Differentially Private Stochastic Convex Optimization in (Non)-Euclidean Space Revisited

Abstract

In this paper, we revisit the problem of Differentially Private Stochastic Convex Optimization (DP-SCO) in Euclidean and general ℓ_p^d spaces. Specifically, we focus on three settings that are still far from well understood: (1) DP-SCO over a constrained and bounded (convex) set in Euclidean space; (2) unconstrained DP-SCO in ℓ_p^d space; (3) DP-SCO with heavy-tailed data over a constrained and bounded set in ℓ_p^d space. For problem (1), for both convex and strongly convex loss functions, we propose methods whose outputs could achieve (expected) excess population risks that are only dependent on the Gaussian width of the constraint set, rather than the dimension of the space. Moreover, we also show the bound for strongly convex functions is optimal up to a logarithmic factor. For problems (2) and (3), we propose several novel algorithms and provide the first theoretical results for both cases when $1 < p < 2$ and $2 \leq p \leq \infty$.

1 INTRODUCTION

Learning from data that contains sensitive information has become a critical consideration. It enforces machine learning algorithms to not only learn effectively from the training data but also provide a certain level of guarantee on privacy preservation. To address the privacy concern, as a rigorous notion for statistical data privacy, differential privacy (DP) [Dwork et al., 2006] has received much attention in the past few years and has become a de facto technique for private data analysis.

As the two most fundamental models in machine learning, Stochastic Convex Optimization (SCO) [Vapnik, 1999] with its empirical form, Empirical Risk Minimization (ERM), can find numerous applications, such as biomedicine and healthcare. However, as these applications always involve sensitive data, it is essential to design DP algorithms for SCO and ERM, which corresponds to the problem of DP-SCO and DP-ERM, respectively. DP-SCO and DP-ERM have been extensively studied for over a decade, starting from Chaudhuri and Monteleoni [2008]. For example, Bassily et al. [2014] presents the optimal rates of general DP-ERM for both convex and strongly loss functions. Bassily et al. [2019], Feldman et al. [2020] later study the optimal rates of general DP-SCO, which is later extended by Su et al. [2022], Asi et al. [2021b] to loss functions that satisfy the growth condition. Bassily et al. [2021], Asi et al. [2021a] provide the first study on DP-SCO over non-Euclidean space, i.e., the ℓ_p space with $1 \leq p \leq \infty$.

While there are a vast number of studies on DP-SCO/DP-ERM, there are still several open problems left, especially the constrained case in Euclidean space where the convex constraint set has some specific geometric structures, and the case where the space is non-Euclidean. In detail, while it has been shown that the optimal rate of DP-ERM over ℓ_2-norm ball depends on $O(\sqrt{d})$ and $O(d)$ for convex and strongly convex loss, respectively [Bassily et al., 2014], Talwar et al. [2014] show that for general constraint set $\mathcal{C}$, the bound on d could be improved to $O(G_{\mathcal{C}})$ and $O(G_{\mathcal{C}}^2)$ for these two classes of functions, where $G_{\mathcal{C}}$ is the Gaussian width of set $\mathcal{C}$ (see Definition 12 for details), which could be far less than the dimension d. However, compared to DP-ERM with Gaussian width, DP-SCO with Gaussian width is far from well understood. The best-known result even cannot recover the optimal rate of the ℓ_2-norm ball case [Amid et al., 2022]. For the non-Euclidean case, Bassily et al. [2021] only study the constrained case where the constrained set has a bounded diameter. Theoretical behaviors for the unconstrained case are still unknown. Moreover, In the Euclidean case, recently, there has been a line of work focusing on DP-SCO where the distribution of loss gradients is heavy-tailed rather than uniformly bounded [Wang et al., 2020, Hu et al., 2022, Kamath et al., 2022]. However, non-Euclidean DP-SCO with heavy-tailed data has not been studied so far.

In this paper, we study the theoretical behaviors of three problems: (1) DP-SCO (with Lipschitz loss) over a convex constraint set $\mathcal{C}$ in Euclidean space; (2) unconstrained DP-SCO in ℓ_p^d space; (3) DP-SCO with heavy-tailed data

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

over a convex constraint set $\mathcal{C}$ in ℓ_p^d space. Specifically, our contributions can be summarized as follows.

1. For problem (1), we consider both convex and strongly convex (smooth) loss functions. We show that for convex functions, there is an (ϵ, δ)-DP algorithm whose output could achieve an (expected) excess population risk of $O(\frac{G_{\mathcal{C}}\sqrt{\log(1/\delta)}}{\epsilon n} + \frac{1}{\sqrt{n}})$, where n is the sample size. The rate could be improved to $O(\frac{G_{\mathcal{C}}^2 \log(1/\delta)}{n^2\epsilon^2} + \frac{1}{n})$ for strongly convex functions. Moreover, we also show that the bound for strongly convex functions is optimal up to a factor of $\text{Poly}(\log d)$ if $\mathcal{C}$ is contained in the unit ℓ_2-norm ball. To the best of our knowledge, this is the first lower bound of DP-SCO that depends on Gaussian width.

2. We then study problem (2). Specifically, when $1 < p < 2$, we propose a novel method named Noisy Regularized Mirror Descent, which adds regularization terms and Generalized Gaussian noise to Mirror Descent. By analyzing its stability, we show the output could achieve an excess population risk of $\tilde{O}(\kappa^{\frac{4}{5}}(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{2}{5}})$, where $\kappa = \min\{\frac{1}{p-1}, 2\log d\}$. We also discuss the case when $2 \le p \le \infty$.

3. Finally, we consider problem (3), assuming that the second-order moment of $\|\cdot\|_*$-norm of the loss gradient is bounded. When $1 < p < 2$, through a noisy, shuffled, and truncated version of Mirror Descent, we show a bound of $\tilde{O}(\frac{\sqrt[4]{\kappa^2 d\log(1/\delta))}}{\sqrt{n\epsilon}})$ in the high privacy regime $\epsilon = \tilde{O}(n^{-\frac{1}{2}})$, and a bound of $O(\frac{\kappa^{\frac{2}{3}}(d\log(1/\delta))^{\frac{1}{6}}}{(n\epsilon)^{\frac{1}{3}}})$ for general $0 < \epsilon < 1$. We also study the case when $2 \le p \le \infty$.

2 RELATED WORK

As there is a long list of work on DP-SCO/DP-ERM, here we just mention the work close to the problems we study in this paper. See Table 1 and 2 for detailed comparisons.

DP-SCO/DP-ERM with Gaussian width. For DP-ERM over ℓ_2-norm ball, although Bassily et al. [2014] show the optimal rate of $O(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})$ and $O(\frac{d\log(1/\delta)}{n^2\epsilon^2})$ for convex and strongly convex loss, respectively, Talwar et al. [2014] show that for general constraint set $\mathcal{C}$ it is possible to improve the factor d to the Gaussian width of $\mathcal{C}$. After that, Kasiviswanathan and Jin [2016] further improve the rate for generalized linear functions, Wang et al. [2017] provide an accelerated algorithm, and Wang and Xu [2019] extend to non-convex loss functions. However, all of them only study the problem of DP-ERM, and their methods cannot be generalized to DP-SCO directly. For DP-SCO, the only known result is given by Amid et al. [2022], which studies general convex loss under the setting where there is some public data. As we can see from Table 1, our result significantly improves theirs. Moreover, we show a nearly optimal rate for strongly convex functions, which is the first lower bound of DP-SCO/DP-ERM that depends on the Gaussian width.

