Differentially Private Non-Convex Optimization under the KL Condition with Optimal Rates

Michael Menart MENART.2@OSU.EDU
Department of Computer Science & Engineering, The Ohio State University

Enayat Ullah ENAYAT@JHU.EDU
Department of Computer Science, The Johns Hopkins University

Raman Arora ARORA@CS.JHU.EDU
Department of Computer Science, The Johns Hopkins University

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

Cristóbal Guzmán CRGUZMANP@MAT.UC.CL
Institute for Mathematical and Computational Engineering, Faculty of Mathematics and School of Engineering, Pontificia Universidad Católica de Chile

Editors: Claire Vernade and Daniel Hsu

Abstract

We study private empirical risk minimization (ERM) problem for losses satisfying the (γ, κ)-Kurdyka-Łojasiewicz (KL) condition, that is, the empirical loss F satisfies $F(w) - \min_w F(w) \leq \gamma^\kappa \|\nabla F(w)\|^\kappa$. The Polyak-Łojasiewicz (PL) condition is a special case of this condition when $\kappa = 2$. Specifically, we study this problem under the constraint of ρ zero-concentrated differential privacy (zCDP). When $\kappa \in [1, 2]$ and the loss function is Lipschitz and smooth over a sufficiently large region, we provide a new algorithm based on variance reduced gradient descent that achieves the rate $\tilde{O}\big(\big(\frac{\sqrt{d}}{n\sqrt{\rho}}\big)^\kappa\big)$ on the excess empirical risk, where n is the dataset size and d is the dimension. We further show that this rate is nearly optimal. When $\kappa \geq 2$ and the loss is instead Lipschitz and weakly convex, we show it is possible to achieve the rate $\tilde{O}\big(\big(\frac{\sqrt{d}}{n\sqrt{\rho}}\big)^\kappa\big)$ with a private implementation of the proximal point method. When the KL parameters are unknown, we provide a novel modification and analysis of the noisy gradient descent algorithm and show that this algorithm achieves a rate of $\tilde{O}\big(\big(\frac{\sqrt{d}}{n\sqrt{\rho}}\big)^{\frac{2\kappa}{4-\kappa}}\big)$ adaptively, which is nearly optimal when $\kappa = 2$. We further show that, without assuming the KL condition, the same gradient descent algorithm can achieve fast convergence to a stationary point when the gradient stays sufficiently large during the run of the algorithm. Specifically, we show that this algorithm can approximate stationary points of Lipschitz, smooth (and possibly nonconvex) objectives with rate as fast as $\tilde{O}\big(\frac{\sqrt{d}}{n\sqrt{\rho}}\big)$ and never worse than $\tilde{O}\big(\big(\frac{\sqrt{d}}{n\sqrt{\rho}}\big)^{1/2}\big)$. The latter rate matches the best known rate for methods that do not rely on variance reduction.

Keywords: Differential Privacy, KL Condition, PL Condition, Non-convex Optimization

© 2024 M. Menart, E. Ullah, R. Arora, R. Bassily & C. Guzmán.

1. Introduction

As modern machine learning techniques have increasingly relied on optimizing non-convex objectives, characterizing our ability to solve such problems has become increasingly important. Due to the inherent limitations of solving non-convex optimization problems, that is, the intractability of approximating global minimizers, work in this area has largely focused on approximating stationary points (Fang et al., 2018; Carmon et al., 2017; Ghadimi and Lan, 2013; Arjevani et al., 2022; Foster et al., 2019), or has imposed further restrictions on the loss function. In the latter camp, numerous possible assumptions have been proposed, such as the restricted secant inequality (Zhang and Yin, 2013) or star/quasar convexity (Hinder et al., 2020). Perhaps the most promising such condition is the Polyak-Łojasiewicz (PL) condition (Polyak, 1963), and its generalization, the Kurdyka-Łojasiewicz (KL) condition (Kurdyka, 1998)[1]. A function $F : \mathbb{R}^d \to \mathbb{R}$ satisfies the (γ, κ)-KL condition if for all $w \in \mathbb{R}^d$ it holds that,

$$F(w) - \min_w\{F(w)\} \leq \gamma^\kappa \left\|\nabla F(w; S)\right\|^\kappa \tag{1}$$

That is, the loss lower bounds the gradient norm. The PL condition is the special case where $\kappa = 2$. Both the KL and PL settings have been the subject of numerous works (Karimi et al., 2016; Foster et al., 2018; Scaman et al., 2022). The KL condition, in addition to being weaker than many of the previously mentioned conditions, has led to a number of strong convergence rate results. Furthermore, an increasingly rich literature has shown that overparameterized models such as neural networks satisfy the KL condition in a number of scenarios (Bassily et al., 2018; Charles and Papailiopoulos, 2018; Liu et al., 2020; Scaman et al., 2022).

On the other hand, the reliance of modern machine learning techniques on large datasets has caused growing concern over user privacy. Overparameterized models are of particular concern due to their ability to memorize training data (Sweeney, 2021; Carlini et al., 2019; Feldman and Zhang, 2020; Brown et al., 2021). In response to this concern, differential privacy (DP) has arisen as the most widely accepted method for ensuring the privacy of individuals present in a dataset. Unfortunately, it has been shown in a variety of settings that differentially private learning has fundamental limitations. As a result, characterizing these limitations has been the subject of numerous recent works.

Non-convex optimization under differential privacy is still not well understood. For example, in regards to the task of approximating stationary points in the DP setting, there are still gaps between existing upper and lower bounds (Arora et al., 2023). Furthermore, for the problem of approximating global minimizers of non-convex loss functions under DP, it has been shown the best possible rate is only $O\big(\frac{d}{n\epsilon}\big)$, even if the optimization algorithm is allowed exponential running time (Ganesh et al., 2023c). In the PL setting however, it has been shown that rates of $\tilde{O}\big(\frac{d}{n^2\epsilon^2}\big)$ on the excess empirical risk are achievable (Wang et al., 2017; Lowy et al., 2023). Interestingly, this matches the optimal rate for DP optimization in the (much more restrictive) *strongly convex* setting, and subsequently lower bounds for this setting show the rate is optimal (Bassily et al., 2014). Given this promising result, the question arises whether such results can be obtained for the more general class of objectives satisfying the KL condition, particularly since recent work has shown this generalization allows one to capture common models outside the reach of the PL condition Scaman et al.

1. These conditions are sometimes also referred to as the gradient domination condition. Further, the KL condition is sometimes phrased in terms of $h(F(w) - \min_w\{F(w)\})$, for some nondecreasing function h, akin to its first appearance (Lezanski, 1962). In our work we instead focus on the (commonly studied) case where h is a monomial.

(2022). In this work, we answer this question in the affirmative, and show that the KL assumption leads to fast rates under differential privacy, even in the absence of convexity. We further provide algorithms which are adaptive in the KL parameters. These results widen the range of non-convex models we can train effectively under DP.

1.1. Contributions

In this work, we develop the first algorithms for differentially private empirical risk minimization (ERM) under the (γ, κ)-KL condition without any convexity assumption. We show that for sufficiently smooth functions it is possible to achieve a rate of $\tilde{O}\big(\big(\frac{\sqrt{d}}{n\epsilon}\big)^{\kappa}\big)$ on the excess empirical risk for any $\kappa \in [1, 2]$. For $\kappa \geq 2$, we give an algorithm which attains the same rate for the strictly larger class of weakly convex functions. This rate is new for any $\kappa \neq 2$. We further show this rate is near optimal when $1 + \Omega(1) \leq \kappa \leq 2$ by leveraging existing lower bounds for convex functions satisfying the growth condition. For $1 \leq \kappa \leq 2$, we obtain our upper bound via a novel variant of the Spider algorithm, first proposed in Fang et al. (2018). This method allows us to leverage the reduced sensitivity of privatizing gradient *differences* to add less noise, an observation that has been leveraged in several other works studying the problem of finding stationary points under differential privacy (Arora et al., 2023; Murata and Suzuki, 2023). We also leverage a novel round structure (i.e. the number of steps before the gradient estimator is reset) for our private Spider algorithm. Whereas previous works have largely used fixed round lengths, our analysis crucially relies on variable round lengths with adaptive stopping. In the case where $\kappa \geq 2$, we obtain our upper bound using a differentially private implementation of the approximate proximal point method.

For both these algorithms, our analysis leverages the fact that the KL condition forces large gradients during the run of the algorithm. We further show that these larger gradient norms allow us to add more noise "for free", and thus better control the privacy budget. We use this observation to run Spider with a higher noise level than, for example, one would see without the KL condition (Arora et al., 2023).

Leveraging this intuition, we further develop a simple variant of noisy gradient descent that automatically scales the noise proportional to the gradient norm. We provide a novel analysis to show this algorithm achieves the rate $\tilde{O}\big(\big(\frac{\sqrt{d}}{n\epsilon}\big)^{\frac{2\kappa}{4-\kappa}} + \big(\frac{1}{n}\big)^{\kappa/2}\big)$ under the (γ, κ)-KL condition when $\kappa \in [1, 2]$. This rate is $\tilde{O}\big(\frac{d}{n^2\epsilon^2}\big)$ when $\kappa = 2$ (i.e. nearly optimal) and is $\tilde{O}\big(\big(\frac{\sqrt{d}}{n\epsilon}\big)^{2/3}\big)$ in the slowest regime ($\kappa = 1$). This result is adaptive and requires no prior knowledge of the KL parameters. We additionally prove that this same gradient descent algorithm can achieve fast convergence guarantees even when the KL condition does not hold. In this case where no KL assumption is made, we settle for convergence to a stationary point as approximating a global minimizer is intractable. Specifically, we show that when the trajectory of noisy gradient descent encounters mostly points with large gradient norm, the algorithm finds a point with gradient norm $\tilde{O}\big(\frac{\sqrt{d}}{n\epsilon}\big)$. We further establish that in the worst case, the algorithm finds a point with gradient norm at most $\tilde{O}\big(\big(\frac{\sqrt{d}}{n\epsilon}\big)^{1/2}\big)$, recovering the best known rate for noisy gradient descent in this setting.

1.2. Related Work

Differentially private optimization by now has a rich literature spanning over a decade. Much of this attention has been directed at the convex setting (Chaudhuri et al., 2011; Jain et al., 2012; Kifer et al., 2012; Bassily et al., 2014; Talwar et al., 2014, 2015; Bassily et al., 2019; Feldman et al.,

2020; Asi et al., 2021a; Bassily et al., 2021b). The study of differentially private optimization in the non-convex setting is comparatively newer, but has nonetheless been growing rapidly (Wang et al., 2017; Ganesh et al., 2023c; Arora et al., 2023; Ganesh et al., 2023b).

Currently, research into DP non-convex optimization under the KL condition specifically has been restricted to the special case of the PL condition. Assuming that the loss is Lipschitz, smooth, and satisfies the PL condition, the works Wang et al. (2017); Lowy et al. (2023) obtained the rate of $\tilde{O}\left(\frac{d}{n^2\epsilon^2}\right)$ on the excess empirical risk. This rate is optimal because of existing lower bounds for the strongly convex setting Bassily et al. (2014). More recently, Yang et al. (2022) studied the (more general) minmax optimization problems under differential privacy when the primal objective is assumed to be PL, although the rates therein are slower. Alternatively, in the *convex* setting, Asi et al. (2021b) characterized the optimal rates for DP optimization under an assumption known as the growth condition. When convexity is assumed, the KL condition and the growth condition are equivalent (Bolte et al., 2017, Theorem 5.2). Convex functions satisfying the growth condition are a strict subset of (general) KL functions.

There are also a number of works which have studied optimization under the KL condition without privacy considerations. The early works Polyak (1963); Lezanski (1962) were the first to show that for gradient descent, linear convergence rates are possible when the objective is smooth and satisfies the PL condition. More recently, Bassily et al. (2018) showed that under an additional assumption known as the interpolation condition, *stochastic* gradient descent also achieves linear convergence. The works Liu et al. (2020); Scaman et al. (2022) studied more general variants of the PL/KL conditions called the PL*/KL* conditions respectively. Specifically, these works study convergence when the condition holds only over a subset of $\mathbb{R}^d$.

2. Preliminaries

Empirical Risk Minimization: Let $\mathcal{X}$ be a data domain and let $S = \{x_1, ..., x_n\} \in \mathcal{X}^n$ be a dataset of n points. Let $f : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$ be a loss function and define the empirical risk/loss as $F(w; S) = \frac{1}{n}\sum_{i=1}^n f(w; x_i)$. We denote the set of global minimizers as $\mathcal{W}^* = \arg\min_w F(w)$, which we assume is nonempty. We assume we are given some starting point $w_0 \in \mathbb{R}^d$ and define the closest global minimizer to w_0 as w^*. That is $w^* = \arg\min_{w\in\mathcal{W}^*}\{\|w_0 - w\|\}$. As $\mathcal{W}^*$ may be non-convex, multiple such minimizers may exist, but it suffices to select one arbitrarily. We consider the problem of minimizing the excess empirical risk, defined at a point w as $F(w; S) - F(w^*; S)$. We assume throughout that f is L_0-Lipschitz continuous over some ball (to be defined later). We will denote the d-dimensional ball centered at w of radius B as $\mathcal{B}_B(w)$.

KL* Condition: Since assuming the loss satisfies the KL condition over all of $\mathbb{R}^d$ is unrealistic in practice (and indeed impossible if Lipschitzness is assumed), several works have proposed the modified KL* condition (Scaman et al., 2022; Liu et al., 2020). The exact definition of this condition varies. We use the following definition.

Definition 1 *A function $F : \mathbb{R}^d \to \mathbb{R}$ satisfies the (γ, κ)-KL* condition on $\mathcal{S} \subset \mathbb{R}^d$ w.r.t. $w' \in \mathbb{R}^d$ if $\forall w \in \mathcal{S}$ it holds that $\gamma^\kappa\|\nabla F(w)\|^\kappa \geq F(w) - F(w')$.*

We will take $w' = w^*$ (i.e. the closest global minimizer to w_0) unless otherwise stated. In this case, under the KL* condition, one equivalently has $\frac{1}{\gamma}(F(w) - F(w^*))^{1/\kappa} \leq \|\nabla F(w)\|$. Prior work studying the PL*/KL* condition has generally further assumed $F(w^*) = 0$, but we will avoid this assumption for the sake of generality (Liu et al., 2020; Scaman et al., 2022). We note that

the condition is often phrased so that the constant γ has no exponent, however this definition will ease notation in our analysis; a conversion to the standard definition is straightforward. For our algorithms, we will show that it is sufficient for the KL* condition to hold in a ball around an initial point w_0. Our guarantees could alternatively be phrased under the condition that the KL* assumptions holds in a ball around w^*, (see Remark 18, Appendix A).

Relevant to our discussion will also be the notion of the (λ, τ)-growth condition, which states that for any $w \in \mathbb{R}^d$, it holds that $F(w) - F(w^*) \geq \lambda^\tau \|w - w^*\|^\tau$. When the loss function is also assumed to be convex, the KL and growth conditions are in fact equivalent up to parameterization. See Appendix A for more details.

Loss bound: We assume throughout that one is given a bound $F_0 \geq 0$ such that $F(w_0; S) - F(w^*; S) \leq F_0$ for some $w_0 \in \mathbb{R}^d$. However, as our results will assume the KL condition holds at w_0, one always has the worst case bound $F_0 \leq \gamma^\kappa L_0^\kappa$ by the fact that the loss is L_0-Lipschitz.

Differential Privacy (DP): We consider primarily the notion of zero concentrated differential privacy (zCDP). For the purpose of referencing existing work, we also define approximate DP.

Definition 2 (ρ-zCDP (Bun and Steinke, 2016)) *An algorithm $\mathcal{A}$ is ρ-zCDP if for all datasets S and S' differing in one data point and all $\alpha \in (1, \infty)$, it holds that $D_\alpha(\mathcal{A}(S)||\mathcal{A}(S')) \leq \rho\alpha$, where D_α is the α-Rényi divergence.*

Definition 3 ((ϵ, δ)-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, $\mathbb{P}(\mathcal{A}(S) \in \mathcal{E}) \leq e^\varepsilon \mathbb{P}(\mathcal{A}(S') \in \mathcal{E}) + \delta$.*

zCDP guarantees imply approximate DP guarantees. Specifically, we note that for any $\delta > 0$ and $\epsilon \leq \sqrt{\log(1/\delta)}$, (ϵ, δ)-DP guarantees can be obtained from our results by setting $\rho = O\left(\epsilon^2 / \log(1/\delta)\right)$ (Bun and Steinke, 2016, Proposition 1.3).

Weak Convexity: A function $F : \mathbb{R}^d \to \mathbb{R}$ is $\tilde{L}_1$-weakly convex w.r.t. $\|\cdot\|$ if for all $0 \leq \lambda \leq 1$ and $w, v \in \mathbb{R}^d$ one has $f(\lambda w + (1-\lambda)v) \leq \lambda f(w) + (1-\lambda) f(v) + \frac{\tilde{L}_1 \lambda(1-\lambda)}{2} \|w - v\|^2$.

3. Optimal Algorithm for $1 \leq \kappa \leq 2$

Algorithm Overview Algorithm 1 is roughly an implementation of noisy Spider with some key differences. Similar to Spider, the algorithm runs over K rounds. At the start of any round k, a noisy minibatch gradient estimate $\nabla_{k,0}$, is computed. Throughout the rest of the round, the gradient is estimated using the change in the gradient between iterates. That is, for some $t \geq 0$, $\nabla_{k,t} = \nabla_{k,0} + \sum_{j=1}^t \Delta_{k,j}$, where each $\Delta_{k,j}$ corresponds to an estimate of a gradient difference. After each gradient estimate is obtained, a standard (normalized) gradient descent update step is performed.

