Proceedings of Machine Learning Research vol 134:1–26, 2021 34th Annual Conference on Learning Theory

Non-Euclidean Differentially Private Stochastic Convex Optimization

Raef Bassily BASSILY.1@OSU.EDU
Department of Computer Science & Engineering
Translational Data Analytics Institute (TDAI)
The Ohio State University

Cristóbal Guzmán C.GUZMAN@UTWENTE.NL
Department of Applied Mathematics
University of Twente
Pontificia Universidad Católica de Chile

Anupama Nandi NANDI.10@OSU.EDU
Department of Computer Science & Engineering
The Ohio State University

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

Differentially private (DP) stochastic convex optimization (SCO) is a fundamental problem, where the goal is to approximately minimize the population risk with respect to a convex loss function, given a dataset of i.i.d. samples from a distribution, while satisfying differential privacy with respect to the dataset. Most of the existing works in the literature of private convex optimization focus on the Euclidean (i.e., ℓ_2) setting, where the loss is assumed to be Lipschitz (and possibly smooth) w.r.t. the ℓ_2 norm over a constraint set with bounded ℓ_2 diameter. Algorithms based on noisy stochastic gradient descent (SGD) are known to attain the optimal excess risk in this setting.

In this work, we conduct a systematic study of DP-SCO for ℓ_p-setups. For $p = 1$, under a standard smoothness assumption, we give a new algorithm with nearly optimal excess risk. This result also extends to general polyhedral norms and feasible sets. For $p \in (1, 2)$, we give two new algorithms, for which a central building block is a novel privacy mechanism, which generalizes the Gaussian mechanism. Moreover, we establish a lower bound on the excess risk for this range of p, showing a necessary dependence on $\sqrt{d}$, where d is the dimension of the space. Our lower bound implies a sudden transition of the excess risk at $p = 1$, where the dependence on d changes from logarithmic to polynomial, resolving an open question in prior work (Talwar et al., 2015) . For $p \in (2, \infty)$, noisy SGD attains optimal excess risk in the low-dimensional regime; in particular, this proves the optimality of noisy SGD for $p = \infty$. Our work draws upon concepts from the geometry of normed spaces, such as the notions of regularity, uniform convexity, and uniform smoothness.

Keywords: Differential privacy, stochastic convex optimization, non-Euclidean norms.

1. Introduction

Stochastic Convex Optimization (SCO) is one of the most fundamental problems in optimization, statistics, and machine learning. In this problem, the goal is to minimize the expectation of a convex loss w.r.t. a distribution given samples from that distribution. In particular, given n i.i.d. samples $z_1, \ldots, z_n$ from a distribution $\mathcal{D}$, we wish to output a solution $x \in \mathcal{X} \subseteq \mathbb{R}^d$ that minimizes the population loss $F_{\mathcal{D}}(x) \triangleq \mathbb{E}_{z\sim\mathcal{D}}[f(x, z)]$ for a convex function f over $\mathcal{X}$. A closely related problem is known as Empirical Risk Minimization (ERM), where the goal to minimize the empirical average of a loss with respect to a dataset S, i.e. output a solution that minimizes the empirical loss, $F_S(x) = \frac{1}{n}\sum_{z\in S} f(x,z)$, subject to $x \in \mathcal{X}$. In this work, we focus on ℓ_p-setups, where we consider the losses are Lipschitz and smooth w.r.t the ℓ_p norm over a constraint set with bounded ℓ_p diameter. (See Section 2 for a more formal description.)

There has been a long line of works that studied the differentially private analogs of these problems known as DP-SCO and DP-ERM, e.g., (Chaudhuri and Monteleoni, 2008; Chaudhuri et al., 2011; Kifer et al., 2012; Jain and Thakurta, 2014; Bassily et al., 2014a; Talwar et al., 2014; Wang et al., 2017; Bassily et al., 2019; Feldman et al., 2020). Nevertheless, the existing theory does not capture a satisfactory understanding of private convex optimization in non-Euclidean settings, and particularly with respect to general ℓ_p norms. Almost all previous works that studied the general formulations of DP-ERM and DP-SCO under general convex losses focused on the *Euclidean setting*, where both the constraint set and the subgradients of the loss are assumed to have a bounded ℓ_2-norm. In this setting, algorithms with optimal error rates are known for DP-ERM (Bassily et al., 2014a) and DP-SCO (Bassily et al., 2019; Feldman et al., 2020; Bassily et al., 2020). On the other hand, (Talwar et al., 2014, 2015) is the only work we are aware of that studied non-Euclidean settings under a fairly general setup in the context of DP-ERM (see "Other Related Work" section below for other works that studied special cases of this problem). However, this work does not address DP-SCO; moreover, for $p > 1$, it only provides upper bounds on the error rate for DP-ERM.

Without privacy constraints, convex optimization in these settings is fairly well-understood in the classical theory. In particular, there exists a universal algorithm that attains optimal rates for ERM as well as SCO over general ℓ_p spaces, namely, the stochastic mirror descent algorithm (Nemirovski and Yudin, 1983; Nemirovski et al., 2009). By contrast, the landscape of private convex optimization is quite unclear in these settings. Closing this gap in the existing theory is not of purely intellectual interest: the flexibility of non-Euclidean norms permits polynomial (in the dimension) acceleration for stochastic first-order methods (see, e.g., discussions in (Sra et al., 2011, Sec. 5.1.1)).

In this work, we focus on DP-SCO in non-Euclidean settings, particularly under the ℓ_p setups. Our work provides upper and lower bounds, which are either nearly optimal, or have small polynomial gaps (see Table 1 for a summary of our bounds). To achieve this, we develop several techniques for differential privacy in non-Euclidean optimization, which we believe could be also useful in other applications of differential privacy.

ℓ_p-setup	Upper Bound	Lower bound
$p = 1$	$\tilde{O}\left(\frac{\log(d)}{\varepsilon\sqrt{n}}\right)$ $(*)$	$\Omega\left(\sqrt{\frac{\log d}{n}}\right)$ (ABRW'12)
$1 < p < 2$	$\tilde{O}\left(\min\left\{\frac{\sqrt{\kappa}d^{1/4}}{\sqrt{n}}, \frac{\kappa}{\sqrt{n}} + \frac{\kappa\sqrt{d}}{\varepsilon n^{3/4}}\right\}\right)$	$\Omega\left(\max\left\{\frac{1}{\sqrt{n}}, \frac{(p-1)\sqrt{d}}{\varepsilon n}\right\}\right)$
$2 < p \leqslant \infty$	$\tilde{O}\left(\frac{d^{1/2-1/p}}{\sqrt{n}} + \frac{d^{1-1/p}}{\varepsilon n}\right)$	$\Omega\left(\min\left\{\frac{d^{1/2-1/p}}{\sqrt{n}}, \frac{1}{n^{1/p}}\right\}\right)$ (NY'83,ABRW'12)
$p = \infty$	$\tilde{O}\left(\sqrt{\frac{d}{n}}\right)$	$\Omega\left(\sqrt{\frac{d}{n}}\right)$ (ABRW'12)

Table 1: Bounds for excess risk of (ε, δ)-DP-SCO. Here d is dimension, n is sample size, and $\kappa = \min\{1/(p-1), e^2[\ln(d)-1]\}$; dependence on other parameters is omitted. $\tilde{O}(\cdot)$ hides logarithmic factors in n and $1/\delta$. Existing lower bounds are for nonprivate SCO: NY'83 (Nemirovski and Yudin, 1983), ABRW'12 (Agarwal et al., 2012). $(*)$: Bound shown for ℓ_1-ball feasible set.

An important instance of this framework is the *polyhedral ℓ_1-setup*, where we consider a polyhedral feasible set with bounded $\|\cdot\|_1$ radius and the losses are convex and smooth w.r.t. $\|\cdot\|_1$. This setting has several practical applications in machine learning especially when sparsity assumptions are invoked (Tibshirani, 1996; Candès et al., 2006). Furthermore, the ℓ_1 setting is the only ℓ_p setting where DP-ERM is known to enjoy nearly dimension-independent rates (Talwar et al., 2015). For this case, we provide an algorithm with nearly optimal excess population risk.

1.1. Overview of Results

We formally study DP-SCO beyond Euclidean setups. More importantly, we identify the appropriate structures that suffice to attain nearly optimal rates in these settings. A crucial ingredient of our algorithms and lower bounds are the concepts of uniform convexity and uniform smoothness in a normed space. More concretely, we use the notion of *κ-regularity* of a normed space (Juditsky and Nemirovski, 2008), which quantifies how (close to) smooth is its squared norm (see Section 2 for a formal definition). This concept has been applied in (nonprivate) convex optimization to design strongly convex regularizers, and to bound the deviations of independent sums and martingales in normed spaces. In this work we make use of these ideas, and we further show that κ-regular spaces have a natural noise addition DP mechanism that we call the *generalized Gaussian mechanism* (see Section 4). We remark that this mechanism may be of independent interest. Now we focus on ℓ_p-setups, and describe our results for the different values of $1 \leqslant p \leqslant \infty$:

Case of $p = 1$: In this case, we provide an algorithm with nearly-optimal excess population risk. Our algorithm is based on the variance-reduced one-sample stochastic Frank-Wolfe algorithm (Zhang et al., 2020a). This algorithm enjoys many attractive features: it is projection free and makes implicit use of gradients through a linear optimization oracle; it uses a single data point per iteration, allowing for larger number of iterations; and it achieves the optimal excess risk in non-private SCO in the Euclidean setting. Despite its advantages, this algorithm does not immediately apply to DP-SCO for ℓ_1-setup. The most important reason being that this algorithm was designed for the ℓ_2-setup, so our first goal is to show that a recursive gradient estimator used in (Zhang et al., 2020a) (which is a variant of the Stochastic Path-Integrated Differential EstimatoR, SPIDER (Fang et al., 2018)) does indeed work in the ℓ_1-setup. This requires controlling the variance of a martingale in ℓ_∞ which is a $O(\ln d)$-regular space. Then, using variance estimates based on κ-regularity, we are able to extend the SFW method to the ℓ_1-setup.
A second challenge comes from the differential privacy requirement. First, we use the fact that at each iteration, only a linear optimization oracle is required, and when the feasible set is polyhedral, we can construct such an oracle privately by the report noisy max mechanism (Dwork and Roth, 2014; Bhaskar et al., 2010). This technique was first used by Talwar et al. (2015) in their construction for the DP-ERM version of this problem which was based on "ordinary" full-batch FW. However, the recursive estimator in our construction is queried multiple times where a growing batch is used each time. In order to certify privacy for the whole trajectory, we carry out a novel privacy analysis for the recursive gradient estimator combined with report noisy max. Our privacy analysis uses the fact that, unlike the non-private version of SFW, our construction uses a large batch in the first iteration and uses a small, constant step size to reduce the sensitivity of the gradient estimate.

Case of $1 < p < 2$: It is interesting to investigate the risk of DP-SCO between $p = 1$ and $p = 2$. This question is intriguing since for $p = 1$ we have shown nearly dimension-independent rates, whereas for $p = 2$ it is known optimal rates grow polynomially with d. In this work, we prove lower bounds for DP-SCO and DP-ERM in this setting, which hold even in the smooth case. Our lower bounds show a surprising phenomenon: there is a *sharp phase transition* of the excess risk around $p = 1$. In fact, when $1+\Omega(1) < p < 2$, our lower bounds are essentially the same as those of the ℓ_2 case. This shows that the dependence on $\sqrt{d}$ is necessary in this regime, thus solving an open question posed in (Talwar et al., 2015). Our proof for the lower bound is based on the fingerprinting code argument due to Bun et al. (2018). In particular, we prove a reduction from DP-ERM in this setting to privately estimating 1-way marginals. Our lower bound does not follow from prior work that used a similar reduction argument, e.g., (Bassily et al., 2014a), as it requires new tools beyond what is readily available in the ℓ_2 setting. In particular, our reduction crucially relies on the strong convexity properties of ℓ_p spaces for $1 < p \leqslant 2$ (Ball et al., 1994).

We complement the lower bound result with upper bounds for DP-ERM and DP-SCO. Our upper bound for DP-ERM is tight. Our upper bound for DP-SCO matches the lower bound when $n \geqslant d^2$, and otherwise, is far from the lower bound by a small polynomial gap. The upper bounds are obtained by two structurally different algorithms. The first is a new noisy stochastic mirror-descent (SMD) algorithm based on our novel generalized Gaussian (GG) mechanism. This algorithm attains the optimal error rate for DP-ERM. For comparison, Talwar et al. (2014) proposed a batch mirror descent method combined with the Gaussian mechanism. Our use of the (GG) mechanism allows to remove the assumption of $\|\cdot\|_2$-Lipschitzness of the loss (used in (Talwar et al., 2014)), which is necessary for the optimality of SMD for DP-ERM. Moreover, using the generalization properties of differential privacy, we show that it yields excess risk $\tilde{O}(d^{1/4}/\sqrt{n})$. Our second algorithm is a variant of noisy SFW where the linear optimization subroutine is based on noisy gradients using the generalized Gaussian mechanism. The resulting excess risk of this algorithm is $O(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n^{3/4}})$, which is strictly better than SMD in the low-dimensional regime, when $n \geqslant d$, and is in fact optimal when $n \geqslant d^2$. Combining the excess-risk upper bounds of these two algorithms gives our upper bound for DP-SCO in Table 1. Sharpening these bounds is an interesting direction for future work.

Case of $2 < p \leqslant \infty$: Another interesting question is what happens in the range of $p > 2$. For comparison, it is known that non-privately the excess risk behaves as $\Theta(\frac{1}{n^{1/p}} + \frac{d^{1/2-1/p}}{\sqrt{n}})$. We show that in the low dimensional regime, $n = \tilde{\Omega}(d)$, the noisy SGD method (Bassily et al., 2014b, 2020) achieves the optimal excess risk. This implies that for $p = \infty$, noisy SGD is essentially optimal.

To conclude our overview, we note that the SFW-based algorithms for the cases $p = 1$ and $1 < p < 2$ run in time *linear in the dataset size* n, which is a desirable property for the large data size regime.

