Private optimization in the interpolation regime: faster rates and hardness results

Hilal Asi*† asi@stanford.edu
Karan Chadha*† knchadha@stanford.edu
Gary Cheng*† chenggar@stanford.edu
John Duchi†‡ jduchi@stanford.edu

November 1, 2022

Abstract

In non-private stochastic convex optimization, stochastic gradient methods converge much faster on interpolation problems—problems where there exists a solution that simultaneously minimizes all of the sample losses—than on non-interpolating ones; we show that generally similar improvements are impossible in the private setting. However, when the functions exhibit quadratic growth around the optimum, we show (near) exponential improvements in the private sample complexity. In particular, we propose an adaptive algorithm that improves the sample complexity to achieve expected error α from $\frac{d}{\varepsilon\sqrt{\alpha}}$ to $\frac{1}{\alpha^\rho} + \frac{d}{\varepsilon}\log\left(\frac{1}{\alpha}\right)$ for any fixed $\rho > 0$, while retaining the standard minimax-optimal sample complexity for non-interpolation problems. We prove a lower bound that shows the dimension-dependent term is tight. Furthermore, we provide a supereffiiciency result which demonstrates the necessity of the polynomial term for adaptive algorithms: any algorithm that has a polylogarithmic sample complexity for interpolation problems cannot achieve the minimax-optimal rates for the family of non-interpolation problems.

1 Introduction

We study differentially private stochastic convex optimization (DP-SCO), where given a dataset $\mathcal{S} = S_1^n \overset{\text{iid}}{\sim} P$ we wish to solve

$$\begin{aligned} &\text{minimize} \;\; f(x) = \mathbb{E}_P[F(x; S)] = \int_\Omega F(x; s) dP(s) \\ &\text{subject to} \;\; x \in \mathcal{X}, \end{aligned} \tag{1}$$

while guaranteeing differential privacy. In problem (1), $\mathcal{X} \subset \mathbb{R}^d$ is the parameter space, Ω is a sample space, and $\{F(\cdot; s) : s \in \mathbb{S}\}$ is a collection of convex losses. We study the interpolation setting, where there exists a solution that simultaneously minimizes all of the sample losses.

Interpolation problems are ubiquitous in machine learning applications: for example, least squares problems with consistent solutions [SV09, NWS14], and problems with over-parametrized models where a perfect predictor exists [MBB18, BHM18, BRT19]. This has led to a great deal of work on the advantages and implications of interpolation [SST10, CSSS11, BHM18, BRT19].

For non-private SCO, interpolation problems allow significant improvements in convergence rates over generic problems [SST10, CSSS11, MBB18, VBS19, WS21]. For general convex functions, [SST10] develop algorithms that obtain $O(\frac{1}{n})$ sub-optimality, improving over the minimax-optimal rate $O(\frac{1}{\sqrt{n}})$

*Equal contribution, author order alphabetical
†Electrical Engineering Department, Stanford University
‡Statistics Department, Stanford University

for non-interpolation problems. Even more dramatic improvements are possible when the functions exhibit growth around the minimizer, as [VBS19] show that SGD achieves exponential rates in this setting compared to polynomial rates without interpolation. [AD19, ACCD20, CCD22] extend these fast convergence results to model-based optimization methods.

Despite the recent progress and increased interest in interpolation problems, in the private setting they remain poorly understood. In spite of the substantial progress in characterizing the tight convergence guarantees for a variety of settings in DP optimization [BST14, BFTT19, FKT20, AFKT21, ALD21], we have little understanding of private optimization in the growing class of interpolation problems.

Given (i) the importance of differential privacy and interpolation problems in modern machine learning, (ii) the (often) paralyzingly slow rates of private optimization algorithms, and (iii) the faster rates possible for non-private interpolation problems, the interpolation setting provides a reasonable opportunity for significant speedups in the private setting. This motivates the following two questions: first, is it possible to improve the rates for DP-SCO in the interpolation regime? And, what are the optimal rates?

1.1 Our contributions

We answer both questions. In particular, we show that

1. **No improvements in general** (Section 3): our first result is a hardness result demonstrating that the rates cannot be improved for DP-SCO in the interpolation regime with general convex functions. More precisely, we prove a lower bound of $\Omega(\frac{d}{n\varepsilon})$ on the excess loss for pure differentially private algorithms. This shows that existing algorithms achieve optimal private rates for this setting.

2. **Faster rates with growth** (Section 4): when the functions exhibit quadratic growth around the minimizer, that is, $f(x) - f(x^\star) \geq \lambda\|x - x^\star\|_2^2$ for some $\lambda > 0$, we propose an algorithm that achieves near-exponentially small excess loss, improving over the polynomial rates in the non-interpolation setting. Specifically, we show that the sample complexity to achieve expected excess loss $\alpha > 0$ is $O(\frac{1}{\alpha^\rho} + \frac{d}{\varepsilon}\log\left(\frac{1}{\alpha}\right))$ for pure DP and $O(\frac{1}{\alpha^\rho} + \frac{\sqrt{d\log(1/\delta)}}{\varepsilon}\log\left(\frac{1}{\alpha}\right))$ for (ε, δ)-DP, for any fixed $\rho > 0$. This improves over the sample complexity for non-interpolation problems with growth which is $O(\frac{1}{\alpha} + \frac{d}{\varepsilon\sqrt{\alpha}})$. We also present new algorithms that improve the rates for interpolation problems with the weaker κ-growth assumption [ALD21] for $\kappa > 2$ where we achieve excess loss $O((\frac{1}{\sqrt{n}} + \frac{d}{n\varepsilon})^{\frac{\kappa}{\kappa-2}})$, compared to the previous bound $O((\frac{1}{\sqrt{n}} + \frac{d}{n\varepsilon})^{\frac{\kappa}{\kappa-1}})$ without interpolation.

3. **Adaptivity to interpolation** (Section 4.3): While these improvements for the interpolation regime are important, practitioners using these methods in practice cannot identify whether the dataset they are working with is an interpolating one or not. Thus, it is crucial that these algorithms do not fail when given a non-interpolating dataset. We show that our algorithms are adaptive to interpolation, obtaining these better rates for interpolation while simultaneously retaining the standard minimax optimal rates for non-interpolation problems.

4. **Tightness** (Section 5): finally, we provide a lower bound and a super-efficiency result that demonstrate the (near) tightness of our upper bounds showing sample complexity $\Omega(\frac{d}{\varepsilon}\log\left(\frac{1}{\alpha}\right))$ is necessary for interpolation problems with pure DP. Moreover, our super-efficiency result shows that the polynomial dependence on $1/\alpha$ in the sample complexity is necessary for adaptive algorithms: any algorithm that has a polylogarithmic sample complexity for interpolation problems cannot achieve minimax-optimal rates for non-interpolation problems.

1.2 Related work

Over the past decade, a lot of works [CMS11, DJW13, ST13, BST14, ACG+16, BFTT19, FKT20, AFKT21, ADF+21, BFGT20] have studied the problem of private convex optimization. [CMS11] and

[BST14] study the closely related problem of differentially private empirical risk minimization (DP-ERM) where the goal is to minimize the empirical loss, and obtain (minimax) optimal rates of $d/n\varepsilon$ for pure DP and $\sqrt{d\log(1/\delta)}/n\varepsilon$ for (ε,δ)-DP. Recently, more papers have moved beyond DP-ERM to privately minimizing the population loss (DP-SCO) [BFTT19, FKT20, AFKT21, ADF+21, BGN21, ALD21]. [BFTT19] was the first paper to obtain the optimal rate $1/\sqrt{n}+\sqrt{d\log(1/\delta)}/n\varepsilon$ for (ε,δ)-DP, and subsequent papers develop more efficient algorithms that achieve the same rates [FKT20, BFGT20]. Moreover, other papers study DP-SCO under different settings including non-Euclidean geometry [AFKT21, ADF+21], heavy-tailed data [WXDX20], and functions with growth [ALD21]. However, to the best of our knowledge, there has not been any work in private optimization that studies the problem in the interpolation regime.

On the other hand, the optimization literature has witnessed numerous papers on the interpolation regime [SST10, CSSS11, MBB18, VBS19, LB20, WS21]. [SST10] propose algorithms that roughly achieve the rate $1/n+\sqrt{f^\star/n}$ for smooth and convex functions where $f^\star=\min_{x\in\mathcal{X}} f(x)$. In the interpolation regime with $f^\star=0$, this result obtains loss $1/n$ improving over the standard $1/\sqrt{n}$ rate for non-interpolation problems. Moreover, [VBS19] studied the interpolation regime for functions with growth and show that SGD enjoys linear convergence (exponential rates). More recently, several papers investigated and developed acceleration-based algorithms in the interpolation regime [LB20, WS21].

2 Preliminaries

We begin with notation that will be used throughout the paper and provide some standard definitions from convex analysis and differential privacy.

Notation We let n denote the sample size and d the dimension. We let x denote the optimization variable and $\mathcal{X}\subset\mathbb{R}^d$ the constraint set. s are samples from Ω, and S is an Ω-valued random variable. For each sample $s\in\Omega$, $F(\cdot;s):\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ is a closed convex function. Let $\partial F(x;s)$ denote the subdifferential of $F(\cdot;s)$ at x. We let Ω^n denote the collection of datasets $\mathcal{S}=(s_1,\dots,s_n)$ with n data points from Ω. We let $f_{\mathcal{S}}(x)\coloneqq\frac{1}{n}\sum_{s\in\mathcal{S}}F(x,s)$ denote the empirical loss and $f(x)\coloneqq\mathbb{E}[F(x;S)]$ denote the population loss. The distance of a point to a set is $\operatorname{dist}(x,Y)=\min_{y\in Y}\|x-y\|_2$. We use $\operatorname{Diam}(\mathcal{X})=\sup_{x,y\in\mathcal{X}}\|x-y\|_2$ to denote the diameter of parameter space $\mathcal{X}$ and use D as a bound on the diameter of our parameter space.

We recall the definition of (ε,δ)-differential privacy.

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

$$P(M(\mathcal{S})\in\mathcal{O})\le e^\varepsilon P(M(\mathcal{S}')\in\mathcal{O})+\delta.$$

We define ε-differential privacy (ε-DP) to be $(\varepsilon,0)$-differential privacy.

We now recall a couple of standard convex analysis definitions.

Definition 2.2.

1. *A function $h:\mathcal{X}\to\mathbb{R}$ is* L-Lipschitz *if for all $x,y\in\mathcal{X}$*

$$|h(x)-h(y)|\le L\|x-y\|_2.$$

   Equivalently, a function is L-Lipschitz *if $\|\nabla f(x)\|_2\le L$ for all $x\in\mathcal{X}$.*

2. *A function h is* H-smooth *if it has H-Lipschitz gradient: for all $x,y\in\mathcal{X}$*

$$\|\nabla h(x)-\nabla h(y)\|_2\le H\|x-y\|_2.$$

3. *A function h is* λ-strongly convex *if for all $x, y \in \mathcal{X}$*

$$h(y) \geq h(x) + \nabla h(x)^T (y - x) + \frac{\lambda}{2} \|y - x\|_2^2 .$$

We formally define interpolation problems:

Definition 2.3 (Interpolation Problem)**.** *Let $\mathcal{X}^\star \coloneqq \operatorname{argmin}_{x \in \mathcal{X}} f(x)$. Then problem (1) is an interpolation problem if there exists $x^\star \in \mathcal{X}^\star$ such that for P-almost all $s \in \Omega$, we have $0 \in \partial F(x^\star; s)$.*

Interpolation problems are common in modern machine learning, where models are overparameterized. One simple example is overparameterized linear regression: there exists a solution that minimizes each individual sample function. Classification problems with margin are another example.

Crucial to our results is the following quadratic growth assumption:

Definition 2.4. *We say that a function f satisfies the quadratic growth condition if for all $x \in \mathcal{X}$*

$$f(x) - \inf_{x' \in \mathcal{X}^\star} f(x') \geq \frac{\lambda}{2} \operatorname{dist}(x, \mathcal{X}^\star)^2 .$$

This assumption is natural with interpolation and holds for many important applications including noiseless linear regression [SV09, NWS14]. Past work ([VBS19, WS21]) uses this assumption with interpolation to get faster rates of convergence for non-private optimization.

Finally, the adaptivity of our algorithms will crucially depend on an innovation leveraging Lipchitzian extensions, defined as follows.

