Adapting to Function Difficulty and Growth Conditions in Private Optimization

Hilal Asi* **Daniel Levy*** **John C. Duchi**
{asi,danilevy,jduchi}@stanford.edu

Abstract

We develop algorithms for private stochastic convex optimization that adapt to the hardness of the specific function we wish to optimize. While previous work provide worst-case bounds for arbitrary convex functions, it is often the case that the function at hand belongs to a smaller class that enjoys faster rates. Concretely, we show that for functions exhibiting κ-growth around the optimum, i.e., $f(x) \geq f(x^\star) + \lambda \kappa^{-1} \|x - x^\star\|_2^\kappa$ for $\kappa > 1$, our algorithms improve upon the standard $\sqrt{d}/n\varepsilon$ privacy rate to the faster $(\sqrt{d}/n\varepsilon)^{\frac{\kappa}{\kappa-1}}$. Crucially, they achieve these rates without knowledge of the growth constant κ of the function. Our algorithms build upon the inverse sensitivity mechanism, which adapts to instance difficulty [2], and recent localization techniques in private optimization [25]. We complement our algorithms with matching lower bounds for these function classes and demonstrate that our adaptive algorithm is *simultaneously* (minimax) optimal over all $\kappa \geq 1 + c$ whenever $c = \Theta(1)$.

1 Introduction

Stochastic convex optimization (SCO) is a central problem in machine learning and statistics, where for a sample space $\mathbb{S}$, parameter space $\mathcal{X} \subset \mathbb{R}^d$, and a collection of convex losses $\{F(\cdot; s) : s \in \mathbb{S}\}$, one wishes to solve

$$\minimize_{x \in \mathcal{X}} f(x) := \mathbb{E}_{S \sim P}[F(x; S)] = \int_{\mathbb{S}} F(x; s) \mathrm{d}P(s) \tag{1}$$

using an observed dataset $\mathcal{S} = S_1^n \overset{\text{iid}}{\sim} P$. While as formulated, the problem is by now fairly well-understood [12, 38, 29, 10, 37], it is becoming clear that, because of considerations beyond pure statistical accuracy—memory or communication costs [45, 26, 13], fairness [23, 28], personalization or distributed learning [35]—problem (1) is simply insufficient to address modern learning problems. To that end, researchers have revisited SCO under the additional constraint that the solution preserves the privacy of the provided sample [22, 21, 1, 16, 19]. A waypoint is Bassily et al. [7], who provide a private method with optimal convergence rates for the related empirical risk minimization problem, with recent papers focus on SCO providing (worst-case) optimal rates in various settings: smooth convex functions [8, 25], non-smooth functions [9], non-Euclidean geometry [5, 4] and under more stringent privacy constraints [34].

Yet these works ground their analyses in worst-case scenarios and provide guarantees for the *hardest* instance of the class of problems they consider. Conversely, they argue that their algorithms are optimal in a minimax sense: for any algorithm, there exists a hard instance on which the error achieved by the algorithm is equal to the upper bound. While valuable, these results are pessimistic—the exhibited hard instances are typically pathological—and fail to reflect achievable performance.

*Equal contribution. Author order determined by coin toss.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

In this work, we consider the problem of adaptivity when solving (1) under privacy constraints. Importantly, we wish to provide private algorithms that *adapt* to the hardness of the objective f. A loss function f may belong to multiple problem classes, each exhibiting different achievable rates, so a natural desideratum is to attain the error rate of the easiest sub-class. As a simple vignette, if one gets an arbitrary 1-Lipschitz convex loss function f, the worst-case guarantee of any ε-DP algorithm is $\Theta(1/\sqrt{n} + d/(n\varepsilon))$. However, if one learns that f exhibits some growth property—say f is 1-strongly convex—the regret guarantee improves to the faster $\Theta(1/n + (d/(n\varepsilon))^2)$ rate with the appropriate algorithm. It is thus important to provide algorithms that achieves the rates of the "easiest" class to which the function belongs [32, 46, 18].

To that end, consider the nested classes of functions $\mathcal{F}^\kappa$ for $\kappa \in [1, \infty]$ such that, if $f \in \mathcal{F}^\kappa$ then there exists $\lambda > 0$ such that for all $x \in \mathcal{X}$,

$$f(x) - \inf_{x' \in \mathcal{X}} f(x') \geq \frac{\lambda}{\kappa} \|x - x^\star\|_2^\kappa.$$

For example, strong convexity implies growth with parameter $\kappa = 2$. This growth assumption closely relates to uniform convexity [32] and the Polyak-Kurdyka-Łojasiewicz inequality [11], and we make these connections precise in Section 2. Intuitively, smaller κ makes the function much easier to optimize: the error around the optimal point grows quickly. Objectives with growth are widespread in machine learning applications: among others, the ℓ_1-regularized hinge loss exhibits sharp growth (i.e. $\kappa = 1$) while ℓ_1- or ℓ_∞-constrained κ-norm regression —i.e. $s = (a, b) \in \mathbb{R}^d \times \mathbb{R}$ and $F(x; s) = |b - \langle a, x\rangle|^\kappa$—has κ-growth for any κ integer greater than 2 [43]. In this work, we provide private adaptive algorithms that adapt to the *actual* growth of the function at hand.

We begin our analysis by examining Asi and Duchi's inverse sensitivity mechanism [2] on ERM as a motivation. While not a practical algorithm, it achieves instance-optimal rates for any one-dimensional function under mild assumptions, quantifying the best bound one could hope to achieve with an adaptive algorithm, and showing (in principle) that adaptive private algorithms can exist. We first show that for any function with κ-growth, the inverse sensitivity mechanism achieves privacy cost $(d/(n\varepsilon))^{\kappa/(\kappa-1)}$; importantly, *without knowledge of the function class $\mathcal{F}^\kappa$, that f belongs to*. This constitutes grounding and motivation for our work in three ways: (i) it validates our choice of sub-classes $\mathcal{F}^\kappa$ as the privacy rate is effectively controlled by the value of κ, (ii) it exhibits the rate we wish to achieve with efficient algorithms on $\mathcal{F}^\kappa$ and (iii) it showcases that for easier functions, privacy costs shrink significantly—to illustrate, for $\kappa = 5/4$ the privacy rate becomes $(d/(n\varepsilon))^5$.

We continue our treatment of problem (1) under growth in Section 4 and develop practical algorithms that achieve the rates of the inverse sensitivity mechanism. Moreover, for approximate (ε, δ)-differential privacy, our algorithms improve the rates, achieving roughly $(\sqrt{d}/(n\varepsilon))^{\kappa/(\kappa-1)}$. Our algorithms hinge on a reduction to SCO: we show that by solving a sequence of increasingly constrained SCO problems, one achieves the right rate whenever the function exhibits growth at the optimum. Importantly, our algorithm only requires a *lower bound* $\underline{\kappa} \leq \kappa$ (where κ is the actual growth of f).

We provide optimality guarantees for our algorithms in Section 5 and show that both the inverse sensitivity and the efficient algorithms of Section 4 are *simultaneously minimax optimal* over all classes $\mathcal{F}^\kappa$ whenever $\kappa = 1 + \Theta(1)$ and $d = 1$ for ε-DP algorithms. Finally, we prove that in *arbitrary dimension*, for both pure- and approximate-DP constraints, our algorithms are also simultaneously optimal for all classes $\mathcal{F}^\kappa$ with $\kappa \geq 2$.

On the way, we provide results that may be of independent interest to the community. First, we develop optimal algorithms for SCO under *pure* differential privacy constraints, which, to the best of our knowledge, do not exist in the literature. Secondly, our algorithms and analysis provide high-probability bounds on the loss, whereas existing results only provide (weaker) bounds on the expected loss. Finally, we complete the results of Ramdas and Singh [40] on (non-private) optimization lower bounds for functions with κ-growth by providing information-theoretic lower bounds (in contrast to oracle-based lower bounds that rely on observing only gradient information) and capturing the optimal dependence on all problem parameters (namely d, L and λ).

1.1 Related work

Convex optimization is one of the best studied problems in private data analysis [16, 19, 41, 7]. The first papers in this line of work mainly study minimizing the empirical loss, and readily establish that

the (minimax) optimal privacy rates are $d/n\varepsilon$ for pure ε-DP and $\sqrt{d\log(1/\delta)}/n\varepsilon$ for (ε,δ)-DP [16, 7]. More recently, several works instead consider the harder problem of privately minimizing the population loss [8, 25]. These papers introduce new algorithmic techniques to obtain the worst-case optimal rates of $1/\sqrt{n}+\sqrt{d\log(1/\delta)}/n\varepsilon$ for (ε,δ)-DP. They also show how to improve this rate to the faster $1/n+d\log(1/\delta)/(n\varepsilon)^2$ in the case of 1-strongly convex functions. Our work subsumes both of these results as they correspond to $\kappa=\infty$ and $\kappa=2$ respectively. To the best of our knowledge, there has been no work in private optimization that investigates the rates under general κ-growth assumptions or adaptivity to such conditions.

In contrast, the optimization community has extensively studied growth assumptions [40, 32, 15] and show that on these problems, carefully crafted algorithms improves upon the standard $1/\sqrt{n}$ for convex functions to the faster $(1/\sqrt{n})^{\kappa/(\kappa-1)}$. [32] derives worst-case optimal (in the first-order oracle model) gradient algorithms in the uniformly convex case (i.e. $\kappa\geq 2$) and provides technique to adapt to the growth κ, while [40], drawing connections between growth conditions and active learning, provides upper and lower bounds in the first-order stochastic oracle model. We complete the results of the latter and provide *information-theoretic* lower bounds that have optimal dependence on d, λ and n—their lower bound only holding for λ inversely proportional to $d^{1/2-1/\kappa}$, when $\kappa\geq 2$. Closest to our work is [15] who studies instance-optimality via local minimax complexity [14]. For one-dimensional functions, they develop a bisection-based instance-optimal algorithm and show that on individual functions of the form $t\mapsto\kappa^{-1}|t|^\kappa$, the local minimax rate is $(1/\sqrt{n})^{\kappa/(\kappa-1)}$.

2 Preliminaries

We first provide notation that we use throughout this paper, define useful assumptions and present key definitions in convex analysis and differential privacy.

Notation. n typically denotes the sample size and d the dimension. Throughout this work, x refers to the optimization variable, $\mathcal{X}\subset\mathbb{R}^d$ to the constraint set and s to elements (S when random) of the sample space $\mathbb{S}$. We usually denote by $F:\mathcal{X}\times\mathbb{S}\to\mathbb{R}$ the (convex) loss function and for a dataset $\mathcal{S}=(s_1,\ldots,s_n)\subset\mathbb{S}$, we define the empirical and population losses

$$f_{\mathcal{S}}(x)\coloneqq\frac{1}{n}\sum_{i\leq n}F(x;s_i)\quad\text{and}\quad f(x)\coloneqq\mathbb{E}_{S\sim P}[F(x;S)].$$

We omit the dependence on P as it is often clear from context. We reserve $\varepsilon,\delta\geq 0$ for the privacy parameters of Definition 2.1. We always take gradients with respect to the optimization variable x. In the case that $F(\cdot;s)$ is not differentiable at x, we override notation and define $\nabla F(x;s)=\operatorname{argmin}_{g\in\partial F(x;s)}\|g\|_2$, where $\partial F(x;s)$ is the subdifferential of $F(\cdot;s)$ at x. We use A for (potentially random) mechanism and S_1^n as a shorthand for $(S_1,\ldots,S_n)$. For $p\geq 1$, $\|\cdot\|_p$ is the standard ℓ_p-norm, $\mathbb{B}_p^d(R)$ is the corresponding d-dimensional p-ball of radius R and $p^\star$ is the dual of p, i.e. such that $1/p^\star+1/p=1$. Finally, we define the Hamming distance between datasets $d_{\mathrm{Ham}}(\mathcal{S},\mathcal{S}')\coloneqq\inf_{\sigma\in\mathfrak{S}_n}\mathbf{1}\{s_i\neq s'_{\sigma(i)}\}$, where $\mathfrak{S}_n$ is the set of permutations over sets of size n.

Assumptions. We first state standard assumptions for solving (1). We assume that $\mathcal{X}$ is a closed, convex domain such that $\mathsf{diam}_2(\mathcal{X})=\sup_{x,y\in\mathcal{X}}\|x-y\|_2\leq D<\infty$. Furthermore, we assume that for any $s\in\mathbb{S}$, $F(\cdot;s)$ is convex and L-Lipschitz with respect to $\|\cdot\|_2$. Central to our work, we define the following κ-growth assumption.

Assumption 1 (κ-growth)**.** *Let $x^\star=\operatorname{argmin}_{x\in\mathcal{X}}f(x)$. For a loss F and distribution P, we say that (F,P) has (λ,κ) growth for $\kappa\in[1,\infty]$ and $\lambda>0$, if the population function satisfies*

$$\text{for all } x\in\mathcal{X},\quad f(x)-f(x^\star)\geq\frac{\lambda}{\kappa}\|x-x^\star\|_2^\kappa.$$

In the case where $\widehat{P}$ is the empirical distribution on a finite dataset $\mathcal{S}$, we refer to (λ,κ)-growth of $(F,\widehat{P})$ as κ-growth of the empirical function $f_{\mathcal{S}}$.

Uniform convexity and Kurdyka-Łojasiewicz inequality. Assumption 1 is closely related to two fundamental notions in convex analysis: uniform convexity and the Kurdyka-Łojasiewicz inequality.

