Journal of Privacy and Confidentiality
Vol. 14 (1) 2024

TPDP 2022

Submitted Feb. 2023 2017
Published Feb. 2024

PRIVATE CONVEX OPTIMIZATION VIA EXPONENTIAL MECHANISM

SIVAKANTH GOPI, YIN TAT LEE, AND DAOGAO LIU

Microsoft Research
e-mail address: sigopi@microsoft.com

University of Washington and Microsoft Research
e-mail address: yintat@uw.edu

University of Washington
e-mail address: dgliu@uw.edu

ABSTRACT. In this paper, we study private optimization problems for non-smooth convex functions $F(x) = \mathbb{E}_i f_i(x)$ on $\mathbb{R}^d$. We show that modifying the exponential mechanism by adding an ℓ_2^2 regularizer to $F(x)$ and sampling from $\pi(x) \propto \exp(-k(F(x) + \mu\|x\|_2^2/2))$ recovers both the known optimal empirical risk and population loss under (ε, δ)-DP. Furthermore, we show how to implement this mechanism using $\widetilde{O}(n \min(d, n))$ queries to $f_i(x)$ for differentially private stochastic convex optimization, where n is the number of samples/users and d is the ambient dimension. We also give a (nearly) matching lower bound $\widetilde{\Omega}(n \min(d, n))$ on the number of evaluation queries.

Our results utilize the following tools that are of independent interest:

- We prove Gaussian differential privacy (GDP) of the exponential mechanism if the loss function is strongly convex and the perturbation is Lipschitz. Our privacy bound is *optimal*: it includes the privacy of Gaussian mechanism as a special case and is proved using the isoperimetric inequality for strongly log-concave measures.
- We show how to sample from $\exp(-F(x) - \mu\|x\|_2^2/2)$ for G-Lipschitz F with η error in total variation (TV) distance using $\widetilde{O}((G^2/\mu)\log^2(d/\eta))$ unbiased queries to $F(x)$. This is the first sampler whose query complexity has *polylogarithmic dependence* on both dimension d and accuracy η.

1. INTRODUCTION

Differential privacy (DP), introduced in [DMNS06, DKM+06], is increasingly becoming the accepted standard in privacy protection. We see an increasing array of adoptions in industry [App17, EPK14, BEM+17, DKY17] and, more recently, the US Census Bureau [Abo16, KCK+18]. Differential privacy allows us to quantify the privacy loss of an algorithm, and is defined as follows.

Key words and phrases: Differential privacy, exponential mechanism, sampling, convex optimization.

The conference version of this paper was published as Gopi, S., Lee, Y. T., & Liu, D. (2022, June). Private convex optimization via exponential mechanism. In Conference on Learning Theory (pp. 1948-1989). PMLR.

In this paper, we say $\mathcal{D}$ and $\mathcal{D}'$ are neighboring databases if they agree on all the user inputs except for a single user's input.

Definition 1.1 ((ε,δ)-DP)**.** A randomized mechanism $\mathcal{M}$ is (ε,δ)-differentially private if for any neighboring databases $\mathcal{D},\mathcal{D}'$ and any subset S of outputs, one has

$$\Pr[\mathcal{M}(\mathcal{D})\in S]\le e^{\varepsilon}\Pr[\mathcal{M}(\mathcal{D}')\in S]+\delta.$$

Privacy concerns are particularly acute in machine learning and optimization using private user data. Suppose we want to minimize some loss function $F(x;\mathcal{D}):\mathcal{K}\to\mathbb{R}$ for some domain $\mathcal{K}$ where $\mathcal{D}$ is some database. We want to output a solution x^{priv} using differentially private mechanism $\mathcal{M}$ such that we minimize the *excess empirical risk*

$$\mathop{\mathbb{E}}_{\mathcal{M}}[F(x^{priv};\mathcal{D})]-F(x^*;\mathcal{D}), \tag{1.1}$$

where $x^*\in\mathcal{K}$ is the true minimizer of $F(x;\mathcal{D})$.

1.1. Exponential Mechanism.

One of the first mechanisms invented in differential privacy, the *exponential mechanism*, was proposed by [MT07]. It involves sampling x^{priv} from the density

$$\pi_{\mathcal{D}}(x)\propto\exp\left(-kF(x;\mathcal{D})\right). \tag{1.2}$$

Here k controls the privacy-vs-utility tradeoff. Large k ensures that we get a good solution but less privacy and small k ensures that we get good privacy but we lose utility. Suppose $\Delta_F=\sup_{\mathcal{D}\sim\mathcal{D}'}\sup_x|F(x;\mathcal{D})-F(x;\mathcal{D}')|$ is the sensitivity of F, where the supremum is over all neighboring databases $\mathcal{D},\mathcal{D}'$. Then choosing $k=\varepsilon/2\Delta_F$, the exponential mechanism satisfies $(\varepsilon,0)$-DP.

The exponential mechanism is widely used both in theory and in practice, such as in mechanism design [HK12], convex optimization [BST14, MV21], statistics [WZ10, WM10, AKRS19], machine learning and AI [ZP19]. Even for infinite and continuous domains, the exponential mechanism can be implemented efficiently for many problems [HT10, CSS13, KT13, BV19, CKS20]. There are also several variants and generalizations of the exponential mechanism that can improve its utility based on different assumptions [TS13, BNS13, RS16, LT19]. See [LT19] for a survey of these results.

1.2. DP Empirical Risk Minimization (DP-ERM).

In many applications, the loss function is given by the average of the loss of each user:

$$F(x;\mathcal{D}):=\frac{1}{n}\sum_{i=1}^{n}f(x;s_i). \tag{1.3}$$

where $\mathcal{D}=\{s_1,s_2,\cdots,s_n\}$ is the collection of users s_i and $f(x;s_i)$ is the loss function of user s_i.

Throughout this paper, we assume $f(\cdot;s)$ is convex, that $f(\cdot;s)-f(\cdot;s')$ is G-Lipschitz for all s,s', and that $\mathcal{K}\subset\mathbb{R}^d$ is convex with diameter D.[1] We call the problem of minimizing the excess empirical risk in (1.3) as DP Empirical Risk Minimization (DP-ERM). This setting is well studied by the DP community [CM08, RBHT12, CMS11, JT14, BST14, KJ16, FTS17,

[1]Some of our results can handle the unconstrained domain, such as $\mathcal{K}=\mathbb{R}^d$.

ZZMW17, Wan18, INS+19, BFTT19, FKT20, KLL21, BGN21, LL21, AFKT21, SSTT21, MBST21, GTU22].[2]

In particular, [BST14] shows that exponential mechanism in (1.2) achieves the optimal excess empirical risk of $O(GDd/\varepsilon)$ under $(\varepsilon, 0)$-DP. On the other hand, [BST14, BFTT19, BFGT20] show that *noisy gradient descent* on $F(x; \mathcal{D})$ achieves an excess empirical risk of

$$O\left(\frac{GD\sqrt{d\log(1/\delta)}}{n\varepsilon}\right) \tag{1.4}$$

under (ε, δ)-DP, which is also shown to be optimal [BST14]. This is a significant $\sqrt{d}$ improvement over the exponential mechanism.

The exponential mechanism is a universally powerful tool in differential privacy. However, nearly all of the previous works on DP-ERM rely on noisy gradient descent or its variants to achieve the significant $\sqrt{d}$ improvement over exponential mechanism under (ε, δ)-DP. One natural question is whether noisy gradient descent has some extra ability that the exponential mechanism lacks or we did not use the exponential mechanism optimally in this setting. This brings us to the first question.

Question 1. Can we obtain the optimal empirical risk in (1.1) under (ε, δ)-DP using exponential mechanism?

1.3. **DP Stochastic Convex Optimization (DP-SCO).** Beyond the privacy and empirical risk guarantees, another important guarantee is the generalization guarantee. Formally, we assume the users are sampled from an unknown distribution $\mathcal{P}$ over convex functions. We define the loss function as

$$\hat{F}(x) = \mathop{\mathbb{E}}_{s\sim\mathcal{P}}[f(x; s)]. \tag{1.5}$$

We want to design a DP mechanism $\mathcal{M}$ that outputs x^{priv} given users $\mathcal{D} = \{s_1, s_2, \ldots, s_n\}$ independently sampled from $\mathcal{P}$ and then to minimize the *excess population loss*

$$\mathop{\mathbb{E}}_{\mathcal{M},\mathcal{D}\sim\mathcal{P}}[\hat{F}(x^{priv})] - \hat{F}(x^*), \tag{1.6}$$

where x^* is the minimizer of $\hat{F}(x)$. We call the problem of minimizing the excess population loss in (1.6) as DP stochastic convex optimization (DP-SCO). By a suitable modification of noisy stochastic gradient descent, [BFTT19, FKT20] show that one can achieve the optimal population loss of

$$O\left(GD\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\right)\right). \tag{1.7}$$

[BFTT19] bounds the generalization error by showing that running SGD on smooth functions is stable and [FKT20] proposes an iterative localization technique. Note that only the algorithm for smooth functions in [BFTT19] can achieve both optimal empirical risk and optimal population loss at the same time, with the price of taking more gradient queries and loss of efficiency. It is unclear to us how one can obtain both using current techniques for non-smooth functions, which brings us to the second question.

[2] Most of the literature uses a stronger assumption that $f(\cdot; s)$ is G-Lipschitz, while some of our results only need to assume the difference $f(\cdot; s) - f(\cdot; s')$ is G-Lipschitz.

Question 2. Can we achieve both the optimal empirical risk and the optimal population loss for non-smooth functions with the same algorithm?

1.4. **Sampling.** Without extra smoothness assumptions on f, currently, there is no optimally efficient algorithm for both problems. For example, with oracle access to gradients of f, the previous best algorithms for DP-SCO use:

- $\widetilde{O}(nd)$ queries to $\nabla f(x;s)$ (by combining [FKT20], Moreau-Yosida regularization and cutting plane methods),
- $\widetilde{O}(\min(n^{3/2}, n^2/\sqrt{d}))$ queries to $\nabla f(x;s)$ [AFKT21],
- $\widetilde{O}(\min(n^{5/4}d^{1/8}, n^{3/2}/d^{1/8}))$ queries to $\nabla f(x;s)$ [KLL21].

Combining these results gives an algorithm for DP-SCO that uses

$$\widetilde{O}(\min(nd, n^{5/4}d^{1/8}, n^{3/2}/d^{1/8}, n^2/\sqrt{d}))$$

many queries to $\nabla f(x;s)$. Although the information lower bound for non-smooth functions with the gradient queries has not been established, it seems unlikely that the answer involves four different cases.

In this paper, we focus on the function value query (zeroth order query) on $f(x;s)$. This query is weaker than the gradient query as it obtains d times less information. They are used in many practical applications, such as clinical trials and ads placement when the gradient is not available, and is also useful in bandit problems. This brings us to the third question.

Question 3. Can we obtain an algorithm with optimal query complexity for DP-SCO for the zeroth order query model?

1.5. **Our Contributions.** In this paper, we answer all three questions positively using the *Regularized Exponential Mechanism*. If we add an ℓ_2^2 regularizer to F and sample x^{priv} from the density

$$\exp\left(-k\left(F(x;\mathcal{D}) + \mu\left\|x\right\|_2^2/2\right)\right), \tag{1.8}$$

then, for a suitable choice of μ and k, we recover the optimal excess risk in (1.4) for DP-ERM and optimal population loss in (1.7) for DP-SCO. Finally, we give an algorithm to sample x^{priv} from the density (1.8) with a nearly optimal number of queries to $f(x;s)$ See Figure 1. To the best of our knowledge, our algorithm is the first whose query complexity has *polylogarithmic dependence* in both dimension and accuracy (TV distance).

Informally, our results are as follows.

Theorem 1.2 (DP-ERM, Informal)**.** *Let $\mathcal{K}$ be a convex set with diameter D and $\{f(\cdot;s)\}$ be a family of convex functions on $\mathcal{K}$, where $f(\cdot;s) - f(\cdot;s')$ is G-Lipschitz for all s, s'. Given a database $\mathcal{D} = \{s_1, s_2, \cdots, s_n\}$, for any $\varepsilon, \delta \in (0, 11/10)$, [3] the regularized exponential mechanism*

$$x^{(priv)} \propto \exp\left(-k \cdot \left(\frac{1}{n}\sum_{i=1}^{n} f(x;s_i) + \frac{\mu}{2}\|x\|_2^2\right)\right)$$

[3]See Theorem 6.2 for general conclusions for all $\varepsilon > 0$

is (ε, δ)-DP with expected excess empirical loss

$$\frac{2GD\sqrt{d\log(1/\delta)}}{\varepsilon n}$$

for appropriate choices of k and μ. Furthermore, if $f(\cdot; s)$ is G-Lipschitz for all s, we can sample $x^{(priv)}$ using $O([\varepsilon^2 n^2/\log(1/\delta)]\log^2(nd/\delta))$ queries in expectation to the values of $f(x; s)$.

Theorem 1.3 (DP-SCO, Informal)**.** *Let $\mathcal{K}$ be a convex set with diameter D and $\{f(\cdot; s)\}$ be a family of convex functions on $\mathcal{K}$ where $f(\cdot; s) - f(\cdot; s')$ is G-Lipschitz for all s, s'. Given a database $\mathcal{D} = \{s_1, s_2, \cdots, s_n\}$ of samples from some unknown distribution $\mathcal{P}$. For any $\varepsilon, \delta \in (0, 1/10)$,[4] the regularized exponential mechanism*

$$x^{(priv)} \propto \exp\left(-k \cdot \left(\frac{1}{n}\sum_{i=1}^{n} f(x; s_i) + \frac{\mu}{2}\|x\|_2^2\right)\right)$$

is (ε, δ)-DP with expected excess population loss

$$\frac{2GD}{\sqrt{n}} + \frac{2GD\sqrt{d\log(1/\delta)}}{\varepsilon n}$$

for some choice of k and μ. Furthermore, if $f(\cdot; s)$ is G-Lipschitz for all s, we can sample $x^{(priv)}$ using $O(\min\{[\varepsilon^2 n^2/\log(1/\delta)], nd\}\log^2(nd/\delta))$ queries in expectation to the values of $f(x; s)$ and the expected number of queries is optimal up to logarithmic terms.

For DP-SCO, we provide a nearly matching information-theoretic lower bound on the number of value queries (Section 7), proving the optimality of our sampling algorithm. Moreover, when f is strongly convex, our proof shows the exponential mechanism (without adding a regularizer) itself simultaneously achieves both the optimal excess empirical risk and optimal population loss.

In a *concurrent and independent* work, [GTU22] study the DP properties of Langevin diffusion, and provide optimal/best known private empirical risk and population loss under both pure-DP ($\delta = 0$) and approximate-DP ($\delta > 0$) constraints. Utility/privacy trade-offs for non-convex functions are also discussed.

2. Techniques

The main contribution of this paper is that adding regularization terms in exponential mechanism leads to optimal algorithms for DP-ERM and DP-SCO. For this, we develop some important tools that could be of independent interest. We now briefly discuss each of the main tools.

[4] See Theorem 6.9 for general conclusions for all $\varepsilon > 0$.

2.1. **Gaussian Differential Privacy (GDP) of Regularized Exponential Mechanism.** To analyze the privacy of the regularized exponential mechanism, we need to bound the privacy curve between a strongly log-concave distribution and its Lipschitz perturbation in the exponent. [MASN16] gave a nearly tight (up to constants) privacy guarantee of exponential mechanism if the distribution $\exp(-kF(x;\mathcal{D}))$ satisfies logarithmic Sobolev inequality (LSI). Since strongly log-concave distributions satisfy LSI, their result immediately gives the (ε,δ)-DP guarantee of our algorithm. However, this gives a sub-optimal privacy bound because it does not take full advantage of strong log-concavity.

Instead, we show directly that the privacy curve between a strongly log-concave distribution and its Lipschitz perturbation in the exponent is bounded above by the privacy curve of an appropriate Gaussian mechanism. This new proof uses the notion of tradeoff function introduced in [DRS19] and the isoperimetric inequality for strongly log-concave distributions.

Theorem 2.1. *Given a convex set $\mathcal{K} \subseteq \mathbb{R}^d$ and μ-strongly convex functions $F, \tilde{F}$ over $\mathcal{K}$, let P, Q be distributions over $\mathcal{K}$ such that $P(x) \propto e^{-F(x)}$ and $Q(x) \propto e^{-\tilde{F}(x)}$. If $\tilde{F} - F$ is G-Lipschitz over $\mathcal{K}$, then for all $\varepsilon > 0$,*

$$\delta(P \parallel Q)(\varepsilon) \le \delta\left(\mathcal{N}(0,1) \,\middle\|\, \mathcal{N}\left(\frac{G}{\sqrt{\mu}}, 1\right)\right)(\varepsilon).$$

This proves that the privacy curve for distinguishing between P, Q is upper bounded the privacy curve of a Gaussian mechanism with sensitivity $G/\sqrt{\mu}$ and noise scale 1.