Definition 2.5 (Lipschitzian extension [HUL93])**.** *The Lipschitzian extension with Lipschitz constant L of a function f is defined as the infimal convolution*

$$f_L(x) \coloneqq \inf_{y \in \mathbb{R}^d} \{f(y) + L \|x - y\|_2\}. \tag{2}$$

The Lipschitzian extension (2) essentially transforms a general convex function into an L-Lipschitz convex function. We now present a few properties of the Lipschitzian extension that are relevant to our development.

Lemma 2.1. *Let $f : \mathcal{X} \to \mathbb{R}$ be convex. Then its Lipschitzian extension satisfies the following:*

1. *f_L is L-Lipschitz.*
2. *f_L is convex.*
3. *If f is L-Lipschitz, then $f_L(x) = f(x)$, for all x.*
4. *Let $y(x) = \operatorname{argmin}_{y \in \mathbb{R}^d} \{f(y) + L \|x - y\|_2\}$. If $y(x)$ is at a finite distance from x, we have*

$$\nabla f_L(x) = \begin{cases} \nabla f(x), & \text{if } \|\nabla f(x)\|_2 \leq L \\ L \frac{x - y(x)}{\|x - y(x)\|_2}, & \text{otherwise.} \end{cases}$$

We use the Lipschitzian extension as a substitute for gradient clipping to ensure differential privacy. Unlike gradient clipping, which may alter the geometry of a convex problem to a non-convex one, the Lipschitzian extension of a function remains convex and thus retains other nice properties that we leverage in our algorithms in Section 4.

3 Hardness of private interpolation

In non-private stochastic convex optimization, for smooth functions it is well known that interpolation problems enjoy the fast rate $O(1/n)$ [SST10] compared to the minimax-optimal $O(1/\sqrt{n})$ without interpolation [Duc18]. In this section, we show that such an improvement is not generally possible with privacy. The same lower bound of private non-interpolation problems, $d/n\varepsilon$, holds for interpolation problems.

To state our lower bounds, we present some notation that we will use throughout of the paper. We let $\mathfrak{S}$ denote the family of function F and dataset $\mathcal{S}$ pairs such that $F : \mathcal{X} \times \Omega \to \mathbb{R}$ is convex and H-smooth in its first argument, $|\mathcal{S}| = n$, and $f_{\mathcal{S}}(y) = \frac{1}{n}\sum_{s\in\mathcal{S}} F(y, s)$ is an interpolation problem (Definition 2.3). We define the constrained minimax risk to be

$$\mathfrak{M}(\mathcal{X}, \mathfrak{S}, \varepsilon, \delta) \coloneqq \inf_{M\in\mathcal{M}^{(\varepsilon,\delta)}} \sup_{(F,\mathcal{S}^n)\in\mathfrak{S}} \mathbb{E}[f_{\mathcal{S}^n}(M(\mathcal{S}^n))] - \inf_{x'\in\mathcal{X}} f_{\mathcal{S}^n}(x').$$

where $\mathcal{M}^{(\varepsilon,\delta)}$ be the collection of (ε, δ)-differentially private mechanisms from Ω^n to $\mathcal{X}$. We use $\mathcal{M}^{(\varepsilon,0)}$ to denote the collection of ε-DP mechanisms from Ω^n to $\mathcal{X}$. Here, the expectation is taken over the randomness of the mechanism, while the dataset $\mathcal{S}^n$ is fixed.

We have the following lower bound for private interpolation problems; the proof is deferred to Appendix B.1.

Theorem 1. *Suppose $\mathcal{X} \subset \mathbb{R}^d$ contains a d-dimensional ℓ_2 ball of diameter D. Then the following lower bound holds for $\delta = 0$*

$$\mathfrak{M}(\mathcal{X}, \mathfrak{S}, \varepsilon, 0) \geq \frac{HD^2 d}{96e^2 n\varepsilon}.$$

Moreover, if $0 < \delta < \varepsilon/6$ and $d = 1$, the following lower bound holds

$$\mathfrak{M}(\mathcal{X}, \mathfrak{S}, \varepsilon, \delta) \geq \frac{HD^2}{16(e+1)n\varepsilon}.$$

Recall the optimal rate for pure DP optimization problems without interpolation is $O(\frac{1}{\sqrt{n}} + \frac{d}{n\varepsilon})$. The first term is the non-private rate, as this is the rate one would get if $\varepsilon = 0$. The second term is the private rate, as this is the price algorithms have to pay for privacy. In modern machine learning, problems are often high dimensional, so we often think of the dimension d scaling with some function of the number of samples n. Thus, the private rate is often thought to dominate the non-private rate. For this reason, in this section, we focus on the private rate. The lower bounds of Theorem 1 show that it is not possible to improve the private rate for interpolation problems in general. Similarly, for approximate (ε, δ)-DP, the lower bound shows that improvements are not possible for $d = 1$. For completeness, as we alluded to earlier, we note that our results do not preclude the possibility of improving the non-private rate from $O(1/\sqrt{n})$ to $O(1/n)$. We leave this as an open problem of independent interest for future work.

Despite this pessimistic result, in the next section we show that substantial improvements are possible for private interpolation problems with additional growth conditions.

4 Faster rates for interpolation with growth

Having established our hardness result for general interpolation problems, in this section we show that when the functions satisfy additional growth conditions, we get (nearly) exponential improvements in the rates of convergence for private interpolation.

Our algorithms use recent localization techniques that yield optimal algorithms for DP-SCO [FKT20, ALD21] where the algorithm iteratively shrinks the diameter of the domain. However, to obtain faster rates for interpolation, we crucially build on the observation that the norm of the gradients is decreasing as we approach the optimal solution, since $\|\nabla F(x; s)\|_2 \leq H \|x - x^\star\|_2$. Hence, by carefully

localizing the domain and shrinking the Lipschitz constant accordingly, our algorithms improve the rates for interpolating datasets.

However, this technique alone yields an algorithm that may not be private for non-interpolation problems, violating that privacy must hold for all inputs: the reduction in the Lipschitz constant may not hold for non-interpolation problems, and thus, the amount of noise added may not be enough to ensure privacy. To solve this issue, we use the Lipschitzian extension (Definition 2.5) to transform our potentially non-Lipschitz sample functions into Lipschitz ones and guarantee privacy even for non-interpolation problems.

We begin in Section 4.1 by presenting our Lipschitzian extension based algorithm, which recovers the standard optimal rates for (non-interpolation) L-Lipschitz functions while still guaranteeing privacy when the function is not Lipschitz. Then in Section 4.2 we build on this algorithm to develop a localization-based algorithm that obtains faster rates for interpolation-with-growth problems. Finally, in Section 4.3 we present our final adaptive algorithm, which obtains fast rates for interpolation-with-growth problems while achieving optimal rates for non-interpolation growth problems.

4.1 Lipschitzian-extension based algorithms

Existing algorithms for DP-SCO with L-Lipschitz functions may not be private if the input function is not L-Lipschitz [BFGT20, FKT20, ALD21]. Given any DP-SCO algorithm $\mathbf{M}^L_{(\varepsilon,\delta)}$, which is private for L-Lipschitz functions, we present a framework that transforms $\mathbf{M}^L_{(\varepsilon,\delta)}$ to an algorithm which is (i) private for all functions, even ones which are not L-Lipschitz functions and (ii) has the same utility guarantees as $\mathbf{M}^L_{(\varepsilon,\delta)}$ for L-Lipschitz functions. In simpler terms, our algorithm essentially feeds $\mathbf{M}^L_{(\varepsilon,\delta)}$ the Lipschitzian-extension of the sample functions as inputs. Algorithm 1 describes our Lipschitzian-extension based framework.

Algorithm 1 Lipschitzian-Extension Algorithm

Require: Dataset $\mathcal{S} = (s_1, \dots, s_n) \in \mathbb{S}^n$;

1: Let $F_L(x; s_i)$ be the Lipschitzian extension of $F(x; s_i)$ for all i.

$$F_L(x; s_i) = \inf_y \{F(y; s_i) + L \|x - y\|_2\}$$

2: Run $\mathbf{M}^L_{(\varepsilon,\delta)}$ over the functions $F_L(\cdot; s_i)$.

3: Let x_{priv} denote the output of $\mathbf{M}^L_{(\varepsilon,\delta)}$.

4: **return** x_{priv}

For this paper we consider $\mathbf{M}^L_{(\varepsilon,\delta)}$ to be Algorithm 2 of [ALD21] (reproduced in Appendix A.2 as Algorithm 5). The following proposition summarizes our guarantees for Algorithm 1.

Proposition 1. *Let $\mathcal{L}_L$ denote the set of sample function-dataset pair (F, S) such that F is L-Lipschitz and let $\mathcal{F}$ denote the set of sample function-dataset pair $(F, \mathcal{S})$ such that $\mathbf{M}^L_{(\varepsilon,\delta)}$ is (ε, δ)-DP for any $(F, \mathcal{S}) \in \mathcal{L}_L \cap \mathcal{F}$. Then*

1. *For any $(F, \mathcal{S}) \in \mathcal{F}$, Algorithm 1 is (ε, δ)-DP.*
2. *For any $(F, \mathcal{S}) \in \mathcal{L}_L \cap \mathcal{F}$, Algorithm 1 achieves the same optimality guarantees as $\mathbf{M}^L_{(\varepsilon,\delta)}$.*

Proof For the first item, note that Lemma 2.1 implies that F_L is L-Lipschitz, i.e. $(F_L, \mathcal{S}) \in \mathcal{L}_L \cap \mathcal{F}$. Since $\mathbf{M}^L_{(\varepsilon,\delta)}$ is (ε, δ)-DP when applied over Lipschitz functions in $\mathcal{F}$, we have that Algorithm 1 is (ε, δ)-DP.

For the second item, Lemma 2.1 implies that $F_L = F$ when F is L-Lipschitz. Thus, in Algorithm 1, we apply $\mathbf{M}^L_{(\varepsilon,\delta)}$ over F itself. □

While clipped DP-SGD does ensure privacy for input functions which are not L-Lipschitz, our algorithm has some advantages over clipped DP-SGD: first, clipping does not result in optimal rates for

Algorithm 2 Domain and Lipschitz Localization algorithm

Require: Dataset $\mathcal{S} = (s_1, \ldots, s_n) \in \mathbb{S}^n$, Lipschitz constant L, domain $\mathcal{X}$, probability parameter β, initial point x_0

Set $L_1 = L$, $D_1 = \mathrm{Diam}(\mathcal{X})$ and $\mathcal{X}_1 = \mathcal{X}$

Partition the dataset into T partitions (denoted by $\{\mathcal{S}_k\}_{k=1}^T$) of size m each; $\mathcal{S}_k = (s_{(k-1)m+1}, \ldots, s_{km})$

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

  $x_i \leftarrow$ Run Algorithm 1 with dataset $\mathcal{S}_i$, constraint set $\mathcal{X}_i$, Lipschitz constant L_i, probability parameter β/T, privacy parameters (ε, δ), initial point x_{i-1},

  Shrink the diameter

$$D_{i+1} = 256 \left(\frac{L_i}{\lambda} \max \left\{ \frac{\sqrt{\log(T/\beta)} \log^{3/2} m}{\sqrt{m}} \ , \ \frac{\min(d, \sqrt{d \log(1/\delta)}) \log(T/\beta) \log m}{m\varepsilon} \right\} \right)$$

  Set $\mathcal{X}_{i+1} = \{x : \|x - x_i\|_2 \le D_{i+1}/2\}$

  Set $L_{i+1} = H D_{i+1}$

end for

return the final iterate x_T

pure DP, and second, clipped DP-SGD results in time complexity $O(n^{3/2})$. In contrast, our Lipschitzian extension approach is amenable to existing linear time algorithms [FKT20] allowing for almost linear time complexity algorithms for interpolation problems. Finally, while clipping the gradients and using the Lipschitzian extension both alter the effective function being optimized, only the Lipschitzian extension is able to preserve the convexity of said effective function (see item 2 in Lemma 2.1). We make a note about the computational efficiency of Algorithm 1. Recall that when the objective is in fact L-Lipschitz, computing gradients for the Lipschitzian extension (say in the context of a first-order method) is only as expensive as computing the gradients for the original function. In particular, one can first compute the gradient of the original function and use item 4 of Lemma 2.1; when the problem is L-Lipschitz, $\|\nabla f(x)\|_2$ is always less than or equal to L and thus the gradient of the Lipschitzian extension is just the gradient of the original function.

