Faster Rates of Convergence to Stationary Points in Differentially Private Optimization

Raman Arora [*1] **Raef Bassily** [*2 3] **Tomás González** [*4] **Cristóbal Guzmán** [*4] **Michael Menart** [*2] **Enayat Ullah** [*1]

Abstract

We study the problem of approximating stationary points of Lipschitz and smooth functions under (ε, δ)-differential privacy (DP) in both the finite-sum and stochastic settings. A point $\widehat{w}$ is called an α-stationary point of a function $F : \mathbb{R}^d \to \mathbb{R}$ if $\|\nabla F(\widehat{w})\| \leq \alpha$. We give a new construction that improves over the existing rates in the stochastic optimization setting, where the goal is to find approximate stationary points of the population risk given n samples. Our construction finds a $\tilde{O}\big(\frac{1}{n^{1/3}} + \big[\frac{\sqrt{d}}{n\varepsilon}\big]^{1/2}\big)$-stationary point of the population risk in time linear in n. We also provide an efficient algorithm that finds an $\tilde{O}\big(\big[\frac{\sqrt{d}}{n\varepsilon}\big]^{2/3}\big)$-stationary point in the finite-sum setting. This improves on the previous best rate of $\tilde{O}\big(\big[\frac{\sqrt{d}}{n\varepsilon}\big]^{1/2}\big)$. Furthermore, under the additional assumption of convexity, we completely characterize the sample complexity of finding stationary points of the population risk (up to polylog factors) and show that the optimal rate on population stationarity is $\tilde{\Theta}\big(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n\varepsilon}\big)$. Finally, we show that our methods can be used to provide dimension-independent rates of $O\big(\frac{1}{\sqrt{n}} + \min\big(\big[\frac{\sqrt{\texttt{rank}}}{n\varepsilon}\big]^{2/3}, \frac{1}{(n\varepsilon)^{2/5}}\big)\big)$ on population stationarity for Generalized Linear Models (GLM), where `rank` is the rank of the design matrix, which improves upon the previous best known rate.

1. Introduction

Protecting users' data in machine learning models has become a central concern in multiple contexts, e.g. those involving financial or health data. In this respect, differential privacy (DP) is the gold standard for rigorous privacy protection (Dwork & Roth, 2014). Therefore, recent research has focused on the limits and possibilities of solving some of the most well-established machine learning problems under the constraint of DP. Despite intensive research, some fundamental problems remain not completely understood. One example is nonconvex optimization; namely, the task of approximating stationary points, which has been heavily studied in recent years in the non-private setting (Fang et al., 2018; Ma et al., 2018; Carmon et al., 2017; Nesterov & Polyak, 2006; Ghadimi & Lan, 2013; Arjevani et al., 2019; Foster et al., 2019). This problem is motivated by the intractability of nonconvex (global) optimization, as well as by a number of settings where stationary points have been shown to be global minima (Ge et al., 2016; Sun et al., 2016).

1.1. Contributions

In this work, we make progress towards resolving the complexity of approximating stationary points in optimization under the constraint of differential privacy, for both empirical and population risks. A summary of our new results is available in Table 1. In what follows, d is the problem dimension, n is the dataset size, and ε, δ are the approximate DP parameters.

Our first set of results pertains to the task of approximating stationary points of the population risk. Results for this problem are scarce. We provide the fastest rate up to date for this problem under DP, of $\tilde{O}\big(\frac{1}{n^{1/3}} + \big[\frac{\sqrt{d}}{n\varepsilon}\big]^{1/2}\big)$, with an algorithm that moreover has oracle complexity n (i.e., is single-pass). This algorithm is a noisy version of the SPIDER algorithm (Fang et al., 2018), whose gradient estimators are built using a tree-aggregation data structure for prefix-sums (Asi et al., 2021).

Next, we focus on the task of approximating stationary points in empirical nonconvex optimization (a.k.a. finite-sum case). In this context, we provide al-

[*]Equal contribution [1]Department of Computer Science, The Johns Hopkins University [2]Department of Computer Science & Engineering, The Ohio State University [3]Translational Data Analytics Institute (TDAI), The Ohio State University [4]Institute for Mathematical and Computational Engineering, Pontificia Universidad Católica de Chile. Correspondence to: Michael Menart <menart.2@osu.edu>, Enayat Ullah <enayat@jhu.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

gorithms with rate $O\big(\big[\frac{\sqrt{d}}{n\varepsilon}\big]^{2/3}\big)$, and oracle complexity[1] $\tilde{O}\big(\max\big\{\big(\frac{n^5\varepsilon^2}{d}\big)^{1/3}, \big(\frac{n\varepsilon}{\sqrt{d}}\big)^2\big\}\big)$. This rate is sharper than the best known for this problem (Wang et al., 2017).

We continue by investigating stationary points for convex losses and give an algorithm based on the recursive regularization technique of (Allen-Zhu, 2018) which achieves the optimal rate of $\tilde{\Theta}\big(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n\varepsilon}\big)$ on population stationarity. To establish optimality, we give a lower bound of $\Omega\big(\frac{\sqrt{d}}{n\varepsilon}\big)$ on empirical stationarity under DP (Theorem 4.3) and a non-private lower bound of $\Omega(\frac{1}{\sqrt{n}})$ on population stationarity (Theorem A.2). We also give a linear-time method, which achieves the optimal rate when the smoothness parameter is not so large. We conclude the paper showing a black-box reduction that converts any DP method for finding stationary points of smooth and Lipschitz losses into a DP method with *dimension-independent rates* for the case of generalized linear models (GLM). Using our proposed method with Private Spiderboost as the base algorithm yields a rate of $\tilde{O}\left(\frac{1}{\sqrt{n}} + \min\left(\big[\frac{\sqrt{\mathrm{rank}}}{n\varepsilon}\big]^{2/3}, \frac{1}{(n\varepsilon)^{2/5}}\right)\right)$ on population stationarity. This improves upon the result of (Song et al., 2021) which proposed a method with $\tilde{O}\big(\big[\frac{\sqrt{\mathrm{rank}}}{n\varepsilon}\big]^{1/2}\big)$ empirical stationarity[2].

1.2. Our Techniques

Our methods combine multiple techniques from optimization and differential privacy in novel ways. The lower bound for the empirical norm of the gradient uses fingerprinting codes to a loss similar to that used for Differentially Private-Empirical Risk Minimization (DP-ERM) (Bassily et al., 2014), crafted to work in the unconstrained case. This lower bound can be extended to the population gradient norm by a known re-sampling argument (Bassily et al., 2019). We also give a non-private lower bound of $\Omega\left(1/\sqrt{n}\right)$ on population stationarity with n samples which holds even in dimension 1, as opposed to previous results (Foster et al., 2019).

Efficient algorithms for (both empirical and population) norm of the gradient are derived using noisy versions of variance-reduced stochastic first order methods, which have proved remarkably useful in DP stochastic optimization (Asi et al., 2021; Bassily et al., 2021b;a). In the case of the empirical risk, we use a noisy version of SpiderBoost (Wang et al., 2019c). We remark that our methods can achieve comparable rates when applied to similar algorithms such as Spider (Fang et al., 2018) and Storm (Cutkosky & Orabona, 2019), but SpiderBoost allows for a larger learning rate which is considered better in practice. For the population risk, it is worth noting that the empirical norm of the gradient does not translate directly into population gradient guarantees, even if the algorithm in use is uniformly stable (Bousquet & Elisseeff, 2002), since this type of guarantee does not enjoy a *stability-implies-generalization* property. Therefore, we opt for single pass methods that combine variance-reduction with tree-aggregation; these techniques are particularly suitable for the classical Spider algorithm (Fang et al., 2018), which is the one we base our method on. For the convex setting, we use recursive regularization (Allen-Zhu, 2018) which was used to achieve the optimal non-private rate by (Foster et al., 2019).

Finally, our method for (non-convex) GLMs uses the Johnson-Lindenstrauss based dimensionality reduction technique similar to (Arora et al., 2022), which focused on the convex setting. Moreover, for population stationarity of GLMs, we give a new uniform convergence result of gradients of Lipschitz functions. This guarantee, unlike the prior work of (Foster et al., 2018), has only poly-logarithmic dependence on the radius of the constraint set, which is crucial for our analysis.

1.3. Related Work

The current work fits within the literature of differentially private optimization, which has primarily focused on the convex case (Chaudhuri et al., 2011; Jain et al., 2012; Kifer et al., 2012; Bassily et al., 2014; Talwar et al., 2014; Jain & Thakurta, 2014; Talwar et al., 2015; Bassily et al., 2019; Feldman et al., 2020; Asi et al., 2021; Bassily et al., 2021b). The culmination of this line of work for the convex smooth case showed that optimal rates are achievable in linear time (Feldman et al., 2020; Asi et al., 2021; Bassily et al., 2021b). Our work shows that in the convex case similar rates are achievable for the norm of the gradient: this result is useful, e.g., for dual formulations of linearly constrained convex programs (Nesterov, 2012), and moreover it has become a problem of independent interest (Allen-Zhu, 2018; Foster et al., 2019).[3]

Regarding stationary points for nonconvex losses, work in DP is far more recent, and primarily focused on the empirical stationarity (Wang et al., 2017; Zhang et al., 2017;

[1] We consider for complexity the first-order oracle model, standard for continuous optimization (Nemirovsky & Yudin, 1983).

[2] This is the rate obtained after fixing a mistake in the proof of Theorem 4.1 in (Song et al., 2021). Specifically, in their proof, the last term in Eq. (14) is missing a factor of T.

[3] To provide a specific example, consider the dual of the regularized discrete optimal transport problem, as discussed in (Diakonikolas & Guzmán, 2023), Section 5.6. If the marginals μ, ν in that model are accessed through i.i.d. samples, then this becomes an SCO problem. Moreover, it is argued in that reference that approximate stationary points provide approximately feasible and optimal transports through duality arguments. Hence, the result is an SCO problem where we require *approximate stationary points.*

Setting	Convergence	Our Rate		Previous best-known rate	
Non-convex	Empirical	$\left(\frac{\sqrt{d}}{n\varepsilon}\right)^{2/3}$	(Thm. 4.2)	$\left(\frac{\sqrt{d}}{n\varepsilon}\right)^{1/2}$	(Wang et al., 2017)
	Population	$\frac{1}{n^{1/3}} + \left(\frac{\sqrt{d}}{n\varepsilon}\right)^{1/2}$	(Thm. 3.2)	$\sqrt{d}\varepsilon + \left(\frac{\sqrt{d}}{n\varepsilon}\right)^{1/2}$	(Zhou et al., 2020)
Convex	Population	$\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n\varepsilon}$	(Thm. 5.1)	None	
Non-convex GLM	Empirical	$\left[\frac{\sqrt{\mathrm{rank}}}{n\varepsilon}\right]^{2/3} \wedge \frac{1}{(n\epsilon)^{2/5}}$	(Cor. 6.2)	$\left(\frac{\sqrt{\mathrm{rank}}}{n\varepsilon}\right)^{1/2}$	(Song et al., 2021)
	Population	$\frac{1}{\sqrt{n}} + \left[\frac{\sqrt{\mathrm{rank}}}{n\varepsilon}\right]^{2/3} \wedge \frac{1}{(n\epsilon)^{2/5}}$	(Cor. 6.2)	None	
Convex GLM	Population	$\frac{1}{\sqrt{n}} + \frac{\sqrt{\mathrm{rank}}}{n\varepsilon} \wedge \frac{1}{\sqrt{n\epsilon}}$	(Cor. 6.2)	None	

Table 1. Results summary: We omit log factors and function-class parameters. The symbol $\wedge$ stands for minimum of the quantities.

Wang & Xu, 2019; Wang et al., 2019a)[4]. Under similar assumptions to ours these works approximate stationary points with rate $\tilde{O}\left(\left[\frac{\sqrt{d}}{n\varepsilon}\right]^{1/2}\right)$, which is slower than ours.

Works addressing population guarantees for the norm of the gradient under DP are scarce. (Zhou et al., 2020) proposed a noisy gradient method, whose population guarantee is obtained by generalization properties of DP. However, the best guarantee obtainable with their analysis is $O\left(\left[\frac{\sqrt{d}}{n\varepsilon}\right]^{1/2} + \sqrt{d}\varepsilon\right)$[5]. Note that for any ε this rate is $\Omega\left([d/n]^{1/3}\right)$. Under additional assumptions (on the Hessian), (Wang & Xu, 2019) obtains a rate of $\tilde{O}(\sqrt{d/(n\varepsilon)})$ by uniform convergence of gradients, which is sharper when ε is constant. By contrast, our rate is much faster than both for $\varepsilon = \Theta(1)$. In particular, in this range, our rates are faster than those obtained by uniform convergence, $O(\sqrt{d/n})$ (Foster et al., 2018). Moreover, our method runs in time linear in n. On the other hand, in the much more restrictive setting where the loss satisfies the Polyak-Łojasiewicz (PL) inequality, (Zhang et al., 2021) provide *population risk* bounds of $\tilde{O}(d/[n\varepsilon]^2)$ under DP.

The work of (Bassily et al., 2021a) studies population guarantees for stationarity in constrained settings, obtaining rates $O\left(\frac{1}{n^{1/3}} + \left[\frac{\sqrt{d}}{n\varepsilon}\right]^{2/5}\right)$ in linear time. Notice first that these guarantees are based on the Frank-Wolfe gap, making those results incomparable to ours. Despite this fact, their rates are slower than ours.[6] On the other hand, they provide results for (close to nearly) stationary points in constrained/unconstrained settings, for a broader class of *weakly convex losses* (possibly nonsmooth). This result is then more general, but the rate of $O\left(\frac{1}{n^{1/4}} + \left[\frac{\sqrt{d}}{n\varepsilon}\right]^{1/3}\right)$ is substantially slower than ours, and their algorithm has oracle complexity which is superlinear in n.

The problem of stationary points in (nonprivate) stochastic optimization has drawn major attention recently (Ghadimi & Lan, 2013; 2016; Fang et al., 2018; Allen-Zhu, 2018; Foster et al., 2018; 2019; Arjevani et al., 2019). To the best of our knowledge, no lower bounds for the sample complexity[7] of this problem are known (beyond those known for the convex case (Foster et al., 2019)). On the other hand, oracle complexity is by now understood: in high dimensions, for (on average) smooth losses the optimal stochastic oracle complexity rate is $O(1/n^{1/3})$ (Arjevani et al., 2019). Although this provides some evidence of the sharpness of our results (see Appendix B.2), note that these lower bounds require very high dimensional constructions (namely, $d = \Omega(1/\alpha^4)$, where α is the rate), which limits their applicability in the private setting.

In an independent and concurrent work, (Tran & Cutkosky, 2022) achieve a rate of $O\left(\left[\frac{\sqrt{d}}{n\epsilon}\right]^{2/3} + \frac{1}{\sqrt{n}}\right)$ on the empirical gradient with gradient complexity $O(n^{7/3}\epsilon^{3/4}/d^{2/3})$ using a DP tree aggregation method. Note that our result removes the $1/\sqrt{n}$ term and improves the oracle complexity to $\tilde{O}\left(\max\left\{\left(\frac{n^5\varepsilon^2}{d}\right)^{1/3}, \left(\frac{n\varepsilon}{\sqrt{d}}\right)^2\right\}\right)$, which is better whenever

[4] Another work, (Wang et al., 2019b), claims to achieve this with improved oracle complexity. However, the analysis therein contains an error which is not easily fixed. Specifically, (Wang et al., 2019b, proof of Theorem 4.1) uses $\sigma_0^2 b_0^2 > 0.7$ to employ privacy amplification via subsampling. This is not true as they set $\sigma_0 = 1/[d^{1/4}\sqrt{n}]$ and $b_0 = \sqrt{n}/d^{1/4}$.

[5] (Zhou et al., 2020) omits the term $\sqrt{d}\varepsilon$, but this omission is only valid when $\varepsilon < 1/[n\sqrt{d}]^{1/3}$.

[6] We believe our methods can be extended to constrained settings using gradient mapping, a guarantee for which is stronger than for Frank-Wolfe gap (Lan, 2020, Section 7.5.1). We defer this extension to future work.

[7] Sample complexity is the fundamental limit on the sample size needed, as a function of α, to achieve α stationarity. This is different from the oracle complexity as one is not limited to first-order methods.

$d \leq n^2\epsilon^{1/4}$ (i.e. essentially whenever the error is nontrivial). Further, we accomplish this with a much simpler analysis.

2. Preliminaries

Let $f : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$ denote a (loss) function taking as input, the model parameter w and data point $x \in \mathcal{X}$. We assume that the function $w \mapsto f(w;x)$ is L_0-Lipschitz and L_1-smooth. That is, for all $x \in \mathcal{X}$ and $w_1, w_2 \in \mathbb{R}^d$, $|f(w_1;x) - f(w_2;x)| \leq L_0 \|w_1 - w_2\|$ and $\|\nabla f(w_1;x) - \nabla f(w_2;x)\| \leq L_1 \|w_1 - w_2\|$. Given a dataset $S \in \mathcal{X}^n$ of n points, we define the empirical risk as $F(w;S) = \frac{1}{n}\sum_{i=1}^n f(w;x_i)$. Assuming that the data points are sampled i.i.d. from an unknown distribution $\mathcal{D}$, the population risk, denoted as $F(w;\mathcal{D})$ is defined as $F(w;\mathcal{D}) = \mathbb{E}_{x\sim\mathcal{D}} f(w;x)$. Furthermore, we define $F_0 = F(0;S) - \min_{w\in\mathbb{R}^d}\{F(w;S)\}$ when discussing the empirical case and similarly for the population loss when discussing stationary points of the population loss. We use w^* to denote the population risk minimizer. Finally, we use the notation $\mathbb{I}_d$ to denote the $d \times d$ identity matrix and use $[a]$ to denote the set $\{1, 2, ..., a\}$ for $a \geq 1$.

Stationary points: Given a dataset S, our goal is to find an α-stationary point $\bar{w}$ of either empirical or population risk; formally, $\|\nabla F(\bar{w};S)\| \leq \alpha$ or $\|\nabla F(\bar{w};\mathcal{D})\| \leq \alpha$, respectively.

Differential Privacy (DP) (Dwork et al., 2006): An algorithm $\mathcal{A}$ is (ε,δ)-differentially private if for all datasets S and S' differing in one data point and all events $\mathcal{E}$ in the range of the $\mathcal{A}$, we have, $\mathbb{P}(\mathcal{A}(S) \in \mathcal{E}) \leq e^{\varepsilon}\mathbb{P}(\mathcal{A}(S') \in \mathcal{E}) + \delta$.

Generalized Linear Models (GLMs): For data domain $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$, a loss function $f : \mathbb{R}^d \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a GLM if $f(w;(x,y)) = \phi_y(\langle w, x\rangle)$ for some function ϕ_y. Our result for GLMs uses random matrices which satisfy the Johnson-Lindenstrauss (JL) property, defined as follows.

Definition 2.1 ((γ,β)-JL property)**.** A random matrix $\Phi \in \mathbb{R}^{k\times d}$ satisfies (γ,β)-JL property if for any $u, v \in \mathbb{R}^d$, $\mathbb{P}[|\langle \Phi u, \Phi v\rangle - \langle u, v\rangle| > \gamma \|u\| \|v\|] \leq \beta$.

3. Stationary Points of Population Risk

For the population gradient, we provide a linear time algorithm; see Algorithm 1 for pseudocode. It is a noisy variant of SPIDER (Fang et al., 2018), and utilizes a variance reduction technique tailored to an underlying binary tree structure. Namely, we run T rounds, where at the beginning of round t we build a binary tree of depth D, whose nodes are denoted by $u_{t,s}$, where $s \in \{0,1\}^D$. Every node $u_{t,s}$ is associated with a parameter vector $w_{t,s}$ and a gradient estimate $\nabla_{t,s}$. Next, we perform a Depth-First-Search traversal of the tree. We denote by $\mathrm{DFS}[D]$ the set of nodes in the visiting order excluding the root, for example: $\mathrm{DFS}[2] = \{u_0, u_{00}, u_{01}, u_1, u_{10}, u_{11}\}$. When a left child node is visited, it receives the same parameter vector and gradient estimator of the parent node.

Algorithm 1 Tree-based Private Spider

Input: $S = (x_1, \ldots, x_n) \in \mathcal{X}^n$: private dataset, (ε,δ): privacy parameters, T: number of rounds, b: batch size at beginning of each round, D: depth of trees at each round, β: step-size parameter, $\tilde{\alpha}$: accuracy parameter.

$w_{0,\ell(2^D-1)} = 0$
for $t = 1$ to T **do**
  Set $w_{t,\varnothing} = w_{t-1,\ell(2^D-1)}$
  Draw a batch $S_{t,\varnothing}$ of b data points, set $S \leftarrow S \setminus S_{t,\varnothing}$.
  Set $\sigma^2_{t,\varnothing} := \frac{8L_0^2\log(1.25/\delta)}{b^2\varepsilon^2}$.
  $\nabla_{t,\varnothing} = \frac{1}{b}\sum_{x\in S_{t,\varnothing}} \nabla f(w_{t,\varnothing};x) + g_{t,\varnothing}$, where $g_{t,\varnothing} \sim \mathcal{N}\left(0, \mathbb{I}_d\sigma^2_{t,\varnothing}\right)$.
  for $u_{t,s} \in \mathrm{DFS}[D]$ **do**
    Let $s = \widehat{s}c$, where $c \in \{0,1\}$.
    if $c = 0$ **then**
      $\nabla_{t,s} = \nabla_{t,\widehat{s}}$
      $w_{t,s} = w_{t,\widehat{s}}$
    else
      Draw a batch $S_{t,s}$ of $\frac{b}{2^{|s|}}$ data points, set $S \leftarrow S \setminus S_{t,s}$.
      Set noise variance $\sigma^2_{t,s} := \frac{8\cdot 2^D\beta^2\log(1.25/\delta)}{b^2\varepsilon^2}$.
      $\Delta_{t,s} = \frac{2^{|s|}}{b}\sum_{x\in S_{t,s}} (\nabla f(w_{t,s};x) - \nabla f(w_{t,\widehat{s}};x)) + g_{t,s}$, where $g_{t,s} \sim \mathcal{N}\left(0, \mathbb{I}_d\sigma^2_{t,s}\right)$.
      $\nabla_{t,s} = \nabla_{t,\widehat{s}} + \Delta_{t,s}$.
    end if
    if $|s| = D$ (i.e, $u_{t,s}$ is a leaf) **then**
      if $\|\nabla_{t,s}\| \leq 2\tilde{\alpha}$ **then**
        Return $w_{t,s}$
      end if
      Let u_{t,s^+} be the next vertex in $\mathrm{DFS}[D]$.
      Set $\eta_{t,s} := \frac{\beta}{2^{D/2}L_1\|\nabla_{t,s}\|}$
      $w_{t,s^+} = w_{t,s} - \eta_{t,s}\nabla_{t,s}$.
    end if
  end for
end for
Return $\overline{w}$, chosen uniformly at random from $\{w_{t,s} : t \in [T], u_{t,s} \text{ is a leaf}\}$.