Tightness: Note that Theorem 2.1 is tight because it contains the privacy of Gaussian mechanism as a special case. If $F(x) = \|x\|_2^2/2$ and $\tilde{F}(x) = \|x-a\|_2^2/2$ for some $a \in \mathbb{R}^d$, then $\tilde{F}(x) - F(x) = -\langle x, a\rangle + \|a\|_2^2/2$ is G-Lipschitz with $G = \|a\|_2$ and $F, \tilde{F}$ are 1-strongly convex. And $P = \mathcal{N}(0, I_d)$ and $Q = \mathcal{N}(a, I_d)$. Therefore,

$$\delta(P \parallel Q) = \delta(\mathcal{N}(0, I_d) \parallel \mathcal{N}(a, I_d)) = \delta(\mathcal{N}(0,1) \parallel \mathcal{N}(\|a\|_2, 1)),$$

which is precisely the upper bound guaranteed by the theorem.

2.2. **Generalization Error of Sampling.** Many important and fundamental problems in machine learning, optimization and operations research are special cases of SCO, and ERM is a classic and widely-used approach to solve it, though their relationships are not well-understood. If one can solve the ERM problem optimally and get the exact optimal solution x^* to minimizing $F(\cdot;\mathcal{D})$ (see Equation 1.3), then [SSSSS09] showed x^* will also be a good solution to the SCO for strongly convex functions. But in most situations, solving ERM optimally costs too much or is even impossible. Can we find a approximately good solution to ERM and hope that it is also a good solution for SCO? [Fel16] provides a negative answer and shows there is no good uniform convergence between $F(\cdot;\mathcal{D})$ and $\hat{F}$, that is there always exists $x \in \mathcal{K}$ such that $|F(x;\mathcal{D}) - \hat{F}(x)|$ is large. This fact forces us to find approximate solution to ERM with very high accuracy, which makes the algorithms inefficient.

Prior works proposed ways to overcome this difficulty, such as the uniform stability in [HRS16] and the iterative localization technique in [AFKT21]. Roughly speaking, uniform stability means that if running algorithms on neighboring datasets lead to similar output distributions, then the generalization error of the ERM algorithm is bounded. Thus a good solution to ERM obtained by a stable algorithm is also a good solution for SCO. [BFTT19] makes use of the stability of running SGD on smooth functions to get a tight bound on the population loss for DP-SCO.

Recall that $F(x;\mathcal{D})$ and $\hat{F}(x)$ are defined in Equation (1.3) and (1.5) respectively. Our result enriches the toolbox of bounding the generalization error and provides new insights for this problem.

Theorem 2.2. *Suppose $\{f_i\}$ that is a family of μ-strongly convex functions over $\mathcal{K}$ and that $f_i - f_{i'}$ is G-Lipschitz for any functions $f_i, f_{i'}$ in the family. For any $k > 0$ and suppose the n samples in data set $\mathcal{D}$ are drawn i.i.d from the underlying distribution, then by sampling $x^{(sol)}$ from density $\propto e^{-kF(x^{(sol)};\mathcal{D})}$, the population loss satisfies*

$$\mathbb{E}[\hat{F}(x^{(sol)})] - \min_{x\in\mathcal{K}} \hat{F}(x) \leq \frac{G^2}{\mu n} + \frac{d}{k}.$$

Considering two neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, our result is based on bounding the Wasserstein distance between the distributions proportional to $e^{-kF(x;\mathcal{D})}$ and $e^{-kF(x;\mathcal{D}')}$, which means the sampling scheme is stable and leads to the $G^2/\mu n$ term in generalization error. The other term d/k is excess empirical loss of the sampling mechanism. One advantage of our result is that it works for both smooth and non-smooth functions. Moreover, we may choose the value k carefully and get a solution with both optimal empirical loss and optimal population loss.

2.3. **Non-smooth Sampling and DP Convex Optimization.** Implementing the exponential mechanism involves sampling from a log-concave distribution. When the negative log-density function F is smooth, i.e. the gradient of F is Lipschitz, there are many efficient algorithms for this sampling tasks such as [Dal17, LSV18, MMW+21, CV19, DMM19, SL19, CDWY20, LST20]. For example, if $F = (1/n)\sum_{i=1}^n f_i$ and each f_i is 1-strongly convex with κ-Lipschitz gradient,[5] we can sample $x \sim \exp(-F(x))$ in $\widetilde{O}(n + \kappa \max(d, \sqrt{nd}) \log(1/\delta))$ iterations with δ error in total variation distance and each iteration involves computing one $\nabla f_i(x)$ [LST21]. Note that this is nearly linear time when $n \gg \kappa^2 d$ and the δ error in total variation distance can be translated to an extra δ error in the (ε, δ)-DP guarantee.

	Complexity	Oracle	Guarantee
[BST14]	$d^{O(1)}$	$F(x)$	$\mathrm{D}_\infty \leq \varepsilon$
[CDJB20]	$G^{O(1)} d^{5/2}/\varepsilon^4$	$\nabla F(x)$	$\mathrm{W}_2 \leq \delta$
[JLLV21] + [Che21]	d^3	$F(x)$	$\mathrm{TV} \leq \delta$
[GT20]	$\alpha^2 G^4 d/\varepsilon^2$	$\nabla F(x)$	$\mathrm{D}_\alpha \leq \varepsilon$
[LC21]	G^2/δ	$\nabla F(x)$	$\mathrm{TV} \leq \delta$
This paper	G^2	$f_i(x)$	$\mathrm{TV} \leq \delta$

Figure 1: The complexity of sampling from $\exp(-F(x))$, where $F = (1/n)\sum_i f_i$ is 1-strongly convex and f_i are G-Lipschitz and convex. For applications in differential privacy, ε is a constant and $\delta = n^{-\Theta(1)}$. Polylogarithmic terms are omitted. Only the last result uses the summation structure and queries only one f_i each step.

[5]For convenience, we used f_i to denote the function $f(\cdot; s_i)$ here and Section 5.

Unfortunately, when the functions f_i are only Lipschitz but not smooth, this problem is more difficult. In Table 1, we summarize some existing results on this topic. They use different guarantees such as Renyi divergence D_α of order α, Wasserstein distance W_2 and total variation distance TV (defined in subsection 3.3). For applications in differential privacy, we need either polynomially small W_2 or TV distance, or ε small D_α distance.

All previous results for non-smooth function use oracle access to F or ∇F (instead of f_i) and have iterative complexity at least d iterations for W_2 or TV distance smaller than $1/d$. Because of this, our algorithm is significantly faster than the previous algorithms and can handle the case when F is expectation of (infinitely many) f_i directly. For example, to get the optimal private empirical loss with typical settings where $\varepsilon = \Theta(1)$ and $\delta = 1/n^{\Theta(1)}$, the previous best samplers use $\widetilde{O}(n^4 d)$ many queries to $\nabla f_i(x)$ by [GT20] or $\widetilde{O}(nd^3)$ many queries to $f_i(x)$ by combining [JLLV21] and [Che21]. In comparison, our algorithm only takes $\widetilde{O}(n^2)$ many $f_i(x)$.

Our result is based on the alternating sampler proposed in [LST21] and a new rejection sampling scheme.

Theorem 2.3. *Let ψ be a μ-strongly convex function defined on a convex set $\mathcal{K} \subseteq \mathbb{R}^d$ and $+\infty$ outside. Given a family of G-Lipschitz convex functions $\{f_i(x)\}_{i\in I}$ defined on $\mathcal{K}$ and an initial point $x_0 \in \mathcal{K}$, define the function $\widehat{F}(x) = \mathbb{E}_{i\in I} f_i(x) + \psi(x)$ and the distance $D = \|x_0 - x^*\|_2$ for some $x^* = \arg\min_{x\in\mathcal{K}} \hat{F}(x)$. Then for any $\delta \in (0, 1/2)$, we can generate a random point x that has δ total variation distance to the distribution proportional to $\exp(-\widehat{F}(x))$ in*

$$T := \Theta\left(\frac{G^2}{\mu}\log^2\left(\frac{G^2(d/\mu + D^2)}{\delta}\right)\right)$$

steps. urthermore, each steps accesses only $O(1)$ many $f_i(x)$ and samples from $\exp(-\psi(x) - \|x - y\|_2^2)/2\eta$ for $O(1)$-many y in expectation, where $\eta = \Theta(G^{-2}/\log(T/\delta))$.

2.4. **Subsequent Work.** Since the preliminary version of our work was published, several further results have emerged. For instance, [AT22] demonstrated that additional iterations do not result in further privacy costs after running DP-SGD on smooth convex functions for a small burn-in period. [GLL+23a, GLL+23b] extended our findings to non-Euclidean geometries, i.e., they extend our results to the setting where the functions $f_i(\cdot; s)$ are G-Lipschitz w.r.t some arbitrary norm $\|\cdot\|_X$. By adding a regularizer that is strongly convex w.r.t. $\|\cdot\|_X$, they show that one can recover similar results as this paper. [FYC23] generalizes the rejection sampler to semi-smooth functions.

3. Preliminaries

3.1. **Differential Privacy.** A DP algorithm $\mathcal{M}$ usually satisfies a collection of (ε, δ)-DP guarantees for each ε, i.e., for each ε there exists some smallest δ for which $\mathcal{M}$ is (ε, δ)-DP. By collecting all of them together, we can form the privacy curve or privacy profile which fully characterizes the privacy of a DP algorithm.

Definition 3.1 (Privacy Curve)**.** Given two random variables X, Y supported on some set Ω, define the privacy curve $\delta(X\|Y) : \mathbb{R}_{\geq 0} \to [0,1]$ as:

$$\delta(X\|Y)(\varepsilon) = \sup_{S \subset \Omega} \Pr[Y \in S] - e^{\varepsilon} \Pr[X \in S].$$

One can explicitly calculate the privacy curve of a Gaussian mechanism as

$$\delta(\mathcal{N}(0,1) \parallel \mathcal{N}(s,1))(\varepsilon) = \Phi\left(-\frac{\varepsilon}{s} + \frac{s}{2}\right) - e^{\varepsilon}\Phi\left(-\frac{\varepsilon}{s} - \frac{s}{2}\right), \tag{3.1}$$

where $\Phi(\cdot)$ is the Gaussian cumulative distribution function (CDF) [BW18].

We say a differentially private mechanism $\mathcal{M}$ has privacy curve $\delta : \mathbb{R}_{\geq 0} \to [0,1]$ if for every $\varepsilon \geq 0$, $\mathcal{M}$ is $(\varepsilon, \delta(\varepsilon))$-differentially private, i.e., $\delta(\mathcal{M}(\mathcal{D}) \| \mathcal{M}(\mathcal{D}'))(\varepsilon) \leq \delta(\varepsilon)$ for all neighbouring databases $\mathcal{D}, \mathcal{D}'$.

We will also need the notion of tradeoff function introduced in [DRS19], which is an equivalent way to describe the privacy curve $\delta(P\|Q)$.

Definition 3.2 (Tradeoff function)**.** Given two (continuous) distributions P, Q, we define the trade-off function[6] $T(P\|Q) : [0,1] \to [0,1]$ as

$$T(P\|Q)(z) = \inf_{S : P(S) = 1 - z} Q(S).$$

It is easy to compute explicitly the tradeoff function for Gaussian mechanism [DRS19],

$$T(\mathcal{N}(0,1)\|\mathcal{N}(s,1))(z) = \Phi(\Phi^{-1}(1-z) - s). \tag{3.2}$$

Note that perfect privacy is equivalent to the tradeoff function $\mathrm{Id}(z) = 1 - z$ and the closer a tradeoff function is to Id, better the privacy. The tradeoff function $T(P\|Q)$ and the privacy curve $\delta(P\|Q)$ are related via convex duality. Therefore to compare privacy curves, it is enough to compare tradeoff curves.

Proposition 3.3 ([DRS19])**.** $\delta(P\|Q) \leq \delta(P'\|Q')$ iff $T(P\|Q) \geq T(P'\|Q')$.

3.2. **Optimization.** Here we collect some properties of functions that are useful for optimization and sampling.

Definition 3.4 (L-Lipschitz continuity)**.** A function $f : \mathcal{K} \to \mathbb{R}$ is L-Lipschitz continuous over the domain $\mathcal{K} \subset \mathbb{R}^d$ if $|f(\omega) - f(\omega')| \leq L\|\omega - \omega'\|_2$ for all $\omega, \omega' \in \mathcal{K}$.

Definition 3.5 (μ-strong convexity)**.** A differentiable function $f : \mathcal{K} \to \mathbb{R}$ is called strongly convex with parameter $\mu > 0$ if $\mathcal{K} \subset \mathbb{R}^d$ is convex

$$f(\omega') \geq f(\omega) + \langle \nabla f(\omega), \omega' - \omega \rangle + \frac{\mu}{2}\|\omega' - \omega\|_2^2.$$

for all points $\omega, \omega' \in \mathcal{K}$.

Definition 3.6 (Log-concave measure and density)**.** A density function $f : \mathcal{K} \to \mathbb{R}_{\geq 0}$ is log-concave if $\int_{\mathcal{K}} f(x)dx = 1$ and $f(x) = \exp(-F(x))$ for some convex function F. We call f is μ-strongly log-concave if F is μ-strongly convex. Similarly, we call π a log-concave measure if its density function is log-concave, and we call π is a μ-strongly log-concave measure if its density function is μ-strongly log-concave.

[6] Tradeoff curves in [DRS19] are defined using type I and type II errors. The definition given here is equivalent to their definition for continuous distributions.

3.3. **Distribution Distance and Divergence.** We present some distribution distances or divergences mentioned or used in this work.

Definition 3.7. [Rén61, Rényi divergence] Suppose $1 < \alpha < \infty$ and π, ν are measures with $\pi \ll \nu$. The Rényi divergence of order α between π and ν is defined as

$$\mathrm{D}_\alpha(\pi \| \nu) = \frac{1}{\alpha} \log \int \left(\frac{\pi(x)}{\nu(x)} \right)^\alpha \nu(x) \mathrm{d}x.$$

We follow the convention that $0/0 = 0$. Rényi divergence of orders $\alpha = 1, \infty$ are defined by continuity. For $\alpha = 1$, the limit in Rényi divergence equals to the Kullback-Leibler divergence of π from ν, as defined next.

Definition 3.8 (Kullback–Leibler divergence)**.** The Kullback–Leibler divergence between probability measures π and ν is defined by

$$\mathrm{D}_{KL}(\pi \| \nu) = \int \log \left(\frac{\pi}{\nu} \right) \mathrm{d}\pi.$$

Definition 3.9 (Wasserstein distance)**.** Let π, ν be two probability distributions on $\mathbb{R}^d$. The second Wasserstein distance W_2 between π and ν is defined by

$$\mathrm{W}_2(\pi, \nu) = \big(\inf_{\gamma \in \Gamma(\pi, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2^2 \mathrm{d}\gamma(x, y) \big)^{1/2},$$

where $\Gamma(\pi, \nu)$ is the set of all couplings of π and ν.

Definition 3.10 (Total variation distance)**.** The total variation distance between two probability measures π and ν on a sigma-algebra $\mathcal{F}$ of subsets of the sample space Ω is defined via

$$\mathrm{TV}(\pi, \nu) = \sup_{S \in \mathcal{F}} |\pi(S) - \nu(S)|.$$

3.4. **Isoperimetric Inequality for Strongly Log-concave Distributions.** The cumulative distribution function (CDF) of the one-dimensional standard Gaussian distribution will be denoted by $\Phi(x) = \Pr_{y \sim \mathcal{N}(0,1)}[y \le x]$. The following lemma relates the expanding property of log-concave measures with Φ.

Proposition 3.11 (Theorem 1.1. in [Led99])**.** Let π be a μ-strongly log-concave measure supported on a convex set $\mathcal{K} \subseteq \mathbb{R}^d$. Let $A \subset \mathcal{K}$ by any subset such that $\pi(A) = z$. For any point $x \in \mathbb{R}^d$, define $d(x, A) = \inf_{y \in A} \|x - y\|_2$. Let $A_r = \{x : d(x, A) \le r\}$. Then if $A_r \subseteq \mathcal{K}$, for every $r \ge 0$,

$$\pi(A_r) \ge \Phi(\Phi^{-1}(z) + r\sqrt{\mu}).$$

The property above implies the concentration of Lipschitz functions over log-concave measures.

