Beyond Uniform Lipschitz Condition in Differentially Private Optimization

Rudrajit Das [*1] **Satyen Kale** [2] **Zheng Xu** [2] **Tong Zhang** [2 3] **Sujay Sanghavi** [1]

Abstract

Most prior results on differentially private stochastic gradient descent (DP-SGD) are derived under the simplistic assumption of uniform Lipschitzness, i.e., the per-sample gradients are uniformly bounded. We generalize uniform Lipschitzness by assuming that the per-sample gradients have sample-dependent upper bounds, i.e., per-sample Lipschitz constants, which themselves may be unbounded. We provide principled guidance on choosing the clip norm in DP-SGD for convex over-parameterized settings satisfying our general version of Lipschitzness when the per-sample Lipschitz constants are bounded; specifically, we recommend tuning the clip norm only till values up to the minimum per-sample Lipschitz constant. This finds application in the private training of a softmax layer on top of a deep network pre-trained on public data. We verify the efficacy of our recommendation via experiments on 8 datasets. Furthermore, we provide new convergence results for DP-SGD on convex and nonconvex functions when the Lipschitz constants are unbounded but have bounded moments, i.e., they are heavy-tailed.

1. Introduction

With the ever-increasing amount of data being used, there is a growing need for the development of privacy-preserving training schemes for machine learning (ML) models. Differential privacy (DP) (Dwork et al., 2006) is a popular privacy-preserving framework that is being incorporated in the training of ML models. We formally define DP in Definition 3.3, but at a high level, DP can be guaranteed by adding Gaussian noise, where the noise scale is determined by the "sensitivity" to an individual's data. There has been copious research on private optimization for private training; in this paper, we focus on DP-SGD (Abadi et al., 2016) which is the default algorithm for private optimization in practice.

We briefly introduce the problem setting and DP-SGD to facilitate further discussion (see Section 3 for more details). We consider empirical risk minimization (ERM) of $f(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{w})$, where each $f_i : \mathbb{R}^d \to \mathbb{R}$. In every iteration of DP-SGD (stated in Algorithm 1), the optimizer receives a noise-perturbed average of the *clipped* per-sample gradients for performing the update; noise is added to guarantee differential privacy. Specifically, at iteration t, the optimizer receives $\boldsymbol{g}_t = \frac{1}{b}\sum_{i\in\mathcal{S}_t} \text{clip}(\nabla f_i(\boldsymbol{w}_t), \tau) + \boldsymbol{\zeta}_t$, where $\mathcal{S}_t$ is a random batch of samples formed by picking each sample in $\{1, \ldots, n\}$ with probability (b/n), $\text{clip}(\boldsymbol{z}, c) := \boldsymbol{z}\min(1, c/\|\boldsymbol{z}\|)$ for a vector $\boldsymbol{z}$, τ is the clipping threshold or clip norm, and $\boldsymbol{\zeta}_t$ is an isotropic Gaussian random vector whose variance is proportional to τ^2 and also depends on the amount of privacy required.

Clipping is employed in DP-SGD to bound the maximum sensitivity of the average gradient to each sample's individual gradient, which is required to set the noise variance. However, clipping can also make $\boldsymbol{g}_t$ a *biased* estimator of $\nabla f(\boldsymbol{w}_t)$, and the amount of bias depends on the clip norm τ – the higher we set τ, the lower is the bias, and vice-versa. As the noise variance is proportional to τ^2, there is an inherent tension between the bias and variance of $\boldsymbol{g}_t$ due to the clip norm τ. This raises a natural question - *how do we set "good" clip norms to balance the bias-variance tradeoff?*

In order to circumvent the analysis of the clipping bias, most prior convergence results for private optimization (Bassily et al., 2014; 2019; Wang et al., 2018; 2019) assume that the loss function is *uniformly Lipschitz* for all samples and model parameters, i.e., the per-sample gradients (w.r.t. the model parameters) have a sample-independent upper bound known as the Lipschitz constant. Under this assumption, the clip norm is set equal to the Lipschitz constant resulting in zero bias as no clipping happens. But in practice, such a choice of clip norm results in very poor performance; see Figure 1 and its caption. Additionally, uniform Lipschitzness does not even hold for simple problems like linear regression with Gaussian data, precluding the existence of a trivial clip norm for analysis.

In this work, we generalize uniform Lipschitzness by instead assuming that the per-sample gradients have sample-

[*]Part of this work was done when Rudrajit was an intern at Google Research. [1]UT Austin [2]Google Research [3]HKUST. Correspondence to: Rudrajit Das <rdas@utexas.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

dependent upper bounds, i.e., per-sample Lipschitz constants, which themselves may be unbounded. Under this generalized assumption, we provide a theoretically-motivated clip norm selection strategy for convex settings. Our method finds direct application in the private training of the softmax layer of pre-trained deep networks which is a popular and effective scheme for private training (De et al., 2022; Mehta et al., 2022). In practice, the clip norm is tuned over multiple runs which is not only computationally inefficient but also increases the privacy cost (Papernot & Steinke, 2021). So it is desirable to have principled methods for setting the clip norm to alleviate these two issues. Additionally, we provide novel convergence results for DP-SGD when the per-sample Lipschitz constants are heavy-tailed.

Before we state our contributions in detail, we need to briefly introduce the metric quantifying convergence, which we call the "optimization risk". Let $\boldsymbol{w}_{\text{priv}}$ be the output of DP-SGD (Alg. 1). If f is convex, the optimization risk is the expected suboptimality gap, i.e., $\mathbb{E}[f(\boldsymbol{w}_{\text{priv}})] - \min_{\boldsymbol{w}} f(\boldsymbol{w})$. If f is nonconvex, the optimization risk is the expected gradient-norm squared, i.e. $\mathbb{E}[\|\nabla f(\boldsymbol{w}_{\text{priv}})\|^2]$. When DP-SGD is (ε, δ)-DP (defined in Definition 3.3), our convergence results are expressed in terms of the following key quantity:

$$\varphi := \sqrt{\nu d \log(1/\delta)}/n\varepsilon, \tag{1}$$

where d is the dimension of the model parameters, n is the number of samples, and ν is an absolute constant. We assume that n is large enough so that $\varphi < 1$, and the number of iterations of DP-SGD is sufficiently large. We now list our **main contributions**, and also summarize the main theoretical results in Tables 1 and 2.

(a) Throughout this work, we generalize uniform Lipschitzness by assuming that the per-sample gradients have sample-dependent upper bounds, i.e., *per-sample Lipschitz constants*, which themselves may be unbounded; we call this *generalized Lipschitzness* (Assumption 4.1).

- In Section 5, we provide a principled clip norm tuning strategy for DP-SGD under generalized Lipschitzness. Specifically, we focus on convex settings with near-interpolation like conditions (Asmp. 5.5), i.e., $\frac{1}{n}\sum_{i=1}^{n}(f_i(\boldsymbol{w}^*) - \min_{\boldsymbol{w}} f_i(\boldsymbol{w})) \approx 0$ where $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} f(\boldsymbol{w})$; over-parameterization is an example of this. For such cases, we recommend *tuning the clip norm only till values up to the minimum per-sample Lipschitz constant* (Remark 5.6), say G_{min}, based on Theorem 5.4 where we show that the optimization risk attains the best bound when the clip norm is $\leq G_{\text{min}}$. This is in contrast to prior theoretical works which set the clip norm equal to the *maximum* per-sample Lipschitz constant, say G_{max}, for ease of analysis.
- Our recommendation for convex settings is of relevance to the private training of the softmax (i.e., last) layer of deep networks pre-trained on public data. In Section 5.2, we corroborate our recommendation with experiments on four[1] vision datasets, viz., Caltech-256 (Griffin et al., 2007), Food-101 (Bossard et al., 2014), CIFAR-100 and CIFAR-10, and two language datasets, viz., TweetEval Emoji (Barbieri et al., 2018) and Emotion (Saravia et al., 2018). As an e.g., for Caltech-256 and Food-101 with $\varepsilon = 2$, the test accuracy obtained by setting the clip norm $\tau = G_{\text{min}}$ is better than that of $\tau = G_{\text{max}}$ by more than 23% and 11%, respectively.

