Proceedings of Machine Learning Research vol 178:1–42, 2022 35th Annual Conference on Learning Theory

Private Convex Optimization via Exponential Mechanism

Sivakanth Gopi SIGOPI@MICROSOFT.COM
Microsoft Research

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

Daogao Liu DGLIU@UW.EDU
University of Washington

Editors: Po-Ling Loh and Maxim Raginsky

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 the DP-SCO 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* as 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.

Keywords: Differential Privacy, Exponential Mechanism, Convex Optimization, Sampling

1. Introduction

Differential Privacy (DP), introduced in Dwork et al. (2006a,b), is increasingly becoming the universally accepted standard in privacy protection. We see an increasing array of adoptions in industry Apple (2017); Erlingsson et al. (2014); Bittau et al. (2017); Ding et al. (2017) and more recently the US census bureau Abowd (2016); Kuo et al. (2018). Differential privacy allows us to quantify the privacy loss of an algorithm and is defined as follows.

Definition 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*

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

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.

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*

$$\mathop{\mathbb{E}}_{\mathcal{M}}[F(x^{priv};\mathcal{D})] - F(x^*;\mathcal{D}), \tag{1}$$

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

Exponential Mechanism One of the first mechanisms invented in differential privacy, the *exponential mechanism*, was proposed by McSherry and Talwar (2007) precisely to solve this. It involves sampling x^{priv} from the density

$$\pi_{\mathcal{D}}(x) \propto \exp\left(-kF(x;\mathcal{D})\right). \tag{2}$$

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 $k = \frac{\varepsilon}{2\Delta_F}$, the exponential mechanism satisfies $(\varepsilon, 0)$-DP.

Exponential mechanism is widely used both in theory and in practice, such as in mechanism design Huang and Kannan (2012), convex optimization Bassily et al. (2014); Mangoubi and Vishnoi (2021), statistics Wasserman and Zhou (2010); Williams and McSherry (2010); Awan et al. (2019), machine learning and AI Zhu and Philip (2019). Even for infinite and continuous domains, exponential mechanism can be implemented efficiently for many problems Hardt and Talwar (2010); Chaudhuri et al. (2013); Kapralov and Talwar (2013); Balcer and Vadhan (2019); Canonne et al. (2020). There are also several variants and generalizations of the exponential mechanism which can improve its utility based on different assumptions Thakurta and Smith (2013); Beimel et al. (2013); Raskhodnikova and Smith (2016); Liu and Talwar (2019). See Liu and Talwar (2019) for a survey of these results.

DP Empirical Risk Minimization (DP-ERM) In many applications, the loss function is given by the average of the loss of each user:

$$F(x;\mathcal{D}) := \frac{1}{n}\sum_{i=1}^{n} f(x;s_i). \tag{3}$$

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 $f(x;s)$ is convex and $f(x;s) - f(x;s')$ is G-Lipschitz for all s, s', and $\mathcal{K} \subset \mathbb{R}^d$ is convex with diameter D.[1] We call the problem of minimizing the excess empirical risk in (3) as DP Empirical Risk Minimization (DP-ERM). This setting is well studied by the DP community with many exciting results Chaudhuri and Monteleoni (2008); Rubinstein et al. (2012); Chaudhuri et al. (2011); Jain and Thakurta (2014); Bassily et al. (2014); Kasiviswanathan and Jin (2016); Fukuchi et al. (2017); Zhang et al. (2017); Wang (2018); Iyengar et al. (2019); Bassily et al. (2019); Feldman et al. (2020); Kulkarni et al. (2021); Bassily et al. (2021); Liu and Lu (2021); Asi et al. (2021); Song et al. (2021); Mangold et al. (2021); Ganesh et al. (2022).[2]

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

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

In particular, Bassily et al. (2014) shows that exponential mechanism in (2) achieves the optimal excess empirical risk of $O\left(\frac{GDd}{n\varepsilon}\right)$ under $(\varepsilon, 0)$-DP. On the other hand, Bassily et al. (2014, 2019, 2020) show that *noisy gradient descent* on $F(x; \mathcal{D})$ achieves an excess empirical risk of

$$O\left(\frac{GD\sqrt{d\log(1/\delta)}}{n\varepsilon}\right) \tag{4}$$

under (ε, δ)-DP, which is also shown to be optimal Bassily et al. (2014). This is a significant $\sqrt{d}$ improvement over the exponential mechanism.

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 exponential mechanism lacks or we didn't use exponential mechanism optimally in this setting. This brings us to the first question.

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

DP Stochastic Convex Optimization (DP-SCO) Beyond the privacy guarantee and the empirical risk guarantee, 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

$$\widehat{F}(x) = \mathop{\mathbb{E}}_{s\sim\mathcal{P}}[f(x; s)]. \tag{5}$$

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

$$\mathop{\mathbb{E}}_{\mathcal{M},\mathcal{D}\sim\mathcal{P}}[\widehat{F}(x^{priv})] - \widehat{F}(x^*) \tag{6}$$

where x^* is the minimizer of $\widehat{F}(x)$. We call the problem of minimizing the excess population loss in (6) as DP Stochastic Convex Optimization (DP-SCO). By a suitable modification of noisy stochastic gradient descent, Bassily et al. (2019); Feldman et al. (2020) show that one can achieve the optimal population loss of

$$O\left(GD\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\right)\right). \tag{7}$$

Bassily et al. (2019) bounds the generalization error by showing that running SGD on smooth functions is stable and Feldman et al. (2020) proposes an iterative localization technique. Note that only the algorithm for smooth functions in Bassily et al. (2019) 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. This brings us to the second question.

Question 1.2 *Can we achieve both the optimal empirical risk and the optimal population loss for non-smooth functions with the same algorithm?*

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 Feldman et al. (2020), Moreau-Yosida regularization and cutting plane methods),
- $\widetilde{O}(\min(n^{3/2}, n^2/\sqrt{d}))$ queries to $\nabla f(x;s)$ Asi et al. (2021),
- $\widetilde{O}(\min(n^{5/4}d^{1/8}, n^{3/2}/d^{1/8}))$ queries to $\nabla f(x;s)$ Kulkarni et al. (2021).

Combining these results, this 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 is open, it is 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 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 1.3 *Can we obtain an algorithm with optimal query complexity for DP-SCO for zeroth order query model?*

1.1. Our Contributions

In this paper, we give a positive answer to all these questions using the *Regularized Exponential Mechanism*. If we add an ℓ_2^2 regularizer to F and sample x^{priv} from the density

$$\exp\left(-k\left(F(x;\mathcal{D}) + \mu\left\|x\right\|_2^2/2\right)\right), \tag{8}$$

then, for a suitable choice of μ and k, we recover the optimal excess risk in (4) for DP-ERM and optimal population loss in (7) for DP-SCO. Finally, we give an algorithm to sample x^{priv} from the density (8) with 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 (in TV distance).