Corollary 3.12. Let π be a μ-strongly log-concave measure supported on a convex set $\mathcal{K} \subseteq \mathbb{R}^d$. Suppose $\alpha : \mathcal{K} \to \mathbb{R}$ is G-Lipschitz. For $z \in [0, 1]$, define $m(z) \in \mathbb{R}$ such that $\Pr_{x \sim \pi}[\alpha(x) \le m(z)] = z$. Then for every $r \ge 0$,

$$\Pr_{x \sim \pi}[\alpha(x) \ge m(z) + r] \le \Phi \left(\Phi^{-1}(1 - z) - \frac{r\sqrt{\mu}}{G} \right)$$

and

$$\Pr_{x\sim\pi}[\alpha(x) \le m(z) - r] \le \Phi\left(\Phi^{-1}(z) - \frac{r\sqrt{\mu}}{G}\right).$$

Proof. Fix some $z \in [0,1]$. Let $A = \{x \in \mathcal{K} : \alpha(x) \le m(z)\}$, so $\pi(A) = z$. Let $A_r = \{x : d(x,A) \le r\}$. Since α is G-Lipschitz, $\alpha(x) \ge m(z) + r$ implies that $d(x,A) \ge r/G$. Therefore $\{x : \alpha(x) \ge m(z) + r\} \subset \{x : d(x,A) \ge r/G\} = \overline{A_{r/G}}$ and so

$$\begin{aligned}\Pr_{x\sim\pi}[\alpha(x) \ge m(z) + r] &\le \pi(\overline{A_{r/G}})\\ &= 1 - \pi(A_{r/G})\\ &\le 1 - \Phi\left(\Phi^{-1}(z) + \frac{r\sqrt{\mu}}{G}\right)\\ &= \Phi\left(-\Phi^{-1}(z) - \frac{r\sqrt{\mu}}{G}\right).\end{aligned}$$

We obtain the other inequality by applying the above inequality to $-\alpha(x)$. □

4. GDP of Regularized Exponential Mechanism

In this section, we prove our DP result (Theorem 2.1). The proof uses the isoperimetric inequality for strongly log-concave measures [Led99]. Intuitively, the privacy loss random variable will be G-Lipschitz under the hypothesis and isoperimetric inequality implies that any Lipschitz function will be as concentrated as a Gaussian with an appropriate standard deviation. This allows us compare the privacy curve $\delta(P \parallel Q)$ to that of a Gaussian mechanism. In our proof, it is actually more convenient to compare tradeoff curves ($T(P \parallel Q)$) which are equivalent to privacy curves via convex duality (Proposition 3.3 and Theorem 2.1).

Theorem 4.1. *Given a convex set $\mathcal{K} \subseteq \mathbb{R}^d$ and μ-strongly convex functions $F, \tilde{F}$ over $\mathcal{K}$, let P, Q be distributions over $\mathcal{K}$ such that $P(x) \propto e^{-F(x)}$ and $Q(x) \propto e^{-\tilde{F}(x)}$. If $\tilde{F} - F$ is G-Lipschitz over $\mathcal{K}$, then for all $z \in [0,1]$,*

$$T(P \parallel Q)(z) \ge T\left(\mathcal{N}(0,1) \,\middle\|\, \mathcal{N}\left(\frac{G}{\sqrt{\mu}}, 1\right)\right)(z).$$

Proof. Let $\gamma = G/\sqrt{\mu}$. Let $\alpha(x) = \tilde{F}(x) - F(x)$ so that $Q(x) \propto e^{-\alpha(x)}P(x)$. Recall that we have $T(P\|Q)(z) = \inf_{S:P(S)=1-z} Q(S)$. Note that the infimum is achieved when we choose $S = \{x \in \mathcal{K} : \alpha(x) \ge m(z)\}$ for some $m(z)$ chosen such that $P(S) = \Pr_{x\sim P}[\alpha(x) \ge m(z)] = 1 - z$ (Neyman-Pearson lemma). Therefore,

$$T(P\|Q)(z) = \int_{x\in S} Q(x)\mathrm{d}x = \frac{\int_{x\in S} e^{-\alpha(x)}P(x)\mathrm{d}x}{\int_{x\in\mathcal{K}} e^{-\alpha(x)}P(x)\mathrm{d}x} = \left(1 + \frac{\mathbb{E}_P[e^{-\alpha}\mathbf{1}_{\bar{S}}]}{\mathbb{E}_P[e^{-\alpha}\mathbf{1}_S]}\right)^{-1}.$$

We now derive a lower bound for $\mathbb{E}_P[e^{-\alpha}\mathbf{1}_S]$. Let the random variable $Y = \alpha(x)$, where $x \sim P$. Let $f_Y(\cdot)$ be the PDF of Y. Then

$$\begin{aligned}
\mathop{\mathbb{E}}_P[e^{-\alpha(x)}\mathbf{1}_S] &= \int_{x:\alpha(x)\geq m(z)} e^{-\alpha(x)}P(x)\mathrm{d}x = \mathbb{E}[e^{-Y}\mathbf{1}(Y \geq m(z))] = \int_{m(z)}^{\infty} e^{-t} f_Y(t)dt \\
&= \int_{t=0}^{\infty} e^{-t-m(z)} \left(-\frac{\mathrm{d}\Pr_{x\sim P}\left[\alpha(x) \geq t+m(z)\right]}{\mathrm{d}t}\right)\mathrm{d}t \\
&= e^{-m(z)}\left(-e^{-t}\Pr_{x\sim P}\left[\alpha(x)\geq t+m(z)\right]\Big|_0^{\infty} - \int_{t=0}^{\infty} e^{-t}\Pr_{x\sim P}\left[\alpha(x)\geq t+m(z)\right]\mathrm{d}t\right) \\
&= (1-z)e^{-m(z)} - e^{-m(z)}\int_{t=0}^{\infty} e^{-t}\Pr_{x\sim P}\left[\alpha(x)\geq t+m(z)\right]\mathrm{d}t \\
&\geq (1-z)e^{-m(z)} - e^{-m(z)}\int_{t=0}^{\infty} e^{-t}\Phi(\Phi^{-1}(1-z)-t/\gamma)\mathrm{d}t \qquad \text{(Corollary 3.12)} \\
&= (1-z)e^{-m(z)} - e^{-m(z)}\left((1-z) - \exp\left(\frac{\gamma^2}{2} - \Phi^{-1}(1-z)\gamma\right)\Phi(\Phi^{-1}(1-z)-\gamma)\right) \\
&\qquad \text{(Claim 4.2)} \\
&= \exp\left(\frac{\gamma^2}{2} + \Phi^{-1}(z)\gamma - m(z)\right)\Phi(-\Phi^{-1}(z)-\gamma).
\end{aligned}$$

We next derive an upper bound for $\mathbb{E}_P[e^{-\alpha}\mathbf{1}_{\bar{S}}]$ in a similar way:

$$\begin{aligned}
\mathop{\mathbb{E}}_P[e^{-\alpha(x)}\mathbf{1}_{\bar{S}}] &= \int_{x:\alpha(x)\leq m(z)} e^{-\alpha(x)}P(x)\mathrm{d}x \\
&= \int_{t=0}^{\infty} e^{-m(z)+t}\left(-\frac{d\Pr_{x\sim P}\left[\alpha(x)\leq m(z)-t\right]}{dt}\right)\mathrm{d}t \\
&= e^{-m(z)}\left(-e^{t}\Pr_{x\sim P}\left[\alpha(x)\leq m(z)-t\right]\Big|_0^{\infty} + \int_{t=0}^{\infty} e^{t}\Pr_{x\sim P}\left[\alpha(x)\leq m(z)-t\right]\mathrm{d}t\right) \\
&= ze^{-m(z)} + e^{-m(z)}\int_{t=0}^{\infty} e^{t}\Pr_{x\sim P}\left[\alpha(x)\leq m(z)-t\right]\mathrm{d}t \\
&\leq ze^{-m(z)} + e^{-m(z)}\int_{t=0}^{\infty} e^{t}\Phi(\Phi^{-1}(z)-t/\gamma)\mathrm{d}t \qquad \text{(Corollary 3.12)} \\
&= ze^{-m(z)} + e^{-m(z)}\left(-z + \exp\left(\frac{\gamma^2}{2} + \Phi^{-1}(z)\gamma\right)\Phi(\Phi^{-1}(z)+\gamma)\right) \quad \text{(Claim 4.2)} \\
&= \exp\left(\frac{\gamma^2}{2} + \Phi^{-1}(z)\gamma - m(z)\right)\Phi(\Phi^{-1}(z)+\gamma).
\end{aligned}$$

Combining the two bounds,

$$\begin{aligned}
T(P\|Q)(z) &= \left(1 + \frac{\mathbb{E}_P[e^{-\alpha}\mathbf{1}_{\bar{S}}]}{\mathbb{E}_P[e^{-\alpha}\mathbf{1}_S]}\right)^{-1} \\
&\geq \left(1 + \frac{\Phi(\Phi^{-1}(z)+\gamma)}{\Phi(-\Phi^{-1}(z)-\gamma)}\right)^{-1} \\
&= \Phi(-\Phi^{-1}(z)-\gamma) \qquad \text{(Using } \Phi(x)+\Phi(-x)=1\text{)} \\
&= T(N(0,1)\parallel N(\gamma,1)). \qquad \text{(Eqn (3.2))}
\end{aligned}$$

□

We finish by calculating the integrals appearing in the proof.

Claim 4.2.
$$\int_0^\infty e^{-t}\Phi\left(a-\frac{t}{\gamma}\right)\mathrm{d}t = \Phi(a) - e^{\frac{\gamma^2}{2}-a\gamma}\Phi(a-\gamma)$$
and
$$\int_0^\infty e^{t}\Phi\left(a-\frac{t}{\gamma}\right)\mathrm{d}t = -\Phi(a) + e^{\frac{\gamma^2}{2}+a\gamma}\Phi(a+\gamma)$$

Proof. We have
$$\begin{aligned}\int_0^\infty e^{-t}\Phi(a-t/\gamma)\mathrm{d}t &= -e^{-t}\Phi(a-t/\gamma)\big|_0^\infty - \int_0^\infty e^{-t}\frac{e^{-(a-t/\gamma)^2/2}}{\gamma\sqrt{2\pi}}\mathrm{d}t\\ &= \Phi(a) - \int_0^\infty e^{\gamma^2/2-a\gamma}\frac{e^{-(t-(\gamma a-\gamma^2))^2/2}}{\gamma\sqrt{2\pi}}\mathrm{d}t\\ &= \Phi(a) - e^{\gamma^2/2-a\gamma}\Phi(a-\gamma).\end{aligned}$$
$$\begin{aligned}\int_0^\infty e^{t}\Phi(a-t/\gamma)\mathrm{d}t &= e^{t}\Phi(a-t/\gamma)\big|_0^\infty + \int_0^\infty e^{t}\frac{e^{-(a-t/\gamma)^2/2}}{\gamma\sqrt{2\pi}}\mathrm{d}t\\ &= -\Phi(a) + \int_0^\infty e^{\gamma^2/2+a\gamma}\frac{e^{-(t-(a\gamma+\gamma^2))^2/2\gamma^2}}{\gamma\sqrt{2\pi}}\mathrm{d}t\\ &= -\Phi(a) + e^{\gamma^2/2+a\gamma}\Phi(a+\gamma).\end{aligned}$$
□

As a corollary to Theorem 4.1, we can bound any divergence measure that decreases under post-processing such as Rényi divergence or KL divergence. In particular, this also implies Rényi differential privacy [Mir17] of our algorithm.

Corollary 4.3. Suppose $F, \tilde{F}$ are μ-strongly convex functions over $\mathcal{K} \subseteq \mathbb{R}^d$, and $F - \tilde{F}$ is G-Lipschitz over $\mathcal{K}$. For any $k > 0$, if we let $P \propto e^{-kF}$ and $Q \propto e^{-k\tilde{F}}$ be two probability distributions on $\mathcal{K}$, then
$$\mathrm{D}(P\|Q) \le \mathrm{D}\left(\mathcal{N}(0,1)\|\mathcal{N}\left(\frac{G\sqrt{k}}{\sqrt{\mu}},1\right)\right)$$
for any divergence measure D that decreases under post-processing. In particular,
$$\mathrm{D}_\alpha(P\|Q) \le \frac{\alpha k G^2}{2\mu} \text{ and } \mathrm{D}_{KL}(P\|Q) \le \frac{kG^2}{2\mu}.$$

Proof. By Theorem 2.10 in [DRS19], if $T(P\|Q) \ge T(X\|Y)$, then there exists a randomized algorithm M such that $M(X) = P$ and $M(Y) = Q$. Therefore for any divergence measure which decreases under post-processing we have,
$$\mathrm{D}(P\|Q) = \mathrm{D}(M(X)\|M(Y)) \le \mathrm{D}(X\|Y).$$
The rest follows from Theorem 4.1. It is well-known that Rényi divergence and KL divergence decrease with post-processing (see [VEH14], for example). We can also compute $\mathrm{D}_\alpha(\mathcal{N}(0,1), \mathcal{N}(s,1)) = \alpha s^2/2$ and $\mathrm{D}_{KL}(\mathcal{N}(0,1), \mathcal{N}(s,1)) = s^2/2$ [Mir17]. □

5. Efficient Non-smooth Sampling

In this section, we will present an efficient sampling scheme for (non-smooth) functions to complement our main results. Specifically, we study the following problem about sampling from a (non-smooth) log-concave distribution.

Problem 5.1. Assume that as given a μ-strongly convex function $\psi(x)$ defined on a convex set $\mathcal{K} \subseteq \mathbb{R}^d$ and $+\infty$ outside and a family of G-Lipschitz convex functions $\{f_i(\cdot)\}_{i\in I}$ defined on $\mathcal{K}$. Our goal is to sample a point $x \in \mathcal{K}$ with probability proportional to $\exp(-\widehat{F}(x))$, where

$$\widehat{F}(x) = \mathop{\mathbb{E}}_{i\in I} f_i(x) + \psi(x).$$

Our sampler is based on the alternating sampling algorithm in [LST21]; see algorithm 1. This algorithm reduces the problem of sampling from $\exp(-\widehat{F}(x))$ to sampling from $\exp(-\widehat{F}(x) - \|x-y\|^2/2\eta)$ for some fixed η and for roughly $1/\eta\mu$ many different y. When the step size η is very small, the later problem is easier because the distribution is almost like a Gaussian distribution. For our problem, we will pick the largest step size η such that we can sample $\exp(-\widehat{F}(x) - \|x-y\|^2/2\eta)$ using only $\widetilde{O}(1)$ many steps.

Algorithm 1: Alternating Sampler

1 **Input:** μ-strongly convex function $\widehat{F}$, step size $\eta > 0$, initial point x_0
2 **for** $t \in [T]$ **do**
3 $\quad y_t \leftarrow x_{t-1} + \sqrt{\eta} \cdot \zeta$, where $\zeta \sim \mathcal{N}(0, I_d)$.
4 $\quad$ Sample $x_t \propto \exp(-\widehat{F}(x) - \|x - y_t\|_2^2/2\eta)$.
5 **end**
6 **Return** x_T

Theorem 5.2 ([LST21, Theorem 1])**.** *Let F be a μ-strongly convex function defined on $\mathcal{K}$ with an initial point x_0. Let the distance $D = \|x_0 - x^*\|_2$ for any $x^* = \arg\min_{x\in\mathcal{K}} \widehat{F}(x)$. Suppose the step size $\eta \leq \frac{1}{\mu}$, the target accuracy $\delta > 0$ and the number of step $T \geq \Theta(\frac{1}{\eta\mu}\log(\frac{d/\mu + D^2}{\eta\delta}))$. Then, Algorithm 1 returns a random point x_T that has δ total variation distance to the distribution proportional to $\exp(-\widehat{F}(x))$.*

Now, we show that Line 4 in Algorithm 1 can be implemented by a simple rejection sampling. The idea is to pick step size η small enough such that $\widehat{F}(x)$ is essentially a constant function for a random $x \sim \mathcal{N}(y, \eta \cdot I_d)$. The precise algorithm is given in Algorithm 2.

Since F has the ψ term, instead of sampling x from $\mathcal{N}(y, \eta \cdot I_d)$, we sample from $\exp(-\psi(x) - \|x-y\|^2/2\eta)$ in Algorithm 2. The following lemma shows how to decompose the distribution $\exp(-\widehat{F}(x) - \|x-y\|^2/2\eta)$ into the distribution mentioned above and the distribution $\exp(-\mathbb{E}_{i\in I} f_i(x))$. It also calculates the distribution given by the algorithm.