In contrast to traditional Spider, at the start of each round $k \in [K]$, a target excess risk threshold, $\hat{\Phi}_k$, is set. The algorithm then uses this threshold to define an adaptive stopping mechanism for the round. The stopping condition is needed for the event where the excess risk of the update iterate falls below $\hat{\Phi}_k$ before the end of the phase. If this happens, the loss lower bound (and hence the gradient norm lower bound) will not be strong enough for the subsequent iterate. Consequently, the noise added for privacy could be too large and cause the trajectory of the algorithm to diverge. As such, we check to see if the loss has fallen below the target threshold before performing any update. We do

Algorithm 1 KL Spider

Require: Dataset $S = \{x_1, ..., x_n\}$, Privacy parameter $\rho > 0$, Failure probability $\beta > 0$, Initial point $w_0 \in \mathbb{R}^d$, Loss bound $F_0 \leq (L_0\gamma)^\kappa$, KL* parameters (γ, κ), Lipschitz parameter L_0, Smoothness parameter L_1

$w_{0,0} = w_0$, $\hat{\Phi}_0 = F_0$

$c = 1 + F_0^{\frac{2-\kappa}{\kappa}} \frac{1}{64\gamma^2 L_1}$

$K = (1 + 64(1/F_0)^{\frac{2-\kappa}{\kappa}}\gamma^2 L_1)\Big[\log(F_0) + \kappa \log\left(\frac{n\sqrt{\rho}}{\gamma L_0\sqrt{d}}\right)\Big]$, $\beta' = \frac{\beta}{K}\left(\frac{\gamma L_0\sqrt{Kd}}{n\sqrt{\rho}F_0^{1/\kappa}}\right)^{2-\kappa}$

$\hat{\sigma} = \frac{L_0\sqrt{K}}{n\sqrt{\rho}}$

for $k = 1, \ldots, K$ **do**

  $\hat{\Phi}_k = \max\left\{\frac{1}{c}\hat{\Phi}_{k-1},\ \min\left\{\left(\frac{32\gamma L_0\sqrt{Kd\log(1/\beta')}}{n\sqrt{\rho}}\right)^\kappa, F_0\right\}\right\}$

  $T_k = (F_0/\hat{\Phi}_k)^{\frac{2-\kappa}{\kappa}}$, $\sigma_k = \frac{\hat{\Phi}_k^{1/\kappa}\sqrt{T_k K}}{\gamma n\sqrt{\rho}}$

  $\nabla_{k,0} = \frac{1}{n}\sum_{i=1}^n \nabla f(w_{k,0}; x_i) + b_{k,0}$ where $b_{k,0} \sim \mathcal{N}(0, \mathbb{I}_d\hat{\sigma}^2)$

  $t = 0$

  while $t \leq T_k$ and $\|\nabla_{t,k}\| \geq \frac{7}{8\gamma}\hat{\Phi}_k^{1/\kappa}$ **do**

    $\eta_{k,t} = \frac{1}{4\gamma L_1\|\nabla_{k,t}\|}\hat{\Phi}_k^{1/\kappa}$

    $w_{k,t+1} = w_{k,t} - \eta_{k,t}\nabla_{k,t}$

    $\Delta_{k,t+1} = \frac{1}{n}\sum_{i=1}^n[\nabla f(w_{k,t+1}; x_i) - \nabla f(w_{k,t}; x_i)] + b_{k,t+1}$, where $b_{k,t+1} \sim \mathcal{N}(0, \mathbb{I}_d\sigma_k^2)$

    $\nabla_{k,t+1} = \nabla_{k,t} + \Delta_{k,t+1}$

    $t = t + 1$

  end while

  $w_{k+1,0} = w_{k,t-1}$

end for

Return $\bar{w} = w_{K+1,0}$

this indirectly by checking the gradient norm and using the KL condition, as bounding the sensitivity of the loss itself (to ensure privacy) is more delicate. Our implementation also uses varying phase lengths such that the length of the k'th phase is roughly $(1/\hat{\Phi}_k)^{(2-\kappa)/\kappa}$ (note the exponent is non-negative since $\kappa \leq 2$). Specifically, the phases get longer as the algorithm progresses. This is due to the fact that as the excess risk decreases, the lower bound on the gradient norm (induced by the KL condition) becomes weaker, leading to progressively slower convergence. We have the following guarantee on the Algorithm.

Theorem 4 *Let $\gamma > 0$, $\kappa \in [1, 2]$. There exists $B = \tilde{O}\big(\frac{F_0^{1/\kappa}}{\gamma L_1} + F_0^{\frac{\kappa-1}{\kappa}}\gamma\big)$ such that if f is L_0-Lipschitz and L_1-smooth over $\mathcal{B}_B(w_0)$, Algorithm 1 is ρ-zCDP. Further, if $F(\cdot; S)$ satisfies the (γ, κ)-KL* condition over $\mathcal{B}_B(w_0)$, with probability at least $1-\beta$ the output of Algorithm 1 satisfies*

$$F(\bar{w}; S) - F(w^*; S) = O\left(\left(\frac{\gamma L_0\sqrt{dK\log(1/\beta')}}{n\sqrt{\rho}}\right)^\kappa\right) = \tilde{O}\left(\left(\frac{\gamma L_0\sqrt{d}\sqrt{1 + (1/F_0)^{\frac{2-\kappa}{\kappa}}\gamma^2 L_1}}{n\sqrt{\rho}}\right)^\kappa\right),$$

where K, β' are as defined in Algorithm 1, namely $K = (1 + 64(1/F_0)^{\frac{2-\kappa}{\kappa}}\gamma^2 L_1)\Big[\log(F_0) + \kappa \log\left(\frac{n\sqrt{\rho}}{\gamma L_0\sqrt{d}}\right)\Big]$, *and* $\beta' = \frac{\beta}{K}\left(\frac{\gamma L_0\sqrt{Kd}}{n\sqrt{\rho}F_0^{1/\kappa}}\right)^{2-\kappa}$.

Note the result can be further simplified by setting $F_0 = (L_0\gamma)^\kappa$ (which is always possible by the KL condition) which makes $(1/F_0)^{\frac{2-\kappa}{\kappa}}\gamma^2 L_1 = \frac{\gamma^\kappa L_1}{L_0^{2-\kappa}}$. We defer the proof of privacy to Appendix B.1, as it is a standard application of the privacy guarantees of the Gaussian mechanism and composition. In the following, we focus on proving the convergence guarantee of the algorithm.

Convergence Proof for Algorithm 1 Our ability to assume loss properties hold only over $\mathcal{B}_B(w_0)$ (rather than $\mathbb{R}^d$) hinges on bounding the trajectory of the algorithm. We assume for the following lemmas that he conditions of Theorem 4 hold.

Lemma 5 *For any $k \in [K]$ and $t \in [T_k]$ corresponding to iterates of Algorithm 1, it holds with probability 1 that $w_{k,t} \in \mathcal{B}_B(w_0)$ for some $B = \tilde{O}\big(\frac{F_0^{1/\kappa}}{\gamma L_1} + F_0^{\frac{\kappa-1}{\kappa}}\gamma\big)$.*

The implication of this result is that the algorithm starts in, and never leaves the KL region around w_0. Thus the KL property holds at every iterate of the algorithm. We provide a proof in Appendix B.3. Note that any L_1-smooth function is also L_1'-smooth for $L_1' > L_1$. Thus the $\frac{F_0^{1/\kappa}}{\gamma L_1}$ term in the distance bound can be made negligible by running the algorithm with $L_1 \geq F_0^{(2-\kappa)/2}/\gamma$ (although this may increase the rate depending on F_0 and γ).

Our utility proof for Algorithm 1, will crucially rely on the following lemma which bounds the gradient error at any step in terms of the excess risk target, $\hat{\Phi}_k$.

Lemma 6 *With probability at least $1-\beta$, for every $k \in [K]$ and $t \in [T_k]$ indexing the iterates of the algorithm, one has that $\|\nabla_{k,t} - \nabla F(w_{k,t}; S)\| \leq \frac{1}{8\gamma}\hat{\Phi}_k^{1/\kappa}$.*

Proof The gradient estimates are generated by using exact gradients plus Gaussian noise, thus

$$\|\nabla_{k,t} - \nabla F(w_{k,t}; S)\|^2$$
$$= \Big\|\nabla F(w_{k,0}; S) + b_{k,0} + \sum_{j=1}^{t}\big[\nabla F(w_{k,j}; S) - \nabla F(w_{k,j-1}; S) + b_{k,j}\big] - \nabla F(w_{k,t}; S)\Big\|^2 = \Big\|\sum_{j=0}^{t} b_{k,j}\Big\|^2.$$

We can use Gaussian concentration results, see (Jin et al., 2019, Lemma 2), to conclude that for any $\tau \geq 0$, $\mathbb{P}\left[\|\nabla_{k,t} - \nabla F(w_{k,t}; S)\| \geq \tau\right] \leq 2\exp\left(-\frac{\tau^2}{2d(\hat{\sigma}^2 + \sum_{j=1}^{T_k}\sigma_k^2)}\right)$. Thus, under the settings of $\hat{\sigma} = \frac{L_0\sqrt{K}}{n\sqrt{\rho}}$ and $\sigma_k = \frac{\hat{\Phi}_k^{1/\kappa}\sqrt{T_k K}}{\gamma n\sqrt{\rho}}$ and $T_k = (F_0/\hat{\Phi}_k)^{\frac{2-\kappa}{\kappa}}$, one has that with probability at least $1-\beta'$ that:

$$\begin{aligned}
\|\nabla_{k,t} - \nabla F(w_{k,t}; S)\| &\leq 2\sqrt{d\log(1/\beta')}(\hat{\sigma} + \sqrt{T_k}\sigma_k) \\
&= \frac{2L_0\sqrt{Kd\log(1/\beta')}}{n\sqrt{\rho}} + \frac{2\sqrt{Kd\log(1/\beta')}}{\gamma n\sqrt{\rho}}\hat{\Phi}_k^{\frac{\kappa-1}{\kappa}}F_0^{\frac{2-\kappa}{\kappa}} \\
&\overset{(i)}{\leq} \frac{2L_0\sqrt{Kd\log(1/\beta')}}{n\sqrt{\rho}} + \frac{2\sqrt{Kd\log(1/\beta')}}{\gamma n\sqrt{\rho}}F_0^{1/\kappa} \\
&\overset{(ii)}{\leq} \frac{4L_0\sqrt{Kd\log(1/\beta')}}{n\sqrt{\rho}} \overset{(iii)}{\leq} \frac{1}{8\gamma}\hat{\Phi}_k^{1/\kappa}.
\end{aligned}$$

Above, (i) uses $\hat{\Phi}_k \leq F_0$. Step (ii) uses that $F_0 \leq (\gamma L_0)^\kappa$ by the KL condition and Lipschitzness. Step (iii) uses the fact that $\hat{\Phi}_k \geq \left(\frac{32\gamma L_0\sqrt{Kd\log(1/\beta')}}{n\sqrt{\rho}}\right)^\kappa$.

Finally, we observe that for all $k \in [K]$, $\hat{\Phi}_k \geq \left(\frac{32\gamma L_0\sqrt{Kd\log(1/\beta')}}{n\sqrt{\rho}}\right)^\kappa$ and $\frac{2-\kappa}{2} \geq 0$. Hence, the total number of iterations of the algorithm is at most

$$\sum_{k=1}^{K} T_k \leq K\left(F_0\left(\frac{n\sqrt{\rho}}{32\gamma L_0\sqrt{Kd\log(1/\beta')}}\right)^\kappa\right)^{\frac{2-\kappa}{\kappa}} \leq K\left(\frac{n\sqrt{\rho}F_0^{1/\kappa}}{\gamma L_0\sqrt{Kd}}\right)^{2-\kappa}$$

Thus by the definition of β', over the run of the algorithm, we have with probability at least $1-\beta$ that every gradient estimate satisfies the desired error bound. ■

We can now prove the main theorem.

Proof [Proof of Theorem 4] In the following, we condition on the high probability event that the gradient errors are bounded, as shown in Lemma 6. Further, recall that by Lemma 5 the trajectory $\{w_{k,t}\}_{k\in[K],t\in[T_k]}$ is contained in $\mathcal{B}_B(w_0)$ with probability 1, and that the (γ,κ)-KL* conditions holds over this set.

We will show that at the end of the the k'th phase (i.e. the k'th iteration of the outer loop), the excess risk is at most $\hat{\Phi}_k$. First, consider the case where at some point during the phase the gradient norm stopping condition is reached. In this case, the condition in the while loop ensures $\|\nabla_{k,t}\| \leq \frac{7}{8\gamma}\hat{\Phi}_k^{1/\kappa}$. Thus by Lemma 6 and a triangle inequality we have $\|\nabla F(w_{k,t};S)\| \leq \frac{7}{8\gamma}\hat{\Phi}_k^{1/\kappa} + \frac{1}{8\gamma}\hat{\Phi}_k^{1/\kappa} = \frac{1}{\gamma}\hat{\Phi}_k^{1/\kappa}$. Then by the KL assumption we have that $F(w_{k,t};S) - F(w^*;S) \leq \gamma^\kappa\|\nabla F(w_{k,t};S)\|^\kappa \leq \gamma^\kappa(\frac{1}{\gamma}\hat{\Phi}_k^{1/\kappa})^\kappa \leq \hat{\Phi}_k$, as desired.

We thus turn towards analyzing the alternative case, where the final iterate of the phase is w_{k,T_k}, using an induction argument. Specifically, under the inductive assumption that $F(w_{k,0};S) - F(w^*;S) \leq \hat{\Phi}_{k-1}$, we will show that $F(w_{k,T_k};S) - F(w^*;S) \leq \hat{\Phi}_k$. For the base case, we clearly have $F(w_{0,0};S) - F(w^*;S) \leq \hat{\Phi}_0 = F_0$. Using smoothness and the setting of $\eta_{k,t}$, we can obtain the following descent inequality,

$$F(w_{k,t};S) - F(w_{k,t+1};S) \geq \frac{1}{16\gamma L_1}\|\nabla_{k,t}\|\hat{\Phi}_k^{1/\kappa} - \frac{1}{4L_1}\|\nabla_{k,t} - \nabla F(w_{k,t};S)\|^2$$

We leave the derivation of the above inequality to Lemma 20 in Appendix B.2. We now can use the fact that updates are only performed when $\|\nabla_{k,t}\| \geq \frac{7}{8\gamma}\hat{\Phi}_k^{1/\kappa}$ and the bound on the gradient estimate error derived in Lemma 6 to obtain

$$F(w_{k,t};S) - F(w_{k,t+1};S) \geq \frac{1}{32\gamma^2 L_1}\hat{\Phi}_k^{2/\kappa} - \frac{1}{256\gamma^2 L_1}\hat{\Phi}_k^{2/\kappa} \geq \frac{1}{64\gamma^2 L_1}\hat{\Phi}_k^{2/\kappa}.$$

Summing over all $T_k = (F_0/\hat{\Phi}_k)^{\frac{\kappa-2}{2}}$ iterations yields

$$F(w_{k,0},S) - F(w_{k,T_k};S) \geq \frac{1}{64\gamma^2 L_1}T_k\hat{\Phi}_k^{2/\kappa} = \frac{1}{64\gamma^2 L_1}F_0^{\frac{2-\kappa}{\kappa}}\hat{\Phi}_k.$$

We then have the following manipulation leveraging the inductive hypothesis,

$$\begin{aligned} F(w_{k,0};S) - F(w^*;S) + F(w^*;S) - F(w_{k,T_k};S) &\geq \frac{1}{64\gamma^2 L_1} F_0^{\frac{2-\kappa}{\kappa}} \hat{\Phi}_k \\ \hat{\Phi}_{k-1} + F(w^*;S) - F(w_{k,T_k};S) &\geq \frac{1}{64\gamma^2 L_1 c} F_0^{\frac{2-\kappa}{\kappa}} \hat{\Phi}_{k-1} \\ \left(1 - F_0^{\frac{2-\kappa}{\kappa}} \frac{1}{64\gamma^2 L_1 c}\right)\hat{\Phi}_{k-1} &\geq F(w_{k,T_k};S) - F(w^*;S) \\ \hat{\Phi}_k &\geq F(w_{k,T_k};S) - F(w^*;S). \end{aligned}$$

The last step follows because $\left(1 - F_0^{\frac{2-\kappa}{\kappa}} \frac{1}{64\gamma^2 L_1 c}\right) = \frac{1}{c}$ and $\frac{1}{c}\hat{\Phi}_k \leq \hat{\Phi}_{k-1}$. We have now shown that final iterate of each phase has excess risk at most $\hat{\Phi}_k$.

Now, all that remains is to show that $\hat{\Phi}_K \leq \left(\frac{32\gamma L_0\sqrt{Kd\log(1/\beta')}}{n\sqrt{\rho}}\right)^\kappa$. Noting that $\hat{\Phi}_K \leq \max\left\{\left(\frac{1}{c}\right)^K F_0, \left(\frac{32\gamma L_0\sqrt{Kd\log(1/\beta')}}{n\sqrt{\rho}}\right)^\kappa\right\}$ it suffices to show that $\left(\frac{1}{c}\right)^K F_0 \leq \left(\frac{\gamma L_0\sqrt{d}}{n\sqrt{\rho}}\right)^\kappa$. The inequality $\left(\frac{1}{c}\right)^K F_0 \leq \left(\frac{\gamma L_0\sqrt{d}}{n\sqrt{\rho}}\right)^\kappa$ is equivalent to $\frac{\log(F_0)+\kappa\log(\frac{n\sqrt{\rho}}{\gamma L_0\sqrt{d}})}{\log(c)} \leq K$. Using the fact that $\log(1+x) \geq \frac{x}{1+x}$ for $x \geq 0$, we can obtain that $\log(c) = \log(1 + 1/[64F_0^{\frac{\kappa-2}{\kappa}}\gamma^2 L_1]) \geq (1 + 64F_0^{\frac{\kappa-2}{\kappa}}\gamma^2 L_1)^{-1}$. It thus suffices to have $K \geq (1+64(1/F_0)^{\frac{2-\kappa}{\kappa}}\gamma^2 L_1)\left[\log(F_0)+\kappa\log\left(\frac{n\sqrt{\rho}}{\gamma L_0\sqrt{d}}\right)\right]$, which is satisfied by the algorithm. ■

3.1. Lower Bound

We now demonstrate a lower bound showing that our upper bound is nearly optimal. To do this, we leverage an existing lower bound from Asi et al. (2021b) for functions exhibiting the growth condition. In Theorem 21 in Appendix B.4, we extend their result to smooth functions and give a lower bound of $\Omega\big((\tau)^{\frac{-1}{\tau-1}}\left(\frac{L_0\sqrt{d}}{\lambda n\epsilon}\right)^{\frac{\tau}{\tau-1}}\big)$ on excess empirical risk of (ϵ,δ)-DP procedures for convex functions satisfying (λ,τ)-growth. Combining this result with the fact that the $(\frac{1}{\gamma},\frac{\kappa}{\kappa-1})$-growth condition and convexity implies the (γ,κ)-KL condition (Theorem 5.2 (ii) in Bolte et al. (2017), restated as Lemma 16), yields the following lower bound.

Corollary 7 *Let $B, L_0, L_1 > 0$ and $1 < \kappa \leq 2$ such that $\kappa = 1+\Omega(1)$. For any ρ-zCDP algorithm, $\mathcal{A}$, there exists a dataset, S, point $w_0 \in \mathbb{R}^d$, and loss function f such that f is L_0-Lipschitz and L_1-smooth over $\mathcal{B}_B(w_0)$ and $F(\cdot;S)$ is (γ,κ)-KL, for which the output of $\mathcal{A}$ has expected excess empirical risk $\tilde{\Omega}\left(\left(\frac{\gamma L_0\sqrt{d}}{n\sqrt{\rho}}\right)^\kappa\right)$.*

Note the bound is independent of B and L_1. More details on how to obtain Corollary 7 from the result of Asi et al. (2021b) can be found in Appendix B.5.