(b) In Section 6, we provide optimization risk bounds for DP-SGD under generalized Lipschitzness, without assuming interpolation, when the per-sample Lipschitz constants have bounded k^{th} moment, i.e., they are *heavy-tailed* (Asmp. 6.1). For private **unconstrained convex** and smooth **non-convex** optimization under this assumption, we derive risk bounds of $\mathcal{O}(\varphi^{1-\frac{2}{k+1}})$ and $\mathcal{O}(\varphi^{1-\frac{1}{2k-1}})$, respectively[2]; see Theorems 6.2 and 6.8. Under an additional mild assumption, we improve the risk bound in the convex case to $\mathcal{O}(\varphi^{1-\frac{1}{k}})$ and also derive a matching lower bound, thereby establishing the *optimality of DP-SGD* in this case; see Assumption 6.4 and Theorems 6.5 and 6.6. These are the first results for private **unconstrained** optimization under the heavy-tailed assumption or anything similar. Our results also imply the *optimality of DP-SGD* in the **unconstrained** convex case under *uniform* Lipschitzness; see Cor. 6.7. To our knowledge, this is the first matching lower bound for the **unconstrained** convex case under *uniform* Lipschitzness.

Table 1. **Summary of our clip norm selection result** (Thm. 5.4) for the convex case under generalized Lipschitzness (Asmp. 5.1) and interpolation (Asmp. 5.5). G_1 and G_n are the *minimum* and *maximum* per-sample Lipschitz constants as per Asmp. 5.1. In the table, $\alpha^* = \alpha(G_1) \geq 1$, where $\alpha(G_1)$ is defined in Definition 5.3, and $B = \mathcal{O}(\|\boldsymbol{w}_0 - \boldsymbol{w}^*\| G_n \varphi)$, where $\boldsymbol{w}_0$ is the initial point, $\boldsymbol{w}^*$ is a minimizer of f and $\varphi = \mathcal{O}(\sqrt{d\log(1/\delta)}/n\varepsilon)$.

Clip Norm τ	Risk Upper Bound when Perfect Interpolation holds ($\Delta(\boldsymbol{w}^*) = 0$ in Asmp. 5.5)
$\in (0, G_1]$ (**this work**)	B/α^* ($\alpha^* \geq 1$)
$\in (G_1, G_n)$ (**this work**)	$\geq B/\alpha^*$ but $\leq B$
G_n (default choice of prior theory)	B

2. Related Work

DP-(S)GD with Clipping: Abadi et al. (2016) introduce the famous DP-SGD algorithm with clipping for differentially private training in practice. There are some papers analyzing

[1]We show results on two more datasets in Appendix A.

[2]The seemingly better result for the nonconvex setting is because of the difference in the risk metrics between the convex and nonconvex cases.

Table 2. **Summary of optimization risk (OR) bounds.** OR is defined in Definition 3.5 and $\varphi = \mathcal{O}(\sqrt{d\log(1/\delta)}/n\varepsilon) < 1$. In Assumption 6.1, we assume that the per-sample gradients have sample-dependent upper bounds with bounded k^{th} moment ($k > 1$). In Assumption 6.4, we assume a mild lower bound on function suboptimality of points far away from the optimum.

Reference	Assumption(s) & Setting	Risk Upper Bound	Matching Lower Bound?
Bassily et al. (2014)	Uniform Lipschitz & Convex Constrained	$\mathcal{O}(\varphi)$	Yes
This work	Uniform Lipschitz & Convex Unconstrained ($\mathcal{W} = \mathbb{R}^d$)	$\mathcal{O}(\varphi)$ (Cor. 6.7)	Yes, for $\delta < e^{-\varepsilon^2}$ (Cor. 6.7)
Kamath et al. (2021), Lowy & Razaviyayn (2023)[1]	Assumption 6.1 & Convex Constrained	$\mathcal{O}(\varphi^{1-\frac{1}{k}})$	Yes
This work	Assumption 6.1 & Convex Unconstrained	$\mathcal{O}(\varphi^{1-\frac{2}{k+1}})$ (Thm. 6.2)	?
This work	Assumptions 6.1, 6.4 & Convex Unconstrained	$\mathcal{O}(\varphi^{1-\frac{1}{k}})$ (Thm. 6.5)	Yes, for $\delta < e^{-\varepsilon^2}$ (Thm. 6.6)
Arora et al. (2022), Tran & Cutkosky (2022)[2]	Uniform Lipschitz & Smooth Nonconvex Unconstrained	$\mathcal{O}(\varphi^{4/3})$	?
This work	Assumption 6.1 & Smooth Nonconvex Unconstrained	$\mathcal{O}(\varphi^{1-\frac{1}{2k-1}})$ (Thm. 6.8)	?

[1]These two works consider the stochastic optimization setting with $(0, \mathcal{O}(\varepsilon^2))$-zCDP and derive bounds for the generalization error. In Appendix I, we show that the same bounds hold for the training error (i.e., OR) in empirical risk minimization with (ε, δ)-DP.

[2]The $\mathcal{O}(\varphi^{4/3})$ bound is attained by algorithms very different from DP-SGD. For DP-SGD like algorithms, Wang et al. (2018) obtain the best known bound of $\mathcal{O}(\varphi)$.

the effect of clipping in DP-(S)GD in different settings such as Chen et al. (2020); Bu et al. (2021); Song et al. (2021). However, these works do not provide any practical insights into how to set the clip norm for DP-SGD, which is a key focus of our work. Related to our focus, there are some variants of DP-SGD such as Andrew et al. (2019); Du et al. (2021); Wu et al. (2021) that adaptively change the clip norm and/or noise variance to improve convergence. However, practitioners typically use a *constant* clip norm which does not change with the iteration number, and so we only focus on how to set a *constant* clip norm used in *vanilla* DP-SGD (Abadi et al., 2016).

Differentially Private Optimization under Uniform Lipschitzness: There are several papers on private empirical risk minimization (ERM) (Chaudhuri & Monteleoni, 2008; Chaudhuri et al., 2011; Kifer et al., 2012; Song et al., 2013; Duchi et al., 2013; Bassily et al., 2014; Talwar et al., 2014; 2015; Wu et al., 2017; Iyengar et al., 2019) and private stochastic optimization (Bassily et al., 2014; 2019; Feldman et al., 2020; Asi et al., 2021a;b; Kulkarni et al., 2021; Bassily et al., 2021) for *convex Lipschitz* objectives within a *bounded set*. The optimal risk bound for private constrained convex ERM over a bounded set in the Lipschitz case is $\mathcal{O}(\varphi)$ (Bassily et al., 2014). Zhang et al. (2017); Wang et al. (2018; 2019); Arora et al. (2022); Tran & Cutkosky (2022) derive convergence results for private unconstrained *nonconvex* ERM with Lipschitz and smooth objectives. Wang et al. (2018; 2019) obtain a risk bound of $\mathcal{O}(\varphi)$ for DP-SGD like algorithms. Arora et al. (2022); Tran & Cutkosky (2022) obtain an improved bound of $\mathcal{O}(\varphi^{4/3})$ but with more advanced algorithms. These papers simply clip the gradients up to the Lipschitz constant, and do not explore the effect of clipping to smaller values. But as shown in Fig. 1, such a choice of clip norm performs poorly in practice and besides, uniform Lipschitzness may not always hold.