Lemma 5.1. *Let π be the distribution proportional to $\exp(-\widehat{F}(x) - \|x-y\|_2^2/2\eta)$ and let $\mathcal{G}$ be the distribution proportional to $\exp(-\psi(x) - \|x-y\|^2)/2\eta$. Then,*

$$\frac{d\pi}{dx} = \frac{d\mathcal{G}}{dx} \cdot \frac{\exp(-\mathbb{E}_{i\in I} f_i(x))}{\mathbb{E}_{x\sim\mathcal{G}} \exp(-\mathbb{E}_{i\in I} f_i(x))}.$$

Algorithm 2: Implementation of Line 4

Input: convex function $\widehat{F}(x) = \mathbb{E}_{i \in I} f_i(x) + \psi(x)$, step size $\eta > 0$, current point y
repeat
  Sample x, z from the distribution $\propto \exp(-\psi(x) - \|x-y\|_2^2/2\eta)$
  Set $\rho \leftarrow 1$
  for $\alpha = 1, 2, \cdots$ **do**
    $\rho \leftarrow \rho + \Pi_{i=1}^{\alpha}(f_{j_i}(z) - f_{j_i}(x))$ where j_i are random indices in I
    With probability $\alpha/(1+\alpha)$, **break**
  end
  Sample u uniformly from $[0, 1]$.
until $u \leq \rho/2$;
Return x

Let $\widetilde{\pi}$ be the distribution returned by Algorithm 2. Then,

$$\frac{d\widetilde{\pi}}{dx} = \frac{d\mathcal{G}}{dx} \cdot \frac{\mathbb{E}(\overline{\rho}|x)}{\mathbb{E}(\overline{\rho})},$$

where $\overline{\rho} = \min(\max(\rho, 0), 2)$ is the truncation of ρ in Algorithm 2 to $[0, 2]$, $\mathbb{E}(\overline{\rho}|x)$ is the expected value of $\overline{\rho}$ conditional on x, and $\mathbb{E}(\overline{\rho}) = \mathbb{E}_{x \sim \mathcal{G}} \mathbb{E}(\overline{\rho}|x)$. Furthermore,

$$\mathbb{E}(\rho|x) = \exp(-\mathop{\mathbb{E}}_{i \in I} f_i(x)) \cdot \mathop{\mathbb{E}}_{z \sim \mathcal{G}} \exp(\mathop{\mathbb{E}}_{i \in I} f_i(z)).$$

Proof. For the true distribution π,

$$\begin{aligned}
\frac{d\pi}{dx} &= \frac{\exp(-\mathbb{E}_{i \in I} f_i(x) - \psi(x) - \frac{1}{2\eta}\|x-y\|_2^2)}{\int \exp(-\mathbb{E}_{i \in I} f_i(x) - \psi(x) - \frac{1}{2\eta}\|x-y\|_2^2) dx} \\
&= \frac{\exp(-\mathbb{E}_{i \in I} f_i(x)) \frac{d\mathcal{G}}{dx}}{\int \exp(-\mathbb{E}_{i \in I} f_i(x)) \frac{d\mathcal{G}}{dx} dx} \\
&= \frac{d\mathcal{G}}{dx} \cdot \frac{\exp(-\mathbb{E}_{i \in I} f_i(x))}{\mathbb{E}_{x \sim \mathcal{G}} \exp(-\mathbb{E}_{i \in I} f_i(x))}.
\end{aligned}$$

For the distribution $\widetilde{\pi}$ by the algorithm, we sample $x \sim \mathcal{G}$, then accept the sample if $u \leq \rho/2$. Hence, we have

$$\frac{d\widetilde{\pi}}{dx} = \frac{d\mathcal{G}}{dx} \frac{\Pr(u \leq \frac{1}{2}\rho|x)}{\Pr(u \leq \frac{1}{2}\rho)}.$$

Since u is uniform between 0 and 1, we have the result.

Finally, for the expectation of ρ, we note that

$$\mathbb{E}\, \Pi_{i=1}^{\alpha}(f_{j_i}(z) - f_{j_i}(x)) = (\mathop{\mathbb{E}}_{i \in I}(f_i(z) - f_i(x)))^{\alpha}$$

and that the probability that the loop pass step α is exactly $\frac{1}{\alpha!}$. Hence,

$$\mathbb{E}(\rho|x, z) = 1 + \sum_{\alpha=1}^{\infty} \frac{1}{\alpha!}(\mathop{\mathbb{E}}_{i \in I}(f_i(z) - f_i(x)))^{\alpha} = \exp(\mathop{\mathbb{E}}_{i \in I}(f_i(z) - f_i(x))).$$

Taking expectation over z gives the result. $\square$

Note that if we always had $0 \leq \rho \leq 2$, then $\mathbb{E}(\overline{\rho}|x) = \mathbb{E}(\rho|x) \propto \exp(-\mathbb{E}_{i\in I} f_i(x))$ and hence $d\pi/dx = \mathrm{d}\widetilde{\pi}/dx$. Therefore, the only thing left is to show that $0 \leq \rho \leq 2$ with high probability and that it does not induces too much error in total variation distance. To do this, we use Gaussian concentration to prove that $\mathbb{E}_{i\in I} f_i(x)$ is almost a constant over random $x \sim \mathcal{G}$.

Lemma 5.2 (Gaussian concentration [Led99, Eq 1.21])**.** *Let $X \sim \exp(-\widehat{F})$ for some $1/\eta$-strongly convex $\widehat{F}$ and let ℓ be a G-Lipschitz function. Then, for all $t \geq 0$,*

$$\Pr[\ell(X) - \mathbb{E}[\ell(X)] \geq t] \leq e^{-t^2/(2\eta G^2)}.$$

Now, we are already to prove our main result. This shows that if $\eta \ll G^{-2}$, then the algorithm indeed implements Line 4 correctly up to small error.

Lemma 5.3. *If $\eta \leq C \log^{-1}(1/\delta_{\text{inner}}) G^{-2}$ for some small enough C and the inner accuracy $\delta_{\text{inner}} \in (0, 1/2)$, then Algorithm 2 returns a random point x that has δ_{inner} total variation distance to the distribution proportional to $\exp(-\widehat{F}(x) - \|x-y\|_2^2/2\eta)$. Furthermore, the algorithm accesses only $O(1)$ many $f_i(x)$ in expectation and samples from $\exp(-\psi(x) - \|x - y\|_2^2/2\eta)$ for $O(1)$ many y.*

Proof. Let π be the distribution given by $c\cdot\exp(-\widehat{F}(x)-\|x-y\|_2^2/2\eta)$ and $\widetilde{\pi}$ is the distribution outputted by the algorithm. By Lemma 5.1,

$$\begin{aligned} d_{\mathrm{TV}}(\pi, \widetilde{\pi}) &= \int_{\mathbb{R}^d} \left| \frac{d\mathcal{G}}{dx} \frac{\exp(-\mathbb{E}_{i\in I} f_i(x))}{\mathbb{E}_{x\sim\mathcal{G}} \exp(-\mathbb{E}_{i\in I} f_i(x))} - \frac{d\mathcal{G}}{dx} \frac{\mathbb{E}(\overline{\rho}|x)}{\mathbb{E}(\overline{\rho})} \right| dx \\ &= \mathop{\mathbb{E}}_{x\sim\mathcal{G}} \left| \frac{\exp(-\mathbb{E}_{i\in I} f_i(x))}{\mathbb{E}_{x\sim\mathcal{G}} \exp(-\mathbb{E}_{i\in I} f_i(x))} - \frac{\mathbb{E}(\overline{\rho}|x)}{\mathbb{E}(\overline{\rho})} \right|. \end{aligned}$$

Let X be the random variable $\mathbb{E}(\rho|x)$ and $\widetilde{X}$ be the random variable $\mathbb{E}(\overline{\rho}|x)$. Lemma 5.1 shows that $X = \exp(-\mathbb{E}_{i\in I} f_i(x)) \cdot \mathbb{E}_{z\sim\mathcal{G}} \exp(\mathbb{E}_{i\in I} f_i(z))$ and hence

$$\frac{\exp(-\mathbb{E}_{i\in I} f_i(x))}{\mathbb{E}_{x\sim\mathcal{G}} \exp(-\mathbb{E}_{i\in I} f_i(x))} = \frac{X}{\mathbb{E}_{x\sim\mathcal{G}} X}.$$

Therefore, noting that $\widetilde{X}$ is nonnegative,

$$d_{\mathrm{TV}}(\pi, \widetilde{\pi}) = \mathbb{E}\left| \frac{X}{\mathbb{E} X} - \frac{\widetilde{X}}{\mathbb{E}\widetilde{X}} \right| \leq \mathbb{E}\left| \frac{X}{\mathbb{E} X} - \frac{\widetilde{X}}{\mathbb{E} X} \right| + \mathbb{E}\left| \frac{\widetilde{X}}{\mathbb{E} X} - \frac{\widetilde{X}}{\mathbb{E}\widetilde{X}} \right| \leq 2\frac{\mathbb{E}|X - \widetilde{X}|}{|\mathbb{E} X|}. \tag{5.1}$$

We simplify the right-hand side using a lower bound on $\mathbb{E} X$. By Lemma 5.2 and the fact that the negative log-density of $\mathcal{G}$ is $1/\eta$-strongly convex, we have that $\mathbb{E}_{i\in I} f_i(z) \geq \mathbb{E}_{x\sim\mathcal{G}} \mathbb{E}_{i\in I} f_i(x) - 2G\sqrt{\eta}$ with probability $\geq 1 - e^{-2}$. Hence, we have

$$\begin{aligned} \mathbb{E} X &= \mathop{\mathbb{E}}_{x\sim\mathcal{G}} \exp(-\mathop{\mathbb{E}}_{i\in I} f_i(x)) \cdot \mathop{\mathbb{E}}_{z\sim\mathcal{G}} \exp(\mathop{\mathbb{E}}_{i\in I} f_i(z)) \\ &\geq \exp(-\mathop{\mathbb{E}}_{x\sim\mathcal{G}} \mathop{\mathbb{E}}_{i\in I} f_i(x)) \cdot \mathop{\mathbb{E}}_{z\sim\mathcal{G}} \exp(\mathop{\mathbb{E}}_{i\in I} f_i(z)) \\ &= \mathop{\mathbb{E}}_{z\sim\mathcal{G}} \exp(\mathop{\mathbb{E}}_{i\in I} f_i(z) - \mathop{\mathbb{E}}_{x\sim\mathcal{G}} \mathop{\mathbb{E}}_{i\in I} f_i(x)) \\ &\geq (1 - e^{-2}) \exp(-2G\sqrt{\eta}). \end{aligned}$$

Using $\eta \leq G^{-2}/8$, we have $\mathbb{E}[X] \geq 2/3$. From this inequality, (5.1), $X = \mathbb{E}(\rho|x)$, and $\widetilde{X} = \mathbb{E}(\overline{\rho}|x)$,

$$d_{\mathrm{TV}}(\pi, \widetilde{\pi}) \leq 3 \cdot \mathbb{E}|X - \widetilde{X}| \leq 3 \cdot \mathbb{E}(|\rho| \cdot 1_{\rho\notin[0,2]}).$$

We split the ρ into two terms $\rho_{\le L}$ and $\rho_{>L}$. The first term $\rho_{\le L}$ is the sum of all terms added to ρ when $\alpha \le L$ (including the initial term 1). The second term $\rho_{>L}$ is the sum when $\alpha > L$. Hence, we have $\rho = \rho_{>L} + \rho_{\le L}$, so that

$$d_{\mathrm{TV}}(\pi, \widetilde{\pi}) \le 3 \cdot \mathbb{E}(|\rho_{>L}| \cdot 1_{\rho \notin [0,2]}) + 3 \cdot \mathbb{E}(|\rho_{\le L}| \cdot 1_{\rho \notin [0,2]}). \tag{5.2}$$

For the term $\rho_{>L}$, by a calculation similar to Lemma 5.1,

$$\mathbb{E}(|\rho_{>L}| \cdot 1_{\rho \notin [0,2]}) \le \mathbb{E}\,|\rho_{>L}| \le \mathop{\mathbb{E}}_{x,z} \Phi(\mathop{\mathbb{E}}_{i \in I} |f_i(z) - f_i(x)|),$$

where $\Phi(t) = \sum_{\alpha=L+1}^{\infty} t^\alpha/\alpha!$ is a power series in t with positive coefficients. By picking $L > C \log(1/\delta_{\mathrm{inner}})$ for some large constant C, we have $\Phi(t) \le \delta_{\mathrm{inner}}/16$ for all $|t| \le 1$. Let Δ be the random variable $\mathbb{E}_{i \in I} |f_i(z) - f_i(x)|$, whose randomness comes from x and z. Then,

$$\mathbb{E}(|\rho_{>L}| \cdot 1_{\rho \notin [0,2]}) \le \frac{\delta_{\mathrm{inner}}}{16} + \mathbb{E}\, e^{\Delta} 1_{\Delta \ge 1} \le \frac{\delta_{\mathrm{inner}}}{16} + \sum_{k=1}^{\infty} e^{k+1} \Pr_{x,z}(\Delta \ge k).$$

Let $h_{x,z}(t) := \Pr_{i \in I}[|f_i(z) - f_i(x)| \ge t]$. Since each f_i is G-Lipschitz, Lemma 5.2 shows that

$$\Pr_{x,z}[|f_i(z) - f_i(x)| \ge t] \le 4e^{-t^2/(8\eta G^2)},$$

which implies that

$$\mathop{\mathbb{E}}_{x,z}[h_{x,z}(t)] = \Pr_{x,z,i}[|f_i(z) - f_i(x)| \ge t] \le 4e^{-t^2/(8\eta G^2)}.$$

By Markov's inequality, for any $k > 0$,

$$\Pr_{x,z}[h_{x,z}(t) \ge e^{-k}] \le 4e^{k - t^2/(8\eta G^2)}.$$

As $|f_i(z) - f_i(x)| \le G\|x - z\|_2$, if $h_{x,z}(t) = \Pr_{i \in I}[|f_i(z) - f_i(x)| \ge t] \le e^{-t^2/(16\eta G^2)}$, we know

$$\mathop{\mathbb{E}}_{i \in I} |f_i(z) - f_i(x)| \le t + e^{-t^2/(16\eta G^2)} \cdot G\|x - z\|_2.$$

Hence, one has

$$\begin{aligned} \Pr_{x,z}\left[\mathop{\mathbb{E}}_{i \in I} |f_i(z) - f_i(x)| \ge t + e^{-t^2/(16\eta G^2)} G\|x - z\|_2\right] &\le \Pr_{x,z}[h_{x,z}(t) \ge e^{-t^2/(16\eta G^2)}] \\ &\le 4e^{-t^2/(16\eta G^2)}. \end{aligned}$$

By Gaussian concentration,

$$\Pr_{x,z}[\|x - z\|_2 \ge t] \le \Pr_{x,z}[\|x - \mathbb{E}\,x\|_2 \ge t/2 \text{ or } \|z - \mathbb{E}\,z\| \ge t/2] \le 2e^{-t^2/(8\eta)}.$$

Thus,

$$\begin{aligned}
&\Pr_{x,z}[\mathop{\mathbb{E}}_{i\in I}|f_i(z)-f_i(x)|\geq 2t] \\
&= Pr_{x,z}[\mathop{\mathbb{E}}_{i\in I}|f_i(z)-f_i(x)|\geq 2t, \|x-z\|_2\geq t/G] \\
&\qquad +\Pr_{x,z}[\mathop{\mathbb{E}}_{i\in I}|f_i(z)-f_i(x)|\geq 2t, \|x-z\|_2< t/G] \\
&\leq 2e^{-t^2/(8G^2\eta)}+\Pr_{x,z}[\mathop{\mathbb{E}}_{i\in I}|f_i(z)-f_i(x)|\geq 2t, \|x-z\|_2< t/G] \\
&\leq 2e^{-t^2/(8G^2\eta)}+\Pr_{x,z}[\mathop{\mathbb{E}}_{i\in I}|f_i(z)-f_i(x)|\geq t+e^{-t^2/(16\eta G^2)}G\|x-z\|_2] \\
&\leq 6e^{-t^2/(16\eta G^2)}.
\end{aligned}$$

Hence, we have $\Pr(\Delta\geq k)\leq 6\exp(-k^2/(64G^2\eta))$ and

$$\mathbb{E}(|\rho_{>L}|\cdot 1_{\rho\notin[0,2]})\leq\frac{\delta_{\text{inner}}}{16}+17\sum_{k=1}^{\infty}e^{k-\frac{k^2}{64G^2\eta}}\leq\frac{\delta_{\text{inner}}}{9}, \tag{5.3}$$

where we used $\eta\leq 2^{-6}G^{-2}/\log(400/\delta_{\text{inner}})$ at the end.