4. Algorithm for $\kappa \geq 2$

In this section, we assume the loss $F(\cdot;S)$ is $\tilde{L}_1$-weakly convex and that the empirical loss satisfies the (γ,κ)-KL* condition for $\kappa \geq 2$. We avoid making a smoothness assumption in this

Algorithm 2 (KL) Proximal Point Method

Require: Dataset S, Privacy parameter ρ, zCDP Optimizer for SC loss $\mathcal{A}$, Initial point $w_0 \in \mathbb{R}^d$, Initial loss bound, $F_0 \geq 0$, Failure probability β, Lipschitz parameter L_0, Weak convexity $\tilde{L}_1$

$T = (1 + 32F_0^{\frac{\kappa-2}{\kappa}}\gamma^2\tilde{L}_1)\Big[\log(F_0) + \kappa\log\left(\frac{n\sqrt{\rho}}{\gamma L_0\sqrt{d}}\right)\Big]$

$\beta' = \frac{\beta}{T}$

for $t = 1 \dots T$ **do**

  $F_t(w; S) := F(w; S) + \tilde{L}_1 \|w - w_{t-1}\|^2$

  $w_t = \mathcal{A}(F_t, w_{t-1}, \frac{\rho}{T}, \beta')$

end for

Return w_T

regime. When $\kappa > 2$ and the KL* condition holds in a region with small excess risk, the loss functions cannot be smooth (unless it is the constant function). To elaborate, one can show that the loss upper bound implied by smoothness and the loss lower bound implied by the KL* condition lead to a contradiction. Instead of smoothness, we consider the (strictly weaker) assumption of weak convexity. As convex functions are weakly convex with $\tilde{L}_1 = 0$, this setting is a strict relaxation of the loss assumptions considered by Asi et al. (2021b). Despite this, we achieve essentially the same rate as theirs. Moreover, in Theorem 23 in Appendix C, we give a lower bound of $\left(\frac{1}{n\epsilon}\right)^\kappa$, which establishes that our rate is tight (at least) for $d = 1$. Its proof adapts the construction in Asi et al. (2021b, Theorem 5) from a lower bound on excess population risk under pure, $(\epsilon, 0)$-DP to that on excess empirical risk under approximate, (ϵ, δ)-DP. The lower bound holds for convex functions satisfying the growth condition and thus satisfying the KL condition, via Lemma 16.

Our algorithm in this case is simply a differentially-private implementation of the approximate proximal point method. This method has been used in prior work for non-KL functions to approximate stationary points (Davis and Grimmer, 2019; Davis and Drusvyatskiy, 2019b; Bassily et al., 2021a). We have the following guarantee for this method.

Theorem 8 *Let $\gamma > 0$, $\kappa \geq 2$, There exists $B = \tilde{O}\big(\frac{L_0}{\tilde{L}_1}(1 + \frac{\sqrt{Td\log\left(n^2\log^2\left(1/\beta'\right)/d\beta'\right)}}{n\sqrt{\rho}}) + L_0F_0^{\frac{\kappa-2}{\kappa}}\gamma^2\big)$ and a subroutine $\mathcal{A}$ such that if f is Lipschitz then Algorithm 2 is ρ-zCDP. If $F(\cdot; S)$ also satisfies the (γ, κ)-KL* condition and $\tilde{L}_1$-weak convexity over $\mathcal{B}_B(w_0)$, then with probability at least $1 - \beta$ the output of Algorithm 2 has excess risk, $F(w_T; S) - F(w^*; S)$, at most*

$$O\left(\left(\frac{\gamma L_0\sqrt{Td\log(n^2\log^2\left(1/\beta'\right)/d\beta')}}{n\sqrt{\rho}}\right)^\kappa\right) = \tilde{O}\left(\left(\frac{\gamma L_0\sqrt{d(1 + F_0^{\frac{\kappa-2}{\kappa}}\gamma^2\tilde{L}_1)}}{n\sqrt{\rho}}\right)^\kappa\right).$$

where $T = (1 + 32F_0^{\frac{\kappa-2}{\kappa}}\gamma^2\tilde{L}_1)\Big[\log(F_0) + \kappa\log\left(\frac{n\sqrt{\rho}}{\gamma L_0\sqrt{d}}\right)\Big]$ and $\beta' = \beta/T$, as defined in Algorithm 2.

Note the term $\frac{\sqrt{Td\log\left(n^2\log^2\left(1/\beta'\right)/d\beta'\right)}}{n\sqrt{\rho}}$ in the radius bound will be $o(1)$ in regime where the convergence guarantees are nontrivial. The privacy of Algorithm 2 is straightforward since the subroutine $\mathcal{A}$ is ρ-zCDP by the assumption. Algorithm 2 is then private by post processing and composition. To prove the convergence result, we will use the following fact about the strength of differentially private optimizers for strongly convex loss functions.

Lemma 9 *There exists an implementation of $\mathcal{A}$ which is ρ-zCDP and with probability at least $1-\beta'$ the output of the algorithm has excess risk $O\left(\frac{L_0^2 d \log\left(n^2 \log^2\left(1/\beta'\right)/d\beta'\right)}{\tilde{L}_1 n^2 \rho}\right)$.*

We provide the details for this result in Appendix E. Furthermore, as in Section 3, we only need the KL condition to hold over the trajectory of the algorithm. The following lemma allows us to utilize the KL property at every iterate generated by the algorithm.

Lemma 10 *Assume $\mathcal{A}$ is as described by Lemma 9 above. With probability at least $1-\beta$, $w_1, ..., w_T \in \mathcal{B}_B(w_0)$ for some $B = \tilde{O}\big(\frac{L_0}{\tilde{L}_1}\big(1 + \frac{\sqrt{Td\log\left(n^2\log^2\left(1/\beta'\right)/d\beta'\right)}}{n\sqrt{\rho}}\big) + L_0 F_0^{\frac{\kappa-2}{\kappa}}\gamma^2\big)$.*

The proof is deferred to Appendix C.1. The proof of Theorem 8 now follows similar steps to those used in Theorem 4, but is overall much simpler. One key difference is that, for each $t \in [T]$, we need to use the KL condition to lower bound $\|w_t - w_{t-1}\|$, rather than $\|\nabla F(w_t; S)\|$. For this, note that the optimality conditions of F_t imply $2\tilde{L}_1\|w_t^* - w_{t-1}\| = \|\nabla F(w_t^*; S)\| \geq \frac{1}{\gamma}(F(w_t^*; S) - F(w^*; S))^{1/\kappa}$. The inequality comes from the KL condition. The full proof of Theorem 8 is in Appendix C.2.

5. Adapting to KL condition

In this section, we present an alternative algorithm for ERM under the KL* condition. At the cost of weaker rates when $\kappa < 2$, our algorithm automatically adapts to κ and γ. This is in contrast to the Spider method presented previously which requires prior knowledge of κ and γ. Furthermore, we are able to obtain this result with a comparatively simple algorithm. That is, our algorithm is a simple modification of the traditional noisy gradient descent algorithm seen frequently in the DP literature (Bassily et al., 2014; Wang et al., 2017; Bassily et al., 2019). Throughout the following,

Algorithm 3 Adaptive Noisy Gradient Descent

Require: Dataset S, Privacy parameter $\rho > 0$, Probability $\beta > 0$, Initial point $w_0 \in \mathbb{R}^d$, Lipschitz parameter L_0, Smoothness parameter L_1

$\eta = \frac{1}{2L_1}$, $t = 0$, $\rho_0 = 0$

while $\sum_{j=0}^{t} \rho_t \leq \frac{\rho}{2}$ **do**

  $N_t = \left\|\frac{1}{n}\sum_{i=1}^n \nabla f(w_t; x_i)\right\| + \hat{b}_t$ where $\hat{b}_t \sim \mathcal{N}(0, \mathbb{I}_d \hat{\sigma}^2)$ and $\hat{\sigma} = \frac{L_0}{\sqrt{n}\rho^{1/4}}$

  $\nabla_t = \frac{1}{n}\sum_{i=1}^n \nabla f(w_t; x_i) + b_t$ where $b_t \sim \mathcal{N}(0, \mathbb{I}_d \sigma_t^2)$ and $\sigma_t = \max\left\{\frac{N_t}{\sqrt{d\log(n\sqrt{\rho}/\beta)}}, \frac{2L_0}{n\sqrt{\rho}}\right\}$

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

  $\rho_t = \min\left\{\frac{L_0^2 d\log(n\sqrt{\rho}/\beta)}{n^2 N_t^2}, \frac{\rho}{2}\right\} + \frac{\sqrt{\rho}}{n}$

  $t = t + 1$

end while

we will use $T+1$ to denote the highest value of t reached during the run of Algorithm 3.

Theorem 11 *Assume f is Lipschitz. Then running Algorithm 3 and releasing $w_0, ..., w_T$ is ρ-zCDP.*

The proof is given in Appendix D.2, and relies on the fully adaptive composition theorem of Whitehouse et al. (2022). Our aim is now to provide convergence guarantees when the loss satisfies the KL* condition over some region $\mathcal{S} \subset \mathbb{R}^d$. We here demonstrate an alternative way of

defining $\mathcal{S}$ which allows us to leverage the KL* condition (in contrast to assuming $\mathcal{S}$ is a ball). Define the threshold $\alpha = \max\left\{F(w_0;S), F(w^*;S) + 2(\gamma^{\kappa/2} + L_0)\left(\frac{L_0^2 \log(n\sqrt{\rho}/\beta)}{n\sqrt{\rho}}\right)^{1/2}\right\}$. Let $\mathcal{I} = \{w : F(w;S) \leq \alpha\}$ denote a lower level set of $F(\cdot;S)$. Note the second term in the max of α only handles the trivial case where w_0 already has small excess risk. Observe that $\mathcal{I}$ may not be a path-connected set, thus we define $\mathcal{S}$ as the path-connected component of $\mathcal{I}$ that contains w_0. That is, $w' \in \mathcal{S}$ if there exists a continuous function $\mathbf{w} : [0,1] \to \mathcal{I}$, such that $\mathbf{w}(0) = w_0$, $\mathbf{w}(1) = w'$. Intuitively, $\mathcal{S}$ is the local "valley" of $F(\cdot;S)$ in which w_0 resides. Furthermore, we can guarantee that the trajectory of Algorithm 3 stays in this valley for the duration of its run.

Lemma 12 *Assume $F(\cdot;S)$ is L_1-smooth and L_0-Lipschitz. If $F(\cdot;S)$ satisfies the (γ,κ)-KL* condition over $\mathcal{S}$ w.r.t. $w^*_{\mathcal{S}} := \arg\min_{w\in\mathcal{S}}\{F(w;S)\}$, then w.p. at least $1-2\beta$, for all $t \in [T]$, $w_t \in \mathcal{S}$.*

The proof is deferred to Appendix D.4. Note we are assuming the KL* condition w.r.t. the minimizer over $\mathcal{S}$ (as opposed to the global minimizer) here. We also remark that an existing work, Ganesh et al. (2023a), argued the importance of public pretraining in the non-convex setting to find some w_0 in a convex subregion before training on private data. Alternatively, our result suggests meaningful convergence if the empirical loss over the localized region is instead KL. This may be more realistic in the overparameterized regime as existing work has shown such models tend to be non-convex (but KL) around the minimizer (Liu et al., 2020). Our convergence result for Algorithm 3 is as follows.

Theorem 13 *Let $\beta, \gamma > 0$, $\kappa \in [1,2]$. Let $\rho \geq 0$ be s.t. $L_0^2 \log(n\sqrt{\rho}/\beta)/(L_1 n) \leq \sqrt{\rho}$. Define $p_{\mathsf{max}} := (1 + 8\gamma^2 L_1)\left[\log(F_0) + \frac{2\kappa}{4-\kappa}\log(n\sqrt{\rho}/[\gamma L_0])\right]$. If $F(\cdot;S)$ is L_1-smooth and L_0-Lipschitz and satisfies the (γ,κ)-KL* condition over $\mathcal{S}$ (as described above) w.r.t. $w^*_{\mathcal{S}}$, then with probability at least $1-2\beta$, Algorithm 3 finds w_T such that $F(w_T;S) - F(w^*_{\mathcal{S}};S)$ is at most*

$$O\left(\left(\frac{\gamma L_0\sqrt{d\log(n\sqrt{\rho}/\beta)p_{\mathsf{max}}}}{n\sqrt{\rho}}\right)^{\frac{2\kappa}{4-\kappa}} + \left(\frac{\max\{\gamma^2,1\}\, L_0^2 \log(n\sqrt{\rho}/\beta)}{\min\{L_1,1\}n\sqrt{\rho}}\right)^{\kappa/2} + \left(\frac{p_{\mathsf{max}}}{n\sqrt{\rho}}\right)^{\frac{\kappa}{2-\kappa}}\right).$$

Ignoring polylogarithmic terms and problem constants we can more simply write $F(w_T;S) - F(w^*_{\mathcal{S}};S) = \tilde{O}\big(\big(\frac{\sqrt{d}}{n\sqrt{\rho}}\big)^{\frac{2\kappa}{4-\kappa}} + \big(\frac{1}{n\sqrt{\rho}}\big)^{\kappa/2}\big)$.

The simplification in the theorem uses the fact that $\frac{\kappa}{2-\kappa} \geq \frac{\kappa}{2}$ for all κ. We defer the proof of Theorem 13 to Appendix D.3. The overarching ideas of the proof are similar to those of Theorem 4. However, the adaptive nature of the algorithm makes the analysis much more delicate.

Observe that for $\kappa = 2$ (i.e. the PL condition) this obtains the rate $\tilde{O}\big(\frac{d}{n^2\rho} + \frac{1}{n\sqrt{\rho}}\big)$ which essentially captures the optimal rate in this setting. The rate slows as κ decreases, and for $\kappa = 1$ we obtain a rate of $\tilde{O}\big(\big(\frac{\sqrt{d}}{n\sqrt{\rho}}\big)^{2/3} + \frac{1}{\sqrt{n}\rho^{1/4}}\big)$.

5.1. Convergence Guarantees without the KL Condition

One of the key properties of Algorithm 3 is that it leverages large gradients to better control the privacy budget. In fact, even in the absence of an explicit KL assumption, we can show that Algorithm 3 obtains strong convergence guarantees when large gradient norms are observed. We provide the following result on Adaptive Gradient Descent's ability to approximate stationary points. Note that we cannot give excess risk guarantees in this case due to the fact finding approximate global minimizers of non-convex functions is intractable in this setting.

Theorem 14 *Assume f is L_1-smooth and L_0-Lipschitz. Let $T+1$ denote the largest value attained by t during the run of Algorithm 3. Let t^* be sampled from $\{0, ..., T\}$ with probability proportional to $\exp\left(-\frac{n\sqrt{\rho}}{2L_0}\|\nabla F(w_t; S)\|\right)$. This algorithm is 2ρ-zCDP and with probability at least $1-3\beta$ satisfies $\|\nabla F(w_{t^*}; S)\| = O\left(\min\left\{\sqrt{\frac{F_0 L_1}{T}}, \left(\frac{L_0\sqrt{F_0 L_1 d}}{n\sqrt{\rho}}\right)^{1/2}\right\} + \frac{L_0\sqrt{\log(n\sqrt{\rho}/\beta)}}{\sqrt{n}\rho^{1/4}}\right)$.*

The proof is given in Appendix D.6. The best case scenario is when most gradients in the run of the algorithm are $\Omega(1)$. In this case, the algorithm attains $T = \tilde{\Theta}\big(\min\big\{n\sqrt{\rho}, \frac{n^2\rho}{d}\big\}\big)$ iterations and the convergence guarantee becomes $\tilde{O}\big(\frac{\sqrt{d}}{n\sqrt{\rho}} + \frac{1}{\sqrt{n}\rho^{1/4}}\big)$. We note an existing work showed a lower bound $\Omega\big(\frac{\sqrt{d}}{n\epsilon}\big)$ for approximating stationary points, although this is not directly comparable as the previously stated upper bound does not hold for all functions. In the worst case, the algorithm will achieve convergence guarantee $\tilde{O}\big(\frac{d^{1/4}}{\sqrt{n}\rho^{1/4}} + \frac{1}{\sqrt{n}\rho^{1/4}}\big)$. By contrast, the best known rate for approximating stationary points is $\tilde{O}\big(\big(\frac{\sqrt{d}}{n\sqrt{\rho}}\big)^{2/3}\big)$ (Arora et al., 2023), and the best known rate for methods which do not rely on variance reduced gradient estimates (as is more typical in practice) is $\tilde{O}\big(\big(\frac{\sqrt{d}}{n\sqrt{\rho}}\big)^{1/2}\big)$ (Wang et al., 2017). Our analysis recovers the $\tilde{O}\big(\big(\frac{\sqrt{d}}{n\sqrt{\rho}}\big)^{1/2}\big)$ rate obtained by noisy gradient descent as a worst case guarantee with minimal modification to the algorithm itself, while also potentially achieving a much stronger rate.

The worst case guarantee comes from balancing the number of iterations that the algorithm performs (which increases when the gradient norms are large) with the minimum gradient norm over the trajectory. For simplicity, consider the scenario where every gradient in the trajectory has the same norm $N > 0$. Then clearly the minimum norm is also N, but in this case $T = \tilde{O}\big(\frac{N^2 n^2 \rho}{d}\big)$. Thus the convergence guarantee implies that $N = \tilde{O}\big(\frac{\sqrt{d}}{Nn\sqrt{\rho}}\big)$, which at worst means $N = \tilde{O}\big(\frac{d^{1/4}}{\sqrt{n}\rho^{1/4}}\big)$. More formal/general details are in the proof in Appendix D.6.

Acknowledgments

RB's and MM's research is supported by NSF CAREER Award 2144532, NSF Award AF-1908281, and NSF Award 2112471. RA's and EU's research is supported, in part, by NSF BIGDATA award IIS-1838139 and NSF CAREER award IIS-1943251. CG'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

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. 199(1–2):165–214, jun 2022. ISSN 0025-5610. doi: 10.1007/s10107-022-01822-7. URL https://doi.org/10.1007/s10107-022-01822-7.

Raman Arora, Raef Bassily, Tomás González, Cristóbal A Guzmán, Michael Menart, and Enayat Ullah. Faster rates of convergence to stationary points in differentially private optimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and

Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 1060–1092. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/arora23a.html.

Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in l1 geometry. In *International Conference on Machine Learning*, pages 393–403. PMLR, 2021a.

Hilal Asi, Daniel Levy, and John C Duchi. Adapting to function difficulty and growth conditions in private optimization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 19069–19081. Curran Associates, Inc., 2021b. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/9f820adf84bf8a1c259f464ba89ea11f-Paper.pdf.

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

Raef Bassily, Mikhail Belkin, and Siyuan Ma. On exponential convergence of sgd in non-convex over-parametrized learning. *ArXiv*, abs/1811.02564, 2018. URL https://api.semanticscholar.org/CorpusID:53232028.