Other Related Work: Before (Talwar et al., 2015), there have been a few works that studied DP-ERM and DP-SCO in special cases of the ℓ_1 setting. Kifer et al. (2012) and Smith and Thakurta (2013) studied DP-ERM for ℓ_1 regression problems; however, they make strong assumptions about the model (e.g., restricted strong convexity). Jain and Thakurta (2014) studied DP-ERM and DP-SCO in the special case of generalized linear models (GLMs). Their bound for DP-ERM was suboptimal, and their generalization error bound relies on the special structure of GLMs, where such a bound can be obtained via a standard uniform-convergence argument. We note that such an argument does not lead to optimal bounds for general convex losses.

In independent and concurrent work, Asi et al. Asi et al. (2021) provide sharp upper bounds for DP-SCO in ℓ_1 setting, for both smooth and nonsmooth objectives. Their algorithm for the smooth case is similar to our polyhedral Stochastic Frank-Wolfe method, where their improvements are obtained by a more careful privacy accounting using a binary-tree technique. On the other hand, their work also provides nearly-optimal risk for the ℓ_p setting, when $1 < p < 2$. Interestingly, their sequential regularization approach can be further refined by using our generalized Gaussian mechanism, removing the additional poly-logarithmic factors in dimension present by their use of the standard Gaussian mechanism. We observe that the optimality of this method is certified by our lower bounds in Section 7. To conclude our comparison, we observe that our Generalized Gaussian mechanism allows us to substantially extend the applicability of the Noisy Mirror-Descent and Noisy Stochastic Frank-Wolfe method (in Section 5) to arbitrary normed spaces with a regular dual.

2. Preliminaries

Normed Spaces and Regularity. Let $(\mathbf{E}, \|\cdot\|)$ be a normed space of dimension d, and let $\langle\cdot,\cdot\rangle$ an arbitrary inner product over $\mathbf{E}$ (not necessarily inducing the norm $\|\cdot\|$). Given $x \in \mathbf{E}$ and $r > 0$, let $\mathcal{B}_{\|\cdot\|}(x, r) = \{y \in \mathbf{E} : \|y - x\| \leqslant r\}$. The dual norm over $\mathbf{E}$ is defined as usual, $\|y\|_* := \max_{\|x\|\leqslant 1}\langle y, x\rangle$. With this definition, $(\mathbf{E}, \|\cdot\|_*)$ is also a d-dimensional normed space. As a main example, consider the case of $\ell_p^d \triangleq (\mathbb{R}^d, \|\cdot\|_p)$, where $1 \leqslant p \leqslant \infty$ and $\|x\|_p \triangleq \big(\sum_{j\in[d]} |x_j|^p\big)^{1/p}$. As a consequence of the Hölder inequality, one can prove that the dual of ℓ_p^d corresponds to ℓ_q^d, where $1 \leqslant q \leqslant \infty$ is the conjugate exponent of p, determined by $\frac{1}{p} + \frac{1}{q} = 1$.

The algorithms we consider in this work can be applied to general spaces whose dual has a sufficiently smooth norm. To quantify this property, we use the notion of *regular spaces*, following Juditsky and Nemirovski (2008). Given $\kappa \geqslant 1$, we say a normed space $(\mathbf{E}, \|\cdot\|)$ is κ-regular, if there exists $1 \leqslant \kappa_+ \leqslant \kappa$ and a norm $\|\cdot\|_+$ such that $(\mathbf{E}, \|\cdot\|_+)$ is κ_+-smooth, i.e.,

$$\|x + y\|_+^2 \leqslant \|x\|_+^2 + \langle\nabla(\|\cdot\|_+^2)(x), y\rangle + \kappa_+\|y\|_+^2 \qquad (\forall x, y \in \mathbf{E}), \tag{1}$$

and $\|\cdot\|$ and $\|\cdot\|_+$ are equivalent with constant $\sqrt{\kappa/\kappa_+}$:

$$\|x\|^2 \leqslant \|x\|_+^2 \leqslant (\kappa/\kappa_+)\|x\|^2 \qquad (\forall x \in \mathbf{E}). \tag{2}$$

As basic example, Euclidean spaces are 1-regular. Other examples of regular spaces are ℓ_q^d where $2 \leqslant q \leqslant \infty$: these spaces are κ-regular with $\kappa = \min\{q-1, e^2[\ln(d)-1]\}$ and $\kappa_+ = \min\{q-1, \ln(d)-1\}$; in this case, $\|x\|_+ = \big(\sum_{j\in[d]} |x_j|^{\kappa_+}\big)^{1/\kappa_+}$ (smoothness of this function is proved e.g. in (Beck, 2017, Example 5.11).) Finally, consider a polyhedral norm $\|\cdot\|$ with unit ball $\mathcal{B}_{\|\cdot\|} = \mathrm{conv}(\mathcal{V})$. Then $(\mathbf{E}, \|\cdot\|_*)$ is $(e^2[\ln|\mathcal{V}|-1])$-regular. More precisely, note that $\|x\|_* = \max_{v\in\mathcal{V}} |\langle v, x\rangle|$, hence the norm $\|x\|_+ := \big(\sum_{v\in\mathcal{V}} |\langle v, x\rangle|^q\big)^{1/q}$, with $q = \ln|\mathcal{V}|$, satisfies (1) with $\kappa_+ = (q-1)$ (e.g., follows from (Beck, 2017, Example 5.11)), and satisfies (2) with $\sqrt{\kappa/\kappa_+} = \exp\{\frac{1}{q}\ln|\mathcal{V}|\} = e$ (using the equivalence of $\|\cdot\|_q$ and $\|\cdot\|_\infty$), thus $\kappa = e^2\kappa_+ = e^2[\ln|\mathcal{V}|-1]$.

Definition 1 (Differential Privacy (Dwork et al., 2006a,b; Dwork and Roth, 2014)) *Let $\varepsilon, \delta > 0$. A (randomized) algorithm $M : \mathcal{Z}^n \to \mathcal{R}$ is (ε, δ)-differentially private if for all pairs of datasets $S, S' \in \mathcal{Z}$ that differ in exactly one entry, and every measurable $\mathcal{O} \subseteq \mathcal{R}$, we have:*

$$\Pr(M(S) \in \mathcal{O}) \leqslant e^\varepsilon \cdot \Pr\big(M(S') \in \mathcal{O}\big) + \delta.$$

When $\delta = 0$, M is referred to as ε-differentially private.

Lemma 2 (Advanced composition (Dwork et al., 2010; Dwork and Roth, 2014)) *For any $\varepsilon > 0, \delta \in [0, 1)$, and $\delta' \in (0, 1)$, the class of (ε, δ)-differentially private algorithms satisfies $(\varepsilon', k\delta + \delta')$-differential privacy under k-fold adaptive composition, for $\varepsilon' = \varepsilon\sqrt{2k\log(1/\delta')} + k\varepsilon(e^\varepsilon - 1)$.*

Differentially Private Stochastic Convex Optimization. Let $(\mathbf{E}, \|\cdot\|)$ be a normed space, and $\mathcal{X} \subseteq \mathbf{E}$ a closed convex set of diameter $M > 0$. Given $L_0, L_1 > 0$, denote $\mathcal{C}(L_0, L_1)$ the class of functions $f : \mathcal{X} \mapsto \mathbb{R}$ which are convex; L_0-Lipschitz, i.e., $f(x) - f(y) \leqslant L_0\|x - y\|$ for all $x, y \in \mathcal{X}$; and L_1-smooth, i.e., $\|\nabla f(x) - \nabla f(y)\|_* \leqslant L_1\|x - y\|$ for all $x, y \in \mathcal{X}$. Given a *loss function* $f : \mathcal{X} \times \mathcal{Z} \mapsto \mathbb{R}$ s.t. $f(\cdot, z) \in \mathcal{C}(L_0, L_1)$ for all $z \in \mathcal{Z}$, and a distribution $\mathcal{D}$ over $\mathcal{Z}$, the SCO problem corresponds to the minimization of the *population risk*, $F_{\mathcal{D}}(x) \triangleq \mathbb{E}_{z\sim\mathcal{D}}[f(x, z)]$ over $\mathcal{X}$. Let $F^*_{\mathcal{D}} := \min_{x\in\mathcal{X}} F_{\mathcal{D}}(x)$. Given an algorithm $\mathcal{A} : \mathcal{Z}^n \mapsto \mathbf{E}$, define its *excess population risk* as

$$\mathcal{R}_{\mathcal{D}}[\mathcal{A}] = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^n,\mathcal{A}} [F_{\mathcal{D}}(\mathcal{A}(S)) - F^*_{\mathcal{D}}]. \tag{3}$$

The DP-SCO problem in the $(\mathbf{E}, \|\cdot\|)$-setup corresponds to the setting above, where algorithms are constrained to satisfy (ε, δ)-differential privacy.

We distinguish the problem above from its empirical counterpart (DP-ERM), where we are interested in minimizing the empirical risk, $\min\left\{F_S(x) = \frac{1}{n}\sum_{z\in S} f(x, z) : x \in \mathcal{X}\right\} =: F^*_S$, and accuracy is measured by the excess empirical risk, $\mathcal{R}_S[\mathcal{A}] := \mathbb{E}_{\mathcal{A}}[F_S(\mathcal{A}(S)) - F^*_S]$.

3. Private Stochastic Frank-Wolfe with Variance Reduction for Polyhedral Setup

In this section, we consider DP-SCO in the *polyhedral setup*. Let K be a positive integer, and consider $(\mathbf{E}, \|\cdot\|)$ a normed space, where the unit ball of the norm, $\mathcal{B}_{\|\cdot\|} = \text{conv}(\mathcal{V})$ is a polytope with at most K vertices. Further, the feasible set $\mathcal{X}$, is a polytope with at most K-vertices and $\|\cdot\|$-diameter $M > 0$. Notice that since the norm its polyhedral, its dual norm is also polyhedral. Moreover, $(\mathbf{E}, \|\cdot\|_*)$ is $O(\ln K)$-regular (see discussion in Section 2).

We give a differentially private stochastic Frank-Wolfe algorithm that is based on the variance reduction approach proposed in (Zhang et al., 2020b). We define the gradient variation for a given sample point $z_t \in \mathcal{Z}$ and $x^t, x^{t-1} \in \mathcal{X}$ as

$$\Delta_t(z_t) \triangleq \nabla f(x^t, z_t) - \nabla f(x^{t-1}, z_t).$$

Note that $\Delta_t(z_t)$ also depends on x^t, x^{t-1}, for notational brevity we will drop this dependence as it is clear from the context. In our algorithm we will construct a private unbiased gradient estimator $\mathbf{d}_t$ of $F_{\mathcal{D}}(x^t)$. At iteration t, for averaging parameter $\rho_t \in [0, 1]$, we use the following recursive gradient estimator:

$$\mathbf{d}_t \triangleq (1 - \rho_t)\left(\mathbf{d}_{t-1} + \Delta_t(z_t)\right) + \rho_t \nabla f(x^t, z_t). \tag{4}$$

Here, for all t we choose $\rho_t = \eta$, where η is the step size. Next, we compute a private version of $\mathbf{d}_t$ via the Report Noisy Max mechanism (Dwork and Roth, 2014; Bhaskar et al., 2010). Then, the next iterate x^{t+1} is obtained via the update step of conditional gradient methods. Hence, given the step size η, gradient estimate $\mathbf{d}_t$, the set of vertices $\mathcal{V}$, and the global sensitivity of $\langle v, \mathbf{d}_t\rangle$, that we call s_t, we have the following private update step:

$$x^{t+1} = (1 - \eta)x^t + \eta\, v_t, \quad \text{where} \quad v_t = \arg\min_{v\in\mathcal{V}} \{\langle v, \mathbf{d}_t\rangle + u^t_v\}, \quad u^t_v \sim \mathsf{Lap}(2s_t\sqrt{n\log(1/\delta)}/\varepsilon).$$

In Algorithm 1 we describe our Private Polyhedral Stochastic Frank-Wolfe Algorithm.
The privacy guarantee and expected excess population risk of Algorithm $\mathcal{A}_{\mathsf{polySFW}}$ are given by the following theorems.

Theorem 3 (Privacy Guarantee of $\mathcal{A}_{\mathsf{polySFW}}$) *Algorithm 1 is (ε, δ)-differentially private.*

Algorithm 1 $\mathcal{A}_{\mathsf{polySFW}}$: Private Polyhedral Stochastic Frank-Wolfe Algorithm

```
Require: Private dataset $S = (z_1, \dots z_n) \in \mathcal{Z}^n$, privacy parameters $(\varepsilon, \delta)$, polyhedral set $\mathcal{X}$ with
    a set of $K$ vertices $\mathcal{V} = (v_1, \dots, v_K)$
1: Set step size $\eta := \frac{\log(n/\log(K))}{n}$
2: Choose an arbitrary initial point $x^0 \in \mathcal{X}$
3: Let $B_0 = (z_1^0, \dots, z_{n/2}^0)$ be an initial batch of $\frac{n}{2}$ data points from $S$
4: Compute $\mathbf{d}_0 = \frac{2}{n}\sum_{i=1}^{n/2} \nabla f(x^0, z_i^0)$
5: $v_0 = \arg\min_{v \in \mathcal{V}} \{\langle v, \mathbf{d}_0 \rangle + u_v^0\}$, where $u_v^0 \sim \mathsf{Lap}\left(\frac{4L_0 M\sqrt{\log(1/\delta)}}{\varepsilon\sqrt{n}}\right)$
6: $x^1 \leftarrow (1-\eta)x^0 + \eta v_0$
7: Let $\widehat{S} = (z_1, \dots, z_{n/2})$ be the remaining $\frac{n}{2}$ data points in $S$ that are not in $B_0$
8: for $t = 1$ to $\frac{n}{2}$ do
9:     Set $s_t := \max\left\{(1-\eta)^t \cdot \frac{2L_0 M}{n}, 2\eta\left(L_1 M^2 + L_0 M\right)\right\}$
10:    Compute $\Delta_t(z_t) = \nabla f(x^t, z_t) - \nabla f(x^{t-1}, z_t)$
11:    $\mathbf{d}_t = (1-\eta)\left(\mathbf{d}_{t-1} + \Delta_t(z_t)\right) + \eta\nabla f(x^t, z_t)$
12:    $\forall v \in \mathcal{V}, \gamma_v \leftarrow \langle v, \mathbf{d}_t \rangle + u_v^t$, where $u_v^t \sim \mathsf{Lap}\left(\frac{2s_t\sqrt{n\log(1/\delta)}}{\varepsilon}\right)$
13:    Compute $v_t = \arg\min_{v \in \mathcal{V}} \gamma_v$
14:    $x^{t+1} \leftarrow (1-\eta)x^t + \eta v_t$
15: Output $x^{\mathsf{priv}} = x^{n/2+1}$
```

Theorem 4 (Accuracy Guarantee of $\mathcal{A}_{\mathsf{polySFW}}$) *Let $\mathcal{D}$ be any distribution over $\mathcal{Z}$. Then, for the polyhedral setup,*

$$\mathcal{R}_{\mathcal{D}}[\mathcal{A}_{\mathsf{polySFW}}] = O\left(M(L_1 M + L_0) \cdot \frac{\log(K)\log\left(n/\log(K)\right)\sqrt{\log(1/\delta)}}{\varepsilon\sqrt{n}}\right).$$

We start by stating and prove the following useful lemma.