Following [39], we say that $h : \mathcal{Z} \subset \mathbb{R}^d \to \mathbb{R}$ is (σ, κ)-uniformly convex with $\sigma > 0$ and $\kappa \geq 2$ if

$$\text{for all } x, y \in \mathcal{Z}, \quad h(y) \geq h(x) + \langle \nabla h(x), y - x \rangle + \frac{\sigma}{\kappa} \|x - y\|_2^\kappa.$$

This immediately implies that (i) sums (and expectations) preserve uniform convexity (ii) if f is uniformly convex with λ and κ, then it has (λ, κ)-growth. This will be useful when constructing hard instances as it will suffice to consider (λ, κ)-uniformly convex functions which are generally more convenient to manipulate. Finally, we point out that, in the general case that $\kappa \geq 1$, the literature refers to Assumption 1 as the Kurdyka-Łojasiewicz inequality [11] with, in their notation, $\varphi(s) = (\kappa/\lambda)^{1/\kappa} s^{1/\kappa}$. Theorem 5-(ii) in [11] says that, under mild conditions, Assumption 1 implies the following inequality between the error and the gradient norm for all $x \in \mathcal{X}$

$$f(x) - \inf_{x' \in \mathcal{X}} f(x') \leq \frac{e}{\lambda^{\frac{1}{\kappa - 1}}} \|\nabla f(x)\|_2^{\frac{\kappa}{\kappa - 1}}, \tag{2}$$

This is a key result in our analysis of the inverse sensitivity mechanism of Section 3.

Differential privacy. We begin by recalling the definition of (ε, δ)-differential privacy.

Definition 2.1 ([22, 21]). *A randomized algorithm* A *is* (ε, δ)*-differentially private (*(ε, δ)*-DP) if, for all datasets* $\mathcal{S}, \mathcal{S}' \in \mathbb{S}^n$ *that differ in a single data element and for all events* $\mathcal{O}$ *in the output space of* A*, we have*

$$\Pr\left(\mathsf{A}(\mathcal{S}) \in \mathcal{O}\right) \leq e^\epsilon \Pr\left(\mathsf{A}(\mathcal{S}') \in \mathcal{O}\right) + \delta.$$

We use the following standard results in differential privacy.

Lemma 2.1 (Composition [20, Thm. 3.16]). *If* $\mathsf{A}_1, \ldots, \mathsf{A}_k$ *are randomized algorithms that each is* (ε, δ)*-DP, then their composition* $(\mathsf{A}_1(\mathcal{S}), \ldots, \mathsf{A}_k(\mathcal{S}))$ *is* $(k\varepsilon, k\delta)$*-DP.*

Next, we consider the Laplace mechanism. We will let $Z \sim \mathsf{Lap}_d(\sigma)$ denote a d-dimensional vector $Z \in \mathbb{R}^d$ such that $Z_i \overset{\text{iid}}{\sim} \mathsf{Lap}(\sigma)$ for $1 \leq i \leq d$.

Lemma 2.2 (Laplace mechanism [20, Thm. 3.6]). *Let* $h : \mathbb{S}^n \to \mathbb{R}^d$ *have* ℓ_1*-sensitivity* Δ*, that is,* $\sup_{\mathcal{S}, \mathcal{S}' \in \mathbb{S}^n : d_{\mathrm{Ham}}(\mathcal{S}, \mathcal{S}') \leq 1} \|h(\mathcal{S}) - h(\mathcal{S}')\|_1 \leq \Delta$*. Then the Laplace mechanism* $\mathsf{A}(\mathcal{S}) = h(\mathcal{S}) + \mathsf{Lap}_d(\sigma)$ *with* $\sigma = \Delta/\varepsilon$ *is* ε*-DP.*

Finally, we need the Gaussian mechanism for (ε, δ)-DP.

Lemma 2.3 (Gaussian mechanism [20, Thm. A.1]). *Let* $h : \mathbb{S}^n \to \mathbb{R}^d$ *have* ℓ_2*-sensitivity* Δ*, that is,* $\sup_{\mathcal{S}, \mathcal{S}' \in \mathbb{S}^n : d_{\mathrm{Ham}}(\mathcal{S}, \mathcal{S}') \leq 1} \|h(\mathcal{S}) - h(\mathcal{S}')\|_2 \leq \Delta$*. Then the Gaussian mechanism* $\mathsf{A}(\mathcal{S}) = h(\mathcal{S}) + \mathsf{N}(0, \sigma^2 I_d)$ *with* $\sigma = 2\Delta \log(2/\delta)/\varepsilon$ *is* (ε, δ)*-DP.*

Inverse sensitivity mechanism. Our goal is to design private optimization algorithms that adapt to the difficulty of the underlying function. As a reference point, we turn to the inverse sensitivity mechanism of [2] as it enjoys general instance-optimality guarantees. For a given function $h : \mathbb{S}^n \to \mathcal{T} \subset \mathbb{R}^d$ that we wish to estimate privately, define the *inverse sensitivity* at $x \in \mathcal{T}$

$$\mathsf{len}_h(\mathcal{S}; x) = \inf_{\mathcal{S}'} \{d_{\mathrm{Ham}}(\mathcal{S}', \mathcal{S}) : h(\mathcal{S}') = x\}, \tag{3}$$

that is, the inverse sensitivity of a target parameter $y \in \mathcal{T}$ at instance $\mathcal{S}$ is the minimal number of samples one needs to change to reach a new instance $\mathcal{S}'$ such that $h(\mathcal{S}') = y$. Having this quantity, the inverse sensitivity mechanism samples an output from the following probability density

$$\pi_{\mathsf{A}_{\mathrm{inv}}(\mathcal{S})}(x) \propto e^{-\varepsilon \mathsf{len}_h(\mathcal{S}; x)}. \tag{4}$$

The inverse sensitivity mechanism preserves ε-DP and enjoys instance-optimality guarantees in general settings [2]. In contrast to (worst-case) minimax optimality guarantees which measure the performance of the algorithm on the hardest instance, these notions of instance-optimality provide stronger per-instance optimality guarantees.

3 Adaptive rates through inverse sensitivity for ε-DP

To understand the achievable rates when privately optimizing functions with growth, we begin our theoretical investigation by examining the inverse sensitivity mechanism in our setting. We show that, for instances that exhibit κ-growth of the empirical function, the inverse sensitivity mechanism privately solves ERM with excess loss roughly $(d/n\varepsilon)^{\frac{\kappa}{\kappa-1}}$.

In our setting, we use a gradient-based approximation of the inverse sensitivity mechanism to simplify the analysis, while attaining similar rates. Following [3] with our function of interest $h(\mathcal{S}) := \operatorname{argmin}_{x\in\mathcal{X}} f_{\mathcal{S}}(x)$, we can lower bound the inverse sensitivity $\mathsf{len}_h(\mathcal{S}; x) \geq n\|\nabla f_{\mathcal{S}}(x)\|_2/2L$ under natural assumptions. We define a ρ-smoothed version of this quantity which is more suitable to continuous domains

$$G^{\rho}_{\mathcal{S}}(x) = \inf_{y\in\mathcal{X}:\|y-x\|_2\leq\rho} \|\nabla f_{\mathcal{S}}(y)\|_2,$$

and define the ρ-smooth gradient-based inverse sensitivity mechanism

$$\pi_{\mathsf{A}_{\mathrm{gr-inv}}(\mathcal{S})}(x) \propto e^{-\varepsilon n G^{\rho}_{\mathcal{S}}(x)/2L}. \tag{5}$$

Note that while exactly sampling from the un-normalized density $\pi_{\mathsf{A}_{\mathrm{gr-inv}}(\mathcal{S})}$ is computationally intractable, analyzing its performance is an important step towards understanding the optimal rates for the family of functions with growth that we study in this work. The following theorem demonstrates the adaptivity of the inverse sensitivity mechanism to the growth of the underlying instance. We defer the proof to Appendix A.

Theorem 1. *Let $\mathcal{S} = (s_1, \ldots, s_n) \in \mathbb{S}^n$, $F(x; s)$ be convex, L-Lipschitz for all $s \in \mathbb{S}$. Let $x^\star = \operatorname{argmin}_{x\in\mathcal{X}} f_{\mathcal{S}}(x)$ and assume $x^\star$ is in the interior of $\mathcal{X}$. Assume that $f_{\mathcal{S}}(x)$ has κ-growth (Assumption 1) with $\kappa \geq \underline{\kappa} > 1$. For $\rho > 0$, the ρ-smooth inverse sensitivity mechanism $\mathsf{A}_{\mathrm{gr-inv}}$ (5) is ε-DP, and with probability at least $1-\beta$ the output $\hat{x} = \mathsf{A}_{\mathrm{gr-inv}}(\mathcal{S})$ has*

$$f_{\mathcal{S}}(\hat{x}) - \min_{x\in\mathcal{X}} f_{\mathcal{S}}(x) \leq \frac{1}{\lambda^{\frac{1}{\kappa-1}}} \left(\frac{2L(\log(1/\beta) + d\log(D/\rho))}{n\varepsilon}\right)^{\frac{\kappa}{\kappa-1}} + L\rho.$$

Moreover, setting $\rho = (L/\lambda)^{\frac{1}{\kappa-1}}(d/n\varepsilon)^{\frac{\kappa}{\kappa-1}}$, we have

$$f_{\mathcal{S}}(\hat{x}) - \min_{x\in\mathcal{X}} f_{\mathcal{S}}(x) \leq \frac{1}{\lambda^{\frac{1}{\kappa-1}}} \widetilde{O}\left(\frac{Ld}{n\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}.$$

The rates of the inverse sensitivity in Theorem 1 provide two main insights regarding the landscape of the problem with growth conditions. First, these conditions allow to improve the worst-case rate $d/n\varepsilon$ to $(d/n\varepsilon)^{\frac{\kappa}{\kappa-1}}$ for pure ε-DP and therefore suggest a better rate $(\sqrt{d\log(1/\delta)}/n\varepsilon)^{\frac{\kappa}{\kappa-1}}$ is possible for approximate (ε, δ)-DP. Moreover, the general instance-optimality guarantees of this mechanism [2] hint that these are the optimal rates for our class of functions. In the sections to come, we validate the correctness of these predictions by developing efficient algorithms that achieve these rates (for pure and approximate privacy), and prove matching lower bounds which demonstrate the optimality of these algorithms.

4 Efficient algorithms with optimal rates

While the previous section demonstrates that there exists algorithms that improve the rates for functions with growth, we pointed out that $\mathsf{A}_{\mathrm{gr-inv}}$ was computationally intractable in the general case. In this section, we develop efficient algorithms—e.g. that are implementable with gradient-based methods—that achieve the same convergence rates. Our algorithms build on the recent localization techniques that Feldman et al. [25] used to obtain optimal rates for DP-SCO with general convex functions. In Section 4.1, we use these techniques to develop private algorithms that achieve the optimal rates for (pure) DP-SCO with high probability, in contrast to existing results which bound the expected excess loss. These results are of independent interest.

In Section 4.2, we translate these results into convergence guarantees on privately optimizing convex functions with growth by solving a sequence of increasingly constrained SCO problems—the high-probability guarantees of Section 4.1 being crucial to our convergence analysis of these algorithms.

4.1 High-probability guarantees for convex DP-SCO

We first describe our algorithm (Algorithm 1) then analyze its performance under pure-DP (Proposition 1) and approximate-DP constraints (Proposition 2). Our analysis builds on novel tight generalization bounds for uniformly-stable algorithms with high probability [24]. We defer the proofs to Appendix B.

Algorithm 1 Localization-based Algorithm

Require: Dataset $\mathcal{S} = (s_1, \dots, s_n) \in \mathbb{S}^n$, constraint set $\mathcal{X}$, step size η, initial point x_0, privacy parameters (ε, δ);

Set $k = \lceil \log n \rceil$ and $n_0 = n/k$

for $i = 1$ to k **do**

  Set $\eta_i = 2^{-4i}\eta$

  Solve the following ERM over $\mathcal{X}_i = \{x \in \mathcal{X} : \|x - x_{i-1}\|_2 \le 2L\eta_i n_0\}$:

$$F_i(x) = \frac{1}{n_0} \sum_{j=1+(i-1)n_0}^{in_0} F(x; s_j) + \frac{1}{\eta_i n_0}\|x - x_{i-1}\|_2^2$$

  Let $\hat{x}_i$ be the output of the optimization algorithm

  if $\delta = 0$ **then**

    Set $\zeta_i \sim \mathsf{Lap}_d(\sigma_i)$ where $\sigma_i = 4L\eta_i\sqrt{d}/\varepsilon_i$

  else if $\delta > 0$ **then**

    Set $\zeta_i \sim \mathsf{N}(0, \sigma_i^2)$ where $\sigma_i = 4L\eta_i\sqrt{\log(1/\delta)}/\varepsilon$

  Set $x_i = \hat{x}_i + \zeta_i$

return the final iterate x_k

Proposition 1. *Let $\beta \le 1/(n+d)$, $\mathsf{diam}_2(\mathcal{X}) \le D$ and $F(x; s)$ be convex, L-Lipschitz for all $s \in \mathbb{S}$. Setting*

$$\eta = \frac{D}{L}\min\left(\frac{1}{\sqrt{n\log(1/\beta)}}, \frac{\varepsilon}{d\log(1/\beta)}\right)$$

then for $\delta = 0$, Algorithm 1 is ε-DP and has with probability $1 - \beta$

$$f(x) - f(x^\star) \le LD \cdot O\left(\frac{\sqrt{\log(1/\beta)}\log^{3/2} n}{\sqrt{n}} + \frac{d\log(1/\beta)\log n}{n\varepsilon}\right).$$

Similarly, by using a different choice for the parameters and noise distribution, we have the following guarantees for approximate (ε, δ)-DP.