On the other hand, when a right child node is visited, it receives a fresh set of samples and uses it to update the gradient estimator coming from the parent node. Every time a leaf node is reached, a gradient step is performed using the gradient estimator associated to the leaf. Finally, the parameter vector of a right child node comes from the gradient step performed at the right-most leaf in the left sub-

tree of it. The use of the binary tree structure is benefitial because every gradient estimator is updated at most D times within a round of 2^D optimization steps, as opposed to the original SPIDER algorithm where the gradient estimators are updated at every optimization step. This way, we are able to perform the same number of optimization steps but adding substantially smaller amounts of noise, leading to a faster rate than the one we would get without using the tree. In the following, we denote by $\ell(k)$ the binary representation of any number $k \in [0, 2^D - 1]$ and by $|s|$ the depth of $u_{t,s}$ for any $t \in [T]$.

The proposed algorithm is similar to the one in Section 5 of (Bassily et al., 2021b) for constrained Differentially Private-Stochastic Convex Optimization (DP-SCO), with the key difference that Algorithm 1 executes each round with fixed depth trees, which is key for our convergence analysis, whereas the prior work leverages convexity to construct trees that increase depth by one at each round. In addition, to choose the step-size in (Bassily et al., 2021b) the authors leverage the bounded diameter of the domain, while our step-size is chosen as that of (Fang et al., 2018), i.e. normalized by the norm of the gradient estimator and proportional to the target accuracy. This choice is crucial for controlling the sensitivity of the gradient variation estimator in the unconstrained setting, and consequently for the privacy analysis as well. Our results are presented below and the proofs are deferred to Appendix C.

Theorem 3.1 (Privacy guarantee). *For any $\varepsilon, \delta \in [0, 1]$, Algorithm 1 is (ε, δ)-DP.*

Theorem 3.2 (Accuracy guarantee). *Let $p \in (0, 1)$, $\varepsilon, \delta > 0$, $b = \max\left\{n^{2/3}, \frac{\sqrt{n}d^{1/4}}{\sqrt{\varepsilon}}\right\}$, D be such that $D2^{D+1} = b$, $T = \frac{n}{b(D/2+1)}$, $\alpha = \sqrt{2}L_0 \max\left\{\frac{1}{n^{1/3}}, \left(\frac{\sqrt{d}}{n\varepsilon}\right)^{1/2}\right\}$, $\beta = \alpha \min\{1, \frac{\sqrt{b}\varepsilon}{\sqrt{d}}\}$, and $\tilde{\alpha} = \tilde{C}\alpha$, where $\tilde{C} = 256 \log\left(\frac{1.25}{\delta}\right) \log\left(\frac{2T2^{D+1}}{p}\right) + \frac{8L_1F_0\sqrt{2D}(D/2+1)}{2L_0^2}$. Then, for any $n \geq \max\{\sqrt{d}(\frac{D}{2}+1)^2/\varepsilon, (\frac{D}{2}+1)^3\}$, with probability $1-p$, Algorithm 1 ends in line 20, returning an iterate $w_{t,s}$ with*

$$\|\nabla F(w_{t,s}; \mathcal{D})\| \leq 3\sqrt{2}L_0\tilde{C} \max\left\{\frac{1}{n^{1/3}}, \left(\frac{\sqrt{d}}{n\varepsilon}\right)^{1/2}\right\}.$$

Furthermore, Algorithm 1 has oracle complexity of n.

4. Stationary Points of Empirical Risk

4.1. Efficient Algorithm with Faster Rate

The algorithm for our upper bound is a noisy version of the SpiderBoost algorithm (Wang et al., 2019c)[8]. The algorithm works by running a series of phases of length q. Each phase starts with a minibatch estimate of the gradient, and subsequent gradient estimates within the phase are then computed by adding an estimate of the gradient variation. The key to the analysis is to bound the error in the gradient estimate at each iteration. Towards this end, we have the following generalization of the (Wang et al., 2019c) Lemma 1, which follows directly from (Fang et al., 2018) Proposition 1.

Lemma 4.1. *Consider Algorithm 2, and for any $t \in \{0, .., T\}$ let $s_t = \left\lfloor \frac{t}{q} \right\rfloor q$. If each ∇_t computed in line 9 is an unbiased estimate of $\nabla F(w_t; S)$ satisfying $\mathbb{E}\left[\|\nabla_{s_t} - \nabla F(w_{s_t}; S)\|^2\right] \leq \tau_1^2$ and each Δ_t computed in line 13 is an unbiased estimate of the gradient variation satisfying $\mathbb{E}\left[\|\Delta_t - [\nabla F(w_t; S) - \nabla F(w_{t-1}; S)]\|^2\right] \leq \tau_2^2 \|w_t - w_{t-1}\|^2$. Then for any $t \geq s_t + 1$, the iterates of Algorithm 2 satisfy*

$$\mathbb{E}\left[\|\nabla_t - \nabla F(w_t)\|^2\right] \leq \tau_2^2 \sum_{k=s_t+1}^{t} \mathbb{E}\left[\|w_k - w_{k-1}\|^2\right] + \tau_1^2.$$

For privacy, using smoothness we observe the sensitivity of the gradient variation estimate at iteration t is proportional to $\beta \|w_t - w_{t-1}\|$. Thus we can apply the above lemma with $\tau_1^2 = \frac{L_0^2}{b_1} + L_0^2\sigma_1^2$ and $\tau_2^2 = \frac{L_1^2}{b_2} + L_1^2\sigma_2^2$ (note the Gaussian noise in line 13 is drawn with variance scale at most $\sigma_2^2 \|w_t - w_{t-1}\|^2$). By carefully balancing the algorithm parameters, we are then able to obtain the following result. The full proof is deferred to Appendix B.1.

Theorem 4.2 (Private Spiderboost ERM). *Let $\varepsilon, \delta \in [0, 1]$. Let $n \geq \max\left\{\frac{(L_0\varepsilon)^2}{F_0L_1d\log(1/\delta)}, \frac{\sqrt{d}\max\{1, \sqrt{L_1F_0}/L_0\}}{\varepsilon}\right\}$. Algorithm 2 is (ε, δ)-DP. Further, there exist settings of T, η, q, b_1, b_2 such that Algorithm 2 has $\mathbb{E}[\|\nabla F(\bar{w}; S)\|]$ bounded as*

$$O\left(\left(\frac{\sqrt{F_0L_1L_0}\sqrt{d\log(1/\delta)}}{n\varepsilon}\right)^{2/3} + \frac{L_0\sqrt{d\log(1/\delta)}}{n\varepsilon}\right)$$

and oracle complexity $\tilde{O}\left(\max\left\{\left(\frac{n^{5/3}\varepsilon^{2/3}}{d^{1/3}}\right), \left(\frac{n\varepsilon}{\sqrt{d}}\right)^2\right\}\right)$.

Note that the restriction on n in the theorem statement is essentially trivial when the upper bound is nontrivial. We remark that the case where the dominant error term is $\alpha = \tilde{O}\left(\left[\frac{\sqrt{d}}{n\varepsilon}\right]^{2/3}\right)$, then we approximately have oracle complexity $\tilde{O}\left(\max\left\{\frac{1}{\alpha^3}, \frac{n}{\alpha}\right\}\right)$.

4.2. Lower Bound

We now show a lower bound for the sample complexity of finding a stationary point under differential privacy in the unconstrained setting, which shows that the $O\left(\frac{L_0\sqrt{d\log(1/\delta)}}{n\varepsilon}\right)$

[8] SpiderBoost itself is essentially the Spider algorithm (Fang et al., 2018) with a different learning rate and analysis.

Algorithm 2 Private SpiderBoost

Input: Dataset: $S \in \mathcal{X}^n$, Function: $f : \mathbb{R}^d \times \mathcal{X} \mapsto \mathbb{R}$, Learning Rate: η, Phase Size: q, Batch Sizes b_1, b_2, Privacy Parameters: (ε, δ), Iterations: T

$w_0 = 0$

$\sigma_1 = \frac{cL_0\sqrt{\log(1/\delta)}}{\varepsilon} \max\left\{\frac{1}{b_1}, \frac{\sqrt{T}}{\sqrt{q}n}\right\}$, where c is a universal constant.

$\sigma_2 = \frac{cL_1\sqrt{\log(1/\delta)}}{\varepsilon} \max\left\{\frac{1}{b_2}, \frac{\sqrt{T}}{n}\right\}$

$\widehat{\sigma}_2 = \frac{2cL_0\sqrt{\log(1/\delta)}}{\varepsilon} \max\left\{\frac{1}{b_2}, \frac{\sqrt{T}}{n}\right\}$

for $t = 0, \ldots, T$ **do**

  if $\mod (t, q) = 0$ **then**

    Sample batch S_t of size b_1

    Sample $g_t \sim \mathcal{N}(0, \mathbb{I}_d \sigma_1^2)$

    $\nabla_t = \frac{1}{b_1} \sum_{x \in S_t} \nabla f(w_t; x) + g_t$

  else

    Sample batch S_t of size b_2

    $g_t \sim \mathcal{N}\left(0, \mathbb{I}_d \min\left\{\sigma_2^2 \left\|w_t - w_{t-1}\right\|^2, \widehat{\sigma}_2^2\right\}\right)$

    $\Delta_t = \frac{1}{b_2} \sum_{x \in S_t} \left[\nabla f(w_t; x) - \nabla f(w_{t-1}; x)\right] + g_t$

    $\nabla_t = \nabla_{t-1} + \Delta_t$

  end if

  $w_{t+1} = w_t - \eta \nabla_t$

end for

return $\bar{w}$ uniformly at random from $\{w_1, \ldots, w_T\}$

term in the rate given in Theorem 4.2 is necessary. Furthermore, as our lower bound holds for all levels of smoothness, it also shows that our rate in Theorem 4.2 is optimal in the (admittedly uncommon) regime where $L_1 \leq \frac{\sqrt{d}L_0^2}{F_0 n\varepsilon}$. Our lower bound in fact holds even for convex functions. Furthermore, this result implies the same lower bound (up to log factors) for the population gradient using the technique in (Bassily et al., 2019), Appendix C.

Theorem 4.3. *Given $L_0, L_1, n, \varepsilon = O(1), 2^{-\Omega(n)} \leq \delta \leq 1/n^{1+\Omega(1)}$, there exists an L_0-Lispchitz, L_1-smooth (convex) loss $f : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$ and a dataset S of n points such that any (ε, δ)-DP algorithm run on S with output $\bar{w}$ satisfies,*

$$\|\nabla F(\bar{w}; S)\| = \Omega\left(L_0 \min\left(1, \frac{\sqrt{d\log(1/\delta)}}{n\varepsilon}\right)\right).$$

The proof is based on a reduction to DP mean estimation. Specifically, we consider a instance of the Huber loss function for which the minimizer is the empirical mean of the dataset. We then argue that close to the minimizer, the empirical stationarity is lower bounded by DP mean estimation bound (Steinke & Ullman, 2015), and far away, by construction, the empirical stationarity is L_0.

Proof of Theorem 4.3. For any $r > 0$, let $\mathcal{W}_r$ denote the ball of radius r centered at the origin. Let $B = \frac{L_0}{L_1}$. Consider the loss function:

$$f(w; x) = \begin{cases} \frac{L_1}{2}\|w - x\|^2 & \text{if } \|w - x\| \leq B \\ L_0\|w - x\| - \frac{L_0^2}{2L_1} & \text{otherwise} \end{cases}$$

The function $f(w; x)$ is convex, L_1-smooth and L_0-Lispchitz in $\mathbb{R}^d$. We restrict to datasets $S = \{x_i\}_{i=1}^n$ where $x_i \in \mathcal{W}_{B/4}$ for all i, and let $F(w; S) = \frac{1}{n}\sum_{i=1}^n f(w; x_i)$ be the empirical risk on S. The unconstrained minimizer of $F(w; S)$ is $w^* = \frac{1}{n}\sum_{i=1}^n x_i$ which lies in $\mathcal{W}_{B/4}$.

For any $w \in \mathcal{W}_{3B/4}$, w lies in the quadratic region around all data points. Hence, from L_1-strong convexity of $w \mapsto F(w; S)$ on $\mathcal{W}_{3B/4}$, we have that whenever $\bar{w} \in \mathcal{W}_{3B/4}$,

$$\begin{aligned} \|\nabla F(\bar{w}; S)\| \|\bar{w} - w^*\| &\geq \langle \nabla F(\bar{w}; S), w^* - \bar{w}\rangle \\ &\geq F(\bar{w}; S) - F(w^*; S) \\ &\geq \frac{L_1}{2}\|\bar{w} - w^*\|^2. \end{aligned}$$

Let E be the event that $\bar{w} \in \mathcal{W}_{3B/4}$ and let $\mathbb{E}_E$ denote the conditional expectation (conditioned on event E) operator. Then,

$$\begin{aligned} \mathbb{E}_E\|\nabla F(\bar{w}; S)\| &\geq \frac{L_1}{2}\mathbb{E}\|\bar{w} - w^*\| \\ &\geq \frac{L_1}{2}\Omega\left(\left(\frac{L_0}{4L_1}\right)\min\left(1, \frac{\sqrt{d\log(1/\delta)}}{n\varepsilon}\right)\right). \end{aligned}$$

where the last inequality follows from known lower bounds for DP mean estimation (Steinke & Ullman, 2015; Kamath & Ullman, 2020). We remark that the lower bound in the referenced work is for algorithms which produce outputs in the ball of the same radius as the dataset, i.e. $\mathcal{W}_{B/4}$. However, a simple post-processing argument shows that the same lower bound applies to algorithms which produce output in $\mathcal{W}_{3B/4}$. Specifically, assuming the contrary, we simply project the output in $\mathcal{W}_{3B/4}$ to $\mathcal{W}_{B/4}$: privacy is preserved by post-processing and the distance to the mean cannot increase by the non-expansiveness property of projection to convex sets, hence a contradiction. This gives us,

$$\mathbb{E}_E\left[\|\nabla F(\bar{w}; S)\|\right] \geq \Omega\left(L_0 \min\left(1, \frac{\sqrt{d\log(1/\delta)}}{n\varepsilon}\right)\right)$$

Let $\tilde{\mathcal{W}} = \{w : \|w - w^*\| \leq B/2\}$. Since $\tilde{\mathcal{W}} \subseteq \mathcal{W}_{3B/4}$, we have that the above conditional lower bound applies for $\bar{w} \in \tilde{\mathcal{W}}$ as well. We now consider $\bar{w} \notin \tilde{\mathcal{W}}$. Let w' be *any* point on the boundary of $\tilde{\mathcal{W}}$, denoted as $\partial\mathcal{W}$. Note that w' lies in the region where, for any data point, the

corresponding loss is a quadratic function. Hence, by direct computation, $\nabla F(w'; S) = L_1 (w' - w^*)$. Therefore,

$$\langle \nabla F(w'), w' - w^* \rangle = L_1 \|w' - w^*\|^2 = \frac{L_1 B^2}{4}.$$

We now apply gradient monotonicity to obtain the following (see Lemma A.1, Appendix A),

$$\mathbb{E}_{E^c} \|\nabla F(\bar{w}; S)\| \geq \frac{L_1 B^2}{4} \cdot \frac{2}{B} = \frac{L_0}{2},$$

where E^c denotes the complement set of E. We combine the above bounds using the law of total expectation as follows,

$$\begin{aligned}
&\mathbb{E}[\|\nabla F(\bar{w}; S)\|] \\
&= \mathbb{E}_E[\|\nabla F(\bar{w}; S)\|]\mathbb{P}\{\bar{w} \in E\} + \mathbb{E}_{E^c}[\|\nabla F(\bar{w}; S)\|]\mathbb{P}\{\bar{w} \in E^c\} \\
&= \Omega\Big(L_0 \min\Big\{1, \frac{\sqrt{d \log(1/\delta)}}{n\varepsilon}\Big\}\Big)\mathbb{P}(\bar{w} \in E) + \Omega(L_0)\mathbb{P}(\bar{w} \in E^c) \\
&= \Omega\Big(L_0 \min\Big\{1, \frac{\sqrt{d \log(1/\delta)}}{n\varepsilon}\Big\}\Big).
\end{aligned}$$

This completes the proof. □

Challenges for Further Rate Improvements: Given the above lower bound, the question arises as to whether the $\tilde{O}\big(\big[\frac{\sqrt{d}}{n\varepsilon}\big]^{2/3}\big)$ term can be improved. An informal argument using the oracle complexity lower bound of (Arjevani et al., 2019) suggests several major challenges in obtaining further rate improvements. A more detailed version of the following discussion can be found in Appendix B.2.

Consider methods which ensure privacy by directly privatizing the gradient/gradient variation queries. The aim of such methods is to design some private stochastic first order oracle, $\mathcal{O}_{\varepsilon',\delta'}$, such that a set of G queries to $\mathcal{O}_{\varepsilon',\delta'}$ satisfies (ε, δ)-DP, and use this oracle in some optimization algorithm $\mathcal{A}(\mathcal{O}_{\varepsilon',\delta'})$. Such a setup encapsulates numerous results in the convex setting (Bassily et al., 2019; Kulkarni et al., 2021), and is even more dominant in non-convex settings (Wang et al., 2017; Zhou et al., 2020; Abadi et al., 2016). Under advanced composition based arguments, to make G calls to such a private oracle one needs $\varepsilon' \leq \varepsilon/\sqrt{G}$. Now, standard fingerprinting code arguments suggest lower bounds on the level of accuracy of any such private oracle (Steinke & Ullman, 2015). Specifically, without leveraging further problem structure beyond Lipschitzness, one needs the gradient estimation error to be at least $\tau_1 = \Omega\Big(\frac{L_0\sqrt{Gd\log(1/\delta)}}{n\varepsilon}\Big)$. A similar argument suggests the error in the gradient variation between iterates w, w' must at least $\tau_2 \|w - w'\| = \Omega\Big(\frac{L_1\|w-w'\|\sqrt{Gd\log(1/\delta)}}{n\varepsilon}\Big)$. Now consider some optimization algorithm, $\mathcal{A}$, which takes as input a stochastic oracle $\mathcal{O}$ for some smooth function $\mathcal{L}$. The lower bound of (Arjevani et al., 2019) suggests that if $\mathcal{A}$ makes at most G queries to $\mathcal{O}$ (as a black box) the algorithm satisfies $\mathbb{E}\left[\|\nabla \mathcal{L}(\mathcal{A}(\mathcal{O}))\|\right] = \Omega\left(\left(\frac{F_0\tau_2\tau_1}{G}\right)^{1/3} + \frac{\tau_1}{\sqrt{G}}\right)$. If $\mathcal{O}$ is a private oracle satisfying the previously mentioned conditions, we would then have under the setting of τ_1 and τ_2 suggested by privacy that the convergence guarantee for $\mathbb{E}\left[\|\nabla \mathcal{L}(\mathcal{A}(\mathcal{O}))\|\right]$ is lower bounded as

$$\Omega\left(\left(\frac{\sqrt{F_0 L_1 L_0}\sqrt{d\log(1/\delta)}}{n\varepsilon}\right)^{2/3} + \frac{L_0\sqrt{d\log(1/\delta)}}{n\varepsilon}\right).$$

This indicates a substantial challenge for future rate improvements, as alternative methods which avoid private gradients (see e.g. (Feldman et al., 2020)) rely crucially on stability guarantees arising from convexity.

5. Stationary Points in the Convex Setting

Algorithm 3 Recursive Regularization

Input: Dataset S, loss function f, steps T, $\{\lambda_t\}_t$, $\{R_t\}_t$, PrivateSubRoutine, number of steps of sub-routine $\{K_t\}$, selector functions $\{\mathcal{S}_t(\cdot)\}_t$, step size $\{\eta_t\}_t$, noise variances $\{\sigma_t\}_t$

$w_0 = 0$, $n_0 = 1$
Define function $(w, x) \mapsto f^{(0)}(w; x) = f(w; x) + \frac{\lambda_0}{2}\|w - w_0\|^2$
for $t = 1$ to $T - 1$ **do**
  $n_t = n_{t-1} + \left\lfloor \frac{|S|}{T} \right\rfloor$
  $\bar{w}_t = \mathsf{PrivateSubRoutine}(S_{n_{t-1}:n_t}, f^{(t-1)}, R_t, K_t, \eta_t, \mathcal{S}_t(\cdot), \sigma_t)$
  Define function $(w, x) \mapsto f^{(t)}(w; x) = f^{(t-1)}(w; x) + \frac{\lambda_t}{2}\|w - \bar{w}_t\|^2$
end for

Output: $\bar{w} = \bar{w}_T$

In this section, we additionally assume that the loss function is convex. The motivation for this is two-fold: firstly, this setting has recently gained attention in a non-private setting (Nesterov, 2012; Allen-Zhu, 2018; Foster et al., 2019). Secondly, in this setting we are able to establish tightly the sample complexity of approximate stationary points.

