Characterizing Private Clipped Gradient Descent on Convex Generalized Linear Problems

Shuang Song
Google Research - Brain
shuangsong@google.com

Om Thakkar
Google
omthkkr@google.com

Abhradeep Thakurta
Google Research - Brain
athakurta@google.com

Abstract

Differentially private gradient descent (DP-GD) has been extremely effective both theoretically, and in practice, for solving private empirical risk minimization (ERM) problems. In this paper, we focus on understanding the impact of the clipping norm, a critical component of DP-GD, on its convergence. We provide the first formal convergence analysis of clipped DP-GD.

More generally, we show that the value which one sets for clipping really matters: *done wrong, it can dramatically affect the resulting quality; done properly, it can eliminate the dependence of convergence on the model dimensionality*. We do this by showing a dichotomous behavior of the clipping norm. First, we show that if the clipping norm is set smaller than the optimal, even by a constant factor, the excess empirical risk for convex ERMs can increase from $O(1/n)$ to $\Omega(1)$, where n is the number of data samples. Next, we show that, regardless of the value of the clipping norm, clipped DP-GD minimizes a well-defined convex objective over an unconstrained space, as long as the underlying ERM is a generalized linear problem. Furthermore, if the clipping norm is set within at most a constant factor higher than the optimal, then one can obtain an excess empirical risk guarantee that is independent of the dimensionality of the model space.

Finally, we extend our result to non-convex generalized linear problems by showing that DP-GD reaches a first-order stationary point as long as the loss is smooth, and the convergence is independent of the dimensionality of the model space.

1 Introduction

Over the past few years, there has been tremendous progress in differentially private convex empirical risk minimization (ERM) [9, 7, 38, 1, 6, 29, 44, 21, 33, 40, 15]. We know an almost-complete characterization of this problem in terms of upper and lower bounds [7, 6], both for excess empirical risk and excess population risk. Differentially private gradient descent (DP-GD) [39] (or its close variant, differentially private stochastic gradient descent (DP-SGD) [7, 38, 1]) provides the tightest upper bounds.

One important assumption in both the convergence and the privacy guarantees for DP-(S)GD is that the loss functions for the ERM problem are ℓ_2-Lipschitz with an explicitly known Lipschitz constant. When the Lipschitz constant is unknown or nonexistent, to guarantee privacy, the gradients of the *individual* loss functions are "clipped" to a bounded ℓ_2-norm, typically referred to as the *clipping norm* [1, 29, 33, 40]. We denote this variant by the *clipped DP-(S)GD*. While there has been empirical progress on adaptively adjusting the clipping norm to maximize the signal-to-noise ratio [33, 40], the fundamental impact of clipping norm on DP-(S)GD has not yet been studied. In this paper, we provide the *first convergence analysis of clipped DP-GD*[1] for convex generalized linear problems (defined in Section 2). We show that the clipping norm has a significant impact. *If set wrong, it can dramatically affect utility;*

[1] While the results in this paper extend to DP-SGD, for brevity, we will only focus on DP-GD.

if set properly, it can eliminate the dependence of convergence on the model dimensionality. We show a dichotomous behavior:

i) **Lower bound:** If the clipping norm is smaller (even by a constant factor) than the maximum ℓ_2-norm of the gradient for any of the individual loss function, then the excess empirical risk can increase from $O(1/n)$ to $\Omega(1)$, where n is the number of data samples. Furthermore, in certain ERM formulations (e.g., multiclass softmax regression), we show that the clipped gradients may not correspond to the gradient field of any "natural" convex/non-convex function.

ii) **Upper bound:** We provide the first formal convergence guarantees for clipped DP-GD. For unconstrained convex generalized linear problems, we show that clipped DP-GD minimizes a well-defined objective function (which still has the generalized linear problem structure, but may not correspond to the original ERM objective). Furthermore, we show that its convergence does not have any explicit dependence on the number of model parameters: If the clipping norm is within a constant multiple of the maximum ℓ_2-norm of the gradient for any of the individual loss function, then one can obtain an excess empirical risk guarantee of $\widetilde{O}(\sqrt{k}/\varepsilon n)$ for the original ERM objective, where k is the rank of the feature matrix, and ε is the privacy parameter.

In the following, we formally introduce the problem, and state our contributions. We note that there is a line of work on the practice and theory of gradient clipping [17, 31, 32, 45]. Despite the similarity in name, these algorithms are different as they clip the *averaged* gradient in each step, while in clipped DP-(S)GD, we need the *individual* gradient to be clipped to get a reasonable privacy-utility trade-off.

Problem setup: Given a data set $D = \{d_1, \ldots, d_n\}$ and an objective function $\mathcal{L}(\theta; D) = \frac{1}{n}\sum\limits_{i=1}^{n} \ell(\theta; d_i)$ with ℓ being some loss function, DP-GD (Algorithm 1) is an iterative procedure that optimizes $\mathcal{L}(\theta; D)$ as follows. At every time step t, 1) Compute $\boldsymbol{g}_t = \frac{1}{|n|}\sum\limits_{d \in D} \mathsf{clip}\left(\nabla \ell(\theta_t; d)\right)$, (an approximation of) the gradient at the current model θ_t. Here, $\mathsf{clip}(\boldsymbol{v}) = \boldsymbol{v} \cdot \min\left\{\frac{L}{\|\boldsymbol{v}\|_2}, 1\right\}$, and L is called the *clipping norm*, 2) Update model parameters as $\theta_{t+1} \leftarrow \theta_t - \eta \cdot \left(\boldsymbol{g}_t + \mathcal{N}(0, \sigma^2 \mathbb{1}_p)\right)$, where η is the learning rate, and $\mathcal{N}(0, \sigma^2 \mathbb{1}_p)$ is the random noise to guarantee privacy, with p denoting the number of model parameters. The variance σ^2 controls the strength of the achieved DP guarantee. Throughout the paper, our privacy guarantees are for (ε, δ)-DP (Definition 2.1) and our accuracy guarantees are for excess empirical risk $R(\theta) = \mathcal{L}(\theta; D) - \min\limits_{\theta'} \mathcal{L}(\theta'; D)$.

1.1 Our Contributions

In this paper, we provide upper and lower bounds on the convergence of clipped DP-GD on generalized linear problems. The results below, in particular, demonstrate the importance of choosing the right clipping norm. While the design of algorithms to choose the clipping norm optimally is beyond the scope of this paper, in Appendix D.1 we provide a discussion of the prior work on effectively choosing the clipping norm.

1. **Lower bound on the excess empirical risk:** In Section 3.1, we first provide a lower bound for *binary logistic regression* showing that the excess empirical risk can increase from $\widetilde{O}(1/n)$ to $\widetilde{\Omega}(1)$ if the clipping norm is *smaller* (even by a constant factor) than its optimal value. This lower bound holds for *unconstrained optimization*, i.e., where the model θ is allowed to be anywhere in $\mathbf{R}^p$. In particular, the lower bound only depends on the norm of the feature vectors, and does not depend on any bound on the model space. Consider the setting where the feature vectors have ℓ_2-norm bounded by 1. For any clipping norm $L < 1/4$, and for any n larger than a constant $n_0(L)$, we can construct a dataset with size n where the excess empirical risk of binary logistic regression for any DP algorithm is $\Omega\left(\log \frac{1}{L}\right)$. The proof of this lower bound follows by carefully exploiting the structure of a locally quadratic region in the logistic loss, and demonstrating that this region is destroyed if the clipping norm is not chosen properly.

   Additionally, to formalize our argument, we prove a structural lemma (Lemma 3.1) to precisely quantify the underlying optimization problem that clipped DP-GD solves. We show that clipping roughly corresponds to the Huberization operation [19] commonly used in robust statistics. More importantly, we show that clipping *does not* *impact convexity* for a common class of problems called generalized linear problems (which include binary logistic regression and linear regression). For any clipping norm, there exists a well-defined convex problem which clipped DP-GD optimizes. We use this lemma to upper bound the convergence of clipped DP-GD on convex generalized linear problems as well.

2. **Clipping introduces non-convexity on multiclass softmax regression:** In Section 3.2, we show that for softmax regression with more than two classes, there *does not exist* any "natural" function (convex or non-convex) that clipped DP-GD optimizes. Specifically, let $G(\theta)$ denote the clipped gradient of cross-entropy loss for the softmax regression. We show that there does not exist any function f which is differentiable everywhere except for a closed set with zero Lebesgue measure such that $\nabla_\theta f(\theta) = G(\theta)$ for all θ where f is differentiable. As a result, any of the excess empirical risk guarantees for private convex ERMs [7, 5] or private non-convex optimization [42] cease to hold.

3. **Dimension-independent excess empirical risk bounds for convex generalized linear problems:** We consider generalized linear problems, a class of problems with loss function $\ell(\langle\theta, x\rangle; y)$, where x is the feature vector and y is the response variable. If ℓ is convex in the first parameter, we call it a convex generalized linear problem. In Section 4.1, using the structural lemma mentioned earlier (Lemma 3.1), for each ERM problem $\mathcal{L}(\theta; D) = \frac{1}{n}\sum\limits_{i=1}^{n} \ell(\langle\theta, \mathbf{x}_i\rangle; y_i)$ and clipping norm L, we can find an objective function $\mathcal{L}^{(L)}_{\mathsf{clipped}}(\theta; D) = \frac{1}{n}\sum\limits_{i=1}^{n} \ell^{(L)}_{\mathsf{clipped}}(\langle\theta, \mathbf{x}_i\rangle; y_i)$ that the clipped DP-GD actually optimizes. The loss function $\ell^{(L)}_{\mathsf{clipped}}$ is still convex in its first parameter.

   We show that if the optimization is over an unconstrained space, then for objective function $\mathcal{L}^{(L)}_{\mathsf{clipped}}$, one can achieve an excess empirical risk of $\widetilde{O}(L\sqrt{\mathsf{rank}}/(\varepsilon \cdot n))$, where $\mathsf{rank} \leq n$ is the rank of the feature matrix. Furthermore, if the original loss function ℓ is B-Lipschitz w.r.t. the ℓ_2-norm, and the clipping norm $L \geq B$, then the above excess empirical risk corresponds to the original objective function $\mathcal{L}$. To the best of our knowledge, *this is the first formal convergence guarantee of clipped DP-GD for any objective function*.

   Existing lower bounds for constrained private convex learning [7] show that for excess empirical risk, a polynomial dependence on the dimensionality of the model space is necessary. In contrast, our bound only depends on rank. Our main insight is that, for DP-GD on generalized linear problems, the gradients lie in a low-rank subspace, and the noisy gradients that DP-GD operates on do not significantly impact this low-rank structure due to the spherical nature of the noise. Our results seamlessly extend to the local differentially private (LDP) variant of DP-GD (shown in Appendix D), albeit with an increase by a $\sqrt{n}$ factor in the excess empirical risk, which is necessary [10].

   While [22] proved a similar dimension-independent risk guarantee for two other differentially private algorithms, namely output perturbation [9] and objective perturbation [9, 25], our result is notable in the following aspects. First, we provide a more fine-grained control over the rank parameter. The result in [22] only provides guarantees where rank is upper-bounded by n. Second, [22] crucially relies on the existence of a centralized data source, whereas our result extends seamlessly to the LDP setting. Third, unlike the algorithms in [22], 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 [7, 1]. Depending on the optimization profile, we may still observe a dimension-independent convergence. We provide more evidence of this phenomenon below.

4. **Dimension-independent convergence to a first-order stationary point:** In Section 4.2, 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. Such problems appear commonly in robust regression [3, 28, 27]. We show that for this class of problems, DP-GD converges to a first-order stationary point (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 rank of the feature matrix. Specifically, we show that if the loss function for the non-convex generalized linear problem is smooth and B-Lipschitz in the ℓ_2 norm, then DP-GD (with mild modification) with clipping norm $L \geq B$ outputs a model θ_{priv} such that the gradient of the objective function at θ_{priv} has ℓ_2-norm $\widetilde{O}\left(L\sqrt{\mathsf{rank}}/(\varepsilon n)\right)$.