4.2 Faster non-adaptive algorithm

Building on the Lipschitzian-extension framework of the previous section, in this section, we present our epoch based algorithm, which obtains faster rates in the interpolation-with-growth regime. It uses Algorithm 1 with $\mathbf{M}^L_{(\varepsilon,\delta)}$ as Algorithm 5 (reproduced in Appendix A.2) as a subroutine in each epoch, to localize and shrink the domain as the iterates get closer to the true minimizer. Simultaneously, the algorithm also reduces the Lipschitz constant, as the interpolation assumption implies that the norm of the gradient decreases for iterates near the minimizer. The detailed algorithm is given in Algorithm 2 where D_i denotes the effective diameter and L_i denotes the effective Lipschitz constant in epoch i.

The following theorem provides our upper bounds for Algorithm 2, demonstrating near-exponential rates for interpolation problems; we present the proof in Appendix C.

Theorem 2. *Assume each sample function F is L-Lipschitz and H-smooth, and let the population function f satisfy quadratic growth (Definition 2.4). Let Problem (1) be an interpolation problem. Then Algorithm 2 is (ε, δ)-DP. For $\delta = 0$, $\beta = \frac{1}{n^\mu}$, $m = 256 \log^2 n \frac{H \log(1/\beta)}{\lambda} \max\left\{\frac{256H}{\lambda}, \frac{d}{\varepsilon\sqrt{\log n}}\right\}$, $T = n/m$ and any $\mu > 0$, Algorithm 2 returns x_T such that*

$$\mathbb{E}[f(x_T) - f(x^\star)] \le LD \left(\frac{1}{n^\mu} + \exp\left(-\widetilde{\Theta}\left(\frac{n\lambda^2}{H^2} \right) \right) + \exp\left(-\widetilde{\Theta}\left(\frac{\lambda n \varepsilon}{Hd} \right) \right) \right). \tag{3}$$

For $\delta > 0$, $\beta = \frac{1}{n^\mu}$, $m = 256 \log^2 n \frac{H \log(1/\beta)}{\lambda} \max\left\{\frac{256H}{\lambda}, \frac{\sqrt{d}\log(1/\delta)}{\varepsilon\sqrt{\log n}}\right\}$, $T = n/m$ and any $\mu > 0$,

Algorithm 2 returns x_T such that

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq LD\left(\frac{1}{n^\mu} + \exp\left(\widetilde{\Theta}\left(\frac{n\lambda^2}{H^2}\right)\right) + \exp\left(-\widetilde{\Theta}\left(\frac{\lambda n\varepsilon}{H\sqrt{d\log(1/\delta)}}\right)\right)\right). \tag{4}$$

The exponential rates in Theorem 2 show a significant improvement in the interpolation regime over the minimax-optimal $O((\frac{1}{\sqrt{n}} + \frac{d}{n\varepsilon})^2)$ without interpolation [FKT20, ALD21]. To get the linear convergence rates, we run roughly $n/\log n$ epochs with $\log n$ samples each. Thus, each call of the subroutine runs the algorithm on only logarithmic number of samples compared to the number of epochs. Intuitively, growth conditions improves the performance of the sub-algorithm, while growth and interpolation conditions reduce the search space. This in tandem leads to faster rates.

To better illustrate the improvement in rates compared to the non-private setting, the next corollary states the private sample complexity required to achieve error α in the interpolation regime.

Corollary 4.1. *Let the conditions of Theorem 2 hold. For $\delta = 0$, Algorithm 2 is ε-DP and requires*

$$n = \widetilde{O}\left(\frac{1}{\alpha^\rho} + \frac{d}{\rho\varepsilon}\log\left(\frac{1}{\alpha}\right)\right)$$

samples to ensure $\mathbb{E}[f(x_T) - f(x^\star)] \leq \alpha$ for any fixed $\rho > 0$, where $\widetilde{O}$ ignores only polyloglog factors in $1/\alpha$.
Moreover, for $\delta > 0$, Algorithm 2 is (ε, δ)-DP and requires

$$n = \widetilde{O}\left(\frac{1}{\alpha^\rho} + \frac{\sqrt{d\log(1/\delta)}}{\rho\varepsilon}\log\left(\frac{1}{\alpha}\right)\right)$$

samples to ensure $\mathbb{E}[f(x_T) - f(x^\star)] \leq \alpha$, for any fixed $\rho > 0$, where $\widetilde{O}$ ignores polyloglo factors in $1/\alpha$.

As the sample complexity of DP-SCO to achieve expected error α on general quadratic growth problems is [ALD21]

$$\Theta\left(\frac{1}{\alpha} + \frac{d}{\varepsilon\sqrt{\alpha}}\right),$$

Corollary 4.1 shows that we are able to improve the polynomial dependence on $1/\alpha$ in the sample complexity to (nearly) logarithmic for interpolation problems.

Remark 1. *In contrast to Corollary 4.1, we can tune the failure probability parameter β to get the sample complexity $\frac{d}{\varepsilon}\log^2\left(\frac{1}{\alpha}\right)$. Even though this sample complexity does not have the polynomial factor, it may be worse than $\frac{1}{\alpha^\rho} + \frac{d}{\varepsilon}\log\left(\frac{1}{\alpha}\right)$, because generally the dimension term is the dominant one.*

We end this section by considering growth conditions that are weaker than quadratic growth.

Remark 2. *(interpolation with κ-growth) We can extend our algorithms to work for the weaker κ-growth condition [ALD21], i.e., $f(x) - f(x^\star) \geq \frac{\lambda}{\kappa}\|x - x^\star\|_2^\kappa$. We present the full details of these algorithms in Appendix C.1 (see Algorithm 6). In this setting, we obtain excess loss*

$$O\left(\left(\frac{1}{\sqrt{n}} + \frac{d}{n\varepsilon}\right)^{\frac{\kappa}{\kappa-2}}\right),$$

for interpolation problems, improving over the minimax-optimal loss for non-interpolation problems which is

$$O\left(\left(\frac{1}{\sqrt{n}} + \frac{d}{n\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right).$$

As an example, when $\kappa = 3$, this corresponds to an improvement from roughly $(d/n\varepsilon)^{3/2}$ to $(d/n\varepsilon)^3$. Like our previous results, we are again able to show similar improvements for (ε, δ)-DP with better dependence on the dimension. Finally, we note that we have not provided lower bounds for the interpolation-with-κ-growth setting for $\kappa > 2$. We leave this question as a direction for future research.

Algorithm 3 Algorithm that adapts to interpolation

Require: Dataset $\mathcal{S} = (s_1, \ldots, s_n) \in \mathbb{S}^n$, Lipschitz constant L, domain $\mathcal{X}$, probability parameter β, initial point x_0

1: Partition the dataset into 2 partitions $S_1 = (s_1, \ldots, s_{n/2})$ and $S_2 = (s_{(n/2)+1}, \ldots, s_n)$

2: $x_1 \leftarrow$ Run Algorithm 1 with dataset S_1, constraint set $\mathcal{X}_i$, Lipschitz constant L_i, probability parameter $\beta/2$, privacy parameters (ε, δ), initial point x_{i-1},

3: Shrink the diameter

$$D_{\text{int}} = \frac{128L}{\lambda} \cdot \left(\frac{\sqrt{\log(2/\beta)} \log^{3/2} n}{\sqrt{n}} + \frac{\min\{d, \sqrt{d \log(1/\delta)}\} \log(2/\beta) \log n}{n\varepsilon} \right)$$

4: $\mathcal{X}_{\text{int}} = \{x : \|x - x_1\|_2 \leq D_{\text{int}}/2\}$

5: $x_{\text{adapt}} \leftarrow$ Run Algorithm 2 with dataset S_2, diameter D_{int}, Lipschitz constant L, domain $\mathcal{X}_{\text{int}}$, smoothness parameter H, tail probability parameter $\beta/2$, growth parameter λ, initial point x_1

6: **return** the final iterate x_{adapt}.

4.3 Adaptive algorithm

Though Algorithm 2 is private and enjoys faster rates of convergence in the interpolation regime, it is not necessarily adaptive to interpolation, i.e. it may perform poorly given a non-interpolation problem. In fact, since the shrinkage of the diameter and Lipschitz constants at each iteration hinges squarely on the interpolation assumption, the new domain may not include the optimizing set $\mathcal{X}^\star$ in the non-interpolation setting, so our algorithm may not even converge. Since in general we do not know a priori whether a dataset is interpolating, it is important to have an algorithm which adapts to interpolation.

To that end, we present an adaptive algorithm that achieves faster rates for interpolation-with-growth problems while simultaneously obtaining the standard optimal rates for general growth problems. The algorithm consists of two steps. In the first step, our algorithm privately minimizes the objective without assuming it is an interpolation problem. Next, we run our non-adaptive interpolation algorithm of Section 4.2 over the localized domain returned by the first step. If our problem was an interpolating one, the second step recovers the faster rates in Section 4.2. If our problem was not an interpolating one, the first localization step ensures that we at least recover the non-interpolating convergence rate. We stress that the privacy of Algorithm 3 requires that the call to Algorithm 2 remains private even if the problem is non-interpolating. This is ensured by using our Lipschitzian extension based algorithm with $\mathbf{M}^L_{(\varepsilon,\delta)}$ as Algorithm 5. The Lipschitzian extension allows us to continue preserving privacy. We present the full details of this algorithm in Algorithm 3.

The following theorem (Theorem 3) states the convergence guarantees of our adaptive algorithm (Algorithm 3) in both the interpolation and non-interpolation regimes for the pure DP setting. The results for approximate DP are similar and can be obtained by replacing d with $\sqrt{d \log(1/\delta)}$; we give the full details in Appendix C.

Theorem 3. *Let each sample function F be L-Lipschitz and H-smooth, and let the population function f satisfy quadratic growth (Definition 2.4) with coefficient λ. Let x_{adapt} be the output of Algorithm 3. Then*

1. *Algorithm 3 is ε-DP.*
2. *Without any additional interpolation assumption, x_{adapt} satisfies*

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq LD \cdot \widetilde{O}\left(\frac{1}{\sqrt{n}} + \frac{d}{n\varepsilon} \right)^2.$$

3. *Let problem (1) be an interpolation problem. Thenx_{adapt} satisfies*

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq LD \left(\frac{1}{n^\mu} + \exp\left(-\widetilde{\Theta}\left(\frac{n\lambda^2}{H^2} \right) \right) + \exp\left(-\widetilde{\Theta}\left(\frac{\lambda n \varepsilon}{Hd} \right) \right) \right).$$

Proof The privacy of Algorithm 3 follows from the privacy of Algorithms 1 and 2 and post-processing.

To prove the convergence guarantees, we first need to show that the optimal set $\mathcal{X}^\star$ is in the shrinked domain $\mathcal{X}_{\text{int}}$. Using the high probability guarantees of Algorithm 1, we know that with probability $1-\beta/2$, we have

$$f(x_1) - f(x^\star) \le \frac{2^{12}L}{\lambda} \cdot \left(\frac{\sqrt{\log(2/\beta)}\log^{3/2} n}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}\log(2/\beta)\log n}{n\varepsilon} \right).$$

Using the quadratic growth condition, we immediately have $\|x^\star - x_1\|_2 \le D_{\text{int}}/2$ and hence $\mathcal{X}^\star \subset \mathcal{X}_{\text{int}}$.

Using smoothness, we have that for any $x \in \mathcal{X}_{\text{int}}$,

$$f(x) - f(x^\star) \le \frac{HD_{\text{int}}^2}{2}.$$

Since Algorithm 2 always outputs a point in its input domain (in this case $\mathcal{X}_{\text{int}}$), even in the non-interpolation setting that

$$\mathbb{E}[f(x_T) - f(x^\star)] \le LD \cdot \widetilde{O}\left(\frac{1}{\sqrt{n}} + \frac{d}{n\varepsilon} \right)^2.$$

In the interpolation setting, the guarantees of Algorithm 2 hold and result is immediate. □

5 Optimality and Superefficiency

We conclude this paper by providing a lower bound and a super-efficiency result that demonstrate the tightness of our upper bounds. Recall that our upper bound from Section 4 is roughly (up to constants)

$$\frac{1}{n^c} + \exp\left(-\widetilde{\Theta}\left(\frac{n\varepsilon}{d} \right) \right), \tag{5}$$

for any arbitrarily large c. We begin with an exponential lower bound showing that the second term in (5) is tight. We then prove a superefficiency result that demonstrates that any private algorithm which avoids the first term in (5) cannot be adaptive to interpolation, that is, it can not achieve the minimax optimal rate for the family of non-interpolation problems.

