Evading the Curse of Dimensionality in Unconstrained Private GLMs

Shuang Song
Google

Thomas Steinke
Google

Om Thakkar
Google

Abhradeep Thakurta
Google

Abstract

We revisit the well-studied problem of differentially private empirical risk minimization (ERM). We show that for unconstrained convex generalized linear models (GLMs), one can obtain an excess empirical risk of $\widetilde{O}\left(\sqrt{\texttt{rank}}/\varepsilon n\right)$, where $\texttt{rank}$ is the rank of the feature matrix in the GLM problem, n is the number of data samples, and ε is the privacy parameter. This bound is attained via differentially private gradient descent (DP-GD). Furthermore, via the *first lower bound for unconstrained private ERM*, we show that our upper bound is tight. In sharp contrast to the constrained ERM setting, there is no dependence on the dimensionality of the ambient model space (p). (Notice that $\texttt{rank} \leq \min\{n, p\}$.) Besides, we obtain an analogous excess population risk bound which depends on $\texttt{rank}$ instead of p.

For the smooth non-convex GLM setting (i.e., where the objective function is non-convex but preserves the GLM structure), we further show that DP-GD attains a dimension-independent convergence of $\widetilde{O}\left(\sqrt{\texttt{rank}}/\varepsilon n\right)$ to a first-order-stationary-point of the underlying objective.

Finally, we show that for convex GLMs, a variant of DP-GD commonly used in practice (which involves clipping the individual gradients) also exhibits the same dimension-independent convergence to the minimum of a well-defined objective. To that end, we provide a structural lemma that characterizes the effect of clipping on the optimization profile of DP-GD.

1 Introduction

Differentially private empirical risk minimization (ERM) is a well-studied area in the privacy literature [Chaudhuri et al., 2011, Kifer et al., 2012, Song et al., 2013, Bassily et al., 2014, Jain and Thakurta, 2014, Abadi et al., 2016, McMahan et al., 2017, Wu et al., 2017, Bassily et al., 2019b, Iyengar et al., 2019, Pichapati et al., 2019, Thakkar et al., 2019, Feldman et al., 2020]. In the constrained setting, where the model space is bounded by $\mathcal{C} \subsetneq \mathbf{R}^p$, tight upper and lower bounds are known for both excess empirical risk [Bassily et al., 2014], and excess population risk [Bassily et al., 2019a, Feldman et al., 2020]. Surprisingly, in the arguably simpler unconstrained setting where $\mathcal{C} = \mathbf{R}^p$, the problem space is much less explored. To our knowledge, the only prior work that distinguishes between the constrained and the unconstrained case is that of Jain and Thakurta [2014]. They show dimension-independent upper bounds for population risk (under differential privacy) in the convex generalized linear models (GLMs) case, which alludes to a separation between constrained and unconstrained settings. In contrast, the lower bound of Bassily et al. [2014] shows that an explicit dependence on the dimensionality (p) is necessary in the constrained setting, even for GLMs.

In this work, we revisit the private unconstrained ERM setting, and close the gap between upper and lower bounds for the GLM case. We first show that for convex GLMs, one can attain an excess empirical risk of $\widetilde{O}\left(\sqrt{\texttt{rank}}/(\varepsilon \cdot n)\right)$[1], where $\texttt{rank}$ is the rank of the feature matrix, and ε is the privacy parameter. In comparison, Jain and Thakurta [2014] provide a much worse upper bound on the excess population risk for the unconstrained setting as $\widetilde{O}\left(1/\varepsilon\sqrt{n}\right)$. If interpreted in terms of excess empirical risk, their bound matches ours only when $\texttt{rank} = n$. Our upper bound is in sharp contrast to the lower bound of $\widetilde{\Omega}\left(\sqrt{p}/(\varepsilon \cdot n)\right)$ for the constrained setting [Bassily et al., 2014] (p being the dimensionality of the model space), as $\texttt{rank} \leq \min\{n, p\}$ always holds, and $\texttt{rank}$ may be much smaller. Our upper bound is achieved via differentially private gradient descent (DP-GD) [Bassily et al., 2014, Song et al., 2013, Talwar et al., 2014, Abadi et al., 2016]. While our guarantees extend to the stochastic variant of DP-GD, we focus on the full gradient version for brevity.

We further show that our bound on the excess empirical

[1] $\widetilde{O}(\cdot)$ hides $\mathrm{polylog}\left(1/\delta\right)$, where δ is a privacy parameter.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

risk is essentially tight. It is worth mentioning that ours is *the first lower bound* on excess empirical risk for any unconstrained ERM problem (including GLMs). The lower bound is based on fingerprinting codes [Bun et al., 2018, Steinke and Ullman, 2017].

In subsequent works [Kairouz et al., 2020, Zhou et al., 2020], there have been extensions of our results for convex GLMs to general convex ERMs via adaptive preconditioners. However, these results are more restrictive in their guarantees, and *do not imply our results*. They require the existence of public data to identify the subspace where the gradient of the objective function lies.

Going beyond convex GLMs, we show that the dimension-independent convergence holds even for *non-convex* GLMs (i.e., the loss function has the GLM structure, but can be non-convex). Such problems appear commonly in robust regression [Amid et al., 2019, Masnadi-Shirazi and Vasconcelos, 2009, Masnadi-Shirazi et al., 2010]. Given an objective function $\mathcal{L}(\theta; D) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; d_i)$, where dataset $D = \{d_1, \ldots, d_n\}$, we show that DP-GD reaches a first-order stationary point (FOSP) of $\mathcal{L}(\theta; D)$ at a rate of $\widetilde{O}\left(\sqrt{\texttt{rank}}/(\varepsilon \cdot n)\right)$ as long as the individual loss functions ℓ are smooth in the model parameter.

All our upper bounds are primarily based on DP-GD, which in general, requires a bound on the ℓ_2-norm of the subgradients of individual loss functions in $\mathcal{L}(\theta; D)$. In practice, however, such a bound is seldom known a priori for complex models. As a result, a variant of DP-GD, called *clipped* DP-GD [Abadi et al., 2016, Papernot et al., 2020], is commonly used. It scales the subgradients down if the ℓ_2-norm crosses a predefined threshold, a.k.a. the clipping norm ($B > 0$). In this work, we show that no matter what the clipping norm is, for convex GLMs, clipped DP-GD still has an excess empirical risk of $\widetilde{O}(\sqrt{\texttt{rank}}/(\varepsilon \cdot n))$ with respect to a well-defined objective function $\mathcal{L}^{(B)}_{\text{clipped}}(\theta; D)$ (in contrast to the original objective $\mathcal{L}(\theta; D)$). The function $\mathcal{L}^{(B)}_{\text{clipped}}$ still satisfies the convex GLM property. While there are other contemporary works [Chen et al., 2020] that study the effect of clipping on DP-GD, they are orthogonal to the results in this paper. We focus on formal excess empirical risk guarantees, whereas Chen et al. [2020] focus on understanding the gradient profile generated by DP-GD due to clipping.

1.1 Our Contributions

Dimension-independent excess empirical risk bounds for convex generalized linear models (GLMs): In Section 3.1, we consider a class of problems with loss functions of the form $\ell(\langle \theta, \mathbf{x} \rangle; y)$, where $\mathbf{x} \in \mathcal{X} \subset \mathbf{R}^p$ is the feature vector, $y \in \mathbf{R}$ is the response variable, and ℓ is convex in the first parameter. We show that if the optimization is over an unconstrained space (i.e., $\theta \in \mathbf{R}^p$), then for an objective function $\mathcal{L}(\theta; D) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\langle \theta, \mathbf{x}_i \rangle; y_i\right)$, one can achieve an excess empirical risk of $\widetilde{O}\left(L\sqrt{\texttt{rank}}/(\varepsilon \cdot n)\right)$, where $\texttt{rank} \leq \min\{n, p\}$ is the rank of the feature matrix $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$. To the best of our knowledge, *this is the first rank-based excess empirical risk bound for private convex GLMs*. Notice that the bound does not have any explicit dependence on the dimensionality p. We achieve this bound by optimizing on $\mathcal{L}(\theta; D)$ using *differentially private gradient descent* (DP-GD) [Bassily et al., 2014, Talwar et al., 2014, Song et al., 2013]. We also obtain an excess population risk of the form $\widetilde{O}(L \cdot \min\{1/\sqrt{n}, \sqrt{\texttt{rank}}/(\varepsilon \cdot n)\})$, which is equivalent to the optimal excess population risk obtained by prior works [Bassily et al., 2019a, 2020, 2019b], except that the ambient dimensionality p is replaced by $\texttt{rank}$[2].

