-\frac{\varepsilon}{2}},
\end{aligned} \tag{17}$$

where the second Inequality holds by Inequality (16) since $(1-\Delta)\theta \in \mathrm{int}_{\Delta r}(K)$ by Lemma B.1. The third inequality holds because $\Delta \leq \frac{\min(1,\varepsilon)}{16d}$. The last inequality holds since $\Delta \leq \frac{\varepsilon}{16LR}$ since $K \subseteq B(0,R)$ implies that $\|\Delta\theta\| \leq \Delta R$.

Finally, Corollary B.4 implies that

$$\begin{aligned}
\nu(\theta) &\geq \mathbb{P}(\tau < \tau_{\max}) \times \nu^*(\theta) \\
&\overset{\text{Corr. } B.4}{\geq} \left(1 - \left(\frac{2}{3}\right)^{\tau_{\max}}\right) \times \nu^*(\theta) \\
&\geq e^{-\frac{\varepsilon}{2}} \times \nu^*(\theta) \\
&\overset{\text{Eq. 17}}{\geq} e^{-\varepsilon} \times \nu(\theta),
\end{aligned}$$

where the third inequality holds since $\tau_{\max} \geq \log(\frac{4}{\varepsilon})$ and $\varepsilon \leq 1$.

■

Lemma B.6 *For every $\theta \in K$ we have*

$$\nu^*(\theta) \leq e^{\frac{\varepsilon}{2}} \pi(\theta),$$

and

$$\nu(\theta) \leq e^{\varepsilon} \pi(\theta).$$

Proof: For all $\theta \in K$, we have

$$\begin{aligned}
\tilde{\nu}(\theta) &= \frac{1}{\mathrm{Vol}(B(0,\Delta r))} \int_{w \in B(0,\Delta r)} \mu(\theta + w) \mathrm{d}w \\
&\leq \frac{1}{\mathrm{Vol}(B(0,\Delta r))} \left[\int_{w \in B(0,\Delta r)} \pi(\theta + w) \mathrm{d}w + \delta \right] \\
&\leq \pi(\theta) e^{L\Delta r} + \frac{\delta}{\mathrm{Vol}(B(0,\Delta r))} \\
&= \pi(\theta) e^{L\Delta r} + \frac{\delta}{\mathrm{Vol}(B(0,\Delta r))} \times \pi(\theta) \times \frac{1}{\pi(\theta)} \\
&\leq \pi(\theta) e^{L\Delta r} + \frac{\delta}{\mathrm{Vol}(B(0,\Delta r))} \times \pi(\theta) \times \left(\frac{\max_{w \in K} \pi(w)}{\min_{w \in K} \pi(w)} \times \mathrm{Vol}(B(0,R)) \right) \\
&\leq \pi(\theta) e^{L\Delta r} + \left(\frac{R}{\Delta r}\right)^d \times \delta \times \pi(\theta) \times e^{LR} \\
&= \pi(\theta) \times \left[e^{L\Delta r} + \left(\frac{R}{\Delta r}\right)^d \times \delta \times e^{LR} \right] \\
&\leq \pi(\theta) \times e^{\frac{\varepsilon}{8}},
\end{aligned} \tag{18}$$

where the last inequality holds since $\Delta \leq \frac{\varepsilon}{16Lr}$ and $\delta \leq (e^{\frac{\varepsilon}{16}} - e^{\frac{\varepsilon}{8}}) \times (\frac{R}{\Delta r})^{-d} e^{-LR}$.

Moreover, by Corollary B.3 and Inequality (15) we have that

$$\begin{aligned}
\mathbb{P}_{Z\sim\tilde{\nu}}(Z \in \mathrm{int}_{\Delta r}((1-\Delta)K)) &\overset{\text{Eq. (15)}}{\geq} e^{\frac{-\varepsilon}{8}} \times \mathbb{P}_{Z\sim\pi}(Z \in \mathrm{int}_{\Delta}((1-\Delta)K)) \\
&\overset{\text{Corr. } B.3}{\geq} e^{\frac{-\varepsilon}{8}} \times [(1-\Delta)^d e^{-L\Delta R}]^2 \\
&\geq e^{\frac{-\varepsilon}{4}}, \qquad (19)
\end{aligned}$$

where the last inequality holds since $\Delta \leq \frac{\min(1,\varepsilon)}{64d}$ and $\Delta \leq \frac{\varepsilon}{128LR}$. Thus, Inequality (19) implies that

$$\int_{\hat{K}} \nu^*(\theta)\mathrm{d}\theta \leq e^{\frac{\varepsilon}{4}} \int_{\hat{K}} \tilde{\nu}(\theta)\mathrm{d}\theta. \qquad (20)$$