Theorem 4 below presents our exponential lower bounds for private interpolation problems with growth. We use the notation and proof structure of Theorem 1. We let $\mathfrak{S}^\lambda \subset \mathfrak{S}$ be the subcollection of function, data set pairs which also have functions $f_{\mathcal{S}^n}$ that have λ-quadratic growth (Definition 2.4). The proof of Theorem 4 is found in Appendix D.1.

Theorem 4. *Let $\mathcal{X} \subset \mathbb{R}^d$ contain a d-dimensional ℓ_2-ball of diameter D. Then*

$$\mathfrak{M}(\mathcal{X}, \mathfrak{S}^\lambda, \varepsilon, 0) \ge \frac{\lambda D^2}{96} \exp\left(-\frac{2\lambda n\varepsilon}{Hd} \right).$$

This lower bound addresses the second term of (5); we now turn to our superefficiency results to lower bound the first term of (5). We start with defining some notation and making some simplifying assumptions. For a fixed function $F : \mathcal{X}, \Omega \to \mathbb{R}$ which is convex, H-smooth with respect to the first argument, let $\mathfrak{S}_\lambda^L(F)$ be the set of datasets $\mathcal{S}$ of n data points sampled from Ω such that $f_{\mathcal{S}}(x) := \frac{1}{n}\sum_{s\in\mathcal{S}^n} F(x,s)$ is L-Lipschitz and have λ-strongly convex objectives. For simplicity, we will assume that 1. $\inf_{x\in\mathcal{X}} F(x;s) = 0$ for all $s \in \Omega$, 2. $\mathcal{X} = [-D, D] \subset \mathbb{R}$, and 3. the codomain of F is $\mathbb{R}_+$. With this setup, we present the formal statement of our result; the proof of Theorem 5 is found in Appendix D.2.

Theorem 5. *Suppose we have some $\mathcal{S} \in \mathfrak{S}_\lambda^L(F)$ with $L = 2HD$ such that $(F, \mathcal{S})$ satisfy Definition 2.3. Suppose there is an ε-DP estimator M such that*

$$\mathbb{E}[f_{\mathcal{S}}(M(\mathcal{S}))] - \inf_{x\in\mathcal{X}} f_{\mathcal{S}}(x) \le cD^2 e^{-\Theta((n\varepsilon)^t)}$$

for some $t > 0$ and absolute constant c. Then, for sufficiently large n, there exists another dataset $\mathcal{S}' \in \mathfrak{S}_\lambda^L(F)$, where $(F, \mathcal{S}')$ may ***not*** *satisfy Definition 2.3, such that*

$$\mathbb{E}[f_{\mathcal{S}'}(M(\mathcal{S}'))] - \inf_{x \in \mathcal{X}} f_{\mathcal{S}'}(x) = \Omega\left(\frac{D^2}{(n\varepsilon)^{2(1-t)}}\right)$$

To better contextualize this result, suppose there exists an algorithm which atttains a $\exp(-\widetilde{\Theta}(n\varepsilon/d))$ convergence rate on interpolation problems; i.e., the algorithm is able to avoid the $1/n^c$ term in (5). Then Theorem 5 states that there exists some strongly convex, non-interpolation problem on which the aforementioned algorithm will optimize very poorly; in particular, the algorithm will only be able to return a solution that attains, on average, constant error on this "hard" problem. More generally, recall that in the non-interpolation quadratic growth setting, the optimal error rate is on the order of $1/(n\varepsilon)^2$ [ALD21]. Theorem 5 shows that attaining better-than-polynomial error complexity on quadratic growth interpolation problems implies that the algorithm cannot be minimax optimal in the non-interpolation quadratic growth setting. Thus, the rates our adaptive algorithms attain are the best we can hope for if we want an algorithm to perform well on both interpolation and non-interpolation quadratic growth problems.

References

[ACCD20] Hilal Asi, Karan Chadha, Gary Cheng, and John C. Duchi. Minibatch stochastic approximate proximal point methods. In *Advances in Neural Information Processing Systems 33*, 2020.

[ACG+16] Martin Abadi, Andy Chu, Ian Goodfellow, Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *23rd ACM Conference on Computer and Communications Security (ACM CCS)*, pages 308–318, 2016.

[AD19] Hilal Asi and John C. Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

[AD20] Hilal Asi and John Duchi. Near instance-optimality in differential privacy. *arXiv:2005.10630 [cs.CR]*, 2020.

[ADF+21] Hilal Asi, John Duchi, Alireza Fallah, Omid Javidbakht, and Kunal Talwar. Private adaptive gradient methods for convex optimization. In *Proceedings of the 38th International Conference on Machine Learning*, pages 383–392, 2021.

[AFKT21] Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in ℓ_1 geometry. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[ALD21] Hilal Asi, Daniel Levy, and John C. Duchi. Adapting to function difficulty and growth conditions in private optimization. In *Advances in Neural Information Processing Systems 34*, 2021.

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

[BFTT19] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems 32*, 2019.

[BGN21] Raef Bassily, Cristóbal Guzmán, and Anupama Nandi. Non-euclidean differentially private stochastic convex optimization. In *Proceedings of the Thirty Fourth Annual Conference on Computational Learning Theory*, pages 474–499, 2021.

[BHM18] Mikhail Belkin, Daniel Hsu, and Partha Mitra. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems 31*, pages 2300–2311. Curran Associates, Inc., 2018.

[BRT19] Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. Does data interpolation contradict statistical optimality? In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619, 2019.

[BST14] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *55th Annual Symposium on Foundations of Computer Science*, pages 464–473, 2014.

[CCD22] Karan Chadha, Gary Cheng, and John Duchi. Accelerated, optimal and parallel: Some results on model-based stochastic optimization. In *International Conference on Machine Learning*, pages 2811–2827. PMLR, 2022.

[CMS11] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

[CSSS11] Andrew Cotter, Ohad Shamir, Nati Srebro, and Karthik Sridharan. Better mini-batch algorithms via accelerated gradient methods. In *Advances in Neural Information Processing Systems 24*, 2011.

[DJW13] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *54th Annual Symposium on Foundations of Computer Science*, pages 429–438, 2013.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3 & 4):211–407, 2014.

[Duc18] John C. Duchi. Introductory lectures on stochastic convex optimization. In *The Mathematics of Data*, IAS/Park City Mathematics Series. American Mathematical Society, 2018.

[FKT20] Vitaly Feldman, Tomer Koren, and Kunal 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.

[HUL93] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I & II*. Springer, New York, 1993.

[LB20] Chaoyue Liu and Mikhail Belkin. Accelerating sgd with momentum for over-parameterized learning. In *International Conference on Learning Representations*, 2020.

[MBB18] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[NWS14] Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. In *Advances in Neural Information Processing Systems 27*, pages 1017–1025, 2014.

[SSSSS09] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *Proceedings of the Twenty Second Annual Conference on Computational Learning Theory*, 2009.

[SST10] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *nips2010*, pages 2199–2207, 2010.

[ST13] Adam Smith and Abhradeep 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.

[SV09] T. Strohmer and Roman Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

[VBS19] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[WS21] Blake E. Woodworth and Nathan Srebro. An even more optimal stochastic optimization algorithm: Minibatching and interpolation learning. In *Advances in Neural Information Processing Systems 34*, 2021.

[WXDX20] Di Wang, Hanshen Xiao, Srini Devadas, and Jinhui Xu. Private stochastion differentially private stochastic convex optimization with heavy-tailed data. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

A Results from previous work

A.1 Proof of Lemma 2.1

1. Follows from Proposition IV.3.1.4 of [HUL93].
2. Follows from Proposition IV.3.1.4 of [HUL93].
3. Follows since for L-lipschitz functions $0 \in \nabla f(x) + L\mathbb{B}_2$.
4. Follows from Section VI.4.5 of [HUL93].

A.2 Algorithms from [ALD21]

A.3 Theoretical results from [ALD21]

We first reproduce the high probability guarantees of Algorithm 4 as proved in [ALD21].

Proposition 2. *Let $\beta \le 1/(n+d)$, $D_2(\mathcal{X}) \le D$ 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 4 is ε-DP and has with probability $1-\beta$

$$f(x) - f(x^\star) \le 128LD \cdot \left(\frac{\sqrt{\log(1/\beta)}\log^{3/2} n}{\sqrt{n}} + \frac{d\log(1/\beta)\log n}{n\varepsilon}\right).$$

Proposition 3. *Let $\beta \le 1/(n+d)$, $D_2(\mathcal{X}) \le D$ and $F(x;s)$ be convex, L-Lipschitz for all $s \in \mathbb{S}$. Setting*

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

Algorithm 4 Localization based Algorithm

Require: Dataset $D = (s_1, \ldots, s_n) \in \mathbb{S}^n$, constraint set $\mathcal{X}$, step size η, initial point x_0, Lipschitz (clipping) constant L, 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 $\mathcal{X}_i = \{x \in \mathcal{X} : \|x - x_{i-1}\|_2 \leq 2L\eta_i n_0\}$:

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

  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 $\sigma_i = 4L\eta_i\sqrt{\log(1/\delta)}/\varepsilon$
  end if
  Set $x_i = \hat{x}_i + \zeta_i$
end for
return the final iterate x_k

Algorithm 5 Epoch-based algorithms for κ-growth

Require: Dataset $\mathcal{S} = (s_1, \ldots, s_n) \in \mathbb{S}^n$, constraint set $\mathcal{X}$, Lipschitz (clipping) constant L, initial point x_0, number of iterations T, probability parameter β, privacy parameters (ε, δ);

Set $n_0 = n/T$ and $D_0 = \mathrm{diam}(\mathcal{X})$
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**
  Set

$$\eta_0 = \frac{D_0}{2L} \min\left\{\frac{1}{\sqrt{n_0 \log(n_0)\log(1/\beta)}}, \frac{\varepsilon}{\sqrt{d\log(1/\delta)}\log(1/\beta)}\right)$$

end if
for $i = 0$ to $T - 1$ **do**
  Let $\mathcal{S}_i = (s_{1+(i-1)n_0}, \ldots, s_{in_0})$
  Set $D_i = 2^{-i}D_0$ and $\eta_i = 2^{-i}\eta_0$
  Set $\mathcal{X}_i = \{x \in \mathcal{X} : \|x - x_i\|_2 \leq D_i\}$
  Run Algorithm 4 on dataset $\mathcal{S}_i$ with starting point x_i, Lipschitz (clipping) constant L, 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
end for
return x_T

then for $\delta > 0$, Algorithm 4 is (ε, δ)-DP and has with probability $1 - \beta$

$$f(x) - f(x^\star) \leq 128LD \cdot \left(\frac{\sqrt{\log(1/\beta)} \log^{3/2} n}{\sqrt{n}} + \frac{\sqrt{d \log(1/\delta)} \log(1/\beta) \log n}{n\varepsilon} \right).$$

Now, we reproduce the high probability convergence guarantees of Algorithm 5.

