How to Make the Gradients Small Privately: Improved Rates for Differentially Private Non-Convex Optimization

Andrew Lowy [1] **Jonathan Ullman** [2] **Stephen J. Wright** [1]

Abstract

We provide a simple and flexible framework for designing differentially private algorithms to find approximate stationary points of non-convex loss functions. Our framework is based on using a private approximate risk minimizer to "warm start" another private algorithm for finding stationary points. We use this framework to obtain improved, and sometimes optimal, rates for several classes of non-convex loss functions. First, we obtain improved rates for finding stationary points of smooth non-convex empirical loss functions. Second, we specialize to quasar-convex functions, which generalize star-convex functions and arise in learning dynamical systems and training some neural nets. We achieve the *optimal* rate for this class. Third, we give an *optimal* algorithm for finding stationary points of functions satisfying the Kurdyka-Łojasiewicz (KL) condition. For example, over-parameterized neural networks often satisfy this condition. Fourth, we provide new state-of-the-art rates for stationary points of non-convex *population* loss functions. Fifth, we obtain improved rates for non-convex generalized linear models. A modification of our algorithm achieves nearly the same rates for *second-order* stationary points of functions with Lipschitz Hessian, improving over the previous state-of-the-art for each of the above problems.

1. Introduction

The increasing prevalence of machine learning (ML) systems, such as large language models (LLMs), in societal contexts has led to growing concerns about the privacy of these models. Extensive research has demonstrated that ML models can leak the training data of individuals, violating their privacy (Shokri et al., 2017; Carlini et al., 2021). For instance, individual training examples were extracted from GPT-2 using only black-box queries (Carlini et al., 2021). *Differential privacy* (DP) (Dwork et al., 2006) provides a rigorous guarantee that training data cannot be leaked. Informally, it guarantees that an adversary cannot learn much more about an individual piece of training data than they could have learned had that piece never been collected.

Differentially private optimization has been studied extensively over the last 10–15 years (Bassily et al., 2014; 2019; Feldman et al., 2020; Asi et al., 2021; Lowy & Razaviyayn, 2023b). Despite this large body of work, certain fundamental and practically important problems remain open. In particular, for minimizing *non-convex* functions, which is ubiquitous in ML applications, we have a poor understanding of the optimal rates achievable under DP.

In this work, we measure the performance of an algorithm for optimizing a non-convex function g by its ability to find an *α-stationary point*, meaning a point w such that

$$\|\nabla g(w)\| \leqslant \alpha.$$

We want to understand the smallest α achievable. There are several reasons to study stationary points. First, finding approximate global minima is intractable for general non-convex functions (Murty & Kabadi, 1985), but finding an approximate stationary point is tractable. Second, there are many important non-convex problems for which all stationary (or second-order stationary) points are global minima (e.g. phase retrieval (Sun et al., 2018), matrix completion (Ge et al., 2016), and training certain classes of neural networks (Liu et al., 2022)). Third, even for problems where it is tractable to find approximate global minima, the stationarity gap may be a better measure of quality than the excess risk (Nesterov, 2012; Allen-Zhu, 2018).

Stationary Points of Empirical Loss Functions. A fundamental open problem in DP optimization is determining the sample complexity of finding stationary points of non-

[1]University of Wisconsin-Madison, Wisconsin Institute of Discovery, Madison, WI, USA [2]Northeastern University, Khoury College of Computer Sciences, Boston, MA, USA. Correspondence to: Andrew Lowy <alowy@wisc.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

convex *empirical loss functions*

$$\widehat{F}_X(w) := \frac{1}{n}\sum_{i=1}^{n} f(w, x_i),$$

where $X = (x_1, \ldots, x_n)$ denotes a fixed data set. For *convex* loss functions, the minimax optimal complexity of DP empirical *risk minimization* is $\widehat{F}_X(w) - \min_{w'} \widehat{F}_X(w') = \widetilde{\Theta}(\sqrt{d\ln(1/\delta)}/\varepsilon n)$ (Bun et al., 2014; Bassily et al., 2014; Steinke & Ullman, 2016). Here d is the dimension of the parameter space and ε, δ are the privacy parameters. However, the algorithm of Bassily et al. (2014) was suboptimal in terms of finding DP stationary points. This gap was recently closed by (Arora et al., 2023), who showed that the optimal rate for stationary points of *convex* $\widehat{F}_X$ is $\mathbb{E}\|\nabla \widehat{F}_X(w)\| = \widetilde{\Theta}(\sqrt{d\ln(1/\delta)}/\varepsilon n)$. For *non-convex* $\widehat{F}_X$, the best known rate prior to 2022 was $O((\sqrt{d\ln(1/\delta)}/\varepsilon n)^{1/2})$ (Zhang et al., 2017; Wang et al., 2017; 2019). In the last two years, a pair of papers made progress and obtained improved rates of $\widetilde{O}((\sqrt{d\ln(1/\delta)}/\varepsilon n)^{2/3})$ (Arora et al., 2023; Tran & Cutkosky, 2022). Arora et al. (2023) gave a detailed discussion of the challenges of further improving beyond the $\widetilde{O}((\sqrt{d\ln(1/\delta)}/\varepsilon n)^{2/3})$ rate. Thus, a natural question is:

Question 1. Can we improve the $\widetilde{O}((\sqrt{d\ln(1/\delta)}/\varepsilon n)^{2/3})$ rate for DP stationary points of smooth non-convex empirical loss functions?

Contribution 1. We answer Question 1 affirmatively, giving a novel DP algorithm that finds a $\widetilde{O}((\sqrt{d\ln(1/\delta)}/\varepsilon n)d^{1/6})$-stationary point. This rate improves over the prior state-of-the-art whenever $d < n\varepsilon$.

Contribution 2. We provide algorithms that achieve the *optimal* rate $\widetilde{O}((\sqrt{d\ln(1/\delta)}/\varepsilon n))$ for two subclasses of non-convex loss functions: *quasar-convex* functions (Hinder et al., 2020), which generalize *star-convex* functions (Nesterov & Polyak, 2006), and *Kurdyka-Łojasiewicz* (KL) functions (Kurdyka, 1998), which generalize Polyak-Łojasiewicz (PL) functions (Polyak, 1963). Quasar-convex functions arise in learning dynamical systems and training recurrent neural nets (Hardt et al., 2018; Hinder et al., 2020). Also, the loss functions of some neural networks may be quasar-convex in large neighborhoods of the minimizers (Kleinberg et al., 2018; Zhou et al., 2019). On the other hand, the KL condition is satisfied by overparameterized neural networks in many scenarios (Bassily et al., 2018; Liu et al., 2020; Scaman et al., 2022). This is *the first time that the optimal rate has been achieved without assuming convexity*. To the best of our knowledge, no other DP algorithm in the literature would be able to get the optimal rate for either of these function classes.

Second-Order Stationary Points. Recently, Wang & Xu (2021); Gao & Wright (2023); Liu et al. (2023) provided DP algorithms for finding α-*second-order stationary points* (SOSP) of functions g with ρ-Lipschitz Hessian. A point w is an α-SOSP of g if w is an α-FOSP *and*

$$\nabla^2 g(w) \succeq -\sqrt{\alpha\rho}\,\mathbf{I}_d.$$

The state-of-the-art rate for α-SOSPs of empirical loss functions is due to Liu et al. (2023): $\alpha = \widetilde{O}((\sqrt{d\ln(1/\delta)}/\varepsilon n)^{2/3})$, which matches the state-of-the-art rate for FOSPs (Arora et al., 2023; Tran & Cutkosky, 2022).

Contribution 3. Our framework readily extends to SOSPs and achieves an improved $\widetilde{O}((\sqrt{d\ln(1/\delta)}/\varepsilon n)d^{1/6})$ second-order-stationarity guarantee.

Stationary Points of Population Loss Functions. Moving beyond empirical loss functions, we also consider finding stationary points of *population loss* functions

$$F(w) := \mathbb{E}_{x\sim\mathcal{P}}[f(w, x)],$$

where $\mathcal{P}$ is some unknown data distribution and we are given n i.i.d. samples $X \sim \mathcal{P}^n$. The prior state-of-the-art rate for finding SOSPs of F is $\widetilde{O}(1/n^{1/3} + (\sqrt{d}/\varepsilon n)^{3/7})$ (Liu et al., 2023).

Contribution 4. We give an algorithm that improves over the state-of-the-art rate for SOSPs of the population loss in the regime $d < n\varepsilon$. When $d = \Theta(1) = \varepsilon$, our algorithm is *optimal* and matches the *non-private* lower bound $\Omega(1/\sqrt{n})$.

We also specialize to (non-convex) *generalized linear models* (GLMs), which have been studied privately in (Song et al., 2021; Bassily et al., 2021a; Arora et al., 2022; 2023; Shen et al., 2023). GLMs arise, for instance, in robust regression (Amid et al., 2019) or when fine-tuning the last layers of a neural network. Thus, this problem has applications in privately fine-tuning LLMs (Yu et al., 2021; Li et al., 2021). Denoting the rank of the design matrix X by $r \leqslant \min(d, n)$, the previous state-of-the-art rate for finding FOSPs of GLMs was $O(1/\sqrt{n} + \min\{(\sqrt{r}/\varepsilon n)^{2/3}, 1/(\varepsilon n)^{2/5}\})$ (Arora et al., 2023).

Contribution 5. We provide improved rates of finding first- and second-order stationary points of the *population loss* of GLMs. Our algorithm finds a $\widetilde{O}(1/\sqrt{n} + \min\{(\sqrt{r}/\varepsilon n)r^{1/6}, 1/(\varepsilon n)^{2/7}\}$-stationary point, which is better than Arora et al. (2023) when $r < n\varepsilon$.

A summary of our main results is given in Table 1.