Algorithm 1 DP-GD: Differentially private gradient descent

```
**Input:** Data set $D = \{d_1, \cdots, d_n\}$, loss function: $\ell : \mathbf{R}^p \times \mathcal{D} \to \mathbf{R}$, clipping norm: $L$, constraint set: $\mathcal{C} \subseteq \mathbf{R}^p$, noise multiplier: $\lambda$, number of iterations: $T$, noise variance: $\sigma^2$, learning rate: $\eta$.
$\theta_0 \leftarrow \mathbf{0}$.
for $t = 0, \ldots, T-1$ do
    $\boldsymbol{g}_t = \frac{1}{n} \sum\limits_{i=1}^{n} \mathsf{clip}\left(\nabla \ell(\theta_t; d_i)\right)$, where $\mathsf{clip}(\boldsymbol{v}) = \boldsymbol{v} \cdot \min\left\{1, \frac{L}{\|\boldsymbol{v}\|_2}\right\}$.
    $\theta_{t+1} \leftarrow \Pi_{\mathcal{C}}\left(\theta_t - \eta\left(\boldsymbol{g}_t + \mathcal{N}\left(0, \sigma^2\right)\right)\right)$, where $\Pi_{\mathcal{C}}(\boldsymbol{v}) = \arg\min\limits_{\theta \in \mathcal{C}} \|\boldsymbol{v} - \theta\|_2$.
end for
return $\theta^{\texttt{priv}} = \frac{1}{T} \sum\limits_{t=1}^{T} \theta_t$.
```

While there has been work on understanding the convergence of variants of DP-(S)GD on non-convex losses [42], this is the first result to demonstrate a dimension-independent convergence. At the heart of our result is a simple folklore argument stated in [2] that shows first-order convergence of gradient descent for non-convex objectives. We conjecture that our result can be extended to second-order convergence (analogous to [42]) under additional assumptions on the loss function. A natural direction would be to modify the argument in [23] to be amenable to DP-GD.

2 Preliminaries

Differential Privacy: Throughout the paper, we focus on approximate differential privacy [12, 11].

Definition 2.1 (Differential privacy [12, 11])**.** *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 present in one set, and absent in another), 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 $n = |D|$.

Generalized Linear Problems: For a major part of this paper, we focus on a special class of ERM problems called generalized linear problems [35], where the loss function $\ell(\theta; d)$ takes a special inner-product form $\ell(\langle \theta, \mathbf{x} \rangle; y)$. Here, $\mathbf{x} \in \mathbf{R}^p$ is usually denoted as a feature vector, and $y \in \mathbf{R}$ is the response. A data element d corresponds to a tuple $(\mathbf{x}, y)$. Instead of the original feature vector in the data, $\mathbf{x}$ can also be extended to represent a mapped value $\phi(\mathbf{x})$ of original feature vector. We do not make this distinction here.

A more comprehensive preliminaries is in Appendix A.

Differentially Private Gradient Descent: Now, we provide a formal version of Differentially Private Gradient Descent (DP-GD) (Algorithm 1). The version mentioned here is the one where the gradient $\boldsymbol{g}_t$ is computed over the complete data set, and the final model $\theta^{\texttt{priv}}$ is an average of the models obtained so far. In practice, we may instead use DP stochastic gradient descent (DP-SGD), where $\boldsymbol{g}_t$ is computed over a random minibatch of the data, and the final model $\theta^{\texttt{priv}}$ 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 2.2 (From [1, 30])**.** *Differentially private gradient descent (Algorithm 1) is (ε, δ)-differentially private, if one sets the noise variance as $\sigma^2 = \frac{2L^2 T \log(1/\delta)}{(n\varepsilon)^2}$.*

Theorem 2.3 (From [7] and [39])**.** *If the constraint set $\mathcal{C}$ is convex, the loss function $\ell(\theta; d)$ is convex in the first parameter, $\|\nabla_\theta \ell(\theta; d)\| \leq B$ for all $\theta \in \mathcal{C}$ and $d \in D$, and the clipping norm $L \geq B$, then for objective function*

$\mathcal{L}(\theta; D) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; d_i)$, for appropriate choices of the learning rate and the number of iterations in differentially private gradient descent (Algorithm 1), we have

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

where $\theta^* = \arg\min_{\theta\in\mathcal{C}} \mathcal{L}(\theta; D)$ is the optimizer of $\mathcal{L}$ and $\theta_0 \in \mathcal{C}$ is the initialization of θ in DP-GD. The corresponding high-probability version is as follows: with probability at least $1-\beta$,

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

3 Negative Effects of Gradient Clipping

In this section, we demonstrate a two-fold impact of clipping for convex generalized linear problems. In Section 3.1, we show that clipping can increase the excess empirical risk from $O(1/n)$ to $\Omega(1)$, even for fairly simple tasks like binary logistic regression. Next, Section 3.2 shows that for more complex tasks like multiclass softmax regression, clipping can completely destroy the underlying convexity property. We show that DP-GD with gradient clipping does not even correspond to optimizing any objective function which is differentiable almost everywhere. This class, in particular, includes any convex function.

3.1 Aggressive Clipping Increases Excess Empirical Risk

We highlight the importance of choosing an appropriate clipping strategy by computing the explicit error that clipping introduces in the excess empirical risk. We first provide an analytical tool (Lemma 3.1) to precisely quantify the objective function DP-GD optimizes when the underlying loss function is a generalized linear problem. This is a fairly natural problem class including linear and logistic regression. Using Lemma 3.1, we then construct a lower bound for logistic regression that quantifies the bias introduced by clipping.

Let $\partial f(y)$ denote the subdifferential of f at y and $\partial_\theta \ell(\theta'; \mathbf{x}')$ denote the partial subdifferential of ℓ with respect to θ at $(\theta, \mathbf{x}) = (\theta', \mathbf{x}')$.

Lemma 3.1. *Let $f : \mathbf{R} \to \mathbf{R}$ be any convex function and $L \in \mathbf{R}_+$ be any positive value. For any $\mathbf{x} \neq \mathbf{0}$, let $Y_1 = \left\{y : u < -\frac{L}{\|\mathbf{x}\|_2}\ \forall u \in \partial f(y)\right\}$ and $Y_2 = \left\{y : u > \frac{L}{\|\mathbf{x}\|_2}\ \forall u \in \partial f(y)\right\}$. If Y_1 is non-empty, let $y_1 = \sup Y_1$; otherwise, let $y_1 = -\infty$. If Y_2 is non-empty, let $y_2 = \inf Y_2$; otherwise, let $y_2 = \infty$. Let $g_{\mathbf{x}} : \mathbf{R} \to \mathbf{R}$ be*

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

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*

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

Note: Lemma 3.1 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 [7, 16, 15, 5] to DP-GD on convex generalized linear problems. We use it for both Theorem 3.2 and Theorem 4.1.

The proof is based on the fact that clipping does not affect the monotonicity property of the derivative of one-dimensional convex function. It can be found at Appendix B.

Using Lemma 3.1, in Theorem 3.2, we show that running DP-GD with aggressive clipping can result in a constant excess empirical risk for logistic regression, in contrast to the best achievable excess empirical risk of $O\left(1/n\right)$. The proof can be found in Appendix B.2.

For a dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $x_i \in \mathbf{R}^p$ is the feature and $y_i \in \{+1, -1\}$ is the label, and for a convex set $\mathcal{C}$, logistic regression is defined as solving for $\theta^* := \arg\min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D)$ where $\mathcal{L}(\theta; D) = \frac{1}{n}\sum_{i=1}^n \ell(\theta; (x_i, y_i))$ with $\ell(\theta; (x, y)) = \log\left(1 + e^{-y\langle\theta, x\rangle}\right)$.

Theorem 3.2. *Consider the objective function $\mathcal{L}(\theta, D)$ for logistic regression as defined above. Let θ^{priv} be the output of DP-GD on $\mathcal{L}(\theta, D)$ with clipping norm L. For any $L < 1/4$, there exists a positive integer $n_0(L)$ such that for any $n \geq n_0(L)$, there exists a dataset $D = \{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \{x \in \mathbf{R}^p : \|x\|_2 \leq 1\}$ and $y_i \in \{+1, -1\}$, such that*

$$\mathbb{E}\left[\mathcal{L}(\theta^{priv}; D)\right] - \min_{\theta \in \mathbf{R}^p} \mathcal{L}(\theta; D) = \Omega\left(\log(1/L)\right).$$

Note: The lower bound construction does not require constraining $\mathcal{C}$. With $\mathcal{C}$ being the whole space $\mathbf{R}^p$, the optimization problem considered here is unconstrained. Also, notice that $L < 1/4$ is not a strong requirement, as for any (x_i, y_i), the gradient of logistic loss is upper bounded by $\|x_i\|_2 \leq 1$. So, $L = 1$ is already equivalent to no clipping.

We can show a similar lower bound for linear regression where the objective function is $\mathcal{L}(\theta; D) = \frac{1}{n}\sum_{i=1}^n \left(y_i - \langle\theta, x_i\rangle\right)^2$ and y_i and x_i are bounded. The proof follows the same strategy as in the lower bound for logistic regression. However, to get an upper bound B on the gradient norm for Theorem 2.3, the construction needs $\mathcal{C}$ to have bounded radius. The details are in Appendix B.3.

For both logistic regression and the linear regression, it is obvious that if we set the clipping norm L to be higher than the upper bound of the gradient norm, which exists in both cases, then we can still get $\tilde{O}(1/n)$ excess empirical risk. Therefore, we can conclude that picking a proper L is critical in convex optimization problems.

3.2 Clipped Multiclass Softmax Regression Doesn't Correspond to any "Natural" Function

We have shown that for any loss function ℓ that is convex in $\langle\theta, \mathbf{x}\rangle$, optimizing ℓ with DP-GD is equivalent to optimizing another convex function, though they may have different minimizers. Does the same apply to other common loss functions, such as the cross-entropy loss for softmax regression with more than two classes? The answer is it might not. In this section, we will show that for softmax regression, there does not exist any function whose subgradient is the clipped gradient of the cross-entropy loss as long as the function is required to be differentiable almost everywhere, which includes any convex function.

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 $\theta = [\theta^{(1)}, \ldots, \theta^{(K)}]$,

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

We then have the gradient of ℓ as

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

and the clipped gradient as $G(\theta) := \min\left(1, \frac{L}{\|\nabla_\theta(\theta;(x,y))\|_2}\right) \cdot \nabla_\theta\left(\theta; (x, y)\right)$ for any $\theta \in \mathbf{R}^{p \times K}$ where $\nabla_\theta\left(\theta; (x, y)\right) = \left[\nabla_{\theta^{(1)}}\left(\theta; (x, y)\right), \ldots, \nabla_{\theta^{(K)}}\left(\theta; (x, y)\right)\right]$.

Theorem 3.3. *Consider any sample (x, y) with $x \in \mathbf{R}^p \backslash \{\mathbf{0}\}$, $y \in [K]$ (for $K \geq 3$) and any $L > 0$ such that $\Theta = \{\theta : \|\nabla_\theta \ell(\theta; (x, y))\|_2 > L\}$ 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 $\mathcal{C}_N \subseteq \mathcal{C}$ such that $\mathcal{C}_N$ 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 differentiable almost everywhere [34, Theorem 25.5], if f is convex, we only need to require $\mathcal{C}_N$ to be a closed set. Notice that the ℓ_1 regularizer $\|\theta\|_1$ is only non-differentiable on a closed set, and the hinge loss, $\ell_{\mathsf{hinge}}(\theta; (x, y)) = \max(0, 1 - y\langle\theta, x\rangle)$, is also non-differentiable on a closed set. Theorem 3.3 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 or ℓ_2 regularization, and in non-convex models like neural networks with either smooth or non-smooth activation functions. The proof of Theorem 3.3 is in Appendix B.4.