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

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

Raef Bassily, Cristobal Guzman, and Anupama Nandi. Non-euclidean differentially private stochastic convex optimization. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 474–499. PMLR, 15–19 Aug 2021b. URL https://proceedings.mlr.press/v134/bassily21a.html.

Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165:471–507, 2017.

Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. When is memorization of irrelevant training data necessary for high-accuracy learning? In *Proceedings of the 53rd annual ACM SIGACT symposium on theory of computing*, pages 123–132, 2021.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In Martin Hirt and Adam Smith, editors, *Theory of Cryptography*, pages 635–658, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg. ISBN 978-3-662-53641-4.

Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. STOC '14, page 1–10, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327107. doi: 10.1145/2591796.2591877. URL https://doi.org/10.1145/2591796.2591877.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, 2019.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. "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, page 654–663. JMLR.org, 2017.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 745–754. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/charles18a.html.

Kamalika Chaudhuri and Daniel Hsu. Convergence rates for differentially private statistical estimation. In *Proceedings of the... International Conference on Machine Learning. International Conference on Machine Learning*, volume 2012, page 1327. NIH Public Access, 2012.

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

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019a. doi: 10.1137/18M1178244. URL https://doi.org/10.1137/18M1178244.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.*, 29(1):207–239, 2019b. doi: 10.1137/18M1178244. URL https://doi.org/10.1137/18M1178244.

Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019. doi: 10.1137/17M1151031. URL https://doi.org/10.1137/17M1151031.

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

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal nonconvex optimization via stochastic path-integrated differential estimator. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/1543843a4723ed2ab08e18053ae6dc5b-Paper.pdf.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, page 439–449, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369794. doi: 10.1145/3357713.3384335. URL https://doi.org/10.1145/3357713.3384335.

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

Dylan J Foster, Ayush Sekhari, Ohad Shamir, Nathan Srebro, Karthik Sridharan, and Blake Woodworth. The complexity of making the gradient small in stochastic convex optimization. In *Conference on Learning Theory*, pages 1319–1345. PMLR, 2019.

Arun Ganesh, Mahdi Haghifam, Milad Nasr, Sewoong Oh, Thomas Steinke, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Why is public pretraining necessary for private model training? In *International Conference on Machine Learning*, 2023a. URL https://api.semanticscholar.org/CorpusID:257038349.

Arun Ganesh, Daogao Liu, Sewoong Oh, and Abhradeep Thakurta. Private (stochastic) non-convex optimization revisited: Second-order stationary points and excess risks, 2023b. URL https://arxiv.org/abs/2302.09699.

Arun Ganesh, Abhradeep Thakurta, and Jalaj Upadhyay. Universality of langevin diffusion for private optimization, with applications to sampling from rashomon sets. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1730–1773. PMLR, 2023c.

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

Nicholas JA Harvey, Christopher Liaw, and Sikander Randhawa. Simple and optimal high-probability bounds for strongly-convex stochastic gradient descent. *arXiv preprint arXiv:1909.00843*, 2019.

Oliver Hinder, Aaron Sidford, and Nimit Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1894–1938. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/hinder20a.html.

Prateek Jain, Pravesh Kothari, and Abhradeep Thakurta. Differentially private online learning. In *25th Annual Conference on Learning Theory (COLT)*, pages 24.1–24.34, 2012.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In Paolo Frasconi, Niels Landwehr, Giuseppe Manco, and Jilles Vreeken, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 795–811, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46128-1.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1, 2012.

Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'Institut Fourier*, 48:769–783, 1998. URL https://api.semanticscholar.org/CorpusID:3751297.

T. Lezanski. Über das minimumproblem von funktionalen in banachschen räumen. *Bull. Acad. Pol. Sci.*, 1962.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning. *CoRR*, abs/2003.00307, 2020. URL https://arxiv.org/abs/2003.00307.

Andrew Lowy, Ali Ghafelebashi, and Meisam Razaviyayn. Private non-convex federated learning without a trusted server. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5749–5786. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/lowy23a.html.

Tomoya Murata and Taiji Suzuki. DIFF2: differential private optimization via gradient differences for nonconvex distributed learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 25523–25548. PMLR, 2023. URL https://proceedings.mlr.press/v202/murata23b.html.

Boris Polyak. Gradient methods for the minimisation of functionals. *Ussr Computational Mathematics and Mathematical Physics*, 3:864–878, 12 1963. doi: 10.1016/0041-5553(63)90382-3.

Kevin Scaman, Cedric Malherbe, and Ludovic Dos Santos. Convergence rates of non-convex stochastic gradient descent under a generic lojasiewicz condition and local smoothness. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19310–19327. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/scaman22a.html.

Latanya Sweeney. Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine and Ethics*, 25(2-3):98–110, 2021. doi: 10.1111/j.1748-720X.1997.tb01885.x.

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

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

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

Justine Whitehouse, Aaditya Ramdas, Ryan M. Rogers, and Zhiwei Steven Wu. Fully adaptive composition in differential privacy. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:247362627.

Zhenhuan Yang, Shu Hu, Yunwen Lei, Kush R Vashney, Siwei Lyu, and Yiming Ying. Differentially private sgda for minimax problems. In James Cussens and Kun Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 2192–2202. PMLR, 01–05 Aug 2022. URL https://proceedings.mlr.press/v180/yang22a.html.

Hui Zhang and Wotao Yin. Gradient methods for convex minimization: better rates under weaker conditions. *ArXiv*, abs/1303.4645, 2013. URL https://api.semanticscholar.org/CorpusID:2499404.

Appendix A. Relationship between Growth Condition and KL Condition

Definition 15 ((λ, τ)-growth) *A function $F : \mathbb{R}^d \to \mathbb{R}$ satisfies (λ, τ)-growth if the set of minimizers $\mathcal{W}^* := \arg\min_w F(w)$ is non-empty, and*

$$F(w) - F(w_p) \geq \lambda^\tau \|w - w_p\|^\tau$$

where w_p be the projection of w onto $\mathcal{W}^*$.

Lemma 16 (Theorem 5.2 (ii) in Bolte et al. (2017)) *Let $\kappa \geq 1$ and $\gamma > 0$. If $F : \mathbb{R}^d \to \mathbb{R}$ is convex and satisfies $\left(\gamma^{-1}, \frac{\kappa}{\kappa-1}\right)$ growth condition, then it satisfies (γ, κ)-KL condition.*

It is proven in (Karimi et al., 2016, Appendix A) that the KL condition with $\kappa = 2$ (i.e., the PL condition) implies quadratic growth. We present the following generalized version of this argument.

Lemma 17 *Assume $F : \mathbb{R}^d \to \mathbb{R}$ satisfies the (γ, κ)-KL condition for $\kappa \geq 1$ and $\gamma > 0$. Let $w \in \mathbb{R}^d$ and let w_p be the projection of w onto the set of optimal solutions, $\mathcal{W}^* := \arg\min_w F(w)$. Then it holds that $F(w) - F(w_p) \geq \left[\frac{1}{\gamma} \cdot \frac{\kappa-1}{\kappa}\right]^{\frac{\kappa}{\kappa-1}} \|w - w_p\|^{\frac{\kappa}{\kappa-1}}$.*

Proof Define $F^* = \min_{w \in \mathbb{R}^d} \{F(w)\}$ and $g(w) = \frac{1}{1-1/\kappa}[F(w) - F^*]^{1-\frac{1}{\kappa}}$. We have

$$\|\nabla g(w)\|^2 = \left\| \frac{\nabla F(w)}{[F(w) - F^*]^{1/\kappa}} \right\|^2 \tag{2}$$

$$= \frac{\|\nabla F(w)\|^2}{[F(w) - F^*]^{2/\kappa}} \tag{3}$$

$$= \left(\frac{\|\nabla F(w)\|^\kappa}{[F(w) - F^*]} \right)^{2/\kappa} \geq \frac{1}{\gamma^2} \tag{4}$$

Consider the gradient flow starting at a point w_0 given by

$$\frac{d\boldsymbol{w}(t)}{dt} = -\nabla g(w(t)), \qquad \boldsymbol{w}(t)|_{t=0} = w_0$$

Note F is invex (i.e. its stationary points are global minimizers) because it is KL. Thus g is an invex function because it is the composition of monotonically increasing function and invex function. Further, because g is bounded from below (by 0), the path described above eventually reaches the minimum thus there exists $T < +\infty$ such that $F(\boldsymbol{w}(T)) = F(w^*)$.

Note the length of the path is at least $\|w_0 - w_p\|$. We then have

$$\begin{aligned} g(w_0) - g(w_T) &= -\int_0^T \langle \nabla g(\boldsymbol{w}(t)), \frac{d\boldsymbol{w}(t)}{dt} \rangle dt \\ &= \int_0^T \|\nabla g(\boldsymbol{w}(t))\|^2 dt \\ &\overset{(i)}{\geq} \frac{1}{\gamma} \int_0^T \|\nabla g(\boldsymbol{w}(t))\| \\ &\overset{(ii)}{\geq} \frac{1}{\gamma} \|w_0 - w_p\|, \end{aligned}$$

where (i) uses Eqn. (2) and (ii) uses the lower bound on the path length. Plugging in the definition of g then gives

$$F(w) - F(w_p) \geq \left[\frac{1 - 1/\kappa}{\gamma}\|w - w_p\|\right]^{\frac{1}{1-1/\kappa}}.$$

Note the bound is non-negative if $\kappa \geq 1$. Finally, observing that $\frac{1}{1-1/\kappa} = \frac{\kappa}{\kappa-1}$ establishes the claim. ■

Remark 18 *Using the above result one can observe that if the KL condition holds over a ball of radius $B \geq \frac{\kappa}{\kappa-1}\gamma F_0^{\frac{\kappa-1}{\kappa}}$, then $w^* \in \mathcal{B}_B(w_0)$. Then for some $w' \in \mathbb{R}^d$, a triangle inequality can then be used to obtain $\|w' - w^*\| \leq \|w_0 - w'\| + \|w_0 - w^*\|$. This would allow one to phrase our results in terms of a ball centered at w^*.*

Appendix B. Missing Proofs from Section 3

B.1. Privacy of Algorithm 1

Lemma 19 *Assume f is L_0-Lipschitz and L_1-smooth over $\mathcal{B}_B(w^*)$ (where B is as given in Theorem 4). Then Algorithm 1 is 2ρ-zCDP.*

Proof First, by Lemma 5, every $w_{k,t}$, $k \in [K]$,$t \in [T_k]$, is in $\mathcal{B}_B(w^*)$, and thus the loss is Lipschitz and smooth at the iterates generated by the algorithm. The sensitivity of the minibatch gradient estimates (made in the outer loop) is then $\frac{L_0}{n}$, and at most K such estimates are made. Smoothness guarantees the sensitivity of the gradient difference estimates (made in the inner loop) at some $k \in [T]$, $t \in [T_k]$ is $\frac{\eta_{k,t}L_1}{n}\|\nabla_{k,t}\| \leq \frac{1}{n}\hat{\Phi}_k^{1/\kappa}$ since $\eta_{k,t} = \frac{1}{4\gamma L_1\|\nabla_{k,t}\|}\hat{\Phi}_k^{1/\kappa}$. Note at most T_k such estimates are made.

The zCDP guarantees of the Gaussian mechanism ensures that the process of generating each $\nabla_{k,0}$ is $\hat{\rho}$-CDP with $\hat{\rho} = \frac{1}{K}$. Similarly, we have that the process of generating each $\Delta_{k,t}$, $t > 0$, is ρ_k-zCDP with $\rho_k = \frac{\rho}{KT_k}$. By the composition theorem for zCDP we then have the overall privacy, is at most $\sum_{k=1}^{K}\left(\frac{\rho}{K} + \sum_{t=1}^{T_k}\frac{\rho}{KT_k}\right) = 2\rho$. ■

B.2. Descent Equation for Algorithm 1

Lemma 20 *With probability at least $1 - \beta$, for every $k \in [K]$ and $t \in [T_k]$ indexing iterates of the algorithm it holds that $F(w_{k,t}; S) - F(w_{k,t+1}; S) \geq \frac{1}{16\gamma L_1}\|\nabla_{k,t}\|\hat{\Phi}_k^{1/\kappa} - \frac{1}{4L_1}\|\nabla_{k,t} - \nabla F(w_{k,t}; S)\|^2$*

Proof We start with a standard descent analysis. Since $F(\cdot; S)$ is L_1-smooth, we have

$$\begin{aligned}
F(w_{k,t}; S) - F(w_{k,t+1}; S) &\geq \langle \nabla F(w_{k,t}; S), w_{k,t} - w_{k,t+1} \rangle - \frac{L_1}{2} \|w_{k,t+1} - w_{k,t}\|^2 \\
&= \eta_{k,t} \langle \nabla F(w_{k,t}; S), \nabla_{k,t} \rangle - \frac{L_1 \eta_{k,t}^2}{2} \|\nabla_{k,t}\|^2 \\
&= \eta_{k,t} \left(1 - \frac{\eta_{k,t} L_1}{2}\right) \|\nabla_{k,t}\|^2 + \eta_{k,t} \langle \nabla F(w_t; S) - \nabla_{k,t}, \nabla_{k,t} \rangle \\
&\overset{(i)}{\geq} \eta_{k,t} \left(\frac{1}{2} - \frac{\eta_{k,t} L_1}{2}\right) \|\nabla_{k,t}\|^2 - \frac{\eta_{k,t}}{2} \|\nabla_{k,t} - \nabla F(w_{k,t}; S)\|^2. \\
&\overset{(ii)}{\geq} \frac{\eta_{k,t}}{4} \|\nabla_{k,t}\|^2 - \frac{1}{4L_1} \|\nabla_{k,t} - \nabla F(w_{k,t}; S)\|^2 \\
&\overset{(iii)}{=} \frac{1}{16 \gamma L_1} \|\nabla_{k,t}\| \hat{\Phi}_k^{1/\kappa} - \frac{1}{4L_1} \|\nabla_{k,t} - \nabla F(w_{k,t}; S)\|^2
\end{aligned}$$

Step (i) uses Young's inequality. Step (ii) uses the fact that $\eta_{k,t} \leq \frac{1}{2L_1}$. This is because $\eta_{k,t} = \frac{1}{4\gamma L_1 \|\nabla_{k,t}\|} \hat{\Phi}_k^{1/\kappa}$ and updates are only performed when $\|\nabla_{k,t}\| \geq \frac{7}{8\gamma} \hat{\Phi}_k^{1/\kappa}$. Step (iii) uses the setting of η_t. ■

B.3. Proof of Lemma 5

Due to the step size and the phases lengths, with probability 1, we have that,

$$\|w_{k,t} - w_0\| \leq \sum_{k=1}^{K} \frac{1}{4\gamma L_1} \hat{\Phi}_k^{1/\kappa} T_k \quad \leq \sum_{k=1}^{K} \frac{1}{4 F_0^{\frac{\kappa-2}{\kappa}} \gamma L_1} \hat{\Phi}_k^{1/\kappa} \hat{\Phi}_k^{\frac{\kappa-2}{\kappa}} = \frac{F_0^{\frac{2-\kappa}{\kappa}} F_0^{\frac{\kappa-1}{\kappa}}}{4\gamma L_1} \sum_{k=1}^{K} \left(\frac{1}{c^{\frac{\kappa-1}{\kappa}}}\right)^k$$

Above, we use the fact that $\frac{\kappa-1}{\kappa} \geq 0$ (since $\kappa \geq 1$) to bound $\hat{\Phi}_k \leq F_0$. Since $c > 1$ we have, recalling $K = (1 + 64 F_0^{\frac{\kappa-2}{\kappa}} \gamma^2 L_1) \Big[\log(F_0) - \kappa \log\left(\frac{L_0 \sqrt{d}}{n\sqrt{\rho}}\right) \Big]$,

$$\|w_{k,t} - w_0\| \leq \frac{K F_0^{1/\kappa}}{4\gamma L_1} = \left(\frac{F_0^{1/\kappa}}{4\gamma L_1} + 16 F_0^{\frac{\kappa-1}{\kappa}} \gamma\right) \Big[\log(F_0) + \kappa \log\left(\frac{n\sqrt{\rho}}{L_0 \sqrt{d}}\right) \Big].$$

B.4. Lower Bound for Smooth Losses Satisfying Growth Condition

We provide the following extension of the lower bound on excess risk in Asi et al. (2021b). Our extension yields a lower bound for losses which satisfy (λ, τ)-growth and are $L_1 \geq 0$-smooth over a ball $\mathcal{B}_R(w_0)$, for any smoothness parameter $L_1 \geq 0$ and radius $R > 0$. In contrast, the setting of Asi et al. (2021b) did not have the above smoothness and existence of a (large) ball $\mathcal{B}_R(w_0)$ assumption (over which smoothness and Lipschitzness holds). Further, Asi et al. (2021b) provide a lower bound for *constrained* DP procedures, which is is based on a reduction from convex ERM over a constrained set of *any* diameter D (Bassily et al., 2014). In contrast, we are interested in lower bound for unconstrained procedures. Therefore, in Theorem 22, we extend the lower bound

of Bassily et al. (2014) to the unconstrained setting. We then provide a reduction, closely following Asi et al. (2021b), from unconstrained convex ERM to unconstrained optimization of functions satisfying a growth condition. Finally, we note that our unconstrained lower bound in Theorem 22 holds pointwise for all values of the norm of optimal solution D, so it suffices to construct a reduction for *some* choice of D. We show that for any given setting of problem parameters, there is a choice of D, for which the reduced instance satisfies the requisite properties.