1.1. Our Approach

Our algorithmic approach is inspired by Nesterov, who proposed the following method for finding stationary points in non-private convex optimization: first run T steps of accelerated gradient descent (AGD) to obtain w_0, and then run T steps of gradient descent (GD) initialized at w_0 (Nesterov,

Loss Function	Previous SOTA	New SOTA	Lower Bound
Non-Convex Empirical	$\left(\frac{\sqrt{d}}{\varepsilon n}\right)^{2/3}$ (Liu et al., 2023)	$\left(\frac{\sqrt{d}}{\varepsilon n}\right)^{2/3} \wedge \frac{\sqrt{d}}{\varepsilon n} d^{1/6}$ (Cor. 4.4)	$\frac{\sqrt{d}}{\varepsilon n}$ (Arora et al., 2023)
Quasar-Convex Empirical	$\left(\frac{\sqrt{d}}{\varepsilon n}\right)^{2/3}$ (Liu et al., 2023)	$\frac{\sqrt{d}}{\varepsilon n}$ (Cor. 5.3 & Remark 5.4) (Optimal)	$\frac{\sqrt{d}}{\varepsilon n}$ (Arora et al., 2023)
KL Empirical	$\left(\frac{\sqrt{d}}{\varepsilon n}\right)^{2/3}$ (Liu et al., 2023)	$\frac{\sqrt{d}}{\varepsilon n}$ (Cor. 6.3 & Remark 6.4) (Optimal)	$\frac{\sqrt{d}}{\varepsilon n}$ (Lemma 6.5)
Non-Convex Population	$\frac{1}{n^{1/3}} + \left(\frac{\sqrt{d}}{\varepsilon n}\right)^{3/7}$ (Liu et al., 2023)	$\left(\frac{\zeta}{n}\right)^{1/3} + \left(\frac{\zeta\sqrt{d}}{\varepsilon n}\right)^{3/7}$ (Cor. 7.1) for $\zeta \leq 1$ defined in caption	$\frac{1}{n^{1/2}} + \frac{\sqrt{d}}{\varepsilon n}$ (Arora et al., 2023)
GLM Population	$\frac{1}{n^{1/2}} + \left(\frac{\sqrt{r}}{\varepsilon n}\right)^{2/3} \wedge \frac{1}{(\varepsilon n)^{2/5}} =: \alpha$ (Arora et al., 2023) (FOSP)	$\alpha \wedge \left[\frac{1}{n^{1/2}} + \left(\frac{\sqrt{r}}{\varepsilon n} r^{1/6} \wedge \frac{1}{(\varepsilon n)^{3/7}}\right)\right]$ (Cor. 8.2 & Remark 8.3)	N/A

Figure 1. Summary of results for second-order stationary points (SOSP). All bounds should be read as $\min(1, \ldots)$. SOTA = state-of-the-art. $\zeta := 1 \wedge \left(\frac{d}{\varepsilon n} + \sqrt{\frac{d}{n}}\right)$. $r := \text{rank}(X)$. We omit logarithms, Lipschitz and smoothness paramaters. The GLM algorithm of (Arora et al., 2023) only finds FOSP, not SOSP.

2012). Nesterov's approach provided improved stationary guarantees for convex loss functions, compared to running either AGD or GD alone.

We generalize and extend Nesterov's approach to private non-convex optimization. We first observe that there is nothing special about AGD or GD that makes his approach work. As we will see, one can obtain improved (DP) stationarity guarantees by running algorithm $\mathcal{B}$ after algorithm $\mathcal{A}$, provided that: (a) $\mathcal{A}$ moves us in the direction of a global minimizer, and (b) the stationarity guarantee of $\mathcal{B}$ benefits from a small initial suboptimality gap. Intuitively, the algorithm $\mathcal{A}$ functions as a "warm start" that gets us a bit closer to a global minimizer, which allows $\mathcal{B}$ to converge faster.

1.2. Roadmap

Section 2 contains relevant definitions, notations, and assumptions. In Section 3, we describe our general algorithmic framework and provide privacy and stationarity guarantees. The remaining sections contain applications of our algorithmic framework to non-convex empirical losses (Section 4), quasar-convex losses (Section 5), KL losses (Section 6), population losses (Section 7), and GLMs (Section 8).

2. Preliminaries

We consider loss functions $f : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is a data universe. For a data set $X \in \mathcal{X}^n$, let $\widehat{F}_X(w) := \frac{1}{n}\sum_{i=1}^n f(w, x_i)$ denote the empirical loss function. Let $F(w) := \mathbb{E}_{x\sim P}[f(w,x)]$ denote the population loss function with respect to some unknown data distribution P.

Assumptions and Notation.

Definition 2.1 (Lipschitz continuity)**.** Function $g : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz if $|g(w) - g(w')| \leqslant L\|w - w'\|_2$ for all $w, w' \in \mathbb{R}^d$.

Definition 2.2 (Smoothness)**.** Function $g : \mathbb{R}^d \to \mathbb{R}$ is β-smooth if g is differentiable and has β-Lipschitz gradient: $\|\nabla g(w) - \nabla g(w')\|_2 \leqslant \beta\|w - w'\|_2$.

We assume the following throughout:

Assumption 2.3.

1. $f(\cdot, x)$ is L-Lipschitz for all $x \in \mathcal{X}$.
2. $f(\cdot, x)$ is β-smooth for all $x \in \mathcal{X}$.
3. $\widehat{F}^*_X := \inf_w \widehat{F}_X(w) > -\infty$ for empirical loss optimization, or $F^* := \inf_w F(w) > -\infty$ for population.

Definition 2.4 (Stationary Points)**.** Let $\alpha \geqslant 0$. We say w is an *α-first-order-stationary point* (FOSP) of function g if $\|\nabla g(w)\| \leqslant \alpha$. If the Hessian $\nabla^2 g$ is ρ-Lipschitz, then w is an *α-second-order-stationary point* (SOSP) of g if $\|\nabla g(w)\| \leqslant \alpha$ and $\nabla^2 g(w) \succeq -\sqrt{\rho\alpha}\,\mathbf{I}_d$.

For functions $a = a(\theta)$ and $b = b(\phi)$ of input parameter vectors θ and ϕ, we write $a \lesssim b$ if there is an absolute constant $C > 0$ such that $a \leqslant Cb$ for all values of input parameter vectors θ and ϕ. We use $\tilde{O}$ to hide logarithmic factors. Denote $a \wedge b = \min(a, b)$.

Differential Privacy.

Definition 2.5 (Differential Privacy (Dwork et al., 2006))**.** Let $\varepsilon \geqslant 0$, $\delta \in [0, 1)$. A randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$ is *(ε, δ)-differentially private* (DP) if for all pairs of data sets $X, X' \in \mathcal{X}^n$ differing in one sample and all measurable subsets $S \subseteq \mathcal{W}$, we have

$$\mathbb{P}(\mathcal{A}(X) \in S) \leqslant e^{\varepsilon}\mathbb{P}(\mathcal{A}(X') \in S) + \delta.$$

An important fact about DP is that it composes nicely:

Algorithm 1 DP-SPIDER (Arora et al., 2023)

Input: Data $X \in \mathcal{X}^n$, loss function $f(w,x)$, (ε,δ), initialization w_0, stepsize η, iteration number T, phase length q, noise variances $\sigma_1^2, \sigma_2^2, \hat{\sigma}_2^2$, batch sizes b_1, b_2.
for $t = 0, \ldots, T-1$ **do**
  if $q|t$ **then**
    Sample batch S_t of size b_1
    Sample $g_t \sim \mathcal{N}(0, \sigma_1^2 \mathbf{I}_d)$
    $\nabla_t = \frac{1}{b_1}\sum_{x \in S_t} \nabla f(w_t, x) + g_t$
  else
    Sample batch S_t of size b_2
    Sample $h_t \sim \mathcal{N}(0, \min\{\sigma_2^2 \|w_t - w_{t-1}\|^2, \hat{\sigma_2}^2\}\mathbf{I}_d)$
    $\Delta_t = \frac{1}{b_2}\sum_{x \in S_t}[\nabla f(w_t, x) - \nabla f(w_{t-1}, x)] + h_t$
    $\nabla_t = \nabla_{t-1} + \Delta_t$
  end if
  $w_{t+1} = w_t - \eta \nabla_t$
end for
Return: $\hat{w} \sim$ **Unif**$(w_1, \ldots, w_T)$.

Lemma 2.6 (Basic Composition)**.** *If $\mathcal{A}$ is $(\varepsilon_1, \delta_1)$-DP and $\mathcal{B}$ is $(\varepsilon_2, \delta_2)$-DP, then $\mathcal{B} \circ \mathcal{A}$ is $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-DP.*

3. Our Warm-Start Algorithmic Framework

For ease of presentation, we will first present a concrete instantiation of our algorithmic framework for ERM, built upon the DP-SPIDER algorithm of Arora et al. (2023), which is described in Algorithm 1.

For initialization $w_0 \in \mathbb{R}^d$, denote the suboptimality gap by

$$\hat{\Delta}_{w_0} := \widehat{F}_X(w_0) - \widehat{F}_X^*.$$

We recall the guarantees of DP-SPIDER below:

Lemma 3.1. *(Arora et al., 2023) There exist algorithmic parameters such that Algorithm 1 is $(\varepsilon/2, \delta/2)$-DP and returns $\hat{w}$ satisfying*

$$\mathbb{E}\|\nabla \widehat{F}_X(\hat{w})\| \lesssim \left(\frac{\sqrt{\hat{\Delta}_{w_0} L\beta}\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)^{2/3} + \frac{L\sqrt{d\ln(1/\delta)}}{\varepsilon n}. \tag{1}$$

Typically, the first term on the RHS of (1) is dominant.

Our algorithm is based on a simple observation: the stationarity guarantee in Lemma 3.1 depends on the initial suboptimality gap $\hat{\Delta}_{w_0}$. Therefore, if we can privately find a good "warm start" point w_0 such that $\widehat{F}_X(w_0) - \widehat{F}_X^*$ is small with high probability, then we can run DP-SPIDER initialized at w_0 to improve over the $O((\sqrt{d}/\varepsilon n)^{2/3})$ guarantee of DP-SPIDER. More generally, we can apply any DP stationary point finder $\mathcal{B}$ with initialization w_0 after warm starting. Pseudocode for our general meta-algorithm is given in Algorithm 2.

Algorithm 2 Warm-Start Meta-Algorithm for ERM

Input: Data $X \in \mathcal{X}^n$, loss function $f(w,x)$, privacy parameters (ε,δ), warm-start DP-ERM algorithm $\mathcal{A}$, DP-ERM stationary point finder $\mathcal{B}$.
Run $(\varepsilon/2, \delta/2)$-DP $\mathcal{A}$ on $\widehat{F}_X(\cdot)$ to obtain w_0.
Run $\mathcal{B}$ on $\widehat{F}_X(\cdot)$ with initialization w_0 and privacy parameters $(\varepsilon/2, \delta/2)$ to obtain w_{priv}.
Return: w_{priv}.

We have the following guarantee for Algorithm 2 instantiated with $\mathcal{B} =$ Algorithm 1.

Theorem 3.2 (First-Order Stationary Points for ERM: Meta-Algorithm)**.** *Let $\zeta \leqslant \sqrt{d}/\varepsilon n$. Suppose $\mathcal{A}$ is $(\varepsilon/2, \delta/2)$-DP and $\widehat{F}_X(\mathcal{A}(X)) - \widehat{F}_X^* \leqslant \psi$ with probability $\geqslant 1 - \zeta$. Then, Algorithm 2 with $\mathcal{B}$ as DP-SPIDER is (ε, δ)-DP and returns w_{priv} with*

$$\mathbb{E}\|\nabla \widehat{F}_X(w_{priv})\| \lesssim \frac{L\sqrt{d\ln(1/\delta)}}{\varepsilon n} + L^{1/3}\beta^{1/3}\psi^{1/3}\left(\frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)^{2/3}.$$

Proof. Privacy follows from Lemma 2.6, since $\mathcal{A}$ and DP-SPIDER are both $(\varepsilon/2, \delta/2)$-DP.