Proposition 2. *Let $\beta \le 1/(n+d)$, $\mathsf{diam}_2(\mathcal{X}) \le D$ and $F(x; s)$ be convex, L-Lipschitz for all $s \in \mathbb{S}$. Setting*

$$\eta = \frac{D}{L}\min\left(\frac{1}{\sqrt{n\log(1/\beta)}}, \frac{\varepsilon}{\sqrt{d\log(1/\delta)}\log(1/\beta)}\right),$$

then for $\delta > 0$, Algorithm 1 is (ε, δ)-DP and has with probability $1 - \beta$

$$f(x) - f(x^\star) \le LD \cdot O\left(\frac{\sqrt{\log(1/\beta)}\log^{3/2} n}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}\log(1/\beta)\log n}{n\varepsilon}\right).$$

Before presenting our algorithms for functions with growth, we remark that the exact calculation of the ERM solution in step 5 of Algorithm 1 is not necessary; we chose it to clarify the main algorithmic ideas. However, as long as the returned solution $\hat{x}_i$ is sufficiently accurate, say $F_i(\hat{x}_i) - \min_{x_i^\star \in \mathcal{X}} F_i(x_i^\star) \le \Delta$, we have that $\|\hat{x}_i - x_i^\star\|_2 \le \sqrt{2\Delta/\lambda_i}$ where $\lambda_i = 1/(\eta_i n_0)$. This implies that as long as $\Delta \le \frac{L}{n}$, the sensitivity of $\hat{x}_i$ is at most twice the sensitivity of the exact ERM solution $x_i^\star$ and hence multiplying the noise σ_i by a factor of 2 is sufficient to guarantee privacy. As the set $\mathcal{X}_i$ is convex, finding sufficiently accurate $\hat{x}_i$ can be done efficiently using standard optimization methods for minimizing convex functions over convex domains.

4.2 Algorithms for DP-SCO with growth

Building on the algorithms of the previous section, we design algorithms that recover the rates of the inverse sensitivity mechanism for functions with growth, importantly without knowledge of the value of κ. Inspired by epoch-based algorithms from the optimization literature [31, 29], our algorithm iteratively applies the private procedures from the previous section. Crucially, the growth assumption allows to reduce the diameter of the domain after each run, hence improving the overall excess loss by carefully choosing the hyper-parameters. We provide full details in Algorithm 2.

Algorithm 2 Epoch-based algorithms for κ-growth

Require: Dataset $\mathcal{S} = (s_1, \dots, s_n) \in \mathbb{S}^n$, convex set $\mathcal{X}$, initial point x_0, number of iterations T, privacy parameters (ε, δ);

Set $n_0 = n/T$ and $D_0 = \mathsf{diam}_2(\mathcal{X})$
if $\delta = 0$ **then**
  Set $\eta_0 = \frac{D_0}{2L} \min\left(\frac{1}{\sqrt{n_0 \log(n_0)\log(1/\beta)}}, \frac{\varepsilon}{d\log(1/\beta)}\right)$
else if $\delta > 0$ **then**
  $\eta_0 = \frac{D_0}{2L} \min\left(\frac{1}{\sqrt{n_0 \log(n_0)\log(1/\beta)}}, \frac{\varepsilon}{\sqrt{d\log(1/\delta)\log(1/\beta)}}\right)$
for $i = 0$ to $T-1$ **do**
  Let $\mathcal{S}_i = (s_{1+(i-1)n_0}, \dots, s_{in_0})$
  Set $D_i = 2^{-i} D_0$ and $\eta_i = 2^{-i}\eta_0$
  Set $\mathcal{X}_i = \{x \in \mathcal{X} : \|x - x_i\|_2 \le D_i\}$
  Run Algorithm 1 on dataset $\mathcal{S}_i$ with starting point x_i, privacy parameter (ε, δ), domain $\mathcal{X}_i$ (with diameter D_i), step size η_i
  Let x_{i+1} be the output of the private procedure
return x_T

The following theorem summarizes our main upper bound for DP-SCO with growth in the pure privacy model, recovering the rates of the inverse sensitivity mechanism in Section 3. We defer the proof to Appendix B.3.

Theorem 2. *Let $\beta \le 1/(n+d)$, $\mathsf{diam}_2(\mathcal{X}) \le D$ and $F(x;s)$ be convex, L-Lipschitz for all $s \in \mathbb{S}$. Assume that f has κ-growth (Assumption 1) with $\kappa \ge \underline{\kappa} > 1$. Setting $T = \left\lceil \frac{2\log n}{\underline{\kappa}-1} \right\rceil$, Algorithm 2 is ε-DP and has with probability $1-\beta$*

$$f(x_T) - \min_{x\in\mathcal{X}} f(x) \le \frac{1}{\lambda^{\frac{1}{\kappa-1}}} \cdot \widetilde{O}\left(\frac{L\sqrt{\log(1/\beta)}}{\sqrt{n}} + \frac{Ld\log(1/\beta)}{n\varepsilon(\underline{\kappa}-1)}\right)^{\frac{\kappa}{\kappa-1}},$$

where $\widetilde{O}$ hides logarithmic factors depending on n and d.

Sketch of the proof. The main challenge of the proof is showing that the iterate achieves good risk without knowledge of κ. Let us denote by $D\cdot\rho$ the error guarantee of Proposition 1 (or Proposition 2 for approximate-DP). At each stage i, as long as $x^\star = \mathrm{argmin}_{x\in\mathcal{X}} f(x)$ belongs to $\mathcal{X}_i$, the excess loss is of order $D_i \cdot \rho$ and thus decreases exponentially fast with i. The challenge is that, without knowledge of κ, we do not know the index i_0 (roughly $\frac{\log_2 n}{\kappa-1}$) after which $x^\star \notin D_j$ for $j \ge i_0$ and the regret guarantees become meaningless with respect to the original problem. However, in the stages after i_0, as the constraint set becomes very small, we upper bound the variations in function values $f(x_{j+1}) - f(x_j)$ and show that the sub-optimality cannot increase (overall) by more than $O(D_{i_0}\cdot\rho)$, thus achieving the optimal rate of stage i_0.

□

Moreover, we can improve the dependence on the dimension for approximate (ε,δ)-DP, resulting in the following bounds.

Theorem 3. *Let $\beta \le 1/(n+d)$, $\mathsf{diam}_2(\mathcal{X}) \le D$ and $F(x;s)$ be convex, L-Lipschitz for all $s \in \mathbb{S}$. Assume that f has κ-growth (Assumption 1) with $\kappa \ge \underline{\kappa} > 1$. Setting $T = \left\lceil \frac{2\log n}{\underline{\kappa}-1} \right\rceil$ and $\delta > 0$,*

Algorithm 2 is (ε, δ)-DP and has with probability $1-\beta$

$$f(x_T) - \min_{x\in\mathcal{X}} f(x) \le \frac{1}{\lambda^{\frac{1}{\kappa-1}}} \cdot \widetilde{O}\left(\frac{L\sqrt{\log(1/\beta)}}{\sqrt{n}} + \frac{L\sqrt{d\log(1/\delta)}\log(1/\beta)}{n\varepsilon(\underline{\kappa}-1)}\right)^{\frac{\kappa}{\kappa-1}},$$

where $\widetilde{O}$ hides logarithmic factors depending on n and d.

5 Lower bounds

In this section, we develop (minimax) lower bounds for the problem of SCO with κ-growth under privacy constraints. Note that taking $\varepsilon \to \infty$ provides lower bound for the unconstrained minimax risk. For a sample space $\mathbb{S}$ and collection of distributions $\mathcal{P}$ over $\mathbb{S}$, we define the function class $\mathcal{F}^\kappa(\mathcal{P})$ as the set of convex functions from $\mathbb{R}^d \to \mathbb{R}$ that are L-Lipschitz and has κ-growth (Assumption 1). We define the *constrained* minimax risk [6]

$$\mathfrak{M}_n(\mathcal{X}, \mathcal{P}, \mathcal{F}^\kappa, \varepsilon, \delta) := \inf_{\widehat{x}_n \in \mathcal{A}^{\varepsilon,\delta}} \sup_{(F,P)\in\mathcal{F}^\kappa\times\mathcal{P}} \mathbb{E}\left[f(\widehat{x}_n(S_1^n)) - \inf_{x'\in\mathcal{X}} f(x')\right], \tag{6}$$

where $\mathcal{A}^{\epsilon,\delta}$ is the collection of (ε,δ)-DP mechanisms from $\mathbb{S}^n$ to $\mathcal{X}$. When clear from context, we omit the dependency on $\mathcal{P}$ of the function class and simply write $\mathcal{F}^\kappa$. We also forego the dependence on δ when referring to pure-DP constraints, i.e. $\mathfrak{M}_n(\mathcal{X}, \mathcal{P}, \mathcal{F}^\kappa, \varepsilon, \delta = 0) =: \mathfrak{M}_n(\mathcal{X}, \mathcal{P}, \mathcal{F}^\kappa, \varepsilon)$. We now proceed to prove tight lower bounds for ε-DP in Section 5.1 and (ε,δ)-DP in Section 5.2.

5.1 Lower bounds for pure ε-DP

Although in Section 4 we show that the same algorithm achieves the optimal upper bounds for all values of $\kappa > 1$, the landscape of the problem is more subtle for the lower bounds and we need to delineate two different cases to obtain tight lower bounds. We begin with $\kappa \ge 2$, which corresponds to uniform convexity and enjoys properties that make the problem easier (e.g., closure under summation or addition of linear terms). The second case, $1 < \kappa < 2$, corresponds to sharper growth and requires a different hard instance to satisfy the growth condition.

κ-growth with $\kappa \ge 2$. We begin by developing lower bounds under pure DP for $\kappa \ge 2$

Theorem 4 (Lower bound for ε-DP, $\kappa \ge 2$)**.** *Let $d \ge 1$, $\mathcal{X} = \mathbb{B}_2^d(R)$, $\mathbb{S} = \{\pm e_j\}_{j\le d}$, $\kappa \ge 2$ and $n \in \mathbb{N}$. Let $\mathcal{P}$ be the set of distributions on $\mathbb{S}$. Assume that*

$$2^{\kappa-1} \le \frac{L}{\lambda}\frac{1}{R^{\kappa-1}} \le 2^{\kappa-1}\sqrt{96n} \;\; \textit{and} \;\; n\varepsilon \ge \frac{1}{\sqrt{3}}$$

The following lower bound holds

$$\mathfrak{M}_n(\mathcal{X}, \mathcal{P}, \mathcal{F}^\kappa, \epsilon) \ge \frac{1}{\lambda^{\frac{1}{\kappa-1}}}\tilde{\Omega}\left(\left(\frac{L}{\sqrt{n}}\right)^{\frac{\kappa}{(\kappa-1)}} + \left(\frac{Ld}{n\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right). \tag{7}$$

First of all, note that $L \ge \lambda 2^\kappa R^{\kappa-1}$ is not an overly-restrictive assumption. Indeed, for an arbitrary (λ,κ)-uniformly convex and L-Lipschitz function, it always holds that $L \ge \frac{\lambda}{2}R^{\kappa-1}$. This is thus equivalent to assuming $\kappa = \Theta(1)$. Note that when $\kappa \gg 1$, the standard $n^{-1/2} + d/(n\varepsilon)$ lower bound holds. We present the proof in Appendix C.1.1 and preview the main ideas here.

Sketch of the proof. Our lower bounds hinges on the collections of functions $F(x;s) := a\kappa^{-1}\|x\|_2^\kappa + b\langle x, s\rangle$ for $a, b \ge 0$ to be chosen later. These functions are [39, Lemma 4] κ-uniformly convex for any $s \in \mathbb{S}$ and in turn, so is the population function f. We proceed as follows, we first prove an information-theoretic (non-private) lower bound (Theorem 8 in Appendix C.1.1) which provides the statistical term in (7). With the same family of functions, we exhibit a collection of datasets and prove by contradiction that if an estimator were to optimize below a certain error it would have violated ε-DP—this yields a lower bound on ERM for our function class (Theorem 9 in Appendix C.1.1). We conclude by proving a reduction from SCO to ERM in Proposition 4. □

κ-growth with $\kappa \in (1, 2]$. As the construction of the hard instance is more intricate for $\kappa < 2$, we provide a one-dimensional lower bound and leave the high-dimensional case for future work. In this case we directly obtain the result with a private version of Le Cam's method [44, 42, 6], however with a different family of functions.

The issue with the construction of the previous section is that the function does not exhibit sharp growth for $\kappa < 2$. Indeed, the added linear function shifts the minimum away from 0 where the function is differentiable and as a result it locally behaves as a quadratic and only achieves growth $\kappa = 2$. To establish the lower bound, we consider a different sample function F that has growth exactly 1 on one side and κ on the other side. This yields the following

Theorem 5 (Lower bound for ε-DP, $\kappa \in (1, 2]$)**.** *Let $d = 1$, $\mathbb{S} = \{-1, +1\}$, $\kappa \in (1, 2]$, $\lambda = 1$, $L = 2$, and $n \in \mathbb{N}$. There exists a collection of distributions $\mathcal{P}$ such that, whenever $n\varepsilon \geq 1/\sqrt{3}$, it holds that*

$$\mathfrak{M}_n([-1, 1], \mathcal{P}, \mathcal{F}_{d=1}^{\kappa}, \epsilon) = \Omega\left\{ \left(\frac{1}{\sqrt{n}}\right)^{\frac{\kappa}{\kappa-1}} + \left(\frac{1}{n\varepsilon}\right)^{\frac{\kappa}{\kappa-1}} \right\}. \tag{8}$$

5.2 Lower bounds under approximate privacy constraints

We conclude our treatment by providing lower bounds but now under *approximate* privacy constraints, demonstrating the optimality of the risk bound of Theorem 3. We prove the result via a reduction: we show that if one solves ERM with κ-growth with error Δ, this implies that one solves arbitrary convex ERM with error $\phi(\Delta)$. Given that a lower bound of $\Omega(\sqrt{d}/(n\varepsilon))$ holds for ERM, a lower bound of $\phi^{-1}(\sqrt{d}/(n\varepsilon))$ holds for ERM with κ-growth. However, for this reduction to hold, we require that $\kappa \geq 2$. Furthermore, we consider κ to be roughly a constant—in the case that κ is too large, standard lower bounds on general convex functions apply.