DP-SCO in ℓ_p^d space. Compared to the Euclidean space case, there is little work on DP-SCO in non-Euclidean (ℓ_p^d) space. Bassily et al. [2021] provide the first study of the problem for $1 \le p \le \infty$ and propose several results for $p = 1$, $1 < p < 2$ and $2 \le p \le \infty$. Later Han et al. [2022] further extend to the online setting. However, all the previous algorithms and utility analyses highly rely on the assumption that the diameter of the constrained set is bounded and known, i.e., their results will not hold in the unconstrained case, which is more difficult than the constrained case. In this paper, we fill the gap by providing the first results for unconstrained DP-SCO in ℓ_p^d space by proposing several new methods.

3 PRELIMINARIES

In this section, we recall some definitions and lemmas that would be used throughout the paper.

Definition 1 (Differential Privacy [Dwork et al., 2006])**.** Given a data universe $\mathcal{X}$, we say that two datasets $D, D' \subseteq \mathcal{X}$ are neighbors if they differ by only one data sample, which is denoted as $D \sim D'$. A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private (DP) if for all neighboring datasets D, D' and for all events S in the output space of $\mathcal{A}$, we have $\Pr(\mathcal{A}(D) \in S) \le e^{\epsilon}\Pr(\mathcal{A}(D') \in S) + \delta$.

Lemma 1 (Advanced Composition Theorem Dwork et al. [2014])**.** Given target privacy parameters $0 < \epsilon < 1$ and $0 < \delta < 1$, to ensure $(\epsilon, T\delta' + \delta)$-DP over T mechanisms, it suffices that each mechanism is (ϵ', δ')-DP, where $\epsilon' = \frac{\epsilon}{2\sqrt{2T\ln(2/\delta)}}$ and $\delta' = \frac{\delta}{T}$.

Definition 2 (DP-SCO in General Normed Space [Bassily et al., 2021])**.** Given a dataset $D = \{x_1, \cdots, x_n\}$ from a data universe $\mathcal{X}$ where $\{x_i = (z_i, y_i)\}_i$ with a feature vector z_i and a label/response y_i are i.i.d. samples from some unknown distribution $\mathcal{D}$, a normed space $(\mathbf{E}, \|\cdot\|)$ of dimension d, a convex constraint set $\mathcal{C} \subseteq \mathbf{E}$, and a convex loss function $\ell : \mathcal{C} \times \mathcal{X} \mapsto \mathbb{R}$. Differentially Private Stochastic Convex Optimization (DP-SCO) is to find θ^{priv} to minimize the population risk, *i.e.,* $\mathcal{L}(\theta) = \mathbb{E}_{x\sim\mathcal{D}}[\ell(\theta, x)]$ with the guarantee of being differentially private.[1] The utility of the algorithm is measured by the (expected) excess population risk, that is $\mathcal{L}(\theta^{\text{priv}}) - \mathcal{L}(\theta^*)$, where $\theta^* = \arg\min_{\theta\in\mathcal{C}} \mathcal{L}(\theta)$. Besides the population risk, we can also measure the *empirical risk* of dataset D: $\hat{\mathcal{L}}(\theta, D) = \frac{1}{n}\sum_{i=1}^n \ell(\theta, x_i)$.

In Definition 2, we consider DP-SCO in general normed space with a convex set $\mathcal{C} \subseteq \mathbf{E}$. In this paper, we mainly focus on two cases: 1) Constraint Euclidean case where $\mathbf{E} = \mathbb{R}^d$, $\|\cdot\|$ is the ℓ_2-norm, and $\mathcal{C}$ is a bounded set whose diameter is denoted as $\|\mathcal{C}\|_2 = \max_{\theta,\theta'\in\mathcal{C}} \|\theta - \theta'\|_2$; 2) ℓ_q^d

[1] Note that in this paper we consider the proper learning case, that is θ^{priv} should be in $\mathcal{C}$.

Methods	Problem	Assumption	Convex Bound	Strongly Convex Bound
[Talwar et al., 2014]	ERM	Lipschitz	$\tilde{O}(\frac{G_{\mathcal{C}}}{n\epsilon})$	$\tilde{O}(\frac{G_{\mathcal{C}}^2}{n^2\epsilon^2})$
[Kasiviswanathan and Jin, 2016]	ERM	Lipschitz and GLM	$\tilde{O}(\frac{\sqrt{G_{\mathcal{C}}}}{\sqrt{n\epsilon}})$	—
Amid et al. [2022]	SCO	Lipschitz	$\tilde{O}(\frac{\sqrt{G_{\mathcal{C}}}}{\sqrt{n}n_{public}^{1/4}} + \frac{1}{\sqrt{n}})$	—
This paper	SCO	Lipschitz	$\tilde{O}(\frac{G_{\mathcal{C}}}{n\epsilon} + \frac{1}{\sqrt{n}})$	$\tilde{O}(\frac{G_{\mathcal{C}}^2}{n^2\epsilon^2} + \frac{1}{n})$ (*)

Table 1: Comparisons on the results for (ϵ, δ) DP-SCO/DP-ERM in Euclidean space with bounded constraint set $\mathcal{C}$ (dependence on other parameters are omitted). Here $G_{\mathcal{C}}$ is the Gaussian width of $\mathcal{C}$, n is the sample size, and n_{public} is the size of public data. $\tilde{O}$ hides other logarithmic factors. (*): We also show such a bound is nearly optimal when $\mathcal{C}$ is contained in unit ℓ_2 ball.

Methods	Constrained	Assumption	Bound for ℓ_p^d $(1 < p < 2)$	Bound for ℓ_p^d $(2 \leq p \leq \infty)$
[Bassily et al., 2021]	Yes	Lipschitz	$\tilde{O}(\sqrt{\frac{\kappa}{n}} + \frac{\kappa\sqrt{d}}{n\epsilon})$	$\tilde{O}(\frac{d^{\frac{1}{2}-\frac{1}{p}}}{\sqrt{n}} + \frac{d^{1-\frac{1}{p}}}{n\epsilon})$
This paper	No	Lipschitz	$\tilde{O}(\kappa^{\frac{4}{5}} \cdot (\frac{\sqrt{d}}{n\epsilon})^{\frac{2}{5}})$	$\tilde{O}(d^{1-\frac{2}{p}}(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\epsilon n}))$
This paper	Yes	Heavy-tailed	$\tilde{O}(\frac{\sqrt[4]{\kappa^2 d})}{\sqrt{n\epsilon}})/\tilde{O}(\frac{\kappa^{\frac{2}{3}}(d)^{\frac{1}{6}}}{(n\epsilon)^{\frac{1}{3}}})$ (*)	$\tilde{O}(\frac{d^{\frac{3}{2}-\frac{1}{p}}}{\sqrt{n}} + \frac{d^{\frac{3}{2}-\frac{1}{2p}}}{\sqrt{n\epsilon}})$

Table 2: Comparisons on the results for (ϵ, δ) DP-SCO in ℓ_p^d space with $1 < p \leq \infty$ (dependence on other parameters are omitted). Here d is the dimension, n is the sample size, and $\kappa = \min\{\frac{1}{p-1}, 2\log d\}$. $\tilde{O}$ hides other logarithmic factors. (*): The first bound is for the case of $\epsilon = \tilde{O}(n^{-\frac{1}{2}})$ and the second one is for general $0 < \epsilon < 1$.

case where $\mathbf{E} = \mathbb{R}^d$ and $\|\cdot\|$ is the ℓ_p-norm $\|\cdot\|_p$ with $1 < p \leq \infty$ (where $||x||_p = (\sum_{j=1}^d |x_j|^p)^{\frac{1}{p}}$), and $\mathcal{C}$ could be either bounded or unbounded. Since ℓ_p^d spaces are regular. To better illustrate our idea, we will introduce regular spaces.