Existing lower bounds for constrained private convex learning [Bassily et al., 2014] (i.e., $\theta \in \mathcal{C} \subsetneq \mathbf{R}^p$) show that for excess empirical risk, an explicit polynomial dependence on the dimensionality of the model space (p) is necessary. In contrast, our bound only depends on the $\texttt{rank}$ of the feature matrix X. Our main insight is that for DP-GD on generalized linear problems, the gradients lie in a low-rank subspace. The noisy gradients that DP-GD uses for state updates do not significantly impact this low-rank structure due to the *spherical* (and *stable*) nature of the Gaussian distribution. Our results extend to the local differentially private (LDP) [Warner, 1965, Evfimievski et al., 2003, Kasiviswanathan et al., 2008] variant of DP-GD, albeit with an increase of a $\sqrt{n}$ factor in the excess empirical risk [Duchi et al., 2018].

While Jain and Thakurta [2014] proved a related dimension-independent risk guarantee for two other differentially private algorithms, namely output perturbation [Chaudhuri et al., 2011] and objective perturbation [Chaudhuri et al., 2011, Kifer et al., 2012], our result is notable in the following aspects. First, we provide a more fine-grained control via the $\texttt{rank}$ parameter. The result in Jain and Thakurta [2014] only provides guarantees where $\texttt{rank}$ is upper-bounded by n. Second, Jain and Thakurta [2014] crucially relies on the existence of a centralized data source, whereas our result extends seamlessly to the LDP setting. Third, unlike the algorithms in Jain and Thakurta [2014], DP-GD does not require convexity to ensure privacy. This is important because even if the overall optimization function is non-convex, DP-GD still ensures differential privacy [Bassily et al., 2014, Abadi et al., 2016]. Depending on the optimization profile, we may still observe a dimension-independent convergence. We provide more evidence of this phenomenon in Section 4.

Additionally, we obtain a population risk guarantee that is asymptotically the same as the optimal excess population

[2] In the context of population risk, $\texttt{rank}$ refers to the rank of the covariance matrix for the data generating distribution.

risk in Bassily et al. [2019b], except that the ambient dimensionality (p) is replaced by $\texttt{rank}$. In Section 6, we provide empirical evidence demonstrating the dimension-independence of DP-GD with Gaussian noise on logistic regression.

Tight lower bound on excess empirical risk for convex GLMs: In Section 3.2, we show that our dimension-independent upper bound on the excess empirical risk for unconstrained convex GLM achieved via DP-GD is tight. This lower bound is in sharp contrast to that in Bassily et al. [2014], where they show that for GLMs in the constrained setting, an explicit dependence on the dimensionality (p) is necessary. It is worth mentioning that *our lower bound is the first for any unconstrained private convex ERM*. Our lower bound is proved by transforming a GLM instance (namely, $\ell(\langle\theta, \mathbf{x}\rangle; y) = |\langle\mathbf{x}, \theta\rangle - y|$) to estimating one-way marginals, to which we can apply fingerprinting techniques [Bun et al., 2018, Steinke and Ullman, 2017].

Dimension-independent convergence to a first-order stationary point for non-convex GLMs: In Section 4, we extend our dimension-independent result to non-convex generalized linear problems, i.e., where the loss function ℓ can be non-convex but preserves the inner-product structure. We show that for this class of problems, DP-GD converges to a first-order stationary point (FOSP) (i.e., where the gradient of the objective function is zero). Again, this convergence guarantee is independent of the model dimensionality, and only depends on $\texttt{rank}$ of the feature matrix. Specifically, we show that if the loss function for the non-convex generalized linear problem is smooth and L-Lipschitz in the ℓ_2 norm, then DP-GD (paired with the exponential mechanism [McSherry and Talwar, 2007]) outputs a model θ_{priv} such that the gradient of the objective function $\mathcal{L}(\theta; D)$ at θ_{priv} has ℓ_2-norm of $\widetilde{O}\left(L\sqrt{\texttt{rank}}/(\varepsilon n)\right)$.

While there has been work on understanding the convergence of variants of DP-GD on non-convex losses [Wang et al., 2019], ours is the first result to demonstrate a dimension-independent convergence. At the heart of our result is a simple folklore argument stated in Allen-Zhu [2018] that shows first-order convergence of GD for non-convex objectives. We conjecture that our result can be extended to second-order convergence (analogous to Wang et al. [2019]) under additional assumptions on the loss function. A natural direction would be to modify the argument of Jin et al. [2017] to make it amenable to DP-GD.

Analysis of clipped differentially private gradient descent (DP-GD) on convex GLMs: While the upper bounds in this paper are achieved by DP-GD, one major caveat for using the algorithm in practice is that it requires a predefined upper bound of L on $\|\partial_\theta \ell(\langle\theta, \mathbf{x}\rangle; y)\|_2$ for all $\mathbf{x}, y$. (Here, ∂_θ corresponds to the subgradients.) In real-world applications, L is almost never known a priori. As a result, a variant of DP-GD (called clipped DP-GD) is used in practice where the individual subgradients corresponding to each data sample $(\mathbf{x}, y)$ are scaled/clipped to ensure that they are upper bounded by a predefined quantity $B > 0$, a.k.a. the clipping norm [Abadi et al., 2016, Papernot et al., 2020, Chen et al., 2020]. In Section 5, we show that no matter what the clipping norm B is, for convex GLMs, clipped DP-GD optimizes a well-defined convex objective (denoted by $\mathcal{L}^{(B)}_{\mathsf{clipped}}(\theta; D)$) corresponding to $\mathcal{L}(\theta; D)$. Furthermore, the excess empirical risk with respect to $\mathcal{L}^{(B)}_{\mathsf{clipped}}$ is $\widetilde{O}\left(B\sqrt{\texttt{rank}}/(\varepsilon \cdot n)\right)$, with $\texttt{rank}$ being the rank of the feature matrix X. Notice the same dimension-independent convergence for clipped DP-GD as that of vanilla DP-GD. We also show that if $B \geq L$, then $\mathcal{L}^{(B)}_{\mathsf{clipped}}(\theta; D)$ equals $\mathcal{L}(\theta; D)$ point-wise. To prove the above bound, we provide a structural lemma (see Section 5.1, which characterizes the clipping operation as a variant of Huberization [Huber and Ronchetti, 1981] for convex GLMs. To our knowledge, this is the *first convergence guarantee for clipped DP-GD*.

As an interlude, in Section 5.3, we show that the convergence guarantees of clipped DP-GD are sensitive to the choice of the clipping norm B. Setting it low (i.e., $B \ll L$) can result in strange behaviors in the optimization profile, ranging from introducing $\Omega(1)$ bias in the excess empirical risk, to generating vectors that do not conform to the gradient field of any "natural" convex function.

We note that there is a line of work on the practice and theory of gradient clipping [Goodfellow et al., 2016, Pascanu et al., 2012, 2013, Zhang et al., 2019]. Despite the similarity in name, these algorithms are different as they clip the *averaged* gradient in each step, while in clipped DP-GD, we need the *individual* gradient to be clipped to get a reasonable privacy/utility trade-off.

2 Preliminaries

In this section, we provide the some of the concepts required in the rest of the paper.

Definition 2.1 (Seminorm). *Given a vector space V over a field F of the real numbers $\mathbf{R}$, a seminorm on V is a nonnegative-valued function $\rho: V \to \mathbf{R}$ with the following properties. For all $a \in F$, and $\boldsymbol{u}, \boldsymbol{v} \in V$:*

1. **Triangle inequality***: $\rho(\boldsymbol{u} + \boldsymbol{v}) \leq \rho(\boldsymbol{u}) + \rho(\boldsymbol{v})$.*
2. **Absolute scalability***: $\rho(a \cdot \boldsymbol{u}) = |a| \cdot \rho(\boldsymbol{u})$.*

Lipschitzness, Convexity, and Smoothness: We additionally require the following definitions to state our results. These properties usually govern the rate of convergence of an algorithm for optimizing ERMs.

Definition 2.2 (ℓ_2-Lipschitz continuity). *A function $f: \mathcal{C} \to \mathbf{R}$ is L-Lipschitz w.r.t. the ℓ_2-norm over a set $\mathcal{C} \subseteq \mathbf{R}^p$*

if the following holds: $\forall \theta_1, \theta_2 \in \mathcal{C}, |f(\theta_1) - f(\theta_2)| \leq L \cdot \|\theta_1 - \theta_2\|_2$.

Definition 2.3 ((Strong) convexity w.r.t. ℓ_2-norm)**.** *A function $f : \mathcal{C} \rightarrow \mathbf{R}$ is Δ-strongly convex w.r.t. the ℓ_2-norm over a set $\mathcal{C} \subseteq \mathbf{R}^p$ if $\forall \alpha \in (0, 1), (\theta_1, \theta_2) \in \mathcal{C} \times \mathcal{C}$:*

$$f(\alpha\theta_1 + (1-\alpha)\theta_2) \leq \alpha f(\theta_1) + (1-\alpha) f(\theta_2) - \Delta \frac{\alpha(\alpha-1)}{2} \|\theta_1 - \theta_2\|_2^2 .$$

Function f is simply convex if the above holds for $\Delta = 0$.