Formally, our result is follows:

Theorem 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 $\varepsilon, \delta \in (0, \frac{1}{10})$,* [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 25 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 some appropriate choices of k and μ. Furthermore, if $f(\cdot; s)$ is G-Lipschitz for all s, we can sample $x^{(priv)}$ using $O(\frac{\varepsilon^2 n^2}{\log(1/\delta)}\log^2(\frac{nd}{\delta}))$ queries in expectation to the values of $f(x; s)$.

Theorem 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 $\varepsilon, \delta \in (0, \frac{1}{10})$,*[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 appropriate choice of k and μ. Furthermore, if $f(\cdot; s)$ is G-Lipschitz for all s, we can sample $x^{(priv)}$ using $O(\min\{\frac{\varepsilon^2 n^2}{\log(1/\delta)}, nd\}\log^2(\frac{nd}{\delta}))$ 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 C), proving the optimality of our sampling algorithm. Moreover, when f is already 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.

2. Techniques

The main contribution of this paper is the discovery 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.

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. Minami et al. (2016) 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 fully take advantage of the strongly log-concave property.

[4] See Theorem 29 for general conclusions for all $\varepsilon > 0$.

Instead, we show directly that the privacy curve between a strongly log-concave distribution and its Lipschitz perturbation in the exponent is upper bounded by the privacy curve of an appropriate Gaussian mechanism. This new proof uses the notion of tradeoff function introduced in Dong et al. (2019) and the isoperimetric inequality for strongly log-concave distribution.

Theorem 4 *Given 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 4 is completely 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 3), then Shalev-Shwartz et al. (2009) 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 even impossible. Can we find a approximately good solution to ERM and hope that it is also a good solution for SCO? Feldman (2016) provides a negative answer and shows there is no good uniform convergence between $F(\cdot; \mathcal{D})$ and $\widehat{F}$, that is there always exists $x \in \mathcal{K}$ such that $|F(x; \mathcal{D}) - \widehat{F}(x)|$ is large. This fact forces us to find approximate solution to ERM with very high accuracy, which makes the algorithms inefficient.

Prior works proposed a few interesting ways to overcome this difficulty, such as the uniform stability in Hardt et al. (2016) and the iterative localization technique in Asi et al. (2021). 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. Bassily et al. (2019) makes use of the stability of running SGD on smooth functions to get a tight bound on the population loss for DP-SCO.

Recall $F(x; \mathcal{D})$ and $\widehat{F}(x)$ are defined in Equation (3) and (5) respectively. Our result enriches the toolbox of bounding the generalization error and provides new insights for this problem.

Theorem 5 *Suppose $\{f_i\}$ is a family of μ-strongly convex functions over $\mathcal{K}$ and $f_i - f_{i'}$ is G-Lipschitz for any two 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*

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

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 $e^{-kF(x;\mathcal{D}')}$. By the Talagrand transportation inequality, this can be done by bounding the KL divergence between these two distributions. Finally, a bound on the KL divergence easily follows from our privacy bounds. Therefore the sampling scheme is stable (in Wasserstein distance) and this leads to the $\frac{G^2}{\mu n}$ term in generalization error. The other term $\frac{d}{k}$ 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 Dalalyan (2017); Lee et al. (2018); Mou et al. (2021); Chen and Vempala (2019); Durmus et al. (2019); Shen and Lee (2019); Chen et al. (2020); Lee et al. (2020). For example, if $F = \frac{1}{n}\sum_{i=1}^{n} f_i$ 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 $\nabla f_i(x)$ Lee et al. (2021). 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
Bassily et al. (2014)	$d^{O(1)}$	$F(x)$	$\mathrm{D}_\infty \le \varepsilon$
Chatterji et al. (2020)	$G^{O(1)} d^{5/2}/\varepsilon^4$	$\nabla F(x)$	$\mathrm{W}_2 \le \delta$
Jia et al. (2021) + Chen (2021)	d^3	$F(x)$	$\mathrm{TV} \le \delta$
Ganesh and Talwar (2020)	$\frac{\alpha^2 G^4 d}{\varepsilon^2}$	$\nabla F(x)$	$\mathrm{D}_\alpha \le \varepsilon$
Liang and Chen (2021)	$\frac{G^2}{\delta}$	$\nabla F(x)$	$\mathrm{TV} \le \delta$
This	G^2	$f_i(x)$	$\mathrm{TV} \le \delta$

Figure 1: The complexity of sampling from $\exp(-F(x))$ where $F = \frac{1}{n}\sum_i f_i$ 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.

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 D.1). For applications in differential privacy, we need either polynomially small W_2 or TV distance, or ε small D_α distance.

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

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 Ganesh and Talwar (2020) or $\widetilde{O}(nd^3)$ many queries to $f_i(x)$ by combining Jia et al. (2021) and Chen (2021). In comparison, our algorithm only takes $\widetilde{O}(n^2)$ many $f_i(x)$.

Our result is based on the alternating sampler proposed in Lee et al. (2021) and a new rejection sampling scheme.

Theorem 6 *Given a μ-strongly convex function $\psi(x)$ 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 $x^* = \arg\min_{x\in\mathcal{K}} F(x)$. 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*

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

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

3. Preliminaries

We present some necessary definitions and background used in the paper. We refer to the Appendix for other basic definitions such as Wasserstein distance and log-concavity.

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 7 (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

$$\delta(\mathcal{N}(0,1) \parallel \mathcal{N}(s,1))(\varepsilon) = \Phi\left(-\frac{\varepsilon}{s} + \frac{s}{2}\right) - e^{\varepsilon}\Phi\left(-\frac{\varepsilon}{s} - \frac{s}{2}\right) \tag{9}$$

where $\Phi(\cdot)$ is the Gaussian cumulative distribution function (CDF) Balle and Wang (2018). 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 Dong et al. (2019) which is an equivalent way to describe the privacy curve $\delta(P\|Q)$.

Definition 8 (Tradeoff function) *Given two (continuous) distributions P, Q, we define the tradeoff 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 Dong et al. (2019)

$$T(\mathcal{N}(0,1)\|\mathcal{N}(s,1))(z) = \Phi(\Phi^{-1}(1-z) - s). \tag{10}$$

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 9 (Dong et al. (2019)) $\delta(P\|Q) \le \delta(P'\|Q')$ *iff* $T(P\|Q) \ge T(P'\|Q')$

3.2. Isoperimetric Inequality for Strongly Log-concave Distributions

The cumulative distribution function (CDF) of 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 10 (Theorem 1.1. in Ledoux (1999)) *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 11 *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 $r \ge 0$,*

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

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

4. GDP of Regularized Exponential Mechanism

In this section, we prove our DP result (Theorem 4). The proof uses the isoperimetric inequality for strongly log-concave measures Ledoux (1999). 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 appropriate standard deviation. This allows us

[6] Tradeoff curves in Dong et al. (2019) are defined using type I and type II errors. The definition given here is equivalent to their definition for continuous distributions.

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 9 and Theorem 4). We now prove our main privacy bound. Assume the following claim holds.