Theorem 6 (Private lower bound for (ε, δ)-DP)**.** *Let $\kappa \geq 2$ such that $\kappa = \Theta(1)$, $\mathcal{X} = \mathbb{B}_2^d(D)$. Let $d \geq 1$ and $\mathbb{S} = \{\pm 1/\sqrt{d}\}^d$. Assume that $n\varepsilon = \Omega(\sqrt{d})$, then for any (ε, δ) mechanism A, there exists $\lambda > 0$, F and $\mathcal{S} \subset \mathbb{S}$ such that*

$$\mathbb{E}[f_{\mathcal{S}}(\mathsf{A}(\mathcal{S}))] - \inf_{x' \in \mathcal{X}} f_{\mathcal{S}}(x') \geq \tilde{\Omega}\left[\frac{1}{\lambda^{\frac{1}{\kappa-1}}} \left(\frac{L\sqrt{d}}{n\varepsilon} \right)^{\frac{\kappa}{\kappa-1}} \right].$$

Theorem 6 implies that the same lower bound (up to logarithmic factors) applies to SCO via the reduction of [8, Appendix C]. Before proving the theorem, let us state (and prove in Appendix C.2) the following reduction: if an (ε, δ)-DP algorithm achieves excess error (roughly) Δ on ERM for any function with κ-growth, there exists an (ε, δ)-DP algorithm that achieves error $\Delta^{(\kappa-1)/\kappa}$ for any convex function. We construct the latter by iteratively solving ERM problems with geometrically increasing $\|\cdot\|_2^{\kappa}$-regularization towards the previous iterate to ensure the objective has κ-growth.

Proposition 3 (Solving ERM with κ-growth implies solving any convex ERM)**.** *Let $\kappa \geq 2$. Assume there exists an (ϵ, δ) mechanism A such that for any L-Lipschitz loss G on $\mathcal{Y}$ and dataset $\mathcal{S}$ such that $g_{\mathcal{S}}(x) \coloneqq \frac{1}{n}\sum_{s \in \mathcal{S}} G(x; s)$ exhibits (λ, κ)-growth, the mechanism achieves excess loss*

$$\mathbb{E}[g_{\mathcal{S}}(\mathsf{A}(\mathcal{S}, G, \mathcal{Y}))] - \inf_{y' \in \mathcal{Y}} g_{\mathcal{S}}(y') \leq \frac{1}{\lambda^{\frac{1}{\kappa-1}}} \Delta(n, L, \epsilon, \delta).$$

Then, we can construct an (ε, δ)-DP mechanism A' such that for any L-Lipschitz loss f, the mechanism achieves excess loss

$$\mathbb{E}[f_{\mathcal{S}}(\mathsf{A}'(\mathcal{S}))] - \inf_{x' \in \mathcal{X}} f_{\mathcal{S}}(x') \leq O\left(D[\Delta(n, L, \epsilon/k, \delta/k)]^{\frac{\kappa-1}{\kappa}} \right),$$

where k is the smallest integer such that $k \geq \log\left[\frac{\kappa^{\frac{1}{\kappa-1}} L^{\frac{\kappa}{\kappa-1}}}{2^{2\kappa-3}\Delta(n, L, \varepsilon/k, \delta/k)} \right]$.

With this proposition, the proof of the theorem directly follows as Bassily et al. [7] prove a lower bound $\Omega(\sqrt{d}/(n\varepsilon))$ for ERM with (ε, δ)-DP.

Discussion

In this work, we develop private algorithms that adapt to the growth of the function at hand, achieving the convergence rate corresponding to the "easiest" sub-class the function belongs to. However, the picture is not yet complete. First of, there are still gaps in our theoretical understanding, the most interesting one being $\kappa = 1$. On these functions, appropriate optimization algorithms achieve linear convergence [43] and raise the question, can we achieve exponentially small privacy cost in our setting? Finally, while our optimality guarantees are more fine-grained than the usual minimax results over convex functions, they are still contigent on some predetermined choice of sub-classes. Studying more general notions of adaptivity is an important future direction in private optimization.

Acknowledgments

The authors would like to thank Karan Chadha and Gary Cheng for comments on an early version of the draft. HA, DL and JCD were supported NSF under CAREER Award CCF-1553086 and HDR 1934578 (Stanford Data Science Collaboratory), Office of Naval Research YIP Award N00014-19-2288 and the Stanford DAWN Consortium.

Competing Interests

JCD has a consulting relationship with Apple. HS has spent internships at Apple during the 2019, 2020 and 2021 summers. DL has spent an internship at Google during the summer of 2020.

References

[1] M. Abadi, A. Chu, I. Goodfellow, B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *23rd ACM Conference on Computer and Communications Security (ACM CCS)*, pages 308–318, 2016.

[2] H. Asi and J. Duchi. Near instance-optimality in differential privacy. *arXiv:2005.10630 [cs.CR]*, 2020.

[3] H. Asi and J. C. Duchi. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. In *Advances in Neural Information Processing Systems 33*, 2020.

[4] H. Asi, J. Duchi, A. Fallah, O. Javidbakht, and K. Talwar. Private adaptive gradient methods for convex optimization. In *Proceedings of the 38th International Conference on Machine Learning*, pages 383–392, 2021.

[5] H. Asi, V. Feldman, T. Koren, and K. Talwar. Private stochastic convex optimization: Optimal rates in ℓ_1 geometry. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[6] R. F. Barber and J. C. Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv:1412.4451 [math.ST]*, 2014.

[7] R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *55th Annual Symposium on Foundations of Computer Science*, pages 464–473, 2014.

[8] R. Bassily, V. Feldman, K. Talwar, and A. Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems 32*, 2019.

[9] R. Bassily, V. Feldman, C. Guzmán, and K. Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems 33*, 2020.

[10] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.

[11] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165:471–507, 2017.

[12] L. Bottou, F. Curtis, and J. Nocedal. Optimization methods for large-scale learning. *SIAM Review*, 60(2):223–311, 2018.

[13] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the Forty-Eighth Annual ACM Symposium on the Theory of Computing*, 2016. URL https://arxiv.org/abs/1506.07216.

[14] T. Cai and M. Low. A framework for estimating convex functions. *Statistica Sinica*, 25: 423–456, 2015.

[15] S. Chatterjee, J. Duchi, J. Lafferty, and Y. Zhu. Local minimax complexity of stochastic convex optimization. In *Advances in Neural Information Processing Systems 29*, 2016.

[16] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

[17] J. C. Duchi. Information theory and statistics. Lecture Notes for Statistics 311/EE 377, Stanford University, 2019. URL http://web.stanford.edu/class/stats311/lecture-notes.pdf. Accessed May 2019.

[18] J. C. Duchi and F. Ruan. Asymptotic optimality in stochastic optimization. *Annals of Statistics*, 49(1):21–48, 2021.

[19] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *54th Annual Symposium on Foundations of Computer Science*, pages 429–438, 2013.

[20] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3 & 4):211–407, 2014.

[21] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology (EUROCRYPT 2006)*, 2006.

[22] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference*, pages 265–284, 2006.

[23] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, pages 214–226, 2012.

[24] V. Feldman and J. Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Proceedings of the Thirty Second Annual Conference on Computational Learning Theory*, pages 1270–1279, 2019.

[25] V. Feldman, T. Koren, and K. Talwar. Private stochastic convex optimization: Optimal rates in linear time. In *Proceedings of the Fifty-Second Annual ACM Symposium on the Theory of Computing*, 2020.

[26] A. Garg, T. Ma, and H. L. Nguyen. On communication cost of distributed statistical estimation and dimensionality. In *Advances in Neural Information Processing Systems 27*, 2014.

[27] M. Hardt and K. Talwar. On the geometry of differential privacy. In *Proceedings of the Forty-Second Annual ACM Symposium on the Theory of Computing*, pages 705–714, 2010. URL http://arxiv.org/abs/0907.3754.

[28] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[29] E. Hazan and S. Kale. An optimal algorithm for stochastic strongly convex optimization. In *Proceedings of the Twenty Fourth Annual Conference on Computational Learning Theory*, 2011. URL http://arxiv.org/abs/1006.2425.

[30] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv:1902.03736 [math.PR]*, 2019.

[31] A. Juditsky and Y. Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. URL http://hal.archives-ouvertes.fr/docs/00/50/89/33/PDF/Strong-hal.pdf, 2010.

[32] A. Juditsky and Y. Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1):44—-80, 2014.

[33] D. Levy and J. C. Duchi. Necessary and sufficient geometries for gradient methods. In *Advances in Neural Information Processing Systems 32*, 2019.

[34] D. Levy, Z. Sun, K. Amin, S. Kale, A. Kulesza, M. Mohri, and A. T. Suresh. Learning with user-level privacy. *Advances in Neural Information Processing Systems 34*, 2021. URL https://arxiv.org/abs/2102.11845.

[35] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

[36] M. Mitzenmacher and E. Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

[37] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

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

[39] Y. Nesterov. Accelerating the cubic regularization of newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

[40] A. Ramdas and A. Singh. Optimal rates for stochastic convex optimization under tsybakov noise condition. In *Proceedings of the 30th International Conference on Machine Learning*, pages 365–373, 2013.

[41] A. Smith and A. Thakurta. Differentially private feature selection via stability arguments, and the robustness of the Lasso. In *Proceedings of the Twenty Sixth Annual Conference on Computational Learning Theory*, pages 819–850, 2013. URL http://proceedings.mlr.press/v30/Guha13.html.

[42] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[43] Y. Xu, Q. Lin, and T. Yang. Stochastic convex optimization: Faster local growth implies faster global convergence. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3821–3830, 2017.

[44] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, 1997.

[45] Y. Zhang, J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Information-theoretic lower bounds for distributed estimation with communication constraints. In *Advances in Neural Information Processing Systems 26*, 2013.

[46] Y. Zhu, S. Chatterjee, J. Duchi, and J. Lafferty. Local minimax complexity of stochastic convex optimization. In *Advances in Neural Information Processing Systems 29*, 2016.

Checklist

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

Potential negative societal impact

The aim of our work is theoretical in essence and as such, we do not expect direct negative societal impact. As DP becomes a more established norm, we believe this research is relevant for practitionners in both industry and government. Indeed, an important obstacle to applying DP is the loss of performance compared to non-private models; our theoretical results suggests that better adaptive algorithms would significantly narrow this performance gap. We wish to point out two potential negative consequences of growing research in privacy. First, a simple but effective method to guarantee privacy is to either delete existing user data or limit data collection in the first place. Paradoxically, the more confident institutions are in DP algorithms, the less they are susceptible to turn to these simpler—and most effective—solutions. Finally, using DP algorithms should not preclude one from (1) carefully choosing ϵ and δ to provide meaningful guarantees for the specific application at hand and (2) developing exhaustive and meticulous evaluation methods of the privacy of deployed models.

A Proofs for Section 3

A.1 Proof of Theorem 1

Theorem 1. *Let $\mathcal{S} = (s_1, \dots, s_n) \in \mathbb{S}^n$, $F(x; s)$ be convex, L-Lipschitz for all $s \in \mathbb{S}$. Let $x^\star = \operatorname{argmin}_{x\in\mathcal{X}} f_{\mathcal{S}}(x)$ and assume $x^\star$ is in the interior of $\mathcal{X}$. Assume that $f_{\mathcal{S}}(x)$ has κ-growth (Assumption 1) with $\kappa \geq \underline{\kappa} > 1$. For $\rho > 0$, the ρ-smooth inverse sensitivity mechanism $\mathsf{A}_{\mathrm{gr-inv}}$ (5) is ε-DP, and with probability at least $1 - \beta$ the output $\hat{x} = \mathsf{A}_{\mathrm{gr-inv}}(\mathcal{S})$ has*

$$f_{\mathcal{S}}(\hat{x}) - \min_{x\in\mathcal{X}} f_{\mathcal{S}}(x) \leq \frac{1}{\lambda^{\frac{1}{\kappa-1}}} \left(\frac{2L(\log(1/\beta) + d\log(D/\rho))}{n\varepsilon} \right)^{\frac{\kappa}{\kappa-1}} + L\rho.$$

Moreover, setting $\rho = (L/\lambda)^{\frac{1}{\underline{\kappa}-1}} (d/n\varepsilon)^{\frac{\underline{\kappa}}{\underline{\kappa}-1}}$, we have

$$f_{\mathcal{S}}(\hat{x}) - \min_{x\in\mathcal{X}} f_{\mathcal{S}}(x) \leq \frac{1}{\lambda^{\frac{1}{\kappa-1}}} \widetilde{O}\left(\frac{Ld}{n\varepsilon} \right)^{\frac{\kappa}{\kappa-1}}.$$

Let us first prove privacy. The sensitivity of $\|\nabla f_{\mathcal{S}}(x)\|_2$ is $2L/n$ as F is L-Lipschitz, therefore following the privacy proof of the smooth inverse sensitivity mechanism [2, Prop. 3.2] we get that $\mathsf{A}_{\mathrm{gr-inv}}$ (5) is ε-DP.