Our method is based on the recursive regularization technique proposed in (Allen-Zhu, 2018), and further improved by (Foster et al., 2019). The main idea, as the name suggests, is to recursively regularize the objective and optimize it via some solver. For the DP setting, the key idea is to use a private sub-routine as the inner solver. Furthermore, while a solver for the unconstrained problem suffices non-privately, we need to carefully increase the radius of the constrained set over which the solver operates.

Theorem 5.1. *Let $L_0, L_1, \varepsilon, \delta > 0$, $d, n \in \mathbb{N}$. Let $w \mapsto f(w; x)$ be an L_0-Lipschitz L_1-smooth convex function for*

all x. *Let* $R_t = \left(\sqrt{2}\right)^t \|w^*\|, \lambda_t = 2^t\lambda$, $\eta_t = \frac{\log(K_t)}{\lambda_t K_t}$, $T = \left\lfloor \log_2\left(\frac{L_1}{\lambda}\right)\right\rfloor$, $\sigma_t^2 = \frac{64L_0^2K_t^2\log(1/\delta)}{n^2\varepsilon^2}$, *and* $\mathcal{S}_t(\{w_k\}_k) = \frac{1}{\sum_{k=1}^{K_t}(1-\eta_t\lambda_t)^{-k}}\sum_{k=1}^{K_t}(1-\eta_t\lambda_t)^{-k}w_k$.

1. *(Optimal rate) Algorithm 3 run with* NoisyGD *(Algorithm 7 in Appendix D) as the PrivateSubRoutine with above parameter settings and* $\lambda = \frac{L_0^2}{L_1\|w^*\|}\min\left(\frac{1}{n}, \frac{d}{n^2\varepsilon^2}\right)$ *and* $K_t = \max\left(\frac{L_1+\lambda_t}{\lambda_t}\log\left(\frac{L_1+\lambda_t}{\lambda_t}\right), \frac{n^2\varepsilon^2\left(L_0^2\lambda+L_1^{3/2}\right)}{T^2\lambda dL_0^2\log(1/\delta)}\right)$ *satisfies* (ε,δ)*-DP, and given a dataset* S *of* n *i.i.d. samples from* $\mathcal{D}$*, outputs* $\bar{w}$ *such that*

$$\mathbb{E}\|\nabla F(\bar{w};\mathcal{D})\| = \tilde{O}\left(\frac{L_0}{\sqrt{n}} + \frac{L_0\sqrt{d}}{n\varepsilon}\right).$$

Furthermore, the above rate is tight up to poly-logarithmic factors.

2. *(Linear time rate) Algorithm 3 run with* PhasedSGD *(Algorithm 5) as the PrivateSubRoutine with with above parameter settings and* $\lambda = \max\left(\frac{L_0^2}{L_1\|w^*\|^2}\min\left(\frac{1}{n}, \frac{d}{n^2\varepsilon^2}\right), \frac{L_1\log(n)}{n}\right)$ *and* $K_t = \lfloor\frac{n}{T}\rfloor$ *satisfies* (ε,δ)*-DP and given a dataset* S *of* n *i.i.d. samples from* $\mathcal{D}$*, in linear time, outputs* $\bar{w}$ *with*

$$\mathbb{E}\|\nabla F(\bar{w};\mathcal{D})\| = \tilde{O}\left(\frac{L_0}{\sqrt{n}} + \frac{L_0\sqrt{d}}{n\varepsilon} + \frac{L_1\|w^*\|}{\sqrt{n}}\right).$$

The proof of the above result is deferred to Appendix D. For the tightness of the rate, the necessity of the second term $\frac{L_0\sqrt{d}}{n\varepsilon}$ is due to our DP empirical stationarity lower bound, Theorem 4.3. For the first "non-private" term $\frac{L_0}{\sqrt{n}}$, even though (Foster et al., 2019) proved a sample complexity lower bound, their instance is not Lipschitz and has $d = \Omega(n\log(n))$, hence not applicable. To remedy this, we give a new lower bound construction with a Lispchitz function in $d = 1$, Theorem A.2 in Appendix A. The polylog dependence on L_1 and $\|w^*\|$ in the upper bounds, is consistent with the non-private sample complexity in (Foster et al., 2019).

The second result is a linear time method which has an additional $L_1\|w^*\|/\sqrt{n}$ term. Firstly, if the smoothness parameter is *small enough*, then there is no overhead; this small-enough smoothness is precisely the regime in which we have linear time methods with optimal rates for smooth DP-SCO (Feldman et al., 2020). More importantly, (Foster et al., 2019) showed that even in the non-private setting, a polynomial dependence on $L_1\|w^*\|$ is necessary in the stochastic oracle model. However, the optimal non-private term, shown in (Foster et al., 2019), is $L_1\|w^*\|/n^2$, achieved by accelerated methods. Improving this dependency, if possible, is an interesting direction for future work.

6. Generalized Linear Models

In this section, we assume that the loss function is a generalized linear model (GLM), $f(w;(x,y)) = \phi_y(\langle w,x\rangle)$. Also, assume the norm of data points x are bounded by $\|\mathcal{X}\|$ and the function $\phi_y : \mathbb{R}\to\mathbb{R}$ is L_0-Lipschitz and L_1-smooth for all y. Furthermore, let `rank` denote the rank of design matrix $X \in \mathbb{R}^{n\times d}$.

Algorithm 4 JL method

Input: Dataset S, function $(z,y)\mapsto\phi_y(z)$, Algorithm $\mathcal{A}$, JL matrix $\Phi\in\mathbb{R}^{k\times d}$, L_0, L_1, $\|\mathcal{X}\|$

1: $\tilde{w} = \mathcal{A}((z,y)\mapsto\phi_y(z), \{(\Phi x_i, y_i)\}_{i=1}^n, 2L_0\|\mathcal{X}\|, 2L_1\|\mathcal{X}\|^2, \varepsilon, \delta/2)$

Output: $\bar{w} = \Phi^\top\tilde{w}$

Algorithm 4 is a generic method which converts *any* for smooth Lipschitz losses with an empirical stationarity guarantee to get dimension-independent rates on population stationarity for smooth Lipschitz GLMs. This algorithm is the JL method from (Arora et al., 2022) used therein to give excess risk bounds for convex GLM. We note that while the JL method there is limited to the Noisy GD method, ours is a black-box reduction. Furthermore, unlike (Arora et al., 2022), we show that the JL method gives finer rank based guarantees by leveraging the fact it acts as an oblivious approximate subspace embedding (see Definition E.1 in Appendix E).

Theorem 6.1. *Let* $\mathcal{A}$ *be an* (ε,δ)*-DP algorithm which when run on a* L_1*-smooth* L_0*-Lipschitz function on a dataset* $S = \{(x_i,y_i)\}_{i=1}^n$ *where* $x_i\in\mathcal{X}\subseteq\mathbb{R}^d$*, guarantees* $\mathbb{E}[\|\nabla F(\mathcal{A}(S);S)\|] \le g(d,n,L_1,L_0,\varepsilon,\delta)$ *and* $\|\mathcal{A}(S)\| \le poly(n,d,L_0,L_1)$ *with probability at least* $1-\frac{1}{\sqrt{n}}$*. Then, Algorithm 4 run with*

$$k = \left\lceil \min\left(\arg\min_{j\in\mathbb{N}}\left(g(j,n,2L_0\|\mathcal{X}\|,2L_1\|\mathcal{X}\|^2,\varepsilon,\delta/2) + \frac{L_0\|\mathcal{X}\|\log(n)}{\sqrt{j}}\right), \mathsf{rank}\log\left(\frac{2n}{\delta}\right)\right)\right\rceil$$

on a L_0*-Lipschitz,* L_1*-smooth GLM loss, is* (ε,δ)*-DP. Furthermore, given a dataset of* n *i.i.d samples from* $\mathcal{D}$*, its output* $\bar{w}$ *has* $\mathbb{E}[\|\nabla F(\bar{w};\mathcal{D})\|]$ *bounded as*

$$\tilde{O}\left(\frac{L_0\|\mathcal{X}\|}{\sqrt{n}} + g(k,n,2L_0\|\mathcal{X}\|,2L_1\|\mathcal{X}\|^2,\varepsilon,\delta/2)\right)$$

The expression for k above comes from the subspace embedding property of JL, and from balancing the dimension of the embedding with respect to the error of $\mathcal{A}$ and the approximation error of the JL embedding. The proof is based on the properties of JL matrices: oblivious subspace embedding and preservation of norms, together with a new

uniform convergence result for gradients of Lipschitz GLMs. The full proof is deferred to Appendix E.

Below, we instantiate the above with our proposed algorithms.

Corollary 6.2. *Under the assumptions of Theorem 6.1, Algorithm 4 run with $\mathcal{A}$ as*

1. *Private Spiderboost (Alg. 2) yields* $\|\nabla F(\bar{w};\mathcal{D})\| = \tilde{O}\left(\frac{1}{\sqrt{n}} + \min\left(\left(\frac{\sqrt{\mathrm{rank}}}{n\varepsilon}\right)^{2/3}, \frac{1}{(n\varepsilon)^{2/5}}\right)\right)$.
2. *Algorithm 3 with NoisyGD as PrivateSubRoutine, under the additional assumption that $w \mapsto f(w;(x,y))$ is convex for all x, y, yields* $\|\nabla F(\bar{w};\mathcal{D})\| = \tilde{O}\left(\frac{1}{\sqrt{n}} + \min\left(\frac{\sqrt{\mathrm{rank}}}{n\varepsilon}, \frac{1}{\sqrt{n\varepsilon}}\right)\right)$.

We remark that the above technique also gives bounds on empirical stationarity. In particular, the first term $\frac{1}{\sqrt{n}}$, in the above guarantees, is the uniform convergence bound and the second term is the bound on empirical stationarity.

Acknowledgements

RA and EU are supported, in part, by NSF BIGDATA award IIS-1838139 and NSF CAREER award IIS-1943251. RB's and MM's research is supported by NSF CAREER Award 2144532 and NSF Award AF-1908281. CG and TG's research was partially supported by INRIA Associate Teams project, FONDECYT 1210362 grant, ANID Anillo ACT210005 grant, and National Center for Artificial Intelligence CENIA FB210017, Basal ANID.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *23rd ACM Conference on Computer and Communications Security*, CCS '16, pp. 308–318, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341394. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

Allen-Zhu, Z. How to make the gradients small stochastically: Even faster convex and nonconvex sgd. *Advances in Neural Information Processing Systems*, 31, 2018.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization, 2019.

Arora, R., Bassily, R., Guzmán, C., Menart, M., and Ullah, E. Differentially private generalized linear models revisited. *arXiv preprint arXiv:2205.03014*, 2022.

Asi, H., Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: Optimal rates in l1 geometry. In *International Conference on Machine Learning*, pp. 393–403. PMLR, 2021.

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

Bassily, R., Feldman, V., Talwar, K., and Guha Thakurta, A. Private stochastic convex optimization with optimal rates. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/3bd8fdb090f1f5eb66a00c84dbc5ad51-Paper.pdf.

Bassily, R., Guzmán, C., and Menart, M. Differentially private stochastic optimization: New results in convex and non-convex settings. *Advances in Neural Information Processing Systems*, 34, 2021a.

Bassily, R., Guzman, C., and Nandi, A. Non-euclidean differentially private stochastic convex optimization. In Belkin, M. and Kpotufe, S. (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 474–499. PMLR, 15–19 Aug 2021b. URL https://proceedings.mlr.press/v134/bassily21a.html.

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

Bun, M., Dwork, C., Rothblum, G. N., and Steinke, T. Composable and versatile privacy via truncated cdp. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, pp. 74–86, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355599. doi: 10.1145/3188745.3188946. URL https://doi.org/10.1145/3188745.3188946.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. "convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 654–663. JMLR.org, 2017.

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

Cohen, M. B. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*, pp. 278–287. SIAM, 2016.

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex sgd. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/b8002139cdde66b87638f7f91d169d96-Paper.pdf.

Diakonikolas, J. and Guzmán, C. Complementary composite minimization, small gradients in general norms, and applications, 2023.

Duchi, J. Lecture notes for statistics 311/electrical engineering 377. *URL: https://stanford.edu/class/stats311/Lectures/full_notes. pdf. Last visited on*, 2:23, 2016.

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

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/1543843a4723ed2ab08e18053ae6dc5b-Paper.pdf.

Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 439–449, 2020.

Foster, D. J., Sekhari, A., and Sridharan, K. Uniform convergence of gradients for non-convex learning and optimization. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/59ab3ba90ae4b4ab84fe69de7b8e3f5f-Paper.pdf.

Foster, D. J., Sekhari, A., Shamir, O., Srebro, N., Sridharan, K., and Woodworth, B. The complexity of making the gradient small in stochastic convex optimization. In *Conference on Learning Theory*, pp. 1319–1345. PMLR, 2019.

Ge, R., Lee, J. D., and Ma, T. Matrix completion has no spurious local minimum. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/7fb8ceb3bd59c7956b1df66729296a4c-Paper.pdf.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Ghadimi, S. and Lan, G. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2016.

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

Jain, P., Kothari, P., and Thakurta, A. Differentially private online learning. In *25th Annual Conference on Learning Theory (COLT)*, pp. 24.1–24.34, 2012.

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

Kamath, G. and Ullman, J. A primer on private statistics. *arXiv preprint arXiv:2005.00010*, 2020.

Kifer, D., Smith, A., and Thakurta, A. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pp. 25–1, 2012.

Kulkarni, J., Lee, Y. T., and Liu, D. Private non-smooth erm and sco in subquadratic steps. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 4053–4064. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/211c1e0b83b9c69fa9c4bdede203c1e3-Paper.pdf.

Lan, G. *First-order and stochastic optimization methods for machine learning*. Springer, 2020.

Ma, C., Wang, K., Chi, Y., and Chen, Y. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix

completion. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3345–3354. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ma18c.html.

Nemirovsky, A. S. and Yudin, D. B. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

Nesterov, Y. How to make the gradients small. *Optima. Mathematical Optimization Society Newsletter*, (88):10–11, 2012.

Nesterov, Y. and Polyak, B. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108:177–205, 2006.

Rudelson, M. and Vershynin, R. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pp. 1576–1602. World Scientific, 2010.

Song, S., Steinke, T., Thakkar, O., and Thakurta, A. Evading the curse of dimensionality in unconstrained private glms. In *International Conference on Artificial Intelligence and Statistics*, pp. 2638–2646. PMLR, 2021.

Steinke, T. and Ullman, J. Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality*, 7, 01 2015. doi: 10.29012/jpc.v7i2.648.

Sun, J., Qu, Q., and Wright, J. A geometric analysis of phase retrieval. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 2379–2383, 2016. doi: 10.1109/ISIT.2016.7541725.

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

Talwar, K., Thakurta, A., and Zhang, L. Nearly optimal private lasso. In *NIPS*, 2015.

Tran, H. and Cutkosky, A. Momentum aggregation for private non-convex erm. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022. URL https://openreview.net/pdf?id=x56v-UN7BjD.

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

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

Wang, D., Chen, C., and Xu, J. Differentially private empirical risk minimization with non-convex loss functions. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6526–6535. PMLR, 09–15 Jun 2019a. URL https://proceedings.mlr.press/v97/wang19c.html.

Wang, L., Jayaraman, B., Evans, D., and Gu, Q. Efficient privacy-preserving nonconvex optimization. *CoRR*, abs/1910.13659, 2019b. URL http://arxiv.org/abs/1910.13659.

Wang, Z., Ji, K., Zhou, Y., Liang, Y., and Tarokh, V. Spiderboost and momentum: Faster variance reduction algorithms. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019c. URL https://proceedings.neurips.cc/paper/2019/file/512c5cad6c37edb98ae91c8a76c3a291-Paper.pdf.

Zhang, J., Zheng, K., Mou, W., and Wang, L. Efficient private erm for smooth objectives. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pp. 3922–3928. AAAI Press, 2017. ISBN 9780999241103.

Zhang, Q., Ma, J., Lou, J., and Xiong, L. Private stochastic non-convex optimization with improved utility rates. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 3370–3376, 2021.

Zhou, Y., Chen, X., Hong, M., Wu, Z. S., and Banerjee, A. Private stochastic non-convex optimization: Adaptive algorithms and tighter generalization bounds. *CoRR*, abs/2006.13501, 2020. URL https://arxiv.org/abs/2006.13501.

A. Lower bounds

A.1. Missing details from DP Empirical Stationarity Lower Bound

Proof of Theorem 4.3. For any $r > 0$, let $\mathcal{W}_r$ denote the ball of radius r centered at the origin. Let $B = \frac{L_0}{L_1}$. Consider the loss function:

$$f(w; x) = \begin{cases} \frac{L_1}{2} \|w - x\|^2 & \text{if } \|w - x\| \leq B \\ L_0 \|w - x\| - \frac{L_0^2}{2L_1} & \text{otherwise} \end{cases}$$

The function $f(w; x)$ is convex, L_1-smooth and L_0-Lispchitz in $\mathbb{R}^d$. We restrict to datasets $S = \{x_i\}_{i=1}^n$ where $x_i \in \mathcal{W}_{B/4}$ for all i, and let $F(w; S) = \frac{1}{n} \sum_{i=1}^n f(w; x_i)$ be the empirical risk on S. The unconstrained minimizer of $F(w; S)$ is $w^* = \frac{1}{n} \sum_{i=1}^n x_i$ which lies in $\mathcal{W}_{B/4}$.

For any $w \in \mathcal{W}_{3B/4}$, w lies in the quadratic region around all data points. Hence, from L_1-strong convexity of $w \mapsto F(w; S)$ on $\mathcal{W}_{3B/4}$, we have that whenever $\bar{w} \in \mathcal{W}_{3B/4}$,

$$\|\nabla F(\bar{w}; S)\| \|\bar{w} - w^*\| \geq \langle \nabla F(\bar{w}; S), w^* - \bar{w} \rangle \geq F(\bar{w}; S) - F(w^*; S) \geq \frac{L_1}{2} \|\bar{w} - w^*\|^2 .$$

Let E be the event that $\bar{w} \in \mathcal{W}_{3B/4}$ and let $\mathbb{E}_E$ denote the conditional expectation (conditioned on event E) operator. Then,

$$\mathbb{E}_E \|\nabla F(\bar{w}; S)\| \geq \frac{L_1}{2} \mathbb{E} \|\bar{w} - w^*\| \geq \frac{L_1}{2} \Omega \left(\left(\frac{L_0}{4L_1} \right) \min \left(1, \frac{\sqrt{d \log(1/\delta)}}{n\varepsilon} \right) \right).$$

where the last inequality follows from known lower bounds for DP mean estimation (Steinke & Ullman, 2015; Kamath & Ullman, 2020). We remark that the lower bound in the referenced work is for algorithms which produce outputs in the ball of the same radius as the dataset, i.e. $\mathcal{W}_{B/4}$. However, a simple post-processing argument shows that the same lower bound applies to algorithms which produce output in $\mathcal{W}_{3B/4}$. Specifically, assuming the contrary, we simply project the output in $\mathcal{W}_{3B/4}$ to $\mathcal{W}_{B/4}$: privacy is preserved by post-processing and the distance to the mean cannot increase by the non-expansiveness property of projection to convex sets, hence a contradiction. This gives us,

$$\mathbb{E}_E \left[\|\nabla F(\bar{w}; S)\| \right] \geq \Omega \left(L_0 \min \left(1, \frac{\sqrt{d \log(1/\delta)}}{n\varepsilon} \right) \right)$$

Let $\tilde{\mathcal{W}} = \{w : \|w - w^*\| \leq B/2\}$. Since $\tilde{\mathcal{W}} \subseteq \mathcal{W}_{3B/4}$, we have that the above conditional lower bound applies for $\bar{w} \in \tilde{\mathcal{W}}$ as well. We now consider $\bar{w} \notin \tilde{\mathcal{W}}$. Let w' be *any* point on the boundary of $\tilde{\mathcal{W}}$, denoted as $\partial\mathcal{W}$. Note that w' lies in the region where, for any data point, the corresponding loss is a quadratic function. Hence, by direct computation, $\nabla F(w'; S) = L_1 (w' - w^*)$. Therefore,

$$\langle \nabla F(w'), w' - w^* \rangle = L_1 \|w' - w^*\|^2 = \frac{L_1 B^2}{4}.$$

We now apply Lemma A.1 which gives us,

$$\mathbb{E}_{E^c} \|\nabla F(\bar{w}; S)\| \geq \frac{L_1 B^2}{4} \cdot \frac{2}{B} = \frac{L_0}{2},$$

where E^c denotes the complement set of E. We combine the above bounds using the law of total expectation as follows,

$$\begin{aligned} \mathbb{E}[\|\nabla F(\bar{w}; S)\|] &= \mathbb{E}_E[\|\nabla F(\bar{w}; S)\|] \mathbb{P}\{\bar{w} \in E\} + \mathbb{E}_{E^c}[\|\nabla F(\bar{w}; S)\|] \mathbb{P}\{\bar{w} \in E^c\} \\ &= \Omega\Big(L_0 \min \Big\{ 1, \frac{\sqrt{d \log(1/\delta)}}{n\varepsilon} \Big\} \Big) \mathbb{P}(\bar{w} \in E) + \Omega(L_0) \mathbb{P}(\bar{w} \in E^c) \\ &= \Omega\Big(L_0 \min \Big\{ 1, \frac{\sqrt{d \log(1/\delta)}}{n\varepsilon} \Big\} \Big). \end{aligned}$$

This completes the proof. □

Lemma A.1. *Let $G, R \geq 0, d \in \mathbb{N}$. Let $\mathcal{W}_R(w_0)$ denote the Euclidean ball around w_0 of radius R and let $\partial\mathcal{W}_R(w_0)$ denote its boundary. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable convex function. Suppose $w_0 \in \mathbb{R}^d$ is such that for every $v \in \partial\mathcal{W}_R(w_0)$, $\langle \nabla f(v), v - w_0 \rangle \geq G$, then for any $w \notin \mathcal{W}_R(w_0)$, we have $\|\nabla f(w)\| \geq \frac{G}{R}$.*