Claim 12

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

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

Theorem 13 *Given 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 $z \in [0,1]$,*

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

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 $S = \{x \in \mathcal{K} : \alpha(x) \geq m(z)\}$ for some $m(z)$ chosen such that $P(S) = \Pr_{x\sim P}[\alpha(x) \geq m(z)] = 1 - z$ (Neyman-Pearson lemma).

Therefore:

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

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

$$\mathbb{E}_P[e^{-\alpha(x)}\mathbf{1}_S] = \int_{x:\alpha(x)\geq m(z)} e^{-\alpha(x)}P(x)\mathrm{d}x = \mathbb{E}[e^{-Y}\mathbf{1}(Y \geq m(z))] = \int_{m(z)}^\infty e^{-t}f_Y(t)dt$$

Moreover,

$$\begin{aligned}
\int_{m(z)}^\infty e^{-t}f_Y(t)dt &= \int_{t=0}^\infty e^{-t-m(z)}\left(-\frac{\mathrm{d}\Pr_{x\sim P}\left[\alpha(x) \geq t + m(z)\right]}{\mathrm{d}t}\right)\mathrm{d}t \\
&= e^{-m(z)}\left(-e^{-t}\Pr_{x\sim P}\left[\alpha(x) \geq t + m(z)\right]\Big|_0^\infty - \int_{t=0}^\infty e^{-t}\Pr_{x\sim P}\left[\alpha(x) \geq t + m(z)\right]\mathrm{d}t\right) \\
&= (1-z)e^{-m(z)} - e^{-m(z)}\int_{t=0}^\infty e^{-t}\Pr_{x\sim P}\left[\alpha(x) \geq t + m(z)\right]\mathrm{d}t \\
&\geq (1-z)e^{-m(z)} - e^{-m(z)}\int_{t=0}^\infty e^{-t}\Phi(\Phi^{-1}(1-z) - t/\gamma)\mathrm{d}t \qquad \text{(Corollary 11)} \\
&= (1-z)e^{-m(z)} - e^{-m(z)}\left((1-z) - \exp\left(\frac{\gamma^2}{2} - \Phi^{-1}(1-z)\gamma\right)\Phi(\Phi^{-1}(1-z) - \gamma)\right) \\
&\qquad \text{(Claim 12)} \\
&= \exp\left(\frac{\gamma^2}{2} + \Phi^{-1}(z)\gamma - m(z)\right)\Phi(-\Phi^{-1}(z) - \gamma)
\end{aligned}$$

We can upper bound

$$\mathop{\mathbb{E}}_{P}[e^{-\alpha}\mathbf{1}_{\overline{S}}] \leq \exp\left(\frac{\gamma^2}{2} + \Phi^{-1}(z)\gamma - m(z)\right)\Phi(\Phi^{-1}(z)+\gamma)$$

in a similar way, and we refer to the Appendix for the full proof. Combining the two bounds, we get:

$$\begin{aligned} T(P\|Q)(z) = \left(1 + \frac{\mathbb{E}_P[e^{-\alpha}\mathbf{1}_{\overline{S}}]}{\mathbb{E}_P[e^{-\alpha}\mathbf{1}_S]}\right)^{-1} &\geq \left(1 + \frac{\Phi(\Phi^{-1}(z)+\gamma)}{\Phi(-\Phi^{-1}(z)-\gamma)}\right)^{-1} \\ &= \Phi(-\Phi^{-1}(z)-\gamma) \qquad \text{(Using } \Phi(x)+\Phi(-x)=1\text{)} \\ &= T(N(0,1)\parallel N(\gamma,1)). \qquad \text{(Eqn (10))} \end{aligned}$$

■

As a corollary to Theorem 13, we can bound any divergence measure that decreases under post-processing such as Renyi divergence or KL divergence. In particular, this also implies Renyi Differential Privacy Mironov (2017) of our algorithm. The proof can be found in the Appendix.

Corollary 14 *Suppose $F, \tilde{F}$ are two 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 we have*

$$\mathrm{D}(P\|Q) \leq \mathrm{D}\left(\mathcal{N}(0,1)\|\mathcal{N}\left(\frac{G\sqrt{k}}{\sqrt{\mu}}, 1\right)\right)$$

for any divergence measure D which decreases under post-processing. In particular,

$$\mathrm{D}_\alpha(P\|Q) \leq \frac{\alpha k G^2}{2\mu} \text{ and } \mathrm{D}_{KL}(P\|Q) \leq \frac{kG^2}{2\mu}.$$

5. Sampling and Optimization Results Overview

Due to the space limit, we briefly discuss our results on efficient non-smooth sampling algorithm and DP convex optimization in this section, and the details can be found in Appendix A and Appendix B.

5.1. Sampling

We study the following problem about sampling from a (non-smooth) log-concave distribution.

Problem 15 *Given a μ-strongly convex function $\psi(x)$ 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 $\mathcal{K}$. Our goal is to sample a point $x \in \mathcal{K}$ with probability proportionally to $\exp(-\widehat{F}(x))$ where*

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

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

Our sampler is based on the alternating sampling algorithm in Lee et al. (2021) (See algorithm 1). This algorithm reduces the problem of sampling from $\exp(-\widehat{F}(x))$ to sampling from $\exp(-\widehat{F}(x) - \frac{1}{2\eta}\|x - y\|^2)$ for some fixed η and for roughly $\frac{1}{\eta\mu}$ 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 $\exp(-\widehat{F}(x) - \frac{1}{2\eta}\|x - y\|^2)$ using only $\widetilde{O}(1)$ many steps.

Theorem 16 *(Lee et al., 2021, Theorem 1) Given a μ-strongly convex function F 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 1 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.

```
Algorithm 2: Implementation of Line 1
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 $\propto \exp(-\psi(x) - \frac{1}{2\eta}\|x - y\|_2^2)$
    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 $\frac{\alpha}{1+\alpha}$, break
    end
    Sample $u$ uniformly from $[0, 1]$.
until $u \leq \frac{1}{2}\rho$;
Return $x$
```

The properties of our sampler are demonstrated in the following theorem.

Theorem 17 *Given a μ-strongly convex function $\psi(x)$ 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 $\mathcal{K}$. 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 $\exp(-\psi(x) - \frac{\|x-y\|_2^2}{2\eta})$ 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*

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

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

5.2. DP Optimization

We analyze the performance of our algorithm for DP-ERM and DP-SCO in this section. As for DP-ERM, briefly speaking, we show sampling from $\exp(-k(F(x;\mathcal{D})+\mu\|x\|_2^2/2))$ for some appropriately chosen k,μ can achieve the optimal DP excess empirical risk. The privacy guarantee follows from our abstract theorem of the Regularized Exponential Mechanism (Theorem 13), and the utility guarantee of sampling schemes follows from the following standard result.

Lemma 18 (Utility Guarantee, (De Klerk and Laurent, 2018, 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 Kalai and Vempala (2006) for any linear function F, and Bassily et al. (2014) extends it to any convex function F with a slightly worse constant.