Theorem 21 *Let $L_0, L_1, B, \lambda \geq 0, \tau \geq 2, \tau = O(1), 0 < \epsilon \leq 1, 2^{-\Omega(n)} \leq \delta < \frac{1}{n}$. For any (ϵ, δ)-DP algorithm $\mathcal{A}$, there exists a set $\mathcal{W} \subset \mathbb{R}^d$ containing a ball of radius B, a dataset S and a convex loss function f such that for all x, the function $w \mapsto f(w; x)$ is L_0-Lipschitz, L_1-smooth over $\mathcal{W}$, the empirical loss $w \mapsto F(w; S)$ satisfies (λ, τ)-growth, and*

$$\mathbb{E}_{\mathcal{A}}[F(\mathcal{A}(S); S) - \inf_{w \in \mathbb{R}^d} F(w)] = \Omega\left(\frac{1}{\tau^{\frac{1}{\tau-1}}}\left(\frac{L_0\sqrt{d}}{\lambda n \epsilon}\right)^{\frac{\tau}{\tau-1}}\right).$$

Proof The key to the proof is the following reduction, based on Proposition 3 of Asi et al. (2021b). Herein, the aim is to show that the existence of a DP optimizer for convex losses satisfying the growth condition implies the existence of an optimizer for general convex losses. More formally, consider a problem instance class where we are given a set $\mathcal{W} \subset \mathbb{R}^d$ containing a ball of radius B, a dataset $S \in \mathcal{X}^n$ for some $\mathcal{X}$, a function $f(w; x)$ where $w \mapsto f(w; x)$ is L_0-Lipschitz, L_1-smooth over $\mathcal{W}$ for all $x \in \mathcal{X}$ and the empirical loss $w \mapsto F(w; S)$ satisfies (λ, τ)-growth. Note since these properties hold over $\mathcal{W}$, they hold over the ball of radius B. If there exists an (ϵ, δ)-DP algorithm $\mathcal{A}$, which for the above problem instance has expected excess empirical risk,

$$\mathbb{E}_{\mathcal{A}}[F(\mathcal{A}(S); S) - \inf_{w} F(w)] = o\left(\left(\tau\lambda^{\tau}\right)^{-\frac{1}{\tau-1}} \Delta(n, d, L_0, L_1, \epsilon, \delta)\right),$$

then for $D = \max\left(\frac{(\Delta(n,d,L_0,L_1,\epsilon,\delta))^{1/\tau} L_1 L_0^{\frac{\tau-2}{\tau-1}}}{c_2(\tau)}, \frac{(\Delta(n,d,L_0,L_1,\epsilon,\delta))^{1/\tau} B}{L_0^{\frac{1}{(\tau-1)}} c_3(\tau)}, \frac{\sqrt{d} L_0}{L_1 n \epsilon}\right)$, where $c_2(\tau) = \Omega(1)$ and $c_3(\tau) = \Omega(1)$ are specified later, there exists an (ϵ, δ)-DP algorithm $\tilde{\mathcal{A}}$, such that for any L_0-Lipschitz, convex, L_1-smooth loss function $w \mapsto \tilde{f}(w; x)$ for all x, with minimizer norm $\|w^*\| = D$, its excess risk is

$$\mathbb{E}_{\tilde{\mathcal{A}}}[\tilde{F}(\tilde{\mathcal{A}}(S); S) - \inf_{w \in \mathbb{R}^d} \tilde{F}(w)] = o\left(D\left(\Delta(n, d, 2L_0, 2L_1, \epsilon/k, \delta/k)\right)^{\frac{\tau-1}{\tau}}\right)$$

where k is the smallest integer larger that $\log\left(\frac{\tau^{\frac{1}{\tau-1}} L_0^{\frac{\tau}{\tau-1}}}{2^{2\tau-3}\Delta(n,d,2L_0,2L_1,\epsilon/k,\delta/k)}\right)$.

The main difference between above and the statement of Asi et al. (2021b) is that unlike Asi et al. (2021b), our reduction is for unconstrained procedures and is tailored to the aforementioned choice of diameter D.

The proof uses the construction of Asi et al. (2021b), verifying that for the provided parameter settings, the assumptions hold. For simplicity of notation, let $\Delta = \Delta(n, L_0, L_1, \epsilon, \delta)$.

Let w_0 be the origin. For a sequence of $\{\lambda_i\}_i$ to be instantiated later, define

$$\tilde{\mathcal{W}}_i = \left\{w : \|w - w_{i-1}\| \leq \left(\frac{L_0}{2\tau\lambda_i^{\tau} 2^{\tau-2}}\right)^{\frac{1}{\tau-1}}\right\}$$
$$\tilde{F}_i(w; S) = F(w; S) + \lambda_i^{\tau} 2^{\tau-2} \|w - w_{i-1}\|^{\tau}$$

where $w_i = \mathcal{A}(\tilde{F}_i, S)$. The function $\tilde{F}_i$ satisfies $(\lambda_i 2^{(\tau-2)/\tau}, \tau)$-growth (over all of $\mathbb{R}^d$). We now inspect its Lipschitzness and smoothness parameters over $\tilde{\mathcal{W}}_i$. By direct calculation, the Lipschitz parameter is bounded by $L_0 + L_0 = 2L_0$ The smoothness parameter is at most,

$$L_1 + \lambda_i^\tau 2^{\tau-2}\tau(\tau-1)\left\|w - w_{i-1}\right\|^{\tau-2} = L_1 + \lambda_i^\tau 2^{\tau-2}\tau(\tau-1)\left(\frac{L_0}{2\tau\lambda_i^\tau 2^{\tau-2}}\right)^{\frac{\tau-2}{\tau-1}}$$
$$= L_1 + (\lambda_i)^{\frac{\tau}{\tau-1}}(L_0)^{\frac{\tau-2}{\tau-1}} c_1(\tau),$$

where $c_1(\tau) = \frac{2^{\frac{\tau-2}{\tau-1}}\tau^{\frac{1}{\tau-1}}(\tau-1)}{2^{\frac{\tau-2}{\tau-1}}} = \tau^{\frac{1}{\tau-1}}(\tau-1)$. In Asi et al. (2021b), λ_i is set as $\lambda_i = 2^{-\left(\frac{\tau-1}{\tau}\right)i}\lambda$ for λ to be specified later. The above smoothness bound is a decreasing function in i, so what suffices is to show that the above bound is smaller than $2L_1$ for the largest λ_i, which is $\lambda_1 = 2^{-\left(\frac{\tau-1}{\tau}\right)}\lambda$. From Asi et al. (2021a), $\lambda = 4^{\frac{(\tau-1)^2}{\tau^2}}\left(\frac{\Delta\tau}{D^\tau(\tau-1)}\right)^{\frac{(\tau-1)}{\tau^2}}$, so we have,

$$(\lambda_i)^{\frac{\tau}{\tau-1}}(L_0)^{\frac{\tau-2}{\tau-1}}c(\tau) = 4^{\frac{\tau-1}{\tau}}\left(\frac{\tau}{\tau-1}\right)^{1/\tau}\frac{\Delta^{1/\tau}}{D}(L_0)^{\frac{\tau-2}{\tau-1}}c_1(\tau) = c_2(\tau)\frac{\Delta^{1/\tau}}{D}(L_0)^{\frac{\tau-2}{\tau-1}},$$

where $c_2(\tau) = 4^{\frac{\tau-1}{\tau}}\left(\frac{\tau}{\tau-1}\right)^{1/\tau}\tau^{\frac{1}{\tau-1}}(\tau-1)$. The choice of $D \geq \frac{L_1\Delta^{1/\tau}(L_0)^{\frac{\tau-2}{\tau-1}}}{c_2(\tau)}$, ensures the above is at most L_1, thereby establishing that the smoothness parameter is at most $2L_1$. The final condition we want to ensure is that all the sets $\tilde{\mathcal{W}}_i$ contain a ball of radius at least B. Since λ_i is decreasing in i, it suffices to consider $i = 1$. We have,

$$\left(\frac{L_0}{2\tau\lambda_1^\tau 2^{\tau-2}}\right)^{\frac{1}{\tau-1}} = \frac{1}{2\tau^{\frac{1}{\tau-1}}}\left(\frac{\tau-1}{\tau}\right)^{\frac{1}{\tau}}\frac{1}{4^{\frac{\tau-1}{\tau}}}L_0^{\frac{1}{\tau-1}}\frac{D}{\Delta^{1/\tau}} = c_3(\tau)L_0^{\frac{1}{\tau-1}}\frac{D}{\Delta^{1/\tau}}$$

where $c_3(\tau) = \frac{1}{2\tau^{\frac{1}{\tau-1}}}\left(\frac{\tau-1}{\tau}\right)^{\frac{1}{\tau}}\frac{1}{4^{\frac{\tau-1}{\tau}}}$. The choice of $D \geq \frac{Bk\Delta^{1/\tau}}{L_0^{\frac{1}{\tau-1}}c_3(\tau)}$ ensures the above is at least B.

The rest of the proof repeats the arguments in Asi et al. (2021b), to get,

$$\mathbb{E}[\tilde{F}(\mathcal{A}'(S); S)] - \min_w \tilde{F}(w; S) = o\left(D\left(\Delta(n, d, 2L_0, 2L_1, \epsilon/k, \delta/k)\right)^{\frac{\tau-1}{\tau}}\right)$$

We now instantiate $\Delta(n, d, L_0, L_1, \epsilon, \delta) = \frac{L_0\sqrt{d}}{n\epsilon}$. This gives us that $\mathbb{E}[\tilde{F}(\mathcal{A}'(S); S)] - \min_w \tilde{F}(w; S) = o\left(\frac{L_0 D\sqrt{d}}{n\epsilon}\right)$. However, this contradicts our lower bound in Theorem 22 for unconstrained DP procedures for convex, L_0-Lispchitz, $\frac{\sqrt{d}L_0}{Dn\epsilon} \leq L_1$-smooth (by our choice of D) losses. ■

B.5. Additional Details for Corollary 7 (Lower Bound)

In Theorem 21 (in Appendix B.4), an extension of (Asi et al., 2021b, Theorem 6), we show that for $\tau \geq 2$ and $\tau = \Theta(1)$, the lower bound on the minimax optimal expected excess empirical risk, α, for (ϵ, δ)-DP ERM of functions which are smooth and Lipschitz over a ball of any finite radius $B > 0$ and globally satisfy convexity and (λ, τ)-growth, is

$$\alpha = \tilde{\Omega}\left(\frac{1}{(\tau)^{\frac{1}{\tau-1}}}\left(\frac{L_0\sqrt{d}}{\lambda n\epsilon}\right)^{\frac{\tau}{\tau-1}}\right).$$

Lemma 16 gives that (λ, τ)-growth and convexity implies (γ, κ)-KL with $\lambda = \gamma^{-1}$ and $\tau = \frac{\kappa}{\kappa-1}$. Further, if $\kappa \leq 2$, then $\tau = \frac{\kappa}{\kappa-1} \geq 2$ and if $\kappa = 1 + \Omega(1)$ then $\tau = O(1)$. Thus we have the lower bound,

$$\alpha = \tilde{\Omega}\left(\left[\frac{\kappa}{\kappa-1}\right]^{1-\kappa}\left(\frac{\gamma L_0\sqrt{d}}{n\epsilon}\right)^{\kappa}\right) = \tilde{\Omega}\left(\left(\frac{\gamma L_0\sqrt{d}}{n\epsilon}\right)^{\kappa}\right).$$

The last step uses the fact that $\kappa = 1 + \Omega(1)$. Finally, the existence of a ρ-zCDP algorithm with rate better than $\tilde{O}\left(\left(\frac{\gamma L_0\sqrt{d}}{n\sqrt{\rho}}\right)^{\kappa}\right)$ would imply the existence of an (ϵ, δ)-DP algorithm (see (Bun and Steinke, 2016, Proposition 1.3)) with rate better than $\tilde{O}\left(\left(\frac{\gamma L_0\sqrt{d}}{n\epsilon}\right)^{\kappa}\right)$, a contradiction.

B.6. Unconstrained Lower Bound for General Loss Functions

In this section, we provide an extension of the lower bound on excess risk of DP procedures for convex Lipschitz functions in the constrained setting (Bassily et al., 2014) to the unconstrained setting. The key idea in the proof is to define a Lipschitz extension of the hard instance in Bassily et al. (2014) using the Huber regularizer. The dataset for our construction, as in Bassily et al. (2014), leverages fingerprinting codes. The exact details of fingerprinting codes are note needed for our proof below, but we defer the interested reader to Bun et al. (2014) for more details. The following result is used in the proof of the lower bound for functions satisfying (λ, τ)-growth for $2 \leq \tau = O(1)$ in Theorem 21.

Theorem 22 *Let $0 < \epsilon \leq 1, 0 < \delta < \frac{1}{n}$, $D, L_0 > 0$. For any (ϵ, δ)-DP algorithm, there exists a dataset S, and a L_0-Lipschitz, $\frac{\sqrt{d}L_0}{n\epsilon D}$-smooth convex loss function $w \mapsto f(w; x)$ for all x, such that its unconstrained minimizer, $w^* = \arg\min_w \{F(w; S)\}$, has norm at most D, and*

$$\mathbb{E}_{\mathcal{A}}[F(\mathcal{A}(S); S) - F(w^*; S)] = \Omega\left(L_0 D \min\left\{\frac{\sqrt{d}}{n\epsilon}, 1\right\}\right).$$

Proof Consider the loss function

$$F(w; S) = \frac{1}{n}\sum_{i=1}^{n}\langle w, x_i\rangle + \lambda H(w), \tag{5}$$

where H is the "Huber regularization" defined as

$$H(w) = \begin{cases} \|w\|^2 & \text{if } \|w\| \leq 4D \\ 4D\,\|w\| & \text{otherwise} \end{cases} \tag{6}$$

Note that if N of the x_i vectors are vectors in $\left\{\pm\frac{L_0}{\sqrt{d}}\right\}^d$ and the rest are the zero vector, we have $\|\sum_{i=1}^n x_i\| \leq NL_0$. The empirical minimizer is $w^* = -\frac{\sum_{i=1}^n x_i}{2n\lambda}$. Thus we set $\lambda = \frac{NL_0}{2nD}$ so that $\|w^*\| \leq D$. We also remark that under this setting of λ that F is Lipschitz with parameter $L_0' = L_0 + 4\lambda D \leq 5L_0$.

Now we will show that any w which achieves small excess risk is close to w^*. Then we will use a lower bound on this distance to lower bound the error (as in Bassily et al. (2014)). For any w such that $\|w\| \leq 4D$ have

$$\begin{aligned}
F(w;S) - F(w^*;S) &= \langle w - w^*, \frac{1}{n}\sum_{i=1}^{n} x_i \rangle + \lambda \left(\|w\|^2 - \|w^*\|^2 \right) \\
&= 2\lambda \langle w - w^*, -w^* \rangle + \lambda \left(\|w\|^2 - \|w^*\|^2 \right) \\
&= 2\lambda (\|w^*\|^2 - \langle w, w^* \rangle) + \lambda \left(\|w\|^2 - \|w^*\|^2 \right) \\
&= 2\lambda \left(\frac{1}{2}\|w^*\|^2 - \frac{1}{2}\|w\|^2 + \frac{1}{2}\|w - w^*\|^2 \right) + \lambda \left(\|w\|^2 - \|w^*\|^2 \right) \\
&= \lambda \|w - w^*\|^2 \\
&= \frac{NL_0}{nD} \|w - w^*\|^2
\end{aligned}$$

where the fourth equality comes from $\langle a, b \rangle = \frac{1}{2}(\|a\|^2 + \|b\|^2 - \|a - b\|^2)$. Now (Bassily et al., 2014, Lemma 5.1) gives that for $N = \min\left\{ \frac{\sqrt{d}}{\epsilon}, n \right\}$ there exists a construction of the non-zero dataset vectors such that the output of any (ϵ, δ)-DP algorithm, $\mathcal{A}(S)$, must satisfy $\mathbb{E}[\|\mathcal{A}(S) - w^*\|] = \Omega\left(\frac{\sqrt{d}D}{N\epsilon} \right)$. Thus we have

$$\mathbb{E}[F(\mathcal{A}(S);S) - F(w^*;S)] = \Omega\left(L_0 D \min\left\{ \frac{\sqrt{d}}{n\epsilon}, 1 \right\} \right).$$

This lower bounds the excess loss for any w such that $\|w\| \leq 4D$. Finally, note that any w' such that $\|w'\| \geq 4D$ (i.e. a point outside the quadratic region of H) would also have high empirical risk because of the regularization term. Specifically, we have for any such w' that

$$F(w';S) \geq -\frac{\|w'\| L_0 N}{n} + 4\lambda D \|w'\| \geq 16\lambda D^2 - \frac{4DL_0N}{n}$$

Further since $\|w^*\| \leq D$, we have $F(w^*;S) \leq \frac{NL_0}{n} + \lambda D^2$. This gives

$$F(w';S) - F(w^*;S) \geq 15\lambda D^2 - \frac{5L_0DN}{n} = \Omega\left(\frac{\sqrt{d}DL_0}{n\epsilon} \right)$$

where the last step follows from the setting of λ. Combining the two cases finishes the proof. ■

Appendix C. Missing Results from Section 4

C.1. Proof of Lemma 10

For any $t \in [T]$, the stationarity conditions of F_t imply $\|\nabla F(w_t^*;S)\| = 2\tilde{L_1}\|w_t^* - w_{t-1}\|$, and so by Lipschitzness $\|w_t^* - w_{t-1}\| \leq \frac{L_0}{2\tilde{L_1}}$. Further, we have by strong convexity and the accuracy guarantee

of $\mathcal{A}$ that with probability $1-\beta$ for any $t \in [T]$ that $\|w_t - w_t^*\| = O\left(\frac{1}{\sqrt{\tilde{L}_1}}\sqrt{F_t(w_t;S) - F_t(w_t^*;S)}\right) = O\left(\frac{L_0\sqrt{Td\log\left(n^2\log^2\left(1/\beta'\right)/d\beta'\right)}}{\tilde{L}_1 n\sqrt{\rho}}\right)$. Thus using the triangle inequality the overall magnitudes of the updates are bounded by $\|w_t^* - w_{t-1}\| = O\left(\frac{1}{\tilde{L}_1}\left(L_0 + \frac{L_0\sqrt{Td\log\left(n^2\log^2\left(1/\beta'\right)/d\beta'\right)}}{n\sqrt{\rho}}\right)\right)$. In the following, let $\tau = \frac{\sqrt{Td\log\left(n^2\log^2\left(1/\beta'\right)/d\beta'\right)}}{n\sqrt{\rho}}$. Since at most T iterations occur, we have

$$\begin{aligned}
\|w_t - w_0\| &= O\left(\frac{T}{\tilde{L}_1}\left(L_0 + \frac{L_0\sqrt{Td\log\left(n^2\log^2\left(1/\beta'\right)/d\beta'\right)}}{n\sqrt{\rho}}\right)\right) \\
&= O\left(\frac{TL_0(1+\tau)}{\tilde{L}_1}\right) \\
&= O\left(\frac{L_0(1+\tau)}{\tilde{L}_1}(1 + F_0^{\frac{\kappa-2}{\kappa}}\gamma^2\tilde{L}_1)\Big[\log(F_0) - \kappa\log\left(\frac{\gamma L_0\sqrt{d\log(1/\beta')}}{n\sqrt{\rho}}\right)\Big]\right) \\
&= O\left(\left(\frac{L_0(1+\tau)}{\tilde{L}_1} + L_0F_0^{\frac{\kappa-2}{\kappa}}\gamma^2\right)\Big[\log(F_0) - \kappa\log\left(\frac{\gamma L_0\sqrt{d\log(1/\beta')}}{n\sqrt{\rho}}\right)\Big]\right).
\end{aligned}$$

C.2. Proof of Theorem 8 (Convergence of PPM under the KL* Condition)

Proof [Proof of Theorem 8] In the following, we condition on the event that every run of $\mathcal{A}$ obtains excess risk at most $\frac{aL_0^2 d\log\left(n^2\log^2\left(1/\beta'\right)/d\beta'\right)}{\tilde{L}_1 n^2\rho}$ for some universal constant a. Since $\beta' = \frac{\beta}{T}$, this event happens w.p. at least $1-\beta$ by Lemma 9. Further, under this same event, the KL condition holds at every w_t, $t \in [T]$, by Lemma 10.