Let $(\mathbf{E}, ||\cdot||)$ be a normed space of dimension d and let $\langle\cdot,\cdot\rangle$ be an arbitrary inner product over $\mathbf{E}$ (not necessarily inducing the norm $\|\cdot\|$). The dual norm over $\mathbf{E}$ is defined as $||y||_* = \max_{||x||\leq 1} \langle y, x\rangle$. So $(\mathbf{E}, ||\cdot||_*)$ is also a d-dimensional normed space. For example, let $\ell_p^d = (\mathbb{R}^d, ||\cdot||_p)$ with $1 \leq p \leq \infty$, the dual norm of ℓ_p^d is ℓ_q^d, where $\frac{1}{p} + \frac{1}{q} = 1$.

We call a normed space regular if its dual norm is sufficiently smooth. In detail, we have the following definition.

Definition 3 (κ-regular Space Juditsky and Nemirovski [2008])**.** Given $\kappa \geq 1$, we say a normed space $(\mathbf{E}, ||\cdot||)$ κ-regular if there exists a κ_+, s.t., $1 \leq \kappa_+ \leq \kappa$ and there exists a norm $||\cdot||_+$ such that $(\mathbf{E}, ||\cdot||_+)$ is κ_+-smooth, i.e., for all $x, y \in \mathbf{E}$,

$$||x+y||_+^2 \leq ||x||_+^2 + \langle \nabla(||\cdot||_+^2)(x), y\rangle + \kappa_+||y||_+^2.$$

And $||\cdot||$ and $||\cdot||_+$ are equivalent with the following constraint: $||x||^2 \leq ||x||_+^2 \leq \frac{\kappa}{\kappa_+}||x||^2$ $(\forall x \in \mathbf{E})$.

For ℓ_p^d space with $2 \leq p \leq \infty$, it is κ-regular with $\kappa = \min\{p-1, 2e\log d\}$. In this case we have $\|x\|_+ = \|x\|_r$ with $r = \min\{p, 2\log d + 1\}$ and $\kappa_+ = (r-1)$ Dümbgen et al. [2010]. So in the ℓ_p spaces with $1 < p < 2$ we focus on, their dual spaces are κ-regular with $\kappa = \min\{\frac{1}{p-1}, 2\ln d\}$.

In the following, we introduce the mechanisms that will be used in the latter sections.

Lemma 2 (Gaussian Mechanism)**.** Given a dataset $D \in \mathcal{X}^n$ and a function $q : \mathcal{X}^n \to \mathbb{R}^d$, the Gaussian mechanism is defined as $q(D) + \xi$ where $\xi \sim \mathcal{N}(0, \frac{2\Delta_2^2(q)\log(1.25/\delta)}{\epsilon^2}\mathbb{I}_d)$, where $\Delta_2(q)$ is the ℓ_2-sensitivity of the function q, *i.e.,* $\Delta_2(q) = \sup_{D\sim D'} \|q(D) - q(D')\|_2$. Gaussian mechanism preserves (ϵ, δ)-DP.

Note that the Gaussian mechanism is tailored for the case where the query has bounded ℓ_2-norm sensitivity. Bassily et al. [2021] propose a Generalized Gaussian mechanism that leverages the regularity of the dual space $(\mathbf{E}, \|\cdot\|_*)$.

Definition 4 (Generalized Gaussian distribution Bassily et al. [2021])**.** Let $(\mathbf{E}, ||\cdot||_*)$ be a d-dimensional κ-regular space with smooth norm $||\cdot||_+$. Define the generalized Gaussian distribution $\mathcal{GG}_{||\cdot||_+}(\mu, \sigma^2)$, as one with density $g(z) = C(\sigma, d) \cdot e^{-\frac{||z-\mu||_+^2}{2\sigma^2}}$, where $C(\sigma, d) = [\text{Area}(\{||x||_+ = 1\})\frac{(2\sigma^2)^{d/2}}{2}\Gamma(\frac{d}{2})]^{-1}$, and the Area is the $d-1$ dimensional surface measure on $\mathbb{R}^d$.

Lemma 3 (Generalized Gaussian mechanism Bassily et al. [2021])**.** Given a dataset $D \in \mathcal{X}^n$, and a query $q : \mathcal{X}^n \to \mathbf{E}$ with bounded $||\cdot||_*$-sensitivity: $s = \sup_{D\sim D'} ||q(D) - q(D^{'})||_*$, the Generalized Gaussian mechanism is defined as $q(D) + \xi$ where $\xi \sim \mathcal{GG}_{||\cdot||_+}(0, \frac{2\kappa \log(1/\delta)s^2}{\epsilon^2})$. The Generalized Gaussian mechanism preserves (ϵ, δ)-DP.

Lemma 4 (Prop 4.2 in [Bassily et al., 2021])**.** For any $m \geq 1$, if $z \sim \mathcal{GG}_{||\cdot||_+}(0, \sigma^2)$, then $\mathbb{E}[\|z\|_+^m] \leq (2\sigma^2)^{\frac{m}{2}}\Gamma(\frac{m+d}{2})/\Gamma(\frac{d}{2})$. Specifically, $\mathbb{E}[\|z\|_*^2] \leq \mathbb{E}[\|z\|_+^2] \leq d\sigma^2$, where $\Gamma(\cdot)$ is the Gamma function.

In the following, we recall some terminologies on the properties of the loss function and the constraint set $\mathcal{C}$.

Definition 5. (L-Lipschitz) Given the loss function $\ell(\cdot,\cdot) : \mathcal{C} \times \mathcal{X} \to \mathbb{R}$. It is L-Lipschitz w.r.t. the norm $||\cdot||$ if for all $x \in \mathcal{X}$ and $w_1, w_2 \in \mathcal{C}$ we have

$$|\ell(w_1, x) - \ell(w_2, x)| \leq L \cdot ||w_1 - w_2||.$$

Definition 6. (β-Smooth) Given the loss function $\ell(\cdot,\cdot) : \mathcal{C} \times \mathcal{X} \to \mathbb{R}$. It is β-smooth w.r.t. the norm $||\cdot||$ if its gradient is β-Lipschitz w.r.t. $||\cdot||$, namely, for all $x \in \mathcal{X}$ and $w_1, w_2 \in \mathcal{C}$ we have

$$||\nabla\ell(w_1, x) - \nabla\ell(w_2, x)||_* \leq \beta \cdot ||w_1 - w_2||.$$

Definition 7. (Strongly convex) Given the loss function $\ell(\cdot,\cdot) : \mathcal{C} \times \mathcal{X} \to \mathbb{R}$, it is α-strongly convex w.r.t. the norm $||\cdot||$ if for all $x \in \mathcal{X}$ and $w_1, w_2 \in \mathcal{C}$,

$$\langle \nabla\ell(w_1, x) - \nabla\ell(w_2, x), w_1 - w_2\rangle \geq \alpha \cdot ||w_1 - w_2||^2.$$

Definition 8. (Bregman divergence) For a convex function $\Phi : \mathbf{E} \to \mathbb{R}$, the Bregman divergence is defined as

$$D_\Phi(y, x) = \Phi(y) - \Phi(x) - \langle \nabla\Phi(x), y - x\rangle.$$

Notice that the Bregman divergence is always positive, and it is convex in the first argument.

Definition 9. (Relative strongly convex [Lu et al., 2018]) A function $f : \mathbf{E} \to \mathbb{R}$ is α-strongly convex **relative** to $\Phi : \mathbf{E} \to \mathbb{R}$ if for all $x, y \in \mathbf{E}$,

$$f(x) + \langle \nabla f(x), y - x\rangle + \alpha D_\Phi(y, x) \leq f(y).$$

Definition 10. (Relative smooth [Lu et al., 2018]) A function $f : \mathbf{E} \to \mathbb{R}$ is β-smooth **relative** to $\Phi : \mathbf{E} \to \mathbb{R}$ if $\forall x, y \in \mathbf{E}$, $f(x) + \langle \nabla f(x), y - x\rangle + \beta D_\Phi(y, x) \geq f(y)$.