Lemma 5 *For Algorithm 1 (Algorithm $\mathcal{A}_{\mathsf{polySFW}}$), let s_t be the global sensitivity of $\langle v, \mathbf{d}_t \rangle$, namely $s_t = \max_{v \in \mathcal{V}} \max_{S \simeq S'} |\langle v, \mathbf{d}_t - \mathbf{d}_t' \rangle|$. Then*

$$s_t \leqslant \max\left\{(1-\eta)^t \cdot \frac{2L_0 M}{n}, 2\eta\left(L_1 M^2 + L_0 M\right)\right\} \qquad (\forall t \in [n/2]).$$

Proof Let $S, S' \in \mathcal{Z}^n$ be neighboring datasets. Let $\mathbf{d}_t$ and $\mathbf{d}_t'$ denote the gradient estimates corresponding to S and S', respectively. Then, $s_t = \max_{v \in \mathcal{V}} \max_{S \simeq S'} |\langle v, \mathbf{d}_t - \mathbf{d}_t' \rangle| \leqslant M\|\mathbf{d}_t - \mathbf{d}_t'\|_*$. Now we upper bound the global $\|\cdot\|_*$ sensitivity of $\mathbf{d}_t$. First, by Step 4 in Algorithm 1, we have that the $\|\cdot\|_*$ sensitivity of $\mathbf{d}_0$ is at most $\frac{2L_0}{n}$. For $t \geqslant 1$, by expanding the recursion (4) we have

$$\mathbf{d}_t = (1-\eta)^t\,\mathbf{d}_0 + (1-\eta)\sum_{j=0}^{t-1}(1-\eta)^j\,\Delta_{t-j}(z_{t-j}) + \eta\sum_{j=0}^{t-1}(1-\eta)^j\,\nabla f(x^{t-j}, z_{t-j}).$$

We have two cases: either the data point where S and S' differ lies in the initial batch B_0 and B_0' (where B_0' is the initial batch when the input dataset is S'), or in the remaining portions of the datasets, denoted as $\widehat{S}$ and $\widehat{S}'$, respectively. If the data point where S and S' differ lies in the initial batch, then $\|\mathbf{d}_0 - \mathbf{d}'_0\|_* \leqslant (1-\eta)^t \;\, 2L_0/n$. Else, suppose that $z_{i^*} \in \widehat{S}$ and $z'_{i^*} \in \widehat{S}'$. Then

$$\|\mathbf{d}_t - \mathbf{d}'_t\|_* = (1-\eta)^{t-i^*} \|(1-\eta)\left(\Delta_{i^*}(z_{i^*}) - \Delta_{i^*}(z'_{i^*})\right) + \eta\left(\nabla f(x^{i^*}, z_{i^*}) - \nabla f(x^{i^*}, z'_{i^*})\right)\|_*.$$

Now, using that $f(\cdot, z)$ is L_0-Lipschitz and L_1-smooth w.r.t $\|\cdot\|$:

$$\|\mathbf{d}_t - \mathbf{d}'_t\|_* \leqslant 2\eta(1-\eta)^{t-i^*+1}\, L_1 M + 2\eta(1-\eta)^{t-i^*}\, L_0 \leqslant 2\eta\left(L_1 M + L_0\right)$$

Hence, in summary we obtain, $s_t \leqslant \max\left\{(1-\eta)^t \cdot \frac{2L_0 M}{n}, 2\eta\left(L_1 M^2 + L_0 M\right)\right\}$. ■

Proof of Theorem 3 By the privacy guarantee of the Report Noisy Max mechanism (Dwork and Roth, 2014; Bhaskar et al., 2010), first note that Step 5 is $\frac{\varepsilon}{\sqrt{n\log(1/\delta)}}$-DP since the global sensitivity of $\langle v, \mathbf{d}_0\rangle$ is $\frac{2L_0 M}{n}$. At any iteration $t \in [\frac{n}{2}]$, we add Laplace noise $u_v^t \sim \mathsf{Lap}\left(\frac{2s_t\sqrt{n\log(1/\delta)}}{\varepsilon}\right)$, where s_t denotes the global sensitivity of $\langle v, \mathbf{d}_t\rangle$ (upper bounded in Lemma 5). Hence, Steps 9-14 are $\frac{\varepsilon}{\sqrt{n\log(1/\delta)}}$-DP. Thus, by Lemma 2, Algorithm $\mathcal{A}_{\mathsf{polySFW}}$ is (ε, δ)-DP. ■

Now we prove the excess risk guarantee in Theorem 4. The proof relies on the following lemma which recursively bounds the variance of the gradient estimator. We defer its proof to Appendix A.

Lemma 6 *Let $\mathcal{D}$ be any distribution over $\mathcal{Z}$. Let $S \sim \mathcal{D}^n$ be the input dataset of Algorithm $\mathcal{A}_{\mathsf{polySFW}}$ (Algorithm 1). For $t \in [0, \frac{n}{2}]$, the recursive gradient estimator $\mathbf{d}_t$ satisfies*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n, \mathcal{A}_{\mathsf{polySFW}}}\left[\|\mathbf{d}_t - \nabla F_{\mathcal{D}}(x^t)\|_*\right] \leqslant \sqrt{\frac{2\log(K)}{n}}2eL_0(1-\eta)^t + 4\eta\sqrt{\log(K)t}\left(L_1 M + L_0\right).$$

Proof of Theorem 4 Let $\alpha_t \triangleq \langle v_t, \mathbf{d}_t\rangle - \min_{v\in\mathcal{V}}\langle v, \mathbf{d}_t\rangle$. By smoothness and convexity of $F_{\mathcal{D}}$:

$$\begin{aligned}
F_{\mathcal{D}}(x^{t+1}) &\leqslant F_{\mathcal{D}}(x^t) + \langle \nabla F_{\mathcal{D}}(x^t), x^{t+1} - x^t\rangle + \tfrac{L_1}{2}\|x^{t+1} - x^t\|^2 \\
&\leqslant F_{\mathcal{D}}(x^t) + \eta\langle \nabla F_{\mathcal{D}}(x^t) - \mathbf{d}_t, v_t - x^t\rangle + \tfrac{L_1\eta^2 M^2}{2} + \eta\langle \mathbf{d}_t, v_t - x^t\rangle \\
&\leqslant F_{\mathcal{D}}(x^t) + \eta M\|\nabla F_{\mathcal{D}}(x^t) - \mathbf{d}_t\|_* + \tfrac{L_1\eta^2 M^2}{2} + \eta\langle \mathbf{d}_t, x^* - x^t\rangle + \eta\alpha_t \\
&\leqslant F_{\mathcal{D}}(x^t) + 2\eta M\|\nabla F_{\mathcal{D}}(x^t) - \mathbf{d}_t\|_* + \tfrac{L_1\eta^2 M^2}{2} + \eta\langle \nabla F_{\mathcal{D}}(x^t), x^* - x^t\rangle + \eta\alpha_t \\
&\leqslant F_{\mathcal{D}}(x^t) + 2\eta M\|\nabla F_{\mathcal{D}}(x^t) - \mathbf{d}_t\|_* + \tfrac{L_1\eta^2 M^2}{2} + \eta\left(F_{\mathcal{D}}(x^*) - F_{\mathcal{D}}(x^t)\right) + \eta\alpha_t.
\end{aligned}$$

Taking expectations, and letting $\Gamma_t = F_{\mathcal{D}}(x^t) - F_{\mathcal{D}}(x^*)$, we obtain the following recursion

$$\mathbb{E}\left[\Gamma_{t+1}\right] \leqslant (1-\eta)\mathbb{E}\left[\Gamma_t\right] + 2\eta M\,\mathbb{E}\left[\|\nabla F_{\mathcal{D}}(x^t) - \mathbf{d}_t\|_*\right] + \tfrac{L_1\eta^2 M^2}{2} + \eta\mathbb{E}\left[\alpha_t\right].$$

Note that $\mathbb{E}\left[\alpha_t\right] \leqslant 2s_t\log(K)\sqrt{n\log(1/\delta)}/\varepsilon$, by standard analysis (Dwork et al., 2014). Hence,

$$\mathbb{E}\left[\Gamma_{t+1}\right] \leqslant (1-\eta)\mathbb{E}\left[\Gamma_t\right] + 2\eta M\,\mathbb{E}\left[\|\nabla F_{\mathcal{D}}(x^t) - \mathbf{d}_t\|_*\right] + \tfrac{L_1\eta^2 M^2}{2} + \tfrac{2\eta s_t\log(K)\sqrt{n\log(1/\delta)}}{\varepsilon}.$$

Hence, by Lemma 5 and Lemma 6, for $t \in [0, \frac{n}{2}]$:

$$\begin{aligned}\mathbb{E}\left[\Gamma_{t+1}\right] \leqslant & (1-\eta)\mathbb{E}\left[\Gamma_t\right] + 4e\eta L_0 M\sqrt{\tfrac{\log(K)}{n}}(1-\eta)^t + 8\eta^2 M\sqrt{\log(K)t}\ (L_1 M + L_0) \\ & + \tfrac{L_1\eta^2M^2}{2} + \tfrac{2\eta\log(K)\sqrt{n\log(1/\delta)}}{\varepsilon}\cdot\max\left\{(1-\eta)^t\,\tfrac{2L_0M}{n}, 2\eta\left(L_1M^2 + L_0M\right)\right\} \\ \leqslant & (1-\eta)\mathbb{E}\left[\Gamma_t\right] + 4e\eta L_0 M\Big(\sqrt{\tfrac{\log(K)}{n}} + \tfrac{\log(K)}{\varepsilon}\sqrt{\tfrac{\log(1/\delta)}{n}}\Big)(1-\eta)^t \\ & + 8\eta^2M(L_1M+L_0)\Big(\sqrt{\log(K)t} + \tfrac{\log(K)\sqrt{n\log(1/\delta)}}{\varepsilon}\Big) + \tfrac{L_1\eta^2M^2}{2}.\end{aligned}$$

Next, by expanding the above recursion we have

$$\begin{aligned}\mathbb{E}\left[\Gamma_{\frac{n}{2}+1}\right] \leqslant & (1-\eta)^{\frac{n}{2}+1}L_0M + \tfrac{4e\eta(1-\eta)^{\frac{n}{2}}L_0M\ n}{2}\Big(\sqrt{\tfrac{\log(K)}{n}} + \tfrac{\log(K)}{\varepsilon}\sqrt{\tfrac{\log(1/\delta)}{n}}\Big) \\ & + \tfrac{1}{\eta}\Big[8\eta^2M(L_1M+L_0)\Big(\sqrt{\log(K)n} + \tfrac{\log(K)\sqrt{n\log(1/\delta)}}{\varepsilon}\Big) + \tfrac{L_1\eta^2M^2}{2}\Big] \\ \leqslant & e^{-\eta\left(\frac{n}{2}+1\right)}L_0M + 2e\cdot e^{-\eta\left(\frac{n}{2}\right)}\eta L_0M\ n\Big(\sqrt{\tfrac{\log(K)}{n}} + \tfrac{\log(K)}{\varepsilon}\sqrt{\tfrac{\log(1/\delta)}{n}}\Big) \\ & + 8\eta M(L_1M+L_0)\Big(\sqrt{\log(K)n} + \tfrac{\log(K)\sqrt{n\log(1/\delta)}}{\varepsilon}\Big) + \tfrac{L_1\eta M^2}{2}.\end{aligned}$$

Choosing $\eta = \frac{1}{n}\log\left(\frac{n}{\log(K)}\right)$, we get

$$\begin{aligned}\mathbb{E}\left[\Gamma_{\frac{n}{2}+1}\right] \leqslant & L_0M\sqrt{\tfrac{\log(K)}{n}} + 2eL_0M\log\left(\tfrac{n}{\log(K)}\right)\left(\tfrac{\log(K)}{n} + \tfrac{\log^{3/2}(K)\sqrt{\log(1/\delta)}}{\varepsilon\ n}\right) \\ & + 8M(L_1M+L_0)\log\left(\tfrac{n}{\log(K)}\right)\left(\sqrt{\tfrac{\log(K)}{n}} + \tfrac{\log(K)}{\varepsilon}\sqrt{\tfrac{\log(1/\delta)}{n}}\right) + \tfrac{L_1M^2}{2n}\log\left(\tfrac{n}{\log(K)}\right).\end{aligned}$$

By assuming $n > \log(K)$ (which is necessary to achieve non-trivial error even in the non-private setting), we obtain $\mathbb{E}\left[\Gamma_{\frac{n}{2}+1}\right] = O\left(\frac{M(L_1M+L_0)}{\varepsilon\sqrt{n}}\cdot\log(K)\log\left(n/\log(K)\right)\sqrt{\log(1/\delta)}\right)$, which is the desired bound on the excess population risk. ■

4. Generalized Gaussian Distribution and Mechanism

One important requirement for the application of DP stochastic first-order methods is designing the proper private mechanism for an iterative method. If we want to achieve privacy by adding noise to gradients, then we need to do it in a way to achieve privacy from *gradient sensitivity*, w.r.t. the dual norm. With this purpose in mind, we design a new noise addition mechanism that leverages the regularity of the *dual space* $(\mathbf{E}, \|\cdot\|_*)$.