Proof. For a unit vector $u \in \mathbb{R}^d$, define directional directive $f'_u(w) = \langle \nabla f(w), u \rangle$. We first show that for any $u \in \mathbb{R}^d : \|u\| = 1$ and any $w' \in \mathbb{R}^d$, the function $f'_u(w' + ru)$ is non-decreasing in $r \in \mathbb{R}_+$. This simply follows from monotonicity of gradients since f is convex. In particular, for any $r' > r > 0$, we have

$$\begin{aligned} f'_u(w' + r'u) - f'_u(w' + ru) &= \langle \nabla f(w' + r'u) - \nabla f(w' + ru), u \rangle \\ &= \frac{1}{r' - r} \langle \nabla f(w' + r'u) - \nabla f(w' + ru), w' + ru - (w' + ru) \rangle \\ &> 0 \end{aligned}$$

We now prove the claim in the lemma statement. Let $w \notin \partial W_R$ and define $u = \frac{w - w_0}{\|w - w_0\|}$. Then from Cauchy-Schwarz inequality and the above monotonicity property, we have,

$$\begin{aligned} \|\nabla f(w)\| \geq \langle \nabla f(w), u \rangle = f'_u(w) \geq f'_u(w_0 + Ru) = \langle \nabla f(w_0 + Ru), u \rangle \\ = \frac{1}{R} \langle \nabla f(v), v - w_0 \rangle \geq \frac{G}{R} \end{aligned}$$

which finishes the proof. □

A.2. Non-private Sample Complexity Lower Bound

Theorem A.2. *For any $L_0, L_1, n, d \in \mathbb{N}$, there exists a distribution $\mathcal{D}$ over some set $\mathcal{X}$ and a L_0-Lipschitz, L_1-smooth (convex) loss function $w \mapsto f(w; x)$ such that given n i.i.d samples from $\mathcal{D}$, the output $\bar{w}$ of any algorithm satisfies,*

$$\mathbb{E} \|\nabla F(\bar{w}; \mathcal{D})\| = \Omega\left(\frac{L_0}{\sqrt{n}}\right)$$

Proof. We construct a hard instance in $d = 1$ dimension. Let $p \in [0, 1]$ be a parameter to be set later and let $v \in \{-1, 1\}$ be chosen by an adversary. Let the data domain $\mathcal{X} = \{-1, 1\}$ and consider the distribution $\mathcal{D}$ on $\mathcal{X}$ as follows:

$$x = \begin{cases} 1 & \text{with probability } \frac{1+vp}{2} \\ -1 & \text{with probability } \frac{1-vp}{2} \end{cases}$$

Note that $\mathbb{E}[x] = vp$. Consider the loss function $f(w; x)$ as

$$f(w; x) = \frac{L_0}{2} wx + \frac{L_1}{2} \Delta(w)$$

where Δ is the Huber regularization function, defined as,

$$\Delta(w) = \begin{cases} |w|^2 & \text{if } |w| \leq \frac{L_0}{2L_1} \\ \frac{L_0 |w|}{L_1} - \frac{L_0^2}{4L_1^2} & \text{otherwise} \end{cases}$$

Note that the loss function $w \mapsto f(w; x)$ is convex, L_0-Lipschitz and L_1-smooth in $\mathbb{R}^d$, for all x. The population risk function is,

$$F(w; \mathcal{D}) = \frac{L_0}{2} wpv + \frac{L_1}{2} \Delta(w)$$

Let $\bar{w}$ be output some algorithm given n i.i.d. samples from $\mathcal{D}$. Consider two cases:

Case 1: $|\bar{w}| > \frac{L_0}{2L_1}$**:** The gradient norm in this case is

$$\begin{aligned}
|\nabla F(\bar{w};\mathcal{D})|^2 &= \left|\frac{L_0}{2}vp + \frac{L_0\bar{w}}{2|\bar{w}|}\right|^2 \\
&= \frac{L_0^2p^2}{4} + \frac{L_0^2}{4} + \frac{L_0^2}{2|\bar{w}|}vp\bar{w} \\
&\geq \frac{L_0^2}{4} - \frac{L_0^2}{2}p \\
&= \frac{L_0^2}{4} - \frac{L_0^2}{8\sqrt{n}} \\
&\geq \frac{L_0^2}{8}
\end{aligned}$$

where the first inequality follows since $v\frac{\bar{w}}{|\bar{w}|} \geq -1$, the third equality follows by setting $p = \frac{1}{\sqrt{16n}}$ and the second inequality follows since $n \geq 1$. We therefore have that $\mathbb{E}\,|\nabla F(\bar{w};\mathcal{D})| \geq \frac{L_0}{2\sqrt{2}}$.

Case 2: $|\bar{w}| \leq \frac{L_0}{2L_1}$**:** In this case, the gradient norm is,

$$|\nabla F(\bar{w};\mathcal{D})|^2 = \left|\frac{L_0}{2}vp + L_1\bar{w}\right|^2$$

Suppose there exists an algorithm with output $\bar{w}$, which, with n samples guarantees that $\mathbb{E}\,|\nabla F(\bar{w};\mathcal{D})| < o\left(\frac{L_0}{\sqrt{n}}\right)$. Then from Markov's inequality, with probability at least 0.9, we have that $|\nabla F(\bar{w};\mathcal{D})|^2 < o\left(\frac{L_0^2}{n}\right)$. Let $\tilde{w} = -\frac{2L_1\bar{w}}{L_0}$, then we have that with probability at least 0.9,

$$|\nabla F(\bar{w};\mathcal{D})|^2 \leq o\left(\frac{L_0^2}{n}\right) \iff |vp - \tilde{w}|^2 < o\left(\frac{1}{n}\right)$$

This contradicts the well-known bias estimation lower bounds, with $p = \frac{1}{\sqrt{16n}}$, using Le Cam's method ((Duchi, 2016), Example 7.7), hence $\mathbb{E}\,|\nabla F(\bar{w};\mathcal{D})| \geq \Omega\left(\frac{L_0}{\sqrt{n}}\right)$. Combining the two cases finishes the proof. □

B. Missing Results for Empirical Stationary Points

B.1. Private Spiderboost

The following lemma largely follows from the analysis in (Wang et al., 2019c). We present a full proof below for completeness.

Lemma B.1. *Let the conditions of Lemma 4.1 be satisfied. Let $\eta \leq \frac{1}{2L_1}$ and $q \leq O\left(\frac{1}{\tau_2^2\eta^2}\right)$. Then the output of Private SpiderBoost, $\bar{w}$ satisfies*

$$\mathbb{E}\left[\|\nabla F(\bar{w};S)\|\right] = O\left(\sqrt{\frac{F_0}{\eta T}} + \tau_1\right). \tag{1}$$

Proof. In the following, for any $t \in [T]$, let $s_t = \left\lfloor \frac{t}{q} \right\rfloor q$ (i.e. the index corresponding to the start of the phase containing iteration t).

By a standard analysis for smooth functions we have (recalling that ∇_t is an unbiased estimate of $\nabla F(w_t;S)$ for any $t \in [T]$)

$$F(w_{t+1};S) \leq F(w_t;S) + \frac{\eta}{2}\|\nabla F(w_t;S) - \nabla_t\|^2 - \left(\frac{\eta}{2} - \frac{L_1\eta^2}{2}\right)\|\nabla_t\|^2.$$

Taking expectation we have the following manipulation using the update rule of Algorithm 2

$$\begin{aligned}\mathbb{E}\left[F(w_{t+1};S)-F(w_t;S)\right] &\leq \frac{\eta}{2}\mathbb{E}\left[\|\nabla F(w_t;S)-\nabla_t\|^2\right]-\left(\frac{\eta}{2}-\frac{L_1\eta^2}{2}\right)\mathbb{E}\left[\|\nabla_t\|^2\right]\\ &\leq \frac{\eta\tau_2^2}{2}\sum_{k=s_t+1}^{t}\mathbb{E}\left[\|w_{k+1}-w_k\|^2\right]+\frac{\eta}{2}\mathbb{E}\left[\|\nabla_{s_t}-F(w_{s_t};S)\|^2\right]\\ &\quad-\left(\frac{\eta}{2}-\frac{L_1\eta^2}{2}\right)\mathbb{E}\left[\|\nabla_t\|^2\right]\\ &\leq \frac{\eta^3\tau_2^2}{2}\sum_{k=s_t+1}^{t}\mathbb{E}\left[\|\nabla_k\|^2\right]+\frac{\eta\tau_1^2}{2}-\left(\frac{\eta}{2}-\frac{L_1\eta^2}{2}\right)\mathbb{E}\left[\|\nabla_t\|^2\right],\end{aligned}$$

where the second inequality follows from Lemma 4.1 and the last inequality follows from the update rule. Note that if $t=s_t$ the sum is empty. Summing over a given phase we have

$$\begin{aligned}\mathbb{E}\left[F(w_{t+1};S)-F(w_{s_t};S)\right] &\leq \frac{\eta^3\tau_2^2}{2}\sum_{k=s_t}^{t}\sum_{j=s_t+1}^{k}\mathbb{E}\left[\|\nabla_j\|^2\right]+\sum_{k=s_t}^{t}\left[\tfrac{\eta\tau_1^2}{2}-\left(\tfrac{\eta}{2}-\tfrac{L_1\eta^2}{2}\right)\mathbb{E}\left[\|\nabla_k\|^2\right]\right]\\ &\leq \frac{\eta^3\tau_2^2 q}{2}\sum_{k=s_t}^{t}\mathbb{E}\left[\|\nabla_k\|^2\right]+\sum_{k=s_t}^{t}\left[\tfrac{\eta\tau_1^2}{2}-\left(\tfrac{\eta}{2}-\tfrac{L_1\eta^2}{2}\right)\mathbb{E}\left[\|\nabla_k\|^2\right]\right]\\ &= -\sum_{k=s_t}^{t}\Bigg[\underbrace{\left(\frac{\eta}{2}-\frac{L_1\eta^2}{2}-\frac{\eta^3\tau_2^2 q}{2}\right)}_{A}\mathbb{E}\left[\|\nabla_k\|^2\right]-\frac{\eta\tau_1^2}{2}\Bigg],\end{aligned}\tag{2}$$

where the second inequality comes from the fact that each gradient appears at most q times in the sum. We now sum over all phases. Let $P=\{p_0,p_1,...,\}=\left\{0,q,2q,...,\left\lfloor\frac{T-1}{q}\right\rfloor q,T\right\}$. We have

$$\begin{aligned}\mathbb{E}\left[F(w_T;S)-F(w_0;S)\right] &\leq \sum_{i=1}^{|P|}\mathbb{E}\left[F(w_{p_i};S)-F(w_{p_{i-1}};S)\right]\\ &\leq -\sum_{t=0}^{T}A\,\mathbb{E}\left[\|\nabla_k\|^2\right]+\frac{T\eta\tau_1^2}{2}.\end{aligned}$$

Rearranging the above yields

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\left[\|\nabla_k\|^2\right]\leq\frac{F_0}{TA}+\frac{\eta\tau_1^2}{2A}.\tag{3}$$

Now let i^* denote the index of $\bar{w}$ selected by the algorithm. Note that

$$\mathbb{E}\left[\|\nabla F(w_{i^*};S)\|^2\right]\leq 2\mathbb{E}\left[\|\nabla F(w_{i^*};S)-\nabla_{i^*}\|^2\right]+2\mathbb{E}\left[\|\nabla_{i^*}\|^2\right].\tag{4}$$

The second term above can be bounded via inequality (3). To bound the first term we have by Lemma 4.1 that

$$\begin{aligned}\mathbb{E}\left[\|\nabla_{i^*} - \nabla F(w_{i^*}; S)\|^2\right] &\leq \tau_2^2 \sum_{k=s_{t^*}+1}^{t^*} \mathbb{E}\left[\|w_k - w_{k-1}\|^2\right] + \tau_1^2 \\ &= \eta^2\tau_2^2 \sum_{k=s_{t^*}+1}^{t^*} \mathbb{E}\left[\|\nabla_k\|^2\right] + \tau_1^2 \\ &\leq \frac{q\eta^2\tau_2^2}{T}\sum_{k=0}^{T}\mathbb{E}\left[\|\nabla_k\|^2\right] + \tau_1^2 \\ &\leq \frac{\tau_2^2\eta^2 q F_0}{TA} + \frac{\eta^3 q\tau_2^2}{2A}\tau_1^2 + \tau_1^2,\end{aligned}$$

where the last inequality comes from inequality (3) and the expectation over i^*. Plugging into inequality (4) one can obtain

$$\mathbb{E}\left[\|\nabla F(w_{i^*}; S)\|^2\right] \leq \frac{2F_0}{TA}(1 + \tau_2^2\eta^2 q) + \left(\frac{\eta}{A} + 2 + \frac{\tau_2^2\eta^3 q}{A}\right)\tau_1^2. \tag{5}$$

Now recall $A = \frac{\eta}{2} - \frac{L_1\eta^2}{2} - \frac{\eta^3\tau_2^2 q}{2}$. Since $q \leq O\left(\frac{1}{\tau_2^2\eta^2}\right)$ and $\eta \leq \frac{1}{2L_1}$ we have $A = \Theta(\eta)$. Thus plugging into inequality (5) and again using the fact that $q \leq O\left(\frac{1}{\tau_2^2\eta^2}\right)$ we have

$$\mathbb{E}\left[\|\nabla F(w_{i^*}; S)\|^2\right] = O\left(\frac{F_0}{T\eta}(1 + \tau_2^2\eta^2 q) + \left(3 + \frac{\tau_2^2\eta^3 q}{A}\right)\tau_1^2\right) = O\left(\frac{F_0}{T\eta} + \tau_1^2\right).$$

The claim then follows from the Jensen inequality. □

For privacy, we will rely on the moments accountant analysis of (Abadi et al., 2016). This roughly gives the same analysis as using privacy amplification via subsampling and the advanced composition theorem, but allows for improvements in log factors. We provide the following theorem implicit in (Abadi et al., 2016) Theorem 1 below. The same result can be obtained using the analysis for (Kulkarni et al., 2021) Theorem 3.1 which uses the truncated central differential privacy guarantees of the Gaussian mechanism (Bun et al., 2018).

Theorem B.2 ((Abadi et al., 2016; Kulkarni et al., 2021)). *Let $\varepsilon, \delta \in (0, 1]$ and c be a universal constant. Let $D \in \mathcal{Y}^n$ be a dataset over some domain $\mathcal{Y}$, and let $h_1, ..., h_T : \mathcal{Y} \mapsto \mathbb{R}^d$ be a series of (possibly adaptive) queries such that for any $y \in \mathcal{Y}$, $t \in [T]$, $\|h_t(y)\|_2 \leq \lambda_t$. Let $\sigma_t = \frac{c\lambda_t\sqrt{\log(1/\delta)}}{\varepsilon}\max\left\{\frac{1}{b}, \frac{\sqrt{T}}{n}\right\}$. Then the algorithm which samples batches of size $B_1, .., B_t$ of size b uniformly at random and outputs $\frac{1}{n}\sum_{y\in B_t} h_t(y) + g_t$ for all $t \in [T]$ where $g_t \sim \mathcal{N}(0, \mathbb{I}\sigma_t^2)$, is (ε, δ)-DP.*

We note that the original statement of the Theorem in (Abadi et al., 2016) requires $\sigma_t \geq \frac{c\lambda_t\sqrt{T\log(1/\delta)}}{n\varepsilon}$ and $T \geq \frac{n^2\varepsilon}{b^2}$ (or $T \geq \frac{n^2}{b^2}$ so long as $\varepsilon \leq 1$). However, in the case where $T \leq \frac{n^2}{b^2}$, one can simply consider the meta algorithm that does run $T' = \frac{n^2}{b^2}$ steps and only outputs the first T results. This algorithm is at least as private as the algorithm which outputs every result, and under the setting T' the scale of noise is $\frac{8\lambda_t\sqrt{\log(1/\delta)}}{b\varepsilon}$.

We can now prove the main result for Private Spiderboost, restated below. We note that the setting of b_2 given below will always be less than n under required conditions. More details are provided in the proof below.

Theorem B.3 (Private Spiderboost). *Let $n \geq \max\left\{\frac{(L_0\varepsilon)^2}{F_0L_1d\log(1/\delta)}, \frac{\sqrt{d}\max\{1, \sqrt{L_1F_0}/L_0\}}{\varepsilon}\right\}$. Private Spiderboost run with parameter settings $\eta = \frac{1}{2L_1}$, $b_1 = n$, $b_2 = \left\lfloor\max\left\{\left(\frac{L_0n\varepsilon}{\sqrt{F_0L_1d\log(1/\delta)}}\right)^{2/3}, \frac{(L_0nd\log(1/\delta))^{1/3}}{(L_1F_0)^{1/6}\varepsilon^{2/3}}\right\}\right\rfloor$, $T = \left\lfloor\max\left\{\left(\frac{(F_0L_1)^{1/4}n\varepsilon}{\sqrt{L_0d\log(1/\delta)}}\right)^{4/3}, \frac{n\varepsilon}{\sqrt{d\log(1/\delta)}}\right\}\right\rfloor$, and $q = \left\lfloor\frac{n^2\varepsilon^2}{L_1^2Td\log(1/\delta)}\right\rfloor$ satisfies*

$$\mathbb{E}\left[\|\nabla F(\tilde{w})\|\right] = O\left(\left(\frac{\sqrt{F_0L_1L_0d\log(1/\delta)}}{n\varepsilon}\right)^{2/3} + \frac{\sqrt{d\log(1/\delta)}L_0}{n\varepsilon}\right)$$

is (ε, δ)*-DP and has oracle complexity* $\tilde{O}\left(\max\left\{\left(\frac{n^{5/3}\varepsilon^{2/3}}{d^{1/3}}\right), \left(\frac{n\varepsilon}{\sqrt{d}}\right)^2\right\}\right)$.

Proof. For privacy, we rely on the moment accountant analysis of the Gaussian mechanism as per Theorem B.2. Note that each gradient estimate computed in line 9 has elements with ℓ_2-norm at most L_0, and this estimate is computed at most $\frac{T}{q}$ times. Similarly, for a gradient variation at step t in line 13 we have norm bound $L_1 \|w_t - w_{t-1}\|$, and have that at most T such estimates are computed. As such, the scale of noise in both cases ensures the overall algorithm is (ε, δ)-DP by Theorem B.2.

We now prove the convergence result. To simplify notation in the following, we define $\bar{\alpha} = \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}$. If $b_1 = n$ (full batch gradient), the conditions of Lemma 4.1 are satisfied with $\tau_1^2 = O\left(\frac{L_0^2 T\bar{\alpha}^2}{q}\right)$ and $\tau_2^2 = O\left(\frac{L_1^2}{b_2} + L_1^2 T\bar{\alpha}^2\right)$ and some setting of q so long as $T \geq q\frac{n^2}{b_1^2} = q$ and $T \geq \frac{n^2}{b_2^2}$. Further, if $b_2 \geq \frac{1}{T\bar{\alpha}^2}$ then $\tau_2^2 = O\left(L_1^2 T\bar{\alpha}^2\right)$. Thus the condition on q in Lemma B.1 is satisfied with $q = \frac{L_1^2}{\tau_2^2} = \frac{1}{T\bar{\alpha}^2}$ since $\eta = \frac{1}{2L_1}$

Plugging into Eqn. (1) we obtain

$$\begin{aligned}\mathbb{E}\left[\|\nabla F(\tilde{w})\|\right] &= O\left(\sqrt{\frac{F_0 L_1}{T}} + \frac{L_0\sqrt{T}\bar{\alpha}}{\sqrt{q}}\right) \\ &= O\left(\sqrt{\frac{F_0 L_1}{T}} + L_0 T\bar{\alpha}^2\right). \end{aligned} \tag{6}$$

We now consider the setting of T. Since $q = \frac{1}{T\bar{\alpha}^2}$, it suffices to set $T \geq \frac{1}{\bar{\alpha}}$ to ensure $T \geq q$. We now set $T = \max\left\{\left(\frac{(L_1F_0)^{1/4}}{\sqrt{L_0}\bar{\alpha}}\right)^{4/3}, \frac{1}{\bar{\alpha}}\right\}$. Using Eqn. (6) above we have

$$\mathbb{E}\left[\|\nabla F(\tilde{w})\|\right] = O\left(\left(\sqrt{F_0 L_1 L_0}\bar{\alpha}\right)^{2/3} + L_0\bar{\alpha}\right).$$

The claimed rate now follows if there exists a valid setting for b_2 satisfying the previously stated conditions. The restrictions on the batch size implied by T imply we need $b_2 \geq \frac{n}{\sqrt{T}}$ and thus it suffices to have $b_2 \geq \frac{L_0^{1/3} n\bar{\alpha}^{2/3}}{(L_1F_0)^{1/6}}$ to satisfy this condition since $T \geq \left(\frac{(L_1F_0)^{1/4}}{\sqrt{L_0}\bar{\alpha}}\right)^{4/3}$. We recall that for the setting of q to be valid we also require $b_2 \geq \frac{1}{T\bar{\alpha}^2}$ and because $T \geq \left(\frac{(L_1F_0)^{1/4}}{\sqrt{L_0}\bar{\alpha}}\right)^{4/3}$ it suffices that $b_2 \geq \left(\frac{L_0}{\sqrt{F_0L_1}\bar{\alpha}}\right)^{2/3}$. Thus we need $b_2 = \max\left\{\left(\frac{L_0}{\sqrt{F_0L_1}\bar{\alpha}}\right)^{2/3}, \frac{L_0^{1/3}n\bar{\alpha}^{2/3}}{(L_1F_0)^{1/6}}\right\}$.