Next, we introduce some basic concepts on Minkowski norm of a symmetric, closed, and convex set $\mathcal{C}$.

Definition 11 (Minkowski norm)**.** For a centrally symmetric convex set $\mathcal{C} \subseteq \mathbb{R}^d$, the Minkowski norm (denoted by $||\cdot||_\mathcal{C}$) is defined as follows. For any vector $v \in \mathbb{R}^d$,

$$||\cdot||_\mathcal{C} = \min\{r \in \mathbb{R}^+ : v \in r\mathcal{C}\}.$$

The dual norm of $||\cdot||_\mathcal{C}$ is denoted as $||\cdot||_{\mathcal{C}^*}$, and for any vector $v \in \mathbb{R}^d$, $||v||_{\mathcal{C}^*} = \max_{w\in\mathcal{C}} |\langle w, v\rangle|$. Note that by Holder's inequality, for any pair of dual norms $||\cdot||$ and $||\cdot||_*$, and any $x, y \in \mathbb{R}^d$, $|\langle x, y\rangle| \leq ||x|| \cdot ||y||_*$. So we have $|\langle x, y\rangle| \leq ||x||_\mathcal{C} \cdot ||y||_{\mathcal{C}^*}$.

In the constrained Euclidean case, our work relies on appropriately quantifying the size of a convex body, which leads to the following definition of Gaussian width.

Definition 12. (Gaussian width) Let $\xi \sim \mathcal{N}(0, \mathbb{I}_d)$ be a Gaussian random vector in $\mathbb{R}^d$, for a set $\mathcal{C}$, the Gaussian width is defined as $G_\mathcal{C} = \mathbb{E}_\xi[\sup_{w\in\mathcal{C}} \langle \xi, w\rangle]$.

Compared to dimension d, the Gaussian width of a convex set $\mathcal{C} \subset \mathbb{R}^d$ could be much smaller. For example, when $\mathcal{C}$ is the unit ℓ_1-norm ball, $G_\mathcal{C} = O(\sqrt{\log d})$; and when $\mathcal{C}$ is the set of of all unit s-sparse vectors on $\mathbb{R}^d$, $G_\mathcal{C} = O(\sqrt{s \log \frac{d}{s}})$. We refer readers to Talwar et al. [2014] for details.

4 DP-SCO IN EUCLIDEAN SPACE

In this section, we focus on the Euclidean case with a closed, bounded, and convex constraint set $\mathcal{C}$, and the loss function could be either convex or strongly convex.

4.1 GENERAL CONVEX CASE

Before showing our idea, we need to discuss the weakness of previous approaches. Note that since our goal is getting an upper bound that depends on the Gaussian width of the constrained set $\mathcal{C}$, we will not discuss the approaches that achieve upper bounds that are polynomial in d.

In general, all methods can be categorized into two classes: gradient perturbation and objective function perturbation. In gradient perturbation methods [Talwar et al., 2014], the key idea is modifying the Mirror Descent by adding noise to gradients. While this approach could achieve satisfactory bounds for the empirical risk [Wang et al., 2017, Wang and Xu, 2019], however, when considering the population risk we need to use batched gradients at each iteration, which will induce a sub-optimal rate [Amid et al., 2022]. Instead of perturbing the gradient, Talwar et al. [2014] show that the objective function perturbation method in Chaudhuri et al. [2011] could also achieve an upper bound that only depends on the Gaussian width, instead of d. However, this approach has two weaknesses: First, Talwar et al. [2014] only shows the bound for the empirical risk, and whether its excess population risk is satisfactory or not is unknown; Secondly, it is well-known that the objective perturbation approach needs to exactly get the minimizer of the perturbed objective function, which is inefficient in practice.

Motivated by the objective perturbation method in Talwar et al. [2014], our algorithm is an approximate version proposed in Bassily et al. [2019]. See detailed procedures in

Algorithm 1. In detail, first, similar to the standard objective perturbation, we add a random and linear term $\frac{\langle \mathbf{G}, \theta \rangle}{n}$ with Gaussian noise $\mathbf{G}$ and an ℓ_2 regularization to the original empirical risk function to obtain a new objective function $\mathcal{J}(\theta, D)$. Then we obtain an approximate minimizer θ_2 of the perturbed empirical risk $\mathcal{J}(\theta, D)$ via any efficient optimization method (such as proximal SVRG Xiao and Zhang [2014] or projected SGD) to ensure that $\mathcal{J}(\theta_2, D) - \min_{\theta \in \mathcal{C}} \mathcal{J}(\theta, D)$ is at most α. Formally, we can define such an optimization method as an optimizer function $\mathcal{O} : \mathcal{F} \times [0, 1] \to \mathcal{C}$, where $\mathcal{F}$ is the class of objectives and the other argument is the optimization accuracy. Finally, we perturb θ_2 with Gaussian noise to fuzz the difference between θ_2 and the true minimizer, we then project the perturbed θ_2 onto set $\mathcal{C}$.

Since the algorithm itself is not new, here we highlight our contributions: First, with some specific parameters, we show such an algorithm could achieve an excess population risk of $O(\frac{G_{\mathcal{C}}}{n\epsilon} + \frac{1}{\sqrt{n}})$, while Bassily et al. [2019] only show an upper bound of $O(\frac{\sqrt{d}}{n\epsilon} + \frac{1}{\sqrt{n}})$; Second, we extend the algorithm to the strongly convex case (see Section 4.2 for details). In the following, we will show the theoretical guarantees of our algorithm. First, we need the following assumption on the loss function ℓ.

Assumption 1. The loss function ℓ is twice differentiable, L-Lipschitz and β-smooth w.r.t. the Euclidean norm $||\cdot||_2$ over $\mathcal{C}$.

Algorithm 1 $\mathcal{A}_{\text{App-ObjP}}$: Approximate Objective perturbation

1: **Input:** Dateset D, loss function ℓ, regularization parameter λ, optimizer $\mathcal{O} : \mathcal{F} \times [0, 1] \to \mathcal{C}$, where $\mathcal{F}$ is the class of objectives, and the other argument is the optimization accuracy. $\alpha \in [0, 1]$: optimization accuracy.

2: Sample $\mathbf{G} \sim \mathcal{N}(0, \sigma_1^2 \mathbb{I}_d)$ where $\sigma_1^2 = \frac{128 L^2 \log(2.5/\delta)}{\epsilon^2}$. Set $\lambda \geq \frac{r\beta}{\epsilon n}$, where $r = \min\{d, 2 \cdot \text{rank}(\nabla^2 \ell(\theta, x))\}$ with $\text{rank}(\nabla^2 \ell(\theta, x))$ being the maximal rank of the Hessian of ℓ for all $\theta \in \mathcal{C}$ and $x \sim \mathcal{P}$.

3: Let $\mathcal{J}(\theta, D) = \hat{\mathcal{L}}(\theta, D) + \frac{\langle \mathbf{G}, \theta \rangle}{n} + \lambda ||\theta||_2^2$.

4: **return** $\hat{\theta} = \text{Proj}_{\mathcal{C}}[\mathcal{O}(\mathcal{J}, \alpha) + \mathbf{H}]$ where $\mathbf{H} \sim \mathcal{N}(0, \sigma_2^2 \mathbb{I}_d)$ with $\sigma_2^2 = \frac{64\alpha \log(2.5/\delta)}{\lambda \epsilon^2}$

Theorem 1. Suppose that Assumption 1 holds and that the smoothness parameter β satisfies $\beta \leq \frac{\epsilon n \lambda}{r}$. Then for any $0 < \epsilon, \delta < 1$, $\mathcal{A}_{\text{App-ObjP}}$ (Algorithm 1) is (ϵ, δ)-DP.