Theorem 6. *Let $\beta \leq 1/(n+d)$, $D_2(\mathcal{X}) \leq D$ and $F(x; s)$ be convex, L-Lipschitz for all $s \in \Omega$. Assume that f has κ-growth (Assumption 2.4) with $\kappa \geq \underline{\kappa} > 1$. Setting $T = \left\lceil \frac{2 \log n}{\underline{\kappa} - 1} \right\rceil$, Algorithm 5 is ε-DP and has with probability $1 - \beta$*

$$f(x_T) - \min_{x \in \mathcal{X}} f(x) \leq \frac{4032}{\lambda^{\frac{1}{\kappa - 1}}} \cdot \left(\frac{L\sqrt{\log(1/\beta)} \log^{3/2} n}{\sqrt{n}} + \frac{Ld \log(1/\beta) \log n}{n\varepsilon(\underline{\kappa} - 1)} \right)^{\frac{\kappa}{\kappa - 1}}.$$

Theorem 7. *Let $\beta \leq 1/(n+d)$, $D_2(\mathcal{X}) \leq D$ and $F(x; s)$ be convex, L-Lipschitz for all $s \in \Omega$. Assume that f has κ-growth (Assumption 2.4) with $\kappa \geq \underline{\kappa} > 1$. Setting $T = \left\lceil \frac{2 \log n}{\underline{\kappa} - 1} \right\rceil$ and $\delta > 0$, Algorithm 5 is (ε, δ)-DP and has with probability $1 - \beta$*

$$f(x_T) - \min_{x \in \mathcal{X}} f(x) \leq \frac{4032}{\lambda^{\frac{1}{\kappa - 1}}} \cdot \left(\frac{L\sqrt{\log(1/\beta)} \log^{3/2} n}{\sqrt{n}} + \frac{L\sqrt{d \log(1/\delta)} \log(1/\beta) \log n}{n\varepsilon(\underline{\kappa} - 1)} \right)^{\frac{\kappa}{\kappa - 1}}.$$

B Proofs from Section 3

B.1 Proof of Theorem 1

Consider the sample risk function

$$F(x; s) \coloneqq \frac{H}{2} \|x - s\|_2^2 \cdot \mathbf{1}\{s \neq 0\}.$$

We define the datasets $\mathcal{S}_v^n \coloneqq \{0\}^{n-k} \cup \{v\}^k$. We define the corresponding population risk to be $f_v(x) \coloneqq \frac{1}{n} \sum_{s \in \mathcal{S}_v} F(x; s) = \frac{kH}{2n} \|x - v\|_2^2$. We select $\mathcal{V}$ to be a γ-packing (with respect to the ℓ_2 norm) of diameter D ball contained in $\mathcal{X}$. Define the separation between $v, v' \in \mathcal{V}$ with respect to the loss f_v and $f_{v'}$ by

$$d_{\text{opt}}(f_v, f_{v'}) \coloneqq \inf_{x \in \mathcal{X}} \frac{f_v(x)}{2} + \frac{f_{v'}(x)}{2} \geq c \coloneqq \frac{kH}{8n} \gamma^2.$$

For the sake of contradiction, suppose that $\mathbb{E}[f_v(M(\mathcal{S}_v))] \leq \tau$ for $\tau < \frac{kH\gamma^2}{8n(1 + e^{k\varepsilon} 2^d \gamma^d / D^d)}$ for all $v \in \mathcal{X}$. Then by Markov's inequality, $\mathbb{P}(f_v(M(\mathcal{S}_v)) > c) \leq \frac{\tau}{c}$ and $\mathbb{P}(f_v(M(\mathcal{S}_v)) \leq c) \geq 1 - \frac{\tau}{c}$ for all v, and so

$$\begin{aligned}
\frac{\tau}{c} &\overset{(i)}{\geq} \mathbb{P}(f_v(M(\mathcal{S}_v)) > c) \\
&\overset{(ii)}{\geq} \mathbb{P}(\cup_{v' \in \mathcal{V} \setminus \{v\}} f_{v'}(M(\mathcal{S}_v)) \leq c) \\
&\overset{(iii)}{\geq} e^{-k\varepsilon} \sum_{v' \in \mathcal{V} \setminus \{v\}} \mathbb{P}(f_{v'}(M(\mathcal{S}_{v'})) \leq c) \\
&\overset{(i)}{\geq} e^{-k\varepsilon} (|\mathcal{V}| - 1) \left(1 - \frac{\tau}{c}\right),
\end{aligned}$$

where inequality (ii) follows from the definition of the separation, and (iii) follows from privacy and the disjoint nature of the events in the union. Rearranging, we get that

$$\tau \geq \frac{kH\gamma^2}{8n(1 + e^{k\varepsilon}(|\mathcal{V}| - 1)^{-1})},$$

which is a contradiction. By standard packing inequalities [Wai19], we know that $|\mathcal{V}| \geq (D/2\gamma)^d$. Setting $k = d/\varepsilon$ and $\gamma = D/2e$ and using the fact that $x/(x-1)$ is decreasing in x gives

$$\tau \geq \frac{dHD^2}{32n\varepsilon e^2(1+e^d(e^d-1)^{-1})} \geq \frac{HD^2 d}{96e^2 n\varepsilon}.$$

We now prove the (ε, δ)-DP lower bound. Consider the following sample risk function

$$F(x;s) \coloneqq \frac{H}{2}(x-s)^2 \mathbf{1}\{s \neq 0\}.$$

We define the datasets $\mathcal{S}_v^n \coloneqq \{0\}^{n-k} \cup \{v\}^k$ inducing the corresponding population risk $f_v(x) \coloneqq \frac{1}{n}\sum_{s\in\mathcal{S}_v} F(x;s) = \frac{kH}{2n}(x-v)^2$. We select two points v, v' contained within the diameter D ball contained in $\mathcal{X}$ such that $|v - v'| = D$. Define the separation between $v, v' \in \mathcal{V}$ with respect to the loss f_v and $f_{v'}$ as

$$d_{\text{opt}}(f_v, f_{v'}) \coloneqq \inf_{x\in\mathcal{X}} \frac{f_v(x)}{2} + \frac{f_{v'}(x)}{2} \geq c \coloneqq \frac{kH}{8n}D^2$$

For the sake of contradiction, suppose that $\mathbb{E}[f_v(M(\mathcal{S}_v))] \leq \tau$ for $\tau < \frac{kHD^2}{8n}\left(\frac{e^{-k\varepsilon} - ke^{-\varepsilon}\delta}{1+e^{-k\varepsilon}}\right)$ for all $v \in \mathcal{X}$. Then by Markov's inequality, $\mathbb{P}(f_v(M(\mathcal{S}_v)) > c) \leq \frac{\tau}{c}$ and $\mathbb{P}(f_v(M(\mathcal{S}_v)) \leq c) \geq 1 - \frac{\tau}{c}$ for all v, and so

$$\begin{aligned}
\frac{\tau}{c} &\overset{(i)}{\geq} \mathbb{P}(f_v(M(\mathcal{S}_v)) > c) \\
&\overset{(ii)}{\geq} \mathbb{P}(f_{v'}(M(\mathcal{S}_v)) \leq c) \\
&\overset{(iii)}{\geq} e^{-k\varepsilon}\mathbb{P}(f_{v'}(M(\mathcal{S}_{v'})) \leq c) - ke^{-\varepsilon}\delta \\
&\overset{(i)}{\geq} e^{-k\varepsilon}\left(1 - \frac{\tau}{c}\right) - ke^{-\varepsilon}\delta,
\end{aligned}$$

where inequality (ii) follows from the definition of the separation, and (iii) follows from group privacy of (ε, δ)-privacy [DR14]. Rearranging, we get that

$$\tau \geq \frac{kHD^2}{8n}\left(\frac{e^{-k\varepsilon} - ke^{-\varepsilon}\delta}{1+e^{-k\varepsilon}}\right),$$

which is a contradiction. Setting $k = 1/\varepsilon$ and using the fact $\delta \leq \varepsilon e^{\varepsilon-1}/2$ gives the first result.

C Proofs from Section 4

We first prove a lemma that each time we shrink the domain size, the set of interpolating solutions still lies in the new domain with high probability, and the new Lipschitz constant we define is a valid Lipschitz constant for the loss defined on the new domain. We prove it in generality for κ-growth.

Lemma C.1. *Let $\mathcal{X}^\star$ denote the set of interpolating solutions of problem (1). Then $\mathcal{X}^\star \subset \mathcal{X}_i$ for all $i \in [T]$ with probability $1-\beta$, and $\|\nabla F(y;s)\|_2 \leq L_i$ for all $y \in \mathcal{X}_i$.*

Proof We prove this lemma for the case when $\delta = 0$, the case when $\delta > 0$ follows similarly. For epoch i, using Theorem 2 of [ALD21], we have with probability $1 - \beta/T$,

$$f(\hat{x}_i) - f(x^\star) \leq \frac{C_\kappa}{\lambda^{\frac{1}{\kappa-1}}} \max\left\{\frac{L_i\sqrt{\log(T/\beta)}\log^{3/2} m}{\sqrt{m}}, \frac{L_i d \log(T/\beta)\log m}{m\varepsilon}\right\}^{\frac{\kappa}{\kappa-1}}$$

Using the growth condition on $f(\cdot)$, we have

$$\|\hat{x}_i - x^\star\|_2 \le \sqrt[\kappa]{\frac{\kappa(f(\hat{x}_i) - f(x^\star))}{\lambda}} \le (C_\kappa \kappa)^{1/\kappa} \max\left\{ \frac{L_i \sqrt{\log(T/\beta)} \log^{3/2} m}{\lambda \sqrt{m}}, \frac{L_i d \log(T/\beta) \log m}{\lambda m \varepsilon} \right\}^{\frac{1}{\kappa - 1}},$$

Using $c_\kappa = 2(C_\kappa \kappa)^{1/\kappa}$, we get $\|\hat{x}_i - x^\star\|_2 \le D_{i+1}/2$ with probability $1 - \beta/T$. Thus, for each epoch i, with probability $1 - \beta/T$, each point in the set $\mathcal{X}^\star$ of optimizers lies in the domain $\mathcal{X}_i$. Using a union bound on all epochs, we have $\mathcal{X}^\star \subset \mathcal{X}_i$ for all $i \in [T]$ with probability $1 - \beta$.

We now prove the second part of the lemma. Using the smoothness of $F(\cdot; s)$ and that $\nabla F(x^\star; s) = 0$ for all $x^\star \in \mathcal{X}^\star$, we have

$$\|\nabla F(y; s)\|_2 = \|\nabla F(y; s) - \nabla F(x^\star; s)\|_2 \le H \|y - x^\star\|_2 \le H \left(\|y - \hat{x}_i\|_2 + \|\hat{x}^\star - \hat{x}_i\|_2 \right) \le H D_i = L_i$$

as desired. □

We now restate and prove the convergence rate of Algorithm 2

Theorem 2. *Assume each sample function F is L-Lipschitz and H-smooth, and let the population function f satisfy quadratic growth (Definition 2.4). Let Problem (1) be an interpolation problem. Then Algorithm 2 is (ε, δ)-DP. For $\delta = 0$, $\beta = \frac{1}{n^\mu}$, $m = 256 \log^2 n \frac{H \log(1/\beta)}{\lambda} \max\left\{ \frac{256H}{\lambda}, \frac{d}{\varepsilon \sqrt{\log n}} \right\}$, $T = n/m$ and any $\mu > 0$, Algorithm 2 returns x_T such that*

$$\mathbb{E}[f(x_T) - f(x^\star)] \le LD \left(\frac{1}{n^\mu} + \exp\left(-\widetilde{\Theta}\left(\frac{n\lambda^2}{H^2} \right) \right) + \exp\left(-\widetilde{\Theta}\left(\frac{\lambda n \varepsilon}{Hd} \right) \right) \right). \tag{3}$$

For $\delta > 0$, $\beta = \frac{1}{n^\mu}$, $m = 256 \log^2 n \frac{H \log(1/\beta)}{\lambda} \max\left\{ \frac{256H}{\lambda}, \frac{\sqrt{d} \log(1/\delta)}{\varepsilon \sqrt{\log n}} \right\}$, $T = n/m$ and any $\mu > 0$, Algorithm 2 returns x_T such that

$$\mathbb{E}[f(x_T) - f(x^\star)] \le LD \left(\frac{1}{n^\mu} + \exp\left(-\widetilde{\Theta}\left(\frac{n\lambda^2}{H^2} \right) \right) + \exp\left(-\widetilde{\Theta}\left(\frac{\lambda n \varepsilon}{H \sqrt{d \log(1/\delta)}} \right) \right) \right). \tag{4}$$

Proof First we prove the privacy guarantee of the algorithm. Each sample impacts only one of the iterates $\hat{x}_i$, thus Algorithm 2 satisfies the same privacy guarantee as Algorithm 5 by postprocessing.

We divide the utility proof into 2 main parts; first is to check the validity of the assumptions while applying Algorithm 5 and second is using its high probability convergence guarantees to get the final rates. To check this, we ensure that the optimum set lies in the new domain $\mathcal{X}_i$ at step i and that the Lipschitz constant L_i defined with respect to the domain is a valid lipschitz constant. This follows from Lemma C.1.