Now we bound the generalization error of the Regularized Exponential Mechanism for DP-SCO. By Corollary 14, for neighboring databases $\mathcal{D},\mathcal{D}'$ we can bound $\mathrm{D}_{KL}(\mathcal{A}(\mathcal{D}),\mathcal{A}(\mathcal{D}'))\le\frac{kG^2}{2n^2\mu}$ where $\mathcal{A}(\mathcal{D})$ is sampling from $\exp(-k(F(x;\mathcal{D})+\mu\left\|x\right\|_2^2/2))$. As the distributions are strongly log-concave, we can get an upper bound on the Wasserstein distance due to Talagrand transportation inequality. Recall for two probability distributions ν_1,ν_2, the Wasserstein distance is defined as

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

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

Theorem 19 (Talagrand transportation inequality, Theorem 1 in Otto and Villani (2000)) *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 w.r.t. π and with finite moments of order 2, we have*

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

From this we get $W_2(\mathcal{A}(\mathcal{D}),\mathcal{A}(\mathcal{D}'))\le\sqrt{\frac{2}{k\mu}\mathrm{D}_{KL}(\mathcal{A}(\mathcal{D}),\mathcal{A}(\mathcal{D}'))}\le\frac{G}{n\mu}$. A small bound on Wasserstein distance in some sense means the Regularized Exponential Mechanism is stable and thus has a small generalization error by the following Lemma.

Lemma 20 (Lemma 7 in Bousquet and Elisseeff (2002)) *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*

$$\mathop{\mathbb{E}}_{\mathcal{D}}[\widehat{F}(\mathcal{A}(\mathcal{D})) - F(\mathcal{A}(\mathcal{D}); \mathcal{D})] = \mathop{\mathbb{E}}_{\mathcal{D}, s' \sim \mathcal{P}, \mathcal{A}} \Big[f(\mathcal{A}(\mathcal{D}); s') - f(\mathcal{A}(\mathcal{D}'); s') \Big]. \tag{11}$$

To see how stability in the Wasserstein metric implies a good bound on the generalization error, suppose $f(\,\cdot\,; s)$ is G-Lipschitz for any s, then the RHS of (11) can be upper bounded by $G \cdot W_2(\mathcal{A}(\mathcal{D}), \mathcal{A}(\mathcal{D}'))$. Combining these ideas, we prove our main result on the generalization error.

Theorem 21 *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'. Suppose we sample our solution from density $\pi_{\mathcal{D}}(x) \propto e^{-kF(x;\mathcal{D})}$. 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 $s' \sim \mathcal{P}$, then $\mathrm{W}_2(\pi_{\mathcal{D}}, \pi_{\mathcal{D}'}) \leq \frac{G}{n\mu}$. We can bound the excess population loss as:*

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

Lastly, one can efficiently implement the sampling by our efficient sampler. In Appendix C, we complement our algorithmic results with a nearly matching information-theoretic lower bound on the zeroth order query complexity for DP-SCO.

Roadmap

In Appendix A, we present our efficient non-smooth sampling algorithm. In Appendix B, we show how to make use of our results to achieve optimal empirical risk for DP-ERM and analyze the generalization error to get optimal DP-SCO population loss. In Appendix C, we give information-theoretic lower bounds on the zeroth order query complexity for DP-SCO and (non-private) sampling scheme, which nearly match our upper bounds. Some omitted definitions and proofs can be found in Appendix D and Appendix E.

Acknowledgments

Lee and Liu are supported by NSF awards CCF-1749609, DMS-1839116, DMS-2023166, CCF-2105772, a Microsoft Research Faculty Fellowship, a Sloan Research Fellowship, and a Packard Fellowship.

References

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

Differential Privacy Team Apple. Learning with privacy at scale. *Technical report, Apple*, 2017.

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.

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.

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.

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

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.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

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

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.

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

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.

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.

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

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.

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

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.

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

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

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

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

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

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

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.

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.

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

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

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

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of langevin monte carlo via convex optimization. *The Journal of Machine Learning Research*, 20(1):2666–2711, 2019.

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, 2006a.

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, 2006b.

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.

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

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.

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.

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.

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

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

Michel Ledoux. Concentration of measure and logarithmic sobolev inequalities. In *Seminaire de probabilites XXXIII*, pages 120–216. Springer, 1999.

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.

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.

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.

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

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

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.

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

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.

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.

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

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

Wenlong Mou, Yi-An Ma, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. High-order langevin diffusion yields an accelerated mcmc algorithm. *J. Mach. Learn. Res.*, 22:42–1, 2021.

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.

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.

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.

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.

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

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.

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.

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.

Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

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

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

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

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

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.

Appendix A. Efficient Non-smooth Sampling

In this section, we finish the proof of the main result on our efficient sampler. Recall we want to sample from the probability proportionally to $\exp(-\widehat{F}(x))$ where $\widehat{F}(x) = \mathbb{E}_{i\in I} f_i(x) + \psi(x)$.

Since F has the ψ term, instead of sampling x from $\mathcal{N}(y, \eta \cdot I_d)$, we sample from $\exp(-\psi(x) - \frac{1}{2\eta}\|x-y\|^2)$ in Algorithm 2. The following lemma shows how to decompose the distribution $\exp(-\widehat{F}(x) - \frac{1}{2\eta}\|x-y\|^2)$ 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 22 *Let π be the distribution proportional to $\exp(-\widehat{F}(x) - \frac{1}{2\eta}\|x-y\|_2^2)$ and let $\mathcal{G}$ be the distribution proportional to $\exp(-\psi(x) - \frac{1}{2\eta}\|x-y\|^2)$. Then, we have that*

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

Let $\widetilde{\pi}$ be the distribution returns by Algorithm 2. Then, we have that

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

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, we have that

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

Proof For the true distribution π, we have

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

For the distribution $\widetilde{\pi}$ by the algorithm, we sample $x \sim \mathcal{G}$, then accept the sample if $u \leq \frac{1}{2}\rho$. 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, we have

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

Taking expectation over z gives the result. ■

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 $\frac{d\pi}{dx} = \frac{d\widetilde{\pi}}{dx}$. 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 23 (Gaussian concentration (Ledoux, 1999, Eq 1.21)) *Let $X \sim \exp(-\widehat{F})$ for some $1/\eta$-strongly convex $\widehat{F}$ and ℓ is 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 1 correctly up to small error.