For the stationarity guarantee, let E be the high-probability good event that $\widehat{F}_X(\mathcal{A}(X)) - \widehat{F}_X^* \leqslant \psi$. Then, by Lemma 3.1, we have

$$\mathbb{E}\left[\|\nabla \widehat{F}_X(w_{\text{priv}})\| | E\right] \lesssim \left(\frac{\sqrt{\psi L\beta}\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)^{2/3} + \frac{L\sqrt{d\ln(1/\delta)}}{\varepsilon n}.$$

On the other hand, if E does not hold, then we still have $\|\nabla \widehat{F}_X(w_{\text{priv}})\| \leqslant L$ by Lipschitz continuity. Thus, taking total expectation yields

$$\begin{aligned}\mathbb{E}\|\nabla \widehat{F}_X(w_{\text{priv}})\| &\leqslant \mathbb{E}\left[\|\nabla \widehat{F}_X(w_{\text{priv}})\| | E\right](1-\zeta) + L\zeta \\ &\lesssim \left(\frac{\sqrt{\psi L\beta}\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)^{2/3} + \frac{L\sqrt{d\ln(1/\delta)}}{\varepsilon n} + L\zeta.\end{aligned}$$

Since $\zeta \leqslant \sqrt{d}/\varepsilon n$, the result follows. $\square$

Note that if we instantiate Algorithm 2 with any DP $\mathcal{B}$, we can obtain an algorithm that improves over the stationarity guarantee of $\mathcal{B}$ as long as the stationarity guarantee of $\mathcal{B}$ scales with the initial suboptimality gap $\hat{\Delta}_{w_0}$. In particular, our framework allows for improved rates of finding *second-order* stationarity points, by choosing $\mathcal{B}$ as the DP SOSP finder of Liu et al. (2023) (which is built on DP-SPIDER). We recall the privacy and utility guarantees of this algorithm—which we refer to as *DP-SPIDER-SOSP*—below in Lemma 3.3. For convenience, denote

$$\alpha := \left(\frac{\sqrt{\hat{\Delta}_{w_0} L\beta}\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)^{2/3} + \frac{L\sqrt{d\ln(1/\delta)}}{\varepsilon n} + \frac{\beta}{n\sqrt{\rho}}\left(\frac{\sqrt{\hat{\Delta}_{w_0} L\beta}\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)^{1/3}.$$

Lemma 3.3. *(Liu et al., 2023) Assume that $f(\cdot, x)$ has ρ-Lipschitz Hessian $\nabla^2 f(\cdot, x)$. Then, there is an $(\varepsilon/2, \delta/2)$-DP Algorithm (DP-SPIDER-SOSP), that returns $\hat{w}$ such that with probability $\geqslant 1-\zeta$, $\hat{w}$ is a $\widetilde{O}(\alpha)$-SOSP of $\widehat{F}_X$.*

Next, we provide the guarantee of Algorithm 2 instantiated with $\mathcal{B}$ as DP-SPIDER-SOSP:

Theorem 3.4 (Second-order Stationary Points for ERM: Meta-Algorithm)**.** *Suppose $\mathcal{A}$ is $(\varepsilon/2, \delta/2)$-DP and $\widehat{F}_X(\mathcal{A}(X)) - \widehat{F}_X^* \leqslant \psi$ with probability $\geqslant 1-\zeta$. Then, Algorithm 2 with $\mathcal{B}$ as DP-SPIDER-SOSP is (ε, δ)-DP, and with probability $\geqslant 1-2\zeta$ has output w_{priv} satisfying*

$$\|\nabla \widehat{F}_X(w_{priv})\| \leqslant \tilde{\alpha} := \widetilde{O}\left(\frac{L\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right) + \widetilde{O}\left(L^{1/3}\beta^{1/3}\psi^{1/3}\left(\frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)^{2/3}\right) + \widetilde{O}\left(\frac{\beta^{7/6}L^{1/6}\psi^{1/6}}{n\sqrt{\rho}}\left(\frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)^{1/3}\right),$$

and

$$\nabla^2 \widehat{F}_X(w_{priv}) \succeq -\sqrt{\rho\tilde{\alpha}}\,\mathbf{I}_d.$$

The proof is similar to the proof of Theorem 3.2, and is deferred to Appendix C.

With Algorithm 2, we have reduced the problem of finding an approximate stationary point w_{priv} to finding an approximate excess risk minimizer w_0. The next question is: *What should we choose as our warm-start algorithm $\mathcal{A}$?* In general, one should choose $\mathcal{A}$ that achieves the smallest possible risk for a given function class.[1] In the following sections, we consider different classes of non-convex functions and instantiate Algorithm 2 with an appropriate warm-start $\mathcal{A}$ for each class to obtain new state-of-the-art rates.

[1]In particular, if there exists a DP algorithm with *optimal* risk, then this algorithm is the optimal choice of warm starter.

4. Improved Rates for Stationary Points of Non-Convex Empirical Losses

In this section, we provide improved rates for finding (first-order and second-order) stationary points of smooth non-convex empirical loss functions. For the non-convex loss functions satisfying Assumption 2.3, we propose using the *exponential mechanism* (McSherry & Talwar, 2007) as our warm-start algorithm $\mathcal{A}$ in Algorithm 2.

We now recall the exponential mechanism. Assume that there is a compact set $\mathcal{W} \subset \mathbb{R}^d$ containing an approximate global minimizer w^* such that $\widehat{F}_X(w^*) - \widehat{F}_X^* \leqslant LD\frac{d}{\varepsilon n}$, and that $\|w - w'\|_2 \leqslant D$ for all $w, w' \in \mathcal{W}$. Note that there exists a finite $D\frac{d}{\varepsilon n}$-net for $\mathcal{W}$, denoted $\widetilde{\mathcal{W}} = \{w_1, \ldots, w_N\}$, with $N := |\widetilde{\mathcal{W}}| \leqslant \left(\frac{2D\varepsilon n}{d}\right)^d$. In particular, $\min_{i\in[N]} \widehat{F}_X(w_i) - \widehat{F}_X^* \leqslant 2LD\frac{d}{\varepsilon n}$.

Definition 4.1 (Exponential Mechanism for ERM)**.** Given inputs $\widehat{F}_X, \widetilde{\mathcal{W}}$, the exponential mechanism $\mathcal{A}_E$ selects and outputs some $w \in \widetilde{\mathcal{W}}$. The probability that a particular w is selected is proportional to $\exp\left(\frac{-\varepsilon n \widehat{F}_X(w)}{4LD}\right)$.

The following lemma specializes (Dwork & Roth, 2014, Theorem 3.11) to our ERM setting:

Lemma 4.2. *The exponential mechanism $\mathcal{A}_E$ is ε-DP. Moreover, $\forall t > 0$, we have with probability at least $1 - \exp(-t)$ that*

$$\widehat{F}_X(\mathcal{A}_E) - \widehat{F}_X(w^*) \leqslant \frac{4LD}{\varepsilon n}\ln\left(\left(\frac{2\varepsilon n}{d}\right)^d + t\right) + 2LD\frac{d}{\varepsilon n}.$$

First-Order Stationary Points. For convenience, denote

$$\gamma := \frac{L\sqrt{d\ln(1/\delta)}}{\varepsilon n} + \widetilde{O}\left(L^{2/3}\beta^{1/3}D^{1/3}\frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n}d^{1/6}\right). \tag{2}$$

By substituting $\varepsilon/2$ for ε and then choosing $t = \ln(\varepsilon n/2\sqrt{d})$ in Lemma 4.2, the $\varepsilon/2$-exponential mechanism returns a point w_0 such that

$$\widehat{F}_X(w_0) - \widehat{F}_X^* \leqslant 20LD\frac{d}{\varepsilon n}\ln(\varepsilon n/\sqrt{d}) =: \psi \tag{3}$$

with probability at least $1 - 2\frac{\sqrt{d}}{\varepsilon n}$. By plugging the above ψ into Theorem 3.2, we obtain:

Corollary 4.3 (First-Order Stationary Points for Non-Convex ERM)**.** *There exist algorithmic parameters such that*

Algorithm 2 with $\mathcal{A} = \mathcal{A}_E$ and $\mathcal{B}$ =DP-SPIDER is (ε, δ)-DP and returns w_{priv} such that

$$\mathbb{E}\|\nabla \widehat{F}_X(w_{priv})\| \lesssim \gamma.$$

If L, β, D are constants, then Corollary 4.3 gives $\mathbb{E}\|\nabla \widehat{F}_X(w_{\text{priv}})\| = \widetilde{O}\left(\frac{\sqrt{d \ln(1/\delta)}}{\varepsilon n} d^{1/6}\right)$. This bound is bigger than the lower bound by a factor of $d^{1/6}$ and improves over the previous state-of-the-art $O\left(\frac{\sqrt{d \ln(1/\delta)}}{\varepsilon n}\right)^{2/3}$ whenever $d < n\varepsilon$ (Arora et al., 2023). If $d \geqslant n\varepsilon$, then one should simply run DP-SPIDER. Combining these two algorithms gives a new state-of-the-art bound for DP stationary points of non-convex empirical loss functions:

$$\mathbb{E}\|\nabla \widehat{F}_X(w_{\text{priv}})\| \lesssim \frac{\sqrt{d \ln(1/\delta)}}{\varepsilon n} d^{1/6} \wedge \left(\frac{\sqrt{d \ln(1/\delta)}}{\varepsilon n}\right)^{2/3}.$$

Challenges of Further Rate Improvements. We believe that it is not possible for Algorithm 2 to achieve a better rate than Corollary 4.3 by choosing $\mathcal{A}$ differently. The exponential mechanism is optimal for non-convex Lipschitz empirical risk minimization (Ganesh et al., 2023). Although the lower bound function in Ganesh et al. (2023) is not β-smooth, we believe that one can smoothly approximate it (e.g. by piecewise polynomials) to extend the same lower bound to smooth functions. For large enough β, their lower bound extends to smooth losses by simple convolution smoothing. Thus, a fundamentally different algorithm may be needed to find $O(\sqrt{d \ln(1/\delta)}/\varepsilon n)$-stationary points for general non-convex empirical losses.

Second-Order Stationary Points. If we assume that f has Lipschitz continuous Hessian, then we can instantiate Algorithm 2 with $\mathcal{B}$ as DP-SPIDER-SOSP to obtain:

Corollary 4.4 (Second-Order Stationary Points for Non–Convex ERM)**.** *Let $\zeta > 0$. Suppose $\nabla^2 f(\cdot, x)$ is ρ-Lipschitz $\forall x$. Then, Algorithm 2 with $\mathcal{A} = \mathcal{A}_E$ and $\mathcal{B}$ = DP-SPIDER-SOSP is (ε, δ)-DP and with probability $\geqslant 1 - \zeta$, returns a ω-SOSP, where*

$$\omega := \gamma + \widetilde{O}\left(\frac{L^{1/3} D^{1/6} \beta^{7/6}}{\sqrt{\rho} n} \left(\frac{\sqrt{d \ln(1/\delta)}}{\varepsilon n}\right)^{1/2} d^{1/12}\right),$$

If L, β, D and ρ are constants, then Corollary 4.4 implies that Algorithm 2 finds a $\widetilde{O}(d^{1/6}\sqrt{d \ln(1/\delta)}/\varepsilon n)$-*second-order* stationary point of $\widehat{F}_X$. This result improves over the previous state-of-the-art (Liu et al., 2023) when $d < n\varepsilon$.

5. Optimal Rate for Quasar-Convex Losses

In this section, we specialize to *quasar-convex* loss functions (Hardt et al., 2018; Hinder et al., 2020) and show, for the first time, that it is possible to attain the optimal (up to logs) rate $\widetilde{O}(\sqrt{d \ln(1/\delta)}/\varepsilon n)$ for stationary points, *without assuming convexity*.