Bounded Gradient Moments: Our assumption of per-sample Lipschitz constants having bounded k^{th} moment, i.e. Assumption 6.1, generalizes the "heavy-tailed" assumption of Wang et al. (2020); Hu et al. (2021) for private stochastic convex optimization (SCO) with bounded second moment. The bounded k^{th} moment assumption has been also analyzed in Kamath et al. (2021); Lowy & Razaviyayn (2023) for private SCO. However, these papers focus only on *constrained convex* optimization. In practice, however, unconstrained minimization is usually performed while training ML models. In this paper, we focus on the *more practical and harder* (w.r.t. analysis) case of private *unconstrained* convex as well as nonconvex optimization (which has not been considered before) under this assumption. We discuss the assumptions of these papers in more detail after Assumption 6.1. Note that none of these papers have any result like our theoretically-motivated clip norm selection method *for use in practice* with *our level of experimental verification*.

3. Preliminaries

Notation: Vectors and matrices are in bold face. For any $n \in \mathbb{N}$, the set $\{1, \ldots, n\}$ is denoted by $[n]$, and the uniform distribution over $\{0, \ldots, n\}$ is denoted by $\text{unif}[0, n]$. $\|.\|$ denotes the ℓ_2 norm throughout this work. For a function h and any point $\boldsymbol{x} \in \mathcal{X}$, the "suboptimality gap" (at $\boldsymbol{x}$) over $\mathcal{X}$ means $h(\boldsymbol{x}) - \min_{\boldsymbol{x}' \in \mathcal{X}} h(\boldsymbol{x}')$. The function $\text{clip}(.,.)$:

$\mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^d$ is defined as:

$$\text{clip}(\boldsymbol{z}, c) := \boldsymbol{z}\min(1, c/\|\boldsymbol{z}\|). \tag{2}$$

Definition 3.1 (Lipschitz). $h : \mathcal{T} \to \mathbb{R}$ is to said to be G-Lipschitz if $\sup_{\boldsymbol{t}\in\mathcal{T}} \|\nabla h(\boldsymbol{t})\| \le G$.

Definition 3.2 (Smooth). $h : \mathcal{T} \to \mathbb{R}$ is to said to be L-smooth if for all $\boldsymbol{t}, \boldsymbol{t}' \in \mathcal{T}$, $\|\nabla h(\boldsymbol{t}) - \nabla h(\boldsymbol{t}')\| \le L\|\boldsymbol{t} - \boldsymbol{t}'\|$.

Definition 3.3 (Differential Privacy (Dwork et al., 2014)). Suppose each sample $\in \mathcal{S}$. Given a query function $h : \mathcal{S}^n \to \mathcal{X}$, a randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ is said to be (ε, δ)-DP if for any two datasets $\mathcal{D}, \mathcal{D}' \in \mathcal{S}^n$ differing in exactly one sample and for any measurable $\mathcal{R} \in \mathcal{Y}$, $\mathbb{P}(\mathcal{M}(h(D)) \in \mathcal{R}) \le e^{\varepsilon}\mathbb{P}(\mathcal{M}(h(D')) \in \mathcal{R}) + \delta$.

The customary way to guarantee DP is to add zero-mean Gaussian noise to the output of $h(.)$ above; this is known as the Gaussian mechanism (Dwork et al., 2014).

Problem Setting and DP-SGD: Suppose we are given a dataset of n i.i.d. samples (features and corresponding labels) $\mathcal{Z} := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ drawn from some distribution $\mathcal{D}$. We wish to train a model, parameterized by $\boldsymbol{w} \in \mathcal{W} \subseteq \mathbb{R}^d$, on the data via DP-SGD such that the whole training process is (ε, δ)-DP. We use a loss function $\ell(\boldsymbol{w}, .)$ (for e.g., the squared loss or cross-entropy loss with some regularization possibly) to learn the model. Let $f_i(\boldsymbol{w}) := \ell(\boldsymbol{w}, \boldsymbol{x}_i, y_i)$; then, we are trying to privately minimize

$$f(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{w}). \tag{3}$$

DP-SGD is summarized in Algorithm 1. Gradient clipping is employed to bound the sensitivity of the average gradient to each sample's individual gradient. Gaussian noise is added to guarantee DP. In Abadi et al. (2016), the last iterate $\boldsymbol{w}_T$ is returned; in contrast, we return a random iterate. We now specify the value of σ_n^2 required to make Alg. 1 (ε, δ)-DP using the moments accountant method of Abadi et al. (2016); we provide a short proof in Appendix G.

Theorem 3.4 (Moments Accountant (Abadi et al., 2016)). *For $\varepsilon < \mathcal{O}\big(\frac{b^2}{n^2}T\big)$, Algorithm 1 is (ε, δ)-DP for $\sigma_n^2 = \frac{\nu T \log(\frac{1}{\delta})\tau^2}{n^2\varepsilon^2}$, where ν is an absolute constant.*

Finally, we define our convergence metric for DP-SGD which we call the *optimization risk*.

Definition 3.5 (Optimization Risk). Recall $\boldsymbol{w}_{\text{priv}}$ is the output of DP-SGD (Alg. 1) after T iterations.
(i) Suppose f is convex. We define the *convex optimization risk* as $\text{OR}(T) := \big(\mathbb{E}[f(\boldsymbol{w}_{\text{priv}})] - f(\boldsymbol{w}^*)\big)$, where $\boldsymbol{w}^* \in \text{argmin}_{\boldsymbol{w}\in\mathcal{W}} f(\boldsymbol{w})$.

(ii) Suppose f is nonconvex & $\mathcal{W} = \mathbb{R}^d$. We define the *nonconvex optimization risk* as $\text{OR}(T) := \mathbb{E}[\|\nabla f(\boldsymbol{w}_{\text{priv}})\|^2]$.

Algorithm 1 DP−SGD (Abadi et al., 2016)

1: **Input:** Domain of parameters $\mathcal{W}$, initial point $\boldsymbol{w}_0 \in \mathcal{W}$, number of iterations T, learning rates $\{\eta_t\}_{t=0}^{T-1}$, sample selection probability (b/n), clip norm τ and noise variance σ_n^2.
2: **for** $t = 0, \ldots, T-1$ **do**
3: Form a random mini-batch $\mathcal{S}_t$ by picking each sample independently of the others with probability b/n.
4: Get the noisy mini-batch stochastic gradient $\boldsymbol{g}_t = \frac{1}{b}\sum_{i\in\mathcal{S}_t} \text{clip}(\nabla f_i(\boldsymbol{w}_t), \tau) + \boldsymbol{\zeta}_t$, where $\boldsymbol{\zeta}_t \sim \mathcal{N}(\vec{0}_d, \sigma_n^2 \mathrm{I}_d)$ and clip() is defined in (2).
5: Let $\boldsymbol{z}_{t+1} \leftarrow \boldsymbol{w}_t - \eta_t \boldsymbol{g}_t$. Update $\boldsymbol{w}_{t+1} \leftarrow \Pi_{\mathcal{W}}(\boldsymbol{z}_{t+1})$, where $\Pi_{\mathcal{W}}(\boldsymbol{z})$ is the projection of $\boldsymbol{z}$ onto $\mathcal{W}$. (Note that $\Pi_{\mathbb{R}^d}(\boldsymbol{z}) = \boldsymbol{z}$.)
6: **end for**
7: Return $\boldsymbol{w}_{\text{priv}} = \boldsymbol{w}_{\widehat{t}}$, where $\widehat{t} \sim \text{unif}[0, T-1]$.

Note that the expectations above are w.r.t. the randomness of Algorithm 1 (in particular, conditioned on the dataset $\mathcal{Z}$).

Also recall the key quantity $\varphi = \frac{\sqrt{\nu d \log(1/\delta)}}{n\varepsilon} < 1$ defined in Equation (1). Our bounds on the optimization risk will be in terms of φ. For brevity, we only present abridged versions of our results in the main paper and provide the full versions and proofs in the Appendix.

4. Generalized Lipschitzness

Here we introduce our *proposed* generalization to the commonly used uniform Lipschitzness assumption.