Finally, we need $b_2 \leq n$ whenever $q \geq 1$. Note that by the setting of q and T we have $q \leq \left(\frac{L_0}{\sqrt{F_0L_1}\bar{\alpha}}\right)^{2/3}$ and thus $q \geq 1 \implies \left(\frac{\sqrt{L_1F_0}\bar{\alpha}}{L_0}\right) \leq 1$. Under this same condition we have $\frac{L_0^{1/3}n\bar{\alpha}^{2/3}}{(L_1F_0)^{1/6}} \leq n$. We further have $\left(\frac{L_0}{\sqrt{F_0L_1}\bar{\alpha}}\right)^{2/3} \leq n$ under the assumption $n \geq \frac{(L_0\varepsilon)^2}{F_0L_1 d\log(1/\delta)}$ given in the theorem statement. It can also be verified that under the condition on n given in the theorem statement that $q \geq 1$. Thus the parameter settings obtain the claimed rate.

Note the number of gradient computations is bounded by

$$\begin{aligned} O\left(Tb_2 + \frac{Tb_1}{q}\right) &= \tilde{O}\left(\left(\frac{n\varepsilon}{\sqrt{d}}\right)^{4/3}\max\left\{\left(\frac{n\varepsilon}{\sqrt{d}}\right)^{2/3}, \frac{(nd)^{1/3}}{\varepsilon^{2/3}}\right\} + n\left(\frac{n\varepsilon}{\sqrt{d}}\right)^{2/3}\right) \\ &= \tilde{O}\left(\max\left\{\left(\frac{n\varepsilon}{\sqrt{d}}\right)^2, \frac{n^{5/3}\varepsilon^{2/3}}{d^{1/3}}\right\}\right). \end{aligned}$$

□

B.2. Additional Discussion of Rate Improvement Challenges

We here give a more detailed version of the informal discussion in Section 4.2. We want to emphasize that the goal of the following discussion is not to provide a universal lower bound, but rather to inform future research.

Let $\mathcal{L}:\mathbb{R}^d\mapsto\mathbb{R}$ be a loss function. We say the randomized mapping $\mathcal{O}:\mathbb{R}^d\times(\mathbb{R}^d\cup\perp)\mapsto\mathbb{R}^d$, is a (τ_1,τ_2)-accurate oracle for $\mathcal{L}$ if $\forall w,w'\in\mathbb{R}^d$

$$\mathop{\mathbb{E}}_{\mathcal{O}}\left[\mathcal{O}(w,\perp)\right]=\nabla\mathcal{L}(w), \qquad \mathop{\mathbb{E}}_{\mathcal{O}}\left[\mathcal{O}(w,w')\right]=\nabla\mathcal{L}(w)-\nabla\mathcal{L}(w')$$
$$\mathop{\mathbb{E}}_{\mathcal{O}}\left[\left\|\mathcal{O}(w,\perp)-\nabla\mathcal{L}(w)\right\|^2\right]\le\tau_1^2, \qquad \mathop{\mathbb{E}}_{\mathcal{O}}\left[\left\|\mathcal{O}(w,w')\right\|^2\right]\le\tau_2^2\left\|w-w'\right\|^2.$$

In short, $\mathcal{O}$ is an unbiased and accurate gradient/gradient variation oracle for $\mathcal{L}$. Define

$$\mathfrak{m}(G,L_1,\mathcal{L}_0,\tau_1,\tau_2)=\inf_{\mathcal{A}}\sup_{\mathcal{O},\mathcal{L}}\inf\left\{\alpha:\mathbb{E}\left[\|\nabla\mathcal{L}(\mathcal{A}(\mathcal{O},L_1,\mathcal{L}_0,\tau_1,\tau_2)\|\right]\le\alpha\right\},$$

where the supremum is taken over L_1-smooth functions $\mathcal{L}$ satisfying $\mathcal{L}(0)-\arg\min_{w\in\mathbb{R}^d}\{\mathcal{L}(w)\}\le\mathcal{L}_0$, and (τ_1,τ_2)-accurate oracles for $\mathcal{L}$. The infimum is taken over algorithms which make at most G calls to $\mathcal{O}$.

We have the following lower bound on $\mathfrak{m}$ (i.e. a lower bound on the accuracy of optimization algorithms which make at most G queries to the oracle) following from (Arjevani et al., 2019, Theorem 3) and the fact that the oracle model described above is a special case of the multi-query oracles considered by (Arjevani et al., 2019).

Theorem B.4 ((Arjevani et al., 2019)). *Let $G,\mathcal{L}_0,L_1,\tau_1,\tau_2\ge0$ and define $\alpha=\left(\frac{\mathcal{L}_0\tau_2\tau_1}{G}\right)^{1/3}+\frac{\tau_1}{\sqrt{G}}$. If $d=\tilde{\Omega}\left(\left[\frac{\mathcal{L}_0L_1}{\alpha^2}\right]^2\right)$, then $\mathfrak{m}(G,L_1,\mathcal{L}_0,\tau_1,\tau_2)=\Omega\left(\alpha\right)$.*

Now consider $\mathcal{L}$ such that $\mathcal{L}(w)=\frac{1}{n}\sum_{x\in S}\ell(w;x)$ for some L_0-Lipschitz and L_1-smooth loss $\ell:\mathbb{R}^d\times\mathcal{X}\mapsto\mathbb{R}$ and $S\in\mathcal{X}^n$. We are interested in designing some $(\widehat{\tau}_1,\widehat{\tau}_2)$-accurate and differentially private oracle, $\widehat{\mathcal{O}}$, which can then be used by an optimization algorithm, $\mathcal{A}$, to obtain an approximate stationary point $\bar{w}=\mathcal{A}(\widehat{\mathcal{O}},L_1,\mathcal{L}_0,\widehat{\tau}_1,\widehat{\tau}_2)$. Specifically, we want $\widehat{\mathcal{O}}$ to be capable of answering G queries under (ε,δ)-DP. A common method for achieving this is to ensure each query to $\mathcal{O}$ is at least $(\frac{\varepsilon}{\sqrt{G}},\delta)$-DP and use advanced composition (or the more refined moment accountant) analysis. Such a setup encapsulates numerous results in the convex setting (Bassily et al., 2019; Kulkarni et al., 2021), and is even more dominant in non-convex settings (Wang et al., 2017; Zhou et al., 2020; Abadi et al., 2016).

Our key observation is that under such a setup, any increase in the number of oracle calls to G must be met with a proportional increase in the accuracy parameters $(\widehat{\tau}_1,\widehat{\tau}_2)$. Thus, if such an oracle, $\widehat{\mathcal{O}}$ is applied in a black box fashion to a stochastic optimization algorithm $\mathcal{A}$, one can obtain a lower bound on the accuracy of the overall algorithm independent of G.

Specifically, since estimating the gradient and gradient variation can be viewed as mean estimation problems on n vectors, we can use fingerprinting code arguments to lower bound $\widehat{\tau}_1$ and $\widehat{\tau}_2$ (Steinke & Ullman, 2015). In Lemma B.5 below, we prove that any $(\widehat{\tau}_1,\widehat{\tau}_2)$-accurate oracle which ensures that any query is $(\frac{\varepsilon}{\sqrt{G}},\delta)$-DP must have $\widehat{\tau}_1=\Omega\left(\frac{L_0\sqrt{Gd\log(1/\delta)}}{n\varepsilon}\right)$ and $\widehat{\tau}_2=\Omega\left(\frac{L_1\sqrt{Gd\log(1/\delta)}}{n\varepsilon}\right)$. Now, observe that by Theorem B.4, we have

$$\mathfrak{m}(G,L_1,\mathcal{L}_0,\widehat{\tau}_1,\widehat{\tau}_2)=\Omega\left(\left(\frac{\sqrt{F_0L_1L_0}\sqrt{d\log\left(1/\delta\right)}}{n\varepsilon}\right)^{2/3}+\frac{L_0\sqrt{d\log\left(1/\delta\right)}}{n\varepsilon}\right),$$

which matches our upper bound.

We now remark on several ways the above barrier could be circumvented. The first and most obvious possibility is to employ a different privatization method than private oracles. However, this is particularly difficult in the nonconvex setting as existing methods which avoid private gradients (see e.g. (Feldman et al., 2020) for several such methods) rely crucially on stability guarantees arising from convexity. Other possible ways to beat the above rate is by designing a stochastic optimization algorithm which leverages the structure of the noise used in private implementations of the oracle or makes use of additional assumptions to beat the $\Omega\left(\left(\frac{\mathcal{L}_0\tau_2\tau_1}{G}\right)^{1/3}+\frac{\tau_1}{\sqrt{G}}\right)$ non-private lower bound.

Additional Details on Fingerprinting Bound We conclude by giving a concrete construction for the fingerprinting argument mentioned above.

Lemma B.5. *Let $L_0, L_1 \geq 0$, $\varepsilon = O(1)$, $2^{-\Omega(n)} \leq \delta \leq \frac{1}{n^{1+\Omega(1)}}$ and $\sqrt{d\log(1/\delta)}/(n\varepsilon) = O(1)$. Let $\ell, \mathcal{L}, S$ satisfy the assumptions above. Then there exists ℓ, S such that for any oracle, $\mathcal{O}$, which is (τ_1, τ_2)-accurate for $\mathcal{L}$ it holds that*

$$\tau_1 = \Omega\left(\frac{L_0\sqrt{d\log(1/\delta)}}{n\varepsilon}\right) \qquad \text{and} \qquad \tau_2 = \Omega\left(\frac{L_1\sqrt{d\log(1/\delta)}}{n\varepsilon}\right).$$

Proof. In the following, we use u_j to denote the j'th component of some vector u. Let $B = \frac{L_0}{L_1\sqrt{d}}$ and define $h : \mathbb{R} \mapsto \mathbb{R}$ as

$$h(z) = \begin{cases} \frac{L_1}{2}w^2 & \text{if} |w| \leq B \\ \frac{L_0}{\sqrt{d}}|w| - \frac{L_0^2}{2dL_1} & \text{otherwise} \end{cases}$$

Define $d' = \frac{d}{2}$ (assume d is even for simplicity) and for any vector $u \in \mathbb{R}^d$ let $u^{(1)} = [u_1, ..., u_{d'}]^\top$ and $u^{(2)} = [u_{d'+1}, ..., u_d]^\top$. Define $\ell(w; x) = \ell_1(w; x) + \ell_2(w; x)$ where

$$\ell_1(w; x) = \frac{L_0}{\sqrt{d}}\left\langle w^{(1)}, x^{(1)}\right\rangle, \qquad \ell_2(w; x) = \frac{1}{2}\sum_{j=d'+1}^{d} h(w_j)x_j.$$

Let $\mathcal{W} = \{w : \|w\|_\infty \leq B\}$ and note for any $w \in \mathcal{W}$ we have

$$\nabla\ell(w; x) = [\frac{x_1}{\sqrt{d}}, ..., \frac{x_{d'}}{\sqrt{d}}, w_{d'+1}x_{d'+1}, ..., w_d x_d]^\top, \qquad \nabla^2\ell_2(w; x) = L_1 \cdot \mathsf{Diag}(0, ..., 0, x_{d'+1}, ..., x_d)$$

That is, the Hessian of $\ell_2(w; x)$ is a diagonal matrix with entries from x. Thus one can observe that for any $x \in \{\pm 1\}^d$ we have that $\ell(\cdot; x)$ is L_0-Lipschitz and L_1-smooth over $\mathbb{R}^d$.

To prove a lower bound on τ_1 and τ_2, it suffices to show that for any (ε, δ)-DP implementation of $\mathcal{O}$ there exists $w \in \mathbb{R}^d$ such that $\mathop{\mathbb{E}}_{\mathcal{O}}\left[\|\mathcal{O}(w; \perp) - \nabla\mathcal{L}(w)\|^2\right] \geq \tau_1^2$ and there exist $w, w' \in \mathbb{R}^d$ such that $\mathop{\mathbb{E}}_{\mathcal{O}}\left[\|\mathcal{O}(w, w')\|^2\right] \geq \tau_2^2\|w - w'\|^2$. For sake of generality, we will show that these properties hold for a set of w, w'.

Note that to lower bound the gradient error, it suffices to lower bound the error with respect to the first d' components. We thus argue using ℓ_1, and will in fact show a lower bound for any $w \in \mathbb{R}^d$. Let $w \in \mathbb{R}^d$. We have for any (ε, δ)-DP oracle $\mathcal{O}$ there exists a dataset $S \subseteq \{\pm 1\}^d$, where $|S| = n$, of fingerprinting codes such that

$$\mathop{\mathbb{E}}_{\mathcal{O}}[\|\mathcal{O}(w; \perp) - \nabla\mathcal{L}(w)\|] \geq \mathop{\mathbb{E}}_{\mathcal{O}}\left[\left\|\mathcal{O}(w; \perp)^{(1)} - \frac{1}{n}\sum_{x\in S} x^{(1)}\right\|\right] = \Omega\left(\frac{L_0\sqrt{d\log(1/\delta)}}{n\varepsilon}\right).$$

The bound follows from standard fingerprinting code arguments. See (Bassily et al., 2014, Lemma 5.1) for a lower bound and (Steinke & Ullman, 2015, Theorem 1.1) for a group privacy reduction that obtains the additional $\sqrt{\log(1/\delta)}$ factor. This fingerprinting result also induces the parameter constraints in the theorem statement. We thus have $\tau_1 = \Omega\left(\frac{L_0\sqrt{d\log(1/\delta)}}{n\varepsilon}\right)$.

Similarly, we will argue a bound on the gradient variation using ℓ_2. Let $w, w' \in \mathcal{W}$ and $u = (w - w')^{(2)}$. In what follows, we only use the second half of the components for each vector, and thus omit the superscript $^{(2)}$ from all vectors for readability. We have $\nabla\ell_2(w; x) - \nabla\ell_2(w'; x) = L_1[u_1x_1, ..., u_{d'}x_{d'}]^\top$. Then for any $c \in (0, \frac{2L_0}{L_1\sqrt{d}}]$ and $u \in \{\pm c\}^2$ we

have

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{O}}\left[\|\mathcal{O}(w,w')-(\nabla\mathcal{L}(w)-\nabla\mathcal{L}(w'))\|^2\right] &= L_1^2\cdot\mathop{\mathbb{E}}_{\mathcal{O}}\left[\sum_{j=1}^{d'}\left(\mathcal{O}(w,w')_j-\frac{u_j}{n}\sum_{x\in S}x_j\right)^2\right]\\ &= L_1^2\cdot\mathop{\mathbb{E}}_{\mathcal{O}}\left[\sum_{j=1}^{d'}\left(u_j\Big(\frac{\mathcal{O}(w,w')_j}{u_j}-\frac{1}{n}\sum_{x\in S}x_j\Big)\right)^2\right]\\ &= L_1^2\cdot\mathop{\mathbb{E}}_{\mathcal{O}}\left[c^2\sum_{j=1}^{d'}\left(\frac{\mathcal{O}(w,w')_j}{u_j}-\frac{1}{n}\sum_{x\in S}x_j\right)^2\right]\\ &= \Omega\left(L_1^2c^2\frac{d^2\log(1/\delta)}{n^2\varepsilon^2}\right),\end{aligned}$$

where the last step again comes from fingerprinting results. Note that the extra factor of d as compared to the previous bound comes from the fact that we are considering fingerprinting codes with norm larger by a factor of $\sqrt{d}$. We also use the fact that the vector $\mathcal{O}(w,w')$ transformed using u is (ε,δ)-DP by post processing. Now since $c=\frac{\|w-w'\|}{\sqrt{d}}$ we have

$$\mathop{\mathbb{E}}_{\mathcal{O}}\left[\|\mathcal{O}(w,w')-(\nabla\mathcal{L}(w)-\nabla\mathcal{L}(w'))\|\right]=\left(L_1\|w-w'\|\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon}\right).$$

Finally, noting that $\mathop{\mathbb{E}}_{\mathcal{O}}\left[\|\mathcal{O}(w,w')-(\nabla\mathcal{L}(w)-\nabla\mathcal{L}(w'))\|^2\right]\le\mathop{\mathbb{E}}_{\mathcal{O}}\left[\|\mathcal{O}(w,w')\|^2\right]$ we obtain $\tau_2=\Omega\Big(\frac{L_1\sqrt{d\log(1/\delta)}}{n\varepsilon}\Big)$. This completes the proof. □

We remark that the accuracy lower bound for the gradient variation can hold for a much more general set of vectors than that given in the proof. Specifically, the same result can be obtained for any $u=w-w'$ such that u has $\Theta(d)$ components which are $\Omega(\frac{\|u\|}{\sqrt{d}})$ (i.e. any sufficiently spread out vector). This uses the fact that it suffices to bound the number of components which disagree in sign with the fingerprinting mean and that fingerprinting codes are sampled using a product distribution, and thus the tracing attack used by fingerprinting constructions holds over any sufficiently large subset of dimensions.

C. Missing Results for Population Stationary Points

Here we present the proof of privacy and accuracy for Algorithm 1. We start by proving the privacy guarantee.

Proof of Theorem 3.1. By parallel composition of differential privacy, and since the used batches are disjoint, it suffices to prove that each step in lines 6 and 15 of the algorithm is (ε,δ)-DP. Note that the gradient estimator in step 6 has ℓ_2-sensitivity $2L_0/b$, so by the Gaussian mechanism this step is (ε,δ)-DP.

For step 15, suppose $S_{t,s}$ and $S'_{t,s}$ are neighboring datasets that differ in at most one element: $x_{i^*}\neq x'_{i^*}$, and let η_{t,s_i} and η'_{t,s_i} the respective stepsizes used in step 23. Then

$$\|\Delta_{t,s}-\Delta'_{t,s}\|=\frac{2^{|s|}}{b}\|\nabla f\left(w_{t,s};x_{i^*}\right)-\nabla f\left(w_{t,\widehat{s}};x_{i^*}\right)-\left(\nabla f\left(w_{t,s};x'_{i*}\right)-\nabla f\left(w_{t,\widehat{s}};x'_{i^*}\right)\right)\|,$$

and note between the parent node $u_{t,\widehat{s}}$ and $u_{t,s}$ there are $2^{D-|s|}$ iterates generated by the algorithm, which we denote as

$w_{t,\widehat{s}} = w_{t,s_0}, w_{t,s_1}, ..., w_{t,s_{2^{|D|-s}}} = w_{t,s}$. Then, by smoothness of f and the triangle inequality

$$\begin{aligned}
&\|\Delta_{t,s} - \Delta'_{t,s}\| \\
&= \frac{2^{|s|}}{b}\|\nabla f\left(w_{t,s}; z_{i^*}\right) - \nabla f\left(w_{t,\widehat{s}}; z_{i^*}\right) - \left(\nabla f\left(w_{t,s}; z'_{i*}\right) - \nabla f\left(w_{t,\widehat{s}}; z'_{i^*}\right)\right)\| \\
&\leq \sum_{i=1}^{2^{D-|s|}} \frac{2^{|s|}}{b}\left[\|\nabla f\left(w_{t,s_i}; z_{i^*}\right) - \nabla f\left(w_{t,s_{i-1}}; z_{i^*}\right)\| + \|\left(\nabla f\left(w_{t,s_i}; z'_{i*}\right) - \nabla f\left(w_{t,s_{i-1}}; z'_{i^*}\right)\right)\|\right] \\
&\leq \sum_{i=1}^{2^{D-|s|}} \frac{2^{|s|}}{b} L_1 \eta_{t,s_{i-1}}\|\nabla_{t,s_{i-1}}\| + \sum_{i=1}^{2^{D-|s|}} \frac{2^{|s|}}{b} L_1 \eta'_{t,s_{i-1}}\|\nabla'_{t,s_{i-1}}\| \\
&= 2\sum_{i=1}^{2^{D-|s|}} \frac{2^{|s|}}{b}\frac{\beta}{2^{D/2}} = \frac{2\beta 2^{D/2}}{b}.
\end{aligned}$$

The Gaussian mechanism combined with our choice of $\sigma_{t,s}$ certifies privacy of this step. □

To prove Theorem 3.2 we will need some technical lemmas. Define $(\mathcal{T}, \mathcal{S})$ as a random stopping time that indicates when Algorithm 1 ends. Also, we say $(t_1, s_1) \preceq_2 (t_2, s_2)$ whenever w_{t_1,s_1} comes before w_{t_2,s_2} in the algorithm iterates.

Lemma C.1 (Gradient estimation error, extension of Lemma 6 in (Fang et al., 2018))**.** *Let $p \in (0,1)$. Then, with probability $1-p$ the event*

$$\mathcal{E} = \{\|\nabla_{t,s} - \nabla F(w_{t,s}; \mathcal{D})\|^2 \leq \alpha \cdot \tilde{\alpha} \quad \forall (t,s) \preceq_2 (\mathcal{T}, \mathcal{S})\}$$

holds, under the parameter setting of $\sigma_{t,\varnothing}, \sigma_{t,s}$ and $\eta_{t,s}$ in Algorithm 1, for

$$\alpha^2 \geq \left(\frac{L_0^2}{b} + \frac{\beta^2 D 2^D}{b}\right)\max\left\{1, \frac{(d+1)}{b\varepsilon^2}\right\} \quad \text{and} \quad \tilde{\alpha} \geq 256\log\left(\frac{1.25}{\delta}\right)\log\left(\frac{2T2^{D+1}}{p}\right)\alpha.$$

Proof. Recall the gradient estimate associated to a left child node is the same as that of the parent node. Hence, the gradient estimate of a non-leaf node is the same as that of the left-most leaf of its left sub-tree. In addition, we only need to control the gradient estimation error when we perform a gradient step, which occurs at the leaves. Then, to prove the claim, it suffices to prove that we can control the gradient estimation error at the leaves. Since, the number of iterations (and leaves) is at most $T2^{D-1}$, to prove event $\mathcal{E}$ happens with probability $1-p$, by the union bound it suffices to prove that $\mathbb{P}[\|\nabla_{t,s} - \nabla F(w_{t,s}; \mathcal{D})\|^2 > \alpha \cdot \tilde{\alpha}] \leq \frac{p}{T2^{D-1}}$ for every $(t,s) \preceq_2 (\mathcal{T}, \mathcal{S})$ where $u_{t,s}$ is a leaf.