One might ask if the problem could be resolved by "per-class" clipping, i.e., clipping $\nabla_{\theta^{(k)}} \ell$ individually for each k? The answer is negative. We provide more details in Appendix B.4.1.

4 Convergence of Clipped DP-GD on Generalized Linear Problems

In Section 4.1, we provide the first convergence guarantee for DP-GD with clipped gradients. We show that there exist a well defined ERM problem which clipped DP-GD optimizes when operating on convex generalized linear problems. Furthermore, if the clipping norm is L, the loss function $\ell(\cdot;\cdot)$ is ℓ_2-Lipschitz bounded by parameter $B \leq L$, then clipped DP-GD optimizes the original ERM problem. The convergence is independent of the number of model parameters p.

In Section 4.2, we extend this guarantee to *non-convex* generalized linear problems with smooth losses, and show that DP-GD approximately converges to a first-order-stationary point (FOSP) with similar dimension-independent guarantee, as long as $B \leq L$. This is the first dimension independent convergence guarantee for any non-convex differentially private learning task.

4.1 Excess Empirical Risk Guarantees of Clipped DP-GD

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 $\mathbf{x}_i \in \mathbf{R}^p$ and $y_i \in \mathbf{R}$ for all $i \in [n]$. Assume the loss function $\ell(\langle\theta, \mathbf{x}\rangle; y)$ is convex in its first parameter and is B-Lipschitz (w.r.t. the ℓ_2-norm) over all $\theta \in \mathbf{R}^p$ and for all $\mathbf{x}$ and y. The objective is to output $\theta^{\mathtt{priv}}$ that approximately solves $\arg\min\limits_{\theta \in \mathbf{R}^p} \mathcal{L}(\theta; D)$ while satisfying DP. Now, we show a utility/privacy trade-off for DP-GD with clipped gradients.

From Lemma 3.1 we know that for a given clipping norm L, DP-GD optimizes $\mathcal{L}^{(L)}_{\mathsf{clipped}}(\theta; D) = \frac{1}{n} \sum\limits_{i=1}^{n} \ell^{(L)}_{\mathsf{clipped}}(\langle\theta, \mathbf{x}_i\rangle; y_i)$, where $\ell^{(L)}_{\mathsf{clipped}}(\langle\theta, \mathbf{x}_i\rangle; y_i)$ can be obtained from Lemma 3.1. We show the following. The proof can be found in Appendix C.

Theorem 4.1. *Let $\theta_0 = \mathbf{0}^p$ be the initial point of DP-GD. Let $\theta^* = \arg\min\limits_{\theta \in \mathbf{R}^p} \mathcal{L}^{(L)}_{\mathsf{clipped}}(\theta; D)$, and M be the projector to the eigenspace of the matrix $\sum\limits_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$. Setting the constraint set $\mathcal{C} = \mathbf{R}^p$, clipping norm L, and running DP-GD on $\mathcal{L}(\theta; D)$ for $T = n^2\varepsilon^2$ steps with appropriate learning rate η, we get:*

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

In particular, if B is the Lipscthiz constant for the loss function $\ell(\cdot;\cdot)$ and $L \geq B$, then $\mathcal{L}^{(L)}_{\mathsf{clipped}}(\theta; D) = \mathcal{L}(\theta; D)$, i.e., there is no effect of clipping.

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

We present our result for excess empirical risk, but it can be translated to excess population risk guarantees via standard stability-based arguments [7, 22]. The crux of our proof technique in Theorem 4.1 is to work in the subspace generated by the feature vectors for generalized linear problem. We proved the guarantees only for DP-GD that returns the average of the models generated during training. Our proof would extend seamlessly (by modifying the proofs of Theorems 1 and 2 in [36]) to settings where the updates are over stochastic gradients computed over mini-batches of the data.

The lower-bound in [7] shows that if one performs constrained optimization with differential privacy, 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 perform *unconstrained optimization*, 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. It is an open question whether an analogous result as in Theorem 4.1 is possible for general unconstrained convex optimization. Furthermore, it will be interesting to see if the dependence of $O(1/\sqrt{n})$ in Theorem 4.1 can be reduced to $O(1/n)$, or it is tight. We leave this problem for future work.

The guarantee for the case $L \leq B$ is of the same flavor as in [22], wherein such a result was shown for two different differentially private algorithms, namely, *output perturbation*, and *objective perturbation* [9, 25] for *convex* optimization problems. Our result for DP-GD improves the state-of-the-art in the following ways. First, unlike output perturbation and objective perturbation, DP-GD *does not* require convexity to ensure differential privacy. As a result, DP-GD can be applied to non-convex losses and enjoys the same dimension-independence behavior as in the convex case. Second, our results for DP-GD almost seamlessly transfer to the local differential privacy (LDP) setting (see Definition D.1). This is the *first* dimension-independent excess risk guarantee in the LDP setting. Output perturbation and objective perturbation are fundamentally incompatible with LDP, as they require a centralized dataset to operate. This result is detailed in Appendix D.

4.2 Reaching Approximate Stationary Points for Non-convex Generalized Linear Problems

In this section, we provide an extension to Theorem 4.1 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 [28], Tangent loss [27], and tempered loss [3]. We show that as long as $\ell(z;\cdot)$ is β-smooth (see Definition A.3), DP-GD approximately reaches a stationary point on the objective function $\mathcal{L}(\theta; D)$, where θ is called a stationary point if $\nabla\mathcal{L}(\theta; D) = 0$. As in all the above-stated results in this section, the convergence guarantee will have no explicit dependence on the number of dimensions. We use a folklore argument stated in [2] to prove our result.

Theorem 4.2. *Recall the notation in Theorem 4.1. Let* $t_{\mathsf{priv}} \leftarrow \underset{0\leq t\leq T-1}{\arg\min} \left\|\nabla\mathcal{L}(\theta_t; D)\right\|_M + \mathsf{Lap}\left(\frac{4L}{n}\right)$. *Then, the algorithm that outputs* $\theta_0, \ldots, \theta_T$ *in conjunction with* t_{priv} *is* $(2\varepsilon, \delta)$*-differentially private. Furthermore, as long as* $T \geq \frac{\beta n^2 \varepsilon^2 \cdot \mathcal{L}(\mathbf{0}^p; D)}{2L^2 \log\left(\frac{1}{\delta}\right)}$ *and* $L \geq B$, *we have with probability at least* $1-\gamma$,

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

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

Theorem 4.2 does not immediately imply converging 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 [23]. However, that would require an additional assumption beyond smoothness, i.e., *Lipschitz continuity* of the Hessian. The proof of Theorem 4.2 is in Appendix C.

Acknowledgements

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

References

[1] 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.

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

[3] 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.

[4] K. Amin, A. Kulesza, A. Munoz, and S. Vassilvtiskii. Bounding user contributions: A bias-variance trade-off in differential privacy. In *International Conference on Machine Learning*, pages 263–271, 2019.

[5] 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, 2019.

[6] 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, 2019.

[7] 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.

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

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

[10] 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.

[11] 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, 2006.

[12] 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, 2006.

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

[14] 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.

[15] 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.

[16] 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.

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

[18] A. Gupta, K. Ligett, F. McSherry, A. Roth, and K. Talwar. Differentially private combinatorial optimization. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 1106–1125. SIAM, 2010.

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

[20] P. J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley New York, 1981.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] J. Liu and K. Talwar. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 298–309, 2019.

[27] 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.

[28] 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.

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

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

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

[32] 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.

[33] 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.

[34] R. T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1970.

[35] 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.

[36] 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.

[37] A. Smith, A. Thakurta, and J. Upadhyay. Is interaction necessary for distributed private learning? In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 58–77. IEEE, 2017.

[38] 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.

[39] 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.

[40] O. Thakkar, G. Andrew, and H. B. McMahan. Differentially private learning with adaptive clipping. *CoRR*, abs/1905.03871, 2019.

[41] A. Thakurta. Beyond worst case sensitivity in private data analysis. In *Encyclopedia of Algorithms*, pages 192–199. 2016.

[42] 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.

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

[44] 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.

[45] 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.

A Background on Convex Learning

Empirical risk minimization (ERM): Let $D = \{d_1, \cdots, d_n\} \subseteq \mathcal{D}$ be a data set of n-samples drawn from the domain $\mathcal{D}$, and let $\ell : \mathcal{C} \times \mathcal{D} \to \mathbf{R}$ be a loss function with $\mathcal{C} \subseteq \mathbf{R}^p$. Then the empirical risk over the data set D is defined as $\mathcal{L}(\theta; D) = \frac{1}{n} \sum\limits_{i=1}^{n} \ell(\theta; d_i)$. 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*. By stability-based arguments [7, 35], one can easily translate excess empirical risk for differentially private algorithms to their corresponding *excess population risk*, which can be defined for a model $\theta \in \mathcal{C}$ as $\mathbb{E}_{d \sim \mathcal{T}} [\ell(\theta; d)]$ where $\mathcal{T}$ is a given distribution over $\mathcal{D}$.

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 A.1 (ℓ_2-Lipschitz continuity)**.** *A function $f : \mathcal{C} \to \mathbf{R}$ is B-Lipschitz w.r.t. the ℓ_2-norm over a set $\mathcal{C} \subseteq \mathbf{R}^p$ if the following holds: $\forall \theta_1, \theta_2 \in \mathcal{C}, |f(\theta_1) - f(\theta_2)| \leq B \cdot \|\theta_1 - \theta_2\|_2$.*

Definition A.2 ((Strong) convexity w.r.t. ℓ_2-norm)**.** *A function $f : \mathcal{C} \to \mathbf{R}$ 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 A.3 (Smoothness)**.** *A function $f : \mathcal{C} \to \mathbf{R}$ 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*

$$f(\theta_2) \leq f(\theta_1) + \langle \nabla f(\theta_1), \theta_2 - \theta_1 \rangle + \frac{\beta}{2} \|\theta_1 - \theta_2\|_2^2.$$

Definition A.4 (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**: $\rho(a \cdot \boldsymbol{u}) = |a| \cdot \rho(\boldsymbol{u})$.

B Omitted Proof from Section 3

B.1 Generic Tool for Understanding Clipping

We first define some notations.

- For any vector v and positive scalar I, let $[v]_I$ denote $\min\left\{\frac{I}{\|v\|_2}, 1\right\} \cdot x$, i.e., x projected onto the ℓ_2-ball of radius I. If v is a scalar, then $[v]_I = \max\{\min\{v, I\}, -I\}$. Also, for scalar, we use $[v]_{I^+}$ to denote $\min\{v, I\}$, and $[v]_{I^-}$ to denote $\max\{v, -I\}$.
- For a set S of scalar or vector, let $[S]_I$ denote $\{[v]_I : v \in S\}$. For a set S of scalar, let $[S]_{I^+} = \{[v]_{I^+} : v \in S\}$ and $[S]_{I^-} = \{[v]_{I^-} : v \in S\}$.
- For a set S of scalar, we write $S > I$ if $\forall u \in S$, $u > I$; similar for $<, \geq$ and $\leq$.

Proof of Lemma 3.1. We consider any fixed $\mathbf{x}$, and for simplicity we use g to denote $g_{\mathbf{x}}$. We first show g is convex and $\partial g(y) = [\partial f(y)]_{L'}$ for $y \in \mathbf{R}$ using the following claims. And then apply that to ℓ_f and ℓ_g to prove the theorem.