Assumption 4.1 (Generalized Lipschitzness). For any $\boldsymbol{w} \in \mathcal{W}$ and $(\boldsymbol{x}, y) \sim \mathcal{D}$, the following holds for some sample-dependent function $G(\boldsymbol{x}, y)$:

$$\|\nabla_{\boldsymbol{w}}\ell(\boldsymbol{w}, \boldsymbol{x}, y)\| \le G(\boldsymbol{x}, y),$$

where ℓ is our loss function (mentioned before Equation (3)). We call $G(\boldsymbol{x}, y)$ the "per-sample Lipschitz constant".[3]

Note that we are *not* imposing the condition that $G(\boldsymbol{x}, y)$ be itself bounded for all $(\boldsymbol{x}, y)$. In fact, if we do impose that, then we recover uniform Lipschitzness. We now provide an important example where generalized Lipschitzness holds.

Logistic regression: Consider doing logistic regression for multi-class classification with the cross-entropy loss, where m is the number of classes. Suppose $\boldsymbol{x} \sim \mathcal{F}$ (with a '1' appended to account for the bias term) is the feature and $y \in [m]$ is the corresponding class number. In Appendix D, we show that $\|\nabla\ell_{\boldsymbol{w}}(\boldsymbol{w}, \boldsymbol{x}, y)\| \le \sqrt{2}\|\boldsymbol{x}\|$. Thus, Assumption 4.1 holds with $G(\boldsymbol{x}, y) = \sqrt{2}\|\boldsymbol{x}\|$ *for any* $\mathcal{W}$.

[3]This assumption extends to non-differentiable functions by replacing the gradient with sub-gradient.

5. Clip Norm Selection for DP-SGD under Generalized Lipschitzness

Here, we will provide a principled strategy to select the constant clip norm τ (which does not change across iterations) in DP-SGD (Algorithm 1) under generalized Lipschitzness when the maximum per-sample Lipschitz constant is bounded, i.e., uniform Lipschitzness holds. To that end, we assume the following.

Assumption 5.1. Assumption 4.1 holds for the dataset $\mathcal{Z} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ that we receive. Let $G_i = G(\boldsymbol{x}_i, y_i)$ for $i \in [n]$. Also, without loss of generality, the sample indices are arranged so that $0 < G_1 < \ldots < G_n < \infty$[4].

Thus, $\{G_i\}_{i=1}^n$ are the per-sample Lipschitz constants for the dataset $\mathcal{Z}$. For e.g., logistic regression with cross-entropy loss satisfies Assumption 5.1 with $G_i = \sqrt{2}\|\boldsymbol{x}_i\|$ (see discussion on logistic regression after Assumption 4.1).

Under Assumption 5.1, if we follow the approach of prior theoretical works such as Bassily et al. (2014), then we would choose G_n as the clip norm τ – this is associated with zero bias (as no actual clipping occurs) but high noise variance, yielding a risk bound of $\mathcal{O}(G_n\varphi)$ in the convex case. While the dependence on φ is tight (Bassily et al., 2014), it is not clear if $\tau = G_n$ leads to the best **constant factors** in the risk bound – which is what makes a difference in practice. Using Thm. 5.4 of this work, *we show that the best constant factors are obtained by choosing $\tau \leq G_1$ in convex problems with near-interpolation like conditions* (i.e., the training data can be almost perfectly fitted) while retaining the $\mathcal{O}(G_n\varphi)$ dependence. This is consistent with experiments in Section 5.2, where clip norms $\leq G_1$ perform the best. Intuitively, this happens because the high noise variance associated with large clip norms is more detrimental to convergence than the bias associated with small clip norms. Let us now talk about this result in more detail.

5.1. Convex Case

The results here are for a general convex constraint set $\mathcal{W}$ which can be $\mathbb{R}^d$ too. Before we can state our result, we need to introduce some definitions first. Let $\Psi := \arg\min_{\boldsymbol{w}\in\mathcal{W}} f(\boldsymbol{w})$ and $f_i^* = \min_{\boldsymbol{w}\in\mathcal{W}} f_i(\boldsymbol{w})$.

Definition 5.2. For any $\boldsymbol{w}^* \in \Psi$, define $\Delta(\boldsymbol{w}^*) := \frac{1}{n}\sum_{i\in[n]}(f_i(\boldsymbol{w}^*) - f_i^*)$.

Definition 5.3. Suppose Assumption 5.1 holds. For clip norm $\tau \in (0, G_n]$, define:

$$\alpha(\tau) := \inf_{\boldsymbol{w}\in\mathcal{W}\setminus\Psi} \frac{\frac{1}{n}\sum_{i\in[n]} \min\left(\frac{1}{\tau}, \frac{1}{G_i}\right)\left(f_i(\boldsymbol{w}) - f_i^*\right)}{\frac{1}{n}\sum_{i\in[n]} \frac{f_i(\boldsymbol{w}) - f_i^*}{G_n}}.$$

[4]We assume strict inequalities here because the probability measure of equality holding is zero. Also, we assume $G_1 > 0$, as otherwise f_1 is a constant function which is trivially minimized everywhere, and we can minimize $\frac{1}{n-1}\sum_{i=2}^n f_i(\boldsymbol{w})$ instead.

Note that:
(i) $\alpha(\tau) \geq 1$ for all $\tau \in (0, G_n]$ and $\alpha(G_n) = 1$.
(ii) $\alpha(\tau)$ is a non-increasing function of τ.
(iii) $\alpha(\tau) = \alpha(G_1)$ for all $\tau \in (0, G_1]$.
(iv) $\alpha(G_1)$ is strictly greater than 1 *unless* there exists a $\widetilde{\boldsymbol{w}}^*$ such that $\widetilde{\boldsymbol{w}}^*$ is a minimizer of $\{f_i\}_{i=1}^{n-1}$ but not of f_n.
(v) $\frac{G_n}{\tau\alpha(\tau)} \geq 1$ for all $\tau \in (0, G_n]$.
(vi) $\frac{G_n}{\tau\alpha(\tau)}$ is a non-increasing function of τ.

Let us see why $\alpha(\tau) \geq 1$ in Definition 5.3. By definition, $f_i(\boldsymbol{w}) - f_i^* \geq 0$ for all $\boldsymbol{w} \in \mathcal{W}$. Now since $G_1 < \ldots < G_n$ (as per Assumption 5.1) and $\tau \leq G_n$, we have that:

$$\min\left(\frac{1}{\tau}, \frac{1}{G_i}\right)\left(f_i(\boldsymbol{w}) - f_i^*\right) \geq \frac{f_i(\boldsymbol{w}) - f_i^*}{G_n} \;\forall i \in [n]. \quad (4)$$

Thus, $\alpha(\tau) \geq 1$ for all $\tau \leq G_n$. (ii) and (iii) are easy to verify using properties of $\min()$. Let us now discuss why (iv) must be true. For $\tau = G_1$, the only way equality will hold in Equation (4) for some $\boldsymbol{w} \notin \Psi$ is if $f_i(\boldsymbol{w}) = f_i^* \;\forall\; i \in [n-1]$ but $f_n(\boldsymbol{w}) > f_n^*$; (iv) follows from this. (v) and (vi) can be checked by just writing out the expression for $\frac{G_n}{\tau\alpha(\tau)}$ and then using properties of $\min()$. We are now ready to present our result for the convex case.

Theorem 5.4 (Convex Case). *Suppose each f_i is convex, $\mathcal{W}$ is a convex set (which can be $\mathbb{R}^d$) and Assumption 5.1 holds. Fix some $C > 0$. In Alg. 1, set $T = \frac{1}{3\varphi^2}$ and $\eta_t = \frac{3C\varphi}{2\tau} \;\forall\; t$ for clip norm τ. Then, DP-SGD has the following optimization risk bound as a function of $\tau \in (0, G_n]$:*

$$\text{OR}(T) \leq \underbrace{\frac{1}{\alpha(\tau)}\left(\left(\frac{\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2}{C} + 2C\right)G_n\varphi\right)}_{\text{(A)}} + \underbrace{\left(\frac{G_n}{\tau\alpha(\tau)} - 1\right)\Delta(\boldsymbol{w}^*)}_{\text{(B)}}, \quad (5)$$

where $\Delta(\boldsymbol{w}^*) \geq 0$ and $\alpha(\tau) \geq 1$ are as defined in Definitions 5.2 and 5.3, respectively.