Now define $c = 1 + F_0^{\frac{2-\kappa}{\kappa}}\frac{1}{32\gamma^2 L_1}$, $\hat{\Phi}_0 = F_0$ and

$$\hat{\Phi}_t = \max\left\{\frac{1}{c}\hat{\Phi}_{t-1}, \min\left\{\left(\frac{a\gamma L_0\sqrt{Td\log(n^2\log^2\left(1/\beta'\right)/d\beta')}}{n\sqrt{\rho}}\right)^{\kappa}, F_0\right\}\right\}.$$

We will first prove by induction that $F(w_t;S) - F(w^*;S) \le \hat{\Phi}_t$ under the assumption that $F(w_{t-1};S) - F(w^*;S) \le \hat{\Phi}_{t-1}$. Clearly the base case is satisfied for $\hat{\Phi}_0$.

To prove the induction step, we will proceed by contradiction. That is, assume by contradiction that $F(w_t;S) - F(w^*;S) > \hat{\Phi}_t$.

Note F_t is $\tilde{L}_1$-strongly convex since it is the sum of a $\tilde{L}_1$ weakly convex function and a $2\tilde{L}_1$ strongly convex function (Davis and Drusvyatskiy, 2019a). Let τ be an upper bound on the excess risk achieved by $\mathcal{A}$ on the strongly convex objective F_t. Then

$$\begin{aligned}
F(w_t;S) = F_t(w_t;S) \le F_t(w_t^*;S) + \tau &\overset{(i)}{\le} F(w_{t-1};S) - \frac{\tilde{L}_1}{2}\|w_{t-1} - w_t^*\|^2 + \tau \\
\implies F(w_{t-1};S) - F(w_t;S) &\ge \frac{\tilde{L}_1}{4}\|w_{t-1} - w_t^*\|^2 - \tau \qquad (7)
\end{aligned}$$

Inequality (i) uses the fact that $F_t(w_t^*; S) = F(w_t^*; S) + \frac{\tilde{L_1}}{2}\|w_t^* - w_{t-1}\|^2 \leq F_t(w_{t-1}; S) = F(w_{t-1}; S)$, which implies $F_t(w_t^*; S) \leq F(w_{t-1}; S) - \frac{\tilde{L_1}}{2}\|w_t^* - w_{t-1}\|^2$. Recall we have $\tau \leq \frac{aL_0^2 d \log(n^2 \log^2(1/\beta')/d\beta')T}{\tilde{L}_1 n^2 \rho}$ by Lemma 9. Further, note by stationarity conditions for the regularized objective we have

$$\|\nabla F(w_t^*; S)\| = 2\tilde{L_1}\|w_t^* - w_{t-1}\|. \tag{8}$$

By the assumption that $F(w_t^*; S) - F(w^*; S) \geq \hat{\Phi}_t$ and the KL condition we have $\|\nabla F(w_t^*; S)\| \geq \frac{1}{\gamma}\hat{\Phi}_t^{1/\kappa}$, and thus by Eqn. (8) we have $\|w_t^* - w_{t-1}\| \geq \frac{1}{2\gamma\tilde{L_1}}\hat{\Phi}_t^{1/\kappa}$. Applying Eqn. (7) gives

$$F(w_{t-1}; S) - F(w_t; S) \geq \frac{1}{16\gamma^2\tilde{L_1}}\hat{\Phi}_t^{2/\kappa} - \tau \overset{(i)}{\geq} \frac{1}{32\gamma^2\tilde{L_1}}\hat{\Phi}_t^{2/\kappa},$$

where inequality (i) comes from the setting $\hat{\Phi}_t \geq \left(\frac{a\gamma L_0\sqrt{Td\log(n^2\log^2(1/\beta')/d\beta')}}{n\sqrt{\rho}}\right)^\kappa$. Adding and subtracting $F(w^*; S)$ on the left hand side and rearranging obtains

$$\begin{aligned}
F(w_t; S) - F(w^*; S) &\leq F(w_{t-1}; S) - F(w^*; S) - \frac{1}{32c^{2/\kappa}\gamma^2\tilde{L_1}}\hat{\Phi}_{t-1}^{2/\kappa} \\
&\overset{(i)}{\leq} \hat{\Phi}_{t-1} - \frac{1}{32c^{2/\kappa}\gamma^2\tilde{L_1}}\hat{\Phi}_{t-1}^{2/\kappa} \\
&= \left(1 - \Phi_{t-1}^{\frac{2-\kappa}{\kappa}}\frac{1}{32c^{2/\kappa}\gamma^2\tilde{L_1}}\right)\hat{\Phi}_{t-1} \\
&\overset{(ii)}{\leq} \left(1 - F_0^{\frac{2-\kappa}{\kappa}}\frac{1}{32c\gamma^2\tilde{L_1}}\right)\hat{\Phi}_{t-1} = \frac{1}{c}\hat{\Phi}_{t-1} \leq \hat{\Phi}_t.
\end{aligned}$$

Step (i) uses the inductive assumption that $F(w_{t-1}) - F(w^*; S) \leq \hat{\Phi}_{t-1}$. Inequality (ii) uses the fact that $\kappa \geq 2$, $c \geq 1$, and $\Phi_{t-1} \leq F_0$. This establishes a contradiction and thus completes the induction argument. We have now proven that $F(w_t; S) - F(w^*; S) \leq \hat{\Phi}_t$ for all $t \in \{0, ..., T\}$.

All that remains to prove convergence is to show that $\hat{\Phi}_T \leq \left(\frac{\gamma L_0\sqrt{Td\log(n^2\log^2(1/\beta')/d\beta')}}{n\sqrt{\rho}}\right)^\kappa$. We have $\hat{\Phi}_T \leq \max\left\{\left(\frac{1}{c}\right)^T F_0, \left(\frac{\gamma L_0\sqrt{Td\log(n^2\log^2(1/\beta')/d\beta')}}{n\sqrt{\rho}}\right)^\kappa\right\}$ and

$$\left(\frac{1}{c}\right)^T F_0 \leq \left(\frac{\gamma L_0\sqrt{d\log(n^2\log^2(1/\beta')/d\beta')}}{n\sqrt{\rho}}\right)^\kappa \iff T \geq \frac{\left[\log(F_0) + \kappa\log\left(\frac{n\sqrt{\rho}}{\gamma L_0\sqrt{d\log(n^2\log^2(1/\beta')/d\beta')}}\right)\right]}{\log(c)}$$

Using the fact that $\log(c) = \log\left(1 + F_0^{\frac{2-\kappa}{\kappa}}\frac{1}{32\gamma^2\tilde{L_1})}\right) \geq (1 + 32F_0^{\frac{\kappa-2}{\kappa}}\gamma^2\tilde{L_1})^{-1}$, the setting of $T = (1 + 32F_0^{\frac{\kappa-2}{\kappa}}\gamma^2\tilde{L_1})\left[\log(F_0) + \kappa\log\left(\frac{n\sqrt{\rho}}{\gamma L_0\sqrt{d}}\right)\right]$ suffices to ensure convergence. ■

C.3. Lower bound for $\kappa \geq 2$

We give a lower bound on excess empirical risk for settings where the empirical risk satisfies $(1, \kappa)$-KL for $\kappa \geq 2$, under approximate differential privacy.

Theorem 23 *Let $\kappa \geq 2, 0 < \epsilon \leq \ln 2, 0 < \delta \leq \frac{1}{16}\left(1 - e^{-\epsilon}\right), d \in \mathbb{N}$ and $B > 0$. For any (ϵ, δ)-DP procedure $\mathcal{A}$, there exists a data space $\mathcal{X}$, a set $\mathcal{W} \subseteq \mathbb{R}^d$ containing a ball of radius B, a dataset S and a convex loss function $w \mapsto f(w; x)$ which is 1-Lipschitz over $\mathcal{W}$, the empirical loss $w \mapsto F(w; S)$ satisfies $(1, \kappa)$-KL, and*

$$\mathbb{E}_{\mathcal{A}}\left[F(\mathcal{A}(S); S) - \inf_{w \in \mathbb{R}^d} F(w; S)\right] \geq \frac{1}{4}\left(\frac{1}{n\epsilon}\right)^{\kappa}$$

Proof The proof adapts the construction of Asi et al. (2021b), Theorem 5 from a lower bound on excess population risk under pure DP setting to that on excess empirical risk under approximate DP. We first prove a lower bound for $(1, \tau)$-growth functions, for $\tau \in (1, 2]$. We recall the one-dimensional, unconstrained (so $\mathcal{W} = \mathbb{R}^d$) construction in Asi et al. (2021b), Theorem 5. The data space $\mathcal{X} = \{-1, 1\}$, and for $a \in [0, 1]$ to be specified later, define functions

$$f(w; 1) = \begin{cases} |w - a| & w \leq a \\ |w - a|^{\tau} & w > a \end{cases} \quad \text{and} \quad f(w; -1) = \begin{cases} |w + a|^{\tau} & w \leq -a \\ |w + a| & w > -a \end{cases}$$

The functions above are 1-Lipschitz. Consider two datasets S and S' such that S contains $\left(\frac{1+\rho}{2}\right)$ fraction of 1's and the rest -1's. Similarly, S' contains $\left(\frac{1-\rho}{2}\right)$ fraction of 1's and the rest -1's. The number of points differing between S and S' is thus $n\rho$. We set $\rho = 1/n\epsilon$ to get $\frac{1}{\epsilon}$ differing points. The corresponding empirical risk functions are,

$$F(w; S) = \left(\frac{1+\rho}{2}\right) f(w; 1) + \left(\frac{1-\rho}{2}\right) f(w; -1)$$
$$F(w; S') = \left(\frac{1-\rho}{2}\right) f(w; 1) + \left(\frac{1+\rho}{2}\right) f(w; -1)$$

In the construction of Asi et al. (2021b), Theorem 5, the above are their population risk functions "$f_1(x)$" and "$f_{-1}(x)$". Their minimizers are $w^*_S = a$ and $w^*_{S'} = -a$, with values $(1 - \rho)a$ and $(1 + \rho)a$ respectively. Note that the above functions are convex. Further, with $a = \frac{\rho^{\frac{1}{(\tau-1)}}}{2} = \frac{1}{2(n\epsilon)^{\frac{1}{\tau-1}}}$, Asi et al. (2021b) showed that both functions $w \mapsto F(w; S)$ and $w \mapsto F(w; S')$ exhibit

$(1,\tau)$-growth over all of $\mathbb{R}$. For any (ϵ,δ)-DP algorithm $\mathcal{A}$, we have that,

$$\begin{aligned}
&\sup_{\tilde{S}\in\{S,S'\}} \mathbb{E}_{\mathcal{A}}[F(\mathcal{A}(\tilde{S});\tilde{S}) - \inf_w F(w;\tilde{S})] \\
&\geq \frac{1}{2}\mathbb{E}_{\mathcal{A}}\left[F(\mathcal{A}(S);\tilde{S}) - F(w_S^*;S) + F(\mathcal{A}(S');S') - F(w_{S'}^*;S')\right] \\
&\geq \mathbb{E}_{\mathcal{A}}\left[|\mathcal{A}(S) - w_S^*|^\tau + \left|\mathcal{A}(S') - w_{S'}^*\right|^\tau\right] \\
&\geq \frac{1}{2}\left(\mathbb{E}_{\mathcal{A}}\left[|\mathcal{A}(S) - w_S^*| + \left|\mathcal{A}(S') - w_{S'}^*\right|\right]\right)^\tau \\
&\geq \frac{1}{4}\left(\mathbb{E}_{\mathcal{A}}\left[|w_S^* - w_{S'}^*|\right]\right)^\tau \\
&\geq \frac{1}{4}\left(\frac{1}{n\epsilon}\right)^{\frac{\tau}{\tau-1}}
\end{aligned}$$

where the second inequality uses the growth condition, the third uses that for $1 \leq \tau \leq 2$, $|u+v|^\tau \leq 2(|u|^\tau + |v|^\tau)$ and Jensen's inequality; the fourth uses Lemma 2 of Chaudhuri and Hsu (2012) and the final inequality plugs in computed distance between minimizers. Finally, the fact (Lemma 16) that convexity and $\left(1, \frac{\kappa}{\kappa-1}\right)$-growth implies $(1,\kappa)$-KL establishes the $\frac{1}{4}\left(\frac{1}{n\epsilon}\right)^\kappa$ lower bound for $(1,\kappa)$-KL functions. ■

Appendix D. Missing Results from Section 5

D.1. Gradient Error of Algorithm 3

Lemma 24 *Let $T+1$ denote the final value of t reached during the run of Algorithm 3. With probability at least $1-2\beta$ under the randomness of Algorithm 3, for any $t \in [T]$ s.t. $\sigma_t = \frac{N_t}{\sqrt{d\log(n\sqrt{\rho}/\beta)}}$, it holds that*

$$\|\nabla_t - \nabla F(w_t;S)\| \leq N_t \leq \|\nabla F(w_t;S)\| + \frac{L_0\sqrt{\log(n\sqrt{\rho}/\beta)}}{\sqrt{n}\rho^{1/4}}.$$

Further, if for any $t \in [T]$, $\sigma_t = \frac{2L_0}{n\sqrt{\rho}}$ then $t = T$ and with probability at least $1-2\beta$ the above condition holds as well as $\|\nabla_T - \nabla F(w_T;S)\| \leq \frac{L_0\sqrt{d\log(n\sqrt{\rho}/\beta)}}{n\sqrt{\rho}}$.

Proof Condition on the high probability event that for all $t \in [T]$, $\|\hat{b}_t\| \leq \sqrt{\log(n\sqrt{\rho}/\beta)}\hat{\sigma}_t = \frac{L_0\sqrt{\log(n\sqrt{\rho}/\beta)}}{\sqrt{n}\rho^{1/4}}$ and $\|b_t\| \leq \sqrt{d\log(n\sqrt{\rho}/\beta)}\sigma_t = N_t$. This event happens with probability at least $1-2\beta$ due to the concentration properties of Gaussian noise and the fact that at most $n\sqrt{\rho}$ iterations are performed. Under this event we then have the following bound on the gradient error,

$$\|\nabla_t - \nabla F(w_t;S)\| \leq N_t \leq \|\nabla F(w_t;S)\| + \frac{L_0\sqrt{\log(n\sqrt{\rho}/\beta)}}{\sqrt{n}\rho^{1/4}}.$$

The second part of the lemma statement comes from the fact that when $\sigma_t = \frac{2L_0}{n\sqrt{\rho}}$, $\rho_t \geq \frac{\rho}{2}$ and the stopping condition is triggered. The second error bound result again comes from the concentration of Gaussian noise. ■

D.2. Proof of Theorem 11 (Privacy of Algorithm 3)

Proof Denote $T+1$ as the highest value attained by the variable t during the run of the algorithm. Consider any round $t \in [T]$. We consider the privacy of the round conditional on w_{t-1}. Specifically, for the process of generating the gradient and gradient norm estimates at the t'th step, the scale of Gaussian noise ensures this process is ρ_t-zCDP. Specifically,

$$\rho_t = \left(\frac{L_0}{n\hat{\sigma}}\right)^2 + \left(\frac{L_0}{n\sigma_t}\right)^2 = \frac{\sqrt{\rho}}{n} + \min\left\{\frac{L_0^2 d\log(n\sqrt{\rho}/\beta)}{n^2 N_t^2}, \frac{\rho}{2}\right\}.$$

The $\frac{\rho}{2}$-zCDP guarantee of releasing the first $T-1$ iterates is then certified by the stopping condition (i.e. $\sum_{j=0}^{t}\rho_t \leq \frac{\rho}{2}$) and the fully adaptive composition properties of zCDP. That is, (Whitehouse et al., 2022, Theorem 1) guarantees the privacy of the overall process even if the privacy bound at each iteration is chosen adaptively (rather than fixed a-priori as with standard composition theorems). Releasing the T'th iterate is also $\frac{\rho}{2}$-zCDP because $\sigma_T \geq \frac{2L_0}{n\sqrt{\rho}}$ and the sensitivity of any gradient estimate is at most $\frac{L_0}{n}$. Thus the overall algorithm is ρ-zCDP by composition. ■

D.3. Proof of Theorem 13 (Convergence of Adaptive GD under KL* Condition)

In the following we define $T+1$ as the highest attained value of t during the run of Algorithm 3 and define $c := 1 + \frac{1}{8\gamma^2 L_1}$.

Before proceeding with the main proof, it will be useful to first show that in the event that for some $t > 0$ one has $\sigma_t = \frac{L_0}{n\sqrt{\rho}}$, the algorithm has reached its convergence criteria and stops.

Lemma 25 *Let $t > 0$ and assume $\sigma_t = \frac{2L_0}{n\sqrt{\rho}}$. Then Algorithm 3 stops at iteration t and with probability at least $1-2\beta$ one has*

$$F(w_t;S) - F(w_{\mathcal{S}}^*;S) = O\left(\left(\frac{\gamma L_0\sqrt{d\log(n\sqrt{\rho}/\beta)}}{n\sqrt{\rho}}\right)^{\kappa} + \left(\frac{\gamma^2 L_0^2\log(n\sqrt{\rho}/\beta)}{n\sqrt{\rho}}\right)^{\kappa/2}\right)$$

Importantly, this rate is strictly faster than the convergence claimed by Theorem 8. The proof is given in Appendix D.5 and follows straightforwardly from the concentration of Gaussian noise and the KL condition.

Given this fact, we can proceed with the rest of the proof only considering the case where $\sigma_t = \frac{N_t}{\sqrt{d\log(n\sqrt{\rho}/\beta)}}$ for all $t \in \{0,\ldots T\}$. We will first prove (under the stated assumption) the following useful lemma which roughly states that the excess risk is monotonically nonincreasing up to a certain threshold. Note in the following, we use Φ, to denote *exact* excess loss quantities. This in contrast to the analysis of Section 3 where $\hat{\Phi}$ was used to indicate target excess risk loss thresholds. For the rest of this section, we assume $\frac{L_0^2\log(n\sqrt{\rho}/\beta)}{L_1 n} \leq \sqrt{\rho}$, as per the statement of Theorem 13.