Definition 7 (Generalized Gaussian distribution and mechanism) *Let $(\mathbf{E}, \|\cdot\|_*)$ be a d-dimensional κ-regular space with smooth norm $\|\cdot\|_+$. We define the generalized Gaussian (GG) distribution $\mathcal{GG}_{\|\cdot\|_+}(\mu, \sigma^2)$, as the one with density $g(z) = C(\sigma, d)\exp\{-\|z-\mu\|_+^2/[2\sigma^2]\}$, where $C(\sigma, d) = \left[Area(\{\|x\|_+ = 1\})\frac{(2\sigma^2)^{d/2}}{2}\Gamma(d/2)\right]^{-1}$, and Area is the $((d-1)$-dim.) surface measure on $\mathbb{R}^d$.*

Given an algorithm $\mathcal{A} : S^n \mapsto \mathbf{E}$ with bounded $\|\cdot\|_*$-sensitivity: $\sup_{S\simeq S'}\|\mathcal{A}(S) - \mathcal{A}(S')\|_* \leqslant s$, we define the generalized Gaussian mechanism *of $\mathcal{A}$ with noise variance σ^2 as*

$$\mathcal{A}_{\mathcal{GG}}(S) \sim \mathcal{GG}_{\|\cdot\|_+}(\mathcal{A}(S), \sigma^2).$$

We observe that, despite the generality of the norm $\|\cdot\|_+$, when integrating on level sets we obtain explicit formulae for the moments of the distribution. On the other hand, the privacy properties of this mechanism can be established by leveraging the smoothness of its negative log-density.

Proposition 8

(a) If $z \sim \mathcal{GG}_{\|\cdot\|_+}(\mu, \sigma^2)$, then $\mathbb{E}[\|z\|_*^2] \leqslant \mathbb{E}[\|z\|_+^2] \leqslant d\sigma^2$.

(b) The generalized Gaussian mechanism applied to a function with $\|\cdot\|_*$-sensitivity bounded by $s > 0$ is (ε, δ)-DP for $\sigma^2 = 2\kappa \log(1/\delta)s^2/\varepsilon^2$.

Proof For (a) we refer to Appendix B. For (b), let $\mathbb{P} = \mathcal{GG}(\mu_1, \sigma^2)$ and $\mathbb{Q} = \mathcal{GG}(\mu_2, \sigma^2)$. Then[1]

$$\begin{aligned} \exp\{(\alpha-1)D_\alpha(\mathbb{P}||\mathbb{Q})\} &= C(\sigma,d)\textstyle\int_{\mathbb{R}^d} \left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)^\alpha d\mathbb{Q} \\ &= C(\sigma,d)\textstyle\int_{\mathbb{R}^d} \exp\Big\{-\frac{\alpha}{2\sigma^2}\|z-\mu_1\|_+^2 + \frac{\alpha-1}{2\sigma^2}\|z-\mu_2\|_+^2\Big\}dz \\ &= C(\sigma,d)\textstyle\int_{\mathbb{R}^d} \exp\Big\{-\frac{\alpha}{2\sigma^2}\|z-\mu_1+\mu_2\|_+^2 + \frac{\alpha-1}{2\sigma^2}\|z\|_+^2\Big\}dz. \end{aligned}$$

Let now $\mu = \mu_1 - \mu_2$ and $p(\cdot) = \|\cdot\|_+^2$. Now, by convexity and smoothness of $\|\cdot\|_+^2$

$$-\alpha\|z-\mu\|_+^2 \leqslant -\alpha\|z\|_+^2 + \langle \nabla p(z), \alpha\mu\rangle \leqslant -\alpha\|z\|_+^2 + [\|z\|_+^2 - \|z-\alpha\mu\|_+^2 + \kappa_+\|\alpha\mu\|_+^2].$$

Plugging this in the integral above, we get

$$\exp\{(\alpha-1)D_\alpha(\mathbb{P}||\mathbb{Q})\} \leqslant \exp\big\{\tfrac{\kappa_+\alpha^2}{2\sigma^2}\|\mu\|_+^2\big\}C(\sigma,d)\textstyle\int_{\mathbb{R}^d}\exp\big\{-\frac{\|z-\alpha\mu\|_+^2}{2\sigma^2}\big\}dz \leqslant \exp\big\{\frac{\kappa\alpha^2}{2\sigma^2}\|\mu\|_*^2\big\},$$

i.e., $D_\alpha(\mathbb{P}||\mathbb{Q}) \leqslant \frac{\kappa\alpha^2}{2\sigma^2(\alpha-1)}\|\mu\|_*^2$. The result can be obtained now by using a known reduction from Rényi DP to DP (Mironov, 2017) (for details see Appendix B). ■

5. Differentially Private SCO: ℓ_p-setup for $1 < p < 2$

Our goal now is to upper bound the excess risk of DP-SCO in the ℓ_p-setup when $1 < p < 2$. For this, we will prove the following upper bound on the excess risk, using two different algorithms, explored in Sections 5.1 and 5.2, respectively.

Theorem 9 *Let $1 < p < 2$ and $\kappa = \min\{1/(p-1), 2\ln d\}$. The expected excess population risk of DP-SCO in the ℓ_p-setup is upper bounded by*[2]

$$O\left(\min\left\{\frac{L_0 M\sqrt{\kappa}[d\log(1/\delta)]^{1/4}}{\sqrt{n}}, \frac{\kappa L_0 M + L_1 M^2 \log n}{\sqrt{n}} + \frac{\kappa L_1 M^2 \log n\sqrt{d\log(1/\delta)}}{\varepsilon n^{3/4}}\right\}\right).$$

[1] Here $D_\alpha(\mathbb{P}||\mathbb{Q})$ is the α-Rényi divergence between $\mathbb{P}$ and $\mathbb{Q}$.

[2] To simplify the bound, here we assume $\varepsilon = \Omega\left(\frac{\sqrt{\kappa}\,[d\,\log(1/\delta)]^{1/4}}{\sqrt{n}}\right)$.

5.1. Noisy Stochastic Mirror-Descent

As a first application of the GG mechanism, we provide a simple algorithm for DP-SCO, in the ℓ_p-setup when $1 < p < 2$ (this algorithm also works for spaces with κ-regular dual). This is the *noisy stochastic mirror-descent* (SMD) method, which turns out to be optimal for DP-ERM, and using the generalization properties of differential privacy we can derive bounds for DP-SCO. We also note that this algorithm does not require smoothness of the objectives, and thus works in the nonsmooth convex setting as well. Here we provide a pseudocode of the algorithm in Algorithm 2. The analysis of this approach is an adaptation of techniques used in Bassily et al. (2019) and Bassily et al. (2014a) for noisy SGD, which we defer to Appendix C. We emphasize the importance of the GG mechanism for this result. For comparison, the sequential regularization approach in Asi et al. (2021) uses the Gaussian mechanism, hence requiring Lipschitzness w.r.t. $\|\cdot\|_2$, which leads to additional poly-logarithmic in d factors in the ℓ_p-setup. Moreover, using Generalized Gaussian mechanism we can extend the applicability of the noisy SMD method to arbitrary normed spaces with a κ-regular dual.

Algorithm 2 $\mathcal{A}_{\mathsf{noisySMD}}$: Noisy minibatch SMD for nonsmooth convex losses

```
Require: Private dataset $S = (z_1, \dots, z_n) \in \mathcal{Z}^n$; step size $\eta$; privacy parameters $\varepsilon \leqslant 1$, $\delta \ll 1/n$
Set noise variance $\sigma^2 = 8\kappa L_0^2 \log(1/\delta)/(\varepsilon n)^2$
Set batch size $m := \max\{n\sqrt{\varepsilon/(4T)}, 1\}$
Choose an arbitrary initial point $x^1 \in \mathcal{X}$
for $t = 1$ to $T-1$ do
    Sample $B_t = \{z_i\}_{i \in I_t}$ with $I_t \sim \mathrm{Unif}([n]^m)$ (i.e., $B_t$ is sampled with replacement.)
    $x^{t+1} := \arg\min_{x \in \mathcal{X}}\{\langle \eta[\nabla F_{B_t}(x^t) + g_t], x - x^t\rangle + \Phi(x) - \Phi(x^t) - \langle \nabla\Phi(x^t), x - x^t\rangle\}$,
    where $g_t \sim \mathcal{GG}(0, \sigma^2)$ drawn independently each iteration
return $\overline{x}^T = \frac{1}{T}\sum_{t=1}^T x^t$
```

Theorem 10 *Let $1 < p < 2$, $\kappa = \min\{\frac{1}{p-1}, 2\ln(d)\}$, and $\kappa_* = \kappa/(\kappa-1)$, and consider the ℓ_p-setup of DP-SCO. Then $\mathcal{A}_{\mathsf{noisySMD}}$ (Algorithm 2) with regularizer $\Phi(\cdot) = \frac{\kappa}{2}\|\cdot\|_{\kappa_*}^2$, $T = \lfloor \frac{(\varepsilon n)^2}{16 d \kappa \log(1/\delta)} \rfloor$, and stepsize $\eta = \frac{M}{L_0}\sqrt{\frac{\kappa}{2T}}$ is (ε, δ)-DP. Moreover, for any dataset $S \in \mathcal{Z}^n$,*

$$\mathcal{R}_S[\mathcal{A}_{\mathsf{noisySMD}}] = O\Big(L_0 M \cdot \frac{\kappa\sqrt{d\log(1/\delta)}}{\varepsilon n}\Big),$$

and for any distribution $\mathcal{D}$ supported on $\mathcal{Z}$,

$$\mathcal{R}_{\mathcal{D}}[\mathcal{A}_{\mathsf{noisySMD}}] = O\left(L_0 M \cdot \left(\max\left(\frac{\sqrt{\kappa}[d\log(1/\delta)]^{1/4}}{\sqrt{n}}, \frac{\kappa\sqrt{d\log(1/\delta)}}{\varepsilon n}\right)\right)\right).$$

5.2. Noisy Variance-Reduced Stochastic Frank-Wolfe

In this section, we describe another variant of the variance-reduced stochastic Frank-Wolfe algorithm algorithm, $\mathcal{A}_{\mathsf{noisySFW}}$, (Algorithm 3). This algorithm differs from the polyhedral SFW (Algorithm 1) in various ways: first, it uses gradient noise addition as privacy-preserving mechanism, perticularly our GG mechanism; second, it uses large minibatches of size $\Omega(\sqrt{n})$ across iterations, which is crucial to control the sensitivity of gradient queries; and third, the recursive gradient estimator is closer to the original SPIDER estimator (Fang et al., 2018), which didn't use averaging factors $0 < \rho < 1$, and simply accumulates the gradient variations.

In Algorithm 3, first we compute the initial gradient estimate using an initial batch of size $n/2$ as given in Step 5. Next, in Step 11 we compute the private version of the gradient variation (Δ_t) with respect to a mini-batch of size $\sqrt{n}/2$, by adding generalized Gaussian noise to it. Hence, the recursive gradient estimator is given by Step 12. Finally, the next iterate is given by $x^{t+1} = (1-\eta)x^t + \eta \arg\min_{v\in\mathcal{X}} \langle \tilde{\nabla}_t, v\rangle$.

Algorithm 3 $\mathcal{A}_{\mathsf{noisySFW}}$: Noisy Private Stochastic Frank-Wolfe Algorithm

```
Require: Private dataset: $S = (z_1, \dots z_n) \in \mathcal{Z}^n$, privacy parameters: $(\varepsilon, \delta)$
 1: Set step size $\eta := \frac{\log(n)}{2\sqrt{n}}$, $\kappa := \min\left\{\frac{1}{p-1}, e^2 \ln(d)\right\}$.
 2: Choose an arbitrary initial point $x^0 \in \mathcal{X}$.
 3: Let $B_0 = (z_1^0, \dots, z_{n/2}^0)$ be an initial batch of $\frac{n}{2}$ data points from $S$.
 4: Set $\sigma_0^2 := \frac{32\kappa L_0^2 \log(1/\delta)}{n^2\varepsilon^2}$.
 5: Compute $\tilde{\nabla}_0 = \frac{2}{n}\sum_{i=1}^{n/2} \nabla f(x^0, z_i^0) + \mathbf{g}_0$, where $\mathbf{g}_0 \sim \mathcal{GG}_{\|\cdot\|_+}(\mathbf{0}, \sigma_0^2)$.
 6: $x^1 \leftarrow (1-\eta)x^0 + \eta \arg\min_{v\in\mathcal{X}} \langle \tilde{\nabla}_0, v\rangle$.
 7: Let $\widehat{S} = (z_1, \dots, z_{n/2})$ be the remaining $\frac{n}{2}$ data points in $S$ that are not in $B_0$.
 8: Set noise variance $\sigma^2 := \frac{32\kappa L_1^2 M^2 \eta^2 \log(1/\delta)}{n\varepsilon^2}$.
 9: for $t = 1$ to $\sqrt{n}$ do
10:     Let $B_t = (z_1^t, \dots, z_{\sqrt{n}/2}^t)$ be a batch of $\frac{\sqrt{n}}{2}$ data points from $\widehat{S}$
11:     Compute $\tilde{\Delta}_t = \frac{2}{\sqrt{n}}\sum_{i=1}^{\sqrt{n}/2}\left(\nabla f(x^t, z_i^t) - \nabla f(x^{t-1}, z_i^t)\right) + \mathbf{g}_t$, where $\mathbf{g}_t \sim \mathcal{GG}_{\|\cdot\|_+}(\mathbf{0}, \sigma^2)$.
12:     $\tilde{\nabla}_t = \tilde{\nabla}_{t-1} + \tilde{\Delta}_t$.
13:     Compute $v_t = \arg\min_{v\in\mathcal{X}} \langle \tilde{\nabla}_t, v\rangle$.
14:     $x^{t+1} \leftarrow (1-\eta)x^t + \eta v_t$.
15: Output $x^{\mathsf{priv}} = x^{\sqrt{n}+1}$.
```

Theorem 11 (Privacy Guarantee of $\mathcal{A}_{\mathsf{noisySFW}}$) *Algorithm 3 is (ε, δ)-differentially private.*

The proof of the above theorem follows from a global sensitivity bound on the gradient queries, together with the privacy guarantee of the generalized Gaussian mechanism (Proposition 8) and parallel composition. We defer the formal proof of Theorem 11 to Appendix D.1.