Next, we use the high probability convergence guarantees of the subalgorithm Algorithm 5 to get convergence rates for Algorithm 2. We prove it for the case when $\delta = 0$, the case when $\delta > 0$ is similar. We know that

$$\begin{aligned} L_i &= H D_i \\ &= c_2 \frac{H L_{i-1}}{\lambda} \max\left\{ \frac{\sqrt{\log(T/\beta)} \log^{3/2} m}{\sqrt{m}}, \frac{d \log(T/\beta) \log m}{m \varepsilon} \right\}. \end{aligned}$$

Thus we have

$$L_T = \left(c_2 \frac{H}{\lambda} \max\left\{ \frac{\sqrt{\log(T/\beta)} \log^{3/2} m}{\sqrt{m}}, \frac{d \log(T/\beta) \log m}{m \varepsilon} \right\} \right)^{T-1} L_1.$$

Using Theorem 2 of [ALD21] on the last epoch, we have with probability $1-\beta$ that

$$\begin{aligned} f(\hat{x}_T) - f(x^\star) &\le C_2 \frac{L_T^2}{\lambda} \max\left\{ \frac{\log(T/\beta)\log^{3/2} m}{m}, \frac{d^2 \log^2(T/\beta)\log m}{m^2\varepsilon^2} \right\} \\ &= \left(c_2^2 \frac{H^2}{\lambda^2} \max\left\{ \frac{\log(T/\beta)\log^{3/2} m}{m}, \frac{d^2 \log^2(T/\beta)\log m}{m^2\varepsilon^2} \right\} \right)^T \frac{C_2 L_1^2 \lambda}{H^2 c_2^2} \\ &= \left(c_2^2 \frac{H^2}{\lambda^2} \max\left\{ \frac{\log(T/\beta)\log^{3/2} m}{m}, \frac{d^2 \log^2(T/\beta)\log m}{m^2\varepsilon^2} \right\} \right)^T \frac{L_1^2 \lambda}{8H^2}. \end{aligned}$$

Let $m = k\log^2 n$ and $T = n/m$ for some k such that

$$\left(c_2^2 \frac{H^2}{\lambda^2} \max\left\{ \frac{\log(n/(\beta k \log^2 n))\log^{3/2}(k\log^2 n)}{k\log^2 n}, \frac{d^2 \log^2(n/(\beta k\log^2 n))\log(k\log^2 n)}{(k\log^2 n)^2\varepsilon^2} \right\} \right) \le \frac{1}{e}.$$

This holds for example for

$$k = 256 \frac{H\log(1/\beta)}{\lambda} \max\left\{ \frac{256H}{\lambda}, \frac{d}{\varepsilon\sqrt{\log n}} \right\},$$

for sufficiently large n. Using these values of m and T, we have

$$f(\hat{x}_T) - f(x^\star) \le \frac{C_2 L^2 \lambda}{H^2 c_2^2} \exp\left(-\frac{n}{k\log^2 n} \right) = \frac{L_1^2 \lambda}{8H^2} \exp\left(-\frac{n}{k\log^2 n} \right). \tag{6}$$

To get the convergence results in expectation, let A denote the "bad" event with tail probability β, where $f(\hat{x}_T) - f(x^\star) > \frac{L_1^2\lambda}{8H^2}\exp\left(-\frac{n}{k\log^2 n}\right)$. Now,

$$\begin{aligned} \mathbb{E}[f(\hat{x}_T) - f(x^\star)] &\le \beta \frac{HD^2}{2} + (1-\beta)\mathbb{E}[f(\hat{x}_T) - f(x^\star) \mid A^c] \\ &\le \beta \frac{HD^2}{2} + \mathbb{E}[f(\hat{x}_T) - f(x^\star) \mid A^c] \end{aligned}$$

Substituting $\beta = \frac{1}{n^\mu}$ and using Equation (6), we get the result. ☐

C.1 Algorithm for general κ

Remark c_κ is an absolute constant dependent on the high probability performance guarantees of Algorithm 5. We can calculate that C_κ is at most $2^{12}(\sim 4000)$ and hence $c_\kappa \le 2(2^{12}\kappa)^{1/\kappa} \le 4 \cdot 2^{12/\kappa}$.

Theorem 8. *Assume each sample function F be L-Lipschitz and H-smooth, and let the population function f satisfy quadratic growth (Definition 2.4). Let Problem (1) be an interpolation problem. Then, Algorithm 6 is (ε,δ)-DP. For $\delta = 0$, Algorithm 6 with $T = \log n$ and $m = \frac{n}{\log n}$, we have*

$$f(\hat{x}_T) - f(x^\star) \le \widetilde{O}\left(\frac{1}{\sqrt{n}} + \frac{d}{n\varepsilon} \right)^{\frac{\kappa}{\kappa-2}},$$

with probability $1-\beta$. For $\delta > 0$, Algorithm 6 when run using $T = \log n$ and $m = n/\log n$ achieves error

$$f(\hat{x}_T) - f(x^\star) \le \widetilde{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}\log(1/\delta)}{n\varepsilon} \right)^{\frac{\kappa}{\kappa-2}},$$

with probability $1-\beta$.

Algorithm 6 Epoch based epoch based epoch based clipped-GD

Require: number of epochs: T, samples in each round: $m = n/T$, Diameter at the start: D_1, lipschitz constant at the start L_1, domain $\mathcal{X}_1$, initial point $\hat{x}_0$

1: **for** $i = 1$ to T **do**

2: $\hat{x}_i \leftarrow$ Output of Algorithm 5 when run on domain $\mathcal{X}_i$ (diameter D_i), with lipschitz constant L_i using m samples.

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

4:

$$\text{Set } D_{i+1} = c_\kappa \left(\frac{L_i}{\lambda} \max \left\{ \frac{\sqrt{\log(T/\beta)} \log^{3/2} m}{\sqrt{m}}, \frac{d \log(T/\beta) \log m}{m\varepsilon} \right\} \right)^{\frac{1}{\kappa-1}}$$

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

6:

$$\text{Set } D_{i+1} = c_\kappa \left(\frac{L_i}{\lambda} \max \left\{ \frac{\sqrt{\log(T/\beta)} \log^{3/2} m}{\sqrt{m}}, \frac{\sqrt{d \log(1/\delta)} \log(T/\beta) \log m}{m\varepsilon} \right\} \right)^{\frac{1}{\kappa-1}}$$

7: **end if**

8: Set $\mathcal{X}_{i+1} = \{\hat{x} : \|\hat{x} - \hat{x}_i\|_2 \leq D_{i+1}/2\}$

9: Set $L_{i+1} = HD_{i+1}$

10: **end for**

11: **return** the final iterate x_T

Proof The privacy guarantee follows from the proof of Theorem 2. We divide the utility proof into 2 main parts; first is to check the validity of the assumptions while applying Algorithm 5 and second is using its high probability convergence guarantees to get the final rates. To check this, we ensure that the optimum set lies in the new domain defined at every step and that the lipschitz constant defined with respect to the domain is a valid lipschitz constant. This follows from Lemma C.1.

Next, we use the high probability convergence guarantees of the subalgorithm Algorithm 5 to get convergence rates for Algorithm 2.

We prove it for the case when $\delta = 0$, the case when $\delta > 0$ is similar. We know that

$$\begin{aligned} L_i &= HD_i \\ &= c_\kappa H \left(\frac{L_{i-1}}{\lambda} \max \left\{ \frac{\sqrt{\log(T/\beta)} \log^{3/2} m}{\sqrt{m}}, \frac{d \log(T/\beta) \log m}{m\varepsilon} \right\} \right)^{\frac{1}{\kappa-1}}. \end{aligned}$$

Thus, we have

$$L_T = (c_\kappa H)^{\frac{\kappa-1}{\kappa-2}\left(1-\frac{1}{(\kappa-1)^{T-1}}\right)} \left(\frac{1}{\lambda} \max \left\{ \frac{\sqrt{\log(T/\beta)} \log^{3/2} m}{\sqrt{m}}, \frac{d \log(T/\beta) \log m}{m\varepsilon} \right\} \right)^{\frac{1}{\kappa-2}\left(1-\frac{1}{(\kappa-1)^{T-1}}\right)} L_1^{\frac{1}{(\kappa-1)^{T-1}}}.$$

We note that for $T \sim \log n$, $\frac{1}{(\kappa-1)^{T-1}} \approx \frac{1}{n^{\log(\kappa-1)}}$ and thus for large n, we ignore the terms of the form $a^{-\frac{1}{n^{\log(\kappa-1)}}}$ since they are ≈ 1. Ignoring these terms by including an additional constant C' we can write

$$L_T = C'(c_\kappa H)^{\frac{\kappa-1}{\kappa-2}} \left(\frac{1}{\lambda} \max \left\{ \frac{\sqrt{\log(T/\beta)} \log^{3/2} m}{\sqrt{m}}, \frac{d \log(T/\beta) \log m}{m\varepsilon} \right\} \right)^{\frac{1}{\kappa-2}} L_1^{\frac{1}{(\kappa-1)^{T-1}}}.$$

Using Theorem 2 of [ALD21] on the last epoch, we have with probability $1-\beta$ that

$$f(\hat{x}_T) - f(x^\star) \leq \frac{C_\kappa}{\lambda^{\frac{1}{\kappa-1}}} \max\left\{ \frac{L_T\sqrt{\log(T/\beta)}\log^{3/2} m}{\sqrt{m}}, \frac{L_T d \log(T/\beta)\log m}{m\varepsilon} \right\}^{\frac{\kappa}{\kappa-1}}$$
$$= \frac{(C')^{\frac{\kappa}{\kappa-1}} C_\kappa (c_\kappa H)^{\frac{\kappa}{\kappa-2}}}{\lambda^{\frac{2}{\kappa-2}}} \max\left\{ \frac{\sqrt{\log(T/\beta)}\log^{3/2} m}{\sqrt{m}}, \frac{d\log(T/\beta)\log m}{m\varepsilon} \right\}^{\frac{\kappa}{\kappa-2}} L_1^{\frac{\kappa}{(\kappa-1)^T}}.$$

Choosing $T = \log n$ and $m = n/\log n$, we have

$$f(\hat{x}_T) - f(x^\star) \leq \frac{(C')^{\frac{\kappa}{\kappa-1}} C_\kappa (c_\kappa H)^{\frac{\kappa}{\kappa-2}}}{\lambda^{\frac{2}{\kappa-2}}} \max\left\{ \frac{\sqrt{\log(\log n/\beta)}\log^{3/2}(n/\log n)}{\sqrt{n/\log n}}, \frac{d\log(\log n/\beta)\log(n/\log n)}{\varepsilon n/\log n} \right\}^{\frac{\kappa}{\kappa-2}} L_1^{\frac{\kappa}{n}}.$$

Now we write results in terms of sample complexity required to achieve a particular error. The sufficient number of samples. To ensure $f(\hat{x}_T) - f(x^\star) < \alpha$, it is sufficient to ensure

$$\frac{(C')^{\frac{\kappa}{\kappa-1}} C_\kappa (c_\kappa H)^{\frac{\kappa}{\kappa-2}}}{\lambda^{\frac{2}{\kappa-2}}} \max\left\{ \frac{\sqrt{\log(T/\beta)}\log^{3/2} m}{\sqrt{m}}, \frac{d\log(T/\beta)\log m}{m\varepsilon} \right\}^{\frac{\kappa}{\kappa-2}} L_1^{\frac{\kappa}{(\kappa-1)^T}} < \alpha.$$

Choosing $n = \tilde{O}\left(\max\{(\frac{1}{\alpha^2})^{\frac{\kappa-2}{\kappa}}, (\frac{d}{\varepsilon\alpha})^{\frac{\kappa-2}{\kappa}}\}\right)$ ensures error $\leq \alpha$. □

Corollary C.1. *Under the conditions of Theorem 8, for $\delta = 0$, the expected error of the output of algorithm is upper bounded by*

$$\mathbb{E}[f(\hat{x}_T) - f(x^\star)] \leq \widetilde{O}\left(\frac{1}{\sqrt{n}} + \frac{d}{n\varepsilon}\right)^{\frac{\kappa}{\kappa-2}},$$

for arbitrarily large μ. For $\delta > 0$, the expected error of the output of algorithm is upper bounded by

$$\mathbb{E}[f(\hat{x}_T) - f(x^\star)] \leq \widetilde{O}\left(\frac{1}{\sqrt{n}} + \frac{d}{n\varepsilon}\right)^{\frac{\kappa}{\kappa-2}},$$

for arbitrarily large μ.

C.2 (ε, δ) version of Theorem 3

Theorem 9. *Assume each sample function F be L-Lipschitz and H-smooth, and let the population function f satisfy quadratic growth (Definition 2.4) with coefficient λ. Let x_{adapt} be the output of Algorithm 3. Then,*

1. *Algorithm 3 is ε-DP.*
2. *Without any additional interpolation assumption, we have that the expected error of the x_{adapt} is upper bounded by*

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq LD \cdot \widetilde{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\varepsilon}\right)^2.$$

3. *Let problem (1) be an interpolation problem. Then, the expected error of the x_{adapt} is upper bounded by*

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq LD\left(\frac{1}{n^\mu} + \exp\left(-\widetilde{\Theta}\left(\frac{n\lambda^2}{H^2}\right)\right)\right.$$
$$\left. + \exp\left(-\widetilde{\Theta}\left(\frac{\lambda n\varepsilon}{H\sqrt{d\log(1/\delta)}}\right)\right)\right).$$

Proof First, we note that the privacy of Algorithm 3 follows from the privacy of Algorithm 2 and Algorithm 1 and post-processing.