Let us now prove the claim about utility. Denote $\hat{x} = \mathsf{A}_{\mathrm{gr-inv}}(\mathcal{S})$ and $E = \frac{2LK}{n\varepsilon}$ with K to be chosen presently. We argue that it is enough to show that $\Pr(G_\rho(\hat{x}) \geq E) \leq \beta$. Indeed then with probability at least $1 - \beta$ we have $G_\rho(\hat{x}) \leq E$, which implies there is y such that $\|\hat{x} - y\|_2 \leq \rho$ and $\|\nabla f_{\mathcal{S}}(y)\|_2 \leq E$, hence using the Kurdyka-Łojasiewicz inequality (2)

$$\begin{aligned} f_{\mathcal{S}}(\hat{x}) - f_{\mathcal{S}}(x^\star) &= f_{\mathcal{S}}(\hat{x}) - f_{\mathcal{S}}(y) + f_{\mathcal{S}}(y) - f_{\mathcal{S}}(x^\star) \\ &\leq L\rho + \frac{e}{\lambda^{\frac{1}{\kappa-1}}} \|\nabla f_{\mathcal{S}}(y)\|_2^{\frac{\kappa}{\kappa-1}} \\ &\leq L\rho + \frac{e}{\lambda^{\frac{1}{\kappa-1}}} E^{\frac{\kappa}{\kappa-1}}. \end{aligned}$$

It remains to prove that $\Pr(G_\rho(\hat{x}) \geq E) \leq \beta$. Let $S_0 = \{x \in \mathbb{R}^d : \|x - x^\star\|_2 \leq \rho\}$ and $S_1 = \{x \in \mathbb{R}^d : G_\rho(x) \geq E\}$. Note that $G_\rho(x) = 0$ for any $x \in S_0$ as $x^\star$ is in the interior of $\mathcal{X}$ which implies $\nabla f_{\mathcal{S}}(x^\star) = 0$. Hence the definition of the smooth inverse sensitivity mechanism (5) implies

$$\begin{aligned} \Pr(\mathsf{A}_{\mathrm{gr-inv}}(\mathcal{S}) \in S_1) &\leq \frac{\mathsf{Vol}(\{x \in \mathbb{R}^d : \|x - x^\star\|_2 \leq D + \rho\}) e^{-\frac{n\varepsilon}{2L}E}}{\mathsf{Vol}(\{x \in \mathbb{R}^d : \|x - x^\star\|_2 \leq \rho\})} \\ &\leq e^{-K} \left(1 + \frac{D}{\rho}\right)^d \leq \beta, \end{aligned}$$

where the last inequality follows by choosing $K = \log(1/\beta) + d\log(1 + D/\rho)$.

B Proofs for Section 4

We need to the following result on the generalization properties of uniformly stable algorithms [24].

Theorem 7. *[24, Cor. 4.2] Assume* $\mathsf{diam}_2(\mathcal{X}) \le D$. *Let* $\mathcal{S} = (S_1, \dots, S_n)$ *where* $S_1^n \overset{\text{iid}}{\sim} P$ *and* $F(x; s)$ *is* L*-Lipschitz and* λ*-strongly convex for all* $s \in \mathbb{S}$. *Let* $\hat{x} = \operatorname{argmin}_{x \in \mathcal{X}} f_{\mathcal{S}}(x)$ *be the empirical minimizer. For* $0 < \beta \le 1/n$, *with probability at least* $1 - \beta$

$$f(\hat{x}) - f(x^\star) \le \frac{cL^2 \log(n) \log(1/\beta)}{\lambda n} + \frac{cLD\sqrt{\log(1/\beta)}}{\sqrt{n}}.$$

B.1 Proof of Proposition 1

Proposition 1. *Let* $\beta \le 1/(n+d)$, $\mathsf{diam}_2(\mathcal{X}) \le D$ *and* $F(x; s)$ *be convex,* L*-Lipschitz for all* $s \in \mathbb{S}$. *Setting*

$$\eta = \frac{D}{L} \min\left(\frac{1}{\sqrt{n \log(1/\beta)}}, \frac{\varepsilon}{d \log(1/\beta)} \right)$$

then for $\delta = 0$, *Algorithm 1 is* ε*-DP and has with probability* $1 - \beta$

$$f(x) - f(x^\star) \le LD \cdot O\left(\frac{\sqrt{\log(1/\beta)} \log^{3/2} n}{\sqrt{n}} + \frac{d \log(1/\beta) \log n}{n\varepsilon} \right).$$

We begin by proving the privacy claim. We show that each iterate is ε-DP and therefore post-processing implies the claim as each sample is used in exactly one iterate. To this end, let $\lambda_i = 1/\eta_i n_0$ and note that the minimizer $\hat{x}_i$ has ℓ_2 sensitivity $2L/\lambda_i n_0 \le 4L\eta_i$ [25], hence the ℓ_1-sensitivity is at most $4L\eta_i\sqrt{d}$. Standard properties of the Laplace mechanism [20] now imply that x_i is ε-DP which give the claim about privacy.

Now we proceed to prove utility which follows similar arguments to the localization-based proof in [25]. Letting $\hat{x}_0 = x^\star$, we have:

$$f(x_k) - f(x^\star) = \sum_{i=1}^{k} f(\hat{x}_i) - f(\hat{x}_{i-1}) + f(x_k) - f(\hat{x}_k).$$

First, by using standard properties of Laplace distributions [17], we know that for $\zeta_i \sim \mathsf{Lap}(\sigma_i)$,

$$\Pr(\|\zeta_i\|_2 \ge t) \le \Pr(\|\zeta_i\|_\infty \ge t/\sqrt{d}) \le d e^{-t/\sqrt{d}\sigma_i},$$

which implies (as $\beta \le 1/(n+d)$) that with probability $1 - \beta/2$ we have $\|\zeta_i\|_2 \le 10\sqrt{d}\sigma_i \log(1/\beta)$ for all $1 \le i \le k$. Hence

$$\begin{aligned} f(x_k) - f(\hat{x}_k) &\le L\|x_k - \hat{x}_k\|_2 \\ &\le L\sigma_k \sqrt{d} \log(1/\beta) \\ &\le 4L^2 d \frac{\eta_i}{\varepsilon} \\ &\le 4L^2 d \frac{\eta}{\varepsilon 2^{4i}} \le \frac{4LD}{n^2}, \end{aligned}$$

where the last inequality follows since $\eta = \frac{D\varepsilon}{Ld \log(k/\beta)}$. Now we use high-probability generalization guarantees of uniformly-stable algorithms. We use Theorem 7 with $F(x; s_j) + \frac{\|x - x_{i-1}\|_2^2}{\eta_i n_0}$ to get that with probability $1 - \beta/2$ for each i

$$f(\hat{x}_i) - f(\hat{x}_{i-1}) \le \frac{\|\hat{x}_{i-1} - x_{i-1}\|_2^2}{\eta_i n_0} + cL^2 \log(n) \log(1/\beta) \eta_i + \frac{cLD\sqrt{\log(1/\beta)}}{\sqrt{n_0}}.$$

Thus,

$$\begin{aligned}
\sum_{i=1}^{k} f(\hat{x}_i) - f(\hat{x}_{i-1}) &\le \sum_{i=1}^{k} \left\{ \frac{\|\hat{x}_{i-1} - x_{i-1}\|_2^2}{\eta_i n_0} + cL^2 \log(n)\log(1/\beta)\eta_i + \frac{cLD\sqrt{\log(1/\beta)}}{\sqrt{n_0}} \right\} \\
&\le \frac{D^2}{\eta n_0} + \left[\sum_{i=2}^{k} \frac{\sigma_{i-1}^2 d \log^2(1/\beta)}{\eta_i n_0} \right] + 2cL^2 \log(n)\log(1/\beta)\eta + \frac{cLD\sqrt{\log(1/\beta)}k}{\sqrt{n_0}} \\
&= \frac{D^2}{\eta n_0} + \left[\sum_{i=2}^{k} \frac{CL^2 \eta_{i-1} d^2 \log^2(1/\beta)}{n_0 \varepsilon^2} \right] + 2cL^2 \log(n)\log(1/\beta)\eta + \frac{cLD\sqrt{\log(1/\beta)}k}{\sqrt{n_0}} \\
&= \frac{D^2}{\eta n_0} + \frac{CL^2 \eta d^2 \log^2(1/\beta)}{n_0 \varepsilon^2} \left[\sum_{i=2}^{k} 2^{-i} \right] + 2cL^2 \log(n)\log(1/\beta)\eta + \frac{cLD\sqrt{\log(1/\beta)}k}{\sqrt{n_0}} \\
&\le LD \cdot O\left(\frac{\sqrt{\log(1/\beta)\log(n)} + \sqrt{\log(1/\beta)}\log^{3/2}(n)}{\sqrt{n}} + \frac{d\log(1/\beta)\log(n)}{n\varepsilon} \right),
\end{aligned}$$

where the last inequality follows by choosing $\eta = \frac{D}{L}\min\left(\frac{1}{\sqrt{n\log(1/\beta)}}, \frac{\varepsilon}{d\log(1/\beta)}\right)$

B.2 Proof of Proposition 2

Proposition 2. *Let $\beta \le 1/(n+d)$, $\mathsf{diam}_2(\mathcal{X}) \le D$ and $F(x;s)$ be convex, L-Lipschitz for all $s \in \mathbb{S}$. Setting*

$$\eta = \frac{D}{L}\min\left(\frac{1}{\sqrt{n\log(1/\beta)}}, \frac{\varepsilon}{\sqrt{d\log(1/\delta)\log(1/\beta)}}\right),$$

then for $\delta > 0$, Algorithm 1 is (ε,δ)-DP and has with probability $1-\beta$

$$f(x) - f(x^\star) \le LD \cdot O\left(\frac{\sqrt{\log(1/\beta)}\log^{3/2} n}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}\log(1/\beta)\log n}{n\varepsilon} \right).$$

The proof is similar to the proof of Proposition 1. For privacy, we show in the proof of Proposition 1 that the ℓ_2-sensitivity of $\hat{x}_i$ is upper bounded by $2L/\lambda_i n_0 \le 4L\eta_i$ hence standard properties of the Gaussian mechanism [20] now imply that x_i is (ε,δ)-DP which implies the final algorithm is (ε,δ)-DP using post-processing.

The utility proof follows the same arguments as in the proof of Proposition 1, except that for $\zeta_i \sim \mathsf{N}(0,\sigma_i^2)$ we have [30] (since ζ_i is $2\sqrt{2}\sigma_i\sqrt{d}$-norm-sub-Gaussian)

$$\Pr(\|\zeta_i\|_2 \ge t\sqrt{d}) \le 2e^{-\frac{t^2}{16\sigma_i^2}},$$

implying that $\|\zeta_i\|_2 \le 4\sqrt{d}\sigma_i \log(4/\beta)$ for all $1 \le i \le k$ with probability $1-\beta/2$.

B.3 Proofs of Theorems 2 and 3

We first restate Theorems 2 and 3.

Theorem 2. *Let $\beta \le 1/(n+d)$, $\mathsf{diam}_2(\mathcal{X}) \le D$ and $F(x;s)$ be convex, L-Lipschitz for all $s \in \mathbb{S}$. Assume that f has κ-growth (Assumption 1) with $\kappa \ge \underline{\kappa} > 1$. Setting $T = \left\lceil \frac{2\log n}{\underline{\kappa}-1} \right\rceil$, Algorithm 2 is ε-DP and has with probability $1-\beta$*

$$f(x_T) - \min_{x\in\mathcal{X}} f(x) \le \frac{1}{\lambda^{\frac{1}{\kappa-1}}} \cdot \widetilde{O}\left(\frac{L\sqrt{\log(1/\beta)}}{\sqrt{n}} + \frac{Ld\log(1/\beta)}{n\varepsilon(\underline{\kappa}-1)} \right)^{\frac{\kappa}{\kappa-1}},$$

where $\widetilde{O}$ hides logarithmic factors depending on n and d.

Theorem 3. *Let $\beta \le 1/(n+d)$, $\mathsf{diam}_2(\mathcal{X}) \le D$ and $F(x;s)$ be convex, L-Lipschitz for all $s \in \mathbb{S}$. Assume that f has κ-growth (Assumption 1) with $\kappa \ge \underline{\kappa} > 1$. Setting $T = \left\lceil \frac{2\log n}{\underline{\kappa}-1} \right\rceil$ and $\delta > 0$, Algorithm 2 is (ε,δ)-DP and has with probability $1-\beta$*

$$f(x_T) - \min_{x\in\mathcal{X}} f(x) \le \frac{1}{\lambda^{\frac{1}{\kappa-1}}} \cdot \widetilde{O}\left(\frac{L\sqrt{\log(1/\beta)}}{\sqrt{n}} + \frac{L\sqrt{d\log(1/\delta)}\log(1/\beta)}{n\varepsilon(\underline{\kappa}-1)} \right)^{\frac{\kappa}{\kappa-1}},$$

where $\widetilde{O}$ hides logarithmic factors depending on n and d.

We start by proving privacy. Since each sample s_i is used in exactly one iterate, we only need to show that each iterate is (ε,δ)-DP, which will imply the main claim using post-processing. The privacy of each iterate follows directly from the privacy guarantees of Algorithm 1. We proceed to prove utility.

We will prove the utility claim assuming the subroutine used in Algorithm 2 satisfies the following: the output x_{k+1} has error

$$f(x_{k+1}) - \min_{x\in\mathcal{X}} f(x) \le D_k \cdot \rho,$$

for some $\rho > 0$. Note that in our setting, Proposition 1 implies that $\rho \le L \cdot O(\frac{\sqrt{\log(1/\beta)}\log n_0}{\sqrt{n_0}} + \frac{d\log(1/\beta)}{n_0\varepsilon})$ for pure-DP and similarly Proposition 2 gives the corresponding ρ for (ε,δ)-DP.

The proof has two stages. In the first stage (Lemma B.1), we prove that as long as $i \le i_0$ for some $i_0 > 0$, then $x^\star \in \mathcal{X}_i$ and the performance of the algorithm keeps improving. We show that at the end of this stage, the points x_{i_0+1} has optimal excess loss. Then, in the second stage (Lemma B.2), we show that the iterates would not move much as the radius D_i of the domain is sufficiently small, hence the final accumulated error along these iterations is small.

Let us begin with the first stage. Let i_0 be the largest i such that $D_i \ge (\frac{\kappa 2^\kappa \rho}{\lambda})^{\frac{1}{\kappa-1}}$. We prove that $x^\star \in \mathcal{X}_i$ for all $0 \le i \le i_0$ where we recall that $\mathcal{X}_i = \{x \in \mathcal{X} : \|x - x_i\|_2 \le D_i\}$ and $D_i = 2^{-i}D_0$.