Denote by $\mathcal{F}_t$ the sigma algebra generated by randomness in the algorithm until the end of round t. Fix $(t,s) \preceq_2 (\mathcal{T}, \mathcal{S})$ such that $u_{t,s}$ is leaf, and let $u_{t,s_\varnothing} = u_{t,s_0}, u_{t,s_1}, ..., u_{t,s_k} = u_{t,s}$ be the path from the root to s. Next, extract a sub-sequence of it including only the root and the nodes that are right children, obtaining $u_{t,s_\varnothing} = u_{t,s_{a_0}}, u_{t,s_{a_1}}, ..., u_{t,s_{a_m}} = u_{t,s}$. Now we can write

$$\begin{aligned}
\nabla_{t,s} - \nabla F(w_{t,s}; \mathcal{D}) &= \sum_{i=0}^{m} g_{t,s_{a_i}} + \sum_{x \in S_{t,\varnothing}} \underbrace{\frac{1}{b}\left(\nabla f(w_{t,\varnothing}; x) - \nabla F(w_{t,\varnothing}; \mathcal{D})\right)}_{\gamma_{1,x}} \\
&+ \sum_{i=1}^{m}\sum_{x \in S_{t,s_{a_i}}} \underbrace{\frac{2^{|s_{a_i}|}}{b}\left[\left(\nabla f(w_{t,s_{a_i}}; x) - \nabla f(w_{t,s_{a_{i-1}}}; x)\right) - \left(\nabla F(w_{t,s_{a_i}}; \mathcal{D}) - \nabla F(w_{t,s_{a_{i-1}}}; \mathcal{D})\right)\right]}_{\gamma_{2,x,i}}.
\end{aligned}$$

To bound the estimation error, we note that

$$\begin{aligned}
&\mathbb{P}[\|\nabla_{t,s} - \nabla F(w_{t,s}; \mathcal{D})\|^2 > \alpha \cdot \tilde{\alpha} | \mathcal{F}_{t-1}] \\
&\leq \mathbb{P}\Big[\Big\|\sum_{i=0}^{m} g_{t,s_{a_i}}\Big\|^2 > \frac{\alpha \cdot \tilde{\alpha}}{4}\Big|\mathcal{F}_{t-1}\Big] + \mathbb{P}\Big[\Big\|\sum_{x \in S_{t,\varnothing}} \gamma_{1,x} + \sum_{i=1}^{m}\sum_{x \in S_{t,s_{a_i}}} \gamma_{2,x,i}\Big\|^2 > \frac{\alpha \cdot \tilde{\alpha}}{4}\Big|\mathcal{F}_{t-1}\Big].
\end{aligned}$$

and proceed to bound each term on the right hand side separately. By vector subgaussian concentration (see Lemma 1 in (Jin et al., 2019)) and noting that the gaussians are independent of $\mathcal{F}_{t-1}$, we know that

$$\mathbb{P}\left[\left\|\sum_{i=0}^{m} g_{t,s_{a_i}}\right\|^2 > \frac{\alpha\cdot\tilde{\alpha}}{4}\right] \leq 4^d \exp\left(-\frac{\alpha\cdot\tilde{\alpha}}{32(\sigma_{t,\varnothing}^2+\sum_{i=1}^{m}\sigma_{t,s_{a_i}}^2)}\right),$$

and in order to bound this probability by $\frac{p}{2T2^{D-1}}$, since $m \leq D$, it suffices that

$$\begin{aligned}\alpha\cdot\tilde{\alpha} &> 32\log\left(\frac{4^d T2^D}{p}\right)\left[\frac{8L_0^2\log(1.25/\delta)}{b^2\varepsilon^2}+\frac{8D2^D\beta^2\log(1.25/\delta)}{b^2\varepsilon^2}\right]\\ &= 256\log\left(\frac{1.25}{\delta}\right)\left[d\log(4)+\log\left(\frac{T2^D}{p}\right)\right]\left[\frac{L_0^2}{b^2\varepsilon^2}+\frac{D2^D\beta^2}{b^2\varepsilon^2}\right].\end{aligned}$$

Now, noting that surely

$$\|\gamma_{1,x}\| \leq \frac{2L_0}{b} \quad \text{and} \quad \|\gamma_{2,x,i}\| \leq \frac{2\beta 2^{D/2}}{b},$$

where the second bound comes from following similar steps as in the privacy analysis in Theorem 3.1, we have that $\sum_{x\in S_{t,\varnothing}}\gamma_{1,x}+\sum_{i=1}^{m}\sum_{x\in S_{t,s_{a_i}}}\gamma_{2,x,i}$ is a sum of bounded martingale differences when conditioned on $\mathcal{F}_{t-1}$, thus by concentration of martingale-difference sequences in ℓ_2 (see Proposition 2 in (Fang et al., 2018)), and using the fact that $|S_{t,\varnothing}| = b$ and $|S_{t,s_{a_i}}| = b/2^{|s_{a_i}|}$ it follows that

$$\mathbb{P}\left[\left\|\sum_{x\in S_{t,\varnothing}}\gamma_{1,x}+\sum_{i=1}^{m}\sum_{x\in S_{t,s_{a_i}}}\gamma_{2,x,i}\right\|^2 > \frac{\alpha\cdot\tilde{\alpha}}{4}\mid\mathcal{F}_{t-1}\right] \leq 4\exp\left(-\frac{\alpha\cdot\tilde{\alpha}}{16\left[\frac{4L_0^2}{b}+\sum_{i=1}^{m}\frac{4\beta^2 2^D}{2^{|s_{a_i}|}b}\right]}\right).$$

Repeating a similar argument as before, to bound this term by $\frac{p}{2T2^{D-1}}$, it suffices that

$$\alpha\cdot\tilde{\alpha} \geq 64\log\left(\frac{2T2^{D+1}}{p}\right)\left[\frac{L_0^2}{b}+\frac{\beta^2 D2^D}{b}\right].$$

Finally, both conditions hold simultaneously for

$$\alpha^2 \geq \left(\frac{L_0^2}{b}+\frac{\beta^2 D2^D}{b}\right)\max\left\{1,\frac{(d+1)}{b\varepsilon^2}\right\}$$

and

$$\tilde{\alpha} \geq 256\log\left(\frac{1.25}{\delta}\right)\log\left(\frac{2T2^{D+1}}{p}\right)\alpha.$$

□

Lemma C.2 (Descent lemma; Lemma 7 in (Fang et al., 2018))**.** *Under the assumption that the event $\mathcal{E}$ from Lemma C.1 occurs and $\beta \leq 2^{D/2}\tilde{\alpha}$, we have that if Algorithm 1 reaches the last line, then*

$$F(w_{T,\ell(2^D)};\mathcal{D}) - F(0;\mathcal{D}) \leq -(T2^{D-1})\frac{\beta\cdot\tilde{\alpha}}{4\cdot 2^{D/2}L_1}.$$

where $w_{T,\ell(2^D)}$ is the last iterate in the T-th tree of Algorithm 1.

We provide the proof of Lemma C.2 adapted to our case for completeness.

Proof. By standard analysis for smooth functions we have

$$F(w_{t,s^+};\mathcal{D}) \le F(w_{t,s};\mathcal{D}) - \frac{\eta_{t,s}}{2}(1-\eta_{t,s}L_1)\|\nabla_{t,s}\|^2 + \frac{\eta_{t,s}}{2}\|\nabla_{t,s} - \nabla F(w_{t,s};\mathcal{D})\|^2,$$

where $\eta_{t,s} = \frac{\beta}{2^{D/2}L_1\|\nabla_{t,s}\|}$ and u_{t,s^+} is the node after $u_{t,s}$ in the tree. Since $\beta \le 2^{D/2}\tilde{\alpha}$ and $\|\nabla_{t,s}\| > 2\tilde{\alpha}$, we have that $(1-\eta_{t,s}L_1) \ge 1/2$. Using this inequality, the definition of $\eta_{t,s}$ and the fact that we are assuming $\mathcal{E}$ occurs, we obtain

$$\begin{aligned} F(w_{t,s^+};\mathcal{D}) - F(w_{t,s};\mathcal{D}) &\le -\frac{\beta}{4\cdot 2^{D/2}L_1\|\nabla_{t,s}\|}\|\nabla_{t,s}\|^2 + \frac{\beta}{2\cdot 2^{D/2}L_1\|\nabla_{t,s}\|}\alpha\cdot\tilde{\alpha} \\ &\le -\frac{\beta}{4\cdot 2^{D/2}L_1}\cdot\tilde{\alpha}, \end{aligned}$$

where the second inequality comes from $\|\nabla_{t,s}\| > 2\tilde{\alpha}$ and $\alpha \le \tilde{\alpha}$. Then telescoping over all $T2^{D-1}$ iterations provides the claimed bound.

□

We are now ready to prove the convergence guarantee of Algorithm 1.

Proof of Theorem 3.2. From Lemma C.1, we know that $\|\nabla_{t,s} - \nabla F(w_{t,s};\mathcal{D})\|^2 \le \alpha\cdot\tilde{\alpha}$ with probability $1-p$ when

$$\alpha = \sqrt{2}L_0\max\left\{\tfrac{1}{n^{1/3}}, \left(\tfrac{\sqrt{d}}{n\varepsilon}\right)^{1/2}\right\}, \tilde{\alpha} = \left(256\log\left(\tfrac{1.25}{\delta}\right)\log\left(\tfrac{2T2^{D+1}}{p}\right) + \tfrac{8L_1F_0\sqrt{2D}(D/2+1)}{2L_0^2}\right)\alpha.$$

Indeed, using our parameter setting, and noting that $d > b\varepsilon^2$ if and only if, $d > n^{2/3}\varepsilon^2$, yields

$$\begin{aligned} \alpha^2 &\ge \frac{L_0^2}{b}\max\left\{1, \frac{(d+1)}{b\varepsilon^2}\right\} + \frac{\beta^2}{2}\max\left\{1, \frac{(d+1)}{b\varepsilon^2}\right\} \\ &= L_0^2\left(\frac{1}{n^{2/3}}\mathbb{1}_{\{d+1\le n^{2/3}\varepsilon^2\}} + \frac{\sqrt{d}}{n\varepsilon}\mathbb{1}_{\{d+1> n^{2/3}\varepsilon^2\}}\right) + \frac{\alpha^2}{2}\min\left\{1, \frac{b\varepsilon^2}{d}\right\}\max\left\{1, \frac{(d+1)}{b\varepsilon^2}\right\} \\ &\ge L_0^2\max\left\{\frac{1}{n^{2/3}}, \frac{\sqrt{d}}{n\varepsilon}\right\} + \frac{\alpha^2}{2}, \end{aligned}$$

which shows our values of α and $\tilde{\alpha}$ are valid for controlling the gradient estimation error with high probability, as claimed in Lemma C.1.

Now, suppose for the sake of contradiction that Algorithm 1 does not end in line 20 under $\mathcal{E}$. This means it performs $T2^{D-1}$ gradient updates. We'll show this implies $(T2^{D-1})\frac{\beta\cdot\tilde{\alpha}}{4\cdot 2^{D/2}L_1} > F_0$ and thus contradicts Lemma C.2, which claims that $F_0 \ge -[F(w_{T,\ell(2^D)};\mathcal{D}) - F(w_{0,\ell(2^D)};\mathcal{D})] \ge (T2^{D-1})\frac{\beta\cdot\tilde{\alpha}}{4\cdot 2^{D/2}L_1}$. Indeed, note that by our parameter setting:

$$\begin{aligned} (T2^{D-1})\frac{\beta\cdot\tilde{\alpha}}{4\cdot 2^{D/2}L_1} > F_0 &\iff \beta\cdot\tilde{\alpha} > \frac{8L_1F_0}{T2^{D/2}} \\ &\iff \alpha\min\left\{1, \frac{\sqrt{b}\varepsilon}{\sqrt{d}}\right\}\cdot\tilde{\alpha} > \frac{8L_1F_0\sqrt{2D}}{T\sqrt{b}} \\ &\iff \alpha\cdot\tilde{\alpha} > \frac{8L_1F_0\sqrt{2D}(D/2+1)\sqrt{b}}{n}\max\left\{1, \frac{\sqrt{d}}{\sqrt{b}\varepsilon}\right\} \\ &\iff \alpha\cdot\tilde{\alpha} > 8L_1F_0\sqrt{2D}(D/2+1)\max\left\{\frac{\sqrt{b}}{n}, \frac{\sqrt{d}}{n\varepsilon}\right\}, \end{aligned}$$

and noting that by the setting of b we have $\max\left\{\frac{\sqrt{b}}{n}, \frac{\sqrt{d}}{n\varepsilon}\right\} = \max\left\{\frac{1}{n^{2/3}}, \frac{\sqrt{d}}{n\varepsilon}\right\}$, we conclude the following

$$(T2^{D-1})\frac{\beta\cdot\tilde{\alpha}}{4\cdot 2^{D/2}L_1} > F_0 \iff \alpha\cdot\tilde{\alpha} > 8L_1F_0\sqrt{2D}(D/2+1)\max\left\{\frac{1}{n^{2/3}}, \frac{\sqrt{d}}{n\varepsilon}\right\}$$
$$\iff \alpha\cdot\tilde{\alpha} > \frac{8L_1F_0\sqrt{2D}(D/2+1)}{2L_0^2}\alpha^2.$$

Finally, note $\alpha\cdot\tilde{\alpha} = \left(256\log(1.25/\delta)\log\left(2T2^{D+1}/p\right) + \frac{8L_1F_0\sqrt{2D}(D/2+1)}{2L_0^2}\right)\alpha^2$ and thus the last inequality holds under our parameter setting. Since this is equivalent to $(T2^{D-1})\frac{\beta\cdot\tilde{\alpha}}{4\cdot 2^{D/2}L_1} > F_0$, we are done with the contradiction. It follows that with high probability, Algorithm 1 ends in line 20 returning $w_{t,s}$ such that $\|\nabla_{t,s}\| \le 2\tilde{\alpha}$. Also, by Lemma C.1 we have $\|\nabla F(w_{t,s};\mathcal{D}) - \nabla_{t,s}\| < \tilde{\alpha}$, so the returned iterate satisfies by the triangle inequality

$$\|\nabla F(w_{t,s};\mathcal{D})\| < 3\tilde{\alpha}.$$

In addition, the linear time oracle complexity follows from the fact that at each binary tree we use b samples at the root, and then $b/2$ in levels 1 to D. This gives a total of $b(D/2+1)$ samples used at every round. Since we run the algorithm for $T = \frac{n}{b(D/2+1)}$ rounds, we compute exactly n gradients. To conclude, note the condition $n \ge \max\{\sqrt{d}(D/2+1)^2/\varepsilon, (D/2+1)^3\}$ implies the number of rounds T is at least 1. Besides, since the definition of D implies $2^D < b$, the size of the mini-batches are well-defined (meaning Algorithm 1 uses batches with at least 1 sample). This concludes the proof. □

D. Missing Results for Stationary Points in the Convex Setting

We first give pseudo-codes of algorithms used in the section.

Algorithm 5 Phased SGD$(S, (w,x)\mapsto f(w;x)), R, \eta, \mathcal{S}(\cdot), \sigma)$

Input: Dataset S, loss function $f(\cdot;x))$, radius R of the constraint set $\mathcal{W}$, steps T, η, Selection function $\mathcal{S}$, Noise variance σ

$w_1 = 0$
$K = \lceil \log(|S|)\rceil$ and $T_0 = 1$
for $k = 1$ to $K-1$ **do**
  $T_k = 2^{-k}|S|, \eta_k = 4^{-k}\eta, \sigma_k = \eta_k\sigma$
  $w_{k+1} = \mathsf{OutputPerturbedSGD}(w_k, S_{T_{k-1}+1:T_k}, R, \eta_k, \sigma_k, \mathcal{S}(\cdot))$
end for

Output: $\bar{w} = w_K$

Algorithm 6 OutputPerturbedSGD$(w_1, S, (w,x)\mapsto f(w;x), \Delta(\cdot), R, \eta, \mathcal{S}(\cdot)$

Input: Dataset S, loss function $f(\cdot;x))$, regularizer $\Delta(\cdot)$, radius R of the constraint set $\mathcal{W}$, steps T, η, Selection function $\mathcal{S}$, Noise variance σ

for $t = 1$ to $|S|-1$ **do**
  $w_{t+1} = \Pi_{\mathcal{W}}\left(w_t - \eta\left(\nabla f(w_t;x_t)\right)\right)$
end for
$\xi \sim \mathcal{N}(0, \sigma^2\mathbb{I})$
$\tilde{w} = \mathcal{S}\left(\{w_t\}_{t=1}^{|S|}\right)$

Output: $\bar{w} = \tilde{w} + \xi$

Proof of Theorem 5.1. The privacy guarantee, in both cases, follows from the privacy guarantees of Algorithm 7 and Algorithm 5, in Lemmas D.3 and D.6 respectively, together with parallel composition.

Algorithm 7 Noisy GD$(S, (w,x) \mapsto f(w;x)), R, T, \eta, \mathcal{S}(\cdot), \sigma)$

Input: Dataset S, loss function $(w,x) \mapsto f(w;x)$, radius R of the constraint set $\mathcal{W}$, steps T, η, Selection function $\mathcal{S}$, Noise variance σ

$w_1 = 0$
for $t = 1$ to $T-1$ **do**
  $\xi_t \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$
  $w_{t+1} = \Pi_{\mathcal{W}}\left(w_t - \eta\left(\nabla F(w_t; S) + \xi_t\right)\right)$
end for

Output: $\bar{w} = \mathcal{S}\left(\{w_t\}_{t=1}^T\right)$

We now proceed to the utility part. For simplicity of notation, let $R = \|w^*\|$. Recall the definition of the regularized losses $f^{(t)}(w,x)$ in Algorithm 3. Let $\{\alpha_t\}_t$ be such that $\mathbb{E}[F^{(t-1)}(\bar{w}_t; \mathcal{D})] - F^{(t-1)}(w^*_{t-1}; \mathcal{D}) \leq \alpha_t$ where $\bar{w}_t$ are the iterates produced in the algorithm and $w^*_{t-1} = \arg\min_{w \in \mathbb{R}^d} F^{(t-1)}(w; \mathcal{D})$. Following (Allen-Zhu, 2018; Foster et al., 2019), we first establish a general result which will be useful for both parts of the result.

$$
\begin{aligned}
\mathbb{E}\left\|\nabla F(\bar{w}_T; \mathcal{D})\right\| &= \mathbb{E}\left\|\nabla F^{(T-1)}(\bar{w}_T; \mathcal{D}) + \lambda \sum_{t=0}^{T} 2^t \left(\bar{w}_t - \bar{w}_T\right)\right\| \\
&\leq \mathbb{E}\left\|\nabla F^{(T-1)}(\bar{w}_T; \mathcal{D})\right\| + \lambda \sum_{t=0}^{T-1} 2^t \mathbb{E}\left(\left\|\bar{w}_t - w^*_{T-1}\right\| + \left\|\bar{w}_T - w^*_{T-1}\right\|\right) \\
&\leq 2\mathbb{E}\left\|\nabla F^{(T-1)}(\bar{w}_T; \mathcal{D})\right\| + \lambda \sum_{t=1}^{T-1} 2^t \mathbb{E}\left\|\bar{w}_t - w^*_{T-1}\right\| + \lambda \mathbb{E}\left\|w_0 - w^*_{T-1}\right\| \\
&\leq 2\mathbb{E}\left\|\nabla F^{(T-1)}(\bar{w}_T; \mathcal{D})\right\| + 4 \sum_{t=1}^{T-1} \sqrt{\lambda 2^t \alpha_t} + \lambda R_{T-1} \\
&\leq 4\sqrt{L_1 \alpha_T} + 4 \sum_{t=1}^{T-1} \sqrt{\lambda 2^{t+1} \alpha_t} + \lambda 2^{T/2} R \\
&\leq 4 \sum_{t=1}^{T} \sqrt{\lambda 2^{t+1} \alpha_t} + \sqrt{\lambda L_1} R
\end{aligned}
$$

where the third and fourth inequality follows from strong convexity of $F^{(T-1)}(\cdot; \mathcal{D})$ and Lemma D.2 respectively. The last inequality follows from the setting of T since we have that $F^{(T-1)}$ is $L_1 + \sum_{t=1}^{T-1} 2^t \lambda \leq L_1 + \lambda 2^T \leq 2L_1$ smooth. Note that the definition of R_t and Lemma D.1, $\left\|w^*_{T-1}\right\| \leq R_{T-1}$, so the unconstrained minimizer lies in the constraint set. Therefore $\mathbb{E}\left\|\nabla F^{(T-1)}(\bar{w}_T; \mathcal{D})\right\| = \mathbb{E}\left\|\nabla F^{(T-1)}(\bar{w}_T; \mathcal{D}) - \nabla F^{(T-1)}(w^*_{T-1}; \mathcal{D})\right\| \leq 2\sqrt{L_1 \alpha_T}$.

Observe that from the setting of T, $F^{(T)}$ is $4L_1$ smooth for all t. Furthermore, the radius of the constraint set in the t-th round is $R_t = 2^{T/2} R$. Hence, the Lipschitz constant $G_t \leq L_0 + 8L_1 R_t \leq O\left(L_0 + L_1 2^{T/2}\right)$. Now we instantiate α_t, which is the excess population risk bound of the DP-SCO sub-routine.

Optimal rate: The excess population risk guarantee of Algorithm 7 is in Lemma D.3, with (in context of the notation in the Lemma) Lipschitz parameter L_0 being the same and $G_\Delta = O\left(L_1 2^{T/2}\right)$. Therefore, we have $\alpha_t = \tilde{O}\left(\frac{G^2}{\lambda_t n} + \frac{dG^2}{\lambda_t n^2 \varepsilon^2}\right)$. Plugging in the above estimate, we get,

$$
\mathbb{E}\left\|\nabla F(\bar{w}; \mathcal{D})\right\| = \tilde{O}\left(\frac{G}{\sqrt{n}} + \frac{\sqrt{d}G}{n\varepsilon} + \sqrt{\frac{\lambda}{L_1}} R\right) = \tilde{O}\left(\frac{G}{\sqrt{n}} + \frac{\sqrt{d}G}{n\varepsilon}\right)
$$

where the last step follows by setting of λ.