Definition 5.1 (Quasar-convex functions)**.** Let $q \in (0, 1]$ and let w^* be a minimizer of differentiable function $g : \mathbb{R}^d \to \mathbb{R}$. g is *q-quasar convex* if for all $w \in \mathbb{R}^d$, we have

$$g(w^*) \geqslant g(w) + \frac{1}{q}\langle \nabla g(w), w^* - w \rangle.$$

Quasar-convex functions generalize star-convex functions (Nesterov & Polyak, 2006), which are quasar-convex functions with $q = 1$. Smaller values of $q < 1$ allow for a greater degree of non-convexity.

Proposition 5.2 shows that returning a uniformly random iterate of DP-SGD (Algorithm 3) attains essentially the same (optimal) rate for quasar-convex ERM as for convex ERM:

Algorithm 3 DP-SGD for Quasar-Convex

1: **Input:** Loss function f, data X, iteration number T noise variance σ^2, step size η, batch size b.
2: Initialize $w_1 \in \mathbb{R}^d$.
3: **for** $t \in \{1, 2, \cdots, T\}$ **do**
4: Sample batch S_t of size b from X
5: Sample $u_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$
6: $\nabla_t = \frac{1}{b}\sum_{x \in S_t} \nabla f(w_t, x) + u_t$
7: $w_{t+1} = w_t - \eta \nabla_t$
8: **end for**
9: **Output:** $\hat{w} \sim$ **Unif**$(w_1, \ldots, w_T)$.

Proposition 5.2. *Let $\widehat{F}_X$ be q-quasar convex and $\|w_1 - w^*\| \leqslant D$ for $w^* \in \mathrm{argmin}_w \widehat{F}_X(w)$. Then, there are algorithmic parameters such that Algorithm 3 is (ε, δ)-DP, and returns $\hat{w}$ such that*

$$\mathbb{E}\widehat{F}_X(\hat{w}) - \widehat{F}_X^* \lesssim LD\frac{\sqrt{d \ln(1/\delta)}}{\varepsilon n q}.$$

Further, $\forall\ \zeta > 0$, there is an (ε, δ)-DP variation of Algorithm 3 that returns $\tilde{w}$ s.t. with probability at least $1 - \zeta$,

$$\widehat{F}_X(\tilde{w}) - \widehat{F}_X^* = \widetilde{O}\left(LD\frac{\sqrt{d \ln(1/\delta)}}{\varepsilon n q}\right).$$

See Appendix D for a proof. The same proof works for non-smooth quasar-convex losses if we replace gradients by subgradients in Algorithm 3. As a byproduct, our proof yields a novel non-private optimization result: SGD achieves the optimal $O(1/\sqrt{T})$ rate for Lipschitz non-smooth quasar-convex stochastic optimization. To our knowledge, this result was only previously recorded for smooth losses (Gower et al., 2021) or convex losses (Nesterov, 2013).

By combining Proposition 5.2 with Theorem 3.2, we obtain:

Corollary 5.3 (Quasar-Convex ERM)**.** *Let $\widehat{F}_X$ be q-quasar convex and $\|w_1 - w^*\| \leqslant D$ for some $w_1 \in \mathbb{R}^d, w^* \in \operatorname{argmin}_w \widehat{F}_X(w)$. Then, there are algorithmic parameters such that Algorithm 2 with $\mathcal{A}$ = Algorithm 3 and $\mathcal{B}$ = DP-SPIDER is (ε, δ)-DP and returns w_{priv} such that*

$$\mathbb{E}\|\nabla \widehat{F}_X(w_{priv})\| \lesssim L\frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n} + \widetilde{O}\left(L^{2/3}\beta^{1/3}D^{1/3}\frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n q}\right).$$

If q is constant and $\beta D \lesssim L$, then this rate is optimal up to a logarithmic factor, since it matches the convex (hence quasar-convex) lower bound of Arora et al. (2023).

Remark 5.4. One can obtain a second-order stationary point with essentially the same (near-optimal) rate by appealing to Theorem 3.4 instead of Theorem 3.2.

6. Optimal Rates for KL* Empirical Losses

In this section, we derive optimal rates (up to logarithms) for functions satisfying the Kurdyka-Łojasiewicz* (KL*) condition (Kurdyka, 1998):

Definition 6.1. Let $\gamma, k > 0$. Function $g : \mathbb{R}^d \to \mathbb{R}$ satisfies the (γ, k)-*KL** condition on $\mathcal{W} \subset \mathbb{R}^d$ if

$$g(w) - \inf_{w' \in \mathbb{R}^d} g(w') \leqslant \gamma^k \|\nabla g(w)\|^k,$$

for all $w \in \mathcal{W}$. If $k = 2$ and $\gamma = \sqrt{1/2\mu}$, say g satisfies the μ-*PL** condition on $\mathcal{W}$.

The KL* (PL*) condition relaxes the KL (PL) condition, by requiring it to only hold on a *subset* of $\mathbb{R}^d$.

Near-optimal *excess risk* guarantees for the KL* class were recently provided in (Menart et al., 2023):

Lemma 6.2. *(Menart et al., 2023, Theorem 1) Assume $\widehat{F}_X$ satisfies the (γ, k)-KL* condition for some $k \in [1, 2]$ on a centered ball $B(0, D)$ of diameter $D = \frac{\hat{\Delta}_0^{1/k}}{\gamma\beta} + \hat{\Delta}_0^{(k-1)/k}\gamma$. Then, there is an $(\varepsilon/2, \delta/2)$-DP algorithm with output w_0 such that with probability at least $1 - \zeta$,*

$$\widehat{F}_X(w_0) - \widehat{F}_X^* \leqslant \widetilde{O}\left(\left[\frac{\gamma L\sqrt{d\ln(1/\delta)}}{\varepsilon n}\sqrt{1 + \left(1/\hat{\Delta}_0\right)^{(2-k)/k}\gamma^2\beta}\right]^k\right)$$

The KL* condition implies that any approximate stationary point is an approximate excess risk minimizer, but the converse is false. The algorithm of Menart et al. (2023) does not lead to (near-optimal) guarantees for stationary points. However, using it as the warm-start algorithm $\mathcal{A}$ in Algorithm 2 gives near-optimal rates for stationary points:

Corollary 6.3 (KL* ERM)**.** *Grant the assumptions in Lemma 6.2. Then, Algorithm 2 with $\mathcal{A}$ = the algorithm in Lemma 6.2 and $\mathcal{B}$ = DP-SPIDER is (ε, δ)-DP and returns w_{priv} such that*

$$\mathbb{E}\|\nabla \widehat{F}_X(w_{priv})\| \lesssim \frac{L\sqrt{d\ln(1/\delta)}}{\varepsilon n} + \widetilde{O}\left(\frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)^{\frac{k+2}{3}}\left(L^{k+1}\beta\gamma^k\right)^{\frac{1}{3}}\left(1 + \frac{(\gamma\sqrt{\beta})^{k/3}}{\hat{\Delta}_0^{\frac{2-k}{6}}}\right).$$

In particular, if $(\gamma\sqrt{\beta})^{k/3}/\hat{\Delta}_0^{\frac{2-k}{6}} \lesssim 1$ and $\left(\frac{\beta\gamma^k}{L^{2-k}}\right)^{1/(k-1)} \lesssim n\varepsilon/\sqrt{d\ln(1/\delta)}$, then

$$\mathbb{E}\|\nabla \widehat{F}_X(w_{priv})\| = \widetilde{O}\left(\frac{L\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right).$$

Proof. Algorithm 2 is (ε, δ)-DP by Theorem 3.2. Further, combining Theorem 3.2 with Lemma 6.2 implies Corollary 6.3: plug the right-hand-side of the risk bound in Corollary 6.3 for ψ in Theorem 3.2. ☐

As an example: If $\widehat{F}_X$ is μ-PL* for $\beta/\mu \lesssim (\varepsilon n/\sqrt{d\ln(1/\delta)})$, then our algorithm achieves $\mathbb{E}\|\nabla \widehat{F}_X(w_{\text{priv}})\| = \widetilde{O}(L\sqrt{d\ln(1/\delta)}/\varepsilon n)$.

Remark 6.4. If $L, \beta, \gamma, \hat{\Delta}_0$ are constants, then we get the same rate as Corollary 6.3 for *second-order* stationary points by using Algorithm 2 with $\mathcal{B}$ as DP-SPIDER-SOSP instead of DP-SPIDER.

We show next that Corollary 6.3 is optimal up to logarithms:

Lemma 6.5 (Lower bound for KL*)**.** *Let $D, L, \beta, \gamma > 0$ and $k \in (1, 2]$ such that $k = 1 + \Omega(1)$. For any (ε, δ)-DP algorithm $\mathcal{M}$, there exists a data set X and L-Lipschitz, β-smooth $f(\cdot, x)$ that is (γ, k)-KL over $B(0, D)$ such that*

$$\mathbb{E}\|\nabla \widehat{F}_X(\mathcal{M}(X))\| = \widetilde{\Omega}\left(L\min\left\{1, \frac{\sqrt{d}}{\varepsilon n}\right\}\right).$$

In contrast to the excess risk setting of Lemma 6.2, larger k does not allow for faster rates of stationary points. Lemma 6.5 is a consequence of the KL* excess risk lower bound (Menart et al., 2023, Corollary 1) and Definition 6.1.

7. Improved Rates for Stationary Points of Non-Convex Population Loss

Suppose that we are given n i.i.d. samples from an unknown distribution $\mathcal{P}$ and our goal is to find an α-second-order stationary point of the population loss $F(w) = \mathbb{E}_{x\sim\mathcal{P}}[f(w, x)]$. Our framework for finding DP approximate stationary points of F is described in Algorithm 4. It is

Algorithm 4 Warm-Start Meta-Algorithm for Pop. Loss

1: **Input:** Data $X \in \mathcal{X}^n$, loss function $f(w,x)$, privacy parameters (ε,δ), warm-start DP risk minimization algorithm $\mathcal{A}$, DP stationary point finder $\mathcal{B}$.
2: Run $(\varepsilon/2,\delta/2)$-DP $\mathcal{A}$ to obtain $w_0 \approx \operatorname{argmin}_w F(w)$.
3: Run $\mathcal{B}$ with initialization w_0 and privacy parameters $(\varepsilon/2,\delta/2)$ to obtain w_{priv}.
4: **Return:** w_{priv}.

a population-loss analog of the warm-start meta-Algorithm 2 for stationary points of $\widehat{F}_X$.