It is notable that although we need to assume β is not large enough, as we will see in Theorem 2, the assumption will always hold when n is sufficiently large.

Theorem 2. Suppose that Assumption 1 holds. When n is large enough such that $n \geq \frac{r^2 \beta^2 ||\mathcal{C}||_2^2}{\epsilon^2 L^2}$ and $n \geq O\left(\frac{\sqrt{d \log(1/\delta)}}{\epsilon}\right)$, take $\lambda = \frac{L}{\sqrt{n} ||\mathcal{C}||_2}$ and $\alpha \leq \min\left\{\frac{L||\mathcal{C}||_2}{n^{\frac{3}{2}}}, \frac{\epsilon^2 L ||\mathcal{C}||_2^3}{G_{\mathcal{C}}^2 \log(1/\delta) n^{\frac{5}{2}}}\right\}$ in Algorithm 1, we have

$$\mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathcal{L}(\theta^*) \leq O\left(\frac{L \cdot G_{\mathcal{C}} \sqrt{\log(1/\delta)}}{\epsilon n} + \frac{L||\mathcal{C}||_2}{\sqrt{n}}\right),$$

where the expectation is taken over the internal randomness of the algorithm.

Remark 1. While we consider the same algorithm as in Bassily et al. [2019], there are several crucial differences. First, to achieve the upper bound of $O(\frac{\sqrt{d}}{n\epsilon} + \frac{1}{\sqrt{n}})$, Bassily et al. [2019] only need to set $\alpha \leq O(\frac{1}{n^2} \max\{\frac{1}{\sqrt{n}}, \frac{d}{n\epsilon}\})$ while we need to be more aggressive by choosing $\alpha \leq O(\epsilon^2 n^{-\frac{5}{2}})$. This is reasonable as we aim to get an improved upper bound. Thus we have to get a more accurate estimation. Secondly, besides enforcing the perturbed approximation to lie in the set $\mathcal{C}$ as it does in Bassily et al. [2019], the projection operator in Step 4 of Algorithm 1 plays a more critical role in achieving a bound that depends on $G_{\mathcal{C}}$ in our analysis, i.e., the bound in Bassily et al. [2019] will still hold even there is no projection step, while this is not true for our case. Specifically, although the noise $\mathbf{H}$ is a d-dimensional Gaussian noise, we can show that due to the projection operator, the error introduced by the noise depends only on $G_{\mathcal{C}}$ rather than $\sqrt{d}$, i.e., $||\hat{\theta} - \theta_2||_2^2 \leq O(\sqrt{\frac{\alpha \log(1/\delta)}{\lambda}} \cdot \frac{G_{\mathcal{C}}}{\epsilon})$. A similar idea has also been used in privately answering multiple linear queries [Nikolov et al., 2016].

4.2 STRONGLY CONVEX CASE

We aim to extend our above idea to the strongly convex case. First, we impose the following assumption.

Assumption 2. We assume the loss is twice differentiable, L-Lipschitz and β-smooth w.r.t. $||\cdot||_2$, and it is Δ-strongly convex w.r.t. $||\cdot||_{\mathcal{C}}$ over the set $\mathcal{C}$.

Note that we can relax the assumption to strongly convex w.r.t $\|\cdot\|_2$ as $\|v\|_2 \geq \mathcal{C}_{\min} \cdot \|v\|_{\mathcal{C}}$, where $\mathcal{C}_{\min}$ is in Theorem 5. See the proof of Theorem 5 for details.

Our method is shown in Algorithm 2. Note that, compared with Algorithm 1, the main difference is the regularization parameter λ. This is because the loss function is already Δ-strongly convex, thus smaller λ will be sufficient to make $\mathcal{J}$ to be $\frac{r\beta}{\epsilon n}$-strongly convex. Moreover, when n is large enough, we can see $\lambda = 0$, indicating that we can get an improved excess population risk compared to the convex case.

Theorem 3. If the loss function satisfies Assumption 2. Then for any $0 < \epsilon, \delta < 1$, $\mathcal{A}_{\text{App-ObjP-SC}}$ (Algorithm 2) is (ϵ, δ)-DP.

Algorithm 2 $\mathcal{A}_{\text{App-ObjP-SC}}$: Approximate Objective perturbation for strongly convex function

1: **Input:** Dateset D, loss function ℓ, regularization parameter λ, optimizer $\mathcal{O} : \mathcal{F} \times [0,1] \to \mathcal{C}$, where $\mathcal{F}$ is the class of objectives and the other argument is the optimization accuracy. $\alpha \in [0,1]$: optimization accuracy.

2: Sample $\mathbf{G} \sim \mathcal{N}(0, \sigma_1^2 \mathbb{I}_d)$ where $\sigma_1^2 = \frac{128L^2 \log(2.5/\delta)}{\epsilon^2}$. Set $\lambda = \max\left\{\frac{r\beta}{\epsilon n} - \Delta, 0\right\}$, where $r = \min\{d, 2 \cdot \text{rank}(\nabla^2 \ell(\theta, x))\}$ with $\text{rank}(\nabla^2 \ell(\theta, x))$ being the maximal rank of the Hessian of ℓ for all $\theta \in \mathcal{C}$ and $x \sim \mathcal{P}$.

3: Let $\mathcal{J}(\theta, D) = \hat{\mathcal{L}}(\theta, D) + \frac{\langle \mathbf{G}, \theta \rangle}{n} + \lambda ||\theta||_2^2$.

4: **return** $\hat{\theta} = \text{Proj}_{\mathcal{C}}[\mathcal{O}(\mathcal{J}, \alpha) + \mathbf{H}]$ where $\mathbf{H} \sim \mathcal{N}(0, \sigma_2^2 \mathbb{I}_d)$ with $\sigma_2^2 = \frac{64\alpha \log(2.5/\delta) \cdot ||\mathcal{C}||_2^2}{\Delta \epsilon^2}$

Theorem 4. Suppose that Assumption 2 holds. If n is large enough such that $n \geq O(\max\{\frac{L^2||\mathcal{C}||_2^2}{\Delta^2}, \frac{||\mathcal{C}||_2^2 r^2 \beta^2}{L^2 \epsilon^2}\})$ and $n \geq O\left(\frac{\sqrt{d \log(1/\delta)}}{\epsilon}\right)$, then by setting $\alpha \leq O\left(\min\left\{\frac{L^2||\mathcal{C}||_2^2}{\Delta n^2}, \frac{L^4 \cdot ||\mathcal{C}||_2^6 \epsilon^2}{\Delta^3 n^4 G_{\mathcal{C}}^2 \log(1/\delta)}\right\}\right)$, we have

$$\mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathcal{L}(\theta^*) \leq O\left(\frac{L^2||\mathcal{C}||_2^2}{\Delta n \epsilon} + \frac{G_{\mathcal{C}}^2 L^2 \log(1/\delta)}{\Delta n^2 \epsilon^2}\right),$$

where the expectation is taken over the internal randomness of the algorithm.

Remark 2. First, it is notable that an objective perturbation method for strongly convex loss has also been presented by Talwar et al. [2014]. However, there are two major differences: (1) the method in Talwar et al. [2014] needs to solve the perturbed objective function exactly, indicating it is inefficient; (2) Talwar et al. [2014] only provide the excess empirical risk. It is unknown whether their method could achieve the same bound as ours for the excess population risk. Secondly, when $\mathcal{C}$ is an ℓ_2-norm ball, the bounds in Theorem 2 and Theorem 4 will recover the optimal rate of DP-SCO over ℓ_2-norm ball for convex and strongly convex loss functions, respectively [Bassily et al., 2019]. Thirdly, the terms of $O(\frac{G_{\mathcal{C}}}{n\epsilon})$ and $O(\frac{G_{\mathcal{C}}^2}{n^2\epsilon^2})$ match the best-known results of excess empirical risk for the convex and strongly convex case, respectively [Talwar et al., 2014].