Lemma 24 *If the step size $\eta \leq C\log^{-1}(1/\delta_{\text{inner}})G^{-2}$ 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 $\exp(-\widehat{F}(x) - \frac{1}{2\eta}\|x-y\|_2^2)$. Furthermore, the algorithm accesses only $O(1)$ many $f_i(x)$ in expectation and samples from $\exp(-\psi(x) - \frac{1}{2\eta}\|x-y\|_2^2)$ for $O(1)$ many y.*

Proof Let π be the distribution given by $c \cdot \exp(-\widehat{F}(x) - \frac{1}{2\eta}\|x-y\|_2^2)$ and $\widetilde{\pi}$ is the distribution outputted by the algorithm. By Lemma 22, we have

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

Let X be the random variable $\mathbb{E}(\rho|x)$ and $\widetilde{X}$ be the random variable $\mathbb{E}(\overline{\rho}|x)$. Lemma 22 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, we have

$$d_{\text{TV}}(\pi,\widetilde{\pi}) = \mathbb{E}\left|\frac{X}{\mathbb{E}\,X} - \frac{\widetilde{X}}{\mathbb{E}\,\widetilde{X}}\right| \leq \mathbb{E}\left|\frac{X}{\mathbb{E}\,X} - \frac{\widetilde{X}}{\mathbb{E}\,X}\right| + \mathbb{E}\left|\frac{\widetilde{X}}{\mathbb{E}\,X} - \frac{\widetilde{X}}{\mathbb{E}\,\widetilde{X}}\right| \leq 2\frac{\mathbb{E}\,|X-\widetilde{X}|}{|\,\mathbb{E}\,X|}. \tag{12}$$

We simplify the right hand side by lower bounding $\mathbb{E}\,X$. By Lemma 23 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 $\mathbb{E}[X] \geq \frac{2}{3}$. Using this, (12), $X = \mathbb{E}(\rho|x)$ and $\widetilde{X} = \mathbb{E}(\overline{\rho}|x)$, we have

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

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

$$d_{\mathrm{TV}}(\pi, \widetilde{\pi}) \leq 3 \cdot \mathbb{E}(|\rho_{>L}| \cdot 1_{\rho \notin [0,2]}) + 3 \cdot \mathbb{E}(|\rho_{\leq L}| \cdot 1_{\rho \notin [0,2]}). \tag{13}$$

For the term $\rho_{>L}$, by a calculation similar to Lemma 22, we have

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

where $\Phi(t) = \sum_{\alpha=L+1}^{\infty} \frac{t^\alpha}{\alpha!}$ is a power series in t with all positive coefficients. By picking $L > C \log(1/\delta_{\mathrm{inner}})$ for some large constant C, we have $\Phi(t) \leq \frac{\delta_{\mathrm{inner}}}{16}$ for all $|t| \leq 1$. 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, we have

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

Denote a function $h_{x,z}(t) := \Pr_{i \in I}[|f_i(z) - f_i(x)| \geq t]$. Since each f_i is G-Lipschitz, Lemma 23 shows that

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

which implies

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

By Markov inequality, for any $k > 0$, we know

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

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

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

Hence, one has

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

By Gaussian Concentration, we know

$$\begin{aligned}\Pr_{x,z}[\|x - z\|_2 \geq t] &\leq \Pr_{x,z}[\|x - \mathbb{E}\,x\|_2 \geq t/2 \text{ or } \|z - \mathbb{E}\,z\| \geq t/2] \\ &\leq 2e^{-t^2/(8\eta)}.\end{aligned}$$

Thus we know

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

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

$$\mathbb{E}(|\rho_{>L}|\cdot 1_{\rho\notin[0,2]})\leq\frac{\delta_{\text{inner}}}{16}+17\sum_{k=1}^{\infty}e^{k-\frac{k^2}{64G^2\eta}}\leq\frac{\delta_{\text{inner}}}{9}, \tag{14}$$

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

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

$$\begin{aligned}
&\mathbb{E}(|\rho_{\leq L}|\cdot 1_{\rho\notin[0,2]})\\
=&\mathbb{E}(|\rho_{\leq L}|\cdot 1_{\rho\notin[0,2]}\cdot 1_{|\rho_{\leq L}|\leq 2^L})+\mathbb{E}(|\rho_{\leq L}|\cdot 1_{\rho\notin[0,2]}\cdot 1_{|\rho_{\leq L}|\geq 2^L})\\
\leq&\Pr[\rho\notin[0,2]]\cdot 2^L+\sum_{k=1}^{\infty}2^{(k+1)L}\Pr(|\rho_{\leq L}|\geq 2^{kL}).
\end{aligned} \tag{15}$$

Note that the term $\rho_{\leq L}$ involves only less than $\frac{L^2}{2}$ many $f_i(x)$ and $f_i(z)$. Lemma 23 shows that for any i, we have

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

By union bound, this shows

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

Under the event $|f_i(x)-f_i(z)|\leq\frac{1}{3}2^k$ for all i appears in $\rho_{\leq L}$, we have

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

Therefore, we have $\Pr(|\rho_{\leq L}|>2^{kL})\leq L^2\exp(-\frac{4^k}{32\eta G^2})$ 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 $\eta\leq 2^{-8}G^{-2}L^{-1}$, we have that

$$\sum_{k=1}^{\infty}2^{(k+1)L}\Pr(|\rho_{\leq L}|>2^{kL})\leq\sum_{k=1}^{\infty}2^{4kL}\exp(-2\cdot 4^kL)\leq\sum_{k=1}^{\infty}2^{-kL}\leq\frac{\delta_{\text{inner}}}{9} \tag{16}$$

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 $\frac{1}{L!}$. Hence we know $\Pr[\rho \notin [0,2]] \leq \frac{1}{L!} + \Pr[\rho_{\leq L} \notin [0,2]]$. Similarly, by Gaussian Concentration and union bound, we have

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

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

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

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

$$\Pr[\rho \notin [0,2]]\cdot 2^L \leq 2^L(L^2\exp(-\frac{1}{8\eta G^2}) + \frac{1}{L!}) \leq \frac{\delta_{\text{inner}}}{9}. \tag{17}$$

Combining (13), (14), (15), (16) and (17), we have the result $d_{\text{TV}}(\pi, \widetilde{\pi}) \leq \delta_{\text{inner}}$.

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

Combining Theorem 16 and Lemma 24, we have the following result:

Theorem 17 *Given a μ-strongly convex function $\psi(x)$ 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 $\mathcal{K}$. 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 $\exp(-\psi(x) - \frac{\|x-y\|_2^2}{2\eta})$ 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*

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

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

Proof This follows from applying Lemma 24 to implement Line 1. 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 16 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 $\exp(-\psi(x) - \frac{1}{2\eta}\|x-y\|_2^2)$ is a truncated Gaussian distribution, and there are many results on how to sample from truncated Gaussian, e.g. Kotecha and Djuric (1999). For more general case, there are also efficient algorithms to do the sampling, such as the Projected Langevin Monte Carlo Bubeck et al. (2018). 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 Duchi et al. (2015) with some modifications. See Section C for a detailed discussion.

Appendix B. DP Convex Optimization

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

B.1. DP-ERM

In this subsection, we state our result for the DP-ERM problem (3). Briefly speaking, our main result (Theorem 4) shows that sampling from $\exp(-k(F(x;\mathcal{D})+\frac{\mu}{2}\|x\|_2^2))$ for some appropriately chosen k and μ is (ε,δ)-DP and achieves the optimal empirical risk in (4). Our sampling scheme in Section A provides an efficient implementation. We start with the following lemma which shows the utility guarantee for the sampling mechanism which was mentioned before.