We present the guarantees for Algorithm 4 with generic $\mathcal{A}$ and $\mathcal{B}$ (analogous to Theorem 3.4) in Theorem E.2 in Appendix E. By taking $\mathcal{A}$ to be the $\varepsilon/2$-DP exponential mechanism and $\mathcal{B}$ to be the $(\varepsilon/2,\delta/2)$-DP-SPIDER-SOSP of Liu et al. (2023), we obtain a new state-of-the-art rate for privately finding second-order stationary points of the population loss:

Corollary 7.1 (Second-Order Stationary Points of Population Loss - Simple Version)**.** *Let $nd \geqslant 1/\varepsilon^2$. Assume $\nabla^2 f(\cdot,x)$ is 1-Lipschitz and that L,β, and D are constants, where $D=\|w^*\|$ for some $w^* \in \operatorname{argmin}_w F(w)$. Then, Algorithm 4 is (ε,δ)-DP and, with probability at least $1-\zeta$, returns a κ-second-order-stationary point, where*

$$\kappa \leqslant \widetilde{O}\left(\frac{1}{n^{1/3}}\left[\frac{d}{\varepsilon n}+\sqrt{\frac{d}{n}}\right]^{1/3}\right) + \widetilde{O}\left(\left(\frac{\sqrt{d}}{\varepsilon n}\right)^{3/7}\left[\frac{d}{\varepsilon n}+\sqrt{\frac{d}{n}}\right]^{3/7}\right).$$

See Appendix E for a precise statement of this corollary, and the proof. The proof combines a (novel, to our knowledge) high-probability excess population risk guarantee for the exponential mechanism (Lemma E.3) with Theorem E.2.

The previous state-of-the-art rate for this problem is $\widetilde{O}(1/n^{1/3}+(\sqrt{d}/\varepsilon n)^{3/7})$ (Liu et al., 2023). Thus, Corollary 7.1 strictly improves over this rate whenever $d/(\varepsilon n)+\sqrt{d/n}<1$. For example, if d and ε are constants, then $\kappa=\widetilde{O}(1/\sqrt{n})$, which is *optimal* and matches the *non-private* lower bound of Arora et al. (2023). (This lower bound holds even with the weaker *first-order* stationarity measure.) If $d > n\varepsilon$, then one should run the algorithm of Liu et al. (2023). Combining the two bounds results in a new state-of-the-art bound for stationary points of non-convex population loss functions.

8. Improved Rate for Stationary Points of Non-Convex GLMs

In this section, we restrict attention to *generalized linear models* (GLMs): loss functions of the form $f(w,(x,y)) = \phi_y(\langle w,x\rangle)$ for some $\phi_y:\mathbb{R}^d\to\mathbb{R}$ that is L-Lipschitz and β-smooth for all $y\in\mathbb{R}$. Assume that the data domain $\mathcal{X}$ has bounded ℓ_2-diameter $\|\mathcal{X}\|=O(1)$ and that the design matrix $X\in\mathbb{R}^{n\times d}$ has $r:=\operatorname{rank}(X)$.

Arora et al. (2022) provided a black-box method for obtaining dimension-independent DP stationary guarantees for non-convex GLMs. Their method applies a DP Johnson-Lindenstrauss (JL) transform to the output of a DP algorithm for finding approximate stationary points of non-convex empirical loss functions.

Lemma 8.1. *(Arora et al., 2023) Let $\mathcal{M}$ be an (ε,δ)-DP algorithm which guarantees $\mathbb{E}\|\nabla\widehat{F}_X(\mathcal{M}(X))\| \leqslant g(d,n,\beta,L,D,\varepsilon,\delta)$ and $\|\mathcal{M}(X)\|\leqslant poly(n,d,\beta,L,D)$ with probability at least $1-1/\sqrt{n}$, when run on an L-Lipschitz, β-smooth $\widehat{F}_X$ with $\|\operatorname{argmin}_w \widehat{F}_X(w)\|\leqslant D$. Let $k=\operatorname{argmin}_{j\in\mathbb{N}}\left[g(j,n,\beta,L,D,\varepsilon,\delta/2)+\frac{L}{\sqrt{j}}\right]\wedge r$. Then, the JL method, run on L-Lipschitz, β-smooth GLM loss G with $\|\operatorname{argmin}_w G(w)\|\leqslant D$ is (ε,δ)-DP. Further, given n i.i.d. samples, the method outputs w_{priv} s.t.*

$$\mathbb{E}\|\nabla F(w_{priv})\| = \widetilde{O}\left(\frac{L}{\sqrt{n}}+g(k,n,\beta,L,D,\varepsilon,\delta/2)\right).$$

Arora et al. (2022) used Lemma 8.1 with DP-SPIDER as $\mathcal{M}$ to obtain a stationarity guarantee for non-convex GLMs: $\widetilde{O}\left(1/\sqrt{n}+\min\{(\sqrt{r}/\varepsilon n)^{2/3},1/(n\varepsilon)^{2/5}\}\right)$ when $L,\beta=O(1)$. If we apply their JL method to the output of our Algorithm 2, then we obtain an improved rate:

Corollary 8.2 (Non-Convex GLMs)**.** *Let $f(w,(x,y))$ be a GLM loss function with $\beta,L,D=O(1)$. Then, the JL method applied to the output of $\mathcal{M}$ = Algorithm 2 (with $\mathcal{A}$ = Exponential Mechanism and $\mathcal{B}$ = DP-SPIDER) is (ε,δ)-DP and, given n i.i.d. samples, outputs w_{priv} s.t.*

$$\mathbb{E}\|\nabla F(w_{priv})\| \leqslant \widetilde{O}\left(\frac{1}{\sqrt{n}}\right)+\widetilde{O}\left(\frac{\sqrt{r}}{\varepsilon n}r^{1/6}\wedge\frac{1}{(\varepsilon n)^{3/7}}\right).$$

See Appendix F for the proof. Corollary 8.2 improves over the state-of-the-art (Arora et al., 2023) if $r < n\varepsilon$.

Remark 8.3. We can obtain essentially the same rate for *second-order* stationary points by substituting DP-SPIDER-SOSP for DP-SPIDER.

9. Preliminary Experiments

In this section, we conduct an empirical evaluation of our algorithm as a proof of concept. We run a small simulation

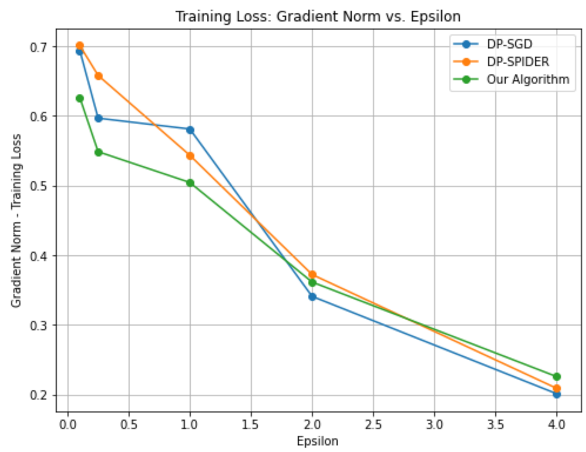

Figure 2. Training Loss: Gradient Norm vs. ε

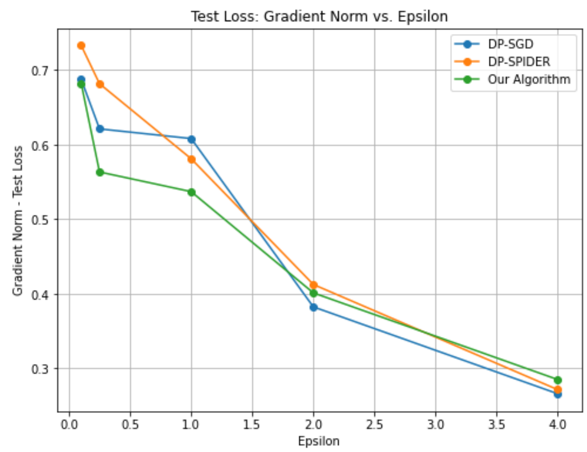

Figure 3. Test Loss: Gradient Norm vs. ε

with a non-convex loss function and synthetic data.[2]

Loss function and data: $f(w, x) = \frac{1}{2}\left[\|w\|^2 + \sin(\|w\|^2)\right] + x^T w$, where x is drawn uniformly from $\mathbb{B}$, the unit ball in $\mathbb{R}^d$ and $\mathcal{W} = 2\mathbb{B}$. Note that $f(\cdot, x)$ is non-convex, 6-smooth, and 5-Lipschitz on $\mathcal{W}$.

Our algorithm: $(\varepsilon_2, \delta/2)$-DP-SPIDER after warm-starting with $(\varepsilon_1, \delta/2)$-DP-SGD. (Recall that this algorithm is optimal for quasar-convex functions and $\varepsilon_1 = \varepsilon_2 = \varepsilon/2$.) We run T_1 iterations of DP-SGD and T_2 iterations of DP-SPIDER. $\varepsilon_1, \varepsilon_2, T_1$ and T_2 are all hyperparameters that we tune. We require $T_1 + T_2 = 50$ and $\varepsilon_1 + \varepsilon_2 = \varepsilon$.

Baselines: We compare against DP-SGD and DP-SPIDER, each run for 100 iterations. We carefully tune all hyperparameters (e.g. step size and phase length). We list the hyperparameters that we used to obtain each point in the plots in Appendix G.

Results: Our results are reported in Figures 2 and 3. *Our algorithm outperforms both baselines in the high privacy regime* $\varepsilon \leqslant 1$. For $\varepsilon \in \{2, 4\}$, the performance of all 3 algorithms is relatively similar and there is no apparent benefit from warm-starting.

Problem parameters: $n = d = 100$, $\delta = 1/n^{1.5}$. We vary $\varepsilon \in \{0.1, 0.25, 1, 2, 4\}$.

For each ε, we ran 10 trials with fresh, independently drawn data and reported average results. We projected the iterates onto $\mathcal{W}$ to ensure that the smoothness and Lipschitz bounds hold in each iteration.

10. Conclusion

We provided a novel framework for designing private algorithms to find (first- and second-order) stationary points of non-convex (empirical and population) loss functions. Our framework led to improved rates for general non-convex loss functions and GLMs, and optimal rates for important subclasses of non-convex functions (quasar-convex and KL).

Our work opens up several interesting avenues for future exploration. First, for general non-convex empirical and population losses, there remains a gap between our improved upper bounds and the lower bounds of Arora et al. (2023)—which hold even for *convex* functions. In light of our improved upper bounds (which are optimal when $d = O(1)$), we believe that the convex lower bounds are attainable for non-convex losses. Second, from a practical perspective, it would be useful to understand whether improvements over the previous state-of-the-art bounds are achievable with more computationally efficient algorithms. Finally, it would be fruitful for future empirical work to have more extensive, large-scale experiments to determine the most effective way to leverage our algorithmic framework in practice.

Acknowledgements

AL and SW's research is supported by NSF grant 2023239 and the AFOSR award FA9550-21-1-0084. SW also acknowledges support of the NSF grant 2224213. JU's research is supported by NSF awards CNS-2232629 and CNS-2247484. AL thanks Hilal Asi for helpful discussions in the beginning phase of this project.

[2]Code for the experiments is available at `https://github.com/lowya/How-to-Make-the-Gradients-Small-Privately/tree/main`.

Impact Statement

We develop algorithms for protecting the privacy of individuals who contribute training data. While this paper is primarily motivated by theoretical questions about the minimax optimal sample complexity of DP non-convex optimization, we acknowledge the potential broader impacts of our work.

We hope that our private optimization algorithms enable the development of machine learning models that can operate on sensitive datasets without compromising individual privacy. This impact extends to applications such as medical research, financial analysis, LLMs, and other domains where data privacy is paramount. We believe that the deployment of differentially private optimization techniques fosters a climate where organizations and decision-makers can harness the power of machine learning without sacrificing data privacy. This encourages a broader adoption of data-driven decision-making across industries, leading to more informed and accurate outcomes while respecting the confidentiality of sensitive information.