Theorem 12 (Accuracy Guarantee of $\mathcal{A}_{\mathsf{noisySFW}}$) *Let $p \in (1,2)$, and $\kappa = \min\left\{\frac{1}{p-1}, e^2 \ln(d)\right\}$, and consider the ℓ_p-setup of DP-SCO. Then, for any distribution $\mathcal{D}$ supported on $\mathcal{Z}$,*

$$\mathcal{R}_{\mathcal{D}}[\mathcal{A}_{\mathsf{noisySFW}}] = O\left(\frac{L_1 M^2 \log(n) + L_0 M \kappa}{\sqrt{n}} + \frac{\kappa L_1 M^2 \log(n)\sqrt{d \log(1/\delta)}}{\varepsilon\, n^{3/4}}\right).$$

The proof of the above theorem follows similar lines to the proof of Theorem 4. Due to space considerations, we defer the full proof of the theorem to Appendix D.2.

6. Differentially Private SCO: ℓ_p-setup for $2 < p \leqslant \infty$

The proposed analyses, when applied to ℓ_p-settings, only appear to provide useful bounds when $1 \leqslant p \leqslant 2$. This limitation comes from the fact that when $p > 2$ the regularity constant of the dual, ℓ_q, grows polynomially on the dimension, more precisely as $d^{1-2/p}$. This additional factor in the analysis substantially degrades the resulting excess risk bounds, unless $p \approx 2$. This leaves the question of what are the optimal rates for DP-SCO in such settings.

It is instructive to recall the optimal excess risk bounds for nonprivate SCO (Nemirovski and Yudin, 1983; Agarwal et al., 2012). These bounds have the form $\Theta(\min\{\frac{d^{1/2-1/p}}{\sqrt{n}}, \frac{1}{n^{1/p}}\})$, and are attained by the combination of two different algorithms: first, stochastic gradient descent, for the low dimensional $d < n$ regime, with rate $O(\frac{d^{1/2-1/p}}{\sqrt{n}})$; and second, stochastic mirror descent (with regularizer $\frac{1}{p}\|x\|_p^p$), for the high dimensional $d \geqslant n$ regime, with rate $O(\frac{1}{n^{1/p}})$. We now show that in the low dimensional case, the multipass noisy SGD method is essentially optimal (Bassily et al., 2014a, 2020). The proof of this simple result is deferred to Appendix E.

Proposition 13 *Consider the problem of DP-SCO in the $\ell_p = (\mathbb{R}^d, \|\cdot\|_p)$-setup, with $p > 2$. Then the multipass noisy SGD method (Bassily et al., 2020, Algorithm 2) attains excess population risk $O\big(L_0 M\big(\frac{d^{1/2-1/p}}{\sqrt{n}} + \frac{d^{1-1/p}\sqrt{\log(1/\delta)}}{\varepsilon n}\big)\big)$.*

We conclude this section observing that in the low-dimensional regime: $n \geqslant d\log(1/\delta)/\varepsilon^2$, the above upper bound is optimal since it matches the optimal non-private lower bound of $\Omega\left(\frac{d^{1/2-1/p}}{\sqrt{n}}\right)$ (Agarwal et al., 2012). Note that in the ℓ_∞ setting, the regime $n > d$ is the only interesting regime since the excess risk is $\Omega(1)$ if $n \leqslant d$. Hence, our result implies that SGD attains essentially optimal excess risk for DP-SCO in the ℓ_∞ setting. We formally state this observation below.

Corollary 14 *Let $2 < p \leqslant \infty$ and $\mathcal{X} = \mathcal{B}_{\|\cdot\|_p}(0, M)$. If $d\log(1/\delta)/\varepsilon^2 \leqslant n$, then Multipass Noisy SGD attains the optimal excess population risk for DP-SCO in the ℓ_p-setup.*

7. Lower Bound for DP-ERM and DP-SCO in the ℓ_p setup for $1 < p < 2$

We provide lower bounds on the excess risk for DP-ERM and DP-SCO in the ℓ_p setting for $1 < p < 2$. In our argument, we first prove a lower bound on DP-ERM, then use the reduction in (Bassily et al., 2019, Appendix C) to assert that essentially the same lower bound (up to a logarithmic factor in $1/\delta$) holds for DP-SCO. Our final lower bound for DP-SCO follows from combining this bound with the non-private $\Omega(1/\sqrt{n})$ lower bound for SCO when $1 < p < 2$ (Nemirovski and Yudin, 1983). Below, we formally state our lower bound for DP-SCO and provide an outline of our argument. We defer the full details of our construction and the statement of the lower bound for DP-ERM to Appendix F. We remark that our lower bound for DP-ERM (Theorem 19 in Appendix F) implies that our upper bound for DP-ERM resulting from the noisy SMD algorithm (Theorem 10) is tight when $1 + \Omega(1) < p < 2$.

Theorem 15 (Lower Bound for DP-SCO for $p \in (1, 2)$) *Let $p \in (1, 2)$ and $n, d \in \mathbb{N}$. Let $\varepsilon > 0$ and $0 < \delta < \frac{1}{n^{1+\Omega(1)}}$. Let $\mathcal{X} = \mathcal{B}_p^d$, where $\mathcal{B}_p^d$ is the unit ℓ_p ball in $\mathbb{R}^d$, and $\mathcal{Z} = \{-\frac{1}{d^{1/q}}, \frac{1}{d^{1/q}}\}^d$,* *where $q = \frac{p}{p-1}$. There exists a distribution $\mathcal{D}$ over $\mathcal{Z}$ such that for any (ε, δ)-DP-SCO algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{X}$, we have*

$$\mathcal{R}_{\mathcal{D}}[\mathcal{A}] = \tilde{\Omega}\left(\max\left(\frac{1}{\sqrt{n}}, (p-1)\frac{\sqrt{d}}{\varepsilon n}\right)\right).$$

As mentioned earlier, to prove this theorem, it suffices to construct a lower bound for DP-ERM. To do so, we consider an a linear instance of the loss given by $f(x,z) = -\langle x, z\rangle$, $x \in \mathcal{X}, z \in \mathcal{Z}$, where $\mathcal{X}$ and $\mathcal{Z}$ as defined in the above theorem. For any dataset $S = (z_1, \ldots, z_n) \in \mathcal{Z}^n$, let $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i$, and let x^* denote the minimizer of the empirical risk $F_S(x), x \in \mathcal{X}$ with respect to the loss f. The first step of our proof is to show that for any $S \in \mathcal{Z}^n$ and any $\alpha > 0$ if we have $\widehat{x} \in \mathcal{X}$ such that $F_S(\widehat{x}) - F_S(x^*) \leqslant \alpha$ then $\|\widehat{x} - x^*\|_p = O\left(\sqrt{\frac{\alpha}{(p-1)\|\bar{z}\|_q}}\right)$. This step is a crucial ingredient in our proof since it allows us to transform a lower bound on the ℓ_p distance to the minimizer into a lower bound on the excess risk. We remark that this step requires new tools than what is readily available in the Euclidean setting (considered in the lower bound of Bassily et al. (2014a)). In particular, it relies on the strong convexity property of the ℓ_p spaces for $1 < p < 2$.

Next, we show the existence of dataset S with $\|\bar{z}\|_q = \Omega\left(\sqrt{d\log(1/\delta)}/(\varepsilon n)\right)$ for which any (ε, δ)-DP-ERM algorithm $\mathcal{A}$ must satisfy $\|\mathcal{A}(S) - x^*\|_p = \Omega(1)$. Note that, given the first step above, this would then imply the desired lower bound on the excess risk. To do so, we use the fingerprinting code argument from (Bun et al., 2018) that shows the existence of a dataset of n elements from $\{-1, 1\}^d$ such that any (ε, δ)-differentially private algorithm for estimating the empirical average of S (1-way marginals) must make error $\Omega(\sqrt{d\log(1/\delta)}/(\varepsilon n))$ in $\Omega(d)$ coordinates of the resulting d-dimensional vector. Via a careful argument that uses the properties of this dataset and those of our instance of the optimization problem, we show that solving the DP-ERM problem w.r.t. a normalized version of this dataset implies privately estimating the 1-way marginals based on this dataset. This leads us to the $\Omega(1)$ lower bound on the distance to the minimizer, which suffices to prove our lower bound on the excess risk as described above.

Acknowledgements

We would like to thank Digvijay Boob, Adam Smith and Abhradeep Guha Thakurta for the insightful discussions in the early stages of this work. RB's and AN's research is supported by NSF Award AF-1908281, Google Faculty Research Award, and the OSU faculty start-up support. CG's research is partially supported by INRIA through the INRIA Associate Teams project, CORFO through the Clover 2030 Engineering Strategy – 14ENI-26862, and FONDECYT 1210362 project.

References

Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Trans. Inf. Theory*, 58(5):3235–3249, 2012. doi: 10.1109/TIT.2011.2182178. URL https://doi.org/10.1109/TIT.2011.2182178.

Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in l1 geometry. *CoRR*, abs/2103.01516, 2021. URL https://arxiv.org/abs/2103.01516.

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

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS 2014). (arXiv preprint arXiv:1405.7085)*, pages 464–473. 2014a.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Differentially private empirical risk minimization: Efficient algorithms and tight error bounds. *FOCS 2014. Also, arXiv preprint arXiv:1405.7085*, 2014b.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *STOC*, 2016.

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. In *Advances in Neural Information Processing Systems 33, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Amir Beck. *First-order methods in optimization / Amir Beck, Tel-Aviv University, Tel-Aviv, Israel.* MOS-SIAM series on optimization. Society for Industrial and Applied Mathematics, Mathematical Optimization Society, Philadelphia, 2017. ISBN 9781611974980.

Raghav Bhaskar, Srivatsan Laxman, Adam Smith, and Abhradeep Thakurta. Discovering frequent patterns in sensitive data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 503–512, 2010.

Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 634–649. IEEE, 2015.

Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. *SIAM Journal on Computing*, 47(5):1888–1938, 2018.

Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou, editors, *NIPS*. MIT Press, 2008.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

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

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, 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.

Cynthia Dwork, Guy N. Rothblum, and Salil P. Vadhan. Boosting and differential privacy. In *FOCS*, 2010.

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.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *NIPS*, 2015.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 687–697, 2018.

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.

Prateek Jain and Abhradeep Thakurta. (near) dimension independent risk bounds for differentially private learning. In *ICML*, 2014.

Anatoli Juditsky and Arkadi Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. Rapport de recherche hal-00318071, HAL, 2008.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. *Journal of Machine Learning Research*, 1:41, 2012.

Ilya Mironov. Rényi differential privacy. In *30th IEEE Computer Security Foundations Symposium, CSF 2017, Santa Barbara, CA, USA, August 21-25, 2017*, pages 263–275. IEEE Computer Society, 2017. doi: 10.1109/CSF.2017.11. URL https://doi.org/10.1109/CSF.2017.11.

A. Nemirovski and D. Yudin. Problem complexity and method efficiency in optimization. 1983.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. on Optimization*, 19(4):1574–1609, January 2009.

Adam Smith and Abhradeep Thakurta. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *COLT*, 2013.

Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright. *Optimization for Machine Learning*. The MIT Press, 2011. ISBN 026201646X.

Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *arXiv preprint arXiv:1501.06095*, 2015.

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

Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Nearly optimal private lasso. In *NIPS*, 2015.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems*, pages 2722–2731, 2017.

C. Zalinescu. On uniformly convex functions. *Journal of Mathematical Analysis and Applications*, 95:344–374, 1983.

Mingrui Zhang, Zebang Shen, Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. One sample stochastic frank-wolfe. In *International Conference on Artificial Intelligence and Statistics*, pages 4012–4023. PMLR, 2020a.

Mingrui Zhang, Zebang Shen, Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. One sample stochastic frank-wolfe. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4012–4023, Online, 26–28 Aug 2020b. PMLR. URL http://proceedings.mlr.press/v108/zhang20i.html.

Appendix A. Proof of Lemma 6 from Section 3

To bound the variance of the gradient estimator, we first compute

$$\begin{aligned}\mathbf{d}_t - \nabla F_{\mathcal{D}}(x^t) &= (1-\eta)\,\left[\mathbf{d}_{t-1} - \nabla F_{\mathcal{D}}(x^{t-1})\right] + (1-\eta)\nabla F_{\mathcal{D}}(x^{t-1}) - \nabla F_{\mathcal{D}}(x^t)\\ &\quad + (1-\eta)\Delta_t(z_t) + \eta\nabla f(x^t, z_t)\\ &= (1-\eta)\,\left[\mathbf{d}_{t-1} - \nabla F_{\mathcal{D}}(x^{t-1})\right] + (1-\eta)\,\left[\Delta_t(z_t) - \left(\nabla F_{\mathcal{D}}(x^t) - \nabla F_{\mathcal{D}}(x^{t-1})\right)\right]\\ &\quad + \eta\,\left[\nabla f(x^t, z_t) - \nabla F_{\mathcal{D}}(x^t)\right].\end{aligned}$$

For a compact notation, let $\bar{\Delta}_t \triangleq \nabla F_{\mathcal{D}}(x^t) - \nabla F_{\mathcal{D}}(x^{t-1})$. Recall that $\|\cdot\|_*$ is κ-regular, with $\kappa = (e^2 \ln K)$. Denote $\|\cdot\|_+$ the corresponding κ_+-smooth norm, where $\kappa_+ = [\ln(K) - 1]$, and $\kappa/\kappa_+ \leqslant e^2$. First we will bound the variance on $\|\cdot\|_+$, and then we will derive the result using the equivalence property (2). Let $\mathcal{F}_t$ be the σ-algebra generated by the randomness in the data and the algorithm up until iteration t. By property (1), we observe that