Definition 2.4 (Smoothness)**.** *A function $f : \mathcal{C} \rightarrow \mathbf{R}$ is β-smooth on $\mathcal{C} \subseteq \mathbf{R}^p$ if for all $\theta_1 \in \mathcal{C}$ and for all $\theta_2 \in \mathcal{C}$, we have $f(\theta_2) \leq f(\theta_1) + \langle \nabla f(\theta_1), \theta_2 - \theta_1 \rangle + \frac{\beta}{2} \|\theta_1 - \theta_2\|_2^2$.*

Differential Privacy: In this paper, we focus on approximate differential privacy (DP) [Dwork et al., 2006b,a].

Definition 2.5 (Differential privacy [Dwork et al., 2006b,a])**.** *A randomized algorithm $\mathcal{A}$ is (ε, δ)-differentially private if, for any pair of datasets D and D' differing in exactly one data point (i.e., one data point is replaced in the other), and for all events $\mathcal{S}$ in the output range of $\mathcal{A}$, we have*

$$\mathbf{Pr}[\mathcal{A}(D) \in \mathcal{S}] \leq e^{\varepsilon} \cdot \mathbf{Pr}[\mathcal{A}(D') \in \mathcal{S}] + \delta,$$

where the probability is taken over the random coins of $\mathcal{A}$.

For meaningful privacy guarantees, ε is assumed to be a small constant, and $\delta \ll 1/n$ for $n = |D|$.

Empirical risk minimization (ERM): Let $D = \{d_1, \cdots, d_n\} \subseteq \mathcal{D}^n$ be a data set of n samples drawn from the domain $\mathcal{D}$, and for $\mathcal{C} \subseteq \mathbf{R}^p$ being the model space, let $\ell : \mathcal{C} \times \mathcal{D} \rightarrow \mathbf{R}$ be a loss function. Then the empirical risk over the data set D is defined as $\mathcal{L}(\theta; D) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; d_i)$. The objective of an *empirical risk minimization* (ERM) algorithm is to output a model $\theta \in \mathcal{C}$ that approximately minimizes the empirical risk $\mathcal{L}$ over the set $\mathcal{C}$. For the theoretical guarantees in this paper, we will only look at ERM loss, and the *excess empirical risk* $R(\theta) = \mathcal{L}(\theta; D) - \mathcal{L}(\theta^*; D)$, where $\theta^* = \arg\min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D)$. By stability-based arguments [Bassily et al., 2014, Shalev-Shwartz et al., 2009], one can easily translate excess empirical risk for differentially private algorithms to their corresponding *excess population risk*, where the population risk for θ is defined as $\mathbb{E}_{d \sim \mathcal{T}}[\ell(\theta; d)]$, with $\mathcal{T}$ being a given distribution over $\mathcal{D}$.

If $\mathcal{C} \subsetneq \mathbf{R}^p$, we are in the so-called constrained ERM setting, whereas $\mathcal{C} = \mathbf{R}^p$ is denoted as the unconstrained setting. In this paper, we focus on the *unconstrained* setting.

Generalized Linear Models: For most of this paper, we focus on a special class of ERM problems called generalized linear models [Shalev-Shwartz et al., 2009], where the loss function $\ell(\theta; d)$ takes a special inner-product form $\ell(\langle \theta, \mathbf{x} \rangle; y)$ for $d = (\mathbf{x}, y)$. Here, $\mathbf{x} \in \mathbf{R}^p$ is usually called the feature vector and $y \in \mathbf{R}$ the response. Instead of being the feature vector in the original data, $\mathbf{x}$ can also be thought of as representing a mapped value $\phi(\mathbf{x})$ of original feature vector. We do not make the distinction here.

Algorithm 1 $\mathcal{A}_{\text{DP-GD}}$: Diff. private gradient descent

Input: Data set $D = \{d_1, \cdots, d_n\}$, loss function: $\ell : \mathbf{R}^p \times \mathcal{D} \rightarrow \mathbf{R}$, gradient ℓ_2-norm bound: L, constraint set: $\mathcal{C} \subseteq \mathbf{R}^p$, number of iterations: T, noise variance: σ^2, learning rate: η.

1: $\theta_0 \leftarrow \mathbf{0}$.
2: **for** $t = 0, \ldots, T-1$ **do**
3: $\quad \boldsymbol{g}_t^{\texttt{priv}} \leftarrow \frac{1}{n} \sum_{i=1}^{n} \partial_\theta \ell(\theta_t; d_i) + \mathcal{N}\left(0, \sigma^2\right)$.
4: $\quad \theta_{t+1} \leftarrow \Pi_{\mathcal{C}}\left(\theta_t - \eta \cdot \boldsymbol{g}_t^{\texttt{priv}}\right)$, where $\Pi_{\mathcal{C}}(\boldsymbol{v}) = \arg\min_{\theta \in \mathcal{C}} \|\boldsymbol{v} - \theta\|_2$.
5: **end for**
6: **return** $\theta^{\texttt{priv}} = \frac{1}{T} \sum_{t=1}^{T} \theta_t$.

Differentially Private Gradient Descent: We provide a formal description of Differentially Private Gradient Descent (DP-GD) in Algorithm 1. In this version, the gradient $\boldsymbol{g}_t$ is computed over the whole data set, and the output $\theta^{\texttt{priv}}$ is the average of the models over all iterations. In practice, we may instead use differentially private stochastic gradient descent (DP-SGD), where $\boldsymbol{g}_t$ is computed over a random mini-batch, and the output $\theta^{\texttt{priv}}$ is the last model. While our analytical results are for the former setting (due to brevity), they extend to the latter with mild modifications to the proofs.

Theorem 2.6 (From Abadi et al. [2016], Mironov [2017])**.** *If $\|\partial_\theta \ell(\theta; d)\|_2 \leq L$ for any $\theta \in \mathcal{C}$ and $d \in \mathbf{R}^p$, then, Algorithm $\mathcal{A}_{\text{DP-GD}}$ (Algorithm 1) is (ε, δ)-differentially private if the noise variance is $\sigma^2 = \frac{2L^2 T \log(1/\delta)}{(n\varepsilon)^2}$.*

Theorem 2.7 (From Bassily et al. [2014], Talwar et al. [2014])**.** *If the constraint set $\mathcal{C}$ is convex, the loss function $\ell(\theta; d)$ is convex in the first parameter, $\|\partial_\theta \ell(\theta; d)\|_2 \leq L$ for all $\theta \in \mathcal{C}$ and $d \in D$, then for objective function $\mathcal{L}(\theta; D) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; d_i)$, under appropriate choices of the learning rate and the number of iterations in Algorithm $\mathcal{A}_{\text{DP-GD}}$ (Algorithm 1), we have*

$$\mathbb{E}\left[\mathcal{L}\left(\theta^{priv}; D\right)\right] - \mathcal{L}\left(\theta^*; D\right) \leq \frac{L\|\theta_0 - \theta^*\|_2 \sqrt{p \log(1/\delta)}}{\varepsilon n},$$

where $\theta^* = \arg\min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D)$ is the minimizer and $\theta_0 \in \mathcal{C}$ is the initial model.
The corresponding high-probability version is as follows.

With probability at least $1-\beta$, we have

$$\mathcal{L}\left(\theta^{priv};D\right)-\mathcal{L}\left(\theta^*;D\right) \leq \frac{L\|\theta_0-\theta^*\|_2\sqrt{p\log(1/\delta)\log(1/\beta)}}{\varepsilon n}.$$

3 Tight Excess Empirical Risk for Convex Generalized Liner Models (GLMs)

In Section 3.1, we show that algorithm $\mathcal{A}_{\text{DP-GD}}$ (Algorithm 1) converges to an excess empirical risk of $\widetilde{O}(\sqrt{\texttt{rank}}/\varepsilon n)$, which in particular is independent of the ambient dimensionality (p). (Here $\texttt{rank}$ is the rank of the feature matrix for GLMs.) In comparison, prior excess empirical risk of $\widetilde{O}(\sqrt{p}/\varepsilon n)$ (via Theorem 2.7) has an explicit dependence p for Algorithm $\mathcal{A}_{\text{DP-GD}}$.