To prove the convergence guarantees, we first need to show that the optimal set $\mathcal{X}^\star$ is included in the shrinked domain $\mathcal{X}_{\text{int}}$. Using the high probability guarantees of Algorithm 1, we know that with probability $1-\beta/2$, we have

$$f(x_1) - f(x^\star) \leq \frac{2^{12}L}{\lambda} \cdot^{\log(\kappa-1)} \left(\frac{\sqrt{\log(2/\beta)}\log^{3/2} n}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}\log(2/\beta)\log n}{n\varepsilon} \right).$$

Using the quadratic growth condition, we immediately have $\|x^\star - x_1\|_2 \leq D_{\text{int}}/2$ and hence $\mathcal{X}^\star \subset \mathcal{X}_{\text{int}}$.

Using smoothness, we have that for any $x \in \mathcal{X}_{\text{int}}$,

$$f(x) - f(x^\star) \leq \frac{HD_{\text{int}}^2}{2}.$$

Since Algorithm 2 always outputs a point in its input domain (in this case $\mathcal{X}_{\text{int}}$), even in the non-interpolation setting we have that

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq LD \cdot \widetilde{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\varepsilon} \right)^2.$$

In the interpolation setting, the guarantees of Algorithm 2 hold and the result is immediate. □

D Proofs from Section 5

D.1 Proof of Theorem 4

The proof is exactly the same as Theorem 1, except we set $k = \frac{\lambda n}{H}$ to ensure that $f_v(x)$ for any $v \in \mathcal{X}$ has λ-quadratic growth. Finally we set $\gamma = \frac{D}{2}\exp(\frac{-\lambda n\varepsilon}{Hd})$ and use the fact that $e^{\frac{\lambda n\varepsilon}{H}} \geq 2$ and the fact that $x/(x-1)$ is decreasing in x to give the desired lower bound.

D.2 Proof of Theorem 5

The proof of this result hinges on the two following supporting propositions. We first copy Proposition 2.2 from [AD20] (listed as Proposition 4 below) in our notation for convenience. We then state Proposition 5 which gives upper and lower bounds on the modulus of continuity (defined in Proposition 4). We note that the lower bound presented in Proposition 5 is one of the novel contributions of this paper; the proof of Proposition 5 can be found in Appendix D.2.1. We will first assume this to be true and prove Theorem 5 before returning prove its correctness.

Proposition 4. *For some fixed $F : \mathcal{X}, \Omega \to \mathbb{R}$ which is convex and H-smooth with respect to its first argument, let $\mathcal{S} \in \mathfrak{S}_\lambda^L(F)$ for $L = 2HD$. Let $x_{\mathcal{S}}^\star = \text{argmin}_{x' \in \mathcal{X}} f_{\mathcal{S}}(x')$. Define the corresponding modulus of continuity*

$$\omega(\mathcal{S}, 1/\varepsilon) \coloneqq \sup_{\mathcal{S}' \in \mathfrak{S}_\lambda^L(F)} \{|x_{\mathcal{S}}^\star - x_{\mathcal{S}'}^\star| : d_{ham}(\mathcal{S}, \mathcal{S}') \leq 1/\varepsilon\}.$$

Assume the mechanism M is ε-DP and for some $\gamma \leq \frac{1}{2e}$ achieves

$$\mathbb{E}[|M(\mathcal{S}) - x_{\mathcal{S}}^\star|] \leq \gamma\left(\frac{\omega(\mathcal{S}; 1/\varepsilon)}{2}\right).$$

Then there exists a sample $\mathcal{S}' \in \mathfrak{S}_\lambda^L(F)$ where $d_{ham}(\mathcal{S}, \mathcal{S}') \leq \frac{\log(1/2\gamma)}{2\varepsilon}$ such that

$$\mathbb{E}[|M(\mathcal{S}') - x_{\mathcal{S}'}^\star|] \geq \frac{1}{4}\ell\left(\frac{1}{4}\omega\left(\mathcal{S}'; \frac{\log(1/2\gamma)}{2\varepsilon}\right)\right).$$

Proposition 5. *Let $F : \mathcal{X}, \Omega \to \mathbb{R}$ be convex and H-smooth in its first argument and satisfying $\inf_{x\in\mathcal{X}} F(x; s) = 0$ for all $s \in \Omega$. Suppose we have some $\mathcal{S} \in \mathfrak{S}_\lambda^L(F)$ with $L = 2HD$ which also induces an interpolation problem (a problem which satisfies Definition 2.3). With respect to the dataset $\mathcal{S}$, the modulus of continuity $\omega(\mathcal{S}, 1/\varepsilon)$ satsifies*

$$\frac{D}{n\varepsilon} \leq \omega(\mathcal{S}, 1/\varepsilon) \leq \frac{8HD}{\lambda n\varepsilon}$$

With these two results, we can now prove Theorem 5. Restating the conditions of the theorem formally, suppose for some constants c_0 and c_1 there is an ε-DP estimator M such that

$$\mathbb{E}[f_{\mathcal{S}}(M(\mathcal{S}))] - \inf_{x\in\mathcal{X}} f_{\mathcal{S}}(x) \leq c_0 D^2 e^{-c_1(n\varepsilon)^t}.$$

If $t > 1$, set $t = \min(1, t)$, then the bound certainly still holds for large enough n. If we let $x^\star_{\mathcal{S}} = \operatorname{argmin}_{x\in\mathcal{X}} f_{\mathcal{S}}(x)$, using the definition of strong convexity, we have that there exists some c_2 and c_3 such that

$$\mathbb{E}[|M(\mathcal{S}) - x^\star_{\mathcal{S}}|] \leq c_2 D e^{-c_3(n\varepsilon)^t}$$

To satisfy the expression from Proposition 4, we select γ such that

$$\frac{\gamma\omega(\mathcal{S}; 1/\varepsilon)}{2} = c_2 D e^{-c_3(n\varepsilon)^t}.$$

Using Proposition 5 we must have $\frac{\lambda n\varepsilon}{4H} c_2 \exp(-c_3(n\varepsilon)^t) \leq \gamma \leq 2n\varepsilon c_2 \exp(-c_3(n\varepsilon)^t)$. Using Proposition 4, we have that

$$\mathbb{E}[|M(\mathcal{S}') - x^\star_{\mathcal{S}'}|] \geq \omega\left(\mathcal{S}'; \frac{\log(1/2\gamma)}{2\varepsilon}\right)$$

Before performing a further lower bound on this quantity, we first verify that $\frac{\log(1/2\gamma)}{2\varepsilon}$ does not exceed the total size of the dataset, n. Using our bounds on γ, we see that

$$\frac{\log(1/2\gamma)}{2\varepsilon} \leq \frac{1}{2\varepsilon}\left(c_3(n\varepsilon)^t - \log c_2 - \log\left(\frac{\lambda n\varepsilon}{2H}\right)\right)$$

For any $t \in (0, 1]$, for sufficiently large n, this quantity is less than n. We now lower bound the modulus of continuity by using the fact that it is a non-decreasing function in its second argument:

$$\begin{aligned}\mathbb{E}[|M(\mathcal{S}') - x^\star_{\mathcal{S}'}|] &\geq \omega\left(\mathcal{S}'; \frac{\log(1/2\gamma)}{2\varepsilon}\right) \geq \omega\left(\mathcal{S}'; \frac{c_3(n\varepsilon)^t - \log c_2 - \log(4n\varepsilon)}{2\varepsilon}\right) \\ &\geq \frac{D}{2n\varepsilon}\left[c_3(n\varepsilon)^t - \log c_2 - \log(4n\varepsilon)\right].\end{aligned}$$

This is the desired result; the last inequality comes from another application of Proposition 5 but with $\frac{c_3(n\varepsilon)^t - \log c_2 - \log(4n\varepsilon)}{2\varepsilon}$ in place of $1/\varepsilon$.

D.2.1 Proof of Proposition 5

Proof Outline

At a high level, starting with a function $f_{\mathcal{S}}$, we first remove an arbitrary $1/\varepsilon$ fraction to create a function $f_{\mathcal{S}}^{\setminus\varepsilon}$. We then replace the sample functions we removed with $1/\varepsilon$ samples of $\frac{H}{2}(x-D)^2$ and argue how far the minimizer of $f_{\mathcal{S}}^{\setminus\varepsilon} + \frac{H}{2n\varepsilon}(x-D)^2$ is away from the minimizer of $f_{\mathcal{S}}$. We will need many supporting lemmas to complete this proof; we quickly outline how we use these lemmas.

1. We use Lemma D.1 to argue that the minimizers of $f_{\mathcal{S}}^{\setminus\varepsilon}$ are no different that $f_{\mathcal{S}}$.
2. We use Lemma D.2 to argue about the growth of $f_{\mathcal{S}}^{\setminus\varepsilon}$.
3. We use Lemma D.3, Lemma D.4, Lemma D.5 to lower bound how far the minimizer of $f_{\mathcal{S}}^{\setminus\varepsilon} + \frac{H}{2n\varepsilon}(x-D)^2$ has moved from the minimizer of $f_{\mathcal{S}}$.
4. We use Lemma D.6 to upper bound how far the minimizer of $f_{\mathcal{S}}^{\setminus\varepsilon} + \frac{H}{2n\varepsilon}(x-D)^2$ has moved from the minimizer of $f_{\mathcal{S}}$.

We now formally introduce the several supporting lemmas which will aid our proof of Proposition 5. The first ensures that the minimizing set does not change upon the removal of a constant number of samples.

Lemma D.1. *Assume that* $\inf_{x\in\mathcal{X}} F(x;s) = 0$ *for all* $s \in \Omega$. *Suppose* $f_{\mathcal{S}}$ *satisfies Definition 2.3 and has* λ*-quadratic growth. Let* $\mathcal{X}^\star := \operatorname{argmin}_{x\in\mathcal{X}} f_{\mathcal{S}}(x)$. *Let* $\mathcal{S}_\varepsilon \subset \mathcal{S}$ *consist of any (constant not scaling with* n*)* $1/\varepsilon > 0$ *data points. Then, for* $f_{\mathcal{S}}^{\setminus\varepsilon} := \frac{1}{n}\sum_{s\in\mathcal{S}\setminus\mathcal{S}_\varepsilon} F(x;s)$ *we have that* $\mathcal{X}^\star_{\setminus\varepsilon} := \operatorname{argmin}_{x\in\mathcal{X}} f_{\mathcal{S}}^{\setminus\varepsilon}(x) = \mathcal{X}^\star$.

Proof Suppose for the sake of contradiction that $\mathcal{X}^\star \neq \mathcal{X}^\star_{\setminus\varepsilon}$ Since $f_{\mathcal{S}}$ is an interpolation problem, the removal of samples can only increase the size of $\mathcal{X}^\star_{\setminus\varepsilon}$. Suppose that $\mathcal{X}^\star_{\setminus\varepsilon} \setminus \mathcal{X}^\star \neq \emptyset$. There exists at most $1/\varepsilon$ points in $\mathcal{S}$ that have non-zero error on $\mathcal{X}^\star_{\setminus\varepsilon} \setminus \mathcal{X}^\star$. However, by smoothness of each sample function (and the fact that $f(x^\star) = 0$ and $f'(x^\star) = 0$ by construction), we have that for $x \in [a,b]$

$$f_{\mathcal{S}}(x) \leq \frac{H}{n\varepsilon}\operatorname{dist}(x, \mathcal{X}^\star)^2.$$

Since $\lim_{n\to\infty} \frac{H}{n\varepsilon} = 0$, this contradicts λ-quadratic growth. ☐

This second lemma ensures that deleting a constant number of samples does not affect the growth or strong convexity of the population function by too much.

Lemma D.2. *Assume that* $\inf_{x\in\mathcal{X}} F(x;s) = 0$ *for all* $s \in \Omega$. *Suppose* $f_{\mathcal{S}}$ *satisfies Definition 2.3 and has* λ*-quadratic growth (respectively* λ*-strong convexity). Let* $f_{\mathcal{S}}^{\setminus\varepsilon}$ *be defined as in Lemma D.1. Then* $f_{\mathcal{S}}^{\setminus\varepsilon}$ *has* γ*-quadratic growth (respectively* γ*-strong convexity) for any* $\gamma \leq \lambda - \frac{H}{n\varepsilon}$.