Lemma 18 (Utility Guarantee, (De Klerk and Laurent, 2018, 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*

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

Theorem 25 (DP-ERM) *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 $f(x;s)-f(x;s')$ 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 $\frac{d}{k}+\frac{\mu D^2}{2}$. Moreover, if $\{f(\cdot,s)\}_{s\in\mathcal{D}}$ are already 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 $\frac{d}{k}$.

Proof The privacy guarantee follows directly from our main result Theorem 4, and the bound on excess empirical loss can be proved by Lemma 18. ■

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

Lemma 26 *For any constants $1/2>\delta>0$ and $\varepsilon>0$, if $|s|\le\sqrt{2\log(1/(2\delta))+2\varepsilon}-\sqrt{2\log(1/(2\delta))}$, one has*

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

Proof By Equation (9), we know that

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

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

Combining the sampling scheme (Theorem 17) and our analysis on DP-ERM, we can get the efficient implementation results on DP-ERM directly.

Theorem 27 (DP-ERM Implementation) *With same assumptions in Theorem 25, and assume $f(\cdot; s)$ is G-Lipschitz over $\mathcal{K}$ for all s. For any constants $1/10 > \delta > 0$ and $\varepsilon > 0$, there is an efficient sampler to solve DP-ERM which 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 $\frac{2GD\sqrt{d\log(1/\delta)}}{\varepsilon n}$. If $\varepsilon \ge \log(1/\delta)$, the expected excess empirical loss is bounded by $O(\frac{GD\sqrt{d}}{n\sqrt{\varepsilon}})$.*
- *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 26, we can set $s = \sqrt{2\log(3/(4\delta)) + 2\varepsilon} - \sqrt{2\log(3/(4\delta))}$ to make $\delta(\mathcal{N}(0,1) \parallel \mathcal{N}(s,1)) \le 2\delta/3$. For our setting, Theorem 25 shows that we have $s = \frac{G\sqrt{k}}{n\sqrt{\mu}}$ 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 it into the excess empirical loss bound of $\frac{d}{k} + \frac{\mu D^2}{2}$ 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.

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

To make it algorithmic, we apply Theorem 17 with the accuracy on the total variation distance to be $\min\{\delta/3, \frac{1}{cn^c\varepsilon}\}$ for some large enough 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 17. ■

B.2. DP-SCO and Generalization Error

We prove our main result on the generalization error (Theorem 21) first. As mentioned before, one can reduce the DP-SCO (5) to DP-ERM (3) by the iterative localization technique proposed by Feldman et al. (2020). 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 will make use of one famous inequality: *Talagrand transportation inequality*. Recall for two probability distributions ν_1, ν_2, the Wasserstein distance is equivalently defined as

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

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

Now we restate and prove our main result on the generalization error.

Theorem 21 *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'. Suppose we sample our solution from density $\pi_{\mathcal{D}}(x) \propto e^{-kF(x;\mathcal{D})}$. 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 $s' \sim \mathcal{P}$, then $\mathrm{W}_2(\pi_{\mathcal{D}}, \pi_{\mathcal{D}'}) \le \frac{G}{n\mu}$. We can bound the excess population loss as:*

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

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 14, we have

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

By the assumptions, we know both $F(x; \mathcal{D})$ and $F(x; \mathcal{D}')$ are μ-strongly convex and by Theorem 19, we have

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

By Lemma 20 and properties of Wasserstein distance, we have

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

Hence, we know that

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

where the last inequality follows from Lemma 18. ■

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

Theorem 28 (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 $f(x;s) - f(x;s')$ 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 data-set $\mathcal{D}$ are drawn i.i.d. from the underlying distribution $\mathcal{P}$, the excess population loss is bounded by $\frac{G}{n\mu} + \frac{d}{k} + \frac{\mu D^2}{2}$. Moreover, if $\{f(\cdot;s)\}_{s\in\mathcal{D}}$ are already 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 $\frac{G}{n\mu} + \frac{d}{k}$.

Proof The first part about privacy is a restatement of our result on DP-ERM (Theorem 27). The excess population loss (See Equation (6)) follows from the bound on generalization error (Theorem 21) and utility guarantee (Lemma 18). ■

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

Theorem 29 (DP-SCO Implementation) *With same assumptions in Theorem 28, and assume $f(\cdot;s)$ is G-Lipschitz over $\mathcal{K}$ for all s. For $0 < \delta < \frac{1}{10}$ and $0 < \varepsilon < \frac{1}{10}$, there is an efficient algorithm to solve DP-SCO which 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 30 *As for the non-typical case when $\varepsilon \geq 1/10$, one can use the bound in Theorem 27 and the bound on generalization error (Theorem 21) . 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 28, sampling from $\exp(-k(F(x;\mathcal{D}) + \mu\|x\|_2^2/2))$ when $k \leq \frac{\varepsilon^2 n^2 \mu}{2G^2 \log(3/(4\delta))}$ is $(\varepsilon, 2\delta/3)$-DP. Besides, we can set $k = \frac{\mu}{G^2}\min\{\frac{\varepsilon^2 n^2}{2\log(3/(4\delta))}, 2nd\}$ for arbitrarily large constant $c > 0$ to make the mechanism $(\varepsilon, 2\delta/3)$-differentially private, achieving tight population loss and decrease the running time. Then the population loss is upper bounded 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 upper bounded 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 it algorithmic, we also apply Theorem 17 with the accuracy on the total variation distance to be $\min\{\delta/3, \frac{1}{cn^c}\}$ 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 17. ■

Appendix C. 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 some logarithmic terms) the query complexity achieved by our algorithm (in Theorem 29). Our proof is similar to the previous works like Arias-Castro et al. (2012); Duchi et al. (2015) with some modifications.

Before stating the lower bound, we define some 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 using some randomness).

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

And we define the optimality gap

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

where $\widehat{F}(x) = \mathbb{E}_{s\sim\mathcal{P}} f(x; s)$, $\widehat{x}$ 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:

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

The minimax error is defined by

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

Theorem 31 *Let $\mathcal{K}$ be the ℓ_2 ball of diameter D in $\mathbb{R}^d$, then*

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

In particular, for any (randomized) algorithm $\mathcal{A}$ which 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 $\mathbb{E}_{s\sim\mathcal{P}}[\|\nabla f(x, s)\|_2^2] \leq G^2$ for all $x \in \mathcal{K}$, such that the output $\widehat{x}$ of the algorithm satisfies

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

C.1. Proof of Theorem 31

We reduce the optimization problem into a series of binary hypothesis tests. Recall we are considering linear functions $f(x; s) \overset{\text{def}}{=} \langle x, s\rangle$. Let $\mathcal{V} = \{-1, 1\}^d$ be a Boolean hyper-cube and for each $v \in \mathcal{V}$, let $\mathcal{N}_v = \mathcal{N}(\delta v, \sigma^2 I_d)$ be a Gaussian distribution for some parameters to be chosen such that $\widehat{F}_v(x) \overset{\text{def}}{=} \mathbb{E}_{s\sim\mathcal{N}_v}[f(x; s)] = \delta\langle x, v\rangle$. 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 Arias-Castro et al. (2012), we suppose that v is uniformly sampled from $\mathcal{V} = \{-1, 1\}^d$. Note that if we can find a good solution to $\widehat{F}_v(x)$, we need to determine the signs of vector v well. Particularly, we have the following claim:

Claim 32 (Duchi et al. (2015)) *For each $v \in \mathcal{V}$, let x^v minimize $\widehat{F}_v$ over $\mathcal{K}$ and obviously we know that $x^v = -v/\sqrt{d}$. For any solution $\widehat{x} \in \mathbb{R}^d$, we have*

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

where the function $\text{sign}(\cdot)$ *is defined as:*

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

Claim 32 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 $d_H(x, y) = \sum_{j=1} \mathbb{1}\{\text{sign}(x_j) \neq \text{sign}(y_j)\}$, and we have

$$\varepsilon_k^*(\mathcal{P}_G, \mathcal{K}) \ge \frac{\delta}{2\sqrt{d}} \{\inf_{\widehat{v}} \mathbb{E}[d_H(\widehat{v}, v)]\}, \tag{18}$$

where $\widehat{v}$ 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 (18), it suffices to lower bound the value of the testing error $\mathbb{E}[d_H(\widehat{v}, v)]$. As discussed in Arias-Castro et al. (2012); Duchi et al. (2015), the randomness in the algorithm can not 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.

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

Lemma 33 ((Arias-Castro et al., 2012, 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 $\pi_1 \overset{def}{=} 1 - \pi_{-1}$ respectively prior to the experiment. For any algorithm that takes one sample v and outputs $\widehat{i} : v \to \{-1, 1\}$, we define the Bayes risk B be the minimum average probability that algorithm fails (v is not sampled from $H_{\widehat{i}(v)}$). That is $B = \inf_{\widehat{i}} \pi_{-1} \Pr[\widehat{i}(v) = 1 \mid v \sim \mathbb{P}_{-1}] + \pi_1 \Pr[\widehat{i}(v) = 0 \mid v \sim \mathbb{P}_1]$. Then, we have*

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

Lemma 34 *Suppose that v is uniformly sampled from $\mathcal{V} = \{-1, 1\}^d$, then any estimate $\widehat{v}$ obeys*

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

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

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

where the first inequality follows from the definition of Bayes risk B_j, the second inequality follows by Lemma 33 and the last inequality follows by the Cauchy-Schwartz inequality.

To complete the proof, it suffices to show that

$$\sum_{j=1}^{d} \|\mathbb{P}_{1,j} - \mathbb{P}_{-1,j}\|_{\mathrm{TV}}^2 \leq \frac{\delta^2}{\sigma^2} k. \tag{19}$$

Assuming Equation (19) first, which will be established later. Then we know that

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

■

We will complete the proof of Lemma 34 by showing the following bounded total variation distance.

Claim 35

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

Proof Applying Pinsker's inequality, we know $\|\mathbb{P}_{1,j} - \mathbb{P}_{-1,j}\|_{\mathrm{TV}}^2 \le \frac{1}{2}\mathrm{D}_{KL}(\mathbb{P}_{-1,j}\|\mathbb{P}_{1,j})$. 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 $\mathbb{P}_{-1,j,v'}(Y^{[k]}) \overset{\mathrm{def}}{=} \mathbb{P}(Y^{[k]} \mid v_j = -1, v')$ to be the distribution conditional on $v_j = -1$ and v'. We have

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

The 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 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.

As 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 $j \ne j'$, 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 we have

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

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 $S^t \ne s_j$, 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 co-variance $\sigma^2 Q_i^{[t]\top} Q_i^{[t]}$.

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

Hence we have

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

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

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

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

This completes the proof. ■

Having Lemma 34, we can complete the proof of Theorem 31.

Proof of Theorem 31. As discussed before, we know

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

and hence we know that

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

where the last line follows from Lemma 34. We now set $\delta=\frac{\sigma\sqrt{d}}{2\sqrt{k}}$ and $\sigma=\frac{G}{\sqrt{d+d^2/4k}}$, so that $d(\sigma^2+\delta^2)=G^2$. Hence one has

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

Thus we complete the proof. ■

Corollary 36 (Lower bound for DP-SCO) *For any (non-private) algorithm which makes less than $O\left(\min\{\frac{\varepsilon^2 n^2}{\log(1/\delta)}, nd\}\right)$ 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 $f(x;s) \overset{def}{=} \langle x, s\rangle$, such that the output $\widehat{x}$ of the algorithm satisfies that*

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

Proof Note that Theorem 31 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 $\inf_{\mathcal{A}\in\mathbb{A}_k} \sup_{v\in\mathcal{V}} \mathbb{E}_{s\sim\mathcal{N}_v,\mathcal{A}}[\widehat{F}_v(\widehat{x}_k) - \widehat{F}_v^*] \geq \frac{GD}{16} \min\left\{1, \sqrt{\frac{d}{4k}}\right\}$ in Theorem 31, where $\widehat{x}_k$ is the output of $\mathcal{A}$ with k observations $Y^{[k]}$. To prove Corollary 36, we need to modify the distribution of s to satisfy the Lipschitz continuity.

In particularly, for some constant c, we know

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

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

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

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

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

If we use the distributions conditioned on $\max_{s_i\in\mathcal{D}} \|s_i\|_2 \leq cG\sqrt{1+\log(nd)/d}$ 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 $k = O\left(\min\{\frac{\varepsilon^2 n^2}{\log(1/\delta)}, nd\}\right)$ observations, we know

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

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

Corollary 37 (Lower bound for sampling scheme) *Given any $G > 0$ and $\mu > 0$. For any algorithm which takes function values queries less than $O\left(\frac{G^2}{\mu}/(1+\log(G^2/\mu)/d)\right)$ 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 By a similar argument in the proof of Corollary 36, 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

$$\mathbb{E}\left[\widehat{F}_v(\widehat{x}_k) - \widehat{F}_v^*\right] \geq \Omega\left(\frac{GD}{\sqrt{1+\log(k)/d}} \cdot \min\left\{\sqrt{\frac{d}{k}}, 1\right\}\right), \tag{20}$$

where $\widehat{F}_v^* = \min_{x\in\mathcal{K}} \widehat{F}_v(x)$ and $\widehat{x}_k \in \mathcal{K}$ 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 $p(x) := \exp(-\widehat{F}_v(x) - \frac{\mu}{2}\|x\|^2)$ on $\mathcal{K}$ with total variation distance $\eta \leq \min(1/2, \sqrt{d\mu/G^2})$.

Lemma 18 shows that

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

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

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

Note that we set $D = \sqrt{d/\mu}$. Comparing with (20), 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 $k = \Omega(\frac{G^2 d/\mu}{\log(G^2/\mu)})$. 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. ■

Appendix D. Omitted Definitions

D.1. Distribution Distance and Divergence

We present some distribution distances or divergences mentioned or used in this work.

Definition 38 *(Rényi, 1961, 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*

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

We follow the convention that $\frac{0}{0} = 0$. Rényi Divergence of orders $\alpha = 1, \infty$ are defined by continuity.