In Section 3.2 we further show that this bound is tight. It is the *first lower bound on excess empirical risk for any unconstrained Lipschitz optimization problem.*

3.1 Upper-bound via Private Gradient Descent

Consider the following convex optimization problem. Let $\mathcal{L}(\theta;D)=\frac{1}{n}\sum\limits_{i=1}^{n}\ell(\langle\theta,\mathbf{x}_i\rangle;y_i)$ be an objective function defined over the data set $D=\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)\}$ with $\mathbf{x}_i\in\mathcal{X}$ and $y_i\in\mathbf{R}$ for all $i\in[n]$, where $\mathcal{X}\subset\mathbf{R}^p$ is a bounded set. Assume the loss function $\ell(\langle\theta,\mathbf{x}\rangle;y)$ is convex in its first parameter and is L-Lipschitz with respect to the ℓ_2-norm over all $\theta\in\mathbf{R}^p$ and for all $\mathbf{x}$ and y. In other words, the ℓ_2-norm of $\partial_\theta\ell$ is upper bounded by L. The objective is to output θ^{priv} that approximately solves $\arg\min\limits_{\theta\in\mathbf{R}^p}\mathcal{L}(\theta;D)$ while satisfying differential privacy. In the following, we show that the excess empirical risk for $\mathcal{A}_{\text{DP-GD}}$ (Algorithm 1) scales approximately as $\widetilde{O}(\sqrt{\texttt{rank}}/(\varepsilon n))$. Here $\texttt{rank}$ is the rank of the feature matrix $[\mathbf{x}_1|\mathbf{x}_2|\cdots|\mathbf{x}_n]$ formed by stacking the feature vectors as columns. In Section 3.2, we show that this bound is indeed tight.

Theorem 3.1. *Let $\theta_0=\mathbf{0}^p$ be the initial point of $\mathcal{A}_{\text{DP-GD}}$. Let $\theta^*=\arg\min\limits_{\theta\in\mathbf{R}^p}\mathcal{L}(\theta;D)$, and M be the projector to the eigenspace of the matrix $\sum\limits_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T$. Letting L be the gradient ℓ_2-norm bound. Setting the constraint set $\mathcal{C}=\mathbf{R}^p$ and running $\mathcal{A}_{\text{DP-GD}}$ on $\mathcal{L}(\theta;D)$ for $T=n^2\varepsilon^2$ steps with appropriate learning rate η, we get*

$$\mathbb{E}\left[\mathcal{L}\left(\theta^{priv};D\right)\right]-\mathcal{L}\left(\theta^*;D\right) \leq \frac{L\left\|\theta^*\right\|_M\sqrt{1+2\cdot \mathit{rank}(M)\cdot\log(1/\delta)}}{\varepsilon n}.$$

Here, $\mathit{rank}(M)\leq n$ (but can be much smaller), and $\|\cdot\|_M$ is the seminorm w.r.t. the projector M.

Though our result is for excess empirical risk, it can be translated to excess population risk guarantees via standard stability-based arguments [Bassily et al., 2020]. (We provide the details below.) The crux of our proof technique is to work in the subspace generated by the feature vectors for generalized linear problem. We proved the guarantees for DP-GD that uses full gradient and returns the average of the models over all iterations. Our proof would extend seamlessly (by modifying the proofs of Theorems 1 and 2 in Shamir and Zhang [2013]) to settings where stochastic gradients over mini-batches are used and the final model is returned as the output.

Proof of Theorem 3.1. We prove the theorem via the standard template for analyzing SGD methods [Bubeck, 2015]. Recall $\theta^{\text{priv}}=\frac{1}{T}\sum\limits_{t=1}^{T}\theta_t$, where $\{\theta_1,\ldots,\theta_T\}$ are the models in each iterate of DP-GD. Let $\boldsymbol{g}_t$ denote any subgradient in $\partial\mathcal{L}(\theta_t;D)$. By convexity and the standard linearization trick in convex optimization [Bubeck, 2015], we have:

$$\mathcal{L}\left(\theta^{\text{priv}};D\right)-\mathcal{L}\left(\theta^*;D\right)\leq\frac{1}{T}\sum_{t=1}^{T}\langle\boldsymbol{g}_t,\theta_t-\theta^*\rangle \qquad (1)$$

Let V be the eigenbasis of $\sum\limits_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T$ and let $M=VV^T$. M is a positive semidefinite matrix and it defines a seminorm $\|\cdot\|_M$ (by Definition 2.1). Let $\boldsymbol{b}_t$ be the Gaussian noise vector added at time step t. To bound the error in (1), we will use a potential argument w.r.t. the potential function

$$\begin{aligned}\Psi_t(\theta)&=\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_t}\left[\|\theta-\theta^*\|_M^2\right]\\&=\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_{t-1}}\left[\mathbb{E}_{\boldsymbol{b}_t}\left[\|\theta-\theta^*\|_M^2\,\middle|\,\boldsymbol{b}_1,\ldots,\boldsymbol{b}_{t-1}\right]\right].\end{aligned}$$

Recall that the update step in DP-GD is $\theta_{t+1}\leftarrow\theta_t-\eta\left(\boldsymbol{g}_t+\boldsymbol{b}_t\right)$. We get the following by simple algebraic manipulation:

$$\begin{aligned}\Psi_t(\theta_{t+1})&=\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_t}\left[\|(\theta_t-\theta^*)-\eta(\boldsymbol{g}_t+\boldsymbol{b}_t)\|_M^2\right]\\&=\Psi_t(\theta_t)-2\eta\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_t}\left[\langle\boldsymbol{g}_t+\boldsymbol{b}_t,\theta_t-\theta^*\rangle_M\right]\\&\quad+\eta^2\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_t}\left[\|\boldsymbol{g}_t+\boldsymbol{b}_t\|_M^2\right] \qquad (2)\end{aligned}$$

$$\begin{aligned}&=\Psi_t(\theta_t)-2\eta\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_t}\left[\langle\boldsymbol{g}_t+\boldsymbol{b}_t,\theta_t-\theta^*\rangle\right]\\&\quad+\eta^2\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_t}\left[\|\boldsymbol{g}_t+\boldsymbol{b}_t\|_M^2\right] \qquad (3)\\&\leq\Psi_t(\theta_t)-2\eta\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_{t-1}}\left[\langle\boldsymbol{g}_t,\theta_t-\theta^*\rangle\right]\\&\quad+\eta^2\left(L^2+\mathbb{E}_{\boldsymbol{b}_t}\left[\|\boldsymbol{b}_t\|_M^2\right]\right)\\&=\Psi_{t-1}(\theta_t)-2\eta\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_{t-1}}\left[\langle\boldsymbol{g}_t,\theta_t-\theta^*\rangle\right]\\&\quad+\eta^2\left(L^2+\mathbb{E}_{\boldsymbol{b}_t}\left[\|\boldsymbol{b}_t\|_M^2\right]\right)\\&=\Psi_{t-1}(\theta_t)-2\eta\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_{t-1}}\left[\langle\boldsymbol{g}_t,\theta_t-\theta^*\rangle\right]\\&\quad+\eta^2\left(L^2+\texttt{rank}(M)\cdot\sigma^2\right). \qquad (4)\end{aligned}$$

where (3) follows because $\boldsymbol{g}_t$ lies in the subspace M, and (4) follows because $\boldsymbol{b}_t \sim \mathcal{N}(0, \sigma^2 I_p)$ and thus $\mathbb{E}_{\boldsymbol{b}_t}\left[\|\boldsymbol{b}_t\|_M^2\right] = \mathtt{rank}(M)\cdot\sigma^2$. Rearranging the terms in (4), we have the following.

$$\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_{t-1}}\left[\langle \boldsymbol{g}_t, \theta_t - \theta^*\rangle\right] \leq \frac{1}{2\eta}\left(\Psi_{t-1}(\theta_t) - \Psi_t(\theta_{t+1})\right) + \frac{\eta}{2}\left(L^2 + \mathtt{rank}(M)\cdot\sigma^2\right). \tag{5}$$

Let $\Psi(\theta) = \|\theta - \theta^*\|_M^2$. Summing up (5) for all $t \in [T]$, averaging over the T iterations, and combining with (1), we get:

$$\mathbb{E}\left[\mathcal{L}\left(\theta^{\mathrm{priv}}; D\right)\right] - \mathcal{L}\left(\theta^*; D\right) \leq \frac{1}{2T\eta}\Psi(\mathbf{0}) + \frac{\eta}{2}\left(L^2 + \mathtt{rank}(M)\cdot\sigma^2\right) \tag{6}$$

Setting η to minimize the RHS, we have

$$\begin{aligned}&\mathbb{E}\left[\mathcal{L}\left(\theta^{\mathrm{priv}}; D\right)\right] - \mathcal{L}\left(\theta^*; D\right)\\ &\leq \|\theta^*\|_M\sqrt{\frac{L^2 + \mathtt{rank}(M)\cdot\sigma^2}{T}}\\ &= \|\theta^*\|_M\sqrt{\frac{L^2}{T} + \frac{2L^2\log(1/\delta)\cdot\mathtt{rank}(M)}{n^2\varepsilon^2}},\end{aligned}$$

where the equality follows by plugging in $\sigma = \frac{L\sqrt{2T\log(1/\delta)}}{n\varepsilon}$. Now, setting $T = n^2\varepsilon^2$, we have

$$\mathbb{E}\left[\mathcal{L}\left(\theta^{\mathrm{priv}}; D\right)\right] - \mathcal{L}\left(\theta^*; D\right) \leq \frac{L\|\theta^*\|_M\sqrt{1 + 2\cdot\mathsf{rank}(M)\cdot\log(1/\delta)}}{\varepsilon\cdot n}.$$

This completes the proof. □

The lower-bound in Bassily et al. [2014] shows that if one performs constrained optimization with DP, then the excess empirical risk is $\widetilde{\Omega}(\sqrt{p}/(\varepsilon n))$. This lower bound holds true for generalized linear problems as well. However, since we consider *unconstrained* optimization here, the lower bound does not apply to our result. In fact, the lower bound does not hold even for general convex functions, as long as the underlying optimization problem is unconstrained. Subsequent to our work, Kairouz et al. [2020] showed that via adaptive preconditioning methods, one can get dimension independent empirical risk bounds for general convex problems as long as the gradients lie in a low-rank subspace. However, for the specific case of GLMs, their bounds are weaker than ours, and depend on $\max_{t\in T}\|\theta_t - \theta^*\|_2$, rather than $\|\theta^*\|_M$ (as defined in Theorem 3.1).