The optimality claim follows by combining the non-private lower bound in Theorem 5.1, and the DP empirical stationarity lower bound in Theorem 4.3 together with a reduction to population stationarity as in (Bassily et al., 2019, Appendix C).

Linear time rate: The excess population risk guarantee of Algorithm 5 is in Lemma D.6, with Lipschitz parameter L_0 being the same and $G_\Delta = O\left(L_1 2^{T/2}\right)$. This gives us $\alpha_t = \tilde{O}\left(\frac{L_0^2}{\lambda_t n} + \frac{dL_0^2}{\lambda_t n^2 \varepsilon^2}\right)$, and thus

$$\mathbb{E}\left\|\nabla F(\bar{w}; \mathcal{D})\right\| = \tilde{O}\left(\frac{L_0}{\sqrt{n}} + \frac{\sqrt{d}L_0}{n\varepsilon} + \sqrt{\lambda L_1} R\right) = \tilde{O}\left(\frac{L_0}{\sqrt{n}} + \frac{\sqrt{d}L_0}{n\varepsilon} + \frac{L_1 R}{\sqrt{n}}\right)$$

where the last step follows by setting of λ. Finally, note that the Lemma D.6 requires that $n = \tilde{\Omega}\left(\frac{L_1+\lambda_t}{\lambda_t}\right)$ for all t. This can be checked to be satisfied by substituting the value of λ_t. □

D.1. Utility Lemmas

We first present some key results which will be useful in the proofs.

Lemma D.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an L_1-smooth convex function and let $w^* = \arg\min_{w \in \mathbb{R}^d} f(w)$. Let $R = \|w^*\|$ and $w_0 \in \mathbb{R}^d$ such that $\|w_0\| \leq R$. Define $\tilde{f}(w) = f(w) + \frac{\lambda}{2}\|w - w_0\|^2$ and let $\tilde{w} = \arg\min \tilde{f}(w)$. Then for any $\lambda \geq 0$, $\|\tilde{w}\| \leq \sqrt{2}R$.*

Proof. From optimality criterion, $0 = \nabla \tilde{f}(\tilde{w}) = \nabla f(\tilde{w}) + \lambda(\tilde{w} - w_0)$. Therefore, $\nabla f(\tilde{w}) = \lambda(w_0 - \tilde{w})$ and thus $\langle \nabla f(\tilde{w}), w_0 - \tilde{w} \rangle > 0$. Furthermore, since f is convex, from monotonicity, $\langle \nabla f(\tilde{w}), w^* - \tilde{w} \rangle \leq 0$. Since both w_0 and w^* lie in the ball of radius R (say $\mathcal{W}_R$), the above two implies that the hyperplane $H = \{w : \langle \nabla f(\tilde{w}), w - \tilde{w} \rangle = 0\}$ intersects with $\mathcal{W}_R$. Furthermore, since $\nabla f(\tilde{w}) = \lambda(w_0 - \tilde{w})$, we have that $\tilde{w}$ is the projection of w_0 on H i.e. $\Pi_H(w_0)$.

Let $w' = \Pi_H(0)$. We have that $w' \in \mathcal{W}_R$; this is because the hyperplane cuts the hypersphere $\mathcal{W}_R$ creating a spherical cap and w' is the center of the cap. From properties of convex projections $\|\Pi_H(w_0) - \Pi_H(0)\| \leq \|w_0 - 0\| \leq R$. Furthermore, $\Pi_H(0)$ and $\Pi_H(w_0) - \Pi_H(0)$ are orthogonal. Hence $\|\tilde{w}\|^2 = \|\Pi_H(w_0)\|^2 = \|\Pi_H(0)\|^2 + \|\Pi_H(w_0) - \Pi_H(0)\|^2 \leq 2R^2$. □

We state the following result from (Allen-Zhu, 2018; Foster et al., 2019).

Lemma D.2. *Suppose for every $t = 1, 2, \ldots T$, $\mathbb{E}[F^{(t-1)}(\bar{w}_t; \mathcal{D})] - F^{(t-1)}(w^*_{t-1}; \mathcal{D}) \leq \alpha_t$ where $\bar{w}_t$ are the iterates produced in the algorithm, $w^*_{t-1} = \arg\min_{w \in \mathbb{R}^d} F^{(t-1)}(w; \mathcal{D})$ and $\lambda_t = 2^t \lambda$, we have,*

1. *For every $t \geq 1$, $\mathbb{E}[\left\|\bar{w}_t - w^*_{t-1}\right\|^2] \leq \frac{2\alpha_t}{\lambda_{t-1}}$*
2. *For every $t \geq 1$, $\mathbb{E}[\|\bar{w}_t - w^*_t\|^2] \leq \frac{\alpha_t}{\lambda_t}$*
3. *$\mathbb{E}[\sum_{t=1}^T \lambda_t \|\bar{w}_t - w^*_T\|] \leq 4\sum_{t=1}^T \sqrt{\alpha_t \lambda_t}$*

D.2. Lemmas for NoisyGD (Algorithm 7)

Lemma D.3. *Consider a function $f(w; x) = \ell(w; x) + \Delta(w)$, where $w \mapsto \ell(w; x)$ is convex and L_0 Lipschitz for all x, and $\Delta(w)$ is λ strongly convex, G_Δ Lipschitz and H_Δ smooth over a bounded convex set $\mathcal{W}$. Algorithm 6 run with parameters $\eta = \frac{\log(T)}{\lambda T}$, $\sigma^2 = \frac{64 L_0^2 T \log(1/\delta)}{n^2 \varepsilon^2}$, $T = \max\left(\frac{L_1 + H_\Delta}{\lambda} \log\left(\frac{L_1 + H_\Delta}{\lambda}\right), \frac{n^2\varepsilon^2\left(L_0^2 + G_\Delta^2\right)}{dL_0^2 \log(1/\delta)}\right)$ and $\mathcal{S}(\{w_t\}_t) = \frac{1}{\sum_{t=1}^T (1-\eta\lambda)^{-t}} \sum_{t=1}^T (1 - \eta\lambda)^{-t} w_t$ satisfies (ε, δ)-DP and given a dataset S of n i.i.d. points from $\mathcal{D}$, the excess population risk of its output $\bar{w}$ is bounded by,*

$$\mathbb{E}\left[F(\bar{w}; \mathcal{D}) - \min_{w \in \mathcal{W}_R} F(w; \mathcal{D})\right] = O\left(\frac{L_0^2}{\lambda n} + \frac{dL_0^2 \log(1/\delta)}{\lambda n^2 \varepsilon^2}\right).$$

Proof. For the privacy analysis, as in (Bassily et al., 2014), for fixed w, the sensitivity of the gradient update is bounded by $\frac{2L_0}{n}$. Applying advanced composition, we have that $\sigma^2 = \frac{64 L_0^2 T \log(1/\delta)}{n^2 \varepsilon^2}$ suffices for (ε, δ)-DP.

For utility, we first compute a bound on uniform argument stability of the algorithm; let $\{w_t\}$ and $\{w'_t\}$ be sequence of iterates on neighbouring datasets. Note that the function $w \mapsto f(w; x)$ is $L_1 + H_\Delta$-smooth and λ-strongly convex for all x. From the setting of T, we have that the step size $\eta \leq \frac{1}{L_1 + H_\Delta}$, hence from the standard stability analysis,

$$\begin{aligned} w_{t+1} - w'_{t+1} &= w_t - \eta \nabla L(w_t; S) - \eta \nabla \Delta(w_t) - w'_t + \eta \nabla L(w'_t; S') + \eta \nabla \Delta(w'_t) \\ &= w_t - w'_t - \eta \left(\nabla L(w_t; S) + \nabla \Delta(w_t) - \nabla L(w'_t; S) - \eta \nabla \Delta(w'_t) \right) \\ &+ \eta \left(\nabla L(w'_t; S') - \nabla L(w'_t; S) \right) \\ &= \left(\mathbb{I} - \eta \left(\nabla^2 L(\tilde{w}_t; S) + \nabla^2 \Delta(\tilde{w}_t) \right) \right) (w_t - w'_t) \\ &+ \eta \left(\nabla L(w'_t; S') - \nabla L(w'_t; S) \right) \end{aligned}$$

where the last equality follows from Taylor remainder theorem where $\tilde{w}_t$ is some intermediate point on the line joining w_t and w'_t. Using the fact that $\eta \leq \frac{1}{L_1 + H_\Delta}$, we have

$$\left\| w_{t+1} - w'_{t+1} \right\| \leq (1 - \eta\lambda) \left\| w_t - w'_t \right\| + \frac{2\eta L_0}{n} \leq \frac{2L_0}{\lambda n}$$

The above gives the same bound for the iterate using the selector $\mathcal{S}$,

$$\left\| \mathcal{S}(\{w_t\}) - \mathcal{S}(\{w'_t\}) \right\| \leq \frac{2L_0}{\lambda n}$$

Note that the overall Lipschitz constant for the empirical loss is $\tilde{L_0} = L_0 + G_\Delta$. For the excess empirical risk guarantee, we use Lemma 5.2 in (Feldman et al., 2020) to get,

$$\begin{aligned} \mathbb{E}\left[L\left(\bar{w}; S\right) + \Delta(\bar{w}) - L(w^*; S) - \Delta(w^*) \right] &= \mathbb{E}\left[F\left(\bar{w}; S\right) - F(w^*; S) \right] \\ &= \tilde{O}\left(\frac{\tilde{L_0}^2}{\lambda T} \right) \\ &= \tilde{O}\left(\frac{\tilde{L_0}^2 + \sigma^2 d}{\lambda T} \right) \\ &= \tilde{O}\left(\frac{\tilde{L_0}^2}{\lambda T} + \frac{d L_0^2 \log\left(1/\delta\right)}{\lambda n^2 \varepsilon^2} \right) \\ &= O\left(\frac{d L_0^2 \log\left(1/\delta\right)}{\lambda n^2 \varepsilon^2} \right) \end{aligned}$$

where the last step follows from the setting of T. For the population risk guarantee, we have,

$$\begin{aligned} \mathbb{E}\left[F(\bar{w}; \mathcal{D}) - F(w^*; \mathcal{D}) \right] &= \mathbb{E}\left[F(\bar{w}; \mathcal{D}) - F(\bar{w}; S) \right] + \mathbb{E}\left[F(\bar{w}; \mathcal{D}) - F(w^*) \right] \\ &= \mathbb{E}[L(\bar{w}; \mathcal{D}) - L(\bar{w}; S)] + O\left(\frac{d L_0^2 \log\left(1/\delta\right)}{\lambda n^2 \varepsilon^2} \right) \\ &\leq L_0 \mathbb{E} \left\| \bar{w} - \bar{w}' \right\| + O\left(\frac{d L_0^2 \log\left(1/\delta\right)}{\lambda n^2 \varepsilon^2} \right) \\ &= \tilde{O}\left(\frac{L_0^2}{\lambda n} + \frac{d L_0^2 \log\left(1/\delta\right)}{\lambda n^2 \varepsilon^2} \right) \end{aligned}$$

where the inequality follows from Lipschitzness and standard generalization gap to stability argument. □

D.3. Lemmas for PhasedSGD (Algorithm 5)

The following lemma gives population risk guarantees for strongly convex functions under privacy, in terms of variance of stochastic gradients, as opposed to standard Lipschitzness bounds.

Lemma D.4 (Variance based bound for constant step-size SGD for strongly-convex functions)**.** *Consider a function $f(w; x)$ such that $w \mapsto f(w; x)$ is λ strongly convex, L_1 smooth over a convex set $\mathcal{W}$ for all x and let*

$\mathbb{E}_x \|\nabla f(w;x) - \mathbb{E}_x \nabla f(w;x)\|^2 \leq \mathcal{V}^2$ *for all* $w \in \mathcal{W}$. *Let* $\gamma_t = (1-\eta\lambda)^{-t}$. *Given a dataset* $S = \{x_1, x_2, \ldots, x_n\}$ *sampled i.i.d from* $\mathcal{D}$ *and* $\eta \leq \frac{1}{2\beta}$ *as input, for any* $w \in \mathcal{W}$, *the iterates of Algorithm 6 satisfy*

$$\mathbb{E}\left[\frac{1}{\sum_{t=1}^n \gamma_t}\sum_{t=1}^n \gamma_t F(w_t;\mathcal{D})\right] - F(w) \leq \frac{\lambda}{e^{\eta\lambda n}-1}\|w_0 - w\|^2 + \eta\mathcal{V}^2$$

Furthermore, for $n = \Omega\left(\frac{L_1}{\lambda}\log\left(\frac{L_1}{\lambda}\right)\right)$, *with* $\eta = \frac{\log(n)}{\lambda n}$ *and* $\mathcal{S}(\{w_t\}_t) = \frac{1}{\sum_{t=1}^n \gamma_t}\sum_{t=1}^n \gamma_t w_t$, *the excess population risk of* $\tilde{w} = \mathcal{S}(\{w_t\}_t)$ *satisfies*

$$\mathbb{E}\left[F(\tilde{w};\mathcal{D}) - \min_{w\in\mathcal{W}} F(w;\mathcal{D})\right] = O\left(\frac{\mathcal{V}^2\log(n)}{\lambda n}\right)$$

Proof. An equivalent way to write the update in Algorithm 6 is

$$w_{t+1} = \arg\min_{w\in\mathcal{W}}\left(\langle\nabla f(w_t,x_t), w\rangle + \frac{1}{\eta}\|w_t - w\|^2 + \psi(w)\right)$$

where $\psi(w) = 0$ if $w \in \mathcal{W}$, otherwise ∞.

Following standard arguments in convex optimization, for any $w \in \mathcal{W}$, we have

$$\begin{aligned}
&F(w_{t+1};\mathcal{D}) - F(w)\\
&= F(w_{t+1};\mathcal{D}) + \psi(w_{t+1}) - F(w;\mathcal{D}) - \psi(w)\\
&\leq F(w_t) + \langle\nabla F(w_t), w_{t+1} - w_t\rangle + \frac{L_1}{2}\|w_{t+1} - w_t\|^2 + \psi(w_{t+1})\\
&+ F(w;\mathcal{D}) - \psi(w)\\
&\leq \langle\nabla F(w_t), w_{t+1} - w_t\rangle + \langle\nabla F(w_t), w_t - w\rangle - \frac{\lambda}{2}\|w_t - w\|^2 + \frac{L_1}{2}\|w_{t+1} - w_t\|^2\\
&+ \psi(w_{t+1}) + F(w;\mathcal{D}) - \psi(w)\\
&= \mathbb{E}_{z_t}\left[\langle\nabla p(w_t;z_t) - \nabla F(w;\mathcal{D}), w_t - w_{t+1}\rangle + \frac{L_1}{2}\|w_{t+1} - w_t\|^2 + \langle\nabla p(w_t;z_t), w_t - w\rangle\right]\\
&- \frac{\lambda}{2}\|w_t - w\|^2 + \psi(w_{t+1}) + F(w;\mathcal{D}) - \psi(w)\\
&\leq \mathbb{E}_{z_t}\Big[\langle\nabla p(w_t;z_t) - \nabla F(w;\mathcal{D}), w_t - w_{t+1}\rangle - \left(\frac{1}{2\eta} - \frac{L_1}{2}\right)\|w_{t+1} - w_t\|^2\\
&+ \left(\frac{1}{2\eta} - \frac{\lambda}{2}\right)\|w_t - w\|^2 - \frac{1}{2\eta}\|w_{t+1} - w\|^2\Big]\\
&\leq \mathbb{E}_{z_t}\Big[\frac{\eta}{2(1-\eta L_1)}\|\nabla p(w_t;z_t) - \nabla F(w;\mathcal{D})\|^2 + \left(\frac{1}{2\eta} - \frac{\lambda}{2}\right)\|w_t - w\|^2 - \frac{1}{2\eta}\|w_{t+1} - w\|^2\Big]\\
&\leq \eta\mathcal{V}^2 + \mathbb{E}_{z_t}\left[\left(\frac{1}{2\eta} - \frac{\lambda}{2}\right)\|w_t - w\|^2 - \frac{1}{2\eta}\|w_{t+1} - w\|^2\right]
\end{aligned}$$

where the first inequality follows from smoothness, the second from strong convexity, the third from Fact D.1 in (Allen-Zhu, 2018), fourth from AM-GM inequality and the last from the assumption about variance bound on the oracle.

Now, the above is exactly the bound obtained in the proof of Lemma 5.2 in (Feldman et al., 2020) with the second moment on gradient norm replaced by variance. Repeating the rest of the arguments in that Lemma gives us the claimed result. □

Lemma D.5 (Privacy of Algorithm 6)**.** *Consider a function* $f(w;x) = \ell(w;x) + \Delta(w)$ *such that* $w \mapsto \ell(w;x)$ *is convex,* L_0 *Lipschitz,* L_1*-smooth for all* z, *and* $\Delta(\cdot)$ *is* λ *strongly convex,* G_Δ *Lipschitz and* H_Δ *smooth over a bounded set* $\mathcal{W}$. *For* $n = \Omega\left(\frac{L_1+H_\Delta}{\lambda}\log\left(\frac{L_1+H_\Delta}{\lambda}\right)\right)$, *Algorithm 6 with input as function* $(w,x) \mapsto f(w;x)$, $\sigma^2 = \frac{64G^2(\log(n))^2\log(1/\delta)}{\lambda^2 n^2\varepsilon^2}$, $\eta = \frac{\log(n)}{\lambda n}$ *and* $\mathcal{S}\left(\{w_t\}_{t=1}^n\right) = \frac{1}{\sum_{t=1}^n \gamma_t}\sum_{t=1}^n \gamma_t w_t$ *for any weights* γ_t *satisfies* (ε,δ)*-DP.*

Proof. We start with computing the sensitivity of the algorithm's output: let $\{w_t\}$ and $\{w'_t\}$ be sequence of iterates produced by Algorithm 6 on neighbouring datasets. Note that the function $w \mapsto f(w;x)$ is $L'_1 = L_1 + H_\Delta$-smooth and λ-strongly convex for all x. From the assumption on n, we have that the step size $\eta \leq \frac{1}{H+H_\Delta}$. Suppose the differing sample between neighbouring datasets is x_j, then $w_t = w'_t$ for all $t \leq j$. Also,

$$\left\|w_{j+1} - w'_{j+1}\right\| = \eta \left\|\nabla \ell(w_j; x_j) - \nabla \ell(w_j; x'_j)\right\| \leq 2\eta L_0 = \frac{2L_0 \log(n)}{\lambda n}$$

Now, for any $t > j$, as in the standard stability analysis we have,

$$\begin{aligned} w_{t+1} - w'_{t+1} &= w_t - \eta \nabla \ell(w_t; x_t) - \eta \nabla \Delta(w_t) - w_t + \eta \nabla \ell(w'_t; x_t) + \eta \nabla \Delta(w'_t) \\ &= \left(\mathbb{I} - \eta \left(\nabla^2 \ell(\tilde{w}_t; x_t) + \nabla^2 \Delta(\tilde{w}_t)\right)\right)(w_t - w'_t) \end{aligned}$$

where the last equality follows from Taylor remainder theorem where $\tilde{w}_t$ is some intermediate point in the line joining w_t and w'_t. Using the fact that $\eta \leq \frac{1}{L_1+H_\Delta}$ and λ strong convexity, we have

$$\left\|w_{t+1} - w'_{t+1}\right\| \leq (1 - \eta\lambda) \left\|w_t - w'_t\right\| \leq \left\|w_{j+1} - w'_{j+1}\right\| \leq \frac{2L_0 \log(n)}{\lambda n}$$

Applying convexity to the weights in the definition of the selector function $\mathcal{S}$, we get,

$$\left\|\mathcal{S}(\{w_t\}) - \mathcal{S}(\{w'_t\})\right\| \leq \frac{2L_0 \log(n)}{\lambda n}$$

The privacy proof now follows from the Gaussian mechanism guarantee. □

Lemma D.6 (Phased SGD composite guarantee)**.** *Consider a function $f(w;x) = \ell(w;x) + \Delta(w)$ where $w \mapsto \ell(w;x)$ is convex, L_0 Lipschitz, L_1 smooth for all x, and $\Delta(w)$ is λ strongly convex, G_Δ Lipschitz and H_Δ smooth over a bounded set $\mathcal{W}$. For $n = \Omega\left(\frac{K(L_1+H_\Delta)}{\lambda} \log\left(\frac{L_1+H_\Delta}{\lambda}\right)\right)$, Algorithm 6 with $\sigma^2 = \frac{64L_0^2K^2(\log(n))^2\log(1/\delta)}{\lambda^2 n^2 \varepsilon^2}$, satisfies (ε, δ)-DP. Furthermore, with input as function $(w,x) \mapsto f(w;x)$, a dataset S of n samples drawn i.i.d. from $\mathcal{D}$, $\eta = \frac{\log(n)}{\lambda n}$, $K = \ln\ln n$, $\gamma_t = (1-\eta\lambda)^{-t}$ and $\mathcal{S}\left(\{w_t\}_{t=1}^n\right) = \frac{1}{\sum_{t=1}^n \gamma_t} \sum_{t=1}^n \gamma_t w_t$, the excess population risk of output w_K is bounded as*

$$\mathbb{E}\left[F(w_K; \mathcal{D})\right] - \min_{w \in \mathcal{W}} F(w; \mathcal{D}) = \tilde{O}\left(\frac{L_0^2}{\lambda n} + \frac{dL_0^2}{\lambda n^2 \varepsilon^2}\right)$$

Proof. The privacy proof simply follows from parallel composition. For the utility proof, we repeat the arguments in Theorem 5.3 in (Feldman et al., 2020) substituting the variance-based bound from Lemma D.4. Note that the variance of the stochastic gradients used, $\mathcal{V}^2 \leq L_0^2$, this gives us,

$$\mathbb{E}\left[F(w_K; \mathcal{D})\right] - \min_{w \in \mathcal{W}} F(w; \mathcal{D}) = \tilde{O}\left(\frac{L_0^2}{\lambda n} + \frac{dL_0^2}{\lambda n^2 \varepsilon^2}\right)$$

□

E. Missing Results for Generalized Linear Models

We first give the definition of oblivious subspace embedding.