Definition 39 (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 40 (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*

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

Definition 41 (Kullback–Leibler divergence) *The Kullback–Leibler divergence between probability measures π and ν is defined by*

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

D.2. Optimization

Here we collect some properties of functions which are useful for optimization and sampling.

Definition 42 (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 the following holds for all $\omega, \omega' \in \mathcal{K}$: $|f(\omega) - f(\omega')| \leq L\|\omega - \omega'\|_2$.*

Definition 43 (μ-Strongly convex) *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 and the following inequality holds for all points $\omega, \omega' \in \mathcal{K}$,*

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

Definition 44 (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.*

Appendix E. Omitted Proofs

Corollary 11 *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 $\Pr_{x\sim\pi}[\alpha(x) \leq m(z)] = z$. Then for every $r \geq 0$,*

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

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

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

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

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

Claim 12

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

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

Proof

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

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

■

Theorem 13 *Given 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 $z \in [0,1]$,*

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

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 $S = \{x \in \mathcal{K} : \alpha(x) \geq m(z)\}$ for some $m(z)$ chosen such that $P(S) = \Pr_{x\sim P}[\alpha(x) \geq m(z)] = 1 - z$ (Neyman-Pearson lemma).

Therefore:

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

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

$$\begin{aligned} \mathop{\mathbb{E}}_P[e^{-\alpha(x)}\mathbf{1}_S] &= \int_{x:\alpha(x)\geq m(z)} e^{-\alpha(x)}P(x)\mathrm{d}x = \mathbb{E}[e^{-Y}\mathbf{1}(Y \geq m(z))] = \int_{m(z)}^{\infty} e^{-t} f_Y(t)dt \\ &= \int_{t=0}^{\infty} e^{-t-m(z)} \left(-\frac{\mathrm{d}\Pr_{x\sim P}[\alpha(x) \geq t + m(z)]}{\mathrm{d}t}\right)\mathrm{d}t \\ &= e^{-m(z)}\left(-e^{-t}\Pr_{x\sim P}[\alpha(x) \geq t + m(z)]\Big|_0^{\infty} - \int_{t=0}^{\infty} e^{-t}\Pr_{x\sim P}[\alpha(x) \geq t + m(z)]\,\mathrm{d}t\right) \\ &= (1-z)e^{-m(z)} - e^{-m(z)}\int_{t=0}^{\infty} e^{-t}\Pr_{x\sim P}[\alpha(x) \geq t + m(z)]\,\mathrm{d}t \\ &\geq (1-z)e^{-m(z)} - e^{-m(z)}\int_{t=0}^{\infty} e^{-t}\Phi(\Phi^{-1}(1-z) - t/\gamma)\mathrm{d}t \qquad \text{(Corollary 11)} \\ &= (1-z)e^{-m(z)} - e^{-m(z)}\left((1-z) - \exp\left(\frac{\gamma^2}{2} - \Phi^{-1}(1-z)\gamma\right)\Phi(\Phi^{-1}(1-z) - \gamma)\right) \\ &\qquad \text{(Claim 12)} \\ &= \exp\left(\frac{\gamma^2}{2} + \Phi^{-1}(z)\gamma - m(z)\right)\Phi(-\Phi^{-1}(z) - \gamma) \end{aligned}$$

We can upper bound $\mathbb{E}_P[e^{-\alpha}\mathbf{1}_{\overline{S}}]$ in a similar way.

$$\begin{aligned}
\mathop{\mathbb{E}}_{P}[e^{-\alpha(x)}\mathbf{1}_{\overline{S}}] &= \int_{x:\alpha(x)\le m(z)} e^{-\alpha(x)}P(x)\mathrm{d}x \\
&= \int_{t=0}^{\infty} e^{-m(z)+t}\left(-\frac{d\,\mathrm{Pr}_{x\sim P}\left[\alpha(x)\le m(z)-t\right]}{dt}\right)\mathrm{d}t \\
&= e^{-m(z)}\left(-e^t \Pr_{x\sim P}\left[\alpha(x)\le m(z)-t\right]\Big|_0^{\infty} + \int_{t=0}^{\infty} e^t \Pr_{x\sim P}\left[\alpha(x)\le m(z)-t\right]\mathrm{d}t\right) \\
&= ze^{-m(z)} + e^{-m(z)}\int_{t=0}^{\infty} e^t \Pr_{x\sim P}\left[\alpha(x)\le m(z)-t\right]\mathrm{d}t \\
&\le ze^{-m(z)} + e^{-m(z)}\int_{t=0}^{\infty} e^t\Phi(\Phi^{-1}(z)-t/\gamma)\mathrm{d}t && \text{(Corollary 11)} \\
&= ze^{-m(z)} + e^{-m(z)}\left(-z+\exp\left(\frac{\gamma^2}{2}+\Phi^{-1}(z)\gamma\right)\Phi(\Phi^{-1}(z)+\gamma)\right) && \text{(Claim 12)} \\
&= \exp\left(\frac{\gamma^2}{2}+\Phi^{-1}(z)\gamma-m(z)\right)\Phi(\Phi^{-1}(z)+\gamma)
\end{aligned}$$

Combining the two bounds, we get:

$$\begin{aligned}
T(P\|Q)(z) &= \left(1+\frac{\mathbb{E}_P[e^{-\alpha}\mathbf{1}_{\overline{S}}]}{\mathbb{E}_P[e^{-\alpha}\mathbf{1}_S]}\right)^{-1} \\
&\ge \left(1+\frac{\Phi(\Phi^{-1}(z)+\gamma)}{\Phi(-\Phi^{-1}(z)-\gamma)}\right)^{-1} \\
&= \Phi(-\Phi^{-1}(z)-\gamma) && \text{(Using } \Phi(x)+\Phi(-x)=1\text{)} \\
&= T(N(0,1)\ \|\ N(\gamma,1)). && \text{(Eqn (10))}
\end{aligned}$$

■

Corollary 14 *Suppose $F,\tilde{F}$ are two 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 we have*

$$\mathrm{D}(P\|Q)\le \mathrm{D}\left(\mathcal{N}(0,1)\|\mathcal{N}\left(\frac{G\sqrt{k}}{\sqrt{\mu}},1\right)\right)$$

for any divergence measure D *which decreases under post-processing. In particular,*

$$\mathrm{D}_\alpha(P\|Q)\le\frac{\alpha kG^2}{2\mu}\ \textit{and}\ \mathrm{D}_{KL}(P\|Q)\le\frac{kG^2}{2\mu}.$$

Proof By Theorem 2.10 in Dong et al. (2019), 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 13. It is well-known that Renyi divergence and KL divergence decrease with post-processing (see Van Erven and Harremos (2014), for example). We can also compute $\mathrm{D}_\alpha(\mathcal{N}(0,1), \mathcal{N}(s,1)) = \alpha s^2/2$ and $\mathrm{D}_{KL}(\mathcal{N}(0,1), \mathcal{N}(s,1)) = s^2/2$ (Mironov (2017)). ■