That being said, there are also potential negative consequences of privacy-preserving machine learning. For example, there is a potential risk that entities, such as corporations or government bodies, might misuse our algorithms for malicious activities, including the unauthorized gathering of personal information. Moreover, employing models trained with private data may lead to reduced accuracy when compared to their non-private counterparts, potentially resulting in unfavorable outcomes. Nevertheless, we maintain a strong conviction that sharing privacy-preserving machine learning algorithms, alongside an improved comprehension of these algorithms, ultimately provides a positive overall impact on society.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Oct 2016. doi: 10.1145/2976749.2978318. URL http://dx.doi.org/10.1145/2976749.2978318.

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

Amid, E., Warmuth, M. K., Anil, R., and Koren, T. Robust bi-tempered logistic loss based on bregman divergences. *Advances in Neural Information Processing Systems*, 32, 2019.

Amid, E., Ganesh, A., Mathews, R., Ramaswamy, S., Song, S., Steinke, T., Suriyakumar, V. M., Thakkar, O., and Thakurta, A. Public data-assisted mirror descent for private model training. In *International Conference on Machine Learning*, pp. 517–535. PMLR, 2022.

Arora, R., Bassily, R., Guzmán, C., Menart, M., and Ullah, E. Differentially private generalized linear models revisited. *Advances in Neural Information Processing Systems*, 35:22505–22517, 2022.

Arora, R., Bassily, R., González, T., Guzmán, C. A., Menart, M., and Ullah, E. Faster rates of convergence to stationary points in differentially private optimization. In *International Conference on Machine Learning*, pp. 1060–1092. PMLR, 2023.

Asi, H., Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: Optimal rates in l1 geometry. In Meila, M. and Zhang, T. (eds.), *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 393–403. PMLR, PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/asi21b.html.

Asi, H., Feldman, V., Koren, T., and Talwar, K. Near-optimal algorithms for private online optimization in the realizable regime. *arXiv preprint arXiv:2302.14154*, 2023.

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

Bassily, R., Belkin, M., and Ma, S. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

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

Bassily, R., Guzmán, C., and Menart, M. Differentially private stochastic optimization: New results in convex and non-convex settings. *Advances in Neural Information Processing Systems*, 34:9317–9329, 2021a.

Bassily, R., Guzmán, C., and Nandi, A. Non-euclidean differentially private stochastic convex optimization. In *Conference on Learning Theory*, pp. 474–499. PMLR, 2021b.

Boob, D. and Guzmán, C. Optimal algorithms for differentially private stochastic monotone variational inequalities and saddle-point problems. *Mathematical Programming*, pp. 1–43, 2023.

Bun, M., Ullman, J., and Vadhan, S. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 1–10, 2014.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T. B., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6, pp. 2633–2650, 2021.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Dwork, C. and Roth, A. *The Algorithmic Foundations of Differential Privacy*, volume 9. Now Publishers, Inc., 2014.

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

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

Foster, D. J., Sekhari, A., and Sridharan, K. Uniform convergence of gradients for non-convex learning and optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Gao, C. and Wright, S. J. Differentially private optimization for smooth nonconvex erm. *arXiv preprint arXiv:2302.04972*, 2023.

Ge, R., Lee, J. D., and Ma, T. Matrix completion has no spurious local minimum. *Advances in neural information processing systems*, 29, 2016.

Gower, R., Sebbouh, O., and Loizou, N. Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, pp. 1315–1323. PMLR, 2021.

Hardt, M., Ma, T., and Recht, B. Gradient descent learns linear dynamical systems. *The Journal of Machine Learning Research*, 19(1):1025–1068, 2018.

Hinder, O., Sidford, A., and Sohoni, N. Near-optimal methods for minimizing star-convex functions and beyond. In *Conference on learning theory*, pp. 1894–1938. PMLR, 2020.

Jain, P. and Thakurta, A. G. (near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning*, pp. 476–484. PMLR, 2014.

Kang, Y., Liu, Y., Niu, B., and Wang, W. Weighted distributed differential privacy erm: Convex and non-convex. *Computers & Security*, 106:102275, 2021.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Kleinberg, B., Li, Y., and Yuan, Y. An alternative view: When does sgd escape local minima? In *International Conference on Machine Learning*, pp. 2698–2707, 2018.

Kurdyka, K. On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, volume 48, pp. 769–783, 1998.

Li, X., Tramer, F., Liang, P., and Hashimoto, T. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*, 2021.

Liu, C., Zhu, L., and Belkin, M. Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning. *arXiv preprint arXiv:2003.00307*, 7, 2020.

Liu, C., Zhu, L., and Belkin, M. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Liu, D., Ganesh, A., Oh, S., and Thakurta, A. G. Private (stochastic) non-convex optimization revisited: Second-order stationary points and excess risks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Lowy, A. and Razaviyayn, M. Private federated learning without a trusted server: Optimal algorithms for convex losses. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=TVY6GoURrw.

Lowy, A. and Razaviyayn, M. Private stochastic optimization with large worst-case lipschitz parameter: Optimal rates for (non-smooth) convex losses and extension to non-convex losses. In *International Conference on Algorithmic Learning Theory*, pp. 986–1054. PMLR, 2023b.

Lowy, A., Ghafelebashi, A., and Razaviyayn, M. Private non-convex federated learning without a trusted server. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 5749–5786. PMLR, 2023a.

Lowy, A., Gupta, D., and Razaviyayn, M. Stochastic differentially private and fair learning. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=3nM5uhPlfv6.

Lowy, A., Li, Z., Huang, T., and Razaviyayn, M. Optimal differentially private learning with public data. *arXiv preprint:2306.15056*, 2023c.

McSherry, F. and Talwar, K. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pp. 94–103. IEEE, 2007.

Mei, S., Bai, Y., and Montanari, A. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

Menart, M., Ullah, E., Arora, R., Bassily, R., and Guzmán, C. Differentially private non-convex optimization under the kl condition with optimal rates. *arXiv preprint arXiv:2311.13447*, 2023.

Murty, K. G. and Kabadi, S. N. Some np-complete problems in quadratic and nonlinear programming. 1985.

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

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Nesterov, Y. and Polyak, B. T. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Polyak, B. T. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Scaman, K., Malherbe, C., and Dos Santos, L. Convergence rates of non-convex stochastic gradient descent under a generic lojasiewicz condition and local smoothness. In *International Conference on Machine Learning*, pp. 19310–19327. PMLR, 2022.

Shen, H., Wang, C.-L., Xiang, Z., Ying, Y., and Wang, D. Differentially private non-convex learning for multi-layer neural networks. *arXiv preprint arXiv:2310.08425*, 2023.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

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

Steinke, T. and Ullman, J. Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality*, 7(2), 2016.

Sun, J., Qu, Q., and Wright, J. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18:1131–1198, 2018.

Tran, H. and Cutkosky, A. Momentum aggregation for private non-convex erm. *Advances in Neural Information Processing Systems*, 35:10996–11008, 2022.

Wang, D. and Xu, J. Escaping saddle points of empirical risk privately and scalably via dp-trust region method. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III*, pp. 90–106. Springer, 2021.

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

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

Yu, D., Naik, S., Backurs, A., Gopi, S., Inan, H. A., Kamath, G., Kulkarni, J., Lee, Y. T., Manoel, A., Wutschitz, L., et al. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, 2021.

Zhang, J., Zheng, K., Mou, W., and Wang, L. Efficient private erm for smooth objectives, 2017.

Zhang, L., Thekumparampil, K. K., Oh, S., and He, N. Bring your own algorithm for optimal differentially private stochastic minimax optimization. *Advances in Neural Information Processing Systems*, 35:35174–35187, 2022.

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

Zhou, Y., Yang, J., Zhang, H., Liang, Y., and Tarokh, V. Sgd converges to global minimum in deep learning via star-convex path. *arXiv preprint arXiv:1901.00451*, 2019.

Zhou, Y., Chen, X., Hong, M., Wu, Z. S., and Banerjee, A. Private stochastic non-convex optimization: Adaptive algorithms and tighter generalization bounds. *arXiv preprint arXiv:2006.13501*, 2020.

A. Further Discussion of Related Work

Private ERM and stochastic optimization with convex loss functions has been studied extensively (Chaudhuri et al., 2011; Bassily et al., 2014; 2019; Feldman et al., 2020). Beyond these classical settings, differentially private optimization has also recently been studied e.g., in the context of online learning (Jain & Thakurta, 2014; Asi et al., 2023), federated learning (Lowy & Razaviyayn, 2023a), different geometries (Bassily et al., 2021b; Asi et al., 2021), min-max games (Boob & Guzmán, 2023; Zhang et al., 2022), fair and private learning (Lowy et al., 2023b), and public-data assisted private optimization (Amid et al., 2022; Lowy et al., 2023c). Below we summarize the literature on DP *non-convex* optimization.

Stationary Points of Empirical Loss Functions. For non-convex $\widehat{F}_X$, the best known stationarity rate prior to 2022 was $\mathbb{E}\|\nabla \widehat{F}_X(\mathcal{A}(X))\| = O((\sqrt{d\ln(1/\delta)}/\varepsilon n)^{1/2})$ (Zhang et al., 2017; Wang et al., 2017; 2019). In the last two years, a pair of papers made progress and obtained improved rates of $\widetilde{O}((\sqrt{d\ln(1/\delta)}/\varepsilon n)^{2/3})$ (Arora et al., 2023; Tran & Cutkosky, 2022). The work of Lowy et al. (2023a) extended this result to non-convex federated learning/distributed ERM and non-smooth loss functions. The work of Liu et al. (2023) extended this result to *second-order* stationary points. Despite this problem receiving much attention from researchers, it remained unclear whether the $\widetilde{O}((\sqrt{d\ln(1/\delta)}/\varepsilon n)^{2/3})$ barrier could be broken. Our algorithm finally breaks this barrier.

Stationary Points of Population Loss Functions. The literature on stationary points of population loss functions is much sparser than for empirical loss functions. The work of (Zhou et al., 2020) gave a DP algorithm for finding α-FOSP, where $\alpha \lesssim \varepsilon\sqrt{d} + (\sqrt{d}/\varepsilon n)^{1/2}$. Thus, their bound is meaningful only when $\varepsilon \ll 1/\sqrt{d}$. Arora et al. (2022) improved over this rate, obtaining $\alpha = \widetilde{O}(1/n^{1/3} + (\sqrt{d}/\varepsilon n)^{1/2})$. The prior state-of-the-art rate for finding SOSPs of F was $\widetilde{O}(1/n^{1/3} + (\sqrt{d}/\varepsilon n)^{3/7})$ (Liu et al., 2023). We improve over this rate in the present work.