In Remark 2, we showed that our results are optimal when $\mathcal{C}$ is an ℓ_2-norm ball and are comparable to the best results of DP-ERM with Gaussian width. A natural question is whether we can further improve these two upper bounds. In the following, we partially answer the question by providing a lower bound for strongly convex loss functions.

Theorem 5. Let $\mathcal{C}$ be a symmetric body contained in the unit Euclidean ball $\mathcal{B}_2^d$ in $\mathbb{R}^d$ and satisfies $\|\mathcal{C}\|_2 = 1$. For any $n = O(\frac{\sqrt{d \log(1/\delta)}}{\epsilon})$, $\epsilon = O(1)$ and $2^{-\Omega(n)} \leq \delta \leq 1/n^{1+\Omega(1)}$, there exists a loss ℓ which is 1-Lipschitz w.r.t. $\|\cdot\|_2$ and $\mathcal{C}_{\min}^2$-strongly convex w.r.t. $\|\cdot\|_{\mathcal{C}}$, and a dataset $D = \{x_1, \cdots, x_n\} \subseteq \mathcal{C}^n$ such as for any (ϵ, δ)-differentially private algorithm on minimizing the empirical risk function $\hat{\mathcal{L}}(\theta, D)$ over $\mathcal{C}$, its output $\theta^{priv} \in \mathcal{C}$ satisfies

$$\mathbb{E}[\mathcal{L}(\theta^{priv})] - \mathcal{L}(\theta^*) = \Omega\left(\max\left\{\frac{G_{\mathcal{C}}^2 \log(1/\delta)}{(\log(2d))^4 \epsilon^2 n^2}, \frac{1}{n}\right\}\right),$$

where the expectation is taken over the internal randomness of the algorithm $\mathcal{A}$. Here $\mathcal{C}_{\min} = \min\{\|v\|_2 : v \in \partial\mathcal{C}\}$ with $\partial\mathcal{C}$ as the boundary of the set $\mathcal{C}$, i.e., it is the distance between the original point to the boundary of $\mathcal{C}$.

Taking $\Delta = \mathcal{C}_{\min}^2$ and $L = 1$ in Theorem 4, we can see the rate of excess population risk in Theorem 4 for strongly convex loss functions is nearly optimal by a factor of $\tilde{O}(\mathcal{C}_{\min}^{-2})$. It is unknown whether we can further close the gap, and we will leave it as an open problem.

5 DP-SCO IN ℓ_p^d SPACE

In this section, we will focus on DP-SCO in ℓ_p^d space where $1 < p \leq \infty$. As we mentioned in the Introduction section, we study two settings: (1) $\mathcal{C}$ is $\mathbb{R}^d$ and the gradient of the loss function is bounded (i.e., the loss is Lipschitz); (2) $\mathcal{C}$ is bounded, and the distribution of gradient of the loss is heavy-tailed. Similar to the previous study in Bassily et al. [2021], for each setting, there are two cases: $1 < p < 2$ and $2 \leq p \leq \infty$. Notice that, unlike the previous section, we only study the case where the loss functions are convex. The reason is that except for the Euclidean space, for a strongly convex function, the ratio between its smoothness and strong convexity, i.e., the condition number, will depend on the dimension of $\mathbf{E}$. For example, in the ℓ_1^d space, it has been shown that there is no function whose condition number is less than d [Juditsky and Nesterov, 2014].

5.1 UNCONSTRAINED CASE

In this part, we will study Lipschitz loss under the following assumption that is commonly used in the related work on general stochastic convex optimization.

Assumption 3. We assume $\ell(\cdot, x)$ is convex, β-smooth and L-Lipschitz w.r.t. $||\cdot||$ over $\mathbb{R}^d$.

Due to its difficulty, we first consider the case where $1 < p < 2$. See Algorithm 3 for details. Note that Algorithm 3 could be considered as a noisy and regularized version of the standard mirror descent, i.e., at each iteration, we first perform linearization of $\hat{\mathcal{L}}(w_t, D)$, then we add a generalized Gaussian noise to its gradient to privatize the algorithm, a Bregman divergence term and a regularized term $\alpha\Phi(\cdot)$ with some specific α to the linearization term. Then we solve the perturbed and regularized optimization problem. We output a linear combination of the intermediate parameters as the final output.

It is notable that although our method is a noisy modification of Mirror Descent, it is completely different from the previous private Mirror Descent based methods in Talwar et al. [2014], Wang et al. [2017], Bassily et al. [2021], Amid et al. [2022]: First, instead of directly adding noise to the gradient in standard Mirror Descent, here we have an additional regularization term, which is crucial for us to make the algorithm stable, indicating that we can get an excess population risk. To be more specific, first, by the definition of $\|\cdot\|_+$, and the duality between strong convexity and smoothness, we can easily see Φ is 1-strongly convex w.r.t $\|\cdot\|$. This indicates that the function $\hat{\mathcal{L}}(w,D)+\alpha\Phi(w)$ is relatively strongly convex and smooth (note that it is not smooth as the regularization term is not smooth when $1<p<2$). And the update step is just a noisy version of Mirror Descent for $\hat{\mathcal{L}}(w,D)+\alpha\Phi(w)$. Recently, it has been shown that Mirror Descent is stable for relatively strongly convex and smooth functions. Thus, we can also show that Algorithm 3 is stable, indicating that we can get an excess population risk. From the above intuition, we can also see the parameter α need to be carefully tuned to balance the stability and the excess empirical risk. The second difference is that, instead of using the last iterate or the average of iterates, our output is a linear combination of intermediate iterates, which is due to the noise we added. In the following we show the main results.

Theorem 6. For the ℓ_p^d space with $1<p<2$, suppose Assumption 3 holds, then for any $0<\epsilon,\delta<1$, Algorithm 3 is (ϵ,δ)-DP.

Theorem 7. For the ℓ_p^d space with $1<p<2$, suppose Assumption 3 holds. In Algorithm 3, take $\alpha=\frac{4\beta}{T}\log_2\frac{n}{T}$ and $T=O((\frac{n\epsilon\kappa}{\sqrt{d\log(1/\delta)}})^{\frac{2}{5}})$, assume n is sufficiently large such that $n\geq O\left(\frac{\epsilon^4}{(d\log(1/\delta))^2\kappa^{1/2}}\right)$, then we have

$$\mathbb{E}[\mathcal{L}(\hat{w})]-\mathcal{L}(\theta^*)\leq\tilde{O}(\kappa^{\frac{4}{5}}\cdot(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{2}{5}}),$$

where $\tilde{O}$ hides β, L and a factor of $\mathbb{E}_D[\tilde{C}_D^2]$ with $\tilde{C}_D^2=\|\tilde{w}^*\|_{\kappa_+}^2\leq\|\tilde{w}^*\|^2$ and $\tilde{w}^*=\arg\min_{w\in\mathbf{E}}\hat{\mathcal{L}}(w,D)$).

The key idea to prove Theorem 7 is to show that Algorithm 3 is uniformly stable (w.r.t $\|\cdot\|$) by bounding the term $\mathbb{E}[\|w_{t+1}-w'_{t+1}\|]$, where w'_{t+1} is the corresponding iterate of the algorithm when the input data is D', which is a neighboring data of D. To show this, rather than analyzing the stability of w_{t+1} directly via the approach in Hardt et al. [2016], our strategy is bounding $\|w_{t+1}-w_\alpha^*\|$, where $w_\alpha^*=\arg\min\hat{\mathcal{L}}(w,D)+\alpha\Phi(w)$. As the regularized function $\hat{\mathcal{L}}(w,D)+\alpha\Phi(w)$ now is relatively smooth and convex, the stability of w_α^* is $O(\frac{1}{n})$. Thus we can get the sensitivity of w_{t+1}. Then we can bound the sensitivity of $\hat{w}$.