The guarantee in Theorem 3.1 is of the same flavor as in Jain and Thakurta [2014], wherein such a result was shown for two different DP algorithms, namely, *output perturbation*, and *objective perturbation* [Chaudhuri et al., 2011, Kifer et al., 2012]. Our result via $\mathcal{A}_{\mathsf{DP\text{-}GD}}$ improves the state-of-the-art in the following ways. First, the result in Jain and Thakurta [2014] was only for the worst-case setting where $\mathtt{rank}(M) = n$, whereas our results extend to any rank setting. Second, unlike output perturbation and objective perturbation, $\mathcal{A}_{\mathsf{DP\text{-}GD}}$ *does not* require convexity to ensure DP. As a result, $\mathcal{A}_{\mathsf{DP\text{-}GD}}$ can be applied to non-convex losses and may enjoy the same dimension-independent behavior as in the convex case. (We show evidence of this in Section 4.) Third, our results for DP-GD almost seamlessly transfer to the local differential privacy (LDP) setting[3] [Warner, 1965, Evfimievski et al., 2003, Kasiviswanathan et al., 2008]. This is the *first* dimension-independent excess risk guarantee in the LDP setting. Output perturbation and objective perturbation are incompatible with LDP, as they require a centralized dataset to operate.

Obtaining optimal excess population risk guarantee: We translate the excess empirical risk guarantee in Theorem 3.1 to population risk guarantee via the standard approach of uniform stability [Bassily et al., 2019b, 2020]. The argument in Theorem 3.2 is identical to that in [Bassily et al., 2020], except that we operate with $\|\cdot\|_M$ (defined in Theorem 3.2) instead of the ℓ_2-norm. Note the bound is asymptotically the same as the optimal excess population risk in Bassily et al. [2019b], except that the ambient dimensionality (p) is replaced with $\min\{\mathtt{rank}(M), n\}$. We provide the proof of Theorem 3.2 in Appendix A.

Theorem 3.2. *Let $\theta_0 = \mathbf{0}^p$ be the initial point of $\mathcal{A}_{\mathsf{DP\text{-}GD}}$. Let $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be drawn i.i.d. from a distribution τ. Let $\theta^*_{\mathsf{pop}} = \arg\min_{\theta\in\mathbf{R}^p}\mathbb{E}_{(\mathbf{x},y)\sim\tau}\left[\ell(\langle\mathbf{x},\theta\rangle; y)\right]$, and M be the projector to the eigenspace of the matrix $\mathbb{E}_{\mathbf{x}\sim\tau}\left[\mathbf{x}\mathbf{x}^T\right]$. Let L be the gradient ℓ_2-norm bound, and $k = \min\{rank(M), n\}$. Setting the constraint set $\mathcal{C} = \mathbf{R}^p$ and running $\mathcal{A}_{\mathsf{DP\text{-}GD}}$ on $\mathcal{L}(\theta; D)$ for $T = n^2$ steps with appropriate learning rate η, we obtain*

$$\mathbb{E}_{\mathcal{A}_{\mathsf{DP\text{-}GD}}, D, (\mathbf{x},y)\sim\tau}\left[\ell(\langle\theta^{priv}, \mathbf{x}\rangle; y) - \ell(\langle\theta^*_{\mathsf{pop}}, \mathbf{x}\rangle; y)\right] = \left\|\theta^*_{\mathsf{pop}}\right\|_M\cdot O\left(\max\left\{\frac{1}{\sqrt{n}}, \frac{L\sqrt{k\cdot\log(1/\delta)}}{\varepsilon n}\right\}\right).$$

3.2 Lower-bound via Fingerprinting Codes

Next, we prove the optimality of the upper bound in Theorem 3.1 up to a $O(\sqrt{\log(1/\delta)})$ factor[4] by proving a matching lower bound. The lower bound is attained by a simple convex GLM given by $\ell(\langle\theta,\mathbf{x}\rangle; y) = |\langle\theta,\mathbf{x}\rangle - y|$; this loss is L-Lipschitz for $L = \|\mathbf{x}\|_2$ and is always non-negative.

[3] Obtaining LDP guarantee results in scaling up the Gaussian noise in Algorithm $\mathcal{A}_{\mathsf{DP\text{-}GD}}$ by $\sqrt{n}$ factor.

[4] This dependence on the δ parameter can be introduced into the lower bound by a generic group privacy reduction [Steinke and Ullman, 2015].

We emphasize that our lower bound holds in the unconstrained setting where $\mathcal{C} = \mathbf{R}^p$. Prior lower bounds [Bassily et al., 2014] heavily relied on θ being constrained; indeed, if we remove the constraint on θ in prior lower bounds, either there is no minimum value (i.e., $\inf_{\theta\in\mathbf{R}^p} \mathcal{L}(\theta; D) = -\infty$) or, if we avoid this with a regularizer, the loss is no longer Lipschitz.

Theorem 3.3. *Let $\mathcal{A} : (\mathbf{R}^p \times \mathbf{R})^n \to \mathbf{R}^p$ be an arbitrary (ε, δ)-differentially private algorithm with $\varepsilon \le c$ and $\delta \le c\varepsilon/n$ for some universal constant $c > 0$. Let $1 \le d \le p$ be an integer. Let $\ell : \mathbf{R}^p \times (\mathbf{R}^p \times \mathbf{R}) \to \mathbf{R}$ be defined by $\ell(\langle\theta, \mathbf{x}\rangle; y) = |\langle\theta, \mathbf{x}\rangle - y|$. Then there exists a data set $D = ((\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)) \in (\mathbf{R}^p \times \mathbf{R})^n$ such that the following holds. For all $i \in [n]$, $y_i \in \{0, 1\}$ and $\|\mathbf{x}_i\|_2 \le 1$, $\|\partial_\theta \ell(\theta; (\mathbf{x}_i, y_i))\|_2 \le 1$ for all $\theta \in \mathbf{R}^p$. Let $\theta^{priv} = \mathcal{A}(D)$ and $\theta^* = \underset{\theta\in\mathbf{R}^p}{\arg\min}\, \mathcal{L}(\theta; D)$ (breaking ties towards lower $\|\theta\|_2$), where $\mathcal{L}(\theta; D) = \frac{1}{n}\sum_{i=1}^n \ell(\langle\theta, \mathbf{x}_i\rangle; y_i)$. Then we have $\mathtt{rank}(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top) \le d$ and $\|\theta^*\|_2 \le \sqrt{d}$ and*

$$\mathbb{E}\left[\mathcal{L}\left(\theta^{priv}; D\right)\right] - \mathcal{L}\left(\theta^*; D\right) \ge \Omega\left(\min\left\{1, \frac{d}{\varepsilon n}\right\}\right)$$
$$\ge \Omega\left(\min\left\{1, \frac{\|\theta^*\|_2 \cdot \sqrt{\mathtt{rank}\left(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\right)}}{\varepsilon n}\right\}\right).$$

The parameter d determines the complexity of the lower bound instance and controls the rank of the feature matrix.

We now sketch the proof of Theorem 3.3. The full details are in Appendix B. The proof is based on the powerful fingerprinting technique for proving differential privacy lower bounds [Bun et al., 2018, Steinke and Ullman, 2017, Dwork et al., 2015], which was originally developed in the cryptography literature [Boneh and Shaw, 1998, Tardos, 2008]. A fingerprinting code provides a sequence of vectors $\mathbf{z}_1, \cdots, \mathbf{z}_n \in \{0, 1\}^d$, which we use to construct the hard dataset with each $\mathbf{z}_i$ corresponding to the data of individual i. The guarantee of the fingerprinting code is that it is not possible to privately estimate the average $\frac{1}{n}\sum_{i=1}^n \mathbf{z}_i \in [0, 1]^d$ to within a certain level of accuracy (depending on the privacy parameters ε and δ, the dimensionality d, and the number of individuals n).