Recall that $\tilde{\nu}$ is the distribution of the random variable $Z = y + \Delta r\xi$, where $y \sim \mu$. And recall that ν^* is the distribution of $(\frac{1}{1-\Delta})Z$ conditional on the event that $(\frac{1}{1-\Delta})Z \in K$. Therefore, inequality (20) implies that, for all $\theta \in K$,

$$\begin{aligned}
\nu^*(\theta) &= \frac{\int_K \nu^*(z)\mathrm{d}z}{\int_K \tilde{\nu}(z)\mathrm{d}z} \tilde{\nu}((1-\Delta)\theta) \\
&\overset{\text{Eq. 20}}{\leq} e^{\frac{\varepsilon}{4}} \times \tilde{\nu}((1-\Delta)\theta) \\
&\overset{\text{Eq. 18}}{\leq} e^{\frac{\varepsilon}{4}} \times e^{\frac{\varepsilon}{8}} \pi((1-\Delta)\theta) \\
&\leq e^{\frac{3\varepsilon}{8}} \times e^{\Delta LR} \pi(\theta) \\
&\leq e^{\frac{\varepsilon}{2}} \pi(\theta), \qquad (21)
\end{aligned}$$

where the second-to-last inequality holds because $\pi \propto e^{-f}$ where f is L-Lipschitz, and since $\|(1-\Delta)\theta - \theta\| = \Delta\|\theta\| \leq \Delta R$ because $K \subseteq B(0,R)$. And the last inequality holds because $\Delta \leq \frac{\varepsilon}{32LR}$.

Finally, Corollary B.4 implies that

$$\begin{aligned}
\nu(\theta) &\leq \nu^*(\theta) + \mathbb{P}(\tau = \tau_{\max}) \times \frac{1}{B(0,r)} \\
&\overset{\text{Corr. } B.4}{\leq} e^{\frac{\varepsilon}{2}} \pi(\theta) + \left(\frac{2}{3}\right)^{\tau_{\max}} \times \frac{1}{B(0,r)} \\
&\leq e^{\frac{\varepsilon}{2}} \pi(\theta) + \left(\frac{2}{3}\right)^{\tau_{\max}} \times \frac{1}{B(0,r)} \times \pi(\theta) \times \left(\frac{\max_{w\in K} \pi(\theta)}{\min_{w\in K} \pi(\theta)} \times \mathrm{Vol}(B(0,R))\right) \\
&\leq e^{\frac{\varepsilon}{2}} \pi(\theta) + \left(\frac{2}{3}\right)^{\tau_{\max}} \times \left(\frac{R}{r}\right)^d \times e^{LR} \times \pi(\theta) \\
&\leq e^{\varepsilon} \pi(\theta),
\end{aligned}$$

where the last inequality holds because $\tau_{\max} \geq 5d\log(\frac{R}{r}) + 5LR + \varepsilon$.

■

B.5 Proof of Theorem 2.1

Proof: [**of Theorem 2.1**] We implement Algorithm 1, using the Dikin Walk Markov chain in [34] as a subroutine to compute the TV-bounded sampling oracle for the distribution μ.

To apply Algorithm 1 and Theorem 2.2, we require that $\|\mu - \pi\|_{\mathrm{TV}} \leq \delta$, where $\delta = \frac{1}{64}\varepsilon \times (\frac{R}{\Delta r})^{-d} e^{-LR}$, as well as the following hyperparameter values for Algorithm 1:

1. $\Delta = \frac{\varepsilon}{512\tau_{\max}\max(d,LR)}$, and
2. $\tau_{\max} = 5d\log(\frac{R}{r}) + 5LR + \varepsilon$.

To sample from such a distribution μ, we implement the Dikin Walk Markov chain given in Section 3 of [34], with logarithmic-barrier for the polytope K and hyper-parameters specified by their Condition 2, for $\delta = \frac{1}{64}\varepsilon \times (\frac{R}{\Delta r})^{-d} e^{-LR}$. To provide an initial point θ_0 to the Dikin Walk Markov chain, we sample $\theta_0 \sim \text{unif}(B(0,r))$. Since f is L-Lipschitz and $B(0,r) \subseteq K \subseteq B(0,R)$, the distribution μ_0 of the initial point θ_0 satisfies

$$\sup_{z\in K} \frac{\mu_0(z)}{\pi(z)} \leq \frac{1}{\text{Vol}(B(0,r))} \times \left(\frac{\max_{z\in K} \pi(\theta)}{\min_{z\in K} \pi(\theta)} \times \text{Vol}(B(0,R)) \right)$$
$$\leq \left(\frac{R}{r}\right)^d \times e^{RL}.$$

Thus, the distribution μ_0 of the initial point θ_0 is w-warm with respect to the distribution π, for $w = \left(\frac{R}{r}\right)^d \times e^{RL}$.