Remark 3. In the constrained case, Bassily et al. [2021] show that it is possible to achieve an upper bound of $\tilde{O}((M+M^2)(\frac{\sqrt{\kappa}}{\sqrt{n}}+\frac{\kappa\sqrt{d\log 1/\delta}}{n\epsilon}))$, where M is the diameter of set $\mathcal{C}$. Thus, we can see there is still a gap between the unconstrained case and the constrained case.

Algorithm 3 Noisy Regularized Mirror Descent for ℓ_p^d $(1<p<2)$.

1: **Input:** Dateset D, loss function ℓ, smoothness parameter β and parameter α.
2: Take $w_1=0$.
3: **for** $t=1,\cdots,T$ **do**
4: Solve the following optimization problem

$$w_{t+1}=\arg\min_{w\in\mathbf{E}}\{\langle\nabla\hat{\mathcal{L}}(w_t,D)+g_t,w-w_t\rangle+\beta\cdot D_\Phi(w,w_t)+\alpha\Phi(w)\},\quad(1)$$

where $g_t\sim\mathcal{GG}_{||\cdot||_+}(0,\sigma^2)$ with $\sigma^2=\frac{64L^2\kappa T\log(1/\delta)}{n^2\epsilon^2}$ and $||\cdot||_+$ is the smooth norm for $(\mathbf{E},||\cdot||_*)$. $\kappa=\min\{\frac{1}{p-1},2\log d\}$ and $\Phi(x)=\frac{\kappa}{2}||x||_{\kappa_+}^2$ with $\kappa_+=\frac{\kappa}{\kappa-1}$.
5: **end for**
6: **return** $\hat{w}=\frac{\sum_{t=1}^T\left(\frac{2\beta+\alpha}{2\beta}\right)^t\cdot w_{t+1}}{\sum_{t=1}^T\left(\frac{2\beta+\alpha}{2\beta}\right)^t}$.

Next, we study the case where $2\leq p\leq\infty$. The key idea is to reduce the ℓ_p^d space to the Euclidean space by leveraging the relationship between the ℓ_p norm and the Euclidean norm. Thus, here we adopt the Phased DP-SGD algorithm proposed by Feldman et al. [2020]. As the parameters in the original Phased DP-SGD depend on the diameter, we modify them to the unconstrained case. Specifically, we have the following result.

Theorem 8. For the ℓ_p^d space with $2\leq p\leq\infty$, suppose Assumption 3 holds. Then for any $0<\epsilon,\delta<1$, there is an (ϵ,δ)-DP algorithm whose output θ satisfies

$$\mathbb{E}[\mathcal{L}(\theta)]-\mathcal{L}(\theta^*)\leq O(d^{1-\frac{2}{p}}\|\theta^*\|^2(\frac{1}{\sqrt{n}}+\frac{\sqrt{d\log(1/\delta)}}{\epsilon n})).$$

In the constrained case, Bassily et al. [2021] shows the optimal rate of $O(Md^{\frac{1}{2}-\frac{1}{p}}(\frac{1}{\sqrt{n}}+\frac{\sqrt{d\log 1/\delta}}{n\epsilon}))$, where M is the diameter of the set $\mathcal{C}$ w.r.t. $\|\cdot\|$. Thus, we can see there is a difference of $O(d^{\frac{1}{2}-\frac{1}{p}})$. This is because, rather than linear in M in the constrained case, in the Euclidean and unconstrained case, we can show the excess population risk depends on $\|\theta^*\|_2^2$, which is less than $d^{1-\frac{2}{p}}\|\theta^*\|^2$.

5.2 HEAVY-TAILED AND CONSTRAINED CASE

In the above section, we studied DP-SCO with Lipschitz loss functions, i.e., the $\|\cdot\|_*$ norm of the loss gradient is uniformly bounded by L. Next, we will relax this assumption to a heavy-tailed distribution, i.e., we only assume the variance of the loss gradient w.r.t $\|\cdot\|_*$ is finite. As we have

discussed the difficulty of the unconstrained case compared to the constrained case, throughout the section, we focus on the constrained case with the $\|\cdot\|$-norm diameter M.

Assumption 4. We assume $\ell(\cdot, x)$ is convex and β-smooth $||\cdot||$ over $\mathcal{C}$. Moreover, for all $w \in \mathcal{C}$ there exists a known constant $\sigma > 0$ such that $\mathbb{E}[||\nabla\ell(w,x) - \nabla\mathcal{L}(w)||_*^2] \leq \sigma^2$.

It is noteworthy that the heavy-tailedness assumption is commonly used in previous related work, such as Vural et al. [2022]. Besides the norm of gradient, there is another line of work that only assumes the second-order moment of each coordinate of the gradient is bounded [Hu et al., 2022, Kamath et al., 2022, Wang et al., 2020, Wang and Xu, 2022, Tao et al., 2022]. We leave such a relaxed assumption as future work.

Like the previous section, we first study the case where $1 < p < 2$. We present our algorithm in Algorithm 4, which could be considered a shuffled, truncated, and noisy version of one-pass Mirror Descent. Specifically, in the first step, we shuffle the dataset and divide it into several batches (we will use one batch for one iteration). Using the by-now standard method of privacy amplification by shuffling [Feldman et al., 2022], we can amplify the overall privacy guarantee by a factor of $\tilde{O}(\frac{1}{n})$ as compared to the analysis for the unshuffled dataset. Next, motivated by Nazin et al. [2019], at each iteration, we first conduct a truncation step to each sample gradient $\nabla\ell(w_{t-1}, x)$. Such an operator can not only remove outliers, but also upper bound the $\|\cdot\|_*$-sensitivity of the truncated gradients to $O(\beta M + \lambda)$. Then we perform the Mirror Descent update by these perturbed and truncated sample gradients. In the following, we show the privacy and utility guarantees of our algorithm.

Theorem 9. For the ℓ_p^d space with $1 < p < 2$, suppose Assumption 4 holds. Algorithm 4 is (ϵ, δ)-DP if $\epsilon = O(\sqrt{\frac{\log(n/\delta)}{n}})$ and $0 < \delta < 1$.

Theorem 10. For the ℓ_p^d space with $1 < p < 2$, suppose Assumption 4 holds and assume n is sufficiently large such that $n \geq O(\frac{\max\{\beta^2, 1\} M^2 \sqrt{d\kappa^2 \log(1/\delta)}}{\epsilon})$. Given a failure probability $\delta' > 0$, in Algorithm 4, take $T = O(\frac{M^2 n^2 \epsilon^2}{\lambda^2 d \log(1/\delta)})$, $\{\gamma\}_{t=1}^T = \bar{\gamma} = \sqrt{T}$, and $\lambda = O(\frac{\sqrt{n\epsilon}}{\sqrt[4]{\kappa^2 d \log(1/\delta)}})$, then the output $\hat{w}$ satisfies the following with probability $1 - \delta'$

$$\mathbb{E}[\mathcal{L}(\hat{w})] - \mathcal{L}(w^*) \leq \tilde{O}(\frac{M \sqrt[4]{\kappa^2 d \log(1/\delta)} \log(1/\delta')}{\sqrt{n\epsilon}}),$$

where the expectation is taken over the randomness of noise, and the probability is w.r.t. the dataset $D \sim \mathcal{D}^n$.