In Equation (5), term (A) is a non-decreasing function of τ; this follows from (ii) in Definition 5.3. But term (B) ≥ 0 is a non-increasing function of τ; this follows from (v) and (vi) in Definition 5.3. Thus, increasing τ increases (A) but reduces (B), and vice-versa. So in general, there is a **tradeoff** between the values of (A) and (B) while setting τ. To provide a recommendation for the choice of τ, we shall focus on a commonly occurring scenario.

Interpolation: Modern **over-parameterized** ML models are able to perfectly fit all the training data (Zhang et al., 2021; Ma et al., 2018). In such cases, it is reasonable to make the following *relaxed interpolation* assumption (based on Assumption 1 of Ma et al. (2018)).

Assumption 5.5 (**Relaxed Interpolation**). There exists some $\boldsymbol{w}^* \in \text{argmin}_{\boldsymbol{w}\in\mathcal{W}} f(\boldsymbol{w})$ such that $\Delta(\boldsymbol{w}^*) = \frac{1}{n}\sum_{i\in[n]}(f_i(\boldsymbol{w}^*) - f_i^*) \approx 0$.

Assumption 1 of Ma et al. (2018) would imply the existence of some $\boldsymbol{w}^* \in \text{argmin}_{\boldsymbol{w}\in\mathcal{W}} f(\boldsymbol{w})$ such that $\Delta(\boldsymbol{w}^*) = 0$ (as $\boldsymbol{w}^* \in \text{argmin}_{\boldsymbol{w}\in\mathcal{W}} f_i(\boldsymbol{w})\ \forall\ i \in [n]$, per their assumption). One can expect Assumption 5.5 to hold for *separable* classification problems even without over-parameterization.

Under Assumption 5.5, the dominant term in the risk bound of Theorem 5.4 (Eq. (5)) would be (A). As discussed previously, (A) is a non-decreasing function of τ. Thus, the lowest risk bound in Equation (5) is obtained for $\tau \in (0, G_1]$. Also recall that $\alpha(\tau) = \alpha(G_1)\ \forall\ \tau \in (0, G_1]$ (see (iii) in Definition 5.3). Since $\alpha(G_n) = 1$, there is an $\alpha(G_1)$-fold improvement in the risk bound with $\tau \leq G_1$ compared to the naive choice of $\tau = G_n$ when $\Delta(\boldsymbol{w}^*) = 0$.

Remark 5.6 (**Recommendation**). For settings where interpolation holds (such as over-parameterization), we recommend tuning the clip norm only till values up to the minimum per-sample Lipschitz constant.

We assume that we have some prior estimate of the minimum Lipschitz constant, just like prior works on uniform Lipschitzness assume that the maximum Lipschitz constant is known. For e.g., it can be estimated privately by applying the Report Noisy Max method (Dwork et al., 2014) on the negative Lipschitz constants.

5.2. Empirical Results

We consider the base problem of private multinomial logistic regression (a convex problem satisfying Asmp. 5.1) to corroborate our theory and recommendation. But we focus on the more powerful application of privately training (only) the last (softmax) layer of deep networks pre-trained on public data which is equivalent to performing logistic regression with features being the previous layer's outputs. Training a softmax layer over pre-trained networks is a popular and effective scheme for private training (De et al., 2022; Mehta et al., 2022). This is because fine-tuning all the layers may not significantly improve performance but the extra parameters increase the privacy cost. In addition, training only the last layer is computationally much cheaper than fine-tuning all the layers with DP-SGD (due to per-sample clipping).

Our experiments here are conducted on four vision datasets available in Torchvision, viz., Caltech-256 with 257 classes, Food-101 with 101 classes, CIFAR-100 with 100 classes and CIFAR-10 with 10 classes, and two language datasets available in Hugging Face, viz., TweetEval Emoji with 20 classes and Emotion with 6 classes. For Caltech-256 and Food-101 (resp., CIFAR-100 and CIFAR-10), we use 512-dimensional features obtained from the pre-softmax layer of a pre-trained ResNet-34 (resp., ResNet-18) model on ImageNet for logistic regression, which is equivalent to training the last (i.e., softmax) layer of a pre-trained ResNet-34 (resp., ResNet-18) model. For the language datasets, we use 768-dimensional features obtained from a pre-trained DistilBERT model (Sanh et al., 2019) for logistic regression, which is the same as training a linear layer on top of a pre-trained DistilBERT model. As mentioned after Assumption 5.1, the per-sample Lipschitz constant is equal to $\sqrt{2}$ times the norm of the sample's feature vector (with a '1' appended to incorporate the bias term). We consider three privacy levels - $(2, 10^{-5})$-DP, $(4, 10^{-5})$-DP and $(6, 10^{-5})$-DP, with batch size = 500. We test several values of the clip norm τ, viz., the 0th, 10th, 20th, 40th, 80th and 100th percentile of the per-sample Lipschitz constants (as well as some values smaller than the 0th percentile). Note that G_1 and G_n correspond to the 0th and 100th percentiles, respectively. For each value of τ, we tune over several values of the constant learning rate η, viz., $\{0.0001, 0.0003, 0.0006, 0.001, 0.003, 0.006, 0.01, 0.03, 0.06, 0.1, 0.3, 0.6, 1, 3, 6, 10\}$. PyTorch's Opacus library (Yousefpour et al., 2021) is used for private training.

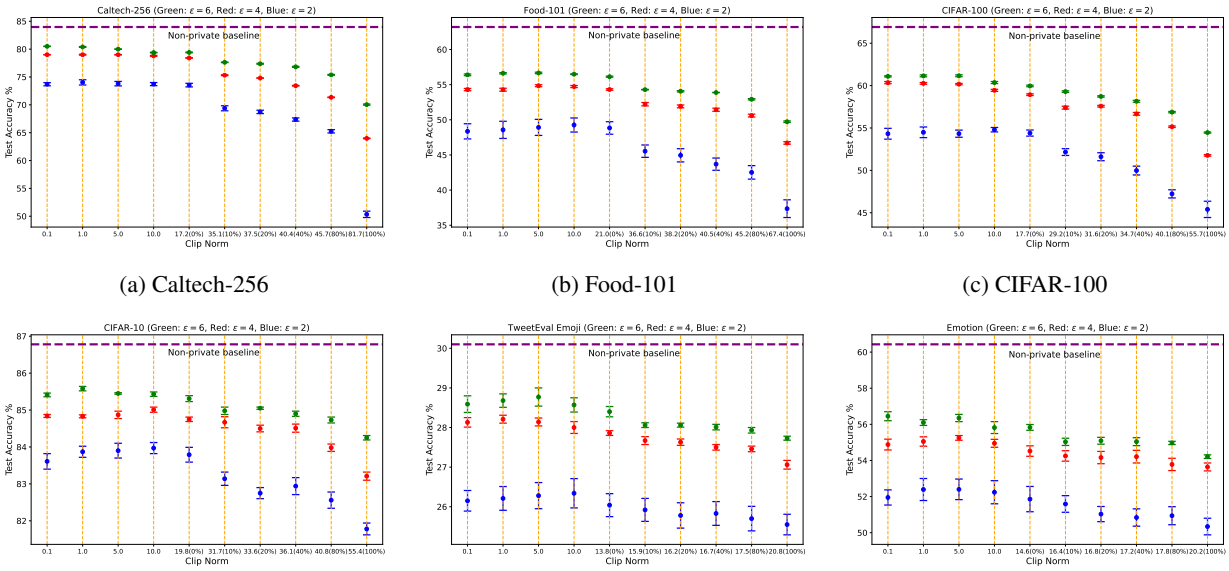
In Figure 1, we plot the best test accuracy obtained for different values of τ (by tuning over η) averaged over the last 5 epochs and across 3 different runs. The figure caption discusses the results. The exact accuracy values are listed in Appendix B. Also, in Appendix A, we show empirical results on two more datasets, viz., EMNIST and Fashion-MNIST.