As for the term $\rho_{\leq L}$,

$$\begin{aligned}
\mathbb{E}(|\rho_{\leq L}|\cdot 1_{\rho\notin[0,2]}) &= \mathbb{E}(|\rho_{\leq L}|\cdot 1_{\rho\notin[0,2]}\cdot 1_{|\rho_{\leq L}|\leq 2^L})+\mathbb{E}(|\rho_{\leq L}|\cdot 1_{\rho\notin[0,2]}\cdot 1_{|\rho_{\leq L}|\geq 2^L}) \\
&\leq \Pr[\rho\notin[0,2]]\cdot 2^L+\sum_{k=1}^{\infty}2^{(k+1)L}\Pr(|\rho_{\leq L}|\geq 2^{kL}). \qquad (5.4)
\end{aligned}$$

Note that the term $\rho_{\leq L}$ involves only fewer than $L^2/2$ many $f_i(x)$ and $f_i(z)$. Lemma 5.2 shows that for any i,

$$\Pr_{x\sim\mathcal{G}}(|f_i(x)-\mathop{\mathbb{E}}_{x\sim\mathcal{G}}f_i(x)|\geq t)\leq 2e^{-t^2/(2\eta G^2)}.$$

By the union bound,

$$\Pr_{x,z\sim\mathcal{G}}(|f_i(x)-f_i(z)|\geq\frac{1}{4}2^k \text{ for any such } i)\leq L^2\exp(-4^k/32\eta G^2).$$

Under the event that $|f_i(x)-f_i(z)|\leq 2^k/3$ for all i appears in $\rho_{\leq L}$,

$$|\rho_{\leq L}|\leq 1+\sum_{\alpha=1}^{L}\Pi_{i=1}^{\alpha}|f_{j_{i,\alpha}}(z)-f_{j_{i,\alpha}}(x)|\leq 1+\sum_{\alpha=1}^{L}(\frac{2^k}{3})^{\alpha}\leq 2^{kL}.$$

Therefore, $\Pr(|\rho_{\leq L}|>2^{kL})\leq L^2\exp(-4^k/32\eta G^2)$ and

$$\sum_{k=1}^{\infty}2^{(k+1)L}\Pr(|\rho_{\leq L}|>2^{kL})\leq\sum_{k=1}^{\infty}2^{(k+1)L}L^2\exp(-\frac{4^k}{32\eta G^2})\leq\sum_{k=1}^{\infty}2^{4kL}\exp(-\frac{4^k}{32\eta G^2}).$$

Picking $\eta\leq 2^{-8}G^{-2}L^{-1}$,

$$\sum_{k=1}^{\infty}2^{(k+1)L}\Pr(|\rho_{\leq L}|>2^{kL})\leq\sum_{k=1}^{\infty}2^{4kL}\exp(-2\cdot 4^kL)\leq\sum_{k=1}^{\infty}2^{-kL}\leq\frac{\delta_{\text{inner}}}{9} \tag{5.5}$$

by picking $L>C\log(1/\delta_{\text{inner}})$ for large enough C.

It remains to bound the term $\Pr[\rho \notin [0,2]] \cdot 2^L$. We know the probability the algorithm enters the $(L+1)$-th phase is at most $1/L!$. Hence we know $\Pr[\rho \notin [0,2]] \leq 1/L! + \Pr[\rho_{\leq L} \notin [0,2]]$. Similarly, by Gaussian Concentration and union bound, we have

$$\Pr_{x,z\sim\mathcal{G}}(|f_i(x) - f_i(z)| \geq 1/2 \text{ for any such } i) \leq L^2 \exp(-\frac{1}{8\eta G^2}).$$

Under the event that $|f_i(x) - f_i(z)| \leq 1/2$ for all i appears in $\rho_{\leq L}$,

$$1 - \sum_{\alpha=1}^{L} \Pi_{i=1}^{\alpha} |f_{j_{i,\alpha}}(z) - f_{j_{i,\alpha}}(x)| \leq \rho_{\leq L} \leq 1 + \sum_{\alpha=1}^{L} \Pi_{i=1}^{\alpha} |f_{j_{i,\alpha}}(z) - f_{j_{i,\alpha}}(x)|,$$

which implies $0 \leq \rho_{\leq L} \leq 2$. Then we know $\Pr[\rho_{\leq L} \notin [0,2]] \leq L^2 \exp(-1/8\eta G^2)$. By our setting of parameters and that $L = C \log(1/\delta_{\text{inner}})$ for some large constant C, we know

$$\Pr[\rho \notin [0,2]] \cdot 2^L \leq 2^L (L^2 \exp(-\frac{1}{8\eta G^2}) + \frac{1}{L!}) \leq \frac{\delta_{\text{inner}}}{9}. \tag{5.6}$$

Combining (5.2), (5.3), (5.4), (5.5) and (5.6), we have the result that $d_{\text{TV}}(\pi, \widetilde{\pi}) \leq \delta_{\text{inner}}$.

Finally, the acceptance probability is given by $\mathbb{E}\,\widetilde{X}/2$ and $\mathbb{E}\,\widetilde{X} \geq \mathbb{E}\,X - \mathbb{E}\,|X - \widetilde{X}| \geq 2/3 - \delta_{\text{inner}}/3 \geq 1/3$. Hence, the number of accesses is $O(1)$. □

Combining Theorem 5.2 and Lemma 5.3, we have the following result:

Theorem 5.3. *Let $\psi(x)$ be a μ-strongly convex function defined on a convex set $\mathcal{K} \subseteq \mathbb{R}^d$ and defined to be $+\infty$ outside. Assume given a family of G-Lipschitz convex functions $\{f_i(x)\}_{i\in I}$ defined on $\mathcal{K}$. Define the function $\widehat{F}(x) = \mathbb{E}_{i\in I} f_i(x) + \psi(x)$ and the distance $D = \|x_0 - x^*\|_2$ for some $x^* = \arg\min_x \widehat{F}(x)$. For any $\delta \in (0, 1/2)$, if we can get samples from $\exp(-\psi(x) - \|x-y\|_2^2/2\eta)$ for any $y \in \mathbb{R}^d$ and $\eta > 0$, we can find a random point x that has δ total variation distance to the distribution proportional to $\exp(-\widehat{F}(x))$ in*

$$T := \Theta(\frac{G^2}{\mu} \log^2(\frac{G^2(d/\mu + D^2)}{\delta}))$$

steps. Furthermore, each steps accesses only $O(1)$ many $f_i(x)$ in expectation and samples from $\exp(-\psi(x) - \|x-y\|_2^2/2\eta$ for $O(1)$ many y with $\eta = \Theta(G^{-2}/\log(T/\delta))$.

Proof. This follows from applying Lemma 5.3 to implement Line 4. Note that the distribution implemented has total variation distance δ_{inner} to the required one. By setting $\delta_{\text{inner}} = \delta/(2T)$, this only gives an extra $\delta/2$ error in total variation distance. Finally, setting $\eta = \Theta(G^{-2}/\log(1/\delta_{\text{inner}}))$, Theorem 5.2 shows that Algorithm 2 outputs the correct distribution up to $\delta/2$ error in total variation distance. This gives the result. □

In the most important case of interest when $\psi(x)$ is ℓ_2^2 regularizer, one can see $\exp(-\psi(x) - \|x-y\|_2^2/2\eta)$ is a truncated Gaussian distribution, and there are many results on how to sample from truncated Gaussian, e.g. [KD99]. For more general case, there are also efficient algorithms to do the sampling, such as projected Langevin Monte Carlo [BEL18]. In fact, our sampling scheme matches the information-theoretical lower bound on the value query complexity up to some logarithmic terms, which can be reduced from the result in [DJWW15] with some modifications. See Section 7 for more detailed discussion.

6. DP Convex Optimization

In this section we present our results about DP-ERM and DP-SCO.

6.1. **DP-ERM.** In this subsection, we state our result for the DP-ERM problem (1.3). Briefly, our main result (Theorem 2.1) shows that sampling from $\exp(-kF(x;\mathcal{D}))$ for some appropriately chosen k is (ε,δ)-DP and achieves the optimal empirical risk in (1.4). Our sampling scheme in Section 5 provides an efficient implementation. We start with the following lemma which shows the utility guarantee for the sampling mechanism.

Lemma 6.1 (Utility Guarantee, [DKL18, Corollary 1])**.** Suppose $k>0$ and F is a convex function over the convex set $\mathcal{K}\subseteq\mathbb{R}^d$. If we sample x according to distribution ν whose density is proportional to $\exp(-kF(x))$, then we have

$$\mathop{\mathbb{E}}_{\nu}[F(x)] \le \min_{x\in\mathcal{K}} F(x) + \frac{d}{k}.$$

This is first shown by [KV06] for any linear function F, and [BST14] extends it to any convex function F, but with a slightly worse constant.

Theorem 6.2 (DP-ERM)**.** *Suppose that $\varepsilon>0$, that $\mathcal{K}\subseteq\mathbb{R}^d$ is a convex set of diameter D, and that $\{f(\cdot;s)\}_{s\in\mathcal{D}}$ is a family of convex functions over $\mathcal{K}$ such that $f(\cdot;s)-f(\cdot;s')$ is G-Lipschitz for all s,s'. For any data-set $\mathcal{D}$ and $k>0$, sampling $x^{(priv)}$ with probability proportional to $\exp\left(-k(F(x;\mathcal{D})+\mu\|x\|_2^2/2)\right)$ is $(\varepsilon,\delta(\varepsilon))$-differentially private, where*

$$\delta(\varepsilon)\le\delta\left(\mathcal{N}(0,1)\,\middle\|\,\mathcal{N}\left(\frac{G\sqrt{k}}{n\sqrt{\mu}},1\right)\right)(\varepsilon).$$

The excess empirical risk is bounded by $d/k+\mu D^2/2$. Moreover, if $\{f(\cdot,s)\}_{s\in\mathcal{D}}$ are μ-strongly convex, then sampling $x^{(priv)}$ with probability proportional to $\exp(-kF(x;\mathcal{D}))$ is $(\varepsilon,\delta(\varepsilon))$-differentially private, where

$$\delta(\varepsilon)\le\delta\left(\mathcal{N}(0,1)\,\middle\|\,\mathcal{N}\left(\frac{G\sqrt{k}}{n\sqrt{\mu}},1\right)\right)(\varepsilon).$$

The excess empirical risk is bounded by d/k.

Proof. The privacy guarantee follows directly from our main result Theorem 2.1, and the bound on excess empirical loss can be proved by Lemma 6.1. $\square$

Before we state the implementation results on DP-ERM, we need the following technical lemma.

Lemma 6.3. For any constants $1/2>\delta>0$ and $\varepsilon>0$, if $|s|\le\sqrt{2\log(1/(2\delta))+2\varepsilon}-\sqrt{2\log(1/(2\delta))}$,

$$\delta(\mathcal{N}(0,1)\,\|\,\mathcal{N}(s,1))\le\delta.$$

Proof. By Equation (3.1), we know that

$$\delta(\mathcal{N}(0,1)\,\|\,\mathcal{N}(s,1))(\varepsilon)\le\Phi\left(-\frac{\varepsilon}{s}+\frac{s}{2}\right).$$

Without loss of generality, we assume $s\ge0$ and want to find an appropriate value of s such that $\Phi\left(-\varepsilon/s+s/2\right)\ leq\delta$. Let $t\stackrel{\text{def}}{=}\Phi^{-1}(1-\delta)$ and since $1-\Phi(t)\le\exp(-t^2/2)/2$ for $t>0$, we know that $t \leq \sqrt{2\log(1/(2\delta))}$. It is equivalent to solve the equation $eps/s - s/2 \geq t$, which is equivalent to $0 \leq s \leq \sqrt{t^2+2\varepsilon} - t$. Note that $\sqrt{t^2+2\varepsilon}-t$ decreases as t increases, which implies that we can set $s \leq \sqrt{2\log(1/(2\delta))+2\varepsilon} - \sqrt{2\log(1/(2\delta))}$. $\square$

Combining the sampling scheme (Theorem 5.3) and our analysis on DP-ERM, we can get the efficient implementation results on DP-ERM directly.

Theorem 6.4 (DP-ERM Implementation)**.** *Impose the assumptions of Theorem 6.2, and assume $f(\cdot;s)$ is G-Lipschitz over $\mathcal{K}$ for all s. Then, for any constants $\delta \in (0,1/10)$ and $\varepsilon > 0$, there is an efficient sampler to solve DP-ERM that has the following guarantees:*

- *The scheme is (ε,δ)-differentially private;*
- *The expected excess empirical loss is bounded by*
$$\frac{GD\sqrt{d}}{n(\sqrt{\log(1/\delta)+\varepsilon}-\sqrt{\log(1/\delta)})}.$$
In particular, if $\varepsilon < 1/10$, the expected excess empirical loss is bounded by $2GD\sqrt{d\log(1/\delta)}/\varepsilon n$. If $\varepsilon \geq \log(1/\delta)$, the expected excess empirical loss is bounded by $O(GD\sqrt{d}/n\sqrt{\varepsilon})$.
- *The scheme takes*
$$\Theta\left(\frac{\varepsilon^2 n^2}{\log(1/\delta)}\log^2(\frac{nd\varepsilon}{\delta})\right)$$
queries to the values on $f(x;s)$ in expectation and takes the same number of samples from some Gaussian restricted to the convex set $\mathcal{K}$.

Proof. By Lemma 6.3, we can set $s = \sqrt{2\log(3/(4\delta))+2\varepsilon} - \sqrt{2\log(3/(4\delta))}$ to make $\delta(\mathcal{N}(0,1)\parallel\mathcal{N}(s,1)) \leq 2\delta/3$. For our setting, Theorem 6.2 shows that $s = G\sqrt{k}/n\sqrt{\mu}$, and hence we can take
$$k = \frac{2\mu n^2\left(\sqrt{\log(3/(4\delta))+\varepsilon}-\sqrt{\log(3/(4\delta))}\right)^2}{G^2}.$$
Putting this into the excess empirical loss bound of $d/k + \mu D^2/2$ and setting
$$\mu = \frac{G\sqrt{d}}{nD\left(\sqrt{\log(3/(4\delta))+\varepsilon}-\sqrt{\log(3/(4\delta))}\right)},$$
we get the result on the empirical loss.

In particular, consider the case that $\varepsilon < 1/10$. We know the excess empirical loss is bounded by $GD\sqrt{d}/n\left(\sqrt{\log(3/(4\delta))+\varepsilon}-\sqrt{\log(3/(4\delta))}\right)$. Note that $1+x/2-x^2/8 \leq \sqrt{1+x} \leq 1+x/2$ for $x \geq 0$. Under the assumption that $\delta,\varepsilon \in (0,1/10)$, we know $GD\sqrt{d}/n(\sqrt{\log(3/(4\delta))+\varepsilon}-\sqrt{\log(3/(4\delta))}) \leq 2GD\sqrt{d\log(4/(5\delta))}/n\varepsilon$. The case when $\varepsilon \geq \log(1/\delta)$ follows similarly.

To make it algorithmic, we apply Theorem 5.3 with the accuracy on the total variation distance to be $\min\{\delta/3, 1/cn^c\varepsilon\}$ for some sufficiently constant c. This leads to (ε,δ)-DP and an extra empirical loss and hence we use $\log(1/\delta)$ rather than $\log(3/(4\delta))$ or $\log(4/(5\delta))$ in the final loss term.

The running time follows from Theorem 5.3. $\square$

6.2. **DP-SCO and Generalization Error.** As mentioned before, one can reduce the DP-SCO (1.5) to DP-ERM (1.3) by the iterative localization technique proposed by [FKT20]. But this method forces us to design different algorithms for DP-ERM and DP-SCO, and may lead to a large constant in the final loss. In this section, we show that the exponential mechanism can achieve both the optimal empirical risk for DP-ERM and the optimal population loss for DP-SCO by simply changing the parameters. The bound on the generalization error works beyond differential privacy and can be useful for other (non-private) optimization settings.

The proof makes use of the *Talagrand transportation inequality.* Recall for two probability distributions ν_1, ν_2, the Wasserstein distance is equivalently defined as

$$W_2(\nu_1, \nu_2) = \inf_{\Gamma} \left(\mathbb{E}_{(x_1,x_2)\sim\Gamma} \|x_1 - x_2\|_2^2 \right)^{1/2},$$

where the infimum is over all couplings Γ of ν_1, ν_2.

Theorem 6.5 (Talagrand transportation inequality)**.** *[OV00, Theorem 1] Let* $\mathrm{d}\pi \propto e^{-F(x)}\mathrm{d}x$ *be a* μ*-strongly log-concave probability measure on* $\mathcal{K} \subseteq \mathbb{R}^d$ *with finite moments of order 2. For all probability measure* ν *absolutely continuous with respect to* π *and with finite moments of order 2,*

$$\mathrm{W}_2(\nu, \pi) \leq \sqrt{\frac{2}{\mu}\mathrm{D}_{KL}(\nu, \pi)}.$$

To prove our main result on bounding the generalization error of sampling mechanism, we need the following lemma.

Lemma 6.6 (Lemma 7 in [BE02])**.** For any learning algorithm $\mathcal{A}$ and dataset $\mathcal{D} = \{s_1, \cdots, s_n\}$ drawn i.i.d from the underlying distribution $\mathcal{P}$, let $\mathcal{D}'$ be a neighboring dataset formed by replacing a random element of $\mathcal{D}$ with a freshly sampled $s' \sim \mathcal{P}$. If $\mathcal{A}(\mathcal{D})$ is the output of $\mathcal{A}$ with $\mathcal{D}$, then

$$\mathbb{E}_{\mathcal{D}}[\hat{F}(\mathcal{A}(\mathcal{D})) - F(\mathcal{A}(\mathcal{D}); \mathcal{D})] = \mathbb{E}_{\mathcal{D}, s'\sim\mathcal{P}, \mathcal{A}} \Big[f(\mathcal{A}(\mathcal{D}); s') - f(\mathcal{A}(\mathcal{D}'); s') \Big].$$

Now state and prove our main result on the generalization error.