Lemma B.1. *For all $0 \le i \le i_0$ we have*

$$x^\star \in \mathcal{X}_i \quad \text{and} \quad f(x_{i_0+1}) - \min_{x\in\mathcal{X}} f(x) \le 6(2^\kappa)^{\frac{1}{\kappa-1}} \frac{1}{\lambda^{\frac{1}{\kappa-1}}} \rho^{\frac{\kappa}{\kappa-1}}.$$

Proof. To prove the first part, we need to show that $\|x_i - x^\star\|_2 \le D_i$. Let $\bar{D}_i = \|x_i - x^\star\|_2$. First, note that the claim is true for $i = 0$. Now we assume it is correct for $0 \le i \le i_0 - 1$ and prove correctness for $i+1$. Note that the growth condition implies

$$\bar{D}_{i+1} \le (\kappa\Delta_i/\lambda)^{1/\kappa},$$

where $\Delta_i = f(x_{i+1}) - \min_{x\in\mathcal{X}} f(x) \le D_i \cdot \rho$. Thus we have

$$\bar{D}_{i+1} \le (\kappa D_i \rho/\lambda)^{1/\kappa} \le D_i/2 = D_{i+1},$$

where the second inequality holds for i that satisfies $D_i \ge (\frac{\kappa 2^\kappa \rho}{\lambda})^{\frac{1}{\kappa-1}}$. This proves the first part of the claim. For the second part, note that the definition of i_0 implies that $D_{i_0} \le 2(\frac{\kappa 2^\kappa \rho}{\lambda})^{\frac{1}{\kappa-1}}$. Therefore, as $x^\star \in \mathcal{X}_{i_0}$ and the algorithm has error $D_i \cdot \rho$, we have

$$\begin{aligned} f(x_{i_0+1}) - \min_{x\in\mathcal{X}} f(x) &\le D_{i_0} \cdot \rho \\ &\le 2(\kappa 2^\kappa/\lambda)^{\frac{1}{\kappa-1}} \rho^{\frac{\kappa}{\kappa-1}}. \end{aligned}$$

The claim now follows as $\kappa^{\frac{1}{\kappa-1}} \le 3$. □

We now proceed to the second stage. The following lemma shows that the accumulated error along the iterates $i > i_0$ is small and therefore x_T obtains the same error as x_{i_0+1} (up to constant factors).

Lemma B.2. *Assume the algorithm has error $D_i \cdot \rho$. Let i_0 be the largest i such that $D_i \geq (\frac{\kappa 2^\kappa \rho}{\lambda})^{\frac{1}{\kappa-1}}$. For all $i \geq i_0 + 1$ we have*

$$f(x_{i+1}) - f(x_i) \leq 2^{-(i-i_0)} D_{i_0} \rho.$$

In particular, for $T \geq i_0 + 1$ we have

$$f(x_T) - \min_{x \in \mathcal{X}} f(x) \leq 12(2^\kappa/\lambda)^{\frac{1}{\kappa-1}} \rho^{\frac{\kappa}{\kappa-1}}.$$

Proof. Note that as $x_i \in \mathcal{X}_i$, the guarantees of the algorithm give

$$f(x_{i+1}) - f(x_i) \leq D_i \rho = 2^{-(i-i_0)} D_{i_0} \rho.$$

For the second part of the claim, we have

$$\begin{aligned} f(x_T) - \min_{x \in \mathcal{X}} f(x) &= f(x_{i_0+1}) - \min_{x \in \mathcal{X}} f(x) + \sum_{i=i_0+1}^{T} f(x_{i+1}) - f(x_i) \\ &\leq D_{i_0}\rho + \sum_{i=i_0+1}^{T} 2^{-(i-i_0)} D_{i_0} \rho \leq 2 D_{i_0} \rho. \end{aligned}$$

The claim now follows as $D_{i_0} \leq 2(\frac{\kappa 2^\kappa \rho}{\lambda})^{\frac{1}{\kappa-1}}$ and $\kappa^{\frac{1}{\kappa-1}} \leq 3$. □

Assuming $T \geq i_0 + 1$, Theorem 2 and Theorem 3 now follow immediately from Lemma B.2. Indeed, for the case of pure-DP ($\delta = 0$), the choice of hyper-parameters in Algorithm 2 and the guarantees of Algorithm 1 (Proposition 1) imply that $\rho \leq L \cdot O(\frac{\sqrt{\log(1/\beta)} \log n_0}{\sqrt{n_0}} + \frac{T d \log(1/\beta)}{n_0 \varepsilon})$, which proves Theorem 2. Similarly, Theorem 3 follows by using the guarantees of of Algorithm 1 for approximate (ε, δ)-DP, that ism Proposition 2, which gives $\rho \leq L \cdot O\left(\frac{\sqrt{\log(1/\beta)} \log n_0}{\sqrt{n_0}} + \frac{T\sqrt{d \log(1/\delta)} \log(1/\beta)}{n_0 \varepsilon}\right)$. Note that our choice of stepsize at each iterate implies that Theorem 2 guarantees the desired utility with probability at least $1 - \beta^2$, hence the final utility guarantee holds with probability at least $1 - T\beta^2 \geq 1 - \beta$.

It remains to verify $T \geq i_0 + 1$. Note that by choosing $T \geq \frac{2 \log(D_0^{\kappa-1} \lambda/\rho)}{\kappa - 1}$, we get that $D_T \leq (\frac{\kappa 2^\kappa \rho}{\lambda})^{\frac{1}{\kappa-1}}$, hence $T \geq i_0 + 1$. As we have $\rho \geq L/\sqrt{n_0}$ (non-private error) and $D_0^{\kappa-1} \leq L/\lambda$ in our setting, we get that choosing $T = \frac{2 \log n}{\kappa - 1}$ gives the claim.

C Proofs of Section 5

In this section, we provide the proofs for our lower bound under privacy constraints for functions with growth. This section is organized as follows: we prove in Appendix C.1, the lower bounds under pure-DP and in Appendix C.2, the lower bounds under approximate-DP. Within Appendix C.1, we distinguish between $\kappa \geq 2$ (Appendix C.1.1) and $\kappa \in (1, 2)$ (Appendix C.1.2).

C.1 Proofs of Section 5.1

C.1.1 Proof of Theorem 4

As we preview in the main text, the proof combines the (non-private) information-theoretic lower bounds of Theorem 8 with the (private) lower bound on ERM of Theorem 9. Finally, we show in Proposition 4 that privately solving SCO is harder than privately solving ERM, concluding the proof of the theorem. We restate the theorem and prove these results in sequence.

Theorem 4 (Lower bound for ε-DP, $\kappa \geq 2$)**.** *Let $d \geq 1$, $\mathcal{X} = \mathbb{B}_2^d(R)$, $\mathbb{S} = \{\pm e_j\}_{j \leq d}$, $\kappa \geq 2$ and $n \in \mathbb{N}$. Let $\mathcal{P}$ be the set of distributions on $\mathbb{S}$. Assume that*

$$2^{\kappa-1} \leq \frac{L}{\lambda} \frac{1}{R^{\kappa-1}} \leq 2^{\kappa-1}\sqrt{96n} \quad \textit{and} \quad n\varepsilon \geq \frac{1}{\sqrt{3}}$$

The following lower bound holds

$$\mathfrak{M}_n(\mathcal{X}, \mathcal{P}, \mathcal{F}^\kappa, \epsilon) \geq \frac{1}{\lambda^{\frac{1}{\kappa-1}}} \tilde{\Omega}\left(\left(\frac{L}{\sqrt{n}} \right)^{\frac{\kappa}{(\kappa-1)}} + \left(\frac{Ld}{n\varepsilon} \right)^{\frac{\kappa}{\kappa-1}} \right). \tag{7}$$

Non-private lower bound We begin the proof of Theorem 4 by proving a (non-private) information-theoretic lower bound for minimizing functions with $\kappa \geq 2$-growth. We use the standard reduction from estimation to testing [see 33, Appendix A.1] in conjunction with Fano's method [42, 44].

Theorem 8 (Non-private lower bound)**.** *Let $d \geq 1$, $\mathcal{X} = \mathbb{B}_2^d(R)$, $\mathbb{S} = \{\pm e_j\}_{j \leq d}$, $\kappa \geq 2$ and $n \in \mathbb{N}$. Let $\mathcal{P}$ be the set of distributions on $\mathbb{S}$. Assume that*

$$2^{\kappa-1} \leq \frac{L}{\lambda} \frac{1}{R^{\kappa-1}} \leq 2^{\kappa-1}\sqrt{96n}.$$

The following lower bound holds

$$\mathfrak{M}_n(\mathcal{X}, \mathcal{P}, \mathcal{F}^\kappa) \gtrsim \frac{1}{\lambda^{\frac{1}{\kappa-1}}} \left(\frac{L}{\sqrt{n}} \right)^{\frac{\kappa}{(\kappa-1)}}.$$

Proof. For $\mathcal{V} \subset \{\pm 1\}^d$ let us consider the following function and distribution

$$F(x; s) := \frac{\lambda 2^{\kappa-2}}{\kappa} \|x\|_2^\kappa + \frac{L}{2}\langle x, s\rangle \text{ and } X \sim P_v \text{ implies } X_j = \begin{cases} v_j e_j & \text{w.p. } \frac{1+\delta}{2} \\ -v_j e_j & \text{w.p. } \frac{1-\delta}{2}. \end{cases}$$

Since the linear term does not affect uniform convexity, Lemma 4 in [39] guarantees that f_v is (λ, κ)-uniformly convex. Furthermore, for $s \in \mathbb{S}$

$$\|\nabla F(x; s)\|_2 \leq \lambda 2^{\kappa-2} R^{\kappa-1} + \frac{L}{2} \leq L,$$

by assumption, so the functions are L-Lipschitz and satisfy Assumption 1.

Computing the separation. As $\mathbb{E}_{P_v} S = \frac{\delta}{d} v$, we have

$$f_v(x) = \frac{\lambda 2^{\kappa-2}}{\kappa} \|x\|_2^\kappa + \frac{L\delta}{2d}\langle x, v\rangle.$$

Note that for $u \in \mathbb{R}^d, \sigma > 0$, it holds that

$$\inf_{x \in \mathbb{R}^d} \sigma \frac{\|x\|_2^\kappa}{\kappa} + \langle x, u\rangle = -\frac{1}{\kappa^\star} \left(\frac{1}{\sigma} \right)^{\frac{1}{\kappa-1}} \|u\|^{\frac{\kappa}{\kappa-1}} \text{ at } x_u^\star = -\left(\frac{1}{\sigma} \right)^{\frac{1}{\kappa-1}} \left(\frac{1}{\|u\|_2} \right)^{\frac{\kappa-2}{\kappa-1}} u.$$

To make sure that $x_u^\star \in \mathbb{B}_2^d(R)$, we require $\|u\|_2 \leq \sigma R^{\kappa-1}$. After choosing δ, we will see that this holds under the assumptions of the theorem. Let us consider the Gilbert-Varshimov packing of the hypercube: there exists $\mathcal{V} \subset \{\pm 1\}^d$ such that $|\mathcal{V}| = \exp(d/8)$ and $d_{\mathrm{Ham}}(v, v') \geq d/4$ for all $v \neq v' \in \mathcal{V}$. Let us compute the separation

$$\inf_{x \in \mathbb{B}_2^d(R)} \frac{f_v(x) + f_{v'}(x)}{2} = -\frac{1}{4\kappa^\star \lambda^{\frac{1}{\kappa-1}}} \left(\frac{L\delta}{d} \right)^{\frac{\kappa}{\kappa-1}} \left\| \frac{v + v'}{2} \right\|_2^{\frac{\kappa}{\kappa-1}}$$

Note that $\|(v + v')/2\|_2 = \sqrt{d - d_{\mathrm{Ham}}(v, v')} \leq \sqrt{3d/4}$. This yields a separation

$$d_{\mathsf{opt}}(v, v', \mathcal{X}) \geq \frac{1 - (3/4)^{\kappa/(2\kappa-2)}}{2\kappa^\star \lambda^{\frac{1}{\kappa-1}}} \left(\frac{L\delta}{\sqrt{d}} \right)^{\frac{\kappa}{\kappa-1}}.$$

Lower bounding the testing error. In the case of a multiple hypothesis test, we use Fano's method and for $V \sim \mathsf{Uni}\{\mathcal{V}\}$ and $S_1^n | V = v \overset{\text{iid}}{\sim} P_v$, Fano's inequality guarantees

$$\inf_{\psi: \mathbb{S}^n \to \mathcal{V}} \Pr(\psi(S_1^n) \neq V) \geq 1 - \frac{\mathsf{I}(S_1^n; V) + \ln 2}{\ln |\mathcal{V}|},$$

where $\mathsf{I}(X;Y)$ is the Shannon mutual information between X and Y. In our case, we have $\ln|\mathcal{V}| \geq d/8$ and $\mathsf{I}(S_1^n;V) \leq n\max_{v\neq v'} D_{\mathrm{kl}}(P_v\|P_{v'}) \leq 3n\delta^2$. In the case $d \geq 48\ln 2$, we choose $\delta = \sqrt{d/(24n)}$. We handle the one-dimensional case thereafter. For this δ, we have

$$\mathfrak{M}_n(\mathcal{X},\mathcal{P},\mathcal{F}^\kappa) \geq \frac{1-\left(\frac{3}{4}\right)^{\frac{\kappa}{2\kappa-2}}}{4\kappa^\star(24)^{\frac{\kappa}{2\kappa-2}}} \frac{1}{\lambda^{\frac{1}{\kappa-1}}}\left(\frac{L^2}{n}\right)^{\frac{\kappa}{2\kappa-2}}.$$

For this choice of δ, the assumption on n ensures that the minimum remains in $\mathbb{B}_2^d(R)$.