Now that we have corroborated our insight from Section 5.1 that clip norms $\leq G_1$ yield the best performance, we shall show some results when G_1 is first estimated *privately* using the Report Noisy Max method (applied on the negative Lipschitz constants) and then used in DP-SGD as the clip norm. We abbreviate the Report Noisy Max method as RNMM subsequently. We show results for the first three datasets in Figure 1, viz., Caltech-256, Food-101 and CIFAR-100. We set $\varepsilon = 0.3$ for RNMM and $\varepsilon = \{1.7, 3.7, 5.7\}$ for DP-SGD so that the *total* privacy budget is $\varepsilon = \{2, 4, 6\}$ for which we showed results earlier *without* RNMM (where we did *not* estimate G_1 privately).[5] Table 3 compares the results with and without RNMM. The table caption discusses the results.

6. Convergence of DP-SGD under Heavy-Tailed Lipschitz Constants

When uniform Lipschitzness holds, the optimization risk is $\mathcal{O}(\varphi)$ in the convex constrained case (Bassily et al., 2014) as well as the smooth nonconvex unconstrained case (Wang et al., 2018). In this section, we shall quantify the optimiza-

[5] We did not optimize the splitting of privacy budget between RNMM and DP-SGD; this is something that can be investigated in future work.



(a) Caltech-256 (b) Food-101 (c) CIFAR-100

(d) CIFAR-10 (e) TweetEval Emoji (f) Emotion

Figure 1. Average test accuracy (depicted by the blobs) $\pm$ 1 standard deviation (depicted by the bars around the blobs) in the last 5 epochs for different values of clip norm τ. "%" in the x-axis stands for percentile. Observe that *clip norms* $\leq G_1$ *(0th percentile or minimum of per-sample Lipschitz constants) generally perform better than clip norms* $> G_1$. Specifically, the performance with $\tau = G_1$ is significantly better than that with $\tau = G_n$ (100th percentile or maximum). Concretely, for Caltech-256 and Food-101 (datasets with the largest no. of classes), in the case of $\varepsilon = 2$, the mean accuracy with $\tau = G_1$ is better than that with $\tau = G_n$ by $> 23\%$ and 11%, respectively; the corresponding improvement in the case of $\varepsilon = 4$ is $> 15\%$ and 7.5%, respectively. These observations are consistent with our theory and recommendation in Section 5.1.

Table 3. Average test accuracy $\pm$ 1 standard deviation in the last 5 epochs (i) *with* and (ii) *without* RNMM as described in the last paragraph of Section 5.2; the latter is the best accuracy (w.r.t. the clip norm) among the results in Figure 1 (which are without RNMM). Note that the difference in accuracy with and without RNMM *reduces as* ε *increases*. Specifically, for $\varepsilon = \{4, 6\}$, RNMM does not harm performance too much.

Dataset	$\varepsilon = 2$ **w/** RNMM	$\varepsilon = 2$ **w/o** RNMM	$\varepsilon = 4$ **w/** RNMM	$\varepsilon = 4$ **w/o** RNMM	$\varepsilon = 6$ **w/** RNMM	$\varepsilon = 6$ **w/o** RNMM
Caltech-256	70.03 ± 0.66	74.03 ± 0.47	78.27 ± 0.05	79.00 ± 0.04	79.53 ± 0.07	80.50 ± 0.04
Food-101	45.84 ± 2.00	49.26 ± 1.00	54.69 ± 0.14	54.85 ± 0.14	56.30 ± 0.12	56.67 ± 0.10
CIFAR-100	52.89 ± 1.02	55.10 ± 0.26	60.19 ± 0.09	60.35 ± 0.13	60.89 ± 0.04	61.16 ± 0.11

tion risk under generalized Lipschitzness (Assumption 4.1) with the relaxation of the per-sample Lipschitz constants being *heavy-tailed*, instead of being bounded; we **do not assume interpolation** here. More specifically, we assume that the per-sample Lipschitz constants have bounded k^{th} uncentered moment, for some $k > 1$, w.r.t. the distribution $\mathcal{D}$. This is formally stated next.

Assumption 6.1 (Bounded k^{th} Moment). Suppose Assumption 4.1 holds. For some $k > 1$ and $G > 0$, we have:

$$\left(\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\left(G(\boldsymbol{x},y)\right)^k\right]\right)^{1/k} \leq G.$$

Let us revisit the logistic regression example that we discussed after Assumption 4.1 to see why Assumption 6.1 is weaker than uniform Lipschitzness. We saw that Assumption 4.1 holds with $G(\boldsymbol{x}, y) = \sqrt{2}\|\boldsymbol{x}\|$ here. Now, if the feature distribution $\mathcal{F}$ is such that its support includes *unbounded* vectors but $\mathbb{E}_{\mathcal{F}}[\|\boldsymbol{x}\|^p] \leq G_{\mathcal{F}}^p < \infty$ for some $p > 1$, then Assumption 6.1 holds here for $k = p$ and $G = \sqrt{2}G_{\mathcal{F}}$ but uniform Lipschitzness does not hold. However, if the support of $\mathcal{F}$ includes only *bounded* vectors, then both Assumption 6.1 and uniform Lipschitzness hold. Also note that uniform Lipschitzness is a special case of Assumption 6.1 with $k = \infty$ and a finite G.

Assumption 6.1[6] is similar to the bounded moment assump-

[6]This is of significance even in non-private optimization; for e.g., Zhang et al. (2020) show that the gradients of Transformer

tion made in Wang et al. (2020); Kamath et al. (2021); Lowy & Razaviyayn (2023) for private *stochastic convex optimization* (SCO). But Wang et al. (2020); Kamath et al. (2021) assume coordinate-wise bounded moments which we do not, Wang et al. (2020) only consider $k = 2$ and Kamath et al. (2021) assume bounded *centered* (around the mean) moment. All these papers provide results for the convex case *within a bounded convex set* (i.e., $\mathcal{W}$ is bounded); we shall focus on the **unconstrained** (i.e., $\mathcal{W} = \mathbb{R}^d$) **convex** case here, which is *more practical and harder* (w.r.t. analysis). For completeness, we include a result for ERM in the *constrained* convex case in Appendix I which matches the bound of Kamath et al. (2021) for SCO. Moreover, we also present a result for the **unconstrained smooth nonconvex** case; these works do not provide any results for the smooth nonconvex case.

Before we present our results for the *unconstrained convex* case, let us first discuss the main technical difficulty compared to the *constrained convex* case. The overall optimization bias in the convex case depends on the bias in mean gradient estimation induced due to clipping as well as on the distance of the current point ($\boldsymbol{w}_t$) from the optimal point ($\boldsymbol{w}^*$), i.e., $\|\boldsymbol{w}_t - \boldsymbol{w}^*\|$ at iteration t. In the constrained convex case, $\|\boldsymbol{w}_t - \boldsymbol{w}^*\|$ can be easily bounded by the diameter of the constraint set. However, in the unconstrained case, bounding $\|\boldsymbol{w}_t - \boldsymbol{w}^*\|$ needs extra work; surprisingly, this has not been analyzed in prior work on DP-SGD. Now, we present our result under Assumption 6.1 in the *unconstrained convex* case.