Theorem 6.7. *Suppose* $\{f(\cdot, s)\}$ *is a family* μ*-strongly convex functions over* $\mathcal{K}$ *such that* $f(x; s) - f(x; s')$ *is* G*-Lipschitz for all* s, s'*. For any* $k > 0$ *and dataset* $\mathcal{D} = \{s_1, s_2, \cdots, s_n\}$ *drawn i.i.d from the underlying distribution* $\mathcal{P}$*, let* $\mathcal{D}'$ *be a neighboring dataset formed by replacing a random element of* $\mathcal{D}$ *with a freshly sampled* $s' \sim \mathcal{P}$*,*

$$\mathrm{W}_2(\pi_{\mathcal{D}}, \pi_{\mathcal{D}'}) \leq \frac{G}{n\mu}.$$

If we sample our solution from density $\pi_{\mathcal{D}}(x) \propto e^{-kF(x;\mathcal{D})}$*, we can bound the excess population loss as:*

$$\mathbb{E}_{\mathcal{D}, x\sim\pi_{\mathcal{D}}}[\hat{F}(x)] - \min_{x\in\mathcal{K}} \hat{F}(x) \leq \frac{G^2}{\mu n} + \frac{d}{k}.$$

Proof. Recall that

$$F(x; \mathcal{D}) = \frac{1}{n} \sum_{s_i \in \mathcal{D}} f(x; s_i).$$

We form a neighboring data set $\mathcal{D}'$ by replacing a random element of $\mathcal{D}$ by a freshly sampled $s' \sim \mathcal{P}$. Let $\pi_{\mathcal{D}} \propto e^{-kF(x;\mathcal{D})}$ and $\pi_{\mathcal{D}'} \propto e^{-kF(x;\mathcal{D}')}$. By Corollary 4.3, we have

$$\mathrm{D}_{KL}(\pi_{\mathcal{D}}, \pi_{\mathcal{D}'}) \leq \frac{G^2 k}{2n^2\mu}.$$

By the assumptions, we know both $F(x;\mathcal{D})$ and $F(x;\mathcal{D}')$ are μ-strongly convex and by Theorem 6.5, we have

$$\mathrm{W}_2(\pi_{\mathcal{D}}, \pi_{\mathcal{D}'}) \leq \sqrt{\frac{2}{k\mu}\mathrm{D}_{KL}(\pi_{\mathcal{D}}, \pi_{\mathcal{D}'})} \leq \frac{G}{n\mu}.$$

By Lemma 6.6 and properties of Wasserstein distance, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{D}, x\sim\pi_{\mathcal{D}}}[\hat{F}(x) - F(x;\mathcal{D})] &= \mathop{\mathbb{E}}_{\mathcal{D}, s'\sim\mathcal{P}}\left[\mathop{\mathbb{E}}_{x\sim\pi_{\mathcal{D}}} f(x;s') - \mathop{\mathbb{E}}_{x'\sim\pi_{\mathcal{D}'}} f(x';s')\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{D}, s'\sim\mathcal{P}}\left[\mathop{\mathbb{E}}_{x\sim\pi_{\mathcal{D}}}\left[f(x;s') - f(x;s'')\right] - \mathop{\mathbb{E}}_{x'\sim\pi_{\mathcal{D}'}}\left[f(x';s') - f(x';s'')\right]\right] \\
&\qquad \text{(where } s'' \text{ is chosen arbitrarily; note that } \mathbb{E}_{\mathcal{D},x\sim\pi_{\mathcal{D}}}[f(x;s'')] = \mathbb{E}_{\mathcal{D}',x'\sim\pi_{\mathcal{D}'}}[f(x';s'')]) \\
&\leq G \cdot \mathrm{W}_2(\pi_{\mathcal{D}}, \pi_{\mathcal{D}'}) \qquad\qquad (f(\cdot;s') - f(\cdot;s'') \text{ is } G\text{-Lipschitz}) \\
&\leq \frac{G^2}{n\mu}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{D}, x\sim\pi_{\mathcal{D}}}[\hat{F}(x)] - \min_{x\in\mathcal{K}} \hat{F}(x) &\leq \mathop{\mathbb{E}}_{\mathcal{D}, x\sim\pi_{\mathcal{D}}}[\hat{F}(x)] - \mathop{\mathbb{E}}_{\mathcal{D}}[\min_{x\in\mathcal{K}} F(x;\mathcal{D})] \\
&\leq \mathop{\mathbb{E}}_{\mathcal{D}, x\sim\pi_{\mathcal{D}}}[\hat{F}(x) - F(x;\mathcal{D})] + \mathop{\mathbb{E}}_{\mathcal{D}, x\sim\pi_{\mathcal{D}}}[F(x;\mathcal{D}) - \min_{x\in\mathcal{K}} F(x;\mathcal{D})] \\
&\leq \frac{G^2}{n\mu} + \mathop{\mathbb{E}}_{\mathcal{D}, x\sim\pi_{\mathcal{D}}}[F(x;\mathcal{D}) - \min_{x\in\mathcal{K}} F(x;\mathcal{D})] \\
&\leq \frac{G^2}{n\mu} + \frac{d}{k},
\end{aligned}$$

where the last inequality follows from Lemma 6.1. □

With the bounds on generalization error, we can get our first result on DP-SCO.

Theorem 6.8 (DP-SCO)**.** *Let $\varepsilon > 0$, $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex set of diameter D and $\{f(\cdot;s)\}_{s\in\mathcal{D}}$ be a family of convex functions over $\mathcal{K}$ such that $f(\cdot;s) - f(\cdot;s')$ is G-Lipschitz for all s, s'. For any data-set $\mathcal{D}$ and $k > 0$, sampling $x^{(priv)}$ with probability proportional to $\exp\left(-k(F(x;\mathcal{D}) + \mu\|x\|_2^2/2)\right)$ is $(\varepsilon, \delta(\varepsilon))$-differentially private, where*

$$\delta(\varepsilon) \leq \delta\left(\mathcal{N}(0,1) \,\middle\|\, \mathcal{N}\left(\frac{G\sqrt{k}}{n\sqrt{\mu}}, 1\right)\right)(\varepsilon).$$

If users in the dataset $\mathcal{D}$ are drawn i.i.d. from the underlying distribution $\mathcal{P}$, the excess population loss is bounded by $G/n\mu + d/k + \mu D^2/2$. Moreover, if $\{f(\cdot;s)\}_{s\in\mathcal{D}}$ are μ-strongly convex, then sampling $x^{(priv)}$ with probability proportional to $\exp(-kF(x;\mathcal{D}))$ is $(\varepsilon, \delta(\varepsilon))$-differentially private where

$$\delta(\varepsilon) \leq \delta\left(\mathcal{N}(0,1) \,\middle\|\, \mathcal{N}\left(\frac{G\sqrt{k}}{n\sqrt{\mu}}, 1\right)\right)(\varepsilon).$$

The excess population loss is bounded by $G/n\mu + d/k$.

Proof. The first part about privacy is a restatement of our result on DP-ERM (Theorem 6.4). The excess population loss (See Equation (1.6)) follows from the bound on generalization error (Theorem 6.7) and utility guarantee (Lemma 6.1). $\square$

We give an implementation result of our DP-SCO result.

Theorem 6.9 (DP-SCO Implementation)**.** *With same assumptions in Theorem 6.8, and assume $f(\cdot;s)$ is G-Lipschitz over $\mathcal{K}$ for all s. For $0 < \delta < 1/10$ and $0 < \varepsilon < 1/10$, there is an efficient algorithm to solve DP-SCO that has the following guarantees:*

- *The algorithm is (ε, δ)-differentially private;*
- *The expected population loss is bounded by*
$$GD\left(\frac{2\sqrt{\log(1/\delta)d}}{\varepsilon n} + \frac{2}{\sqrt{n}}\right),$$
where $c > 0$ is an arbitrary constant to be chosen.
- *The algorithm takes*
$$O\left(\min\left\{\frac{\varepsilon^2 n^2}{\log(1/\delta)}, nd\right\}\log^2\left(\frac{\varepsilon nd}{\delta}\right)\right)$$
queries of the values of $f(\cdot, s_i)$ in expectation and takes the same number of samples from some Gaussian restricted to the convex set $\mathcal{K}$.

Remark 6.10. As for the non-typical case when $\varepsilon \geq 1/10$, one can use the bound in Theorem 6.4 and the bound on generalization error (Theorem 6.7) . Particularly, one can achieve expected population loss
$$O\left(GD\left(\frac{\sqrt{d}/n}{\sqrt{\log(1/\delta)+\varepsilon} - \sqrt{\log(1/\delta)}} + \frac{1}{\sqrt{n}}\right)\right).$$

Proof. By Theorem 6.8, when $k \leq \varepsilon^2 n^2 \mu//2G^2 \log(3/(4\delta))$, sampling from $\exp(-k(F(x;\mathcal{D}) + \mu\|x\|_2^2/2))$ is $(\varepsilon, 2\delta/3)$-DP.

In addition, we can set $k = \frac{\mu}{G^2}\min\{\varepsilon^2 n^2/2\log(3/(4\delta)), 2nd\}$ for anarbitrarily large constant $c > 0$ to make the mechanism $(\varepsilon, 2\delta/3)$-differentially private, achieving tight population loss and decreasing the running time. Then the population loss is bounded above by
$$\frac{d}{k} + \frac{\mu D^2}{2} + \frac{G^2}{\mu n} = \frac{G^2}{\mu}\max\left\{\frac{2\log(3/(4\delta))d}{\varepsilon^2 n^2}, \frac{1}{2n}\right\} + \frac{\mu D^2}{2} + \frac{G^2}{\mu n}.$$

By setting
$$\mu = \frac{G}{D}\sqrt{2(\frac{2\log(3/(4\delta))d}{\varepsilon^2 n^2} + \frac{1}{2n})},$$
the population loss is bounded above by
$$GD\sqrt{\frac{4\log(3/(4\delta))d}{\varepsilon^2 n^2} + \frac{1}{n}} + GD\sqrt{\frac{1}{n}} \leq GD\left(\frac{2\sqrt{\log(3/(4\delta))d}}{\varepsilon n} + \frac{2}{\sqrt{n}}\right).$$

To make the process algorithmic, we also apply Theorem 5.3 with the accuracy on the total variation distance to be $\min\{\delta/3, 1/cn^c\}$ for some large enough constant c. This leads to an extra empirical loss, and hence we use $\log(1/\delta)$ rather than $\log(3/(4\delta))$ in the final loss term. The runtime follows from Theorem 5.3. $\square$

7. Information-theoretic Lower Bound for DP-SCO

In this section, we prove an information-theoretic lower bound for the query complexity required for DP-SCO (with value queries), which matches (up to logarithmic terms) the query complexity achieved by our algorithm (in Theorem 6.9). Our proof is similar to the previous works such as [ACCD12, DJWW15], but with some modifications.

Before stating the lower bound, we define additional notations. Recall that we are given a set $\mathcal{D}$ of n samples (users) $\{s_1, \cdots, s_n\}$. Let $\mathbb{A}_k$ be the collection of all algorithms that observe a sequence of k data points $(Y^1, \cdots, Y^k)$ with $Y^t = f(X^t; S^t)$, where $S^t \in \mathcal{D}$ and $X^t \in \mathcal{K}$ are chosen arbitrarily and adaptively by the algorithm (and possibly introducing randomness).

For the lower bound, we only consider linear functions, that is, we define $f(x; s) \stackrel{\text{def}}{=} \langle x, s\rangle$. Let $\mathcal{P}_G$ be the collection of all distributions such that if $\mathcal{P} \in \mathcal{P}_G$, then $\mathbb{E}_{s\sim\mathcal{P}} \|s\|_2^2 \le G^2$, and define the optimality gap

$$\varepsilon_k(\mathcal{A}, \mathcal{P}, \mathcal{K}) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{\mathcal{D}\sim\mathcal{P}^n,\mathcal{A}}[\hat{F}(\hat{x}(\mathcal{D}))] - \inf_{x\in\mathcal{K}} \hat{F}(x),$$

where $\hat{F}(x) = \mathbb{E}_{s\sim\mathcal{P}} f(x; s)$, $\hat{x}$ is the output the algorithm $\mathcal{A}$ given the input dataset $\mathcal{D}$ and the expectation is over the dataset $\mathcal{D} \sim \mathcal{P}^n$ and the randomness of the algorithm $\mathcal{A}$. Note that we can rewrite the optimality gap as:

$$\begin{aligned}
\varepsilon_k(\mathcal{A}, \mathcal{P}, \mathcal{K}) &= \mathop{\mathbb{E}}_{\mathcal{D}\sim\mathcal{P}^n,\mathcal{A}}[\hat{F}(\hat{x}(\mathcal{D}))] - \inf_{x\in\mathcal{K}} \hat{F}(x) \\
&= \mathop{\mathbb{E}}_{s\sim\mathcal{P}}\left[\mathop{\mathbb{E}}_{\mathcal{D}\sim\mathcal{P}^n,\mathcal{A}} f(\hat{x}(\mathcal{D}); s)\right] - \inf_{x\in\mathcal{K}} \mathop{\mathbb{E}}_{s\sim\mathcal{P}}[f(x; s)] \\
&= \mathop{\mathbb{E}}_{s\sim\mathcal{P},\mathcal{D}\sim\mathcal{P}^n,\mathcal{A}}[\hat{x}(\mathcal{D})^\top s] - \inf_{x\in\mathcal{K}} \mathop{\mathbb{E}}_{s\sim\mathcal{P}}[x^\top s].
\end{aligned}$$

The minimax error is defined by

$$\varepsilon_k^*(\mathcal{P}_G, \mathcal{K}) \stackrel{\text{def}}{=} \inf_{\mathcal{A}\in\mathbb{A}_k} \sup_{\mathcal{P}\in\mathcal{P}_G} \varepsilon_k(\mathcal{A}, \mathcal{P}, \mathcal{K}).$$

Theorem 7.1. *Let $\mathcal{K}$ be the ℓ_2 ball of diameter D in $\mathbb{R}^d$, then*

$$\varepsilon_k^*(\mathcal{P}_G, \mathcal{K}) \ge \frac{GD}{16} \min\left\{1, \sqrt{\frac{d}{4k}}\right\}.$$

In particular, for any (randomized) algorithm $\mathcal{A}$ that can observe a sequence of data points $(Y^1, \cdots, Y^k)$ with $Y^t = f(X^t; S^t)$ where $S^t \in \mathcal{D} = \{s_1, s_2, \ldots, s_n\}$ and $X^t \in \mathcal{K}$ are chosen arbitrarily and adaptively by $\mathcal{A}$, there exists a distribution $\mathcal{P}$ over convex functions such that $\mathbb{E}_{s\sim\mathcal{P}}[\|\nabla f(x, s)\|_2^2] \le G^2$ for all $x \in \mathcal{K}$, such that the output $\hat{x}$ of the algorithm satisfies

$$\mathop{\mathbb{E}}_{s\sim\mathcal{P}}\left[\mathop{\mathbb{E}}_{\mathcal{D}\sim\mathcal{P}^n,\mathcal{A}} f(\hat{x}; s)\right] - \min_{x\in\mathcal{K}} \mathop{\mathbb{E}}_{s\sim\mathcal{P}}[f(x; s)] \ge \frac{GD}{16} \min\left\{1, \sqrt{\frac{d}{4k}}\right\}.$$

7.1. **Proof of Theorem 7.1.** We reduce the optimization problem into a series of binary hypothesis tests. Recall we are considering linear functions $f(x;s) \stackrel{\text{def}}{=} \langle x, s\rangle$. Let $\mathcal{V} = \{-1,1\}^d$ be a Boolean hypercube and for each $v \in \mathcal{V}$, let $\mathcal{N}_v = \mathcal{N}(\delta v, \sigma^2 I_d)$ be a Gaussian distribution for parameters to be chosen such that $\hat{F}_v(x) \stackrel{\text{def}}{=} \mathbb{E}_{s\sim\mathcal{N}_v}[f(x;s)] = \delta\langle x, v\rangle$. Note that

$$\mathop{\mathbb{E}}_{s\sim\mathcal{N}_v}[\|\nabla f(x,s)\|_2^2] = \mathop{\mathbb{E}}_{s\sim\mathcal{N}_v}[\|s\|_2^2] = (\delta^2+\sigma^2)d.$$

Therefore $G = \sqrt{d(\delta^2+\sigma^2)}$.

Clearly the lower bound should scale linearly with D. Therefore without loss of generality, we can assume that the diameter $D = 2$ and define $\mathcal{K} = \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$ to be the unit ball. As in [ACCD12], we suppose that v is uniformly sampled from $\mathcal{V} = \{-1,1\}^d$. Note that if we can find a good solution to $\hat{F}_v(x)$, we need to determine the signs of vector v well. In particular, we have the following claim.