Claim B.1. *The following holds.*

1. *If $y_1 \in \mathbf{R}$, then $-L' \in \partial f(y_1)$, and thus $\partial f(y) \geq -L'$ for all $y > y_1$, $\partial f(y) \leq -L'$ for all $y < y_1$. If $y_2 \in \mathbf{R}$, then $L' \in \partial f(y_2)$, and thus $\partial f(y) \leq L'$ for all $y < y_2$, $\partial f(y) \geq L'$ for all $y > y_2$.*
2. *If $y_1 = -\infty$, then $\partial f(y) \geq -L'$ for all $y \in \mathbf{R}$. If $y_2 = \infty$, then $\partial f(y) \leq L'$ for all $y \in \mathbf{R}$.*
3. *If $y_1 = \infty$, then $\partial f(y) < -L'$ for all $y \in \mathbf{R}$, and thus $y_2 = \infty$. If $y_2 = -\infty$, then $\partial f(y) > L'$ for all $y \in \mathbf{R}$, and thus $y_1 = -\infty$.*

Proof. We consider the three cases separately.

1. By definition of y_2 and monotonicity of subdifferential, for $y > y_2$, we have $\partial f(y) > L'$, and for $y < y_2$, $\min \partial f(y) \leq L'$ (as subdifferential is closed).

   We can find a sequence $y^{(k)} \to y_0^+$, and a sequence $g^{(k)}$ with $g^{(k)} \in \partial f(y^{(k)})$. As subdifferential is monotone, we know $g^{(k)}$ is decreasing. Since $g^{(k)}$ is lower bounded by I, the sequence $g^{(k)}$ converges to a value $\geq L'$. Similarly, we can find a sequence $y^{(k')} \to y_2^-$, and a sequence $g^{(k')}$ with $g^{(k')} = \min \partial f(y^{(k')})$. Since $g^{(k')}$ is increasing and upper bounded by L', the sequence $g^{(k')}$ converges to a value $\leq L'$.

   Recall that subdifferential is continuous. So both $\lim_{k\to\infty} g^{(k)} \geq L'$ and $\lim_{k'\to\infty} g^{(k')} \leq L'$ are contained in $\partial f(y_0)$, and we have $L' \in \partial f(y_2)$ by convexity of subdifferential. Then by monotonicity, for any $y > y_2$, we have $\partial f(y) \geq L'$; for any $y < y_2$, we have $\partial f(y) \leq L'$. Similar argument can be applied to y_1.

2. If Y_1 is non-empty and $y_1 = \infty$, by monotonicity of subdifferential, we have $\partial f(y) < -L'$ for all $y \in \mathbf{R}$. Therefore, $Y_2 = \emptyset$ and we have $y_2 = \infty$. Similar holds for y_2.

3. If Y_1 is empty and $y_1 = -\infty$, then for any y, $\max \partial f(y) \geq -L'$ (as subdifferential is closed). By monotonicity of subdifferential, $\partial f(y) \geq -L'$ for all $y \in \mathbf{R}$.

□

By monotonicity of subdifferential and the definition of y_1 and y_2, we know that $y_1 \leq y_2$ always holds and thus g is well-defined. Let $\mathcal{C} = [y_1, y_2] \cap \mathbf{R}$.

Claim B.2. *We have that g is a convex function when $y_1 \neq \infty$ and $y_2 \neq -\infty$.*

Proof. By Claim B.1, for any $y \in (y_1, y_2)$, $-L' \leq \partial f(y) \leq L'$. Therefore, f is L'-Lipschitz on $[y_1, y_2] \cap \mathbf{R}$. It is obviously also convex on this set. Consider f restricted to $\mathcal{C}$. According to [7, Lemma 6.3], the Lipschitz extension of this function, $\hat{g} : \mathbf{R} \to \mathbf{R}$ with $\hat{g}(y) = \min_{y' \in \mathcal{C}} \{f(y') + L'|y - y'|\}$, is also convex and L'-Lipschitz.

Then we show $g = \hat{g}$. For any $y \in \mathcal{C}$, we have $f(y) \leq f(y') + L'|y - y'|$ by Lipschizness; so $\hat{g}(y) = f(y) = g(y)$ on $\mathcal{C}$. If $y_1 \neq -\infty$, for any $y < y_1$ and any $y' \in \mathcal{C}$, we have $f(y_1) - f(y') \leq L'(y' - y_1)$ by Lipschitzness and $y' \geq y_1$. This translates to $f(y_1) - L'(y - y_1) \leq f(y') - L'(y - y')$, and thus $\hat{g}(y) = f(y_1) - L'(y - y_1) = g(y)$ for $y < y_1$. Similar holds for $y > y_2$ when $y_2 \neq \infty$. □

Claim B.3. *We have $\partial g(y) = [\partial f(y)]_{L'}$ for $y \in \mathbf{R}$.*

Proof. If $y_1 = \infty$, then g is a linear function with coefficient $-L'$, and is obviously convex. By Claim B.1, we have $\partial f(y) < -L'$ on $\mathbf{R}$, and thus $[\partial f(y)]_{L'} = \{-L'\} = \partial g(y)$ for all $y \in \mathbf{R}$. Similar holds for $y_2 = -\infty$.

Now we consider the case where $y_1 \neq \infty$ and $y_2 \neq -\infty$.

- On $(-\infty, y_1)$, g is linear and thus differentiable, so $\partial g(y) = \{-L'\}$. Also, we know from Claim B.1 that $\partial f(y) \leq -L'$ for $y \in (-\infty, y_1)$; so $[\partial f(y)]_{L'} = \{-L'\} = \partial g(y)$. Similar holds for (y_2, ∞).
- On (y_1, y_2), we have $g = f$ and both are convex. Any convex function f on an open subset of $\mathbf{R}$ is semi-differentiable and the subdifferential at point y is of the form $[\partial f_-(y), \partial f_+(y)]$ where $\partial f_-(y)$ is the left derivative and $\partial f_+(y)$ is the right derivative. Therefore, as the left and right derivative of f and g are the same in (y_1, y_2), we have $\partial g(y) = \partial f(y)$. By Claim B.1, $-L' \leq \partial f(y) \leq L'$ on this range, we have $\partial g(y) = [\partial f(y)]_{L'}$.

- At y_1 (if finite), the left derivative is $\partial g_-(y) = -L'$, and the right derivative $\partial g_+(y)$ is $\partial f_+(y)$ if $y_2 > y_1$ and is L' if $y_2 = y_1$. For $y_2 > y_1$, as $-L' \in \partial f(y_1)$, we have $\partial g(y) = [\partial f(y)]_{L'}$. For $y_2 = y_1$, we have $[-L', L'] \subseteq \partial f(y_1)$ and thus $\partial g(y) = [\partial f(y)]_{L'}$. Similar holds for y_2.

□

Then we consider ℓ_g. For a set U of scalar, we use $U \cdot \mathbf{x}$ to denote $\{u\mathbf{x} : u \in U\}$. We have $[U \cdot \mathbf{x}]_L = \{[u\mathbf{x}]_L : u \in U\} = \left\{\min\left\{\frac{L}{\|\mathbf{x}\|_2 u}, 1\right\} u\mathbf{x} : u \in U\right\} = [U]_{L/\|\mathbf{x}\|_2} \cdot \mathbf{x}$. Therefore, $\partial_\theta \ell_g(\theta; \mathbf{x}) = \partial g(\langle \theta, \mathbf{x}\rangle) \cdot \mathbf{x} = [\partial f(\langle \theta, \mathbf{x}\rangle)]_{L/\|\mathbf{x}\|_2} \cdot \mathbf{x} = [\partial f(\langle \theta, \mathbf{x}\rangle) \cdot \mathbf{x}]_L = [\partial_\theta \ell_f(\langle \theta, \mathbf{x}\rangle)]_L$, which completes the proof. □

B.2 Lower Bound for Binary Logistic Regression

Proof of Theorem 3.2. Since $\ell(\theta; (x, y))$ is convex in $\langle \theta, yx\rangle$, as have been shown Lemma 3.1 (with $\mathbf{x}$ there being yx), for any (x, y), there exists another function $\ell_g(\theta; (x, y))$ that is convex in $\langle \theta, yx\rangle$ and $\nabla_\theta \ell_g(\theta; (x, y)) = [\nabla_\theta \ell(\theta; (x, y))]_L$ for any θ. Let $\mathcal{L}_{\texttt{huber}}(\theta; D) = \frac{1}{n}\sum_{i=1}^n \ell_g(\theta; (x_i, y_i))$, which is also convex. For some convex set $\mathcal{C}$, let $\theta^{\texttt{huber}} := \arg\min_{\theta \in \mathcal{C}} \mathcal{L}_{\texttt{huber}}(\theta; D)$ and $\theta^* := \arg\min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D)$. Let $\theta^{\texttt{priv}}$ be the output of DPSGD on objective function $\mathcal{L}$.

To show a lower bound on $\mathbb{E}\left[\mathcal{L}(\theta^{\texttt{priv}}; D)\right] - \mathcal{L}(\theta^*; D)$, we would first show a lower bound on $\|\theta^* - \theta^{\texttt{huber}}\|_2$ and an upper bound on $\|\theta^{\texttt{priv}} - \theta^{\texttt{huber}}\|_2$, which together will give a lower bound on $\|\theta^* - \theta^{\texttt{huber}}\|_2$. Then, using strong convexity property of $\mathcal{L}$, we translate that to lower bound on $\mathcal{L}(\theta^{\texttt{priv}}; D) - \mathcal{L}(\theta^*; D)$.

It is enough to prove the result for dimension $p = 1$, as we can always set the other $p - 1$ dimensions to be 0. Let D be $\{(1/2, +1)\}^{2n} \cup \{(1, -1)\}^n$ and $\mathcal{C} = \mathbf{R}$.

We have

$$\begin{aligned}
&\ell(\theta; (1/2, +1)) = \log\left(1 + e^{-\theta/2}\right), && \text{and} \quad \ell(\theta; (1, -1)) = \log\left(1 + e^{\theta}\right)\\
\Rightarrow &\nabla_\theta \ell(\theta; (1/2, +1)) = -\frac{1/2}{1 + e^{\theta/2}}, && \text{and} \quad \nabla_\theta \ell(\theta; (1, -1)) = \frac{1}{1 + e^{-\theta}}\\
\Rightarrow &\nabla_\theta^2 \ell(\theta; (1/2, +1)) = \frac{1}{4}\frac{1}{(1 + e^{\theta/2})(1 + e^{-\theta/2})}, && \text{and} \quad \nabla_\theta^2 \ell(\theta; (1, -1)) = \frac{1}{(1 + e^{\theta})(1 + e^{-\theta})}
\end{aligned}$$

Given L, we have

$$\ell_g(\theta; (1/2, +1)) = \begin{cases} -L\theta + 2L\log\left(\frac{1}{2L} - 1\right) + \log\frac{1}{1-2L} & \text{for } \theta < 2\log\left(\frac{1}{2L} - 1\right)\\ \log\left(1 + e^{-\theta/2}\right) & \text{for } \theta \geq 2\log\left(\frac{1}{2L} - 1\right)\end{cases}$$

and

$$\ell_g(\theta; (1, -1)) = \begin{cases} \log\left(1 + e^{\theta}\right) & \text{for } \theta \leq -\log\left(\frac{1}{L} - 1\right)\\ L\theta + L\log\left(\frac{1}{L} - 1\right) + \log\frac{1}{1-L} & \text{for } \theta > -\log\left(\frac{1}{L} - 1\right)\end{cases}$$

as the loss function with gradient being $[\nabla_\theta \ell(\theta; (1/2, +1))]_L$ and $[\nabla_\theta \ell(\theta; (1, -1))]_L$.