Theorem 6.2 (**Unconstrained Convex Case**). *Suppose Assumption 6.1 holds, f is convex and $\mathcal{W} = \mathbb{R}^d$. Fix some $\gamma \in (0, 1)$ and $C > 0$. In Algorithm 1, set $T = \varphi^{-2}$, $\tau = \mathcal{O}\big(G\gamma^{-\frac{1}{k}}\varphi^{-\frac{2}{k+1}}\big)$ and $\eta_t = \mathcal{O}\big(C\gamma^{\frac{1}{k}}G^{-1}\varphi^{1+\frac{2}{k+1}}\big)$ for all $t < T$. Then with a probability of at least $(1 - \gamma)$ which is w.r.t. the random dataset $\mathcal{Z}$ that we obtain, DP-SGD (Algorithm 1) has the following guarantee:*

$$\mathrm{OR}(T) \leq \mathcal{O}\left(\frac{G}{\gamma^{1/k}}\left(\frac{\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2}{C} + C\right)\right)\varphi^{(1-\frac{2}{k+1})}.$$

Remark 6.3 (**Comparison with prior results**). As per Thm. 6.2, the optimization risk is $\mathcal{O}(\varphi^{1-\frac{2}{k+1}})$ in the the bounded k^{th} moment **unconstrained** convex case. In comparison, the risk is $\mathcal{O}(\varphi^{1-\frac{1}{k}})$ in the bounded k^{th} moment **constrained** convex case as per Thm. I.1 in Appendix I, and $\mathcal{O}(\varphi)$ in the uniform Lipschitz convex case (Bassily et al., 2014).

The difference in the risk bound between the constrained and unconstrained cases arises because $\|\boldsymbol{w}_t - \boldsymbol{w}^*\| = \mathcal{O}(1)$ (w.r.t. φ) in the constrained case but our bound for $\|\boldsymbol{w}_t - \boldsymbol{w}^*\|$ in the unconstrained case depends on φ. If one is able to improve this bound for the unconstrained case, then the risk for the unconstrained case will also improve; but doing so in the absence of any other assumption seems difficult.

models such as BERT are heavy-tailed.

Under a mild additional assumption, we are able to improve the risk bound to $\mathcal{O}(\varphi^{1-\frac{1}{k}})$ in the unconstrained case, thereby matching the result in the constrained case. We present this extra assumption first, followed by the result.

Assumption 6.4. For any $\boldsymbol{w}$ s.t. $\|\boldsymbol{w} - \boldsymbol{w}^*\| > D = \mathcal{O}(1)$, where $\boldsymbol{w}^* \in \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w})$:

$$f(\boldsymbol{w}) - f(\boldsymbol{w}^*) > \Big(16\varphi^{1-\frac{1}{k}}G\Big)\|\boldsymbol{w} - \boldsymbol{w}^*\|. \tag{6}$$

For large n (our regime), φ is small; thus, $16\varphi^{(1-1/k)}G$ is also small in comparison to the average Lipschitz constant $= \mathcal{O}(G)$ and so, assuming Equation (6) *for points far away from the optimum* is reasonable. Besides, an assumption similar to Assumption 6.4, known as *sharpness* (or the Łojasiewicz inequality), has been used in prior optimization literature (Polyak, 1979; Burke & Ferris, 1993; Bolte et al., 2017; Roulet & d'Aspremont, 2017; Davis et al., 2018). We provide an example for Assumption 6.4 in Appendix E.

Theorem 6.5 (**Unconstrained Convex Case Under Assumption 6.4**). *Suppose Assumptions 6.1 and 6.4 hold, f is convex and $\mathcal{W} = \mathbb{R}^d$. Fix some $C > 0$. In Algorithm 1, set $T = \varphi^{-2}$, $\tau = \mathcal{O}\big(G\varphi^{-\frac{1}{k}}\big)$ and $\eta_t = \mathcal{O}\big(CG^{-1}\varphi^{1+\frac{1}{k}}\big)$ for all $t < T$. Then with a probability of at least $3/4$ which is w.r.t. the random dataset $\mathcal{Z}$ that we obtain, DP-SGD (Alg. 1) has the following **improved** guarantee:*

$$\mathrm{OR}(T) \leq \mathcal{O}\left(G\left(\frac{\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2}{C} + C + D\right)\right)\varphi^{(1-\frac{1}{k})}.$$

Thus, the risk bound under Assumption 6.4 improves to $\mathcal{O}(\varphi^{1-\frac{1}{k}})$, which matches the bound in the constrained case. One caveat of the above result is that unlike Theorem 6.2, the confidence estimate of the high-probability result cannot be controlled by us. Next, we provide a matching lower bound showing the tightness of Theorem 6.5 w.r.t. φ.

Theorem 6.6 (**Lower Bound for Unconstrained Convex Case Under Assumption 6.4**). *Suppose $\delta < \exp(-\varepsilon^2)$. There exists a convex loss function ℓ, such that for every (ε, δ)-DP algorithm $\mathcal{A}$ which tries to solve for $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w})$ where f is the average loss for a dataset $\mathcal{Z}$ of n samples drawn from the data distribution $\mathcal{D}$, there exists a choice of $\mathcal{D}$ such that:*

- *f satisfies Assumptions 6.1 and 6.4 (the latter up to constant terms and w.h.p. w.r.t. $\mathcal{Z}$).*
- *$\mathbb{E}_{\mathcal{Z}\sim\mathcal{D}^n,\mathcal{A}}\Big[f(\boldsymbol{w}_{\mathcal{Z}}^{(\mathcal{A})}) - f(\boldsymbol{w}^*)\Big] \geq \Omega\big(\varphi^{1-\frac{1}{k}}\big)$, where $\boldsymbol{w}_{\mathcal{Z}}^{(\mathcal{A})}$ is the output of algorithm $\mathcal{A}$ on the dataset $\mathcal{Z}$.*

Thus, **DP-SGD is optimal** in the unconstrained convex case under Assumptions 6.1 & 6.4 (for $\delta < \exp(-\varepsilon^2)$). We also have the following corollary under uniform Lipschitzness.

Corollary 6.7 (**Uniform Lipschitz**). *For $k \to \infty$, i.e., under uniform Lipschitzness, the risk in the unconstrained convex case regardless of Assumption 6.4 is $\mathcal{O}(\varphi)$ as per Thm. 6.2 and 6.5. Further, Thm. 6.6 is a lower bound even w/o Assumption 6.4, and it yields $\Omega(\varphi)$ for $k \to \infty$. Thus,* ***DP-SGD is optimal*** *in the unconstrained convex case under uniform Lipschitzness (for $\delta < \exp(-\varepsilon^2)$).*

Regarding the proof of Theorem 6.6: Even though we follow the proof outline of Theorem 6.4 of Kamath et al. (2021), our proof is more involved. First, since we are in the *unconstrained* setting, we had to use a *non-obvious* function (to obtain the lower bound on). Second, since we are in the ERM setting, we have to lower bound the expected training error which is harder than lower bounding the expected generalization error in the SCO setting of Kamath et al. (2021). Finally, since we are dealing with (ε, δ)-DP (whereas Kamath et al. (2021) deal with $(0, \rho)$-zCDP), we had to re-derive an important tool in their analysis. That is also the reason we had to impose $\delta < \exp(-\varepsilon^2)$; removing this constraint is left for future work. Note that this is the first lower bound for (ε, δ)-DP in the unconstrained case. See Appendix L for details.

Finally, we present our result under Assumption 6.1 in the *unconstrained nonconvex* case.

Theorem 6.8 (**Unconstrained Nonconvex Case**). *Suppose Assumption 6.1 holds, f is L-smooth and $\mathcal{W} = \mathbb{R}^d$. Fix some $\gamma \in (0, 1)$. In Algorithm 1, set $T = \varphi^{-2}$, $\tau = \mathcal{O}\big(\gamma^{-\frac{2}{2k-1}} \varphi^{-\frac{1}{2k-1}}\big)$ and $\eta_t = \mathcal{O}\big(\gamma^{\frac{2}{2k-1}} \varphi^{1+\frac{1}{2k-1}}\big)$ for all $t < T$. Then with a probability of at least $(1-\gamma)$ which is w.r.t. the random dataset $\mathcal{Z}$ that we obtain, DP-SGD (Alg. 1) has the following guarantee:*

$$\mathrm{OR}(T) \leq \mathcal{O}\Big(\gamma^{\frac{-2}{2k-1}} \varphi^{(1-\frac{1}{2k-1})}\Big).$$

Remark 6.9 (**Comparison with Prior Results**). As per Thm. 6.8, the risk is $\mathcal{O}(\varphi^{1-\frac{1}{2k-1}})$ in the bounded k^{th} moment nonconvex case. In comparison, Wang et al. (2018; 2019) get a bound of $\mathcal{O}(\varphi)$ in the uniform **Lipschitz** nonconvex case (equivalent to $k = \infty$) with DP-SGD like algorithms.