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{d}_t - \nabla F_{\mathcal{D}}(x^t)\|_+^2 | \mathcal{F}_{t-1}] \\
&\leqslant (1-\eta)^2 \mathbb{E}\left[\|\mathbf{d}_{t-1} - \nabla F_{\mathcal{D}}(x^{t-1})\|_+^2 | \mathcal{F}_{t-1}\right] + \\
&\quad \mathbb{E}\left[\langle \nabla(\|\cdot\|_+^2)\left(\mathbf{d}_{t-1} - \nabla F_{\mathcal{D}}(x^{t-1})\right), (1-\eta)\left(\Delta_t(z_t) - \bar{\Delta}_t\right) + \eta\left(\nabla f(x^t, z_t) - \nabla F_{\mathcal{D}}(x^t)\right)\rangle | \mathcal{F}_{t-1}\right] \\
&\quad + \kappa_+ \mathbb{E}\left[\|(1-\eta)\left(\Delta_t(z_t) - \bar{\Delta}_t\right) + \eta\left(\nabla f(x^t, z_t) - \nabla F_{\mathcal{D}}(x^t)\right)\|_+^2 | \mathcal{F}_{t-1}\right] \\
&= (1-\eta)^2 \mathbb{E}\left[\|\mathbf{d}_{t-1} - \nabla F_{\mathcal{D}}(x^{t-1})\|_+^2 | \mathcal{F}_{t-1}\right] + \\
&\quad + \kappa_+ \mathbb{E}\left[\|(1-\eta)\left(\Delta_t(z_t) - \bar{\Delta}_t\right) + \eta\left(\nabla f(x^t, z_t) - \nabla F_{\mathcal{D}}(x^t)\right)\|_+^2 | \mathcal{F}_{t-1}\right] \\
&\leqslant (1-\eta)^2 \mathbb{E}\left[\|\mathbf{d}_{t-1} - \nabla F_{\mathcal{D}}(x^{t-1})\|_+^2 | \mathcal{F}_{t-1}\right] + 2\kappa_+(1-\eta)^2 \mathbb{E}\left[\|\Delta_t(z_t) - \bar{\Delta}_t\|_+^2 | \mathcal{F}_{t-1}\right] \\
&\quad + 2\kappa_+ \eta^2 \mathbb{E}\left[\|\nabla f(x^t, z_t) - \nabla F_{\mathcal{D}}(x^t)\|_+^2 | \mathcal{F}_{t-1}\right].
\end{aligned}$$

In the second equality we have used the fact that for any $x \in \mathcal{X}$, $\mathbb{E}_{z\sim\mathcal{D}}[\nabla f(x,z)] = \nabla F_{\mathcal{D}}(x)$, and $\mathbb{E}_{z\sim\mathcal{D}}[\Delta_t(z_t)] = \bar{\Delta}_t$, and that the two terms $\left(\mathbf{d}_{t-1} - \nabla F_{\mathcal{D}}(x^{t-1})\right)$ and $(1-\eta)\left(\Delta_t(z_t) - \bar{\Delta}_t\right) + \eta\left(\nabla f(x^t, z_t) - \nabla F_{\mathcal{D}}(x^t)\right)$, conditioned on $\mathcal{F}_{t-1}$ are independent. The last inequality follows by triangle inequality and the fact that $(a+b)^2 \leqslant 2a^2 + 2b^2$ for $a, b \in \mathbb{R}$. Hence, using (2) and that $f(\cdot, z)$ is L_0-Lipschitz and L_1-smooth:

$$\mathbb{E}\left[\|\Delta_t(z_t) - \bar{\Delta}_t\|_+^2 | \mathcal{F}_{t-1}\right] \leqslant \frac{\kappa}{\kappa_+}\mathbb{E}\left[\|\Delta_t(z_t) - \bar{\Delta}_t\|_*^2 | \mathcal{F}_{t-1}\right] \leqslant 4e^2\left(L_1 M\right)^2 \eta^2 \qquad (5)$$

$$\mathbb{E}\left[\|\nabla f(x^t, z_t) - \nabla F_{\mathcal{D}}(x^t)\|_+^2 | \mathcal{F}_{t-1}\right] \leqslant 4e^2 L_0^2. \qquad (6)$$

Let $\mathbf{r}_t \triangleq \|\mathbf{d}_t - \nabla F_{\mathcal{D}}(x^t)\|_+^2$. Thus, by (5) and (6) we get the following recursion:

$$\mathbb{E}[\mathbf{r}_t | \mathcal{F}_{t-1}] \leqslant (1-\eta)^2 \mathbb{E}\left[\mathbf{r}_{t-1} | \mathcal{F}_{t-1}\right] + 8\kappa_+\eta^2\left((1-\eta)^2\left(L_1 M\right)^2 + L_0^2\right). \qquad (7)$$

Next, we show the bound by induction. For the base case $t = 0$, using a similar approach as the above:

$$\begin{aligned}
\mathbb{E}\left[\mathbf{r}_0\right] &= \mathbb{E}\left[\|\tfrac{2}{n}\textstyle\sum_{i=1}^{n/2}\left(\nabla f(x^0, z_i^0) - \nabla F_{\mathcal{D}}(x^0)\right)\|_+^2\right] \\
&\leqslant \tfrac{4}{n^2}\left(\mathbb{E}\left[\|\textstyle\sum_{i=1}^{n/2-1}\left(\nabla f(x^0, z_i^0) - \nabla F_{\mathcal{D}}(x^0)\right)\|_+^2\right] + \kappa_+\mathbb{E}\left[\|\nabla f(x^0, z_n^0/2) - \nabla F_{\mathcal{D}}(x^0)\|_+^2\right]\right) \\
&\leqslant \frac{4\kappa_+}{n^2}\sum_{i=1}^{n/2}\mathbb{E}\left[\|\nabla f(x^0, z_i^0) - \nabla F_{\mathcal{D}}(x^0)\|_+^2\right] \\
&\leqslant \frac{8\kappa L_0^2}{n}.
\end{aligned}$$

Let $C = 8\kappa_+\eta^2\left((1-\eta)^2\left(L_1 M\right)^2 + L_0^2\right)$. Now, for the inductive step use the recursion (7), to obtain

$$\begin{aligned}
\mathbb{E}\left[\mathbf{r}_t\right] &\leqslant (1-\eta)^{2t}\,\mathbb{E}\left[\mathbf{r}_0\right] + C\sum_{j=0}^{t-1}(1-\eta)^{2j} \\
&\leqslant \frac{8\kappa L_0^2}{n}(1-\eta)^{2t} + C\frac{1-(1-\eta)^{2t}}{\eta}.
\end{aligned}$$

Note that since $\eta \in (0,1)$, we have $(1-\eta)^{2t} > 1-2t\eta$. Hence, we get

$$\mathbb{E}[\|\mathbf{d}_t - \nabla F_{\mathcal{D}}(x^t)\|_+^2] \leqslant \frac{8\kappa L_0^2}{n}(1-\eta)^{2t} + 16\kappa_+\eta^2 t\left((1-\eta)^2(L_1M)^2 + L_0^2\right).$$

We obtain the result using the equivalence of norms and Jensen's inequality

$$\begin{aligned}\mathbb{E}[\|\mathbf{d}_t - \nabla F_{\mathcal{D}}(x^t)\|_*] &\leqslant \frac{2\sqrt{2\kappa}L_0}{\sqrt{n}}(1-\eta)^t + 4\eta\sqrt{\kappa_+ t}\,((1-\eta)L_1M + L_0)\\ &\leqslant \frac{2e\sqrt{2\log(K)}L_0}{\sqrt{n}}(1-\eta)^t + 4\eta\sqrt{\log(K)t}\,(L_1M + L_0),\end{aligned}$$

where in the last step we used upper bounds on κ and κ_+.

Appendix B. Missing proofs from Section 4

B.1. Part (a) of Proposition 8

Notice that for any m (below $\Gamma(\cdot)$ is the Gamma function),

$$\int_0^{+\infty} \exp\big\{-\frac{r^2}{2\sigma^2}\big\}r^m dr = \frac{(2\sigma^2)^{m+1/2}}{2}\int_0^{+\infty} e^{-u}u^{m-1/2}du = \frac{(2\sigma^2)^{(m+1)/2}}{2}\Gamma\big(\frac{m+1}{2}\big).$$

This implies that the m-th moment w.r.t. $\|\cdot\|_+$ can be computed as follows

$$\begin{aligned}\mathbb{E}[\|z\|_+^m] &= C(\sigma,d)\int_{\mathbb{R}^d}\|z\|_+^m\exp\Big\{-\frac{\|z\|_+^2}{2\sigma^2}\Big\}dz\\ &= C(\sigma,d)\text{Area}(\{\|x\|_+=1\})\int_0^{\infty} r^{m+d-1}\exp\big\{-\frac{r^2}{2\sigma^2}\big\}dr\\ &= C(\sigma,d)\text{Area}(\{\|x\|_+=1\})\frac{(2\sigma^2)^{(m+d)/2}}{2}\Gamma\big(\frac{m+d}{2}\big)\\ &= (2\sigma^2)^{m/2}\Gamma\big(\frac{m+d}{2}\big)/\Gamma\big(\frac{d}{2}\big).\end{aligned}$$

We conclude using that $\|z\|_* \leqslant \|z\|_+$. On the other hand, the bounds for the second moment are obtained by using that $\Gamma(1+d/2) = (d/2)\Gamma(d/2)$.

B.2. Missing RDP to DP reduction from Proposition 8, part (b)

The missing part of Proposition 8 is a consequence of the following result.

Corollary 16 *The generalized Gaussian mechanism applied to a function with $\|\cdot\|_*$-sensitivity bounded by $s>0$ is (α,ρ)-RDP, where $\rho = \kappa\alpha^2 s^2/[2\sigma^2(\alpha-1)]$. In particular, choosing $\sigma = 2\kappa\log(1/\delta)s^2/\varepsilon^2$, the generalized Gaussian mechanism is (ε,δ)-DP.*

Proof We first prove that the mechanism is (α,ρ)-RDP. For this, consider two neighboring datasets S, S'. Then, by the Rényi divergence estimate proved earlier in part (b)

$$D_\alpha\big(\mathcal{A}_{\|\cdot\|_+}(S)||\mathcal{A}_{\|\cdot\|_+}(S')\big) \leqslant \frac{\kappa\alpha^2 s^2}{2\sigma^2(\alpha-1)}$$

where we used that $\mathcal{A}$ has $\|\cdot\|_*$-sensitivity bounded by s. This proves the mechanism is (α,ρ)-RDP. The second part can be obtained from the first part, together with the DP/RDP reduction in (Mironov, 2017, Proposition 3). ■

Appendix C. Proof of Theorem 10

Proof of Theorem 10 First, this algorithm is (ε,δ)-DP by following known analyses of mini-batch SGD (see, e.g., (Bassily et al., 2019, Thm. 3.1)), together with the privacy guarantees for the GG mechanism in Proposition 8. Next, observe that by definition of $\|\cdot\|_+$ and by the duality between strong convexity and smoothness (Zalinescu, 1983), $\Phi(\cdot)$ is 1-strongly convex w.r.t. $\|\cdot\|$, and $\max_{x,y\in\mathcal{X}}[\Phi(x)-\Phi(y)] \leqslant \frac{\kappa M^2}{2}$, hence by the standard SMD analysis (Nemirovski et al., 2009) (here we use the fact that the minibatches are i.i.d. from the empirical distribution)

$$\mathcal{R}_S(\mathcal{A}_{\mathsf{noisySMD}}) \leqslant \frac{\kappa M^2}{2\eta T} + \frac{\eta L_0^2}{2}[1 + 16\frac{\kappa d T\log(1/\delta)}{(n\varepsilon)^2}].$$

We now use our choices $\eta = \frac{M}{L_0}\sqrt{\frac{\kappa}{2T}}$, and $T = \lfloor\frac{(\varepsilon n)^2}{16\kappa d\log(1/\delta)}\rfloor$, obtaining

$$\mathcal{R}_S(\mathcal{A}_{\mathsf{noisySMD}}) \leqslant 4L_0 M\cdot\frac{\kappa\sqrt{d\log(1/\delta)}}{\varepsilon n}.$$

To bound the excess population risk of the algorithm, we can use the generalization properties of differential privacy Bassily et al. (2016); Dwork et al. (2015) (see also a similar application that appeared earlier in (Bassily et al., 2014a, Lemma F.5)):

$$\mathcal{R}_{\mathcal{D}}[\mathcal{A}_{\mathsf{noisySMD}}] = \mathcal{R}_S[\mathcal{A}_{\mathsf{noisySMD}}] + L_0 M\cdot O(\varepsilon') = L_0 M\cdot O\left(\frac{\kappa\sqrt{d\log(1/\delta)}}{\varepsilon n} + \varepsilon'\right)$$

where, without loss of privacy, we replace ε in Algorithm $\mathcal{A}_{\mathsf{noisySMD}}$ with $\varepsilon' = \min\{\varepsilon, \frac{\sqrt{\kappa}d^{1/4}(\log(1/\delta))^{1/4}}{\sqrt{n}}\}$. Hence, we get

$$\mathcal{R}_{\mathcal{D}}[\mathcal{A}_{\mathsf{noisySMD}}] = L_0 M\cdot O\left(\max\left(\frac{\sqrt{\kappa}d^{1/4}\left(\log(1/\delta)\right)^{1/4}}{\sqrt{n}}, \frac{\kappa\sqrt{d\log(1/\delta)}}{\varepsilon n}\right)\right),$$

which proves the desired bound. Note that when $\varepsilon \geqslant \frac{\sqrt{\kappa}d^{1/4}(\log(1/\delta))^{1/4}}{\sqrt{n}}$, which subsumes the typical setting for the privacy parameter (i.e., $\varepsilon = \Theta(1)$), the above bound yields $L_0 M\cdot O\left(\frac{\sqrt{\kappa}d^{1/4}(\log(1/\delta))^{1/4}}{\sqrt{n}}\right)$. On the other hand, if $\varepsilon < \frac{\sqrt{\kappa}d^{1/4}(\log(1/\delta))^{1/4}}{\sqrt{n}}$, the bound becomes $L_0 M\cdot O\left(\frac{\kappa\sqrt{d\log(1/\delta)}}{\varepsilon n}\right)$, which, given our lower bound in Section 7, is essentially tight when $1+\Omega(1)\leqslant p<2$.

■

Appendix D. Proofs from Section 5.2

D.1. Proof of Theorem 11

Let $\nabla_0 = \frac{2}{n}\sum_{i=1}^{n/2}\nabla f(x^0, z_i^0)$ denote the initial gradient estimate. Note that the global $\|\cdot\|_*$-sensitivity of ∇_0 is bounded by $\frac{4L_0}{n}$. Hence, by Proposition 8 we obtain that Step 5 in Algorithm 3 is (ε,δ)-DP.