Lemma 26 *Define $\Phi_t = F(w_t;S) - F(w_{\mathcal{S}}^*;S)$. Assume $F(\cdot;S)$ is L_1-smooth. Assume $\sigma_t = \frac{N_t}{\sqrt{d\log(n\sqrt{\rho}/\beta)}}$ for all $t \in [T]$. Then with probability at least $1-2\beta$ we have for all $t \in [T]$ that*

$$F(w_t;S) - F(w_{t+1};S) \geq \frac{1}{8L_1}\|\nabla F(w_t;S)\|^2 - \frac{L_0^2\log(n\sqrt{\rho}/\beta)}{L_1 n\sqrt{\rho}} \tag{9}$$

and if $F(\cdot; S)$ is also (γ, κ)-KL then

$$\Phi_{t+1} \leq \max\left\{\Phi_t, 2\left(\frac{\max\left\{\gamma^2, 1\right\} L_0^2 \log(n\sqrt{\rho}/\beta)}{L_1 n\sqrt{\rho}}\right)^{\kappa/2}\right\}$$

Proof Throughout the following we condition on the high probabity event that

$$\|\nabla_t - \nabla F(w_t; S)\| \leq N_t \leq \|\nabla F(w_t; S)\| + \frac{L_0\sqrt{\log(n\sqrt{\rho}/\beta)}}{\sqrt{n}\rho^{1/4}}.$$

which happens with probability at least $1-\beta$ by Lemma 24 (given in Appendix D.1). Now, standard descent lemma analysis yields

$$\begin{aligned}
F(w_t; S) - F(w_{t+1}; S) &\geq \langle \nabla F(w_t; S), w_t - w_{t+1}\rangle - \frac{L_1}{2}\|w_{t+1} - w_t\|^2 \\
&= \frac{1}{4L_1}\|\nabla F(w_t; S)\|^2 - \frac{1}{8L_1}\|\nabla_t - \nabla F(w_t; S)\|^2 \\
&\geq \frac{1}{8L_1}\|\nabla F(w_t; S)\|^2 - \frac{L_0^2 \log(n\sqrt{\rho}/\beta)}{8L_1 n\sqrt{\rho}}.
\end{aligned}$$

This establishes the first claim of the lemma.

Continuing to the second claim, the above implies that if $\Phi_t \geq \left(\frac{\max\left\{\gamma^2,1\right\} L_0^2 \log(n\sqrt{\rho}/\beta)}{L_1 n\sqrt{\rho}}\right)^{\kappa/2}$, we have by the KL condition that

$$\|\nabla F(w_t; S)\|^2 \geq \frac{1}{\gamma^2}\Phi_t^{2/\kappa} \geq \frac{L_0^2 \log(n\sqrt{\rho}/\beta)}{L_1 n\sqrt{\rho}}. \tag{10}$$

Thus we have

$$\Phi_t - \Phi_{t+1} = F(w_t; S) - F(w_{t+1}; S) \geq \frac{1}{8L_1}\|\nabla F(w_t; S)\|^2 - \frac{L_0^2 \log(n\sqrt{\rho}/\beta)}{8L_1 n\sqrt{\rho}} > 0 \tag{11}$$

On the other hand, if $\Phi_t < \left(\frac{\gamma^2 L_0^2 \log(n\sqrt{\rho}/\beta)}{L_1 n\sqrt{\rho}}\right)^{\kappa/2}$ then because using Eqn. (11) and the fact that $\|\nabla F(w_t; S)\| \geq 0$ we obtain

$$\Phi_{t+1} \leq \Phi_t + \frac{L_0^2 \log(n\sqrt{\rho}/\beta)}{L_1 n\sqrt{\rho}} \leq 2\left(\frac{\max\left\{\gamma^2, 1\right\} L_0^2 \log(n\sqrt{\rho}/\beta)}{L_1 n\sqrt{\rho}}\right)^{\kappa/2}.$$

Above we use the assumption that $\frac{L_0^2 \log(n\sqrt{\rho}/\beta)}{L_1 n\sqrt{\rho}} \leq 1$. Thus combining these two inequalities we have $\Phi_{t+1} \leq \max\left\{\Phi_t, 2\left(\frac{\max\left\{\gamma^2,1\right\} L_0^2 \log(n\sqrt{\rho}/\beta)}{L_1 n\sqrt{\rho}}\right)^{\kappa/2}\right\}$. ■

The next lemma establishes how quickly the loss decreases. Specifically, we show that the loss decreases by a constant fraction after a certain number of steps. The smaller the excess risk is, the more steps are required to achieve this decrease. Recall $c := 1 + \frac{1}{8\gamma^2 L_1}$.

Lemma 27 *Let $K > 0$ and $t \in [T]$ and assume the high probability event of Lemma 26 holds. Then for $K \geq (\frac{1}{c}\Phi_t)^{\frac{\kappa-2}{\kappa}} - 1$ it holds that*

$$\Phi_{t+K} \leq \max\left\{\frac{1}{c}\Phi_t, 2\left(\frac{\max\left\{\gamma^2, 1\right\} L_0^2 \log(n\sqrt{\rho}/\beta)}{\min\{L_1, 1\} n\sqrt{\rho}}\right)^{\kappa/2}\right\}.$$

Proof We here condition on the high probability event that Lemma 26 holds (i.e. that the gradient error is bounded for the entire trajectory). We proceed with a proof by contradiction. Assume by contradiction that

$$\Phi_{t+K} > \max\left\{\frac{1}{c}\Phi_t, 2\left(\frac{\max\left\{\gamma^2, 1\right\} L_0^2 \log(n\sqrt{\rho}/\beta)}{\min\{L_1, 1\} n\sqrt{\rho}}\right)^{\kappa/2}\right\}.$$

By Lemma 26, this assumption implies the above inequality also holds for all $\Phi_{t+j}, j \in \{0, \ldots, K\}$,

$$\max\left\{\frac{1}{c}\Phi_t, 2\left(\frac{\max\left\{\gamma^2, 1\right\} L_0^2 \log(n\sqrt{\rho}/\beta)}{\min\{L_1, 1\} n\sqrt{\rho}}\right)^{\kappa/2}\right\} < \Phi_{t+K} \leq \max\left\{\Phi_{t+j}, 2\left(\frac{\max\left\{\gamma^2, 1\right\} L_0^2 \log(n\sqrt{\rho}/\beta)}{L_1 n\sqrt{\rho}}\right)^{\kappa/2}\right\}$$

$$\implies \max\left\{\frac{1}{c}\Phi_t, 2\left(\frac{\max\left\{\gamma^2, 1\right\} L_0^2 \log(n\sqrt{\rho}/\beta)}{\min\{L_1, 1\} n\sqrt{\rho}}\right)^{\kappa/2}\right\} \leq \Phi_{t+j}. \tag{12}$$

The implication above uses the fact that $2\left(\frac{\max\{\gamma^2,1\}L_0^2\log(n\sqrt{\rho}/\beta)}{\min\{L_1,1\}n\sqrt{\rho}}\right)^{\kappa/2} \not< 2\left(\frac{\max\{\gamma^2,1\}L_0^2\log(n\sqrt{\rho}/\beta)}{L_1 n\sqrt{\rho}}\right)^{\kappa/2}$.

We now sum over K steps and using the descent lemma (see Lemma 26, Eqn (9)). We have

$$\begin{aligned}
F(w_t; S) - F(w_{t+K}; S) &\geq \sum_{j=1}^{K}\left(\frac{1}{4L_1}\left\|\nabla F(w_{t+j}; S)\right\|^2 - \frac{L_0^2\log(n\sqrt{\rho}/\beta)}{4L_1 n\sqrt{\rho}}\right) \\
&\overset{(i)}{\geq} \sum_{j=1}^{K}\left(\frac{1}{4\gamma^2 L_1}\Phi_{t+j}^{2/\kappa} - \frac{L_0^2\log(n\sqrt{\rho}/\beta)}{4L_1 n\sqrt{\rho}}\right) \\
&\overset{(ii)}{\geq} \sum_{j=1}^{K}\frac{1}{2\gamma^2 L_1}\Phi_{t+j}^{2/\kappa} \\
&\overset{(iii)}{\geq} \frac{(\frac{1}{c}\Phi_t)^{\frac{\kappa-2}{\kappa}}}{2\gamma^2 L_1}(\frac{1}{c}\Phi_t)^{2/\kappa} = \frac{1}{2c\gamma^2 L_1}\Phi_t
\end{aligned} \tag{13}$$

Step (i) uses the KL condition. Step (ii) uses the fact that Eqn. (12) implies that for all $j \in \{0, \ldots, K\}$, $\Phi_{t+j} \geq \left(\frac{2\gamma^2 L_0^2 A}{n\sqrt{\rho}}\right)^{\kappa/2}$. The second inequality uses the KL condition. Step (iii) uses the fact that $\Phi_{t+j} \geq \frac{1}{c}\Phi_t$, by Eqn. (12), and the setting of K. Manipulating Inequality (13) above we have

$$\begin{aligned}
F(w_t; S) - F(w_{t+K}; S) = \Phi_t - \Phi_{t+K} &\geq \frac{1}{2c\gamma^2 L_1}\Phi_t \\
\implies \Phi_{t+K} &\leq (1 - \frac{1}{2c\gamma^2 L_1})\Phi_t = \frac{1}{c}\Phi_t.
\end{aligned}$$

This establishes the contradiction and thus $\Phi_{t+K} \leq \max\left\{\frac{1}{c}\Phi_t, 2\left(\frac{\max\{\gamma^2,1\}L_0^2\log(n\sqrt{\rho}/\beta)}{\min\{L_1,1\}n\sqrt{\rho}}\right)^{\kappa/2}\right\}$.

■

We can now prove Theorem 13 itself. With the above two lemmas established, our primary concern is analyzing how the stopping conditions affect the convergence of the algorithm.

Proof [Proof of Theorem 13] Condition on the high probability event that Lemma 26 holds (i.e. that the gradient error is bounded for the entire trajectory). We will assume for the rest of the proof we assume that for all $t \in [T]$ that

$$\Phi_t \geq 2\left(\frac{\max\left\{\gamma^2, 1\right\} L_0^2 \log(n\sqrt{\rho}/\beta)}{\min\{L_1, 1\}n\sqrt{\rho}}\right)^{\kappa/2} \tag{14}$$

Note that if for any $t \in [T]$ the above inequality is not satisfied, by Lemma 26 the convergence guarantee of Theorem 13 is satisfied.

We now argue that the algorithm does not stop before convergence is reached by analyzing the stopping condition. It will be helpful to split the run of the algorithm into phases. We denote the first phase as the set of iterates $W_1 = w_0, w_1, ..., w_{K_1}$, where K_1 is the largest integer such that $F(w_{K_1}; S) - F(w_S^*; S) \geq \frac{1}{c}\Phi_0$. Similarly define $W_2 = w_{K_1}, w_{K_1+1}, ..., w_{K_2}$ where K_2 is the largest integer such that $F(w_{K_2}; S) - F(w_S^*; S) \geq \frac{1}{c}\Phi_{K_1}$, and so on for $W_3, W_4, ..., W_p$. Our aim is to show the algorithm does not stop before convergence.

First, we bound the largest value p can obtain without convergence. By Lemma 27 and Eqn. (14) we have $\Phi_{K_p} \leq \frac{1}{c^p}F_0$. Thus for $p \geq p_{\mathsf{max}} := (1 + 8\gamma^2 L_1)\left[\log(F_0) + \frac{2\kappa}{4-\kappa}\log(n\sqrt{\rho}/[\gamma L_0])\right] \geq \frac{\log(F_0)+\frac{2\kappa}{4-\kappa}\log(n\sqrt{\rho}/[\gamma L_0])}{\log(c)}$ we have

$$\Phi_{K_p} \leq \left(\frac{1}{c}\right)^{p_{\mathsf{max}}} F_0 \leq \left(\frac{\gamma L_0}{n\sqrt{\rho}}\right)^{\frac{2\kappa}{4-\kappa}} < \left(\frac{c\gamma L_0\sqrt{p_{\mathsf{max}}dA}}{n\sqrt{\rho}}\right)^{\frac{2\kappa}{4-\kappa}}.$$

Thus if the algorithm has not converged it must be the case that $p \leq p_{\mathsf{max}}$. Let us thus assume $p \leq p_{\mathsf{max}}$ for the following analysis.

The algorithm stops when $\sum_{t=0}^{T}\rho_t > \rho$. We observe (denoting $K_0 = 0$ for convenience)

$$\begin{aligned}
\sum_{t=0}^{T}\rho_t &= \frac{T\sqrt{\rho}}{n} + \frac{L_0^2 d\log(n\sqrt{\rho}/\beta)}{n^2}\sum_{j=1}^{p}\sum_{t=K_{j-1}}^{K_j}\frac{1}{N_t^2} \\
&\leq \frac{T\sqrt{\rho}}{n} + \frac{L_0^2 d\log(n\sqrt{\rho}/\beta)}{n^2}\sum_{j=1}^{p}\sum_{t=K_{j-1}}^{K_j}\left(\|\nabla F(w_t; S)\|^2 - \frac{L_0^2\log(n\sqrt{\rho}/\beta)}{4L_1 n\sqrt{\rho}}\right)^{-1} \\
&\leq \frac{T\sqrt{\rho}}{n} + \frac{L_0^2 d\log(n\sqrt{\rho}/\beta)}{n^2}\sum_{j=1}^{p}\sum_{t=K_{j-1}}^{K_j}\frac{2}{\|\nabla F(w_t; S)\|^2}
\end{aligned}$$

The last step uses the fact that the KL condition and the loss lower bound assumed in Eqn. (14) implies $\|\nabla F(w_t;S)\|^2 \geq \frac{L_0^2 \log(n\sqrt{\rho}/\beta)}{L_1 n\sqrt{\rho}}$. Continuing, by the KL condition we have,

$$\begin{aligned}
\sum_{t=0}^{T} \rho_t &\leq \frac{T\sqrt{\rho}}{n} + \frac{L_0^2 d \log(n\sqrt{\rho}/\beta)}{n^2} \sum_{j=1}^{p} \sum_{t=K_{j-1}}^{K_j} \frac{2\gamma^2}{\Phi_{K_j}^{2/\kappa}} \\
&\leq \frac{T\sqrt{\rho}}{n} + \frac{L_0^2 d \log(n\sqrt{\rho}/\beta)}{n^2} \sum_{j=1}^{p} \Phi_{K_{j-1}}^{\frac{\kappa-2}{\kappa}} \frac{2\gamma^2}{\Phi_{K_j}^{2/\kappa}} \\
&\leq \frac{T\sqrt{\rho}}{n} + \frac{L_0^2 d \log(n\sqrt{\rho}/\beta)}{n^2} \sum_{j=1}^{p} \Phi_{K_j}^{\frac{\kappa-2}{\kappa}} \frac{2\gamma^2}{\Phi_{K_j}^{2/\kappa}} \\
&\leq \frac{T\sqrt{\rho}}{n} + \frac{2\gamma^2 L_0^2 d \log(n\sqrt{\rho}/\beta)}{n^2} p_{\mathsf{max}} \max_{j\in[p]} \left\{ \Phi_{K_j}^{\frac{\kappa-4}{\kappa}} \right\}.
\end{aligned}$$

Thus, if $T \leq \frac{1}{2} n\sqrt{\rho}$, the algorithm has not stopped unless for some $t \in [T]$ we have $\Phi_t = O\left(\left(\frac{\gamma L_0 \sqrt{d \log(n\sqrt{\rho}/\beta) p_{\mathsf{max}}}}{n\sqrt{\rho}} \right)^{\frac{2\kappa}{4-\kappa}} \right)$.

To finish the proof, we consider the convergence when the algorithm stops after $T > \frac{1}{2} n\sqrt{\rho}$. Recall we are assuming the algorithm has run for at most $p \leq p_{\mathsf{max}}$ number of phases (as otherwise the algorithm has converged). The number of iterations during each of each of these phases is at most $\Phi_{K_p}^{\frac{\kappa-2}{\kappa}}$. Thus the algorithm has not stopped unless

$$p_{\mathsf{max}} \Phi_{K_p}^{\frac{\kappa-2}{\kappa}} \geq \frac{1}{2} n\sqrt{\rho} \implies \Phi_{K_p} \leq \left(\frac{2p_{\mathsf{max}}}{n\sqrt{\rho}} \right)^{\frac{\kappa}{2-\kappa}}.$$

To summarize, we now have three different bounds on the excess depending on three possible events. The first case is simply when $\Phi_T \leq 2 \left(\frac{\max\{\gamma^2,1\} L_0^2 \log(n\sqrt{\rho}/\beta)}{\min\{L_1,1\} n\sqrt{\rho}} \right)^{\kappa/2}$. The second case is when $\Phi_T \geq 2 \left(\frac{\max\{\gamma^2,1\} L_0^2 \log(n\sqrt{\rho}/\beta)}{\min\{L_1,1\} n\sqrt{\rho}} \right)^{\kappa/2}$ and $T \leq \frac{1}{2} n\sqrt{\rho}$, in which case we have shown $\Phi_T = O\left(\left(\frac{c\gamma L_0 \sqrt{d \log(n\sqrt{\rho}/\beta) p_{\mathsf{max}}}}{n\sqrt{\rho}} \right)^{\frac{2\kappa}{4-\kappa}} \right)$. The final case is when $\Phi_T \geq 2 \left(\frac{\max\{\gamma^2,1\} L_0^2 \log(n\sqrt{\rho}/\beta)}{\min\{L_1,1\} n\sqrt{\rho}} \right)^{\kappa/2}$ and $T > \frac{1}{2} n\sqrt{\rho}$, in which case we have shown $\Phi_T \leq \left(\frac{p_{\mathsf{max}}}{n\sqrt{\rho}} \right)^{\frac{\kappa}{2-\kappa}}$. Combining these results yields the theorem statement. ■

D.4. Proof of Lemma 12

Proof We will prove the lemma result by induction. For any $t \in 0, ..., T-1$, assuming $w_t \in \mathcal{S}$, we will show that $w_{t+1} \in \mathcal{S}$. The base case for w_0 holds because $\mathcal{S}$ is defined to contain w_0.

Before proceeding, we condition on the event that for all $t \in \{0, ..., T-1\}$ we have that $\|\nabla_t - \nabla F(w_t;S)\| \leq \|\nabla F(w_t;S)\| + \frac{L_0 \sqrt{\log(n\sqrt{\rho}/\beta)}}{\sqrt{n}\rho^{1/4}}$, which happens with probability at least $1 - 2\beta$ by Lemma 24 in Appendix D.1.