Claim 7.2 ([DJWW15])**.** For each $v \in \mathcal{V}$, let x^v minimize $\hat{F}_v$ over $\mathcal{K}$ and note that that $x^v = -v/\sqrt{d}$. For any solution $\hat{x} \in \mathbb{R}^d$, we have

$$\hat{F}_v(\hat{x}) - \hat{F}_v(x^v) \ge \frac{\delta}{2\sqrt{d}} \sum_{j=1}^{d} \mathbb{1}\{\text{sign}(\hat{x}_j) \ne \text{sign}(x_j^v)\},$$

where the function sign$(\cdot)$ is defined as:

$$\text{sign}(\hat{x}_j) = \begin{cases} + & \text{if } \hat{x}_j > 0 \\ 0 & \text{if } \hat{x}_j = 0 \\ - & \text{otherwise} \end{cases}$$

Claim 7.2 provides a method to lower bound the minimax error. Specifically, we define the hamming distance between any two vectors $x, y \in \mathbb{R}^d$ as $d_H(x,y) = \sum_{j=1} \mathbb{1}\{\text{sign}(x_j) \ne \text{sign}(y_j)\}$, and we have

$$\varepsilon_k^*(\mathcal{P}_G, \mathcal{K}) \ge \frac{\delta}{2\sqrt{d}}\{\inf_{\hat{v}} \mathbb{E}[d_H(\hat{v}, v)]\}, \tag{7.1}$$

where $\hat{v}$ denotes the output of any algorithm mapping from the observation $(Y^1, \cdots, Y^k)$ to $\{-1,1\}^d$, and the probability is taken over the distribution of the underlying v, the observation $(Y^1, \cdots, Y^k)$ and any additional randomness in the algorithm.

By Equation (7.1), it suffices to lower bound the value of the testing error $\mathbb{E}[d_H(\hat{v}, v)]$. As discussed in [ACCD12, DJWW15], the randomness in the algorithm cannot help, and we can assume the algorithm is deterministic, i.e., (X^t, S^t) is a deterministic function of $Y^{[t-1]}$.[7] The argument is basically based on the easy direction of Yao's principle.

Now we continue our proof of the lower bound. We will make use of the property of the Bayes risk.

Lemma 7.3 ([ACCD12, Lemma 1])**.** Consider the problem of testing hypothesis $H_{-1} : v \sim \mathbb{P}_{-1}$ and $H_1 : v \sim \mathbb{P}_1$, where H_{-1} and H_1 occur with prior probability π_{-1} and $\pi_1 \stackrel{\text{def}}{=} 1 - \pi_{-1}$ respectively. For any algorithm that takes one sample v and outputs $\hat{i} : v \to \{-1,1\}$, we define the Bayes risk B be the minimum average probability that algorithm fails (v is not

[7] We use $Y^{[t]}$ to denote the first t observations, i.e., $(Y^1, \cdots, Y^t)$

sampled from $H_{\hat{i}(v)}$). That is, $B = \inf_{\hat{i}} \pi_{-1} \Pr[\hat{i}(v) = 1 \mid v \sim \mathbb{P}_{-1}] + \pi_1 \Pr[\hat{i}(v) = 0 \mid v \sim \mathbb{P}_1]$. Then, we have

$$B \geq \min(\pi_{-1}, \pi_1)(1 - \|\mathbb{P}_1 - \mathbb{P}_{-1}\|_{\mathrm{TV}}).$$

Lemma 7.4. Suppose that v is uniformly sampled from $\mathcal{V} = \{-1, 1\}^d$. Then any estimator $\hat{v}$ obeys

$$\mathbb{E}[d_H(\hat{v}, v)] \geq \frac{d}{2}\left(1 - \frac{\delta\sqrt{k}}{\sigma\sqrt{d}}\right).$$

Proof. Let $\pi_{-1} = \pi_1 = 1/2$. For each j, define $\mathbb{P}_{-1,j} = \mathbb{P}(Y^{[k]} \mid v_j = -1)$ and $\mathbb{P}_{1,j} = \mathbb{P}(Y^{[k]} \mid v_j = 1)$ to be distributions over the observations $(Y^1, \cdots, Y^k)$ conditional on $v_j \neq 1$ and $v_j = 1$ respectively. Let B_j be the Bayes risk of the decision problem for j-th coordinate of v between $H_{-1,j} : v_j = -1$ and $H_{1,j} : v_j = 1$. We have that

$$\begin{aligned}
\mathbb{E}[d_H(\hat{v}, v)] \geq & \sum_{j=1}^{d} B_j \\
\geq & \pi_1 \sum_{j=1}^{d} (1 - \|\mathbb{P}_{1,j} - \mathbb{P}_{-1,j}\|_{\mathrm{TV}}) \\
\geq & \frac{d}{2}\left(1 - \frac{1}{\sqrt{d}}\sqrt{\sum_{j=1}^{d} \|\mathbb{P}_{1,j} - \mathbb{P}_{-1,j}\|_{\mathrm{TV}}^2}\right),
\end{aligned}$$

where the first inequality follows from the definition of Bayes risk, the second inequality follows by Lemma 7.3, and the last inequality follows by the Cauchy-Schwartz inequality.

To complete the proof, it suffices to show that

$$\sum_{j=1}^{d} \|\mathbb{P}_{1,j} - \mathbb{P}_{-1,j}\|_{\mathrm{TV}}^2 \leq \frac{\delta^2}{\sigma^2} k, \tag{7.2}$$

which will be established later. Assuming Equation (7.2) to hold,

$$\mathbb{E}[d_H(\hat{v}, v)] \geq \frac{d}{2}(1 - \frac{\delta\sqrt{k}}{\sigma\sqrt{d}}).$$

□

We will complete the proof of Lemma 7.4 by showing the following boundon total variation distance.

Claim 7.5.

$$\sum_{j=1}^{d} \|\mathbb{P}_{1,j} - \mathbb{P}_{-1,j}\|_{\mathrm{TV}}^2 \leq \frac{\delta^2}{\sigma^2} k.$$

Proof. Applying Pinsker's inequality, we know $\|\mathbb{P}_{1,j} - \mathbb{P}_{-1,j}\|_{\mathrm{TV}}^2 \leq \frac{1}{2} \mathrm{D}_{KL}(\mathbb{P}_{-1,j} \| \mathbb{P}_{1,j})$. To bound the KL divergence between $\mathbb{P}_{-1,j}$ and $\mathbb{P}_{1,j}$ over all possible $Y^{[k]}$, consider $v' =$

$(v_1, \cdots, v_{j-1}, v_{j+1}, \cdots, v_d)$, and define $\mathbb{P}_{-1,j,v'}(Y^{[k]}) \stackrel{\text{def}}{=} \mathbb{P}(Y^{[k]} \mid v_j = -1, v')$ to be the distribution conditional on $v_j = -1$ and v'. We have

$$\mathbb{P}_{-1,j}(Y^{[k]}) = \sum_{v'} \Pr[v']\mathbb{P}_{-1,j,v'}(Y^{[k]}).$$

Convexity of the KL divergence suggests that

$$\mathrm{D}_{KL}(\mathbb{P}_{-1,j}\|\mathbb{P}_{1,j}) \le \sum_{v'} \Pr[v']\mathrm{D}_{KL}(\mathbb{P}_{-1,j,v'}\|\mathbb{P}_{1,j,v'}).$$

Fixing any possible v', we want to bound the KL divergence $\mathrm{D}_{KL}(\mathbb{P}_{-1,j,v'}\|\mathbb{P}_{1,j,v'})$.

Recall that we are considering deterministic algorithms and (X^t, S^t) is a deterministic function of $Y^{[t-1]}$. Let $Q_i \in \mathbb{R}^{d\times k}$ be a (random) matrix, which records the set of points the algorithm queries for the user s_i. Specifically, for t-th step, if the algorithm queries (X^t, S^t), then $Q_i^t = X^t$ if $S^t = s_i$, otherwise $Q_i^t = 0$, where Q_i^t is the t-th column of Q_i.

Because we are considering linear functions, without loss of generality we can assume $\langle Q_i^j, Q_i^{j'}\rangle = 0$ for each i and any $j \neq j'$, and $\|Q_i^t\|_2 \in \{0, 1\}$ for any i and t. We name this assumption *Orthogonal Query*. Roughly speaking, for any algorithm, we can modify it to satisfy the Orthogonal Query. Whenever the algorithm wants to query some point, we can use Gram–Schmidt process to query another point and satisfy Orthogonal Query, and recover the function value at the original point queried by the algorithm.

By the chain rule of KL-divergence, if we define $P_{-1,j,v'}(Y^t \mid Y^{[t-1]})$ to be the distribution of tth observation Y^t conditional on v', $v_j = -1$ and $Y^{[t-1]}$, then

$$\begin{aligned}
&\mathrm{D}_{KL}(\mathbb{P}_{-1,j,v'}\|\mathbb{P}_{1,j,v'}) \\
= &\sum_{t=1}^{k} \int_{\mathcal{Y}^{t-1}} \mathrm{D}_{KL}(P_{-1,j,v'}(Y^t \mid Y^{[t-1]} = y)\|P_{1,j,v'}(Y^t \mid Y^{[t-1]} = y)\mathrm{d}P_{-1,j,v'}(y).
\end{aligned}$$

Fix $Y^{[t-1]}$ such that $Y^{[t-1]} = y$. Since the algorithm is deterministic and (X^t, S^t) is fixed given $Y^{[t-1]}$, let $S^t = s_i$ so $X^t = Q_i^t$.

Note that the n users in $\mathcal{D}$ are i.i.d. sampled. Then $\mathrm{D}_{KL}(P_{-1,j,v'}(Y^t \mid Y^{[t-1]} = y)\|P_{1,j,v'}(Y^t \mid Y^{[t-1]} = y)$ only depends on the randomness of s_i and the first t columns of Q_i, which is denoted by $Q_i^{[t]}$. We use Y_j^t to denote the observation corresponding to user s_j for the tth query (if $S^t \neq s_j$, we have $Y_j^t = 0$). Note that the observation $Y_i^{[t]} = Q_i^{[t]\top} s_i$ where $s_i \sim \mathcal{N}(\delta v, \sigma^2 I_d)$. Then we know $Y_i^{[t]}$ is normally distributed with mean $\delta Q_i^{[t]\top} v$ and covariance $\sigma^2 Q_i^{[t]\top} Q_i^{[t]}$.

Recall that the KL divergence between normal distributions is $\mathrm{D}_{KL}(\mathcal{N}(\mu_1, \Sigma)\|\mathcal{N}(\mu_2, \Sigma)) = (\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2)/2$. By the Orthogonal Query assumption, $Q_i^{[t]\top} Q_i^{[t]} \in \{0, 1\}^{t\times t}$ is a diagonal matrix. By the nature of conditional distributions of Gaussian distributions, we know that Y_i^t only depends on the Q_i^t and it is independent of $Q_i^{[t-1]}$. Hence

$$\begin{aligned}
&\mathrm{D}_{KL}(P_{-1,j,v'}(Y^t \mid Y^{[t-1]} = y)\|P_{1,j,v'}(Y^t \mid Y^{[t-1]} = y)) \\
= &\ \mathrm{D}_{KL}(P_{-1,j,v'}(Y_i^t \mid Y^{[t-1]} = y)\|P_{1,j,v'}(Y_i^t \mid Y^{[t-1]} = y)) \\
= &\ \frac{1}{2}(2\delta Q_i^t(j))^2/\sigma^2,
\end{aligned}$$

where $Q_i^t(j)$ is the j-th coordinate of Q_i^t. Summing over the terms,

$$\begin{aligned}\sum_{j=1}^{d}\|\mathbb{P}_{1,j}-\mathbb{P}_{-1,j}\|_{\mathrm{TV}}^2 \leq & \frac{1}{2}\mathrm{D}_{KL}(\mathbb{P}_{-1,j}\|\mathbb{P}_{1,j}) \\ \leq & \frac{1}{2}\sum_{t=1}^{k}\sum_{j=1}^{d}\sum_{i=1}^{n}\mathbb{E}[\frac{1}{2}(2\delta Q_i^t(j))^2/\sigma^2] \\ \leq & \frac{\delta^2}{\sigma^2}k,\end{aligned}$$

where the last line follows from the fact that for each t,$\sum_{i=1}^n \|Q_i^t\|_2^2 = \sum_{i=1}^n \sum_{j=1}^d (Q_i^t(j))^2 = 1$ since we only query one user for t-th step. □

Having Lemma 7.4, we can complete the proof of Theorem 7.1.

Proof. of Theorem 7.1. As discussed before,

$$\hat{F}_v(\hat{x}) - \hat{F}_v(x^v) \geq \frac{\delta}{2\sqrt{d}}\sum_{j=1}^{d}\mathbb{1}\{\mathrm{sign}(\hat{x}_j)\neq\mathrm{sign}(x_j^v)\},$$

and hence

$$\begin{aligned}\varepsilon_k^*(\mathcal{P}_G,\mathcal{K}) \geq & \frac{\delta}{2\sqrt{d}}\inf_{\hat{v}}\mathbb{E}[d_H(\hat{v},v)] \\ \geq & \frac{\delta\sqrt{d}}{4}\left(1-\frac{\delta\sqrt{k}}{\sigma\sqrt{d}}\right),\end{aligned}$$

where the last line follows from Lemma 7.4. We now set $\delta = \sigma\sqrt{d}/2\sqrt{k}$ and $\sigma = G/\sqrt{d+d^2/4k}$, so that $d(\sigma^2+\delta^2) = G^2$. Therefore,

$$\varepsilon_k^*(\mathcal{P}_G,\mathcal{K}) \geq \frac{\delta\sqrt{d}}{8} = \frac{D\delta\sqrt{d}}{16} = \frac{GD}{16\sqrt{1+\frac{4k}{d}}} \geq \frac{GD}{16}\min\left\{1,\sqrt{\frac{d}{4k}}\right\}.$$

□

Corollary 7.6 (Lower bound for DP-SCO)**.** For any (non-private) algorithm which makes less than $O\left(\min\{\varepsilon^2 n^2/\log(1/\delta), nd\}\right)$ function value queries, there exist a convex domain $\mathcal{K}\subset\mathbb{R}^d$ of diameter D, a distribution $\mathcal{P}$ supported on G-Lipschitz linear functions $f(\cdot;s)\overset{\mathrm{def}}{=}\langle x,s\rangle$, such that the output $\hat{x}$ of the algorithm satisfies

$$\mathop{\mathbb{E}}_{s\sim\mathcal{P}}[\langle\hat{x},s\rangle] - \min_{x\in\mathcal{K}}\mathop{\mathbb{E}}_{s\sim\mathcal{P}}[\langle x,s\rangle] \geq \Omega\left(\frac{GD}{\sqrt{1+\log(n)/d}}\cdot\min\left\{\frac{\sqrt{\log(1/\delta)d}}{\varepsilon n}+\frac{1}{\sqrt{n}},1\right\}\right).$$

Proof. Note that Theorem 7.1 almost gives us what we want, except that the Lipschitz constant of the functions in the hard distribution is bounded only on average by G. To get distributions over G-Lipschitz functions, we just condition on the bad event not happening.

Recall that we are considering the set of distributions $\mathcal{N}_v = \mathcal{N}(\delta v, \sigma^2 I_d)$ for which $\mathbb{E}_{s\sim\mathcal{N}_v}\|s\|_2^2 \leq G^2 = d(\delta^2+\sigma^2)$. And we proved that $\inf_{\mathcal{A}\in\mathbb{A}_k}\sup_{v\in\mathcal{V}}\mathbb{E}_{s\sim\mathcal{N}_v,\mathcal{A}}[\hat{F}_v(\hat{x}_k)-\hat{F}_v^*] \geq (GD/16)\min\left\{1,\sqrt{d/4k}\right\}$ in Theorem 7.1, where $\hat{x}_k$ is the output of $\mathcal{A}$ with k observations

$Y^{[k]}$. To prove Corollary 7.6, we need to modify the distribution of s to satisfy the Lipschitz continuity.