For iteration $t \in [\sqrt{n}]$, let B_t denote the mini-batch given in Step 10 in Algorithm 3, and let $\Delta_t = \frac{2}{\sqrt{n}}\sum_{i=1}^{\sqrt{n}/2}\left(\nabla f(x^t, z_i^t) - \nabla f(x^{t-1}, z_i^t)\right)$. Also, let B_t', Δ_t' denote the corresponding quantities in Algorithm $\mathcal{A}_{\mathsf{noisySFW}}$ when the input dataset is S'. Suppose that B_t and B_t' differ in at most one data point, say $z_{i^*} \neq z_{i^*}'$. Then

$$\|\Delta_t - \Delta_t'\|_* = \frac{2}{\sqrt{n}}\|\left(\nabla f(x^t, z_{i^*}) - \nabla f(x^{t-1}, z_{i^*})\right) - \left(\nabla f(x^t, z_{i^*}') - \nabla f(x^{t-1}, z_{i^*}')\right)\|_*$$

Hence, by the smoothness of f w.r.t. $\|\cdot\|$, the global $\|\cdot\|_*$ sensitivity of Δ_t is bounded by $\frac{4\eta L_1 M}{\sqrt{n}}$. Again, using Proposition 8 we have that Step 11 in Algorithm 3 is (ε,δ)-DP. Note that at any given iteration t, the gradient estimate $\widetilde{\nabla}_{t-1}$ from the previous iteration is already computed privately. Since differential privacy is closed under post-processing, the current iteration t is (ε,δ)-DP. Since the batches of the dataset used in different iterations are disjoint, then by parallel composition, Algorithm $\mathcal{A}_{\mathsf{noisySFW}}$ is (ε,δ)-differentially private.

D.2. Proof of Theorem 12

We start by proving a recursive bound on the first moment of the gradient estimator.

Lemma 17 *Let $\mathcal{D}$ be a distribution over $\mathcal{Z}$, and $S \sim \mathcal{D}^n$ be the input to Algorithm $\mathcal{A}_{\mathsf{noisySFW}}$. For $t \in [0, \sqrt{n}]$ and $\kappa = \min\left\{\frac{1}{p-1}, e^2\ln(d)\right\}$, the recursive gradient estimate $\widetilde{\nabla}_t$ satisfies*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n}\left[\|\nabla F_{\mathcal{D}}(x^t) - \widetilde{\nabla}_t\|_*\right] \leqslant 4L_0\sqrt{\frac{\kappa}{n}} + \frac{4\kappa L_1 M\eta\sqrt{t}}{n^{1/4}} + \frac{8\,\kappa\sqrt{d\log(1/\delta)}}{\varepsilon}\left(\frac{L_0}{n} + L_1 M\eta\sqrt{\frac{t}{n}}\right).$$

Proof Consider any iteration $t \geqslant 1$ of $\mathcal{A}_{\mathsf{noisySFW}}$. Similar to the proof of Lemma 6, we first show the second moment bound for the smooth norm $\|\cdot\|_+$ associated with the dual norm, and then we will conclude by the equivalence property (2). Let $\mathbf{r}_t \triangleq \|\nabla F_{\mathcal{D}}(x^t) - \widetilde{\nabla}_t\|_+^2$. Then

$$\mathbf{r}_t = \|\nabla F_{\mathcal{D}}(x^{t-1}) - \widetilde{\nabla}_{t-1} + \nabla F_{\mathcal{D}}(x^t) - \nabla F_{\mathcal{D}}(x^{t-1}) - \tilde{\Delta}_t\|_+^2$$

Let $\mathcal{F}_t$ be the σ-algebra induced by the randomness in the data i.e. the mini-batches $B_0, B_1, \ldots, B_t$ and the the noise vectors $\mathbf{g}_0, \mathbf{g}_1, \ldots, \mathbf{g}_t$ up until iteration t. Note that conditioned on $\mathcal{F}_{t-1}$, $\mathbb{E}\left[\tilde{\Delta}_t|\mathcal{F}_{t-1}\right] = \nabla F_{\mathcal{D}}(x^t) - \nabla F_{\mathcal{D}}(x^{t-1})$. Now, let $\Delta_t = \frac{2}{\sqrt{n}}\sum_{i=1}^{\sqrt{n}/2}\left(\nabla f(x^t, z_i^t) - \nabla f(x^{t-1}, z_i^t)\right)$. Then $\tilde{\Delta}_t = \Delta_t + \mathbf{g}_t$. Thus, by property (1), we observe that

$$\begin{aligned}\mathbb{E}\left[\mathbf{r}_t|\mathcal{F}_{t-1}\right] &\leqslant \mathbb{E}\left[\|\nabla F_{\mathcal{D}}(x^{t-1}) - \widetilde{\nabla}_{t-1}\|_+^2|\mathcal{F}_{t-1}\right] + \kappa_+\mathbb{E}\left[\|\nabla F_{\mathcal{D}}(x^t) - \nabla F_{\mathcal{D}}(x^{t-1}) - \tilde{\Delta}_t\|_+^2|\mathcal{F}_{t-1}\right]\\ &\leqslant \mathbb{E}\left[\mathbf{r}_{t-1}|\mathcal{F}_{t-1}\right] + 2\kappa_+\mathbb{E}\left[\|\nabla F_{\mathcal{D}}(x^t) - \nabla F_{\mathcal{D}}(x^{t-1}) - \Delta_t\|_+^2|\mathcal{F}_{t-1}\right] + 2\kappa_+\mathbb{E}\left[\|\mathbf{g}_t\|_+^2|\mathcal{F}_{t-1}\right].\end{aligned}$$

Using a similar reasoning as given in the proof of Lemma 6, we can show that

$$\begin{aligned}\mathbb{E}\left[\|\nabla F_{\mathcal{D}}(x^t) - \nabla F_{\mathcal{D}}(x^{t-1}) - \Delta_t\|_+^2 | \mathcal{F}_{t-1}\right] &\leqslant \frac{\kappa}{\kappa_+} \\ &\leqslant \frac{8\kappa L_1^2 M^2 \eta^2}{\sqrt{n}}.\end{aligned}$$

Also, by Proposition 8 we have $\mathbb{E}\left[\|\mathbf{g}_t\|_+^2\right] \leqslant \sigma^2 d$, where $\sigma^2 = \frac{32\kappa L_1^2 M^2 \eta^2 \log(1/\delta)}{n\varepsilon^2}$. Thus, given the fact that $\kappa_+ \leqslant \kappa$, we have the following recursion

$$\mathbb{E}\left[\mathbf{r}_t | \mathcal{F}_{t-1}\right] \leqslant \mathbb{E}\left[\mathbf{r}_{t-1} | \mathcal{F}_{t-1}\right] + \frac{16\kappa^2 L_1^2 M^2 \eta^2}{\sqrt{n}} + \frac{64\kappa^2 L_1^2 M^2 \eta^2 d \log(1/\delta)}{n\varepsilon^2}. \tag{8}$$

We proceed by induction: using the same approach as before, for the base case $t = 0$ and noise variance $\sigma_0^2 = \frac{32\kappa L_0^2 \log(1/\delta)}{n^2\varepsilon^2}$ we have

$$\begin{aligned}\mathbb{E}\left[\mathbf{r}_0\right] &= \mathbb{E}\left[\|\nabla F_{\mathcal{D}}(x^0) - \widetilde{\nabla}_0\|_+^2\right] \\ &\leqslant \frac{8\kappa_+}{n^2} \sum_{i=1}^{n/2} \mathbb{E}\left[\|\nabla F_{\mathcal{D}}(x^0) - \nabla f(x^0, z_i^0)\|_+^2\right] + 2\kappa_+ \mathbb{E}\left[\|\mathbf{g}_0\|_+^2\right] \\ &\leqslant \frac{16\kappa L_0^2}{n} + \frac{64\kappa^2 L_0^2 d \log(1/\delta)}{n^2\varepsilon^2}.\end{aligned}$$

Thus, by induction on (8) we obtain

$$\begin{aligned}\mathbb{E}\left[\mathbf{r}_t\right] &\leqslant \mathbb{E}\left[\mathbf{r}_0\right] + \frac{16\kappa^2 L_1^2 M^2 \eta^2 t}{\sqrt{n}} + \frac{64\kappa^2 L_1^2 M^2 t \eta^2 d \log(1/\delta)}{n\varepsilon^2} \\ &\leqslant \frac{16\kappa L_0^2}{n} + \frac{64\kappa^2 L_0^2 d \log(1/\delta)}{n^2\varepsilon^2} + \frac{16\kappa^2 L_1^2 M^2 \eta^2 t}{\sqrt{n}} + \frac{64\kappa^2 L_1^2 M^2 \eta^2 t d \log(1/\delta)}{n\varepsilon^2}.\end{aligned}$$

Therefore, by the equivalence of the norms and the Jensen's inequality

$$\mathbb{E}[\|\nabla F_{\mathcal{D}}(x^t) - \widetilde{\nabla}_t\|_*] \leqslant 4L_0\sqrt{\frac{\kappa}{n}} + \frac{4\kappa L_1 M \eta \sqrt{t}}{n^{1/4}} + \frac{8\,\kappa\sqrt{d\log(1/\delta)}}{\varepsilon}\left(\frac{L_0}{n} + L_1 M \eta \sqrt{\frac{t}{n}}\right).$$

■

Proof of Theorem 12 For $t \in [0, \sqrt{n}]$, by following a similar argument as used in the proof of Theorem 4, we have

$$F_{\mathcal{D}}(x^{t+1}) - F_{\mathcal{D}}^* \leqslant (1-\eta)\left(F_{\mathcal{D}}(x^t) - F_{\mathcal{D}}^*\right) + 2\eta M\|\nabla F_{\mathcal{D}}(x^t) - \widetilde{\nabla}_t\|_* + \frac{L_1\eta^2 M^2}{2}$$

Let $\Gamma_t = F_{\mathcal{D}}(x^t) - F_{\mathcal{D}}^*$, and let $\mathbf{q}_t = \|\nabla F_{\mathcal{D}}(x^t) - \widetilde{\nabla}_t\|_*$. Hence, taking expectation we get

$$\mathbb{E}\left[\Gamma_{t+1}\right] \leqslant (1-\eta)\mathbb{E}\left[\Gamma_t\right] + 2\eta M \mathbb{E}\left[\mathbf{q}_t\right] + \frac{L_1\eta^2 M^2}{2}$$

Thus, for $\sqrt{n}$ iterations we have

$$\mathbb{E}\left[\Gamma_{\sqrt{n}+1}\right] \leqslant (1-\eta)^{\sqrt{n}+1} L_0 M + 2\eta M \sum_{j=0}^{\sqrt{n}} (1-\eta)^j \mathbb{E}\left[\mathbf{q}_{\sqrt{n}-j}\right] + \frac{L_1 \eta M^2}{2}. \tag{9}$$

By Lemma 17 observe that

$$\begin{aligned}
\sum_{j=0}^{\sqrt{n}} (1-\eta)^j \mathbb{E}\left[\mathbf{q}_{\sqrt{n}-j}\right] \leqslant \frac{1}{\eta}&\left(4L_0\sqrt{\frac{\kappa}{n}} + \frac{8\,\kappa L_0 \sqrt{d\log(1/\delta)}}{\varepsilon\, n}\right) \\
&+ 4L_1 M\eta\left(\frac{\sqrt{\kappa}}{n^{1/4}} + \frac{2\kappa\sqrt{d\log(1/\delta)}}{\varepsilon\sqrt{n}}\right)\, n^{1/4}\sum_{j=0}^{\sqrt{n}}(1-\eta)^j \\
\leqslant \frac{1}{\eta}&\left(4L_0\sqrt{\frac{\kappa}{n}} + \frac{8\,\kappa L_0 \sqrt{d\log(1/\delta)}}{\varepsilon\, n}\right) \\
&+ 4L_1 M\left(\sqrt{\kappa} + \frac{2\kappa\sqrt{d\log(1/\delta)}}{\varepsilon\, n^{1/4}}\right).
\end{aligned}$$

Substituting this in (9) and setting $\eta = \frac{\log(n)}{2\sqrt{n}}$, we get

$$\begin{aligned}
\mathbb{E}\left[\Gamma_{\sqrt{n}+1}\right] \leqslant \frac{L_0 M}{\sqrt{n}} + 8L_0 M\sqrt{\frac{\kappa}{n}} + \frac{16\kappa L_0 M\sqrt{d\log(1/\delta)}}{\varepsilon\, n} + 4L_1 M^2\log(n)\sqrt{\frac{\kappa}{n}} \\
+ \frac{16\kappa L_1 M^2 \log(n)\sqrt{d\log(1/\delta)}}{\varepsilon\, n^{3/4}} + \frac{L_1 M^2\log(n)}{4\sqrt{n}}
\end{aligned}$$

Hence, the expected excess risk is

$$\mathcal{R}_{\mathcal{D}}[\mathcal{A}_{\mathsf{noisySFW}}] = O\left(\frac{L_1 M^2\log(n) + L_0 M\kappa}{\sqrt{n}} + \frac{\kappa L_1 M^2\log(n)\sqrt{d\log(1/\delta)}}{\varepsilon\, n^{3/4}}\right).$$

■

Appendix E. Proof of Proposition 13 from Section 6

We start by bounding the $\|\cdot\|_2$-diameter and Lipschitz constant for the ℓ_p-setup. First, since the $\|\cdot\|_p$-diameter of $\mathcal{X}$ is bounded by M, then the $\|\cdot\|_2$-diameter of $\mathcal{X}$ is bounded by $d^{1/2-1/p}M$. Next, if f is L_0-Lipschitz w.r.t. $\|\cdot\|_p$, i.e. $\bigcup_{x\in\mathcal{X}}\partial f(x) \subseteq \mathcal{B}_{\|\cdot\|_{p*}}(0, L_0)$, then f is also L_0-Lipschitz w.r.t. $\|\cdot\|_2$. Therefore, (Bassily et al., 2020, Remark 5.3) implies

$$\begin{aligned}
\mathcal{R}_{\mathcal{D}}[\mathcal{A}_{\mathsf{noisySGD}}] &= d^{\frac{1}{2}-\frac{1}{p}} L_0 M \cdot O\Big(\max\Big\{\tfrac{1}{\sqrt{n}}, \tfrac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\Big\}\Big) \\
&= L_0 M \cdot O\Big(\tfrac{d^{\frac{1}{2}-\frac{1}{p}}}{\sqrt{n}} + \tfrac{d^{1-\frac{1}{p}}\sqrt{\log(1/\delta)}}{\varepsilon n}\Big).
\end{aligned}$$

Appendix F. Full Details of the Lower Bound in Section 7

In this section, we give the full details for our lower bounds on the excess risk of DP-SCO and DP-ERM in the ℓ_p setup when $1 < p < 2$. Our lower bound for DP-SCO is given in Theorem 18. The first term follows directly from the non-private lower bound for SCO in the same setting (Nemirovski and Yudin, 1983). To establish a lower bound of $\tilde{\Omega}((p-1)\sqrt{d}/(\varepsilon n))$, we show a lower bound of essentially the same order (up to a logarithmic factor in $1/\delta$) on the excess empirical error for DP-ERM in the ℓ_p setting (Theorem 19). Given the reduction in (Bassily et al., 2019, Appendix C), this implies the claimed lower bound on DP-SCO.