By Lemma 4 of [34], their Dikin Walk Markov chain, with initial point θ_0, outputs a point $\|\mu - \pi\|_{\text{TV}} \leq \delta$ and takes at most

$$O\left((m^2d^3 + m^2dL^2R^2)\log(\frac{w}{\delta})\right) = O\left((m^2d^3 + m^2dL^2R^2) \times \left[LR + d\log\left(\frac{Rd + LRd}{r\varepsilon}\right)\right]\right)$$

steps. Moreover, each iteration takes $O(1)$ function evaluations and $md^{\omega-1}$ arithmetic operations.

Bounding the infinity-distance. Since the distribution μ of the point $\hat{\theta}$ provided by the Dikin Walk Markov chain satisfies $\|\mu - \pi\|_{\text{TV}} \leq \delta$, we have, by Theorem 2.2, that Algorithm 1 outputs a point $\hat{\theta} \in K$, such that the distribution ν of $\hat{\theta}$ satisfies $\mathrm{d}_\infty(\nu, \pi) \leq \varepsilon$.

Bounding the number of operations. Moreover, by the proof of Theorem 2.2, we also have that Algorithm 1 finishes in τ calls to the sampling oracle (computed via the Dikin Walk Markov chain) and τ calls to the membership oracle, plus $O(\tau d)$ arithmetic operations, where $\mathbb{E}[\tau] \leq 3$ and $\mathbb{P}(\tau \geq t) \leq \left(\frac{2}{3}\right)^t$ for all $\tau \leq \tau_{\max}$.

The membership oracle for K can be computed in md steps since this can be done by checking the inequality $A\theta \leq b$.

Thus, the total number of steps, when implementing Algorithm 1 with the Dikin Walk Markov chain as subroutine is

$$O\left(\tau \times (m^2d^3 + m^2dL^2R^2)\log(\tfrac{w}{\delta})\right) = O\left(\tau \times (m^2d^3 + m^2dL^2R^2) \times \left[LR + d\log\left(\tfrac{Rd+LRd}{r\varepsilon}\right)\right]\right),$$

where $\mathbb{E}[\tau] \leq 3$ and $\mathbb{P}(\tau \geq t) \leq \left(\frac{2}{3}\right)^t$ for all $t \geq 0$, and $\tau \leq \tau_{\max}$ w.p. 1, and where each step takes $O(1)$ function evaluations and $md^{\omega-1}$ arithmetic operations.

■

C Proofs of applications to differentially private optimization

C.1 Proof of Corollary 2.4

To prove Corollary 2.4 we will need the following Lemma about the exponential mechanism from [2]:

Lemma C.1 (Theorems III.1 and III.2 in [2]) *Suppose that $\hat{\theta}$ is sampled from the distribution $\pi(\theta) \propto e^{-\frac{\varepsilon}{2LR} f(\theta,x)}$ where $f(\theta,x) = \sum_{i=1}^n \ell_i(\theta, x_i)$ where each $\ell_i : K \times \mathcal{D} \to \mathbb{R}$ is L-Lipschitz and convex, and $K \subseteq B(0,R)$ is convex. Then $\hat{\theta}$ is ε-differentially private and achieves ERM utility*

$$\mathbb{E}[f(\hat{\theta}, x) - \min_{\theta\in K} f(\theta, x)] = O\left(\frac{dLR}{\varepsilon}\right).$$

Proof: [**of Corollary 2.4**] To prove Corollary 2.4, we first use Algorithm 1 and the Dikin Walk Markov chain from [34] to (approximately) sample from the distribution $\pi(\theta) \propto e^{-\frac{\varepsilon}{2LR} f(\theta,x)}$. However, if Algorithm 1 has not halted after $t = 10\log(\frac{d}{\varepsilon})$ iterations, we stop running Algorithm 1 and instead output $\hat{\theta} = 0 \in K$.