We construct the hard dataset by setting $y_i = \langle\mathbf{z}_i, \mathbf{x}_i\rangle$ for all $i \in [n]$. Intuitively, to obtain low loss $\mathcal{L}(\theta^{\mathrm{priv}}; D) = \frac{1}{n}\sum_{i=1}^n |\langle\theta^{\mathrm{priv}}, \mathbf{x}_i\rangle - y_i| = \frac{1}{n}\sum_{i=1}^n |\langle\theta^{\mathrm{priv}} - \mathbf{z}_i, \mathbf{x}_i\rangle|$ the algorithm must ensure $\theta^{\mathrm{priv}} \approx \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i$, but this is impossible to do privately due to the properties of the fingerprinting code. This is the basis of our lower bound.

The only remaining part the construction is the feature vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \{0, 1\}^d$. These are either standard basis vectors (i.e., one 1) or all zeros. These are chosen so that $\frac{1}{n}\sum_{i=1}^n |\langle\theta^{\mathrm{priv}} - \mathbf{z}_i, \mathbf{x}_i\rangle| = \Theta(\|\theta^{\mathrm{priv}} - \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i\|_1)$, which suffices for the above argument.

Independently and subsequent to our work, Cai et al. [2020] prove lower bounds for estimating the parameters of a GLM. In contrast, our lower bound is stated in terms of the loss. They also give dimension-dependent upper bounds.

4 Dimension Independent First-order Convergence for Non-convex GLMs

In this section, we provide an extension to Theorem 3.1 that captures the setting when the loss function $\ell(z; \cdot)$ may be non-convex in z. Such loss functions appear commonly in robust regression, such as Savage loss [Masnadi-Shirazi and Vasconcelos, 2009], Tangent loss [Masnadi-Shirazi et al., 2010], and tempered loss [Amid et al., 2019]. We show that as long as $\ell(z; \cdot)$ is β-smooth (see Definition 2.4), $\mathcal{A}_{\mathsf{DP\text{-}GD}}$ (Algorithm 1) approximately reaches a stationary point on the objective function $\mathcal{L}(\theta; D)$, where θ is called a stationary point if $\nabla\mathcal{L}(\theta; D) = 0$. Similar to Theorem 3.1, the convergence guarantee in this section will have no explicit dependence on the number of dimensions. We use a folklore argument stated in Allen-Zhu [2018] to prove our result. The proof of Theorem 4.1 is in Appendix C.

Theorem 4.1. *Recall the notation in Theorem 3.1. Let $\Theta = [\theta_0, \ldots, \theta_T]$ be the list of models output by $\mathcal{A}_{\mathsf{DP\text{-}GD}}$. Let $t_{\mathsf{priv}} \leftarrow \underset{t\in\{0,\ldots,T-1\}}{\arg\min} \|\nabla\mathcal{L}(\theta_t; D)\|_M + \mathsf{Lap}\left(\frac{4L}{n}\right)$. Then, the algorithm that outputs the set Θ in conjunction with t_{priv} is $(2\varepsilon, \delta)$-differentially private. Furthermore, as long as $T \ge \frac{\beta n^2\varepsilon^2 \cdot \mathcal{L}(\mathbf{0}^p; D)}{2L^2\log\left(\frac{1}{\delta}\right)}$, we have with probability at least $1 - \gamma$,*

$$\left\|\nabla\mathcal{L}(\theta_{t_{\mathsf{priv}}}; D)\right\|_2 = \left\|\nabla\mathcal{L}(\theta_{t_{\mathsf{priv}}}; D)\right\|_M$$
$$= O\left(\frac{L}{\varepsilon n} \cdot \sqrt{\mathtt{rank}(M) \cdot \log\left(\frac{1}{\delta}\right)\log\left(\frac{T}{\gamma}\right)}\right).$$

Here, L is the Lipschitz constant, β is the smoothness constant of $\mathcal{L}(\theta; D)$. We set a constant learning rate in Algorithm 1 as $\eta = \frac{1}{\beta}$, and $\theta_0 = \mathbf{0}^p$. Notice that $\mathtt{rank}(M) \le n$ always holds but $\mathtt{rank}(M)$ can be much smaller than n.

The algorithm in Theorem 4.1 is a modification of $\mathcal{A}_{\mathsf{DP\text{-}GD}}$ (Algorithm 1), where from the list of models output by $\mathcal{A}_{\mathsf{DP\text{-}GD}}$ $\theta_0, \ldots, \theta_T$, we select the one with approximately minimum gradient norm for the loss $\mathcal{L}(\theta; D)$ via report-noisy-max (a.k.a. the exponential mechanism) [McSherry and Talwar, 2007, Dwork and Roth, 2014].

Theorem 4.1 does not immediately imply convergence to a local minima or a bound on the population risk. However, it demonstrates that convergence of DP-GD can be dimension-independent even in the case of non-convex losses. It is perceivable that this line of argument be extended for convergence to a local minimum using techniques similar to those in Jin et al. [2017]. However, that would require an additional assumption beyond smoothness, i.e., *Lipschitz continuity* of the Hessian.

5 Dimension-Independent Convergence of Clipped Private Gradient Descent

In Sections 3 and 4, we have seen that $\mathcal{A}_{\text{DP-GD}}$ (Algorithm 1) has dimension-independent convergence for unconstrained GLMs (convex or non-convex). Moreover, the bound is tight in the convex case. An issue with $\mathcal{A}_{\text{DP-GD}}$ is that it requires an apriori bound on the ℓ_2-norm of the gradients for all values of θ and d. In practice [Abadi et al., 2016, Papernot et al., 2020], when such a bound is unknown or non-existent, a modified version of $\mathcal{A}_{\text{DP-GD}}$, called *clipped* DP-GD, is used. In this algorithm, the gradients on individual data points are projected to ensure that they conform to a pre-defined ℓ_2-norm bound (denoted by *clipping norm* B). This is done by replacing Line 3 in Algorithm 1 by

$$\boldsymbol{g}_t^{\texttt{priv}} \leftarrow \frac{1}{n}\sum_{i=1}^{n} \texttt{clip}\left(\partial_\theta \ell(\theta_t; d_i)\right) + \mathcal{N}\left(0, \sigma^2\right) \quad (7)$$

where $\texttt{clip}(\boldsymbol{v}) = \boldsymbol{v} \cdot \min\left\{1, \frac{B}{\|\boldsymbol{v}\|_2}\right\}$. We denote this algorithm as $\mathcal{A}_{\text{DP-GD}}^{\text{clipped}}$.

Now, we show that $\mathcal{A}_{\text{DP-GD}}^{\text{clipped}}$ enjoys the same dimension-independent convergence as $\mathcal{A}_{\text{DP-GD}}$ as long as the underlying problem is *convex* GLM. For all values of clipping norm $B > 0$, the convergence is always to the minimum of a well-defined convex objective function. Based on the value of B (i.e., whether $B \leq L$), the convergence may be to the minimum of a different (well-specified) objective function, rather than the original objective $\mathcal{L}(\theta; D)$.

5.1 Characterizing Clipping in Convex GLMs

First, we provide an analytical tool in Lemma 5.1 to precisely quantify the objective function that $\mathcal{A}_{\text{DP-GD}}^{\text{clipped}}$ optimizes when the underlying loss function is a convex GLM.

Lemma 5.1. *Let $f : \mathbf{R} \to \mathbf{R}$ be any convex function and $B \in \mathbf{R}_+$ be any positive value. For any $\mathbf{x} \neq \mathbf{0}$, let $Y_1 = \left\{y : u < -\frac{B}{\|\mathbf{x}\|_2} \, \forall u \in \partial f(y)\right\}$ and $Y_2 = \left\{y : u > \frac{B}{\|\mathbf{x}\|_2} \, \forall u \in \partial f(y)\right\}$. If Y_1 is non-empty, let $y_1 = \sup Y_1$; otherwise $y_1 = -\infty$. If Y_2 is non-empty, let $y_2 = \inf Y_2$; otherwise $y_2 = \infty$. Let $g_{\mathbf{x}} : \mathbf{R} \to \mathbf{R}$ be*

$$g_{\mathbf{x}}(y) = \begin{cases} -\frac{B}{\|\mathbf{x}\|_2}(y - y_1) + f(y_1) & \text{for } y \in (-\infty, y_1) \\ f(y) & \text{for } y \in [y_1, y_2] \cap \mathbf{R} \\ \frac{B}{\|\mathbf{x}\|_2}(y - y_2) + f(y_2) & \text{for } y \in (y_2, \infty) \end{cases}.$$

Then the following holds.