We have $\theta^* = 0$ as

$$\begin{aligned}
\nabla_\theta \mathcal{L}(\theta; D) = 0 &\Leftrightarrow 2\nabla_\theta \ell(\theta; (1/2, +1)) + \nabla_\theta \ell(\theta; (1, -1)) = 0\\
&\Leftrightarrow -\frac{1}{1 + e^{\theta/2}} + \frac{1}{1 + e^{-\theta}} = 0 \Leftrightarrow \theta = 0.
\end{aligned}$$

As for $\theta^{\texttt{huber}}$, we have

$$\nabla_\theta \mathcal{L}_{\texttt{huber}}(\theta; D) = \begin{cases} \frac{2}{3}\frac{-1/2}{1+e^{\theta/2}} + \frac{L}{3} & \text{for } \theta \geq 2\log\left(\frac{1}{2L} - 1\right)\\ -\frac{2L}{3} + \frac{L}{3} & \text{for } \theta \in \left(-\log\left(\frac{1}{L} - 1\right), 2\log\left(\frac{1}{2L} - 1\right)\right)\\ -\frac{2L}{3} + \frac{1}{3}\frac{1}{1+e^{-\theta}} & \text{for } \theta \leq -\log\left(\frac{1}{L} - 1\right)\end{cases},$$

which is equal to 0 at $2\log\left(\frac{1}{L}-1\right)$. So we have $\theta^{\text{huber}} = 2\log\left(\frac{1}{L}-1\right)$.

So we have $\|\theta^* - \theta^{\text{huber}}\|_2 = 2\log\left(\frac{1}{L}-1\right)$.

Now we bound $\|\theta^{\text{huber}} - \theta^{\texttt{priv}}\|_2$. As for each x_i in D, it is given that $\|x_i\|_2 \le 1$, we have $B = 1$. As the initial guess θ_0 is 0, we have $\|\theta_0 - \theta^*\|_2 = 2\log\left(\frac{1}{L}-1\right)$. From Theorem 2.3, with probability $\ge 1-\delta$, $\mathcal{L}_{\texttt{huber}}(\theta^{\texttt{priv}}; D) - \mathcal{L}_{\texttt{huber}}(\theta^{\text{huber}}; D) \le C \cdot \frac{2\log(1/L-1)\sqrt{p\log(1/\delta)\log(1/\beta)}}{n\varepsilon}$ for some positive constant C. We set $\frac{n_0}{3} = \frac{96C\log(1/L-1)\sqrt{p\log(1/\delta)\log(1/\beta)}}{L\varepsilon} > \max\left(20\log\left(\frac{1}{L}-1\right), 96\right)\cdot\frac{C\sqrt{p\log(1/\delta)\log(1/\beta)}}{L\varepsilon}$. As $n \ge \frac{n_0}{3} > \frac{20C\log(1/L-1)\sqrt{p\log(1/\delta)\log(1/\beta)}}{L\varepsilon}$, we have $\mathcal{L}_{\texttt{huber}}(\theta^{\texttt{priv}}; D) - \mathcal{L}_{\texttt{huber}}(\theta^{\text{huber}}; D) < 0.1L$. We now translate this to an upper bound on $\|\theta^{\texttt{priv}} - \theta^{\text{huber}}\|_2$ (with high probability).

Let $\theta_1 = 2\log\left(\frac{1}{2L}-1\right)$ and $\theta_2 = 2\log\left(\frac{2}{L}-1\right)$. We now show $\theta^{\texttt{priv}} \in (\theta_1, \theta_2)$ with probability $\ge 1-\delta$. We know that for $\theta \ge \theta_1$, for $\mathsf{Const} = \frac{1}{3}\left(L\log\left(\frac{1}{L}-1\right) + \log\frac{1}{1-L}\right)$,

$$\mathcal{L}_{\texttt{huber}}(\theta; D) = \frac{1}{3}\left(2\log(1+e^{-\theta/2}) + L\theta\right) + \mathsf{Const},$$

and we thus have

$$\begin{aligned}
\mathcal{L}_{\texttt{huber}}(\theta^{\text{huber}}; D) &= \frac{1}{3}\left(2\log\frac{1}{1-L} + 2L\log\left(\frac{1}{L}-1\right)\right) + \mathsf{Const} \\
\mathcal{L}_{\texttt{huber}}(\theta_1; D) &= \frac{1}{3}\left(2\log\frac{1}{1-2L} + 2L\log\left(\frac{1}{2L}-1\right)\right) + \mathsf{Const} \\
\mathcal{L}_{\texttt{huber}}(\theta_2; D) &= \frac{1}{3}\left(2\log\frac{2}{2-L} + 2L\log\left(\frac{2}{L}-1\right)\right) + \mathsf{Const}.
\end{aligned}$$

So for $L < 1/4$,

$$\begin{aligned}
\mathcal{L}_{\texttt{huber}}(\theta_1; D) - \mathcal{L}_{\texttt{huber}}(\theta^{\text{huber}}; D) &= \frac{2}{3}\Bigg(\log\frac{1}{1-2L} + L\log\left(\frac{1}{2L}-1\right) \\
&\qquad - \log\frac{1}{1-L} - L\log\left(\frac{1}{L}-1\right)\Bigg) \\
&= \frac{2}{3}\left((1-L)\log\frac{1-L}{1-2L} - L\log 2\right) \\
&\ge \frac{2}{3}(1-\log 2)L > 0.2L > 0.1L. \\
\mathcal{L}_{\texttt{huber}}(\theta_2; D) - \mathcal{L}_{\texttt{huber}}(\theta^{\text{huber}}; D) &= \frac{2}{3}\Bigg(\log\frac{2}{2-L} + L\log\left(\frac{2}{L}-1\right) \\
&\qquad - \log\frac{1}{1-L} - L\log\left(\frac{1}{L}-1\right)\Bigg) \\
&= \frac{2}{3}\Bigg(\log\frac{2}{2-L} + L\log\left(\frac{2}{L}-1\right) \\
&\qquad - \log\frac{1}{1-L} - L\log\left(\frac{1}{L}-1\right)\Bigg) \\
&\ge \frac{2}{3}\left(\log(2) - \frac{1}{2}\right)L > 0.1L.
\end{aligned}$$

Notice that $\mathcal{L}_{\texttt{huber}}$ is convex, which means the derivative is monotone and the function is decreasing for $\theta < \theta^{\text{huber}}$ and increasing for $\theta > \theta^{\text{huber}}$. As $\theta_1 < \theta^{\texttt{priv}} < \theta_2$, if $\theta^{\texttt{priv}} \le \theta_1$ or $\theta^{\texttt{priv}} \ge \theta_2$, then $\mathcal{L}_{\texttt{huber}}(\theta^{\texttt{priv}}; D) - \mathcal{L}_{\texttt{huber}}(\theta^{\text{huber}}; D) \ge 0.1L$, which contradicts to the fact that $\mathcal{L}_{\texttt{huber}}(\theta^{\texttt{priv}}; D) - \mathcal{L}_{\texttt{huber}}(\theta^{\text{huber}}; D) < 0.1L$. Therefore, we can conclude that $\theta^{\texttt{priv}} \in (\theta_1, \theta_2)$ with probability $\ge 1-\delta$.

For $\theta \in (\theta_1, \theta_2)$, the 2nd order derivative of $\mathcal{L}_{\texttt{huber}}$ is $\frac{1}{6}\frac{1}{(1+e^{\theta/2})(1+e^{-\theta/2})} \geq \frac{1}{12}\frac{1}{1+e^{\theta/2}}$, which is decreasing and therefore $\geq \frac{L}{24}$. This means $\mathcal{L}_{\texttt{huber}}$ is $\frac{L}{24}$-strongly convex for θ in this range. Therefore, the bound on the difference of the loss translates to a bound on the ℓ_2 distance and we have $\|\theta^{\texttt{priv}} - \theta^{\texttt{huber}}\|_2^2 \leq C \cdot \frac{48\log(2/L-1)\sqrt{p\log(1/\delta)\log(1/\beta)}}{Ln\varepsilon}$. We then have

$$\begin{aligned}
\|\theta^{\texttt{priv}} - \theta^*\|_2 &\geq \|\theta^* - \theta^{\texttt{huber}}\|_2 - \|\theta^{\texttt{priv}} - \theta^{\texttt{huber}}\|_2 \\
&\geq 2\log\left(\frac{1}{L} - 1\right) - \sqrt{\frac{48C\log(2/L-1)\sqrt{p\log(1/\delta)\log(1/\beta)}}{Ln\varepsilon}} \\
&\geq \log\left(\frac{1}{L} - 1\right)
\end{aligned}$$

where the last inequality follows as for $n > \frac{96C}{L}\frac{\sqrt{p\log(1/\delta)\log(1/\beta)}}{\varepsilon}$, we have

$$\sqrt{\frac{48C\log(2/L-1)\sqrt{p\log(1/\delta)\log(1/\beta)}}{Ln\varepsilon}} \leq \log\left(\frac{1}{L} - 1\right) \text{ for any } L < 1/4.$$

Let $\theta_3 = -\log\left(\frac{1}{L} - 1\right)$ and $\theta_4 = \log\left(\frac{1}{L} - 1\right)$. As $\theta^* = 0$, the above inequality implies $\theta^{\texttt{priv}} \in (-\infty, \theta_3] \cup [\theta_4, \infty)$. Similarly, as $\mathcal{L}$ is convex with minimizer θ^*, we know $\mathcal{L}(\theta^{\texttt{priv}}; D) \geq \min\left(\mathcal{L}(\theta_3; D), \mathcal{L}(\theta_4; D)\right)$.

Since

$$\mathcal{L}(\theta; D) = \frac{1}{3}\left(2\log\left(1 + e^{-\theta/2}\right) + \log\left(1 + e^{\theta}\right)\right) = \frac{1}{3}\log\left(\left(1 + e^{\theta/2}\right)^2 + \left(1 + e^{-\theta/2}\right)^2\right)$$

is an even function, we have

$$\mathcal{L}(\theta_3; D) = \mathcal{L}(\theta_4; D) = \frac{1}{3}\log\left(\left(1 + \sqrt{\frac{1-L}{L}}\right)^2 + \left(1 + \sqrt{\frac{L}{1-L}}\right)^2\right) = \frac{2}{3}\log\left(\frac{1}{\sqrt{L}} + \frac{1}{\sqrt{1-L}}\right).$$

As $\mathcal{L}(\theta^*; D) = \log 2$, we have, for $L < 1/4$,

$$\begin{aligned}
\mathcal{L}(\theta^{\texttt{priv}}; D) - \mathcal{L}(\theta^*; D) &\geq \frac{2}{3}\log\left(\frac{1}{\sqrt{L}} + \frac{1}{\sqrt{1-L}}\right) - \log(2) \geq \frac{2}{3}\log\left(1 + \frac{1}{\sqrt{L}}\right) - \log(2) \\
&\geq \frac{2}{3}\log\left(1 + \frac{1}{\sqrt{L}}\right) - \frac{\log(2)}{\log(3)}\log\left(1 + \frac{1}{\sqrt{L}}\right) \geq \frac{1}{30}\log\left(1 + \frac{1}{\sqrt{L}}\right) \\
&\geq \frac{1}{60}\log\frac{1}{L}
\end{aligned}$$

This holds with probability $\geq 1 - \beta$, and we can convert it back to an expectation bound and have

$$\mathcal{L}(\theta^{\texttt{priv}}; D) - \mathcal{L}(\theta^*; D) = \Omega\left(\log\frac{1}{L}\right).$$

□

B.3 Lower Bound for Quadratic Loss and Linear Regression

Let $D = \{x_1, \ldots, x_n\}$ with $x_i \in \mathbf{R}$ and $\mathcal{L}(\theta; D) = \frac{1}{n}\sum_{i=1}^{n}(\theta - x_i)^2$. In Theorem B.4, we show that running DP-SGD with clipping can result in a constant excess empirical risk for this loss as well.