Showing ε-differential privacy. Since $f(\theta, x) := \sum_{i=1}^{n} \ell_i(\theta, x_i)$, where each ℓ_i is an L-Lipschitz function of θ, we have that f is an is a nL-Lipschitz function of θ, and hence that $\frac{\varepsilon}{2LR} f$ is a $\frac{n\varepsilon}{R}$-Lipschitz function of θ.

By Theorem 2.1, Algorithm 1 (with Dikin Walk Markov chain from [34] as subroutine), conditional on Algorithm 1 halting after $t = 10\log(\frac{n\varepsilon}{d})$ iterations, outputs a point $\hat{\theta}$ from a distribution $\hat{\nu}$ where

$$\mathrm{d}_\infty(\hat{\nu}, \pi) < \varepsilon, \tag{22}$$

with probability at least

$$\mathbb{P}(\tau \leq t) \geq 1 - \left(\frac{2}{3}\right)^{t+1}. \tag{23}$$

Otherwise, we output $\hat{\theta} = 0 \in K$. To see why the event when our algorithm outputs $\hat{\theta} = 0$ satisfies ε-differential privacy, from Equation (5) in the proof of Theorem 2.2 we have that the probability $\mathbb{P}(\tau = \tau_{\max})$ that Algorithm 1 will reject at all $\tau_{\max}$ iterations satisfies

$$\left(\frac{1}{2}\right)^{\tau_{\max}} e^{-\frac{\varepsilon}{2}} \leq \mathbb{P}(\tau = \tau_{\max}) \leq \left(\frac{1}{2}\right)^{\tau_{\max}} e^{\frac{\varepsilon}{2}}, \tag{24}$$

But $\mathbb{P}(\tau = \tau_{\max}) = \mathbb{P}(\hat{\theta} = 0)$, since (ignoring events of probability measure zero) Algorithm 1 outputs $\hat{\theta} = 0$ if and only if the number of rejections τ satisfies $\tau = \tau_{\max}$. Therefore, for any dataset $x \in \mathcal{D}^n$, Inequality (24) implies that the probability $\mathbb{P}(\hat{\theta} = 0) \equiv \mathbb{P}(\hat{\theta} = 0|x)$ that Algorithm 1 outputs the point $0 \in K$ satisfies

$$\left(\frac{1}{2}\right)^{\tau_{\max}} e^{-\frac{\varepsilon}{2}} \leq \mathbb{P}(\hat{\theta} = 0|\, x) \leq \left(\frac{1}{2}\right)^{\tau_{\max}} e^{\frac{\varepsilon}{2}}. \tag{25}$$

Therefore, for any $x, x' \in \mathcal{D}^n$, Inequality (25) implies that $\mathbb{P}(\hat{\theta} = 0|\, x) \leq e^{\varepsilon}\mathbb{P}(\hat{\theta} = 0|\, x')$, implying that the event when our Algorithm outputs $\hat{\theta} = 0$ satisfies the definition of ε-differential privacy. Thus, by Equation (22) and Lemma C.1, we have that $\hat{\theta}$ is pure ε-differentially private.

Bounding the ERM utility. Moreover, by Equations (22), (23) , and Lemma C.1, we have that $\hat{\theta}$ achieves ERM utility

$$\begin{aligned}
\mathbb{E}[f(\hat{\theta}, x) - \min_{\theta \in K} f(\theta, x)] &\leq \mathbb{E}_{\xi \sim \hat{\nu}}[f(\xi, x) - \min_{\theta \in K} f(\theta, x)] + \mathbb{P}(\tau > t) \times [f(0, x) - \min_{\theta \in K} f(\theta, x)] \\
&\overset{\text{Eq. 22,23}}{\leq} e^{\varepsilon} \times \mathbb{E}_{z \sim \pi}[f(z, x) - \min_{\theta \in K} f(\theta, x)] + \left(\frac{2}{3}\right)^{t+1} \times 2nLR \\
&\leq e^{\varepsilon} \times \mathbb{E}_{z \sim \pi}[f(z, x) - \min_{\theta \in K} f(\theta, x)] + \left(\frac{2}{3}\right)^{10\log(\frac{n\varepsilon}{d})} \times 2nLR \\
&\overset{\text{Lemma } C.1}{\leq} e^{\varepsilon} \times O\left(\frac{dLR}{\varepsilon}\right) + \frac{dLR}{\varepsilon} \\
&= O\left(\frac{dLR}{\varepsilon}\right),
\end{aligned}$$

where the second inequality holds since f is nL-Lipschitz and $K \subseteq B(0, R)$.