1. *$g_{\mathbf{x}}$ is convex.*

2. *Let $\ell_f : \mathbf{R}^p \times \mathbf{R}^p \to \mathbf{R}$ be $\ell_f(\theta; \mathbf{x}) = f(\langle \theta, \mathbf{x} \rangle)$ for any $\theta, \mathbf{x}$. Let $\ell_g : \mathbf{R}^p \times \mathbf{R}^p \to \mathbf{R}$ be $\ell_g(\theta; \mathbf{x}) = g_{\mathbf{x}}(\langle \theta, \mathbf{x} \rangle)$ for any $\theta, \mathbf{x}$. Then, for any θ, $\mathbf{x}$, we have*

$$\partial_\theta \ell_g(\theta; \mathbf{x}) = \left\{\min\left\{1, \frac{B}{\|\boldsymbol{u}\|_2}\right\} \cdot \boldsymbol{u} : \boldsymbol{u} \in \partial_\theta \ell_f(\theta; \mathbf{x})\right\}.$$

Note: Lemma 5.1 is a generic tool for understanding the effect of clipping. In fact, it can be used to justify the use of standard private convex optimization analysis in [Bassily et al., 2014, Feldman et al., 2018, 2020, Bassily et al., 2019a] to $\mathcal{A}_{\text{DP-GD}}^{\text{clipped}}$ on convex GLMs.

The proof (see Appendix D.1) is based on the fact that clipping does not affect the monotonicity property of the derivative of one-dimensional convex function. A pictorial representation of Lemma 5.1 is given in Figure 1.

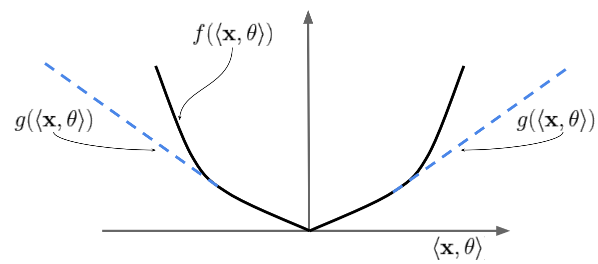

Figure 1: Representing the clipping operation in (7) for Algorithm $\mathcal{A}_{\text{DP-GD}}^{\text{clipped}}$ via Lemma 5.1.

5.2 Rank-based Convergence of Algorithm $\mathcal{A}_{\text{DP-GD}}^{\text{clipped}}$

Now, we provide the rank-based convergence of Algorithm $\mathcal{A}_{\text{DP-GD}}^{\text{clipped}}$. Recall that the original loss function is given by $\mathcal{L}(\theta; D) = \frac{1}{n}\sum_{i=1}^{n} \ell(\langle \theta, \mathbf{x}_i \rangle; y_i)$. Since, by Lemma 5.1 we know that for any clipping level $B > 0$, one can replace each loss function $\ell(\langle \theta, \mathbf{x}_i \rangle; y_i)$ with that obtained from Lemma 5.1. We denote these clipped losses as $\ell_{\text{clipped}}^{(B)}(\langle \theta, \mathbf{x}_i \rangle; y_i)$, and the corresponding objective function as $\mathcal{L}_{\text{clipped}}^{(B)}(\theta; D) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\text{clipped}}^{(B)}(\langle \theta, \mathbf{x}_i \rangle; y_i)$. We immediately have the following corollary from Theorem 3.1.

Corollary 5.2. *Let $\theta_0 = \mathbf{0}^p$ be the initial point for $\mathcal{A}_{\text{DP-GD}}^{\text{clipped}}$, and let $B > 0$ be the corresponding clipping norm. Let $\theta^*_{\text{clipped}} = \arg\min_{\theta \in \mathbf{R}^p} \mathcal{L}_{\text{clipped}}^{(B)}(\theta; D)$, and M be the projector to the eigenspace of matrix $\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$. For constraint set $\mathcal{C} = \mathbf{R}^p$, clipping norm L, and running $\mathcal{A}_{\text{DP-GD}}^{\text{clipped}}$ on $\mathcal{L}(\theta; D)$ for $T = n^2\varepsilon^2$ steps with appropriate learning rate η, we get*

$$\mathbb{E}\left[\mathcal{L}_{\text{clipped}}^{(B)}\left(\theta^{priv}; D\right)\right] - \mathcal{L}_{\text{clipped}}^{(B)}\left(\theta^*_{\text{clipped}}; D\right) \leq \frac{L\left\|\theta^*_{\text{clipped}}\right\|_M \sqrt{1 + 2 \cdot \texttt{rank}(M) \cdot \log(1/\delta)}}{\varepsilon n}.$$

Here, $\texttt{rank}(M) \leq n$ (but can be much smaller), and $\|\cdot\|_M$ is the seminorm w.r.t. the projector M.

From Corollary 5.2, it is immediate that as long as the clipping norm $B \geq L$ (where L is the ℓ_2-Lipschitz constant of the loss functions ℓ), $\mathcal{A}_{\mathsf{DP\text{-}GD}}^{\mathsf{clipped}}$ optimizes the original loss $\mathcal{L}$.

5.3 Interlude: Adverse Effects of Aggressive Clipping

Corollary 5.2 eludes two natural questions: i) Is the output of Algorithm $\mathcal{A}_{\mathsf{DP\text{-}GD}}^{\mathsf{clipped}}$ guaranteed to be good in terms of the empirical loss on the original loss $\mathcal{L}(\theta; D)$ when $B < L$, and ii) When the underlying loss *does not* satisfy the convex GLM property, is the objective function that gets optimized by $\mathcal{A}_{\mathsf{DP\text{-}GD}}^{\mathsf{clipped}}$ still well-defined as for convex GLMs? Our answers to both questions are *negative*.

To answer the first question, we construct an instance of logistic regression $\mathcal{L}(\theta; D)$ with L, the ℓ_2-Lipschitz constant of the individual loss functions, being 1. We show that if the clipping norm $B < 1/4$, then the excess empirical risk of Algorithm $\mathcal{A}_{\mathsf{DP\text{-}GD}}^{\mathsf{clipped}}$ on $\mathcal{L}(\theta; D)$ is $\Omega(1)$. Whereas Algorithm $\mathcal{A}_{\mathsf{DP\text{-}GD}}$ with $L = 1$ has an excess empirical risk of $O(1/n)$, where n is the number of training examples. A formal description of this bound is in Appendix D.2.

To answer the second question, we show that even for simple convex problems like multi-class logistic regression (that does not belong to the traditional GLM class), clipping in $\mathcal{A}_{\mathsf{DP\text{-}GD}}^{\mathsf{clipped}}$ can generate a sequence of "clipped gradient vectors" that do not conform to a gradient field of any "natural" convex function, though the output of $\mathcal{A}_{\mathsf{DP\text{-}GD}}^{\mathsf{clipped}}$ might still minimize the excess empirical risk on $\mathcal{L}(\theta; D)$ practically. We leave this exploration for future work.

Formally, consider a K-class classification problem for $K \geq 3$. Given a sample (x, y) with $x \in \mathbf{R}^p$ and $y \in [K]$, the cross-entropy loss $\ell : \mathbf{R}^{p\times K} \times \mathbf{R}^p \times [K] \to \mathbf{R}$ is, for $\theta = [\theta^{(1)}, \ldots, \theta^{(K)}]$,

$$\ell(\theta; (x, y)) = \sum_{k=1}^{K} \mathbb{1}(y = k) \log \frac{\exp\left(\theta^{(k)} \cdot x\right)}{\sum_{k'=1}^{K} \exp\left(\theta^{(k')} \cdot x\right)}.$$

Theorem 5.3. *Consider any sample (x, y) with $x \in \mathbf{R}^p \backslash \{\mathbf{0}\}$, $y \in [K]$ (for $K \geq 3$) and any $B > 0$ such that $\Theta = \{\theta : \|\nabla_\theta \ell(\theta; (x, y))\|_2 > B\}$ is non-empty. Let $G(\theta)$ be the clipped gradient of $\ell(\theta; (x, y))$ as defined above. Consider any function $f : \mathcal{C} \to \mathbf{R}$, $\mathcal{C} \subseteq \mathbf{R}^{p\times K}$ such that $\mathcal{C}^o \cap \Theta \neq \emptyset$, where $\mathcal{C}^o$ is the interior of set $\mathcal{C}$. If f is differentiable everywhere except for a set $\mathcal{C}_N \subseteq \mathcal{C}$ s.t. $\mathcal{C}_N$ is a closed set on $\mathcal{C}$ and has zero Lebesgue measure, then it is not possible for $\nabla_\theta f(\theta) = G(\theta)$ to hold for all $\theta \in \mathcal{C}^o \backslash \mathcal{C}_N$.*

As convexity implies differentiability almost everywhere [Rockafellar, 1970, Theorem 25.5], if f is convex, we only need $\mathcal{C}_N$ to be a closed set. The ℓ_1 regularizer $\|\theta\|_1$ is non-differentiable on a closed set, and so is the hinge loss $\ell_{\mathsf{hinge}}(\theta; (x, y)) = \max(0, 1 - y\langle\theta, x\rangle)$. Theorem 5.3 essentially rules out the possibility that the field of clipped gradients corresponds to any single objective function in convex models like softmax regression and SVMs with ℓ_1/ℓ_2 regularization, and in non-convex models like neural networks with popular activation functions like logistic sigmoid, tanh, ReLu. We defer the proof of Theorem 5.3 to Appendix D.3. One might further ask if the problem above could be resolved by "per-class" clipping, i.e., clipping $\nabla_{\theta^{(k)}} \ell$ individually for each k? The answer is still negative. We provide more details in Appendix D.3.1.