Proof By Lemma D.1, that the minimizing set of $f_{\mathcal{S}}^{\setminus\varepsilon}$ is the same as $f_{\mathcal{S}}$. Suppose for the sake of contradiction that $f_{\mathcal{S}}^{\setminus\varepsilon}$ does not have γ-quadratic growth. Then there must exist x_1 such that

$$f_{\mathcal{S}}^{\setminus\varepsilon}(x_1) - f_{\mathcal{S}}^{\setminus\varepsilon}(x^\star) < \frac{\gamma}{2}\|x_1 - x^\star\|_2^2.$$

By smoothness and growth we have

$$\frac{H}{2n\varepsilon}\|x_1 - x^\star\|_2^2 + \frac{\gamma}{2}\|x_1 - x^\star\|_2^2 > f_{\mathcal{S}}(x_1) - f_{\mathcal{S}}(x^\star) \geq \frac{\lambda}{2}\|x_1 - x^\star\|_2^2.$$

This implies that $\gamma > \lambda - \frac{H}{n\varepsilon}$, a contradiction.

Suppose for the sake of contradiction that $f_{\mathcal{S}}^{\setminus\varepsilon}$ does not have γ-strong convexity. Then there must exist x_1 and x_2 such that

$$f_{\mathcal{S}}^{\setminus\varepsilon}(x_1) - f_{\mathcal{S}}^{\setminus\varepsilon}(x_2) < \frac{\gamma}{2}\|x_1 - x^\star\|_2^2 + \langle \nabla f_{\mathcal{S}}^{\setminus\varepsilon}(x_2), x_1 - x_2\rangle.$$

By smoothness and strong convexity we have

$$\frac{H}{2n\varepsilon}\|x_1 - x_2\|_2^2 + \frac{\gamma}{2}\|x_1 - x_2\|_2^2 + \langle \nabla f_{\mathcal{S}}(x_2), x_1 - x_2\rangle > f_{\mathcal{S}}(x_1) - f_{\mathcal{S}}(x_2) \geq \frac{\lambda}{2}\|x_1 - x_2\|_2^2 + \langle \nabla f_{\mathcal{S}}(x_2), x_1 - x_2\rangle.$$

However, this implies that $\gamma > \lambda - \frac{H}{n\varepsilon}$ which is a contradiction. □

The next lemma is a standard result on the closure under addition of strongly convex functions.

Lemma D.3. *Let functions h_1 and h_2 be λ and γ strongly convex respectively, then $h_1 + h_2$ is $\lambda + \gamma$ strongly convex.*

This lemma provides some growth conditions on the gradient under smoothness, strong convexity and quadratic growth.

Lemma D.4. *Let $g : \mathcal{X} \to \mathbb{R}_+$ be a convex function with $\mathcal{X}^\star = \operatorname{argmin}_{x\in\mathcal{X}} g(x)$ such that for $x^\star \in \mathcal{X}^\star$, $g(x^\star) = 0$. Suppose g has λ-quadratic growth, then*

$$|g'(x)| \geq \frac{\lambda}{2}\operatorname{dist}(x, \mathcal{X}^\star).$$

If instead g has λ-strong convexity, then

$$|g'(x)| \geq \lambda \operatorname{dist}(x, \mathcal{X}^\star).$$

Alternatively, suppose g has H-smoothness, then

$$|g'(x)| \leq H \operatorname{dist}(x, \mathcal{X}^\star).$$

Proof We note that by first order optimality conditions, for all $x^\star \in \mathcal{X}^\star$, $\nabla g(x^\star) = 0$. To prove the first inequality, we have that for any $x^\star \in \mathcal{X}^\star$, the following is true:

$$\frac{\lambda}{2}\operatorname{dist}(x, \mathcal{X}^\star)^2 \leq g(x) - g^\star \leq |g'(x)||x - x^\star|.$$

In particular, minimizing over $x^\star$ on the right hand side and rearranging gives the desired result. To prove the second result, we know that by strong convexity for any $x^\star \in \mathcal{X}^\star$

$$|g'(x)| = |g'(x) - g'(x^\star)| \geq \lambda|x - x^\star|.$$

To prove the last result, we know that by smoothness for any $x^\star \in \mathcal{X}^\star$

$$|g'(x)| = |g'(x) - g'(x^\star)| \leq H|x - x^\star|.$$

Minimizing over $x^\star$ on the right hand side gives the desired result. □

This lemma controls how much the minimizers of a function can change if another function is added. This will directly be useful in lower bounding the modulus of continuity.

Lemma D.5. *Suppose $h : [-D, D] \to \mathbb{R}_+$ and $g : [-D, D] \to \mathbb{R}_+$. Let $x_h^\star$ be the largest minimizer of h and $x_g^\star$ be the smallest minimizer of g, and assume that $x_h^\star \leq x_g^\star$. Let $x^\star$ be any minimizer of $h + g$. Assume that $h(x_h^\star) = 0$ and $g(x_g^\star) = 0$. If h has λ_h-quadratic growth and g is H_g-smooth, then*

$$x^\star - x_h^\star \leq \frac{H_g(x_g^\star - x_h^\star)}{\frac{\lambda_h}{2} + H_g}.$$

If h is H_h-smooth and g has λ_g-quadratic growth, then

$$\frac{\frac{\lambda_g}{2}(x_g^\star - x_h^\star)}{\frac{\lambda_g}{2} + H_h} \leq x^\star - x_h^\star.$$

The same relation holds with $\lambda_g/2$ and $\lambda_h/2$ replaced with λ_g and λ_h respectively if the above statement is modified such that g and h are λ_g and λ_h strongly convex instead.

Proof If $x_h^\star \neq D$, then the first order condition for optimality implies

$$h'(x_h^\star) + g'(x_h^\star) = g'(x_h^\star) < 0 \quad \text{and} \quad h'(x_g^\star) + g'(x_g^\star) = h'(x_g^\star) > 0.$$

Thus, $x^\star \in (x_h^\star, x_g^\star)$. By the monotonicty of the first derivative of convex functions that for $x^\star \in (x_h^\star, x_g^\star)$, $g'(x^\star) < 0$ and $h'(x^\star) > 0$. Combining this with Lemma D.4, we get

$$\frac{\lambda_h}{2}(x^\star - x_h^\star) \leq h'(x^\star) \leq H_h(x^\star - x_h^\star)$$
$$H_g(x^\star - x_g^\star) \leq g'(x^\star) \leq \frac{\lambda_g}{2}(x^\star - x_g^\star).$$

Combining these facts gives

$$\frac{\lambda_h}{2}(x^\star - x_h^\star) + H_g(x^\star - x_g^\star) \leq h'(x^\star) + g('x^\star) = 0 \leq H_h(x^\star - x_h^\star) + \frac{\lambda_g}{2}(x^\star - x_g^\star)$$

Rearranging these two inequalities gives the desired result. We note that the lower bound only requires that h is H_h-smooth and g has λ_g-quadratic growth, and the upper bound only requires h has λ_h-quadratic growth and g is H_g-smooth. The last statement about strong convexity follows from the same reasoning, except using the strong convexity inequality in Lemma D.4 instead of the quadratic growth inequality. □

The following lemma is a slight modification of Claim 6.1 from [SSSSS09] and will be helpful for us to upper bound the modulus of continuity.

Lemma D.6. *Let $\mathcal{S}'$ consist of n data points where $|\mathcal{S} \triangle \mathcal{S}'| = k$. Suppose that $f_\mathcal{S}$ is λ-strongly convex and satisfies Definition 2.3. Assume the sample function $F : \mathcal{X} \times \Omega \to \mathbb{R}_+$ is L-Lipschitz in its first argument and that $\inf_{x \in \mathcal{X}} F(x; s) = 0$ for all $s \in \Omega$. For $x_\mathcal{S} \in \operatorname{argmin}_{x \in \mathcal{X}} f_\mathcal{S}(x)$ and $x_{\mathcal{S}'} \in \operatorname{argmin}_{x \in \mathcal{X}} f_{\mathcal{S}'}(x)$, we have that*

$$\|x_\mathcal{S} - x_{\mathcal{S}'}\|_2 \leq \frac{4kL}{\lambda n}.$$

Proof By strong convexity, we have that

$$f_\mathcal{S}(x_{\mathcal{S}'}) - f_\mathcal{S}(x_\mathcal{S}) \geq \frac{\lambda}{2} \|x_{\mathcal{S}'} - x_\mathcal{S}\|_2^2,$$

since by first order optimality conditions, we know that $\nabla f_\mathcal{S}(x_\mathcal{S}) = 0$ as a consequence of Definition 2.3. We also have

$$\begin{aligned} f_\mathcal{S}(x_{\mathcal{S}'}) - f_\mathcal{S}(x_\mathcal{S}) &= \frac{1}{n} \sum_{s \in \mathcal{S} \setminus \mathcal{S}'} [F(x_{\mathcal{S}'}; s) - F(x_\mathcal{S}; s)] + \frac{1}{n} \sum_{s \in \mathcal{S} \cap \mathcal{S}'} [F(x_{\mathcal{S}'}; s) - F(x_\mathcal{S}; s)] \\ &= \frac{1}{n} \sum_{s \in \mathcal{S} \setminus \mathcal{S}'} [F(x_{\mathcal{S}'}; s) - F(x_\mathcal{S}; s)] - \frac{1}{n} \sum_{s \in \mathcal{S}' \setminus \mathcal{S}} [F(x_{\mathcal{S}'}; s) - F(x_\mathcal{S}; s)] + f_{\mathcal{S}'}(x_{\mathcal{S}'}) - f_{\mathcal{S}'}(x_\mathcal{S}) \\ &\leq \frac{2kL}{n} \|x_{\mathcal{S}'} - x_\mathcal{S}\|_2, \end{aligned}$$

where the last inequality comes from the Lipschitzness of F and that $x_{\mathcal{S}'} \in \operatorname{argmin}_{x \in \mathcal{X}} f_{\mathcal{S}'}(x)$. □

Armed with these supporting lemmas, we can now bound the modulus of continuity. Let $x_0^\star$ be the largest minimizer of $f_{\mathcal{S}}$ following the steps of the proof outline. Without loss of generality, we assume that $x_0^\star \leq 0$. If $x_0^\star > 0$, by symmetry, it suffices to consider the problem replacing $\frac{H}{2}(x-D)^2$ with $\frac{H}{2}(x+D)^2$ in the following proof. By Lemma D.1, $f_{\mathcal{S}}^{\setminus\varepsilon}$ has the same minimizing set as $f_{\mathcal{S}}$. By Lemma D.2, $f_{\mathcal{S}}^{\setminus\varepsilon}$ has $\lambda - \frac{H}{n\varepsilon}$-strong convexity. Replace the $1/\varepsilon$ datapoints removed with samples that have the loss function $\frac{H}{2}(x-D)^2$; we note that it is clear that $\frac{H}{2}(x-D)^2$ satisifies the desired Lipschitz condition. Our constructed non-interpolation population function is

$$f_{\mathcal{S}}^{\setminus\varepsilon}(x) + \frac{H}{2n\varepsilon}(x-D)^2,$$

which is λ-strongly convex by Lemma D.2 and Lemma D.3 and is $2HD$-Lipschitz. This means that the $\mathcal{S}'$ this function corresponds to belongs to $\mathfrak{S}_\lambda^L(F)$. Let $x^\star$ be the minimizer of $f_{\mathcal{S}}^{\setminus\varepsilon}(x) + \frac{H}{2n\varepsilon}(x-D)^2$.

By the triangle inequality, we have $f_{\mathcal{S}}^{\setminus\varepsilon}$ is $\left(\frac{n-1/\varepsilon}{n}\right)H$-smooth. $\frac{H}{2n\varepsilon}(x-D)^2$ is $\frac{H}{n\varepsilon}$- strongly convex. Thus, by Lemma D.5 setting $h(x) \coloneqq f_{\mathcal{S}}^{\setminus\varepsilon}(x)$ and $g(x) \coloneqq \frac{H}{2n\varepsilon}(x-D)^2$, we have that

$$|x^\star - x_0^\star| = x^\star - x_0^\star \geq \frac{\frac{H}{n\varepsilon}(D - x_0^\star)}{\left(\frac{n-1/\varepsilon}{n}\right)H + \frac{H}{n\varepsilon}} = \frac{D - x_0^\star}{n\varepsilon} = \frac{D + |x_0^\star|}{n\varepsilon} \geq \frac{D}{n\varepsilon}.$$

Here, implicitly, we are using the fact that $x_0^\star$ is also a minimizer of $f_{\mathcal{S}}^{\setminus\varepsilon}$ by Lemma D.2. This completes the proof of the lower bound.

The upper bound follows from Lemma D.6 with $k = 1/\varepsilon$ and $L = 2HD$.