Bounding the number of operations. Moreover, also by Theorem 2.1, the sum T of the number of steps of Dikin Walk [34] over all the times it is called by Algorithm 1 is at most $O(t \times (m^2d^3 + m^2dn^2\varepsilon^2) \times [\varepsilon n + d\log(\frac{Rd + n\varepsilon d}{r\varepsilon})])$ steps. And each step takes $O(md^{\omega-1})$ arithmetic operations, plus one evaluation of the value of f (and hence n evaluations of functions ℓ_i). Thus, the number of steps is at most $T = O((m^2d^3 + m^2dn^2\varepsilon^2) \times [\varepsilon n + d\log(\frac{Rd + n\varepsilon d}{r\varepsilon})] \times \log(\frac{n\varepsilon}{d})$ steps, where each step takes $O(md^{\omega-1})$ arithmetic operations, plus one evaluation of the function f. Thus, Algorithm 1 finishes in at most $T \times md^{\omega-1}$ arithmetic operations plus T evaluations of the function f, where $T = O((m^2d^3 + m^2dn^2\varepsilon^2) \times [\varepsilon n + d] \times \log^2(\frac{n\varepsilon}{d}))$.

■

C.2 Proof of Corollary 2.5

Proof: [of Corollary 2.5]

[24] show that one can find a pure ε-differentially private rank-k projection P such that $E_P[\langle \Sigma, P\rangle] \geq (1-\delta)\sum_{i=1}^k \lambda_i$ whenever $\sum_{i=1}^k \lambda_i \geq \frac{dk}{\varepsilon\delta}\log\frac{1}{\delta}$ for any $\delta > 0$ and some universal constant $C > 0$, where $\lambda_1 \geq \cdots \geq \lambda_d > 0$ denote the eigenvalues of Σ, by generating a sample from a linear (and hence log-Lipschitz) log-concave distribution π on a polytope K with infinity-distance error $O(\varepsilon)$.

Specifically, their linear log-density π has Lipschitz constant $L = d^2(\lambda_1 - \lambda_d)$, (first equation in Section 5.2 in the arXiv version of [24]). Their polytope K is in $\mathbb{R}^{\frac{d(d-1)}{2}}$ and is defined by $m = d(d-1)$ inequalities (Equations (5) and (6) in the arXiv version of [24]). Moreover, K is contained in a ball of radius $R = \sqrt{d}$, ((Lemma 4.7) of [24]) and contains a ball of radius $r = \frac{1}{8d^2}$ ((Lemma 4.8) of [24]).

Applying Theorem 2.1 with the above parameters for m, L, r, R and the dimension $\frac{d(d-1)}{2}$, we obtain a sample from the distribution π with infinity-distance error $O(\varepsilon)$ in a number of arithmetic operations that is *logarithmic* in $\frac{1}{\varepsilon}$ and polynomial in d and $\lambda_1 - \lambda_d$. ■

Remark C.2 (Privacy of running time) *In the proof of Theorem 2.2 we also show that the probability distribution of the number of iterations τ of Algorithm 1 satisfies $\left(\frac{1}{2}\right)^t e^{-\frac{\varepsilon}{2}} \leq \mathbb{P}(\tau = t) \leq \left(\frac{1}{2}\right)^t e^{\frac{\varepsilon}{2}}$ for all $t \leq \tau_{\max}$. This ensures that the number of iterations τ is ε-pure DP; thus a malicious adversary cannot gain much information about the dataset by measuring the time it takes for Algorithm 1 to finish.*

D Performance of Algorithm 1 on simple test functions

In this section we implement Algorithm 1 on two distributions: a simple one-dimensional distribution, and a $d = 100$ dimensional "Dirichlet" distribution. Simulations were performed in Matlab, on a 1.6 GHz Dual-Core Intel Core i5 2019 Macbook Air laptop.

One-dimensional distribution: We first investigate the performance of Algorithm 1 on a simple one-dimensional distribution π, to verify that our algorithm generates points within the required infinity distance. We do this by generating a histogram of the output of Algorithm 1, by running Algorithm 1 10^7 times. For this experiment, we choose a simple one-dimensional distribution as this allows us to obtain a more precise estimate of the infinity distance, since the number of points needed to compute the histogram grows exponentially with the dimension.