Excess Risk of PL and KL Loss Functions. Private optimization of PL loss functions has been considered in (Wang et al., 2017; Kang et al., 2021; Zhang et al., 2021; Lowy et al., 2023a). Prior to the work of (Lowy et al., 2023a), all works on DP PL optimization made the extremely strong assumptions that $f(\cdot, x)$ is Lipschitz and PL on all of $\mathbb{R}^d$. We are not aware of any loss functions that satisfy both these assumptions. This gap was addressed by (Lowy et al., 2023a), who proved near-optimal excess risk bounds for *proximal-PL* (Karimi et al., 2016) loss functions. The proximal-PL condition extends the PL condition to the constrained setting, and allows for functions that are Lipschitz on some compact subset of $\mathbb{R}^d$. The work of Menart et al. (2023) gave near-optimal excess risk bounds under the KL* condition, which generalizes the PL condition. Our work is the first to give optimal bounds for finding approximate stationary points of KL* functions. Note that stationarity is a stronger measure of suboptimality than excess risk for KL* functions, since by definition, the excess risk of these functions is upper bounded by a function of the gradient norm.

Non-Convex GLMs. While DP excess risk guarantees for convex GLMs are well understood (Jain & Thakurta, 2014; Song et al., 2021; Arora et al., 2022), far less is known for stationary points of non-convex GLMs. In fact, we are aware of only one prior work that provides DP stationarity guarantees for non-convex GLMs: Arora et al. (2023) obtains dimension-independent/rank-dependent α-FOSP, where $\alpha \lesssim 1/\sqrt{n} + (\sqrt{r}/\varepsilon n)^{2/3} \wedge (1/\varepsilon n)^{2/5}$ and r is the rank of the design matrix X. We improve over this rate in the present work.

Non-privately, non-convex GLMs have been studied by Mei et al. (2018); Foster et al. (2018).

B. More privacy preliminaries

The following result can be found, e.g. in (Dwork & Roth, 2014, Theorem 3.20).

Lemma B.1 (Advanced Composition Theorem). *Let $\epsilon \geqslant 0, \delta, \delta' \in [0, 1)$. Assume $\mathcal{A}_1, \cdots, \mathcal{A}_T$, with $\mathcal{A}_t : \mathcal{X}^n \times \mathcal{W} \to \mathcal{W}$, are each (ϵ, δ)-DP $\forall t = 1, \cdots, T$. Then, the adaptive composition $\mathcal{A}(X) := \mathcal{A}_T(X, \mathcal{A}_{T-1}(X, \mathcal{A}_{T-2}(X, \cdots)))$ is $(\epsilon', T\delta + \delta')$-DP for $\epsilon' = \sqrt{2T\ln(1/\delta')}\epsilon + T\epsilon(e^{\epsilon} - 1)$.*

C. Second-Order Stationary Points for ERM: Meta-Algorithm

Theorem C.1 (Re-statement of Theorem 3.4). *Suppose $\mathcal{A}$ is $(\varepsilon/2, \delta/2)$-DP and $\widehat{F}_X(\mathcal{A}(X)) - \widehat{F}_X^* \leqslant \psi$ with probability $\geqslant 1 - \zeta$ (for polynomial $1/\zeta$). Then, Algorithm 2 with $\mathcal{B}$ as DP-SPIDER-SOSP (with appropriate parameters) is (ε, δ)-DP,* *and with probability $\geqslant 1 - 2\zeta$ has output w_{priv} satisfying*

$$\|\nabla \widehat{F}_X(w_{priv})\| \leqslant \tilde{\alpha} := \widetilde{O}\left(\frac{L\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right) + \widetilde{O}\left(L^{1/3}\beta^{1/3}\psi^{1/3}\left(\frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)^{2/3}\right) + \widetilde{O}\left(\frac{\beta^{7/6}L^{1/6}\psi^{1/6}}{n\sqrt{\rho}}\left(\frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)^{1/3}\right),$$

and

$$\nabla^2 \widehat{F}_X(w_{priv}) \succeq -\sqrt{\rho\tilde{\alpha}}\,\mathbf{I}_d$$

.

Proof. Let E be the good event that $\widehat{F}_X(\mathcal{A}(X)) - \widehat{F}_X^* \leqslant \psi$ *and* $\mathcal{B}$ satisfies the stationarity guarantees in Lemma 3.3 given input $w_0 = \mathcal{A}(X)$. Then $\mathbb{P}(E) \geqslant 1 - 2\zeta$ by a union bound. Moreover, conditional on E, the stationarity guarantees in Theorem 3.4 hold by applying Lemma 3.3 with parameter $\hat{\Delta}_{w_0}$ replaced by ψ. □

D. Optimal Rate for Quasar-Convex Losses

Proposition D.1 (Precise Statement of Proposition 5.2)**.** *Let $\widehat{F}_X$ be q-quasar convex and $\|w_1 - w^*\| \leqslant D$ for some $w_1 \in \mathbb{R}^d, w^* \in \operatorname{argmin}_w \widehat{F}_X(w)$. Then, Algorithm 3 with*

$$\eta = \frac{D}{\sqrt{T(L^2 + d\sigma^2)}}, \quad T = \frac{\varepsilon^2 n^2}{d\ln(1/\delta)}, \quad b \gtrsim \sqrt{d\varepsilon}, \quad \sigma^2 = \frac{1000L^2 T\ln(1/\delta)}{\varepsilon^2 n^2}$$

is (ε, δ)-DP, and returns $\hat{w}$ such that

$$\mathbb{E}\widehat{F}_X(\hat{w}) - \widehat{F}_X^* \lesssim LD\frac{\sqrt{d\ln(1/\delta)}}{\varepsilon nq}.$$

Moreover, for any $\zeta > 0$, there is an (ε, δ)-DP variation of Algorithm 3 that returns $\tilde{w}$ such that

$$\widehat{F}_X(\hat{w}) - \widehat{F}_X^* = \widetilde{O}\left(LD\frac{\sqrt{d\ln(1/\delta)}}{\varepsilon nq}\right)$$

with probability at least $1 - \zeta$.

Proof. **Privacy:** Privacy of DP-SGD does not require convexity and is an immediate consequence of, e.g. (Abadi et al., 2016, Theorem 1) and our choices of T, b, σ^2.

Expected excess risk: Recall that the updates are given by $w_{t+1} = w_t - \eta\nabla_t$, where $\nabla_t := g_t + u_t := \frac{1}{b}\sum_{x\in S_t}\nabla f(w_t, x) + u_t$ for $u_t \sim \mathcal{N}(0, \sigma^2\mathbf{I}_d)$ and S_t is drawn uniformly with replacement from X with $b = |S_t|$. Thus,

$$\|w_{t+1} - w^*\|^2 = \|w_t - w^*\|^2 - 2\eta\langle\nabla_t, w_t - w^*\rangle + \eta^2\|\nabla_t\|^2.$$

Taking conditional expectation given w_t and using the fact that u_t is mean-zero and independent of w_t gives:

$$\begin{aligned}\mathbb{E}\left[\|w_{t+1} - w^*\|^2 | w_t\right] &= \|w_t - w^*\|^2 - 2\eta\langle\nabla\widehat{F}_X(w_t), w_t - w^*\rangle + \eta^2\left(\|g_t\|^2 + d\sigma^2\right) \\ &\leqslant \|w_t - w^*\|^2 - 2\eta\langle\widehat{F}_X(w_t), w_t - w^*\rangle + \eta^2\left(L^2 + d\sigma^2\right) \\ &\leqslant \|w_t - w^*\|^2 - 2\eta q\left(\widehat{F}_X(w_t) - \widehat{F}_X^*\right) + \eta^2\left(L^2 + d\sigma^2\right),\end{aligned}$$

where the last inequality above used q-quasar-convexity. Now, re-arranging and taking total expectation yields:

$$2\eta q\mathbb{E}[\widehat{F}_X(w_t) - \widehat{F}_X^*] \leqslant \mathbb{E}\left[\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2\right] + \eta^2\left(L^2 + d\sigma^2\right).$$

Telescoping the above inequality from $t = 1$ to T and recalling $\hat{w}_T \sim \textbf{Unif}(\{w_1, \ldots, w_T\})$ yields

$$\mathbb{E}[\widehat{F}_X(\hat{w}_T) - \widehat{F}_X^*] \leqslant \frac{D^2}{2\eta qT} + \frac{\eta(L^2 + d\sigma^2)}{2q}.$$

Plugging in $\eta = \frac{D}{\sqrt{T(L^2+d\sigma^2)}}$ then gives

$$\mathbb{E}[\widehat{F}_X(\hat{w}_T) - \widehat{F}_X^*] \leqslant \frac{2D}{q\sqrt{T}}\left(L + \sqrt{d\sigma^2}\right) \lesssim LD\left(\frac{1}{q\sqrt{T}} + \frac{\sqrt{d\ln(1/\delta)}}{\varepsilon nq}\right).$$

Finally, choosing $T \geqslant \frac{\varepsilon^2 n^2}{d\ln(1/\delta)}$ yields the desired expected excess risk bound.

High-probability excess risk: This is an instantiation of the meta-algorithm described in (Bassily et al., 2014, Appendix D). We run the DP-SGD algorithm above $k = \log(2/\zeta)$ times with privacy parameters $(\varepsilon/2k, \delta/2k)$ for each run. This gives us an $(\varepsilon/2, \delta/2)$-DP list of k vectors, which we denote $\{\hat{w}^1, \ldots, \hat{w}^k\}$. By Markov's inequality, with probability at least $1 - 1/2^k$, there exists $i \in [k]$ such that $\widehat{F}_X(\hat{w}^i) - \widehat{F}_X^* \lesssim \frac{LDk\sqrt{d\ln(k/\delta)}}{\varepsilon n}$. Now we apply the $\varepsilon/2$-DP exponential mechanism (McSherry & Talwar, 2007) to the list $\{\hat{w}^1, \ldots, \hat{w}^k\}$ in order to select the (approximately) best $\hat{w}^i$ with probability at least $1 - \zeta/2$. By a union bound, the output of this mechanism has excess risk bounded by $\widetilde{O}\left(LD\frac{\sqrt{d\ln(1/\delta)}}{q\varepsilon n}\right)$ with probability at least $1 - \zeta$. $\square$

E. Improved Rates for Stationary Points of Non-Convex Population Loss

Denote the initial suboptimality gap of the population loss by

$$\Delta_{w_0} := F(w_0) - F^*.$$

We will need the population stationary guarantees of a variation of DP-SPIDER-SOSP:

Lemma E.1. *(Liu et al., 2023, Theorem 4.6) Let $\zeta \in (0,1)$ and let $\nabla^2 f(\cdot, x)$ be ρ-Lipschitz for all x. Denote*

$$s := \widetilde{O}\left(\left(\frac{L\beta\Delta_{w_0}}{n}\right)^{1/3} + (L\beta^3\Delta_{w_0}^3)^{1/7}\left(\frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)^{3/7}\right),$$

and

$$S := \widetilde{O}\left(s + \frac{\beta}{\sqrt{\rho}}\left(\frac{1}{n\varepsilon} + \frac{1}{\sqrt{n}}\right)\sqrt{s} + L\left(\frac{1}{n\varepsilon} + \frac{1}{\sqrt{n}}\right)\right).$$

Then, there is a $(\varepsilon/2, \delta/2)$-DP variation of DP-SPIDER-SOSP which, given n i.i.d. samples from $\mathcal{P}$, returns a point $\hat{w}$ such that $\hat{w}$ is an S-second-order-stationary point of F with probability at least $1 - \zeta$.