6 Experiments

We conduct experiments to demonstrate the dimension-independence of DP-GD with Gaussian noise. The setup mainly follows that from Jain and Thakurta [2014, Figure 1(c)]. We use a normalized version of the Cod-RNA dataset [Uzilov et al., 2006], and use logistic regression to solve for binary classification. Following Jain and Thakurta [2014], we append zero-valued features to the data samples, such that the accuracy of a non-private classifier does not change. To solve the problem privately, we consider DP-SGD with mini-batch gradient and Gaussian noise. For comparison, we consider DP-SGD with noise drawn from a Gamma distribution. We fix $\varepsilon \approx 5.0$ and $\delta = 10^{-5}$. Additional details on the setup can be found in Appendix E.

In Figure 2, we report the final test accuracy under different data dimensionality after tuning the learning rate. It is clear that DP-SGD with Gaussian noise is dimension-independent, as is predicted by Theorem 3.1. On the other hand, the accuracy of DP-SGD with Gamma noise decreases as data dimensionality increases.

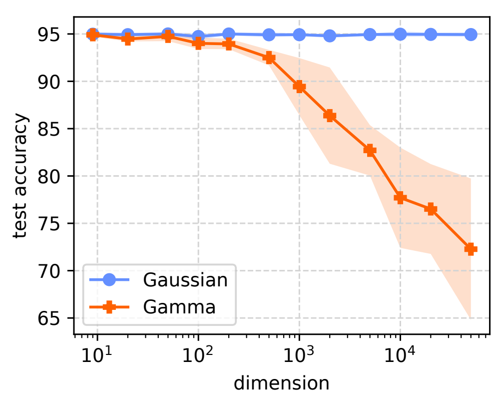

Figure 2: Test accuracy vs. data dimensionality for DP-SGD with i) Gaussian noise, and ii) Gamma noise.

Acknowledgements

The authors would like to thank Vitaly Feldman, Vivek Kulkarni, Monica Ribero, Shadab Shaikh, Adam Smith, Kunal Talwar, Andreas Terzis, and Jalaj Upadhyay for their helpful comments towards improving the paper.

References

M. Abadi, A. Chu, I. J. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security (CCS'16)*, pages 308–318, 2016.

Z. Allen-Zhu. Natasha 2: Faster non-convex optimization than sgd. In *Advances in neural information processing systems*, pages 2675–2686, 2018.

E. Amid, M. K. Warmuth, R. Anil, and T. Koren. Robust bi-tempered logistic loss based on bregman divergences. In *Advances in Neural Information Processing Systems*, pages 14987–14996, 2019.

R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proc. of the 2014 IEEE 55th Annual Symp. on Foundations of Computer Science (FOCS)*, pages 464–473, 2014.

R. Bassily, V. Feldman, K. Talwar, and A. Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, pages 11279–11288, 2019a.

R. Bassily, V. Feldman, K. Talwar, and A. G. Thakurta. Private stochastic convex optimization with optimal rates. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 11279–11288, 2019b.

R. Bassily, V. Feldman, C. Guzmán, and K. Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *arXiv preprint arXiv:2006.06914*, 2020.

D. Boneh and J. Shaw. Collusion-secure fingerprinting for digital data. *IEEE Transactions on Information Theory*, 44(5):1897–1905, 1998.

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

M. Bun, J. Ullman, and S. Vadhan. Fingerprinting codes and the price of approximate differential privacy. *SIAM Journal on Computing*, 47(5):1888–1938, 2018.

T. T. Cai, Y. Wang, and L. Zhang. The cost of privacy in generalized linear models: Algorithms and minimax lower bounds. *arXiv preprint arXiv:2011.03900*, 2020.

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

X. Chen, Z. S. Wu, and M. Hong. Understanding gradient clipping in private sgd: A geometric perspective. *arXiv preprint arXiv:2006.15429*, 2020.

J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

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

C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology—EUROCRYPT*, pages 486–503, 2006a.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proc. of the Third Conf. on Theory of Cryptography (TCC)*, pages 265–284, 2006b. URL `http://dx.doi.org/10.1007/11681878_14`.

C. Dwork, A. Smith, T. Steinke, J. Ullman, and S. Vadhan. Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 650–669. IEEE, 2015.

A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 211–222, 2003.

V. Feldman and T. Steinke. Generalization for adaptively-chosen estimators via stable median. In *Conference on Learning Theory*, 2017. URL `https://arxiv.org/abs/1706.05069`.

V. Feldman, I. Mironov, K. Talwar, and A. Thakurta. Privacy amplification by iteration. In *59th Annual IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 521–532, 2018.

V. Feldman, T. Koren, and K. Talwar. Private stochastic convex optimization: Optimal rates in linear time. In *Proc. of the Fifty-Second ACM Symp. on Theory of Computing (STOC'20)*, 2020.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.

P. Huber and E. M. Ronchetti. *Robust statistics*. Wiley New York, 1981.

R. Iyengar, J. P. Near, D. Song, O. Thakkar, A. Thakurta, and L. Wang. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, 2019.

P. Jain and A. G. Thakurta. (near) dimension independent risk bounds for differentially private learning. In *Inter-*

national Conference on Machine Learning, pages 476–484, 2014.

C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1724–1732. JMLR. org, 2017.

P. Kairouz, M. Ribero, K. Rush, and A. Thakurta. Fast dimension independent private adagrad on publicly estimated subspaces. *arXiv preprint arXiv:2008.06570*, 2020.

S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. D. Smith. What can we learn privately? In *49th Annual IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 531–540, 2008.

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

H. Masnadi-Shirazi and N. Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *Advances in neural information processing systems*, pages 1049–1056, 2009.

H. Masnadi-Shirazi, V. Mahadevan, and N. Vasconcelos. On the design of robust classifiers for computer vision. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 779–786. IEEE, 2010.

H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017.

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

I. Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

N. Papernot, A. Thakurta, S. Song, S. Chien, and Ú. Erlingsson. Tempered sigmoid activations for deep learning with differential privacy. *arXiv preprint arXiv:2007.14191*, 2020.

R. Pascanu, T. Mikolov, and Y. Bengio. Understanding the exploding gradient problem. *CoRR, abs/1211.5063*, 2: 417, 2012.

R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318, 2013.

V. Pichapati, A. T. Suresh, F. X. Yu, S. J. Reddi, and S. Kumar. Adaclip: Adaptive clipping for private sgd. *arXiv preprint arXiv:1908.07643*, 2019.

R. T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1970. ISBN 978-1-4008-7317-3.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009. URL `http://www.cs.mcgill.ca/%7Ecolt2009/papers/018.pdf#page=1`.

O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.

S. Song, K. Chaudhuri, and A. D. Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 245–248. IEEE, 2013.

T. Steinke and J. Ullman. Between pure and approximate differential privacy. *arXiv preprint arXiv:1501.06095*, 2015.

T. Steinke and J. Ullman. Tight lower bounds for differentially private selection. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, 2017. URL `https://arxiv.org/abs/1704.03024`.

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

G. Tardos. Optimal probabilistic fingerprint codes. *Journal of the ACM (JACM)*, 55(2):1–24, 2008.

O. Thakkar, G. Andrew, and H. B. McMahan. Differentially private learning with adaptive clipping. *CoRR*, abs/1905.03871, 2019. URL `http://arxiv.org/abs/1905.03871`.

J. Ullman. CS7880. Rigorous approaches to data privacy, Spring 2017. `http://www.ccs.neu.edu/home/jullman/PrivacyS17/HW1sol.pdf`, 2017.

A. V. Uzilov, J. M. Keegan, and D. H. Mathews. Detection of non-coding rnas on the basis of predicted secondary structure formation free energy change. *BMC bioinformatics*, 7(1):1–30, 2006.

D. Wang, C. Chen, and J. Xu. Differentially private empirical risk minimization with non-convex loss functions. In *International Conference on Machine Learning*, pages 6526–6535, 2019.

S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *J. of the American Statistical Association*, 60(309):63–69, 1965.

X. Wu, F. Li, A. Kumar, K. Chaudhuri, S. Jha, and J. F. Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In S. Salihoglu, W. Zhou, R. Chirkova, J. Yang, and D. Suciu, editors, *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD*, 2017.

J. Zhang, T. He, S. Sra, and A. Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2019.

Y. Zhou, Z. S. Wu, and A. Banerjee. Bypassing the ambient dimension: Private sgd with gradient subspace identification. *arXiv preprint arXiv:2007.03813*, 2020.