Remark 4. First, note that due to the privacy amplification, here the noise added to each sample gradient is $\tilde{O}(\frac{\beta M + \lambda}{\sqrt{n}\epsilon})$ rather than $\tilde{O}(\frac{\beta M + \lambda}{\epsilon})$ if without shuffling. Secondly, note that the truncation step is quite different from the previous work on DP-SCO with heavy-tailed data [Wang et al., 2020], i.e., we enforce the sample gradient to become zero if its norm exceeds the threshold. Finally, compared to the best-known result $O(\sqrt{\frac{\kappa}{n}})$ in the non-private and heavy-tailed case [Nazin et al., 2019] and the bound $\tilde{O}(\sqrt{\frac{\kappa}{n}} + \frac{\kappa\sqrt{d}}{n\epsilon})$ for private and Lipschitz case [Bassily et al., 2021], we can see there may exist a space to improve our bound further.

Algorithm 4 Shuffled Truncated DP Mirror Descent

Input: Dataset D, loss function ℓ, initial point $w_0 = 0$, smooth parameter β and λ.
Randomly permute the data and denote the permuted data as $\{x_1, \cdots, x_n\}$.
Divide the permuted data into T batches $\{B_i\}_{i=1}^T$ where $|B_i| = \frac{n}{T}$ for all $i = 1, \cdots, T$
for $t = 1, \cdots, T$ **do**
  for each $x \in B_t$ **do**
    $$g_x = \begin{cases} \nabla\ell(w_{t-1}, x) & \text{if } ||\nabla\ell(w_{t-1}, x)||_* \leq \beta M + \lambda \\ 0 & \text{otherwise} \end{cases}$$
  end for
  Update
$$w_t = \arg\min_{w \in \mathcal{C}} \left\{ \left\langle \frac{\sum_{x \in B_t} g_x + Z_x^t}{|B_t|}, w \right\rangle + \gamma_t \cdot D_\Phi(w, w_{t-1}) \right\},$$
  where $Z_x^t \sim \mathcal{GG}_{||\cdot||_+}(\sigma_1^2)$ with $\sigma_1^2 = O\left(\frac{\log(\frac{n}{\delta}) \cdot \kappa(\beta M + \lambda)^2 \cdot \log(1/\delta)}{n\epsilon^2}\right)$, $||\cdot||_+$ is the smooth norm for $(\mathbf{E}, ||\cdot||_*)$. $\kappa = \min\{\frac{1}{p-1}, 2\log d\}$ and $\Phi(x) = \frac{\kappa}{2}||x||_{\kappa_+}^2$ with $\kappa_+ = \frac{\kappa}{\kappa - 1}$.
end for
return $\hat{w} = (\sum_{t=1}^T \gamma_t^{-1})^{-1} \cdot \sum_{t=1}^T \gamma_t^{-1} w_t$

There are two limitations in Theorem 10. First, Algorithm 4 is (ϵ, δ) only for $\epsilon = \tilde{O}(n^{-\frac{1}{2}})$, which cannot be generalized to mid or low privacy regime. Secondly, Theorem 10 only holds for the case $1 < p < 2$. To address the first issue, we can slightly modify the algorithm by using batched Mirror Descent without shuffling, while we will get a worse upper bound. For the second one, similar to Theorem 8, we can reduce the problem to the Euclidean case. Informally, we have the following two results.

Theorem 11 (Informal)**.** For the ℓ_p^d space with $1 < p < 2$, suppose Assumption 4 holds. For all $0 < \epsilon, \delta < 1$, there is an (ϵ, δ)-DP algorithm whose output could achieve an excess population risk of $O\left(M^{\frac{4}{3}} \frac{\kappa^{\frac{2}{3}} (d \log(1/\delta))^{\frac{1}{6}}}{(n\epsilon)^{\frac{1}{3}}}\right)$.

Theorem 12 (Informal)**.** For the ℓ_p^d space with $2 \leq p \leq \infty$, suppose Assumption 4 holds (and with some additional mild assumptions). For all $0 < \epsilon, \delta < 1$, there is an (ϵ, δ)-DP algorithm whose output could achieve an expected excess population risk of $O\left(\frac{d^{\frac{3}{2} - \frac{1}{p}}}{\sqrt{n}} + \frac{d^{\frac{3}{2} - \frac{1}{2p}}}{\sqrt{n\epsilon}}\right)$.

References

Ehsan Amid, Arun Ganesh, Rajiv Mathews, Swaroop Ramaswamy, Shuang Song, Thomas Steinke, Vinith M Suriyakumar, Om Thakkar, and Abhradeep Thakurta. Public data-assisted mirror descent for private model training. In *International Conference on Machine Learning*, pages 517–535. PMLR, 2022.

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, 2021a.

Hilal Asi, Daniel Lévy, and John C Duchi. Adapting to function difficulty and growth conditions in private optimization. *Advances in Neural Information Processing Systems*, 34:19069–19081, 2021b.

Amit Attia and Tomer Koren. Uniform stability for first-order empirical risk minimization. In *Conference on Learning Theory*, pages 3313–3332. PMLR, 2022.

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

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

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

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

Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. *Advances in Neural Information Processing Systems*, 21, 2008.

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

Lutz Dümbgen, Sara A Van De Geer, Mark C Veraar, and Jon A Wellner. Nemirovski's inequalities revisited. *The American Mathematical Monthly*, 117(2):138–160, 2010.

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

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

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.

Vitaly Feldman, Audra McMillan, and Kunal Talwar. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 954–964. IEEE, 2022.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Yuxuan Han, Zhicong Liang, Zhipeng Liang, Yang Wang, Yuan Yao, and Jiheng Zhang. On private online convex optimization: Optimal algorithms in ℓ_p-geometry and high dimensional contextual bandits. *arXiv preprint arXiv:2206.08111*, 2022.

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.

Lijie Hu, Shuo Ni, Hanshen Xiao, and Di Wang. High dimensional differentially private stochastic optimization with heavy-tailed data. In *Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 227–236, 2022.

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.

Anatoli Juditsky and Arkadii S Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

Anatoli Juditsky and Yuri Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1): 44–80, 2014.

Gautam Kamath, Xingtu Liu, and Huanyu Zhang. Improved rates for differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pages 10633–10660. PMLR, 2022.

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.

Assimakis Kattis and Aleksandar Nikolov. Lower bounds for differential privacy from gaussian width. *arXiv preprint arXiv:1612.02914*, 2016.

Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

Alexander V Nazin, Arkadi S Nemirovsky, Alexandre B Tsybakov, and Anatoli B Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80:1607–1627, 2019.

Aleksandar Nikolov, Kunal Talwar, and Li Zhang. The geometry of differential privacy: The small database and approximate cases. *SIAM Journal on Computing*, 45(2): 575–616, 2016.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Jinyan Su, Lijie Hu, and Di Wang. Faster rates of private stochastic convex optimization. In *International Conference on Algorithmic Learning Theory*, pages 995–1002. PMLR, 2022.

Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry. *arXiv preprint arXiv:1411.5417*, 2014.

Youming Tao, Yulian Wu, Xiuzhen Cheng, and Di Wang. Private stochastic convex optimization and sparse learning with heavy-tailed data revisited. International Joint Conferences on Artificial Intelligence Organization, 2022.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2(3), 2008.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

Nuri Mert Vural, Lu Yu, Krishna Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In *Conference on Learning Theory*, pages 65–102. PMLR, 2022.

Di Wang and Jinhui Xu. Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1182–1189, 2019.

Di Wang and Jinhui Xu. Differentially private ℓ_1-norm linear regression with heavy-tailed data. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 1856–1861. IEEE, 2022.

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

Di Wang, Hanshen Xiao, Srinivas Devadas, and Jinhui Xu. On differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pages 10081–10091. PMLR, 2020.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.