Problem setup: Let $p \in (1,2)$ and $d \in \mathbb{N}$. Let $\mathcal{X} = \mathcal{B}_p^d$, where $\mathcal{B}_p^d$ is the unit ℓ_p ball in $\mathbb{R}^d$, and let $\mathcal{Z} = \{-\frac{1}{d^{1/q}}, \frac{1}{d^{1/q}}\}^d$ where $q = \frac{p}{p-1}$. Let $f : \mathcal{X} \times \mathcal{Z} \to [-1, 1]$ defined as:

$$f(x,z) = -\langle x, z\rangle,\ x \in \mathcal{X}, z \in \mathcal{Z}.$$

Note that for every $z \in \mathcal{Z}$, $f(\cdot, z)$ is convex, smooth, and 1-Lipschitz w.r.t. $\|\cdot\|_p$ over $\mathcal{X}$. Recall that for any distribution $\mathcal{D}$ over $\mathcal{Z}$, we define the population risk of $x \in \mathcal{X}$ w.r.t. $\mathcal{D}$ as $F_{\mathcal{D}}(x) \triangleq \mathop{\mathbb{E}}_{z\sim\mathcal{D}}[f(x,z)]$, and for any dataset $S = (z_1, \ldots, z_n) \in \mathcal{Z}^n$, we define the empirical risk of $x \in \mathcal{X}$ w.r.t. S as $F_S(x) \triangleq \frac{1}{n}\sum_{i=1}^n f(x, z_i)$.

Our lower bound for DP-SCO is formally stated in the following theorem.

Theorem 18 *Let $p \in (1,2)$ and $n, d \in \mathbb{N}$. Let $\varepsilon > 0$ and $0 < \delta < \frac{1}{n^{1+\Omega(1)}}$. Let $\mathcal{X}, \mathcal{Z}$, and f be as defined in the setup above. There exists a distribution $\mathcal{D}$ over $\mathcal{Z}$ such that for any (ε, δ)-DP-SCO algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{X}$, we have*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n,\mathcal{A}}[F_{\mathcal{D}}(\mathcal{A}(S))] - \min_{x\in\mathcal{X}} F_{\mathcal{D}}(x) = \tilde{\Omega}\left(\max\left(\frac{1}{\sqrt{n}}, (p-1)\frac{\sqrt{d}}{\varepsilon n}\right)\right).$$

As mentioned earlier, given the reduction described in Bassily et al. (2019), it suffices to prove a lower bound of essentially the same order for DP-ERM w.r.t. the problem described above. The remainder of this section will be devoted to this goal. Namely, we will prove the following theorem.

Theorem 19 *Under the same setup in Theorem 18, there exists a dataset $S \in \mathcal{Z}^n$ such that for any (ε, δ)-DP-ERM algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{X}$, we have*

$$\mathop{\mathbb{E}}_{\mathcal{A}}[F_S(\mathcal{A}(S))] - \min_{x\in\mathcal{X}} F_S(x) = \Omega\left((p-1)\frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\right).$$

Proof Let the spaces $\mathcal{X}, \mathcal{Z}$, and the loss function f be as defined in the problem setup above. For any $x \in \mathcal{X}$, let x_j denote the j-th coordinate of x, where $j \in [d]$. Let $S = (z_1, \ldots, z_n) \in \mathcal{Z}^n$, and let z_{ij} denote the j-th coordinate of z_i, where $i \in [n], j \in [d]$. Define $\bar{z} \triangleq \frac{1}{n}\sum_{i=1}^n z_i$, and similarly, let $\bar{z}_j$ denote the j-th coordinate of $\bar{z}$.

Let $x^* \triangleq \arg\min_{x\in\mathcal{X}} F_S(x) = \arg\max_{x\in\mathcal{X}}\langle x, \bar{z}\rangle$. Note that we have

$$x_j^* = \frac{|\bar{z}_j|^{q-1}}{\|\bar{z}\|_q^{q-1}}\mathsf{sign}(\bar{z}_j), \quad j \in [d] \tag{10}$$

where $\|\cdot\|_q$ denote the ℓ_q norm (recall that $q \triangleq \frac{p}{p-1}$). To see this, note that by Hölder's inequality $\forall x \in \mathcal{X}$, $F_S(x) \geqslant -\|\bar{z}\|_q$, and on the other hand, note that $x^* \in \mathcal{X}$ since $\|x^*\|_p = 1$ and $F_S(x^*) = -\|\bar{z}\|_q$. Next, we make the following claim.

Claim 20 *Let $\alpha > 0$. Let $\widehat{x} \in \mathcal{X}$ be such that $F_S(\widehat{x}) - F_S(x^*) \leqslant \alpha$. Then, $\|\widehat{x} - x^*\|_p \leqslant \sqrt{\frac{8\alpha}{(p-1)\|\bar{z}\|_q}}$.*

The proof of this claim relies on the uniform convexity property of the ℓ_p norms for $p \in (1,2]$ (see Ball et al. (1994)). We formally restate this property below:

Fact 21 (see Eq. (1.6) in (Ball et al., 1994)) *Let x, y be any elements of an ℓ_p-normed space $(\mathcal{X}, \|\cdot\|_p)$, where $1 < p \leqslant 2$. We have $\|\frac{x+y}{2}\|_p \leqslant 1 - \frac{p-1}{8}\|x-y\|_p^2$.*

Now, observe that for any $\widehat{x} \in \mathcal{X}$ such that $F_S(\widehat{x}) - F_S(x^*) \leqslant \alpha$, we have

$$1 - \frac{\alpha}{2\|\bar{z}\|_q} \leqslant \langle \frac{\widehat{x} + x^*}{2}, \frac{\bar{z}}{\|\bar{z}\|_q} \rangle \leqslant \|\frac{\widehat{x} + x^*}{2}\|_p \leqslant 1 - \frac{p-1}{8}\|\widehat{x} - x^*\|_p^2,$$

where the last inequality follows from the above fact. Rearranging terms lead to the above claim.

Fix values for ε and δ as in the theorem statement. Next, we will show the existence of a dataset $S = (z_1, \ldots, z_n) \in \mathcal{Z}^n$ with $\|\bar{z}\|_q = \Omega\left(\frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\right)$ such that for any (ε, δ)-DP-ERM algorithm for the above problem that outputs a vector $\widehat{x} \in \mathcal{X}$, we must have $\|\widehat{x} - x^*\|_p = \Omega(1)$ with probability $2/3$ over the algorithm's random coins. Note that, by Claim 20, this implies the desired lower bound. To see this, suppose, for the sake of a contradiction, that there exists an (ε, δ)-DP-ERM algorithm $\mathcal{A}$ that outputs $\widehat{x} \in \mathcal{X}$ such that $\mathbb{E}_{\widehat{x} \leftarrow \mathcal{A}}[F_S(\widehat{x})] - F_S(x^*) = o\left((p-1)\frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\right)$. Then, by Markov's inequality, with probability $\geqslant 0.9$, we have $F_S(\widehat{x}) - F_S(x^*) = o\left((p-1)\frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\right)$. Hence, Claim 20 would imply that, with probability $\geqslant 0.9$, $\|\widehat{x} - x^*\|_p = o(1)$, which contradicts with the claimed $\Omega(1)$ lower bound on $\|\widehat{x} - x^*\|$. Hence, to conclude the proof of Theorem 19, it remains to show the claimed lower bound on $\|\widehat{x} - x^*\|$, which we do next.

In the final step of the proof, we resort to a construction based on the fingerprinting code argument due to Bun et al. (2018). We use the following lemma, which is implicit in the constructions of Bun et al. (2015); Steinke and Ullman (2015).

Lemma 22 *Let $n, d \in \mathbb{N}$. Let $\varepsilon > 0$ and $0 < \delta < \frac{1}{n^{1+\Omega(1)}}$. Let $\mathcal{T} = \{-1,1\}^d$. There exists a dataset $T = (v_1, \ldots, v_n) \in \mathcal{T}^n$ where $\|\frac{1}{n}\sum_{i=1}^n v_i\|_\infty \leqslant c\frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}$ for some universal constant $c > 0$ such that for any (ε, δ)-differentially private algorithm $\mathcal{M} : \mathcal{T}^n \to [-1,1]^d$, the following is true with probability $2/3$ over the random coins of $\mathcal{M}$: $\exists J \subseteq [d]$ with $|J| = \Omega(d)$ such that*

$$(\forall j \in J), \quad \left|\mathcal{M}_j(T) - \frac{1}{n}\sum_{i=1}^n v_{ij}\right| = \Omega\Big(\frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\Big) \quad \textit{and} \quad \left|\sum_{i=1}^n v_{ij}\right| = c \cdot \frac{\sqrt{d\log(1/\delta)}}{\varepsilon n},$$

where $\mathcal{M}_j(T)$ denotes the j-th coordinate of $\mathcal{M}(T)$ and v_{ij} denotes the j-th coordinate of v_i.

We consider a normalized version of the dataset $T = (v_1, \ldots, v_n)$ in the above lemma. Namely, we consider a dataset $S = (z_1, \ldots, z_n) \in \mathcal{Z}^n$, where $z_i = \frac{v_i}{d^{1/q}}$, $i \in [n]$. Note that the above lemma implies the existence of a subset $J \subseteq [d]$ with $|J| = \Omega(d)$ such that for all $j \in J$, $|\bar{z}_j| = \frac{c}{d^{1/q}} \cdot \frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}$ for some universal constant $c > 0$. Note also that $\forall\ j \in [d]\backslash J,\ |\bar{z}_j| \leqslant \frac{c}{d^{1/q}} \cdot \frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}$ since $\|\frac{1}{n}\sum_{i=1}^{n} v_i\|_\infty \leqslant c\frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}$. This implies that

$$\frac{|J|}{d}\, c^q \left(\frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\right)^q \leqslant \|\bar{z}\|_q^q \leqslant c^q \left(\frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\right)^q,$$

which, given the fact that $|J| = \Omega(d)$, implies that $\|\bar{z}\|_q^q = c'\frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}$ for some universal constant $c' > 0$. Hence, by the fact that $q > 2$, we have $\|\bar{z}\|_q = \Theta\left(\frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\right)$. Moreover, note that for all $j \in [J]$, we have $\frac{|\bar{z}_j|^{q-1}}{\|\bar{z}\|_q^{q-1}} = \frac{\left(\frac{c}{c'}\right)^{1-\frac{1}{q}}}{d^{1-\frac{1}{q}}} = \frac{c''}{d^{1/p}}$ for some universal constant c'', where the last equality follows from the fact that $q > 2$ and $\frac{1}{q} = 1 - \frac{1}{p}$.

Let $\bar{v} \triangleq \frac{1}{n}\sum_{i=1}^{n} v_i$, and let $\bar{v}_j$ denote the j-th coordinate of $\bar{v}$ for $j \in [d]$. Given the above observations and the expression of the minimizer x^* in eq. (10), it is not hard to see that for all $j \in J$,

$$x_j^* = \frac{c''}{d^{1/p}}\mathsf{sign}(\bar{v}_j) = \frac{c''}{d^{1/p}} \cdot \frac{\bar{v}_j}{|\bar{v}_j|} = \frac{c''}{c} \cdot \frac{\varepsilon n}{d^{1/2+1/p}\sqrt{\log(1/\delta)}}\bar{v}_j. \tag{11}$$

Let $\mathcal{A}$ be any (ε, δ)-DP-ERM algorithm that takes the dataset S described above as input, and let $\widehat{x} \in \mathcal{X}$ denote its output. Construct an (ε, δ)-differentially private algorithm $\mathcal{M}$ for the dataset T of Lemma 22 by first running $\mathcal{A}$ on $S = \frac{1}{d^{1/q}} \cdot T$, which outputs $\widehat{x}$, then releasing $\mathcal{M}(T) = \frac{c}{c''} \cdot \frac{d^{1/2+1/p}\sqrt{\log(1/\delta)}}{\varepsilon\, n} \cdot \widehat{x}$. Now, using (11) and given the description of $\mathcal{M}$, observe that

$$\begin{aligned}
\|\widehat{x} - x^*\|_p &= \tfrac{c''}{c} \cdot \tfrac{\varepsilon n}{d^{1/2+1/p}\sqrt{\log(1/\delta)}} \cdot \|\mathcal{M}(T) - \bar{v}\|_p \\
&\geqslant \tfrac{c''}{c} \cdot \tfrac{\varepsilon n}{d^{1/2+1/p}\sqrt{\log(1/\delta)}} \cdot \left(\textstyle\sum_{j\in J} |\mathcal{M}_j(T) - \bar{v}_j|^p\right)^{1/p} \\
&= \Omega\left(\tfrac{\varepsilon n}{d^{1/2+1/p}\sqrt{\log(1/\delta)}}\, d^{1/p}\, \tfrac{\sqrt{d\log(1/\delta)}}{\varepsilon\, n}\right) \\
&= \Omega(1)
\end{aligned}$$

where the third step follows from Lemma 22 and the fact that $\mathcal{M}$ is (ε, δ)-differentially private. This establishes the desired lower bound on $\|\widehat{x} - x^*\|_p$, and hence by the argument described earlier, the proof of Theorem 19 is now complete. ■