One-dimensional lower bound with Le Cam's method. Since Fano's method requires $d \geq 48\ln 2$, we finish the proof by providing a lower bound for $d=1$ using Le Cam's method. We use the same family of functions in one dimension, i.e. $\mathbb{S}=\{\pm 1\}$, $v\in\{\pm 1\}$ and for $\delta\in[0,1]$ define

$$F(x;s) = \frac{\lambda 2^{\kappa-2}}{\kappa}|x|^\kappa + \frac{L}{2}s\cdot x \text{ and } X\sim P_v \text{ implies } X = \begin{cases} v & \text{w.p. } \frac{1+\delta}{2} \\ -v & \text{w.p. } \frac{1-\delta}{2}. \end{cases}$$

As this is the one-dimensional analog of the previous construction, F remains L-lipschitz and f has (λ,κ)-growth. A calculation yields that the separation is

$$d_{\mathsf{opt}}(1,-1,\mathcal{X}) \geq \frac{1}{2\lambda^{\frac{1}{\kappa-1}}}(L\delta)^{\frac{\kappa}{\kappa-1}},$$

where we used that $\kappa^\star\in[1,2]$. For $V\sim\mathsf{Uni}\{-1,1\}$ and $S_1^n|V=v \overset{\mathrm{iid}}{\sim} P_v$. Le Cam's lemma in conjunction with Pinsker's inequality yields that

$$\inf_{\psi:\mathbb{S}^n\to\{-1,1\}} \Pr(\psi(S_1^n)\neq V) = \frac{1}{2}(1-\|P_1^n - P_{-1}^n\|_{\mathrm{TV}}) \geq \frac{1}{2}(1-\sqrt{\tfrac{n}{2}D_{\mathrm{kl}}(P_1\|P_{-1})}).$$

In our case, we have $D_{\mathrm{kl}}(P_1\|P_{-1}) = \delta\ln\frac{1+\delta}{1-\delta} \leq 3\delta^2$ for $\delta\in[0,1/2]$. We set $\delta = 1/\sqrt{6n}$, which yields the final result in one dimension

$$\mathfrak{M}_n([-1,1],\mathcal{P},\mathcal{F}_{d=1}^\kappa) = \Omega\left(\frac{1}{\lambda^{\frac{1}{\kappa-1}}}\left(\frac{L}{\sqrt{n}}\right)^{\frac{\kappa}{\kappa-1}}.\right)$$

□

Privatizing the lower bound via a packing argument We now show how this construction yields a private lower bound via a packing argument. For $d\geq 1$, considering the ERM problem, the following private lower bound holds.

Theorem 9 (Private lower bound for ERM)**.** *Let $d\geq 1, \mathcal{X}=\mathbb{B}_2^d(R)$, $\mathbb{S}=\{\pm e_j\}_{j\leq d}$, $\kappa\geq 2$ and $n\in\mathbb{N}$. Let $\mathcal{P}$ be the set of distributions on $\mathbb{S}$. Assume that*

$$2^{\kappa-1} \leq \frac{L}{\lambda}\frac{1}{R^{\kappa-1}} \leq 2^{\kappa-1}\sqrt{96n}.$$

Then any ε-DP algorithm A has

$$\sup_{\mathcal{S}\in\mathbb{S}^n}\mathbb{E}\left[f_{\mathcal{S}}(\mathsf{A}(\mathcal{S})) - \min_{x\in\mathcal{X}} f_{\mathcal{S}}(x)\right] \gtrsim \frac{1}{\lambda^{\frac{1}{\kappa-1}}}\left(\frac{Ld}{n\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}.$$

Proof. First, note that it is enough to prove the following lower bound

$$\sup_{\mathcal{S}\in\mathbb{S}^n}\mathbb{E}\left[\|\mathsf{A}(\mathcal{S})-x^\star\|_2\right] \gtrsim \frac{1}{\lambda^{\frac{1}{\kappa-1}}}\left(\frac{Ld}{n\varepsilon}\right)^{\frac{1}{\kappa-1}}. \tag{9}$$

Indeed, this implies that

$$\begin{aligned}\sup_{\mathcal{S}\in\mathbb{S}^n}\mathbb{E}\left[f(\mathsf{A}(\mathcal{S})) - \min_{x\in\mathcal{X}} f(x)\right] &\geq \frac{\lambda}{\kappa}\sup_{\mathcal{S}\in\mathbb{S}^n}\mathbb{E}\left[\|\mathsf{A}(\mathcal{S})-x^\star\|_2^\kappa\right] \\ &\gtrsim \frac{1}{\kappa\lambda^{\frac{1}{\kappa-1}}}\left(\frac{Ld}{n\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}.\end{aligned}$$

Let us now prove the lower bound (9). To this end, we consider the function $F(x;s) := \frac{\lambda 2^{\kappa-2}}{\kappa}\|x\|_2^\kappa + \frac{L}{2}\langle x, s\rangle$ where $\|s\|_2 \le 1$. We now construct M datasets $\mathcal{S}_1, \dots, \mathcal{S}_M$ as follows. Let $v_1, \dots, v_M \in \left\{\pm\frac{1}{\sqrt{d}}\right\}^d$ be the Gilbert-Varshimov packing of the hypercube: that is, $M \ge \exp(d/8)$ and $d_{\mathrm{Ham}}(v_i, v_j) \ge d/4$ for all $i \ne j$. We define $\mathcal{S}_i = (\underbrace{v_i, \dots, v_i}_{d/20\varepsilon}, 0, \dots, 0)$. Note that $d_{\mathrm{Ham}}(S_i, S_j) \le d/20\varepsilon$ and that $f(x;\mathcal{S}_i) = \frac{\lambda 2^{\kappa-2}}{\kappa}\|x\|_2^\kappa + \frac{L}{2}\frac{d}{20n\varepsilon}\langle x, v_i\rangle$, hence

$$x_i^\star = -\left(\frac{1}{\lambda 2^{\kappa-2}}\right)^{\frac{1}{\kappa-1}}\left(\frac{40n\varepsilon}{Ld}\right)^{\frac{\kappa-2}{\kappa-1}}\frac{Ld}{40n\varepsilon}v_i.$$

Therefore we have

$$\begin{aligned}\|x_i^\star - x_j^\star\|_2^2 &\ge \left(\frac{1}{\lambda 2^{\kappa-2}}\right)^{\frac{2}{\kappa-1}}\left(\frac{40n\varepsilon}{Ld}\right)^{\frac{2(\kappa-2)}{\kappa-1}}\frac{L^2d^2}{1600n^2\varepsilon^2} \\ &\gtrsim \left(\frac{1}{\lambda 2^{\kappa-2}}\right)^{\frac{2}{\kappa-1}}\left(\frac{Ld}{n\varepsilon}\right)^{\frac{2}{\kappa-1}} := \rho^2.\end{aligned}$$

We are now ready to finish the proof using packing-based arguments [27]. Assume by contradiction there is an ε-DP algorithm A such that

$$\sup_{1\le i\le M}\mathbb{E}\left[\|\mathsf{A}(\mathcal{S}_i) - x_i^\star\|_2\right] \le \rho/20.$$

Let $B_i = \{x \in \mathcal{X} : \|x - x_i^\star\|_2 \le \rho/2\}$. Note that the sets B_i are disjoint and that Markov inequality implies

$$\Pr(\mathsf{A}(\mathcal{S}_i) \in B_i) = \Pr(\|\mathsf{A}(\mathcal{S}_i) - x_i^\star\|_2 \le \rho/2) \ge 9/10.$$

Thus, the privacy constraint now gives

$$\begin{aligned}1 &\ge \sum_{i=1}^M \Pr(\mathsf{A}(x_1) \in B_i) \\ &\ge \Pr(\mathsf{A}(x_1) \in B_1) + e^{-d/20}\sum_{i=2}^M \Pr(\mathsf{A}(x_i) \in B_i) \\ &\ge \frac{9}{10}(1 + e^{-d/20}(M-1)),\end{aligned}$$

where the second inequality follows since $d_{\mathrm{Ham}}(\mathcal{S}_i, \mathcal{S}_j) \le d/20\varepsilon$. This gives a contradiction for $d \ge 20$ as $M \ge \exp(d/8)$. For $d = 1$, we can repeat the same arguments with $M = 2$ to get the desired lower bound. □

Reduction from ε-DP ERM to ε-DP SCO We conclude the proof of the theorem by proving that SCO under privacy constraints is strictly harder than ERM. This is similar to Appendix C in [8] but we require it for pure-DP constraints. We make this formal in here.

We have the following lemma.

Proposition 4. *Let $0 < \beta \le 1/n$. Assume A is an $\frac{\varepsilon}{2\log(2/\beta)}$-DP algorithm that for a sample $\mathcal{S} = (S_1, \dots, S_n)$ with $S_1^n \overset{\text{iid}}{\sim} P$ achieves with probability $1 - \beta/2$ error*

$$f(\mathsf{A}(\mathcal{S})) - \min_{x\in\mathcal{X}} f(x) \le \gamma.$$

Then there is an ε-DP algorithm A' such that for any dataset $\mathcal{S} \in \mathbb{S}^n$ has with probability $1 - \beta$,

$$f_{\mathcal{S}}(\mathsf{A}'(\mathcal{S})) - \min_{x\in\mathcal{X}} f_{\mathcal{S}}(x) \le \gamma.$$

Proof. Given the algorithm A, we define A' as follows. For an input $\mathcal{S} \in \mathbb{S}^n$, let $P_{\mathcal{S}}$ be the empirical distribution of $\mathcal{S}$. Then, A' proceeds as follows:

1. Sample a new dataset $\mathcal{S}_1 = (S'_1, \dots, S'_n)$ where $S'_i \sim P_{\mathcal{S}}$
2. If there is a sample S_i that was sampled more than $k = 2\log(2/\beta)$ times, return 0
3. Else, return $\mathsf{A}(\mathcal{S}_1)$

We need to prove that A' is ε-DP and that it has the desired utility. For utility, note that A' returns 0 at step 2 with probability at most $\beta/2$, since we have for every $1 \le i \le n$

$$\begin{aligned}\Pr(s_i \text{ used more than } k \text{ times}) &= \Pr\left(\sum_{j=1}^{n} Z_i \ge k\right) \\ &\le 2^{-k} \le \beta^2/2,\end{aligned}$$

where $Z_j \sim \mathsf{Bernoulli}(p)$ with $p = 1/n$, and the second inequality follows from Chernoff [36, Thm. 4.4] and $\beta \le 1/10$. Applying a union bound over all samples, we get that step 2 returns 0 with probability at most $\beta/2$ as $\beta \le 1/n$. Moreover, Algorithm A fails with probability at most $\beta/2$. Therefore, as $f_{\mathcal{S}}(x) = \mathbb{E}_{S\sim P_{\mathcal{S}}}[F(x;S)]$, we have with probability at least $1-\beta$,

$$f_{\mathcal{S}}(\mathsf{A}'(\mathcal{S})) - \min_{x\in\mathcal{X}} f_{\mathcal{S}}(x) \le \gamma.$$

Let us now prove privacy. Assume we run algorithm A' on two neighboring datasets $\mathcal{S}, \mathcal{S}'$, and let $\mathcal{S}_1, \mathcal{S}'_1$ be the datasets produced at step 1. Let B denote the event that there was a sample s_i that was used more than k times (note that this does not depend on the input). Then for any measurable $\mathcal{O}$,

$$\begin{aligned}\Pr(\mathsf{A}'(\mathcal{S}) \in \mathcal{O}) &= \Pr(\mathsf{A}'(\mathcal{S}) \in \mathcal{O} \mid B)\Pr(B) + \Pr(\mathsf{A}'(\mathcal{S}) \in \mathcal{O} \mid B^c)\Pr(B^c) \\ &\le e^{\varepsilon}\Pr(\mathsf{A}'(\mathcal{S}') \in \mathcal{O} \mid B)\Pr(B) + \Pr(\mathsf{A}'(\mathcal{S}') \in \mathcal{O} \mid B^c)\Pr(B^c) \\ &\le e^{\varepsilon}\Pr(\mathsf{A}'(\mathcal{S}') \in \mathcal{O}),\end{aligned}$$

where the first inequality follows from group privacy since $d_{\mathrm{Ham}}(\mathcal{S}_1, \mathcal{S}'_1) \le k$ and A is ε/k-DP. This completes the proof.

□

C.1.2 Proof of Theorem 5

Theorem 5 (Lower bound for ε-DP, $\kappa \in (1,2]$)**.** *Let $d = 1$, $\mathbb{S} = \{-1, +1\}$, $\kappa \in (1,2]$, $\lambda = 1$, $L = 2$, and $n \in \mathbb{N}$. There exists a collection of distributions $\mathcal{P}$ such that, whenever $n\varepsilon \ge 1/\sqrt{3}$, it holds that*

$$\mathfrak{M}_n([-1,1], \mathcal{P}, \mathcal{F}^{\kappa}_{d=1}, \epsilon) = \Omega\left\{\left(\frac{1}{\sqrt{n}}\right)^{\frac{\kappa}{\kappa-1}} + \left(\frac{1}{n\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right\}. \tag{8}$$

Proof. We follow the same reduction that we used in the proof of Theorem 8. For $\delta \in [0, 1/2]$, we again consider $P_v = 1$ with probability $\frac{1+\delta v}{2}$ and -1 otherwise. For $a \in [0,1]$ to be defined later, we construct the following function

$$F(x;+1) = \begin{cases} |x-a| & \text{if } x \le a \\ |x-a|^{\kappa} & \text{if } x \ge a \end{cases} \quad \text{and} \quad F(x;-1) = \begin{cases} |x+a|^{\kappa} & \text{if } x \le -a \\ |x+a| & \text{if } x \ge -a \end{cases}$$