In particular, for some constant c,

$$\begin{aligned}
&\mathbb{E}[\hat{F}_v(\hat{x}_k) - \hat{F}_v^*] \\
&= \mathbb{E}\left[\hat{F}_v(\hat{x}_k) - \hat{F}_v^* \mid \max_{s_i \in \mathcal{D}} \|s_i\|_2 \le cG\sqrt{1+\log(nd)/d}\right] \Pr\left[\max_{s_i \in \mathcal{D}} \|s_i\|_2 \le cG\sqrt{1+\log(nd)/d}\right] \\
&\quad + \mathbb{E}\left[\hat{F}_v(\hat{x}_k) - \hat{F}_v^* \mid \max_{s_i \in \mathcal{D}} \|s_i\|_2 > cG\sqrt{1+\log(nd)/d}\right] \Pr\left[\max_{s_i \in \mathcal{D}} \|s_i\|_2 > cG\sqrt{1+\log(nd)/d}\right].
\end{aligned}$$

By the concentration of spherical Gaussians, if $s \sim \mathcal{N}(\delta v, \sigma^2 I_d)$, then

$$\Pr\left[\|s - \delta v\|_2^2 \le \sigma^2 d(1 + 2\sqrt{\ln(1/\eta)/d} + 2\ln(1/\eta)/d)\right] \ge 1 - \eta.$$

We can choose the constant c large enough, such that $\Pr[\max_{s_i \in \mathcal{D}} \|s_i\|_2 \le cG\sqrt{1+\log(nd)/d}] \ge 1 - 1/\operatorname{poly}(nd)$, which implies that

$$\inf_{\mathcal{A} \in \mathbb{A}_k} \sup_{v \in \mathcal{V}} \mathbb{E}_{\mathcal{D} \sim \mathcal{N}_v^n, \mathcal{A}} \left[\hat{F}_v(\hat{x}_k) - \hat{F}_v^* \mid \max_{s_i \in \mathcal{D}} \|s_i\|_2 \le cG\sqrt{1+\log(nd)/d}\right] \ge \Omega(GD\frac{\min\{\sqrt{d}, \sqrt{k}\}}{\sqrt{k}}).$$

If we use the distributions conditioned on $\max_{s_i \in \mathcal{D}} \|s_i\|_2 \le cG\sqrt{1+\log(nd)/d}$ rather than the Gaussians, and scale the constant to satisfy the assumption on Lipschitz continuity, we can prove the statement. Particularly, let $G' = cG(\sqrt{1+\log(nd)/d})$. If the algorithm can only make $k = O\left(\min\{[\varepsilon^2 n^2/\log(1/\delta)], nd\}\right)$ observations, then

$$\begin{aligned}
&\inf_{\mathcal{A} \in \mathbb{A}_k} \sup_{v \in \mathcal{V}} \mathbb{E}_{\mathcal{D} \sim \mathcal{N}_v^n, \mathcal{A}} \left[\hat{F}_v(\hat{x}_k) - \hat{F}_v^* \mid \max_{s_i \in \mathcal{D}} \|s_i\|_2 \le G'\right] \\
&\ge \Omega\left(GD \cdot \min\left\{(\frac{\sqrt{\log(1/\delta)d}}{\varepsilon n} + \frac{1}{\sqrt{n}}), 1\right\}\right) \\
&= \Omega\left(\frac{G'D}{\sqrt{1+\log(nd)/d}} \cdot \min\left\{\frac{\sqrt{\log(1/\delta)d}}{\varepsilon n} + \frac{1}{\sqrt{n}}, 1\right\}\right),
\end{aligned}$$

which proves the lower bound claimed in the Corollary statement. □

Corollary 7.7 (Lower bound for sampling scheme)**.** Given any $G > 0$ and $\mu > 0$. For any algorithm which takes function values queries less than $O\left((G^2/\mu)(1 + \log(G^2/\mu)/d)\right)$ times, there is a family of G-Lipschitz linear functions $\{f_i(x)\}_{i \in I}$ defined on some ℓ_2 ball $\mathcal{K} \subset \mathbb{R}^d$, such that the total variation distance between the distribution of the output of the algorithm and the distribution proportional to $\exp(-\mathbb{E}_{i \in I} f_i(x) - \mu\|x\|^2/2)$ is at least $\min(1/2, \sqrt{d\mu/G^2})$.

Proof. As shown in [DJWW15], and similar to Corollary 7.6, for any algorithm which can only make k observations, there are a family of G-Lipschitz linear functions restricted on an ℓ_2 ball $\mathcal{K}$ of diameter D centered at $\mathbf{0}$ such that

$$\mathbb{E}\left[\hat{F}_v(\hat{x}_k) - \hat{F}_v^*\right] \ge \Omega\left(\frac{GD}{\sqrt{1+\log(k)/d}} \cdot \min\left\{\sqrt{\frac{d}{k}}, 1\right\}\right), \tag{7.3}$$

where $\hat{F}_v^* = \min_{x \in \mathcal{K}} \hat{F}_v(x)$ and $\hat{x}_k \in \mathcal{K}$ is the output of $\mathcal{A}$.

Suppose we have a sampling algorithm that takes k queries. We use it to sample from $x^{(sol)}$ proportional to $p(x) := \exp(-\hat{F}_v(x) - mu\|x\|^2/2)$ on $\mathcal{K}$ with total variation distance $\eta \leq \min(1/2, \sqrt{d\mu/G^2})$.

Lemma 6.1 shows that

$$\mathbb{E}[\hat{F}_v(x^{(sol)}) + \frac{\mu}{2}\|x^{(sol)}\|^2] \leq \min_{x\in\mathcal{K}}\left(\hat{F}_v(x) + \frac{\mu}{2}\|x\|^2\right) + O(d) + O(\eta)\cdot(GD+\mu D^2),$$

where the last term involving η is due to the total variation distance between $x^{(sol)}$ and p. Setting $D = \sqrt{d/\mu}$ and using the diameter of $\mathcal{K}$ is D and $\eta \leq \min(1/2, \sqrt{d\mu/G^2})$, we have

$$\begin{aligned}\mathbb{E}[\hat{F}_v(x^{(sol)})] &\leq \min_{x\in\mathcal{K}} \hat{F}_v(x) + \frac{\mu}{2}D^2 + O(d + \eta\cdot(GD+\mu D^2)) \\ &\leq \min_{x\in\mathcal{K}} \hat{F}_v(x) + O(d).\end{aligned}$$

Note that we set $D = \sqrt{d/\mu}$. Comparing with (7.3), we have

$$\frac{G\sqrt{d/\mu}}{\sqrt{1+\log(k)/d}}\min\left\{\sqrt{\frac{d}{k}}, 1\right\} \leq O(d).$$

If $d \leq G^2/\mu \leq \exp(d)$, we have

$$G\sqrt{d/\mu}\sqrt{\frac{d}{k}} \leq O(d)$$

and hence $k = \Omega(G^2/\mu)$. If $G^2/\mu \geq \exp(d)$, we have

$$\frac{G\sqrt{d/\mu}}{\sqrt{\log(k)/d}}\sqrt{\frac{d}{k}} \leq O(d)$$

and hence $k = \Omega((G^2 d/\mu)\log(G^2/\mu))$. If $G^2/\mu \leq d$, we can construct our function only on the first $O(G^2/\mu)$ dimensions to get a lower bound $k = \Omega(G^2/\mu)$. Combining all cases gives the result. □

References

[Abo16] John M. Abowd. The challenge of scientific reproducibility and privacy protection for statistical agencies. *Technical report, Census Scientific Advisory Committee*, 2016.

[ACCD12] Ery Arias-Castro, Emmanuel J Candes, and Mark A Davenport. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481, 2012.

[AFKT21] Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in l1 geometry. In *International Conference on Machine Learning*, pages 393–403. PMLR, 2021.

[AKRS19] Jordan Awan, Ana Kenney, Matthew Reimherr, and Aleksandra Slavković. Benefits and pitfalls of the exponential mechanism with applications to Hilbert spaces and functional PCA. In *International Conference on Machine Learning*, pages 374–384. PMLR, 2019.

[App17] Apple Differential Privacy Team. Learning with privacy at scale, 2017. Technical report, Apple.

[AT22] Jason M Altschuler and Kunal Talwar. Privacy of noisy stochastic gradient descent: More iterations without more privacy loss. *arXiv preprint arXiv:2205.13710*, 2022.

[BE02] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[BEL18] Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected Langevin Monte Carlo. *Discrete & Computational Geometry*, 59(4):757–783, 2018.

[BEM+17] Andrea Bittau, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnes, and Bernhard Seefeld. Prochlo: Strong privacy for analytics in the crowd. In *Proceedings of the 26th Symposium on Operating Systems Principles*, pages 441–459, 2017.

[BFGT20] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33, 2020.

[BFTT19] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, pages 11282–11291, 2019.

[BGN21] Raef Bassily, Cristóbal Guzmán, and Anupama Nandi. Non-Euclidean differentially private stochastic convex optimization. In *Conference on Learning Theory*, pages 474–499. PMLR, 2021.

[BNS13] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 363–378. Springer, 2013.

[BST14] 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.

[BV19] Victor Balcer and Salil Vadhan. Differential privacy on finite computers. *Journal of Privacy and Confidentiality*, 9:2, 2019.

[BW18] Borja Balle and Yu-Xiang Wang. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pages 403–412, 2018.

[CDJB20] Niladri Chatterji, Jelena Diakonikolas, Michael I Jordan, and Peter Bartlett. Langevin Monte Carlo without smoothness. In *International Conference on Artificial Intelligence and Statistics*, pages 1716–1726. PMLR, 2020.

[CDWY20] Yuansi Chen, Raaz Dwivedi, Martin J Wainwright, and Bin Yu. Fast mixing of metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *Journal of Machine Learning Research*, 21:92–1, 2020.

[Che21] Yuansi Chen. An almost constant lower bound of the isoperimetric coefficient in the KLS conjecture. *Geometric and Functional Analysis*, 31(1):34–61, 2021.

[CKS20] Clément L Canonne, Gautam Kamath, and Thomas Steinke. The discrete Gaussian for differential privacy. *Advances in Neural Information Processing Systems*, 33:15676–15688, 2020.

[CM08] Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. In *NIPS*, volume 8, pages 289–296. Citeseer, 2008.

[CMS11] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

[CSS13] Kamalika Chaudhuri, Anand D Sarwate, and Kaushik Sinha. A near-optimal algorithm for differentially-private principal components. *Journal of Machine Learning Research*, 14, 2013.

[CV19] Zongchen Chen and Santosh S Vempala. Optimal convergence rate of Hamiltonian Monte Carlo for strongly logconcave distributions. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[Dal17] 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.

[DJWW15] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[DKL18] Etienne De Klerk and Monique Laurent. Comparison of Lasserre's measure-based bounds for polynomial optimization to bounds obtained by simulated annealing. *Mathematics of Operations Research*, 43(4):1317–1325, 2018.

[DKM+06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer, 2006.

[DKY17] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30, 2017.

[DMM19] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(1):2666–2711, 2019.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

[DRS19] Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2019.

[EPK14] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on Computer and Communications Security*, pages 1054–1067, 2014.

[Fel16] Vitaly Feldman. Generalization of ERM in stochastic convex optimization: The dimension strikes back. *Advances in Neural Information Processing Systems*, 29:3576–3584, 2016.

[FKT20] Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

[FTS17] Kazuto Fukuchi, Quang Khai Tran, and Jun Sakuma. Differentially private empirical risk minimization with input perturbation. In *International Conference on Discovery Science*, pages 82–90. Springer, 2017.

[FYC23] Jiaojiao Fan, Bo Yuan, and Yongxin Chen. Improved dimension dependence of a proximal algorithm for sampling. *arXiv preprint arXiv:2302.10081*, 2023.

[GLL+23a] Sivakanth Gopi, Yin Tat Lee, Daogao Liu, Ruoqi Shen, and Kevin Tian. Algorithmic aspects of the log-Laplace transform and a non-Euclidean proximal sampler. *arXiv preprint arXiv:2302.06085*, 2023.

[GLL+23b] Sivakanth Gopi, Yin Tat Lee, Daogao Liu, Ruoqi Shen, and Kevin Tian. Private convex optimization in general norms. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 5068–5089. SIAM, 2023.

[GT20] Arun Ganesh and Kunal Talwar. Faster differentially private samplers via Rényi divergence analysis of discretized Langevin MCMC. *Advances in Neural Information Processing Systems*, 33:7222–7233, 2020.

[GTU22] Arun Ganesh, Abhradeep Thakurta, and Jalaj Upadhyay. Langevin diffusion: An almost universal algorithm for private Euclidean (convex) optimization. *arXiv preprint arXiv:2204.01585*, 2022.

[HK12] Zhiyi Huang and Sampath Kannan. The exponential mechanism for social welfare: Private, truthful, and nearly optimal. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 140–149. IEEE, 2012.

[HRS16] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.

[HT10] 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.

[INS+19] Roger Iyengar, Joseph P Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 299–316. IEEE, 2019.

[JLLV21] He Jia, Aditi Laddha, Yin Tat Lee, and Santosh Vempala. Reducing isotropy and volume to KLS: an $o(n^3\psi^2)$ volume algorithm. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 961–974, 2021.

[JT14] Prateek Jain and Abhradeep Guha Thakurta. (Near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning*, pages 476–484. PMLR, 2014.

[KCK+18] Yu-Hsuan Kuo, Cho-Chun Chiu, Daniel Kifer, Michael Hay, and Ashwin Machanavajjhala. Differentially private hierarchical count-of-counts histograms. *Proceedings of the VLDB Endowment*, 11(11), 2018.

[KD99] Jayesh H Kotecha and Petar M Djuric. Gibbs sampling approach for generation of truncated multivariate Gaussian random variables. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 3, pages 1757–1760. IEEE, 1999.

[KJ16] Shiva Prasad Kasiviswanathan and Hongxia Jin. Efficient private empirical risk minimization for high-dimensional learning. In *International Conference on*

Machine Learning, pages 488–497. PMLR, 2016.

[KLL21] Janardhan Kulkarni, Yin Tat Lee, and Daogao Liu. Private non-smooth ERM and SCO in subquadratic steps. *Advances in Neural Information Processing Systems*, 34, 2021.

[KT13] Michael Kapralov and Kunal Talwar. On differentially private low rank approximation. In *Proceedings of the Twenty-fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1395–1414. SIAM, 2013.

[KV06] Adam Tauman Kalai and Santosh Vempala. Simulated annealing for convex optimization. *Mathematics of Operations Research*, 31(2):253–266, 2006.

[LC21] Jiaming Liang and Yongxin Chen. A proximal algorithm for sampling from non-smooth potentials. *arXiv preprint arXiv:2110.04597*, 2021.

[Led99] Michel Ledoux. Concentration of measure and logarithmic Sobolev inequalities. In *Séminaire de Probabilités XXXIII*, pages 120–216. Springer, 1999.

[LL21] Daogao Liu and Zhou Lu. Curse of dimensionality in unconstrained private convex ERM. *arXiv preprint arXiv:2105.13637*, 2021.

[LST20] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Logsmooth gradient concentration and tighter runtimes for metropolized Hamiltonian Monte Carlo. In *Conference on Learning Theory*, pages 2565–2597. PMLR, 2020.

[LST21] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted Gaussian oracle. In *Conference on Learning Theory*, pages 2993–3050. PMLR, 2021.

[LSV18] Yin Tat Lee, Zhao Song, and Santosh S Vempala. Algorithmic theory of ODEs and sampling from well-conditioned logconcave densities. *arXiv preprint arXiv:1812.06243*, 2018.

[LT19] Jingcheng Liu and Kunal Talwar. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 298–309, 2019.

[MASN16] Kentaro Minami, HItomi Arai, Issei Sato, and Hiroshi Nakagawa. Differential privacy without sensitivity. In *Advances in Neural Information Processing Systems*, pages 956–964, 2016.

[MBST21] Paul Mangold, Aurélien Bellet, Joseph Salmon, and Marc Tommasi. Differentially private coordinate descent for composite empirical risk minimization. *arXiv preprint arXiv:2110.11688*, 2021.

[Mir17] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

[MMW+21] Wenlong Mou, Yi-An Ma, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. High-order Langevin diffusion yields an accelerated MCMC algorithm. *Journal of Machine Learning Research*, 22:42–1, 2021.

[MT07] 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.

[MV21] Oren Mangoubi and Nisheeth K Vishnoi. Sampling from log-concave distributions with infinity-distance guarantees and applications to differentially private optimization. *arXiv preprint arXiv:2111.04089*, 2021.

[OV00] Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

[RBHT12] Benjamin IP Rubinstein, Peter L Bartlett, Ling Huang, and Nina Taft. Learning in a large function space: Privacy-preserving mechanisms for SVM learning. *Journal of Privacy and Confidentiality*, 4(1):65–100, 2012.

[Rén61] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561. University of California Press, 1961.

[RS16] Sofya Raskhodnikova and Adam Smith. Lipschitz extensions for node-private graph statistics and the generalized exponential mechanism. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 495–504. IEEE, 2016.

[SL19] Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 2100–2111, 2019.

[SSSSS09] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, volume 2(4), page 5, 2009.

[SSTT21] Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Evading the curse of dimensionality in unconstrained private GLMs. In *International Conference on Artificial Intelligence and Statistics*, pages 2638–2646. PMLR, 2021.

[TS13] Abhradeep Guha Thakurta and Adam Smith. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, pages 819–850. PMLR, 2013.

[VEH14] Tim Van Erven and Peter Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[Wan18] Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction and estimation in unbounded domain. *arXiv preprint arXiv:1803.02596*, 2018.

[WM10] Oliver Williams and Frank McSherry. Probabilistic inference and differential privacy. *Advances in Neural Information Processing Systems*, 23:2451–2459, 2010.

[WZ10] Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

[ZP19] Tianqing Zhu and S Yu Philip. Applying differential privacy mechanism in artificial intelligence. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 1601–1609. IEEE, 2019.

[ZZMW17] Jiaqi Zhang, Kai Zheng, Wenlong Mou, and Liwei Wang. Efficient private ERM for smooth objectives. In *IJCAI*, 2017.