Definition E.1 ((r, τ, β)-oblivious subspace embedding)**.** A random matrix $\Phi \in \mathbb{R}^{k \times d}$ is an (r, τ, β)-oblivious subspace embedding if for any r dimensional linear subspace in $\mathbb{R}^d$, say V, we have that with probability at least $1 - \beta$, for all $x \in V$,

$$(1 - \tau)\left\|x\right\|^2 \leq \left\|\Phi x\right\|^2 \leq (1 + \tau)\left\|x\right\|^2$$

It is well-known that JL matrices with embedding dimension $k = O\left(\frac{r\log(2/\beta)}{\tau^2}\right)$ are (r, τ, β)-oblivious subspace embeddings and can be constructed efficiently (Cohen, 2016). A simple example is a scaled Gaussian random matrix, $\Phi = \frac{1}{\sqrt{k}}\mathbf{G}$ where entries of $\mathbf{G}$ are independent and distributed as $\mathcal{N}(0,1)$.

Proof of Theorem 6.1. We first prove privacy. Let $G(S)$ and $H(S)$ be the bounds on the Lipschitz and smoothness constants of the family of loss functions $\{w \mapsto f(w;\Phi x)\}_{x\in S}$. With $k = \Omega(\log(2n/\delta))$, from the JL-property, it follows that with probability at least $1-\delta/2$, $G(S) \le 2L_0\|\mathcal{X}\|$ and $H(S) \le 2L_1\|\mathcal{X}\|^2$. Hence, using the fact that $\mathcal{A}$ is $(\varepsilon, \delta/2)$-DP, we have that Algorithm 4 is (ε,δ)-DP.

We now proceed to the utility part. Let $\tilde{w} \in \mathbb{R}^k$ be the output of the base algorithm in low dimensions. Note that the final output is $\bar{w} = \Phi^\top \tilde{w}$. The transpose of the JL matrix can only increase the norm by the polynomial factor of d and n, hence $\|\bar{w}\| \le \text{poly}(n,d)\|\tilde{w}\|$. By assumption, $\mathbb{P}(\|\tilde{w}\| > \text{poly}(n,d,L_0,L_1)) \le \frac{1}{\sqrt{n}}$. Hence we also have that $\mathbb{P}(\|\bar{w}\| > \text{poly}(n,d,L_0,L_1)) \le \frac{1}{\sqrt{n}}$. Let $\mathcal{W} \subseteq \mathbb{R}^d$ denote the above set with radius $\text{poly}(n,d,L_0,L_1)$.

We now decompose the population stationarity as,

$$\begin{aligned}\mathbb{E}\|\nabla F(\bar{w};\mathcal{D})\| &\le \mathbb{E}\|\nabla F(\bar{w};\mathcal{D}) - \nabla F(\bar{w};S)\| + \|\nabla F(\bar{w};S)\| \\ &\le \mathbb{E}\sup_{w\in\mathcal{W}}\|\nabla F(w;\mathcal{D}) - \nabla F(w;S)\| + \frac{L_0\|\mathcal{X}\|}{\sqrt{n}} + \mathbb{E}\|\nabla F(\bar{w};S)\|, \end{aligned} \tag{7}$$

where the last inequality follows from the above reasoning that that $P(\bar{w}\in\mathcal{W}) \ge 1 - \frac{1}{\sqrt{n}}$. The first term is bounded from uniform convergence guarantee in Lemma E.2 noting that the dependence on $\|\mathcal{W}\|$ in the Lemma is only poly-logarithmic.

$$\mathbb{E}\sup_{w\in\mathcal{W}}\|\nabla F(w;\mathcal{D}) - \nabla F(w;S)\| = \tilde{O}\left(\frac{L_0\|\mathcal{X}\|}{\sqrt{n}}\right) \tag{8}$$

We now prove a bound on the empirical stationarity. Note that it suffices to prove a high-probability (over the random JL matrix) bound because the norm of gradient is bounded in worst case by $L_0\|\mathcal{X}\|$. Thus the expected norm of gradient of the output is bounded by the high probability bound by considering a small enough failure probability.

From the assumption on $\mathcal{A}$, with probability at least $1-\delta/2$,

$$\|\nabla F(\tilde{w};\Phi S)\| = \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \phi'_{y_i}(\langle\tilde{w},\Phi x_i\rangle)\Phi x_i\right\| \le g(k,n,2L_0\|\mathcal{X}\|, 2L_0\|\mathcal{X}\|,\varepsilon,\delta/2)$$

We now use the fact that if $k = O(\mathsf{rank}\log(2n/\delta))$, then the JL transform is an $(\mathtt{rank}, 1/2, \delta/2)$ oblivious subspace embedding (see Definition E.1). Thus, it approximates the norm of any vector in $\mathsf{span}(\{x_i\}_{i=1}^n)$, and hence any gradient. Therefore,

$$\begin{aligned}\mathbb{E}\|\nabla F(\tilde{w};\Phi S)\| &= \mathbb{E}\left\|\Phi\left(\frac{1}{n}\sum_{i=1}^n \phi'_{y_i}(\langle\tilde{w},\Phi x_i\rangle)x_i\right)\right\| \ge \left(1-\sqrt{\frac{\mathsf{rank}}{k}}\right)\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \phi'_{y_i}(\langle\tilde{w},\Phi x_i\rangle)x_i\right\| \\ &\ge \frac{1}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \phi'_{y_i}(\langle\tilde{w},\Phi x_i\rangle)x_i\right\| = \frac{1}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \phi'_{y_i}(\langle\Phi^\top\tilde{w}, x_i\rangle)x_i\right\| = \frac{1}{2}\mathbb{E}\|\nabla F(\bar{w};S)\|\end{aligned}$$

Thus with $k = O(\mathsf{rank}\log(2n/\delta))$, we get

$$\mathbb{E}\|\nabla F(\bar{w};S)\| \le g(k,n,2L_0\|\mathcal{X}\|, 2L_1\|\mathcal{X}\|^2,\varepsilon,\delta) = g(\mathsf{rank},n,2L_0\|\mathcal{X}\|, 2L_1\|\mathcal{X}\|^2,\varepsilon,\delta)$$

For the other bound, let $I_{d-k} \in \mathbb{R}^{d\times k}$ denote the matrix with first k diagonal entries, $(I_{d-k})_{j,j}$ with $j\in[k]$, are 1 and the

rest of the matrix is zero. We have,

$$
\begin{aligned}
&\mathbb{E}\left\|\nabla F(\bar{w}; S)\right\| \\
&= \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\phi'_{y_i}(\langle\Phi^\top\tilde{w}, x_i\rangle)x_i\right\| \\
&\leq \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\phi'_{y_i}(\langle\tilde{w}, \Phi x_i\rangle)I_{d-k}\Phi x_i\right\| + \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\phi'_{y_i}(\langle\tilde{w}, \Phi x_i\rangle)x_i - \frac{1}{n}\sum_{i=1}^{n}\phi'_{y_i}(\langle\tilde{w}, \Phi x_i\rangle)I_{d-k}\Phi x_i\right\|\right] \\
&\leq \mathbb{E}\left\|I_{d-k}\right\|\left\|\frac{1}{n}\sum_{i=1}^{n}\phi'_{y_i}(\langle\tilde{w}, \Phi x_i\rangle)\Phi x_i\right\| + \frac{1}{n}\mathbb{E}\sum_{i=1}^{n}\left|\phi'_{y_i}(\langle\tilde{w}, \Phi x_i\rangle)\right|\left\|\left\|x_i - I_{d-k}\Phi x_i\right\|\right\| \\
&\leq \mathbb{E}\left\|\nabla F(\tilde{w}; \Phi S)\right\| + \frac{1}{n}\mathbb{E}\sum_{i=1}^{n}L_0\left\|I - I_{d-k}\Phi\right\|\left\|x_i\right\| \\
&\leq g(k, n, 2L_0\left\|\mathcal{X}\right\|, 2L_1\left\|\mathcal{X}\right\|^2, \varepsilon, \delta/2) + L_0\left\|\mathcal{X}\right\|\mathbb{E}\left\|I - \mathbf{H}\right\|
\end{aligned}
$$

where the second inequality follows from triangle inequality, the third inequality follows from L_0-Lipschitzness of the GLM, the third inequality follows from the accuracy guarantee of the base algorithm and substituting $\mathbf{H} = I_{d-k}\Phi$. To bound $\mathbb{E}\left\|I - \mathbf{H}\right\|$, we use concentration properties of distribution used in the construction of JL matrices. Specifically, using the scaled Gaussian matrix construction, from concentration of extreme eignevalues of square Gaussian matrices, we have that $\mathbb{E}\left\|I - \mathbf{H}\right\| = \tilde{O}\left(\frac{1}{\sqrt{k}}\right)$ (Rudelson & Vershynin, 2010). This gives us,

$$
\mathbb{E}\left\|\nabla F(\bar{w}; S)\right\| \leq g(k, n, 2L_0\left\|\mathcal{X}\right\|, 2L_1\left\|\mathcal{X}\right\|^2, \varepsilon, \delta/2) + \tilde{O}\left(\frac{L_0\left\|\mathcal{X}\right\|}{\sqrt{k}}\right)
$$

Choosing k to minimize the above yields the bound of $\tilde{O}\left(\frac{L_0\left\|\mathcal{X}\right\|}{\sqrt{k}}\right)$. Combining the two cases, yields the bound of $g(k, n, 2L_0\left\|\mathcal{X}\right\|, 2L_1\left\|\mathcal{X}\right\|^2, \varepsilon, \delta/2)$ on gradient norm. Plugging this and the bound in Eqn. (8) in Inequality (7) gives the claimed bound. □

Lemma E.2. *Let $\mathcal{D}$ be a probability distribution over $\mathcal{X}$ such that $\|x\| \leq \|\mathcal{X}\|$ for all $x \in supp(\mathcal{D})$. Let $f(w; (x, y)) = \phi_y(\langle w, x\rangle)$ be an L_1-smooth L_0-Lipschitz GLM. Then, with probability at least $1 - \beta$, over a draw of n i.i.d. samples S from $\mathcal{D}$, we have*

$$
\sup_{w\in\mathcal{W}}\left\|\nabla F(w; \mathcal{D}) - \nabla F(w; S)\right\| \leq \frac{4L_0\left\|\mathcal{X}\right\|\log\left(2n^{3/2}\left\|\mathcal{W}\right\|L_1\left\|\mathcal{X}\right\|/L_0\right)}{\sqrt{n}} + \frac{4L_0\left\|\mathcal{X}\right\|\sqrt{\log(1/\beta)}}{\sqrt{n}}
$$

Proof. We first give a bound on the expected uniform deviation, $\mathbb{E}_{S\sim\mathcal{D}^n}\sup_{w\in\mathcal{W}}\left\|\nabla F(w; \mathcal{D}) - \nabla F(w; S)\right\|$. The gradient of the loss function is $\nabla f(w; x) = \phi'_x(\langle w, x\rangle)x$. We start with the standard symmetrization trick,

$$
\begin{aligned}
&\mathbb{E}_{S\sim\mathcal{D}^n}\sup_{w\in\mathcal{W}}\left\|\nabla F(w; \mathcal{D}) - \nabla F(w; S)\right\| \\
&= \mathbb{E}_{S\sim\mathcal{D}^n}\sup_{w\in\mathcal{W}}\left\|\mathbb{E}\phi'_y(\langle w, x\rangle)x - \frac{1}{n}\sum_{i=1}^{n}\phi'_{x_i}(\langle w, x_i\rangle)x_i\right\| \\
&= \mathbb{E}_{S\sim\mathcal{D}^n}\sup_{w\in\mathcal{W}}\left\|\mathbb{E}_{\{x'_i\}\sim\mathcal{D}^n}\frac{1}{n}\sum_{i=1}^{n}\phi'_{y'_i}(\langle w, x'_i\rangle)x'_i - \frac{1}{n}\sum_{i=1}^{n}\phi'_{x_i}(\langle w, x_i\rangle)x_i\right\| \\
&\leq \mathbb{E}_{S,S'\sim\mathcal{D}^n}\sup_{w\in\mathcal{W}}\left\|\frac{1}{n}\sum_{i=1}^{n}\phi'_{y'_i}(\langle w, x'_i\rangle)x'_i - \frac{1}{n}\sum_{i=1}^{n}\phi'_{x_i}(\langle w, x_i\rangle)x_i\right\| \\
&= \mathbb{E}_{S,S'\sim\mathcal{D}^n}\mathbb{E}_{\{\sigma_i\}}\sup_{w\in\mathcal{W}}\left\|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(\phi'_{y'_i}(\langle w, x'_i\rangle)x'_i - \phi'_{x_i}(\langle w, x_i\rangle)x_i\right)\right\| \\
&\leq 2\mathbb{E}_{S\sim\mathcal{D}^n}\mathbb{E}_{\{\sigma_i\}}\sup_{w\in\mathcal{W}}\left\|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\phi'_{y_i}(\langle w, x_i\rangle)x_i\right\| \qquad (9)
\end{aligned}
$$

where σ_i are i.i.d. Rademacher random variables. For fixed $\{x_i\}_{i=1}^n$, consider a set $\mathcal{W}_0$ s.t. for all $w \in \mathcal{W}$ and $i \in [n]$, there exists $w_0 \in \mathcal{W}_0$ such that $|\langle w, x_i\rangle - \langle w_0, x_i\rangle| \leq \tau$. Since $\|w\| \leq \|\mathcal{W}\|$ and $\|x_i\| \leq \|\mathcal{X}\|$, we require only $\frac{2n\|\mathcal{W}\|\|\mathcal{X}\|}{\tau}$ points in $\mathcal{W}_0$ to satisfy the above covering condition. Therefore,

$$\begin{aligned}
&\mathbb{E}_{S\sim\mathcal{D}^n}\mathbb{E}_{\{\sigma_i\}}\sup_{w\in\mathcal{W}}\left\|\frac{1}{n}\sum_{i=1}^n\sigma_i\phi'_{y_i}\left(\langle w,x_i\rangle\right)x_i\right\| \\
&= \mathbb{E}_{S\sim\mathcal{D}^n}\mathbb{E}_{\{\sigma_i\}}\sup_{w\in\mathcal{W},w_0\in\mathcal{W}_0}\left\|\frac{1}{n}\sum_{i=1}^n\sigma_i\left(\phi'_{y_i}\left(\langle w,x_i\rangle\right)-\phi'_{y_i}\left(\langle w_0,x_i\rangle\right)+\phi'_{y_i}\left(\langle w_0,x_i\rangle\right)\right)x_i\right\| \\
&\leq \mathbb{E}_{S\sim\mathcal{D}^n}\mathbb{E}_{\{\sigma_i\}}\sup_{w\in\mathcal{W},w_0\in\mathcal{W}_0}\left\|\frac{1}{n}\sum_{i=1}^n\sigma_i\left(\phi'_{y_i}\left(\langle w,x_i\rangle\right)-\phi'_{y_i}\left(\langle w_0,x_i\rangle\right)\right)x_i\right\|+\left\|\frac{1}{n}\sum_{i=1}^n\sigma_i\phi'_{y_i}\left(\langle w_0,x_i\rangle\right)x_i\right\| \\
&\leq \mathbb{E}_{S\sim\mathcal{D}^n}\mathbb{E}_{\{\sigma_i\}}\sup_{w\in\mathcal{W},w_0\in\mathcal{W}_0}L_1\left|\langle w,x_i\rangle-\langle w_0,x_i\rangle\right|\|\mathcal{X}\|+\mathbb{E}_{S\sim\mathcal{D}^n}\mathbb{E}_{\{\sigma_i\}}\sup_{w_0\in\mathcal{W}_0}\left\|\frac{1}{n}\sum_{i=1}^n\sigma_i\phi'_{y_i}\left(\langle w_0,x_i\rangle\right)x_i\right\| \\
&\leq L_1\tau\|\mathcal{X}\|+\mathbb{E}_{S\sim\mathcal{D}^n}\mathbb{E}_{\{\sigma_i\}}\sup_{w_0\in\mathcal{W}_0}\left\|\frac{1}{n}\sum_{i=1}^n\sigma_i\phi'_{y_i}\left(\langle w_0,x_i\rangle\right)x_i\right\| \qquad (10)
\end{aligned}$$

where the second last inequality follows from smoothness and the last from the definition of cover $\mathcal{W}_0$. For fixed w_0, from standard manipulations, we have,

$$\begin{aligned}
\mathbb{E}_{\{\sigma_i\}}\left\|\frac{1}{n}\sum_{i=1}^n\sigma_i\phi'_{y_i}\left(\langle w_0,x_i\rangle\right)x_i\right\| &\leq \sqrt{\mathbb{E}_{\{\sigma_i\}}\left\|\frac{1}{n}\sum_{i=1}^n\sigma_i\phi'_{y_i}\left(\langle w_0,x_i\rangle\right)x_i\right\|^2} \\
&= \sqrt{\frac{1}{n^2}\mathbb{E}_{\{\sigma_i\}}\sum_{i=1}^n\left\|\sigma_i\phi'_{y_i}\left(\langle w_0,x_i\rangle\right)x_i\right\|^2} \\
&\leq \frac{L_0\|\mathcal{X}\|}{\sqrt{n}}
\end{aligned}$$

Using Massart's finite class lemma to handle all $w_0 \in \mathcal{W}_0$, and substituting the above in Eqn. (10), we get,

$$\mathbb{E}_{S\sim\mathcal{D}^n}\mathbb{E}_{\{\sigma_i\}}\sup_{w\in\mathcal{W}}\left\|\frac{1}{n}\sum_{i=1}^n\sigma_i\phi'_{y_i}\left(\langle w,x_i\rangle\right)x_i\right\| \leq L_1\tau\|\mathcal{X}\|+\frac{G\|\mathcal{X}\|\log\left(2n\|\mathcal{W}\|\|\mathcal{X}\|/\tau\right)}{\sqrt{n}}$$

Choosing $\tau = \frac{L_0}{L_1\sqrt{n}}$, we get,

$$\mathbb{E}_{S\sim\mathcal{D}^n}\mathbb{E}_{\{\sigma_i\}}\sup_{w\in\mathcal{W}}\left\|\frac{1}{n}\sum_{i=1}^n\sigma_i\phi'_{y_i}\left(\langle w,x_i\rangle\right)x_i\right\| \leq \frac{2L_0\|\mathcal{X}\|\log\left(2n^{3/2}\|\mathcal{W}\|L_1\|\mathcal{X}\|/L_0\right)}{\sqrt{n}}$$

Finally, substituting the above in Eqn. (9) gives us the following in-expectation bound.

$$\mathbb{E}_{S\sim\mathcal{D}^n}\sup_{w\in\mathcal{W}}\|\nabla F(w;\mathcal{D})-\nabla F(w;S)\| \leq \frac{4L_0\|\mathcal{X}\|\log\left(2n^{3/2}\|\mathcal{W}\|L_1\|\mathcal{X}\|/L_0\right)}{\sqrt{n}}$$

For the high-probability bound, let $\psi(S) = \sup_{w\in\mathcal{W}}\|\nabla F(w;\mathcal{D})-\nabla F(w;S)\|$ and let $w^* \in \mathcal{W}$ achieves the supremum. We can bound the increment between neighbouring datasets S and S' as,

$$\begin{aligned}
|\psi(S)-\psi(S')| &\leq \left|\|\nabla F(w^*;\mathcal{D})-\nabla F(w^*;S)\|-\|\nabla F(w^*;\mathcal{D})-\nabla F(w^*;S')\|\right| \\
&\leq \|\nabla F(w^*;S)-\nabla F(w^*;S')\| \\
&\leq \frac{2L_0\|\mathcal{X}\|}{n}
\end{aligned}$$

Finally, applying McDiarmid's inequality gives the claimed bound. □

Proof of Corollary 6.2. The results follow from Theorem 6.1 provided we show that the conditions on the base algorithm in the Theorem statement are satisfied. The privacy and accuracy claims follow from Theorem 3.2 and 5.1 respectively. We note that even though we are given population stationarity guarantee for the convex case, the same bound for empirical stationarity guarantee simply follows from the re-sampling argument in (Bassily et al., 2019). The only thing left to show is the high-probability bound on the trajectory of the algorithm.

Non-convex setting with Private Spiderboost: From the update in Algorithm 2, we have that for any t

$$\|\nabla_t\| \leq \sum_{i=1}^{t} \|\Delta_i\| + \left\|\sum_{i=1}^{t} g_t\right\| \leq 2tL_0 + \left\|\sum_{i=1}^{t} g_t\right\|$$

where the last inequality follows from the Lipschitzness assumption. Note that $g_t \sim \mathcal{N}(0, \sigma_t^2 \mathbb{I})$ where $\sigma_t \leq O\left(\max\left(\sigma_1, \widehat{\sigma}_2\right)\right) = O\left(\text{poly}(n, d, L_0, L_1)\right)$. Hence $\left\|\sum_{i=1}^{t} g_t\right\| \leq \sqrt{d \log\left(1/\beta'\right)} O\left(\text{poly}(n, d, L_0, L_1)\right)$ with probability at least $1 - \beta'$. Taking a union bound over all $t \in T$ gives us $\|w_t\| \leq \text{poly}(n, d, L_0, L_1, \log\left(\text{poly}(n, d)/\beta\right))$ with probability at least $1 - \beta$. Substituting $\beta = \frac{1}{\sqrt{n}}$ yields the guarantee of Theorem 6.1.

Convex setting with Recursive Regularization: Since the iterates are restricted to the constraint set, the final output, with probability one, lies in the set of radius

$$R_T = 2^{T/2} \|w^*\| = O\left(\sqrt{\frac{L_1}{\lambda}} \|w^*\|\right) = O\left(\frac{L_1 \|w^*\|^{3/2} n}{L_0}\right)$$

which completes the proof. □