Computing the separation. First, let us compute the separation $d_{\mathsf{opt}}(v, v', \mathcal{X})$. We will then choose a to ensure f_v has κ-growth. By symmetry, assume $v = 1$. f_v is increasing on $[a,1]$ and decreasing on $[-1,-a]$, thus the minimum belongs to $[-a,a]$ and by inspection, is attained at $x = a$ with value $a(1-\delta)$. Similarly, the minimum of $f_{+1}(x) + f_{-1}(x)$ is attained on $[-a,a]$ with value $2a$. This yields

$$d_{\mathsf{opt}}(v, v', \mathcal{X}) = 2a - 2a(1-\delta) = 2a\delta.$$

Let us now pick a such that f_v has κ-growth. Again, by symmetry we only treat the $v = 1$ case. We have

$$\text{for } x \ge a,\ f_v(x) - f_v^{\star} = \frac{1+\delta}{2}(x-a)^{\kappa} + \frac{1-\delta}{2}(x+a) - a(1-\delta) = \frac{1+\delta}{2}(x-a)^{\kappa} + \frac{1-\delta}{2}(x-a) \ge |x-a|^{\kappa},$$

where the last inequality is because $(x-a) \leq 1$ and so $(x-a) \geq (x-a)^\kappa$ for $\kappa > 1$. In the second case, we have

$$\text{for } x \in [-a, a], f_v(x) - f_v^\star = \delta(a - x).$$

It holds that $\delta(a-x) \geq (a-x)^\kappa$ for all $x \in [-a, a]$ iff $a \leq \frac{1}{2}\delta^{\frac{1}{\kappa-1}}$. As a result, we set $a = \frac{1}{2}\delta^{\frac{1}{\kappa-1}}$. Finally, for $x \in [-1, -a]$, we define

$$h(x) \coloneqq \frac{1+\delta}{2}|x-a| + \frac{1-\delta}{2}|x+a|^\kappa - a(1-\delta) - \frac{1}{\kappa}|x-a|^\kappa \text{ for } x \in [-1, -a].$$

We wish to prove that $h(x) \geq 0$. First of, note that $h(-a) = \delta^{\frac{\kappa}{\kappa-1}}(\frac{1}{2} + \frac{1}{2} - \frac{1}{\kappa}) > 0$, whenever $\kappa > 1$. Let us show that $h(x)$ is decreasing on $[-1, -a]$ which suffices to conclude the proof. We have

$$h'(x) = -\frac{1+\delta}{2} - \frac{\kappa(1-\delta)}{2}|x+a|^{\kappa-1} + |x-a|^{\kappa-1}.$$

First of, note that $h'(-a) = -\frac{1+\delta}{2} + \delta \leq 0$ and $h'(-1) < 0$, thus it suffices to show that if h' has an extremum then is it negative. An extremum of this function is a point $x^\star$ such that

$$|a - x^\star| = \left(\frac{\kappa(1-\delta)}{2}\right)^{\frac{1}{\kappa-2}} |a + x^\star|,$$

which yields that

$$h'(x^\star) = |a + x^\star|^{\kappa-1}\left(\frac{\kappa(1-\delta)}{2}\right)\left[\left(\frac{\kappa(1-\delta)}{2}\right)^{\frac{1}{\kappa-2}} - 1\right] - \frac{1+\delta}{2} \leq 0,$$

as $\kappa \leq 2$. This calculation shows that f_v has $(1, \kappa)$-growth. Finally note that the function is $\kappa \leq 2$-Lipschitz as desired.

Lower bounding the testing error. It remains to choose the value of δ. Since we require a lower bound under privacy constraints, in contrast to the one-dimensional section of the proof of Theorem 8, we require the following privatized version of Le Cam's lemma from [6]

Proposition 5. *[6, Thm. 2] Let $\mathsf{A} \in \mathcal{A}^\varepsilon$ be an ϵ-DP mechanism from $\mathbb{S}^n \to \mathcal{X}$. It holds that*

$$\inf_{\psi:\mathcal{X}\to\{-1,1\}} \inf_{\mathsf{A}\in\mathcal{A}^\epsilon} \Pr(\psi(\mathsf{A}(S_1^n)) \neq V) \geq \frac{1}{2}\big(1 - \min\big\{2n\varepsilon\|P_1 - P_{-1}\|_{\mathrm{TV}}, \|P_{-1}^n - P_1^n\|_{\mathrm{TV}}\big\}\big).$$

With this result, we set $\delta = \max\{1/\sqrt{6n}, 1/(2\sqrt{3}n\varepsilon)\}$ and lower bound $\max\{a, b\}$ by $a + b$ for readability, which concludes the proof of the theorem. □

C.2 Proof for Section 5.2

Proposition 3 (Solving ERM with κ-growth implies solving any convex ERM)**.** *Let $\kappa \geq 2$. Assume there exists an (ϵ, δ) mechanism A such that for any L-Lipschitz loss G on $\mathcal{Y}$ and dataset $\mathcal{S}$ such that $g_{\mathcal{S}}(x) \coloneqq \frac{1}{n}\sum_{s\in\mathcal{S}} G(x; s)$ exhibits (λ, κ)-growth, the mechanism achieves excess loss*

$$\mathbb{E}[g_{\mathcal{S}}(\mathsf{A}(\mathcal{S}, G, \mathcal{Y}))] - \inf_{y'\in\mathcal{Y}} g_{\mathcal{S}}(y') \leq \frac{1}{\lambda^{\frac{1}{\kappa-1}}}\Delta(n, L, \epsilon, \delta).$$

Then, we can construct an (ε, δ)-DP mechanism A' such that for any L-Lipschitz loss f, the mechanism achieves excess loss

$$\mathbb{E}[f_{\mathcal{S}}(\mathsf{A}'(\mathcal{S}))] - \inf_{x'\in\mathcal{X}} f_{\mathcal{S}}(x') \leq O\left(D[\Delta(n, L, \epsilon/k, \delta/k)]^{\frac{\kappa-1}{\kappa}}\right),$$

where k is the smallest integer such that $k \geq \log\left[\frac{\kappa^{\frac{1}{\kappa-1}} L^{\frac{\kappa}{\kappa-1}}}{2^{2\kappa-3}\Delta(n, L, \varepsilon/k, \delta/k)}\right]$.

Proof of Proposition 3. Let us first show how to construct the mechanism A'. Let $k \in \mathbb{N}$ be such that $k \geq \log\left[\frac{\kappa^{\frac{1}{\kappa-1}} L^{\frac{\kappa}{\kappa-1}}}{2^{2\kappa-3}\Delta(n,L,\varepsilon/k,\delta)}\right]$ and let $\{\lambda_i\}_{i\in[k]}$ be a collection of positive scalars. Set $x_0 \in \mathcal{X}$, for $i \in \{1,\dots,k\}$

$$\text{define } G_i(x;s) = F(x;s) + \frac{\lambda_i \cdot 2^{\kappa-2}}{\kappa}\|x - x_{i-1}\|_2^\kappa, \mathcal{Y}_i := \left\{x \in \mathcal{X} : \|x - x_{i-1}\|_2 \leq \left(\frac{L\kappa}{\lambda_i 2^{\kappa-2}}\right)^{\frac{1}{\kappa-1}}\right\}$$

$$\text{and set } x_i = \mathsf{A}(\mathcal{S}, G_i, \mathcal{Y}_i), \text{ with privacy } (\varepsilon/k, \delta/k).$$

Finally, define $\mathsf{A}'(\mathcal{S}) = x_k$. Standard composition theorems [20] guarantee that A' is (ε,δ)-DP. Let us analyze its utility; we drop the dependence of Δ on other variables when clear from context. First of, since κ is a constant, note that G_i is $c_0 L$-Lipschitz with $c_0 < \infty$ a numerical constant. For simplicity, we define $g_i(x) := \frac{1}{n}\sum_{s\in\mathcal{S}} G_i(x;s)$ and $x_i^\star = \operatorname{argmin}_{x\in\mathcal{Y}_i} g_i(x)$. It holds that g_i is $(\lambda_i 2^{\kappa-2}, \kappa)$-uniformly-convex and thus the following growth condition holds

$$\frac{\lambda_i}{\kappa}\mathbb{E}\|x_i - x_i^\star\|_2^\kappa \leq \mathbb{E}[g_i(x_i)] - g_i(x_i^\star) \leq \frac{1}{\lambda_i^{\frac{1}{\kappa-1}}}\Delta.$$

Also note that for any point $y \in \mathcal{Y}_i$, it holds that

$$f_\mathcal{S}(x_i^\star) - f(y) \leq \frac{\lambda_i 2^{\kappa-2}}{\kappa}\|x_{i-1} - y\|_2^\kappa.$$

Finally, let us bound the distance to the optimum of $f_\mathcal{S}$ at the final iterate. We have

$$\frac{\lambda_k}{\kappa}\|x_k - x_k^\star\|_2^\kappa \leq g_k(x_k) - g_k(x_k^\star) \leq c_0 L\|x_k - x_k^\star\|_2 \text{ which yields } \|x_k - x_k^\star\|_2 \leq \left(\frac{c_0 L\kappa}{\lambda_k}\right)^{\frac{1}{\kappa-1}}.$$

Let us put the pieces together: for $\lambda > 0$ to be determined later and $\nu = \kappa - 1$, set $\lambda_i = 2^{-\nu i}\lambda$. After k rounds and denoting $x_0^\star = \inf_{x\in\mathcal{X}} f_\mathcal{S}(x)$, we have

$$\begin{aligned}
\mathbb{E}[f_\mathcal{S}(x_k)] - f_\mathcal{S}(x^\star) &= \sum_{i=1}^k \mathbb{E}\big[f_\mathcal{S}(x_i^\star) - f_\mathcal{S}(x_{i-1}^\star)\big] + \mathbb{E}[f_\mathcal{S}(x_k) - f_\mathcal{S}(x_k^\star)] \\
&\leq \sum_{i=1}^k \frac{\lambda_i 2^{\kappa-2}}{\kappa}\mathbb{E}\|x_{i-1} - x_{i-1}^\star\|_2^\kappa + L\left(\frac{c_0 L\kappa}{\lambda_k}\right)^{\frac{1}{\kappa-1}} \\
&\leq \frac{\lambda D^\kappa}{\kappa} + \sum_{i=2}^k \frac{\lambda_i 2^{\kappa-2}}{\lambda_{i-1}^{\frac{\kappa}{\kappa-1}}}\Delta + L\left(\frac{c_0 L\kappa}{\lambda_k}\right)^{\frac{1}{\kappa-1}} \\
&= \frac{\lambda D^\kappa}{\kappa} + \frac{\Delta 2^{\kappa-2}}{\lambda^{\frac{1}{\kappa-1}}}\sum_{i=2}^k 2^{-\frac{\nu}{\kappa-1}(i-\kappa)} + L\left(\frac{c_0 L\kappa}{\lambda}\right)^{\frac{1}{\kappa-1}} 2^{-\frac{\nu}{\kappa-1}k} \\
&\leq \frac{\lambda D^\kappa}{\kappa} + 2^{2\kappa-3}\frac{\Delta}{\lambda^{\frac{1}{\kappa-1}}} + \frac{\kappa^{\frac{1}{\kappa-1}}(c_0 L)^{\frac{\kappa}{\kappa-1}} 2^{-k}}{\lambda^{\frac{1}{\kappa-1}}}.
\end{aligned}$$

Finally, note that

$$k \geq \left\lceil \log\left[\frac{\kappa^{\frac{1}{\kappa-1}}(c_0 L)^{\frac{\kappa}{\kappa-1}}}{2^{2\kappa-3}\Delta}\right]\right\rceil \text{ so that } \frac{\kappa^{\frac{1}{\kappa-1}}(c_0 L)^{\frac{\kappa}{\kappa-1}} 2^{-k}}{\lambda^{\frac{1}{\kappa-1}}} \leq 2^{2\kappa-3}\frac{\Delta}{\lambda^{\frac{1}{\kappa-1}}}.$$

It then holds that

$$\mathbb{E}[f_\mathcal{S}(x_k)] - f_\mathcal{S}(x^\star) \leq \lambda\frac{D^\kappa}{\kappa} + 4^{\kappa-1}\Delta\frac{1}{\lambda^{\frac{1}{\kappa-1}}}.$$

It remains to pick λ to minimize the upper bound above. A calculation yields that for $a, b \geq 0$

$$\inf_{\nu \geq 0} a\nu + \frac{b}{\nu^{\frac{1}{\kappa-1}}} = (\kappa-1)^{1/\kappa} a^{1/\kappa} b^{(\kappa-1)/\kappa} \left[\kappa - 1 + \frac{1}{\kappa-1}\right] \text{ at } \nu^\star = \left(\frac{b}{a(\kappa-1)}\right)^{\frac{\kappa-1}{\kappa}}.$$

Setting $\lambda = 4^{\frac{(\kappa-1)^2}{\kappa}} (\frac{\Delta\kappa}{D^\kappa(\kappa-1)})^{(\kappa-1)/\kappa}$ yields the regret bound

$$\mathbb{E}[f_{\mathcal{S}}(x_k)] - f_{\mathcal{S}}(x^\star) \leq O(1) D \Delta^{\frac{\kappa-1}{\kappa}}.$$

□

Proof. Consider the reduction of Proposition 3. For $c_1 < \infty$ to be determined later, assume by contradiction that there exists an (ε, δ) mechanism such that

$$\Delta(n, L, \varepsilon, \delta) \leq c_1 \left(\frac{L\sqrt{d}}{n\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}.$$

Setting $k = \lceil 4 \log(n\varepsilon/\sqrt{d}) \log\log((n\varepsilon/\sqrt{d})^{\kappa/(\kappa-1)}) \rceil$, the condition holds and the result of Proposition 3 guarantees that there exists a numerical constant $c_2 < \infty$ and a mechanism A' such that

$$\mathbb{E}[f_{\mathcal{S}}(\mathsf{A}'(\mathcal{S}))] - \inf_{x' \in \mathcal{X}} f_{\mathcal{S}}(x') \leq c_2 c_1^{\frac{\kappa-1}{\kappa}} kD \frac{L\sqrt{d}}{n\varepsilon}.$$

However, Theorem 5.3 in [7] guarantees that there exists $c_3 > 0$ such that for any (ε, δ)-DP mechanism A'', it must hold

$$c_3 LD \frac{\sqrt{d}}{n\varepsilon} \leq \mathbb{E}[f_{\mathcal{S}}(\mathsf{A}''(\mathcal{S}))] - f_{\mathcal{S}}(x^\star).$$

Setting $c_1 = \frac{1}{2}\left(\frac{c_3}{kc_2}\right)^{\frac{\kappa}{\kappa-1}}$ yields a contradiction and the desired lower bound by noting that k consists only of log factors. □