We consider the target distribution $\pi(\theta) \propto e^{-\frac{1}{2}(3-\theta)}$, with support on $K = [-1, 3]$. And we provide Algorithm 1 with samples from a distribution

$$\mu(\theta) \propto \begin{cases} e^{-\frac{1}{2}(3-\theta)} & \theta \in K \backslash ([0.499, 0.501] \cup [1.999, 2.001] \cup [2.999, 3]) \\ 0 & \text{otherwise.} \end{cases}$$

Note that this distribution μ satisfies $\|\pi - \mu\|_{\mathrm{TV}} \leq \frac{1}{100}$, and yet $\mathrm{d}_\infty(\pi, \mu) = \sup_{\theta \in K} |\log \frac{\nu(\theta)}{\pi(\theta)}| = \infty$ since there are points $\theta \in K$ where $\nu(\theta) = 0$ but $\pi(\theta) > 0$.

We run Algorithm 1 10^7 times (with parameters $\varepsilon = 0.1$, $L = \frac{1}{2}$, $R = 4$, and $\Delta = \frac{\varepsilon}{\max(d, LR)} = 0.05$), to generate a histogram of the distribution ν of the output of Algorithm 1 (Figure 2). We observe that Algorithm 1 terminates after an average of 2.1904 iterations, and generates points from a distribution ν with $\mathrm{d}_\infty(\nu, \pi) = 0.1054$, roughly matching the value of the parameter $\varepsilon = \frac{1}{10}$.

100-*dimensional Dirichlet distribution:* We also implement Algorithm 1 on a $d = 100$ dimensional distribution. Specifically, we consider the Dirichlet distribution $\pi(\theta) \propto \prod_{i=1}^d \theta_i$ with support on the simplex $K = \{\theta \in \mathbb{R}^d : \sum_{i=1}^d \theta_i \leq 1, \theta_i \in [0,1] \forall i \in [d]\}$. And we provide Algorithm 1 with samples from a distribution

$$\mu(\theta) = \begin{cases} \pi(\theta) & \theta \notin B(0, \frac{1}{100}) \\ 0 & \text{otherwise.} \end{cases}$$

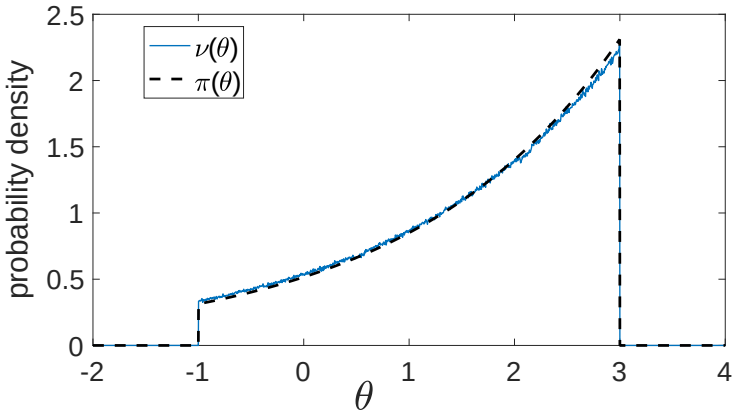

Figure 2: Histogram of the distribution $\nu(\theta)$ of the output of Algorithm 1 (blue curve) for target distribution $\pi(\theta) = e^{-\frac{1}{2}(3-\theta)}$ with support on $K = [-1, 3]$ (dashed black curve), when provided with samples from a distribution μ such that $\|\pi - \mu\|_{\mathrm{TV}} \leq \frac{1}{100}$ and $\mathrm{d}_\infty(\pi, \mu) = \infty$. Algorithm 1 (with parameter $\varepsilon = 0.1$) terminated after an average of 2.1904 iterations, and generated points from a distribution ν with $\mathrm{d}_\infty(\nu, \pi) = 0.1054$, which roughly matches the choice of parameter $\varepsilon = 0.1$.

Note that this distribution μ satisfies $\|\pi - \mu\|_{\mathrm{TV}} < 10^{-d}$, and yet $\mathrm{d}_\infty(\pi, \mu) = \sup_{\theta \in K} |\log \frac{\nu(\theta)}{\pi(\theta)}| = \infty$ since there are points $\theta \in K$ where $\nu(\theta) = 0$ but $\pi(\theta) > 0$. We observe that Algorithm 1 terminates after an average of 1.9935 iterations (with the average taken over 10^5 runs of Algorithm 1). (We do not compute the histogram and infinity distance for the $d = 100$ dimensional Dirichlet distribution, since the number of points needed to compute the histogram grows exponentially with d.)