Theorem B.4. *Consider the objective function $\mathcal{L}(\theta, D)$ as defined above. Let θ^{priv} be the output of DP-SGD on $\mathcal{L}(\theta, D)$ with clipping norm L. There exists n, such that for $\varepsilon = \Theta(1)$ and $\delta = 1/n^{O(1)}$, for any L, there exists a dataset $D = \{x_i\}_{i=1}^{n}$ with $x_i \in \mathbf{R}$, such that*

$$\mathbb{E}\left[\mathcal{L}(\theta^{priv}; D)\right] - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D) = \Omega\left(L^2\right).$$

Proof. For any L, let $D = \{\underbrace{0, \ldots, 0}_{n \text{ of them}}, \underbrace{L, \ldots, L}_{n \text{ of them}}, \underbrace{8L, \ldots, 8L}_{n \text{ of them}}\}$ be a data set of $3n$ elements and let the constraint set $\mathcal{C} = [0, 8L]$. Let $\ell(\theta; x) = (\theta - x)^2$ and we have $\mathcal{L}(\theta; D) = \frac{1}{3n}\sum_{i=1}^{3n} \ell(\theta; x_i)$.

For any x_i, consider the Huberized version [20] of the loss functions $\ell(\cdot, x_i)$: $g_i(\theta) = \begin{cases} (\theta - x_i)^2 \text{ for } |\theta - x_i| \leq L/2 \\ L|\theta - x_i| - \frac{L^2}{4} \text{ otherwise} \end{cases}$. Let $\mathcal{L}_{\texttt{huber}}(\theta; D) = \frac{1}{3n}\sum_i g_i(\theta)$. By Lemma 3.1, running DP-SGD on $\mathcal{L}_{\texttt{huber}}$ without clipping is equivalent to the running DP-SGD on $\mathcal{L}$ with clipping norm L.

First, by equating the gradients to zero, we know that

$$\theta^* := \arg\min_{\theta \in [0, 8L]} \mathcal{L}(\theta; D) = 3L, \text{ and } \theta^{\text{huber}} := \arg\min_{\theta \in [0, 8L]} \mathcal{L}_{\texttt{huber}}(\theta; D) = L$$

Then, by Theorem 2.3 with $B = 8L$ and $\|\mathcal{C}\|_2 = 8L$, we know that with high probability,

$$\mathcal{L}_{\texttt{huber}}(\theta^{\texttt{priv}}; D) - \mathcal{L}_{\texttt{huber}}(\theta^{\text{huber}}; D) = O\left(\frac{L^2\sqrt{\log(1/\delta)}}{n\varepsilon}\right). \tag{1}$$

Notice that at $\theta \in [L/2, 3L/2]$, $\mathcal{L}_{\texttt{huber}}$ is 2/3-strongly convex. This fact, combined with (1), implies the following w.h.p. $\left\|\theta^{\texttt{priv}} - \theta^{\text{huber}}\right\|_2 = O\left(\frac{L\log^{1/4}(1/\delta)}{\sqrt{n\varepsilon}}\right)$. Hence, we can conclude that w.h.p. $\left\|\theta^{\texttt{priv}} - \theta^{\text{huber}}\right\|_2 = o(1)$.

Therefore, w.h.p. $\left\|\theta^{\texttt{priv}} - \theta^*\right\|_2 = 2L \pm o(1)$. Since $\mathcal{L}$ is 2-strongly convex everywhere, we finally conclude that $\mathbb{E}\left[\mathcal{L}(\theta^{\texttt{priv}}; D)\right] - \mathcal{L}(\theta^*; D) = \Omega(L^2)$. □

Because of Theorem B.4, one might wonder if it is at all possible to obtain $o(1)$ excess empirical risk on $\mathcal{L}$, with DP-SGD. A simple modification to the construction of $\mathcal{L}_{\texttt{huber}}$ would do the trick. Huberize the function as follows: $g_i(\theta) = \begin{cases} (\theta - x_i)^2, & |\theta - x_i| \leq 8L \\ 16L|\theta - x_i| - \frac{L^2}{4} & \text{o.w.} \end{cases}$. This construction ensures that $g_i(\theta)$ equals $(\theta - x_i)^2$ in the range $\theta \in [0, 8L]$ and is $16L$-Lipschitz. Hence, running the DP-SGD with $16L$-clipping norm, the constraint set $\mathcal{C} = [0, 8L]$, and the privacy parameters (ε, δ), is equivalent to running DP-SGD on $\mathcal{L}_{\texttt{huber}}$ with the g_i's while keeping the other parameters same. Furthermore, notice that $\mathcal{L}(\theta; D) = \mathcal{L}_{\texttt{huber}}(\theta; D)$ for all $\theta \in \mathcal{C}$, and $\arg\min_{\theta \in \mathcal{C}} \mathcal{L}_{\texttt{huber}}(\theta; D) = \theta^{\text{huber}} = 3L$. By the same argument as in the proof of Theorem B.4, one can conclude that w.h.p. $\left\|\theta^{\texttt{priv}} - \theta^{\text{huber}}\right\|_2 = O\left(\frac{L\log^{1/4}(1/\delta)}{\sqrt{n\varepsilon}}\right)$. Hence, for sufficiently large n one has $\theta^{\texttt{priv}} \in \mathcal{C}$. Combining these observations, we can conclude that with out modified clipping norm,

$$\mathbb{E}\left[\mathcal{L}(\theta^{\texttt{priv}}; D)\right] - \arg\min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D) = O\left(\frac{L^2\sqrt{\log(1/\delta)}}{n\varepsilon}\right). \tag{2}$$

Comparing Theorem B.4 and (2) we observe that clipping norm plays a critical role in the convergence of DP-SGD. In one case, the excess empirical risk is a constant, and in the other case it is $\widetilde{O}(1/n)$. Although, the above observation was for one-dimensional loss functions, it can easily be extended to linear regression in higher-dimensions by formulating the problem as follows: Each loss function ℓ_i is of the form $(y_i - \langle \mathbf{x}_i, \theta \rangle)^2$, where $y_i = x_i$ in the data set D and $\mathbf{x}_i$ has the first coordinate as one and rest all as zeros.

B.4 Clipped Softmax Regression Does Not Correspond to a "Natural" Function

Proof of Theorem 3.3. Without loss of generality, let $y = 1$. Let x be any non-zero vector in $\mathbf{R}^p$.

In the beginning, we state the formulas for the gradient, its norm and the clipped gradient. Let $E^{(k)}(\theta) = \exp\left(\theta^{(k')} \cdot x\right)$. (We omit (θ) when it is clear from the context.) Recall the gradient of the cross-entropy loss is $\nabla_{\theta^{(k)}}\left(\theta; (x, y)\right) = \left(\frac{\exp\left(\theta^{(k)} \cdot x\right)}{\sum_{k'=1}^K \exp\left(\theta^{(k')} \cdot x\right)} - \mathbb{1}\left(y = k\right)\right) \cdot x = \left(\frac{E^{(k)}}{\sum_{k'=1}^K E^{(k')}} - \mathbb{1}\left(y = k\right)\right) \cdot x$, so we have

$$\nabla_{\theta^{(1)}}\ell\left(\theta; (x, y)\right) = -\frac{\sum_{k=2}^K E^{(k)}}{\sum_{k'=1}^K E^{(k')}} \cdot x. \tag{3}$$

$$\text{For } k \geq 2,\ \nabla_{\theta^{(k)}}\ell\left(\theta; (x, y)\right) = \frac{E^{(k)}}{\sum_{k'=1}^K E^{(k')}} \cdot x. \tag{4}$$

The norm of the gradient $\nabla_\theta \ell(\theta; (x, y))$ is thus

$$\left\|\nabla_\theta\left(\theta; (x, y)\right)\right\|_2 = \sqrt{\sum_{k=1}^K \left\|\nabla_{\theta^{(k)}}\left(\theta; (x, y)\right)\right\|_2^2} = \|x\|_2 \cdot \frac{\sqrt{\left(\sum_{k=2}^K E^{(k)}\right)^2 + \sum_{k=2}^K \left(E^{(k)}\right)^2}}{\sum_{k'=1}^K E^{(k')}}, \tag{5}$$

which takes value in $\left(0, \sqrt{\frac{K}{K-1}}\|x\|_2\right)$. Recall that $\Theta = \{\theta : \|\nabla_\theta\left(\theta; (x, y)\right)\|_2 > L\}$.

Recall $G(\theta)$ is the clipped gradient. We also define, for $k \in [K]$, for $\theta \in \mathbf{R}^{p\times K}$,

$$G^{(k)}(\theta) := \min\left(1, \frac{L}{\|\nabla_\theta\left(\theta; (x, y)\right)\|_2}\right) \cdot \nabla_{\theta^{(k)}}\left(\theta; (x, y)\right),$$

so $G(\theta) = \left[G^{(1)}(\theta), \ldots, G^{(K)}(\theta)\right]$.

When $\theta \in \Theta$, we have $G^{(k)}(\theta) = L \cdot \frac{\nabla_{\theta^{(k)}}\ell(\theta; (x,y))}{\|\nabla_\theta(\theta; (x,y))\|_2}$, and thus

$$G^{(1)}(\theta) = -\frac{L}{\|x\|_2}\frac{\sum_{k=2}^K E^{(k)}}{\sqrt{\left(\sum_{k=2}^K E^{(k)}\right)^2 + \sum_{k=2}^K \left(E^{(k)}\right)^2}} \cdot x,$$

$$\text{For } k \geq 2,\ G^{(k)}(\theta) = \frac{L}{\|x\|_2}\frac{E^{(k)}}{\sqrt{\left(\sum_{k=2}^K E^{(k)}\right)^2 + \sum_{k=2}^K \left(E^{(k)}\right)^2}} \cdot x. \tag{6}$$

Notice that for any $k \geq 2$, $\nabla_{\theta^{(1)}} G^{(k)}$ is zero as $G^{(k)}$ does not depend on $\theta^{(1)}$; however, $\nabla_{\theta^{(k')}} G^{(1)}$ may not be zero everywhere as $G^{(1)}$ does not depend on $\theta^{(k')}$ (we will prove this formally).

We will prove the theorem by contradiction. Suppose there exists a function $f : \mathcal{C} \to \mathbf{R}$ such that 1). $\Theta \cap \mathcal{C}^o$ is a non-empty set, 2). f is differentiable except for a set $\mathcal{C}_N$ which is closed on $\mathcal{C}$ and has zero measure, and 3) $G(\theta)$ is a subgradient of f. We will show that on an open subset of $\Theta \cap \mathcal{C}$, f is differentiable but the 2nd derivative is not symmetric, which contradicts the fact that any function with continuous second order partial derivative should have symmetry of 2nd derivative in the interior of its domain.

We use Euclidean topology throughout the proof. When not specified, we talk about Euclidean topology in the space $\mathbf{R}^{p\times K}$. We consider Lebesgue measure on $\mathbf{R}^{p\times K}$ throughout the proof.

1. First, we show Θ is a non-empty open set in $\mathbf{R}^{p\times K}$.

   Recall the formula for $\left\|\nabla_\theta\left(\theta; (x, y)\right)\right\|_2$ in (5), which is obviously a continuous function in $\mathbf{R}^{p\times K}$. Therefore, the preimage of open set (L, ∞) through $\left\|\nabla_\theta\left(\theta; (x, y)\right)\right\|_2$, which is exactly Θ, is an open set in $\mathbf{R}^{p\times K}$. By assumption, Θ is non-empty.

2. Second, let $\Theta_G = \Theta \cap \left\{\theta : \forall k, k',\ \nabla_{\theta^{(k')}} G^{(k)}(\theta) = \nabla_{\theta^{(k)}} G^{(k')}(\theta)\right\}$ be the "good" subset of Θ where the 2nd derivative of f is symmetric if G is the derivative of f. We will show that Θ_G is a closed set in Θ and has Lebesgue measure 0.

   Recall $G^{(k)}(\theta)$ for $\theta \in \Theta$ in (6). For any $k \geq 2$, notice that $G^{(k)}(\theta)$ does not depend on $\theta^{(1)}$, so $\nabla_{\theta^{(1)}} G^{(k)}(\theta) = \mathbf{0}$ for $k \geq 2$.