Theorem E.2 (Second-Order Stationary Points for Population Loss: Meta-Algorithm)**.** *Let $\zeta \in (0,1)$ and let $\nabla^2 f(\cdot, x)$ be ρ-Lipschitz for all x. Suppose $\mathcal{A}$ is $(\varepsilon/2, \delta/2)$-DP and $F(\mathcal{A}(X)) - F^* \leqslant \psi$ with probability $\geqslant 1 - \zeta$. Then, Algorithm 4 with $\mathcal{B}$ as DP-SPIDER-SOSP (with appropriate parameters) is (ε, δ)-DP and, given n i.i.d. samples from $\mathcal{P}$, has output w_{priv} which is a υ-second-order-stationary point of F with probability at least $1 - 2\zeta$, where*

$$\begin{aligned}\upsilon := &\widetilde{O}\left(\left(\frac{L\beta\psi}{n}\right)^{1/3} + (L\psi^3\beta^3)^{1/7}\left(\frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)^{3/7}\right)\\ &+ \widetilde{O}\left(\frac{\beta}{\sqrt{\rho}}\left(\frac{1}{n\varepsilon} + \frac{1}{\sqrt{n}}\right)\left(\frac{(L\beta\psi)^{1/6}}{n^{1/6}} + (L\psi^3\beta^3)^{1/14}\left(\frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)^{3/14}\right)\right)\\ &+ \widetilde{O}\left(L\left(\frac{1}{n\varepsilon} + \frac{1}{\sqrt{n}}\right)\right).\end{aligned}$$

Proof. Privacy is immediate from basic composition.

By assumption, $\mathcal{A}$ returns w_0 such that $\Delta_{w_0} \leqslant \psi$ with probability at least $1-\zeta$. Conditional on this good event happening, then Lemma E.1 implies the desired stationarity guarantee with probability at least $1-\zeta$, by plugging in ψ for Δ_{w_0} in Lemma E.1. By a union bound, we obtain Theorem 3.4. □

In order to obtain Corollary 7.1, we will also need a high-probability excess population risk guarantee for the exponential mechanism:

Lemma E.3 (Excess Population Risk of Exponential Mechanism)**.** *Let $\zeta \in (0,1)$ and let $\mathcal{W}$ be a compact set containing $\tilde{w}$ such that $\|w-\tilde{w}\| \leqslant D$ for all $w \in \mathcal{W}$ and $F(\tilde{w}) - F^* \leqslant LDd/\varepsilon n$. Then, given n i.i.d. samples from $\mathcal{P}$, the ε-DP exponential mechanism of Definition 4.1 outputs w_0 such that, with probability at least $1-\zeta$,*

$$F(w_0) - F^* = \widetilde{O}\left(LD\left(\frac{d}{\varepsilon n} + \sqrt{\frac{d}{n}}\right)\right).$$

Proof. Let $\widetilde{\mathcal{W}} = \{w_1, \ldots, w_N\}$ be a $D\frac{d}{\varepsilon n}$-net for $\mathcal{W}$ with cardinality $N = |\widetilde{\mathcal{W}}| \leqslant \left(\frac{2D\varepsilon n}{d}\right)^d$. Denote the output of the exponential mechanism $w_0 = \mathcal{A}_E(X)$. By Lemma 4.2, we have

$$\widehat{F}_X(w_0) - \widehat{F}_X^* \leqslant \widetilde{O}\left(LD\frac{d}{\varepsilon n}\right) \tag{4}$$

with probability at least $1-\zeta/2$. Now, for any $j \in [N]$, we have

$$\mathbb{P}(|\widehat{F}_X(w_j) - F(w_j)| \leqslant p) \geqslant 1 - 2\exp\left(\frac{-np^2}{2L^2D^2}\right)$$

for any $p \in (0,1)$ by Hoeffding's inequality, since $f(w_j, x) \in [-LD, LD]$ for all x. By a union bound, we have

$$\mathbb{P}\left(\max_{j\in[N]} |\widehat{F}_X(w_j) - F(w_j)| \leqslant p\right) \geqslant 1 - 2N\exp\left(\frac{-np^2}{2L^2D^2}\right). \tag{5}$$

Thus, the following inequalities hold with probability at least $1 - 4N\exp\left(\frac{-np^2}{2L^2D^2}\right) - \zeta/2$:

$$\begin{aligned}
F(w_0) - F^* &\leqslant \widehat{F}_X(w_0) - F^* + p \\
&\leqslant \widehat{F}_X(w_0) - \widehat{F}_X\left(\operatorname*{argmin}_w F(w)\right) + 2p \\
&\leqslant \widehat{F}_X(w_0) - \widehat{F}_X^* + 2p \\
&\leqslant \widetilde{O}\left(LD\frac{d}{\varepsilon n}\right) + 2p.
\end{aligned}$$

Choosing $p = \frac{LD}{\sqrt{n}}\sqrt{\log(8/\zeta) + d}$ ensures that

$$F(w_0) - F^* = \widetilde{O}\left(LD\left(\frac{d}{\varepsilon n} + \sqrt{\frac{d}{n}}\right)\right).$$

with probability at least $1-\zeta$, as desired. □

Note that (Liu et al., 2023, Theorem 5.8) proved a weaker "in-expectation" version of Lemma E.3.

Corollary E.4 (Precise Statement of Corollary 7.1)**.** *Assume $\nabla^2 f(\cdot, x)$ is ρ-Lipschitz and $\mathcal{W}$ is a compact set containing $\tilde{w}$ such that $\|w-\tilde{w}\| \leqslant D$ for all $w \in \mathcal{W}$ and $F(\tilde{w}) - F^* \leqslant LDd/\varepsilon n$. Then, given n i.i.d. samples from $\mathcal{P}$, Algorithm 4* *with $\mathcal{A}$ = Exponential Mechanism and $\mathcal{B}$ = DP-SPIDER-SOSP is (ε, δ)-DP. Moreover, with probability at least $1 - 2\zeta$, the output w_{priv} of Algorithm 4 is a κ-second-order-stationary point of F, where*

$$\kappa \leqslant \widetilde{O}\left(\frac{(L\beta)^{1/3}}{n^{1/3}}\left[(LD)^{1/3}\left(\frac{d}{\varepsilon n}\right)^{1/3}\right]\right) + \widetilde{O}\left(\left[L^4\beta^3 D^3\left(\frac{d}{\varepsilon n} + \sqrt{\frac{d}{n}}\right)^3\right]^{1/7}\left(\frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)^{3/7}\right)$$
$$+ \widetilde{O}\left(\frac{\beta}{\sqrt{\rho}}\left(\frac{1}{n\varepsilon} + \frac{1}{\sqrt{n}}\right)\right)\left[\left(\frac{L\beta}{n}\right)^{1/6}\left(LD\left(\frac{d}{\varepsilon n} + \sqrt{\frac{d}{n}}\right)\right)\right.$$
$$\left. + \left(\frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)^{3/14}(L\beta^3)^{1/4}(LD)^{3/14}\left(\frac{d}{\varepsilon n} + \sqrt{\frac{d}{n}}\right)^{3/14}\right]$$
$$+ L\widetilde{O}\left(\frac{1}{\varepsilon n} + \frac{1}{\sqrt{n}}\right).$$

Proof. Privacy follows from basic composition.

The stationarity result is a consequence of Theorem E.2 and Lemma E.3. Namely, we use Lemma E.3 to plug $\psi = \widetilde{O}\left(LD\left(\frac{d}{\varepsilon n} + \sqrt{\frac{d}{n}}\right)\right)$ into the expression for υ in Theorem E.2. $\square$

Note that Corollary E.4 immediately implies Corollary 7.1.

F. Improved Rates for Stationary Points of Non-Convex GLMs

Corollary F.1 (Re-statement of Corollary 8.2)**.** *Let $f(w, (x, y))$ be a GLM loss function with $\beta, L, D = O(1)$. Then, the JL method applied to the output of $\mathcal{M}$ = Algorithm 2 (with $\mathcal{A}$ = Exponential Mechanism and $\mathcal{B}$ = DP-SPIDER) is (ε, δ)-DP and, given n i.i.d. samples from $\mathcal{P}$, outputs w_{priv} such that*

$$\mathbb{E}\|\nabla F(w_{priv})\| \leqslant \widetilde{O}\left(\frac{1}{\sqrt{n}}\right) + \widetilde{O}\left(\frac{\sqrt{r}}{\varepsilon n} r^{1/6} \wedge \frac{1}{(\varepsilon n)^{3/7}}\right).$$

Proof. The result is a direct consequence of Lemma 8.1 combined with Corollary 4.3. The fact that $\|\mathcal{M}(X)\| \leqslant poly(n, d, \beta, L, D)$ with high probability for $\mathcal{M}$ = Algorithm 2 (with $\mathcal{A}$ = Exponential Mechanism and $\mathcal{B}$ = DP-SPIDER) follows from the proof of (Arora et al., 2023, Corollary 6.2), which showed that $\|\mathcal{B}(X)\| \leqslant poly(n, d, \beta, L, D)$ for any initialization w_0. $\square$

G. Hyperparameters for Experiments

We tuned hyperparameters using the code at `https://github.com/lowya/How-to-Make-the-Gradients-Small-Privately/tree/main`.

The "optimal" hyperparameters that we obtained for each algorithm and each value of ε are listed below (using 10 independent epednent runs of the hyperparameter tuning code with fresh validation data in each run):

$\varepsilon = 0.1$

- $T_1 = 50$
- SPIDER $q = 10$
- Warm-start $q = 100$
- SGD $\eta = 0.0005$
- SPIDER $\eta = 0.005$

- Warm-start $\eta_{sgd} = 0.0005$
- Warm-start $\eta_{spider} = 0.005$
- Warm-start $\varepsilon_1 = \varepsilon/2$

$\varepsilon = 0.25$

- $T_1 = 50$
- SPIDER $q = 5$
- Warm-start $q = 5$
- SGD $\eta = 0.0005$
- SPIDER $\eta = 0.001$
- Warm-start $\eta_{sgd} = 0.05$
- Warm-start $\eta_{spider} = 0.0005$
- Warm-start $\varepsilon_1 = \varepsilon/4$

$\varepsilon = 1$

- $T_1 = 1$
- SPIDER $q = 10$
- Warm-start $q = 10$
- SGD $\eta = 0.0025$
- SPIDER $\eta = 0.0025$
- Warm-start $\eta_{sgd} = 0.001$
- Warm-start $\eta_{spider} = 0.0005$
- Warm-start $\varepsilon_1 = \varepsilon/4$

$\varepsilon = 2$

- $T_1 = 50$
- SPIDER $q = 5$
- Warm-start $q = 5$
- SGD $\eta = 0.0025$
- SPIDER $\eta = 0.0025$
- Warm-start $\eta_{sgd} = 0.0025$
- Warm-start $\eta_{spider} = 0.0025$
- Warm-start $\varepsilon_1 = \varepsilon/4$

$\varepsilon = 4$

- $T_1 = 25$
- SPIDER $q = 5$
- Warm-start $q = 5$
- SGD $\eta = 0.005$
- SPIDER $\eta = 0.005$
- Warm-start $\eta_{sgd} = 0.005$
- Warm-start $\eta_{spider} = 0.005$
- Warm-start $\varepsilon_1 = \varepsilon/100$