To prove the induction step, let $w_t \in \mathcal{S}$. We divide the proof into two cases, depending on $\|\nabla F(w_t; S)\|$. In the first case, assume $\|\nabla F(w_t; S)\| < \left(\frac{L_0^2 \log(n\sqrt{\rho}/\beta)}{n\sqrt{\rho}}\right)^{1/2}$. In this case, we will roughly prove that w_{t+1} is in $\mathcal{S}$ because it has not moved too far from w_t. Since $w_t \in \mathcal{S}$, the KL condition holds at w_t. Thus the gradient norm bound and the KL condition imply $F(w_t; S) - F(w_{\mathcal{S}}^*; S) \leq \left(\frac{\gamma L_0^2 \log(n\sqrt{\rho}/\beta)}{n\sqrt{\rho}}\right)^{\kappa/2}$. Let $R = 2\left(\frac{L_0^2 \log(n\sqrt{\rho}/\beta)}{L_1^2 n\sqrt{\rho}}\right)^{1/2}$ and recall we define the level set threshold as $\alpha = \max\left\{F(w_0; S), F(w_{\mathcal{S}}^*; S) + \max\left\{F(w_0; S), F(w_{\mathcal{S}}^*; S) + 2(\gamma^{\kappa/2} + L_0)\left(\frac{L_0^2 \log(n\sqrt{\rho}/\beta)}{n\sqrt{\rho}}\right)^{1/2}\right\}\right\}$. For any point $w' \in \mathcal{B}(w_t; R)$, by Lipschitzness one has

$$
\begin{aligned}
F(w'; S) &\leq F(w_t; S) - F(w_{\mathcal{S}}^*; s) + F(w_{\mathcal{S}}^*; S) + L_0(\|\nabla F(w_t; S)\| + \|\nabla_t - \nabla F(w_t; S)\|) \\
&\leq F(w_{\mathcal{S}}^*; S) + 2\left(\frac{\gamma L_0^2 \log(n\sqrt{\rho}/\beta)}{n\sqrt{\rho}}\right)^{\kappa/2} + L_0\left(\left(\frac{L_0^2 \log(n\sqrt{\rho}/\beta)}{n\sqrt{\rho}}\right)^{1/2} + \frac{L_0\sqrt{\log(n\sqrt{\rho}/\beta)}}{\sqrt{n}\rho^{1/4}}\right) \\
&\leq F(w_{\mathcal{S}}^*; S) + 2(\gamma^{\kappa/2} + L_0)\left(\frac{L_0^2 \log(n\sqrt{\rho}/\beta)}{n\sqrt{\rho}}\right)^{1/2} \leq \alpha.
\end{aligned}
$$

Thus $\mathcal{B}(w_t; R) \subseteq \mathcal{I}$. Since $\mathcal{B}(w_t; R)$ is path connected and $w_t \in \mathcal{S}$, we have $\mathcal{B}(w_t; R) \subseteq \mathcal{S}$ by the definition of $\mathcal{S}$. Further, with probability at least $1 - \beta$ we have

$$
\begin{aligned}
\|w_t - w_{t+1}\| &\leq \eta(\|\nabla F(w_t; S)\| + \|\nabla_t - \nabla F(w_t; S)\|) \\
&\overset{(i)}{\leq} \eta\left(2\|\nabla F(w_t; S)\| + \frac{L_0\sqrt{\log(n\sqrt{\rho}/\beta)}}{\sqrt{n}\rho^{1/4}}\right) \overset{(ii)}{\leq} \frac{3}{2L_1}\left(\frac{L_0^2 \log(n\sqrt{\rho}/\beta)}{n\sqrt{\rho}}\right)^{1/2} \leq R \\
&\implies w_{t+1} \in \mathcal{B}_R(w_t)
\end{aligned}
$$

Above, step (i) uses that the scale of noise in Algorithm 3 guarantees with high probability that $\|\nabla_t - \nabla F(w_t; S)\|$. Step (ii) uses the assumption that $\|\nabla F(w_t; S)\| \leq \left(\frac{L_0^2 \log(n\sqrt{\rho}/\beta)}{n\sqrt{\rho}}\right)^{1/2}$, the setting of R (above) and $\eta = \frac{1}{2L_1}$. As we have previously show, $\mathcal{B}_R(w_t) \subseteq \mathcal{S}$, so we have shown $w_{t+1} \in \mathcal{S}$.

We now consider the second case where $\|\nabla F(w_t; S)\| \geq \frac{L_0^2 \log(n\sqrt{\rho}/\beta)}{n\sqrt{\rho}}$. Consider the path parameterized by $l \in [0, 1]$ defined by $\mathbf{w}(l) = w_t + l(w_{t+1} - w_t)$. By the update rule of Algorithm 3 and standard descent lemma analysis we have (using the smoothness of $F(\cdot; S)$)

$$
\begin{aligned}
F(w_t; S) - F(\mathbf{w}(l); S) &\geq \langle \nabla F(w_t; S), w_t - \mathbf{w}(l)\rangle - \frac{L_1}{2}\|w_t - \mathbf{w}(l)\|^2 \\
&\geq l\langle \nabla F(w_t; S), w_t - w_{t+1}\rangle - \frac{l^2 L_1}{2}\|w_t - w_{t+1}\|^2 \\
&= \frac{l}{4L_1}\|\nabla F(w_t; S)\|^2 - \frac{l^2}{8L_1}\|\nabla_t - \nabla F(w_t; S)\|^2 \\
&\geq \frac{l}{4L_1}\|\nabla F(w_t; S)\|^2 - \frac{l^2 L_0^2 \log(n\sqrt{\rho}/\beta)}{8L_1 n\sqrt{\rho}} \\
&\overset{(i)}{\geq} l\left[\frac{1}{4L_1}\|\nabla F(w_t; S)\|^2 - \frac{L_0^2 \log(n\sqrt{\rho}/\beta)}{8L_1 n\sqrt{\rho}}\right] \\
&\overset{(ii)}{\geq} \frac{l}{4L_1}\|\nabla F(w_t; S)\|^2 \geq 0.
\end{aligned}
$$

Step (i) uses the fact that $l \leq 1$. Step (ii) uses the assumption that $\|\nabla F(w_t; S)\| \geq \frac{L_0^2 \log(n\sqrt{\rho}/\beta)}{n\sqrt{\rho}}$. We have shown $F(w_t; S) \geq F(\mathbf{w}(l); S)$ for every $l \in [0, 1]$. Thus $\{\mathbf{w}(l)\}_{l\in[0,1]} \subseteq \mathcal{I}$, and because $\{\mathbf{w}(l)\}_{l\in[0,1]}$ is path connected and contains $w_t \in \mathcal{S}$, we have $\{\mathbf{w}(l)\}_{l\in[0,1]} \subseteq \mathcal{S}$ and specifically $w_{t+1} \in \mathcal{S}$. ■

D.5. Proof of Lemma 25

Proof First note that when $\sigma_t = \frac{L_0}{n\sqrt{\rho}}$ then $\rho_t > \rho$ and the algorithm stops. Furthermore, in this case we also have $N_t \leq \frac{L_0\sqrt{d\log(n\sqrt{\rho}/\beta)}}{n\sqrt{\rho}}$, and thus by the concentration of the noise we have with probability at least $1 - \beta$ that

$$\|\nabla F(w_t; S)\| \leq \frac{L_0\sqrt{d\log(n\sqrt{\rho}/\beta)}}{n\sqrt{\rho}} + \frac{L_0\sqrt{\log(n\sqrt{\rho}/\beta)}}{\sqrt{n}\rho^{1/4}}$$

The KL condition then implies that

$$F(w; S) - F(w^*_{\mathcal{S}}; S) = O\left(\left(\frac{\gamma L_0\sqrt{d\log(n\sqrt{\rho}/\beta)}}{n\sqrt{\rho}}\right)^{\kappa} + \left(\frac{\gamma^2 L_0^2\log(n\sqrt{\rho}/\beta)}{n\sqrt{\rho}}\right)^{\kappa/2}\right)$$

■

D.6. Proof of Theorem 14 (Adaptive Gradient Descent without KL Condition)

In the following, we let $T + 1$ denote the largest value attained by t during the run of the algorithm.

Privacy Proof The prove privacy we will use the following lemma.

Lemma 28 *(Bun and Steinke, 2016, Proposition 1.4) Any algorithm which is $(\epsilon, 0)$-DP is also $(\frac{1}{2}\epsilon^2)$-zCDP.*

The process of releasing $w_0, ..., w_T$ is ρ-zCDP by Theorem 11. We thus only need to handel the additional privacy loss incurred via the use of the exponential mechanism to select t^*. Specifically, the exponential mechanism guarantees $(\sqrt{\rho}, 0)$-DP and is thus $\frac{1}{2}\rho$-zCDP by Lemma 28. The overall privacy is then 2ρ-zCDP.

Convergence Proof Recall $t^* \in [T]$ is the index sampled by the exponential mechanism. Let $N^* = \min_{t\in T}\{\|\nabla F(w_t; S)\|\}$. Note that the guarantees of the exponential mechanism (used to sample t^*) and scale of noise added to the gradient norm estimates we have with probability at least $1 - 2\beta$ that,

$$\begin{aligned}\|\nabla F(w_{t^*}; S)\| &= \|\nabla F(w_{t^*}; S)\| - N_{t^*} + N_{t^*} - N^* + N^* \\ &\leq \frac{L_0\log(n\sqrt{\rho}/\beta)}{\sqrt{n}\rho^{1/4}} + \frac{4L_0\log(n\sqrt{\rho}/\beta)}{n\sqrt{\rho}} + N^*, \end{aligned} \tag{15}$$

We will now proceed to bound N^*. First, if for any t one has $\sigma_t = \frac{2L_0}{n\sqrt{\rho}}$, then $N_t \leq \frac{L_0\sqrt{d}}{n\sqrt{\rho}}$ and thus $N^* \leq \frac{L_0\sqrt{d}}{n\sqrt{\rho}}$. The convergence guarantees are then satisfied by Eqn. (15).

We now turn towards the more difficult case where $\sigma_t = \frac{N_t}{\sqrt{d\log(n\sqrt{\rho}/\beta)}}$ for all $t \in [T]$. We start by analyzing the convergence of the algorithm in terms of the number of rounds T. By Lemma 26, Eqn. (9), we have with probability at least $1-\beta$ that,

$$F(w_t; S) - F(w_{t+1}; S) \geq \frac{1}{8L_1} \|\nabla F(w_t; S)\|^2 - \frac{L_0^2 \log(n\sqrt{\rho}/\beta)}{8L_1 n\sqrt{\rho}}.$$

Summing over all iterates and rearranging gives,

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(w_t; S)\| \leq \sqrt{\frac{8F_0L_1}{T}} + \frac{L_0\sqrt{\log(n\sqrt{\rho}/\beta)}}{\sqrt{n}\rho^{1/4}}. \tag{16}$$

We now consider the worst case guarantee for Algorithm 3. Recall $N^* = \min_{t\in T}\{\|\nabla F(w_t; S)\|\}$. We can use N^* to lower bound the number of iterations made by the algorithm. We have,

$$\begin{aligned}\sum_{t=0}^{T} \rho_t &= \frac{T\sqrt{\rho}}{2n} + \frac{L_0^2 d\log(n\sqrt{\rho}/\beta)}{n^2}\sum_{t=1}^{T}\frac{1}{N_t^2} \\ &\leq \frac{T\sqrt{\rho}}{2n} + \frac{TL_0^2 d\log(n\sqrt{\rho}/\beta)}{n^2(N^*)^2}.\end{aligned}$$

Further by the stopping condition we have,

$$\begin{aligned} T\left(\frac{L_0^2 d\log(n\sqrt{\rho}/\beta)}{n^2(N^*)^2} + \frac{\sqrt{\rho}}{2n}\right) &\geq \frac{\rho}{2} \\ \implies T \geq \frac{1}{2}\min\left\{\frac{n^2(N^*)^2\rho}{L_0^2 d\log(n\sqrt{\rho}/\beta)}, n\sqrt{\rho}\right\}&.\end{aligned}$$

By Eqn. (16) we also have,

$$N^* \leq \frac{1}{T}\sum_{t=1}^{T}\|\nabla F(w_t; S)\| \leq \sqrt{\frac{8F_0L_1}{T}} + \frac{L_0\log(n\sqrt{\rho}/\beta)}{\sqrt{n}\rho^{1/4}}.$$

Applying the above lower bound on T to the upper bound on N^* we obtain,

$$\begin{aligned} N^* &\leq \frac{3L_0\sqrt{F_0L_1 d\log(n\sqrt{\rho}/\beta)}}{nN^*\sqrt{\rho}} + \frac{L_0\log(n\sqrt{\rho}/\beta)}{\sqrt{n}\rho^{1/4}} \\ \implies N^* &= \left(\frac{6L_0\sqrt{F_0L_1 d\log(n\sqrt{\rho}/\beta)}}{n\sqrt{\rho}}\right)^{1/2} + \frac{2L_0\log(n\sqrt{\rho}/\beta)}{\sqrt{n}\rho^{1/4}}.\end{aligned}$$

Combining this bound with Eqn. (15) we have with probability at least $1-3\beta$ that,

$$\begin{aligned}\|\nabla F(\bar{w}; S)\| &\leq \min\left\{\sqrt{\frac{8F_0L_1}{T}} + \frac{L_0\sqrt{\log(n\sqrt{\rho}/\beta)}}{\sqrt{n}\rho^{1/4}},\right. \\ &\left. + \left(\frac{6L_0\sqrt{F_0L_1 d\log(n\sqrt{\rho}/\beta)}}{n\sqrt{\rho}}\right)^{1/2} + \frac{3L_0\log(n\sqrt{\rho}/\beta)}{\sqrt{n}\rho^{1/4}} + \frac{4L_0\log(n\sqrt{\rho}/\beta)}{n\sqrt{\rho}}\right\} \\ &= O\left(\min\left\{\sqrt{\frac{F_0L_1}{T}}, \left(\frac{L_0\sqrt{F_0L_1 d}}{n\sqrt{\rho}}\right)^{1/2}\right\} + \frac{L_0\sqrt{\log(n\sqrt{\rho}/\beta)}}{\sqrt{n}\rho^{1/4}}\right).\end{aligned}$$

Appendix E. Regularized Lipschitz Optimization

In this section, we consider a function $\tilde{f}(w;x) = f(w;x) + \tilde{L}_1 \|w - w_0\|^2$, where $w \mapsto f(w;x)$ is L_0-Lipschitz, $\tilde{L}_1$-weakly convex for all $x \in \mathcal{X}$, and $w_0 \in \mathbb{R}^d$. It is well known that in such case, the function $w \mapsto \tilde{f}(w;x)$ is $\tilde{L}_1$ strongly convex (see, e.g. (Davis and Drusvyatskiy, 2019b; Bassily et al., 2021a)). We denote the corresponding empirical risk as $\tilde{F}(w;S) = \frac{1}{n}\sum_{i=1}^n f(w;x_i) + \tilde{L}_1 \|w - w_0\|^2$.

The following result is a rate of $\tilde{O}\left(\frac{L_0^2 d}{\tilde{L}_1 n^2 \rho}\right)$ on excess empirical risk via Noisy Gradient Descent, Algorithm 4. Multiple works have investigated closely related settings (Feldman et al., 2020; Asi et al., 2021a), but due to our specific requirements (i.e. unconstrained setting and only assuming convexity of the regularized loss function) we provide a more tailored result here.

Algorithm 4 Noisy Gradient Descent

Require: Dataset S, zCDP paramter ρ, initial point $w_0 \in \mathbb{R}^d$, probability β, Lipschitz parameter L_0, Weak convexity $\tilde{L}_1$, step size sequence $\{\eta_t\}_t$, number of iterations T, noise standard deviation σ.

1: **for** $t = 1 \dots T-1$ **do**
2: $\quad \xi_t \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$
3: $\quad w_{t+1} = \Pi_{\mathcal{B}_{\frac{L_0}{2\tilde{L}_1}}(w_0)}\left(w_t - \eta_t\left(\nabla \tilde{F}(w;S) + \xi_t\right)\right)$
4: **end for**
5: Return $\bar{w} = \frac{2}{T(T+1)}\sum_{t=1}^T t w_t$

Theorem 29 *Let $\rho > 0$. Algorithm 4 with $T = \frac{n^2\rho \log^2(2/\beta)}{d}$, $\eta_t = \frac{1}{\tilde{L}_1 t}$ and $\sigma^2 = \frac{4L_0^2 T}{n^2\rho}$ satisfies ρ-zCDP. Further, with probability at least $1-\beta$, the excess empirical risk of its output, $\bar{w}$, is bounded as,*

$$\tilde{F}(\bar{w};S) - \tilde{F}(w^*;S) = O\left(\frac{L_0^2 d \log\left(n^2 \log^2(2/\beta)/d\beta\right)}{\tilde{L}_1 n^2 \rho}\right) \tag{17}$$

Proof The privacy proof is based on the observation that, even though the function $w \mapsto \tilde{f}(w;x)$ may not be Lispchitz, the sensitivity of the gradient, in every iteration, is controlled, since it is a sum of a Lipschitz and (data-independent) regularizer. In particular, the sensitivity of gradient at every iteration is bounded by $\frac{2L_0}{n}$. With the stated noise variance, applying the guarantee of Gaussian mechanism for zCDP and composition (Bun and Steinke, 2016), completes the privacy analysis.

The utility proof is based on standard high-probability convergence analysis of (S)GD for strongly convex optimization (Harvey et al., 2019). We first show that the unconstrained minimizer, w^*, lies in the constrained set $\mathcal{B}_{\frac{L_0}{2\tilde{L}_1}}(w_0)$. From the optimality criterion for unconstrained convex optimization, we have that $2\tilde{L}_1 \|w^* - w_0\| = \|\nabla F(w^*;S)\| \leq L_0$. This implies that $\|w^* - w_0\| \leq \frac{L_0}{2\tilde{L}_1}$. This also gives us that the function $w \mapsto \tilde{F}(w;S)$ is $2L_0$-Lipschitz over the constrained set.

From Gaussian concentration (Jin et al., 2019), we have that, with probability, at least $1 - \beta/2$, for all $t \in [T]$, $\|\xi_t\| \leq \sqrt{d \log(2T/\beta)}\sigma$. Further, conditioned on the above, we have from Lemma

4 in Harvey et al. (2019), that with probability at least $1-\beta$,

$$\sum_{t=1}^{T}\langle \xi_t, w_t - w^* \rangle = O\left(\frac{L_0}{\tilde{L}_1}\sqrt{d\log(2T/\beta)}\sigma\log(2/\beta)T\right).$$

The rest of the analysis is repeating the proof of Theorem 3.1 in Harvey et al. (2019). We get,

$$\begin{aligned}
\tilde{F}(\bar{w};S) - \tilde{F}(w^*;S) &= O\left(\frac{L_0^2}{\tilde{L}_1 T} + \frac{\sigma^2 d\log(2T/\beta)}{\tilde{L}_1 T}\right) + \frac{1}{T(T+1)}\sum_{t=1}^{T}\langle \xi_t, w_t - w^*\rangle \\
&= O\left(\frac{L_0^2}{\tilde{L}_1 T} + \frac{\sigma^2 d\log(2T/\beta)}{\tilde{L}_1 T} + \frac{L_0\sqrt{d\log(2T/\beta)}\sigma\log(2/\beta)}{\tilde{L}_1 T}\right) \\
&= O\left(\frac{L_0^2 d\log(n^2\log^2(2/\beta)/d\beta)}{\tilde{L}_1 n^2\rho}\right)
\end{aligned}$$

where the last step follows by setting of σ and T. ■