   Now we look at the derivatives of $G^{(1)}$. Let $D(\theta) = \left(\sum_{k=2}^{K} E^{(k)}\right)^2 + \sum_{k=2}^{K} \left(E^{(k)}\right)^2$. For any $k' \geq 2$,

$$\nabla_{\theta^{(k')}} G^{(1)}(\theta) = -\frac{L E^{(k')}}{\|x\|_2 \left(D(\theta)\right)^{3/2}} \sum_{k=2}^{K} E^{(k)} \left(E^{(k)} - E^{(k')}\right) x x^\top .$$

   As $E^{(k)} > 0$ and $x \neq \mathbf{0}$, we have

$$\begin{aligned}\forall k' \geq 2,\ \nabla_{\theta^{(k')}} G^{(1)}(\theta) = \mathbf{0} &\Leftrightarrow E^{(2)} = \cdots = E^{(K)} \\ &\Leftrightarrow \langle \theta^{(2)}, x\rangle = \cdots = \langle \theta^{(K)}, x\rangle .\end{aligned}$$

   It is also not hard to check that $E^{(2)} = \cdots = E^{(K)}$ is sufficient to guarantee $\nabla_{\theta^{(k')}} G^{(k)}(\theta) = \nabla_{\theta^{(k)}} G^{(k')}(\theta)$ for any $k, k' \geq 2$.

   Therefore, we have $\Theta_G = \left\{\theta \in \Theta : \langle \theta^{(2)}, x\rangle = \cdots = \langle \theta^{(K)}, x\rangle\right\}$. We can define a function a on Θ with $a(\theta) = \sum_{k=3}^{K} |\langle \theta^{(2)}, x\rangle - \langle \theta^{(k)}, x\rangle|$. Since a is continuous on domain Θ, the preimage of the closed set $\{0\}$ through a, which is exactly Θ_G, is a closed set in Θ. Also, Θ_G is obviously a lower dimensional subspace of $\mathbf{R}^{p\times K}$ and thus has measure 0.

3. Third, let the "bad" set be $\Theta_B = \Theta \backslash \Theta_G$. We will show $\Theta_B \cap \mathcal{C}^{\mathrm{o}}$ is a non-empty set and is open on $\mathcal{C}^{\mathrm{o}}$.

   As Θ_G is closed in Θ, Θ_B, its complement, is an open set in Θ. As Θ is open in $\mathbf{R}^{p\times K}$, Θ_B is also open in $\mathbf{R}^{p\times K}$ (since Θ_B is the intersection of two open subsets in $\mathbf{R}^{p\times K}$). So $\Theta_B \cap \mathcal{C}^{\mathrm{o}}$ is open on $\mathcal{C}^{\mathrm{o}}$.

   On the other hand, as Θ and $\mathcal{C}^{\mathrm{o}}$ are open, $\Theta \cap \mathcal{C}^{\mathrm{o}}$ is open. Additionally, by assumption, $\Theta \cap \mathcal{C}^{\mathrm{o}}$ is non-empty. So $\Theta \cap \mathcal{C}^{\mathrm{o}}$ has positive measure. Since Θ_G has measure 0, $\Theta_B \cap \mathcal{C}^{\mathrm{o}} = \Theta \cap \mathcal{C}^{\mathrm{o}} \backslash \Theta_G$ has positive measure and is thus non-empty.

4. Finally, recall that f is differentiable everywhere except for a closed set on $\mathcal{C}_N$ with measure zero. Obviously, $\mathcal{C}_N$ is also closed on $\mathcal{C}^{\mathrm{o}}$. Then f is differentiable on $\Theta'_B := \Theta_B \cap \mathcal{C}^{\mathrm{o}} \backslash \mathcal{C}_N$, which implies that G is the gradient of f on Θ'_B. Also, since $\forall k$, all partial derivatives of $G^{(k)}$ exists and is continuous, we know that f has continuous 2nd derivatives on Θ'_B.

   As $\Theta_B \cap \mathcal{C}^{\mathrm{o}}$ is open and $\mathcal{C}_N$ is closed on $\mathcal{C}^{\mathrm{o}}$, Θ'_B is open on $\mathcal{C}^{\mathrm{o}}$. Also, as $\mathcal{C}_N$ has zero measure, Θ'_B is non-empty.

   By Schwarz's theorem, for any function that has continuous second order partial derivatives, it has symmetry of 2nd derivative in the interior of its domain. So we are supposed to see $\nabla_{\theta^{(k)}} G^{(k')} = \nabla_{\theta^{(k')}} G^{(k)}$ for any pairs of k and k' on Θ'_B (since Θ'_B itself is non-empty and open in $\mathcal{C}^{\mathrm{o}}$). However, this does not hold by definition of Θ_B. We therefore have a contradiction and such f cannot exist.

$\square$

B.4.1 "Per-class" Clipping Does Not Resolve the Problem

Theorem B.5. *Consider any sample (x, y) with $x \in \mathbf{R}^p \backslash \{\mathbf{0}\}$, $y \in [K]$ (for $K \geq 3$) and any $L > 0$ such that $\Theta = \{\theta : \|\nabla_{\theta^{(k)}} \ell(\theta; (x, y))\|_2 > L \text{ for some } k \in [K]\}$ is non-empty. Let $G(\theta)$ be the "per-class" clipped gradient of $\ell(\theta; (x, y))$. Consider any function $f : \mathcal{C} \to \mathbf{R}$, $\mathcal{C} \subseteq \mathbf{R}^{p\times K}$ such that $\Theta \subseteq \mathcal{C}$. If f is differentiable everywhere except for a set $\mathcal{C}_N \subseteq \mathcal{C}$ such that $\mathcal{C}_N$ 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 $\theta \in \mathcal{C}^{\mathrm{o}} \backslash \mathcal{C}_N$.*

Proof. Recall the formula for the gradient in (3) and the definition of $E^{(k)}$, and we have

$$\left\|\nabla_{\theta^{(1)}}\ell(\theta;(x,y))\right\|_2 = \frac{\sum_{k=2}^K E^{(k)}}{\sum_{k'=1}^K E^{(k')}}\left\|x\right\|_2,$$

$$\text{For } k \geq 2,\ \left\|\nabla_{\theta^{(k)}}\ell(\theta;(x,y))\right\|_2 = \frac{E^{(k)}}{\sum_{k'=1}^K E^{(k')}}\left\|x\right\|_2.$$

Obviously, $\left\|\nabla_{\theta^{(1)}}\ell(\theta;(x,y))\right\|_2 = \sum_{k=2}^K \left\|\nabla_{\theta^{(k)}}\ell(\theta;(x,y))\right\|_2$. So $\nabla_{\theta^{(k)}}\ell(\theta;(x,y))$ for $k \geq 2$ is clipped only when $\nabla_{\theta^{(1)}}\ell(\theta;(x,y))$ is clipped. We consider when some of them are clipped, i.e. the set Θ. There are two cases.

1. If all of them are clipped, then $G^{(1)} = -\frac{L}{\|x\|_2}x$ and $G^{(k)} = \frac{L}{\|x\|_2}x$ for $k \geq 2$. G is basically a constant and we have a valid gradient field.

2. If $\nabla_{\theta^{(1)}}\ell$ is clipped and some of $\nabla_{\theta^{(k)}}\ell$ for $k \geq 2$ is not clipped, then $G^{(1)} = -\frac{L}{\|x\|_2}x$ and $G^{(k)} = \frac{E^{(k)}}{\sum_{k'=1}^K E^{(k')}}x$. So we have $\nabla_{\theta^{(k)}}G^{(1)} = 0$ for any $k \geq 2$ and $\nabla_{\theta^{(1)}}G^{(k)} = -\frac{E^{(k)}E^{(1)}}{\left(\sum_{k'=1}^K E^{(k')}\right)^2}x^\top x$ which is always nonzero. So we do not have a valid gradient field when this happens.

   This is the set

$$\begin{aligned}\Theta_B &= \cup_{k_0=2}^K \Theta_{k_0} \\ \text{where } \Theta_{k_0} &= \left\{\theta : \frac{\sum_{k=2}^K E^{(k)}}{\sum_{k'=1}^K E^{(k')}}\left\|x\right\|_2 > L \text{ and } \frac{E^{(k_0)}}{\sum_{k'=1}^K E^{(k')}}\left\|x\right\|_2 \leq L\right\} \\ &\supseteq \left\{\theta : \frac{\sum_{k=2}^K E^{(k)}}{\sum_{k'=1}^K E^{(k')}}\left\|x\right\|_2 > L\right\} \cap \left\{\theta : \frac{E^{(k_0)}}{\sum_{k'=1}^K E^{(k')}}\left\|x\right\|_2 < L\right\}\end{aligned}$$

   It is easy to see that Θ_{k_0} is non-empty for any $k_0 \geq 2$. Also, Θ_{k_0} is an open set as $\frac{\sum_{k=2}^K E^{(k)}}{\sum_{k'=1}^K E^{(k')}}$ and $\frac{E^{(k_0)}}{\sum_{k'=1}^K E^{(k')}}$ are continuous. So Θ_B is an non-empty open set.

   As $\Theta \subseteq \mathcal{C}$, we know f is differentiable on $\Theta \backslash \mathcal{C}_N$ which is an non-empty open set. Then f cannot exists following the similar argument as in the proof of Theorem 3.3.

□

C Missing Proofs from Section 4

C.1 Proof of Theorem 4.1

Proof. We prove the theorem via the standard template for analyzing SGD methods [8]. Recall $\theta^{\texttt{priv}} = \frac{1}{T}\sum_{t=1}^{T}\theta_t$, where $\{\theta_1, \ldots, \theta_T\}$ are the models in each iterate of DP-GD. Let g_t denote any subgradient in $\partial\mathcal{L}^{(L)}_{\textsf{clipped}}(\theta_t; D)$. By convexity and the standard linearization trick in convex optimization [8], we have:

$$\mathcal{L}^{(L)}_{\textsf{clipped}}\left(\theta^{\texttt{priv}}; D\right) - \mathcal{L}^{(L)}_{\textsf{clipped}}\left(\theta^*; D\right) \leq \frac{1}{T}\sum_{t=1}^{T}\langle g_t, \theta_t - \theta^*\rangle \tag{7}$$

Let V be the eigenbasis of $\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T$ and let $M = VV^T$. M is a positive semidefinite matrix and it defines a seminorm $\|\cdot\|_M$ (by Definition A.4). Let $\boldsymbol{b}_t$ be the Gaussian noise vector added at time step t. To bound the error in (7), we will use a potential argument w.r.t. the potential function

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

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

$$\begin{aligned}
\Psi_t(\theta_{t+1}) &= \mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_t}\left[\|(\theta_t - \theta^*) - \eta(g_t + \boldsymbol{b}_t)\|_M^2\right] \\
&= \Psi_t(\theta_t) - 2\eta\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_t}\left[\langle g_t + \boldsymbol{b}_t, \theta_t - \theta^*\rangle_M\right] + \eta^2\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_t}\left[\|g_t + \boldsymbol{b}_t\|_M^2\right] \quad (8) \\
&= \Psi_t(\theta_t) - 2\eta\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_t}\left[\langle g_t + \boldsymbol{b}_t, \theta_t - \theta^*\rangle\right] + \eta^2\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_t}\left[\|g_t + \boldsymbol{b}_t\|_M^2\right] \quad (9) \\
&\leq \Psi_t(\theta_t) - 2\eta\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_t}\left[\langle g_t, \theta_t - \theta^*\rangle\right] + \eta^2\left(L^2 + \mathbb{E}_{\boldsymbol{b}_t}\left[\|\boldsymbol{b}_t\|_M^2\right]\right) \\
&= \Psi_{t-1}(\theta_t) - 2\eta\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_t}\left[\langle g_t, \theta_t - \theta^*\rangle\right] + \eta^2\left(L^2 + \mathbb{E}_{\boldsymbol{b}_t}\left[\|\boldsymbol{b}_t\|_M^2\right]\right) \\
&= \Psi_{t-1}(\theta_t) - 2\eta\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_t}\left[\langle g_t, \theta_t - \theta^*\rangle\right] + \eta^2\left(L^2 + \mathsf{rank}(M)\cdot\sigma^2\right) \quad (10)
\end{aligned}$$

where (9) follows because g_t lies in the subspace M, and (10) follows because $\boldsymbol{b}_t \sim \mathcal{N}(0, \sigma^2 I_p)$ and thus $\mathbb{E}_{\boldsymbol{b}_t}\left[\|\boldsymbol{b}_t\|_M^2\right] = \mathsf{rank}(M)\cdot\sigma^2$. Rearranging the terms in (10), we have the following.

$$\mathbb{E}_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_t}\left[\langle g_t, \theta_t - \theta^*\rangle\right] \leq \frac{1}{2\eta}\left(\Psi_{t-1}(\theta_t) - \Psi_t(\theta_{t+1})\right) + \frac{\eta}{2}\left(L^2 + \mathsf{rank}(M)\cdot\sigma^2\right) \quad (11)$$

Summing up (11) for all $t \in [T]$, averaging over the T iterations, combining with (7), and defining $\Psi(\theta) = \|\theta - \theta^*\|_M^2$, we get:

$$\mathbb{E}\left[\mathcal{L}^{(L)}_{\mathsf{clipped}}\left(\theta^{\mathtt{priv}}; D\right)\right] - \mathcal{L}^{(L)}_{\mathsf{clipped}}\left(\theta^*; D\right) \leq \frac{1}{2T\eta}\Psi(\mathbf{0}) + \frac{\eta}{2}\left(L^2 + \mathsf{rank}(M)\cdot\sigma^2\right) \quad (12)$$

Setting η to minimize the RHS, we have

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

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

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

The last part of the theorem follows from the fact that when $L \geq B$, $\mathcal{L}^{(L)}_{\mathsf{clipped}}(\theta; D) = \mathcal{L}(\theta; D)$ for all $\theta \in \mathbf{R}^p$. This is essentially the regime, where clipping has no effect. □

C.2 Proof of Theorem 4.2

Proof. Recall that M is the projector to the eigenspace of the matrix $\sum\limits_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T$, and $\|\cdot\|_M$ being the corresponding seminorm. Let $\theta_1, \ldots, \theta_T$ be the sequence of models generated in Line 4 of Algorithm 1, and let the constraint set $\mathcal{C} = \mathbf{R}^p$. Also, let $\boldsymbol{b}_t$ be the Gaussian noise added in the t-th iteration. By the smoothness property of $\ell(z; \cdot)$, we have the following:

$$\begin{aligned}
\mathcal{L}(\theta_{t+1}; D) &\leq \mathcal{L}(\theta_t; D) + \langle\nabla\mathcal{L}(\theta_t; D), \theta_{t+1} - \theta_t\rangle_M + \frac{\beta}{2}\|\theta_{t+1} - \theta_t\|_M^2 \\
&= \mathcal{L}(\theta_t; D) - \frac{1}{\beta}\langle\nabla\mathcal{L}(\theta_t; D), \nabla\mathcal{L}(\theta_t; D) + \boldsymbol{b}_t\rangle_M + \frac{1}{2\beta}\|\nabla\mathcal{L}(\theta_t; D) + \boldsymbol{b}_t\|_M^2 \\
&= \mathcal{L}(\theta_t; D) - \frac{1}{2\beta}\|\nabla\mathcal{L}(\theta_t; D)\|_M^2 + \frac{\|\boldsymbol{b}_t\|_M^2}{2\beta} \\
\Leftrightarrow \quad \|\nabla\mathcal{L}(\theta_t; D)\|_M^2 &\leq 2\beta\left(\mathcal{L}(\theta_t; D) - \mathcal{L}(\theta_{t+1}; D)\right) + \|\boldsymbol{b}_t\|_M^2. \quad (13)
\end{aligned}$$

Therefore, averaging over all the $t \in \{0, \dots, T-1\}$, we have the following:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\mathcal{L}(\theta_t; D)\|_M^2 &\leq \frac{2\beta}{T}\left(\mathcal{L}(\mathbf{0}; D) - \mathcal{L}(\theta_T; D)\right) + \frac{1}{T}\sum_{t=0}^{T-1}\|\boldsymbol{b}_t\|_M^2 \\ &\leq \frac{2\beta}{T}\left(\mathcal{L}(\mathbf{0}; D) - \mathcal{L}(\theta^*; D)\right) + \frac{1}{T}\sum_{t=1}^{T-1}\|\boldsymbol{b}_t\|_M^2 . \end{aligned} \tag{14}$$

Using standard Gaussian concentration, w.p. at least $1-\gamma$ over the randomness of $\{\boldsymbol{b}_1, \dots, \boldsymbol{b}_T\}$ in (14), we have the following.

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\mathcal{L}(\theta_t; D)\|_M^2 \leq \frac{2\beta}{T}\left(\mathcal{L}(\mathbf{0}; D) - \mathcal{L}(\theta^*; D)\right) + \frac{8L^2\mathsf{rank}(M)\cdot\log(1/\delta)\log(T/\gamma)}{n^2\varepsilon^2} \tag{15}$$

By an averaging argument, we know there exists $\hat{t} \in \{0, \dots, T-1\}$ s.t.

$$\|\nabla\mathcal{L}(\theta_{\hat{t}}; D)\|_M^2 \leq \frac{2\beta}{T}\left(\mathcal{L}(\mathbf{0}; D) - \mathcal{L}(\theta^*; D)\right) + \frac{8L^2\mathsf{rank}(M)\cdot\log(1/\delta)\log(T/\gamma)}{n^2\varepsilon^2}.$$

As long as $T \geq \frac{\beta n^2\varepsilon^2\cdot\mathcal{L}(\mathbf{0};D)}{2L^2\log(1/\delta)}$, we have $\left\|\nabla\mathcal{L}(\theta_{\hat{t}}; D)\right\|_M \leq \frac{4L\sqrt{\mathsf{rank}(M)\cdot\log(1/\delta)\log(T/\gamma)}}{\varepsilon n}$. Now, notice that the ℓ_2-sensitivity [13] of $\left\|\nabla\mathcal{L}(\theta_{\hat{t}}; D)\right\|_M$ is at most $\frac{2L}{n}$. Therefore, releasing $t_{\mathsf{priv}} \leftarrow \arg\min\limits_{t\in[T]}\left\|\nabla\mathcal{L}(\theta_t; D)\right\|_M + \mathsf{Lap}\left(\frac{4L}{n}\right)$ conditioned on $\theta_1, \dots, \theta_T$ satisfies ε-differential privacy (by the analysis of the report-noisy-max algorithm [13]). Therefore, the whole algorithm is $(2\varepsilon, \delta)$-differentially private.

As for utility, we have w.p. at least $1-\gamma$,

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

Here, we have used the standard concentration property of Laplace random variable. This completes the proof. □

D Dimension-independent Locally-private Empirical Risk Minimization

In this section, we show that the dimension-independence guarantee (Theorem 4.1) seamlessly extends to the setting of local differential privacy (LDP) [43, 14, 24, 37]. Unlike central differential privacy where the data is held by a trusted central curator, in the LDP setting, data is assumed to be distributed and perturbed before sending to any aggregator. The semantics in terms of privacy is that the complete transcript of the interaction with an individual data record should preserve LDP defined as follows.

Definition D.1 ((ε, δ)-Local differential privacy [24, 37])**.** *A randomized algorithm $\mathcal{A}$ is (ε, δ)-locally differentially private (LDP) if, for any pair of data records $d, d' \in \mathcal{D}$, 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 $\mathcal{A}$. A multi-player protocol is (ε, δ)-LDP if for all possible inputs and runs of the protocol, the transcript of player is interactions with the server is (ε, δ)-LDP (for all settings of the remaining data points).

We define algorithm DP-GD$_{\mathsf{LDP}}$ to be Algorithm 1 with the following modifications to Lines 3 and 4:

3. $\boldsymbol{g}_t = \frac{1}{n}\sum\limits_{i=1}^{n}\left(\mathsf{clip}\left(\nabla\ell(\theta_t; d_i)\right) + \mathcal{N}(0, \sigma^2)\right)$.
4. $\theta_{t+1} \leftarrow \Pi_{\mathcal{C}}\left(\theta - \eta_t\cdot\boldsymbol{g}_t\right)$, where $\Pi_{\mathcal{C}}(\boldsymbol{v}) = \arg\min\limits_{\theta\in\mathcal{C}}\|\boldsymbol{v} - \theta\|_2$.

Essentially, at iteration t, each user provides a private version of their data, i.e., $\mathsf{clip}\left(\nabla \ell(\theta_t; d_i)\right) + \mathcal{N}(0, \sigma^2)$, and the central aggregator averages them and updates the model. This fits perfectly in the LDP setting.

Theorem D.2 (Local differential privacy guarantee)**.** *DP-GD*$_{\mathsf{LDP}}$ *is* (ε, δ)*-locally differentially private, if one sets the noise variance as* $\sigma^2 = \frac{2L^2 T \log(1/\delta)}{\varepsilon^2}$*, where* L *is the clipping norm.*

Theorem D.2 follows immediately from the privacy property of the Gaussian mechanism [13, 30].

In the following, we provide a corollary to Theorem 4.1 that highlights the dimension-independence of DP-GD$_{\mathsf{LDP}}$. The proof of Corollary D.3 is identical to that of Theorem 4.1. As long as $\mathsf{rank}(M) \ll p$, this guarantee is tighter than the worst-case guarantee of $\widetilde{\Theta}\left(\frac{\sqrt{p}}{\varepsilon\sqrt{n}}\right)$ [10, 37].

Corollary D.3. *Following the same notation as in Theorem 4.1, setting the constraint set* $\mathcal{C} = \mathbf{R}^p$*, clipping norm* L*, and running DP-GD*$_{\mathsf{LDP}}$ *for* $T = n\varepsilon^2$ *steps with appropriate learning rate* η*, we have*

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

Furthermore, if B *is the Lipscthiz constant for the loss function* $\ell(\cdot;\cdot))$*, and* $L \geq B$*, then we have:*

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

Here, $\mathsf{rank}(M) \leq n$*, but can be much smaller.*

Remark 1. *While the results above are stated for* (ε, δ)*-LDP, they can easily be extended to* ε*-LDP (with the same asymptotics), albeit using a different randomization method from [10].*

D.1 Related Work on Choosing Optimal Clipping Norm

In this work, we show that an optimal choice of the clipping norm in DP-GD is necessary for attaining reasonable excess empirical risk. Choosing the clipping norm “too low” introduces bias by changing the underlying objective, whereas setting it “too high” introduces variance in the model estimate by increasing the noise level. There has been both theoretical [18, 7, 26, 41, 4] and empirical research [40, 33] providing algorithms which can be used for choosing a “near-optimal” value of the clipping norm. For instance, [40, 33] track differentially private estimates of various statistics, like the percentage of the individual gradients getting clipped, or the mean and variance of the noisy gradient estimates across the training, and adaptively adjust the value of the clipping norm. The work in [3] studies the bias-variance trade-off on setting the clipping norm value; our work is tangential in that it provides risk bounds for any value of the clipping norm. The algorithms in [18, 7, 26, 41] come in variety of flavors. One natural and powerful approach [26] is to first compute the excess empirical risk attained by DP-SGD using a “potentially large” candidate set of clipping norms. Then, choose the best clipping norm using a differentially private selection procedure. The strength of this result is that the “cost of privacy” is almost independent of the number of possible clipping norms tried.