7. Limitations

We now discuss some limitations of our work which render interesting directions of future work. We do not have a lower bound for the nonconvex case in Section 6. Our clip norm selection strategy in Section 5 is only for the convex case. Our iteration complexity (T) is worse than some related works because we consider the vanilla DP-SGD algorithm of Abadi et al. (2016) *without any acceleration*, whereas other works consider accelerated variants of DP-SGD. In our case, we had to set $T = \mathcal{O}(1/\varphi^2)$ in order to control the clipping bias. But please note that the goal of this work is to attain smaller optimization risk by properly choosing the clip norm, and *not* accelerate convergence.

8. Conclusion

In this paper, we generalized uniform Lipschitzness by assuming that the per-sample gradients have sample-dependent upper bounds called per-sample Lipschitz constants. Under our generalized Lipschitzness assumption, we provided a theoretically-motivated clip norm tuning recommendation for DP-SGD in the convex case. We showed the effectiveness of our recommendation via extensive experiments. Additionally, we derived novel convergence results for unconstrained DP-SGD when the per-sample Lipschitz constants are heavy-tailed, i.e., they have bounded moments.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Acharya, J., Sun, Z., and Zhang, H. Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pp. 48–78. PMLR, 2021.

Andrew, G., Thakkar, O., McMahan, H. B., and Ramaswamy, S. Differentially private learning with adaptive clipping. *arXiv preprint arXiv:1905.03871*, 2019.

Arora, R., Bassily, R., González, T., Guzmán, C., Menart, M., and Ullah, E. Faster rates of convergence to stationary points in differentially private optimization. *arXiv preprint arXiv:2206.00846*, 2022.

Asi, H., Duchi, J., Fallah, A., Javidbakht, O., and Talwar, K. Private adaptive gradient methods for convex optimization. In *International Conference on Machine Learning*, pp. 383–392. PMLR, 2021a.

Asi, H., Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: Optimal rates in l1 geometry. *arXiv preprint arXiv:2103.01516*, 2021b.

Barbieri, F., Camacho-Collados, J., Ronzano, F., Espinosa-Anke, L., Ballesteros, M., Basile, V., Patti, V., and Saggion, H. Semeval 2018 task 2: Multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 24–33, 2018.

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

Bassily, R., Feldman, V., Talwar, K., and Thakurta, A. Private stochastic convex optimization with optimal rates. *arXiv preprint arXiv:1908.09970*, 2019.

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

Bolte, J., Nguyen, T. P., Peypouquet, J., and Suter, B. W. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Bu, Z., Wang, H., Long, Q., and Su, W. J. On the convergence of deep learning with differential privacy. *arXiv preprint arXiv:2106.07830*, 2021.

Bun, M. and Steinke, T. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pp. 635–658. Springer, 2016.

Burke, J. V. and Ferris, M. C. Weak sharp minima in mathematical programming. *SIAM Journal on Control and Optimization*, 31(5):1340–1359, 1993.

Chaudhuri, K. and Monteleoni, C. Privacy-preserving logistic regression. In *NIPS*, volume 8, pp. 289–296. Citeseer, 2008.

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

Chen, X., Wu, S. Z., and Hong, M. Understanding gradient clipping in private sgd: A geometric perspective. *Advances in Neural Information Processing Systems*, 33, 2020.

Davis, D., Drusvyatskiy, D., MacPhee, K. J., and Paquette, C. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179(3):962–982, 2018.

De, S., Berrada, L., Hayes, J., Smith, S. L., and Balle, B. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

Du, J., Li, S., Feng, M., and Chen, S. Dynamic differential-privacy preserving sgd. *arXiv preprint arXiv:2111.00173*, 2021.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438. IEEE, 2013.

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

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

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

Griffin, G., Holub, A., and Perona, P. Caltech-256 object category dataset. 2007.

Hu, L., Ni, S., Xiao, H., and Wang, D. High dimensional differentially private stochastic optimization with heavy-tailed data. *arXiv preprint arXiv:2107.11136*, 2021.

Iyengar, R., Near, J. P., Song, D., Thakkar, O., Thakurta, A., and Wang, L. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 299–316. IEEE, 2019.

Kamath, G., Liu, X., and Zhang, H. Improved rates for differentially private stochastic convex optimization with heavy-tailed data. *arXiv preprint arXiv:2106.01336*, 2021.

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

Kulkarni, J., Lee, Y. T., and Liu, D. Private nonsmooth empirical risk minimization and stochastic convex optimization in subquadratic steps. *arXiv preprint arXiv:2103.15352*, 2021.

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

Ma, S., Bassily, R., and Belkin, M. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pp. 3325–3334. PMLR, 2018.

Mehta, H., Thakurta, A., Kurakin, A., and Cutkosky, A. Large scale transfer learning for differentially private

image classification. *arXiv preprint arXiv:2205.02973*, 2022.

Papernot, N. and Steinke, T. Hyperparameter tuning with renyi differential privacy. In *International Conference on Learning Representations*, 2021.

Papernot, N., Thakurta, A., Song, S., Chien, S., and Erlingsson, U. Tempered sigmoid activations for deep learning with differential privacy. *arXiv preprint arXiv:2007.14191*, pp. 10, 2020.

Polyak, B. T. Sharp minima. institute of control sciences lecture notes, moscow, ussr, 1979. In *IIASA workshop on generalized Lagrangians and their applications, IIASA, Laxenburg, Austria*, 1979.

Roulet, V. and d'Aspremont, A. Sharpness, restart and acceleration. *Advances in Neural Information Processing Systems*, 30, 2017.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., and Chen, Y.-S. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3687–3697, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1404. URL https://www.aclweb.org/anthology/D18-1404.

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

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

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

Talwar, K., Guha Thakurta, A., and Zhang, L. Nearly optimal private lasso. *Advances in Neural Information Processing Systems*, 28:3025–3033, 2015.

Tran, H. and Cutkosky, A. Momentum aggregation for private non-convex erm. *arXiv preprint arXiv:2210.06328*, 2022.

Wang, D., Ye, M., and Xu, J. Differentially private empirical risk minimization revisited: Faster and more general. *arXiv preprint arXiv:1802.05251*, 2018.

Wang, D., Xiao, H., Devadas, S., and Xu, J. On differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pp. 10081–10091. PMLR, 2020.

Wang, L., Jayaraman, B., Evans, D., and Gu, Q. Efficient privacy-preserving stochastic nonconvex optimization. *arXiv preprint arXiv:1910.13659*, 2019.

Wu, X., Li, F., Kumar, A., Chaudhuri, K., Jha, S., and Naughton, J. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pp. 1307–1322, 2017.

Wu, X., Wang, L., Cristali, I., Gu, Q., and Willett, R. Adaptive differentially private empirical risk minimization. *arXiv preprint arXiv:2110.07435*, 2021.

Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., Cormode, G., and Mironov, I. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

Zhang, J., Zheng, K., Mou, W., and Wang, L. Efficient private erm for smooth objectives. *arXiv preprint arXiv:1703.09947*, 2017.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

Zhang, X., Chen, X., Hong, M., Wu, Z. S., and Yi, J. Understanding clipping for federated learning: Convergence and client-level differential privacy. *arXiv preprint arXiv:2106.13673*, 2021.

Appendix

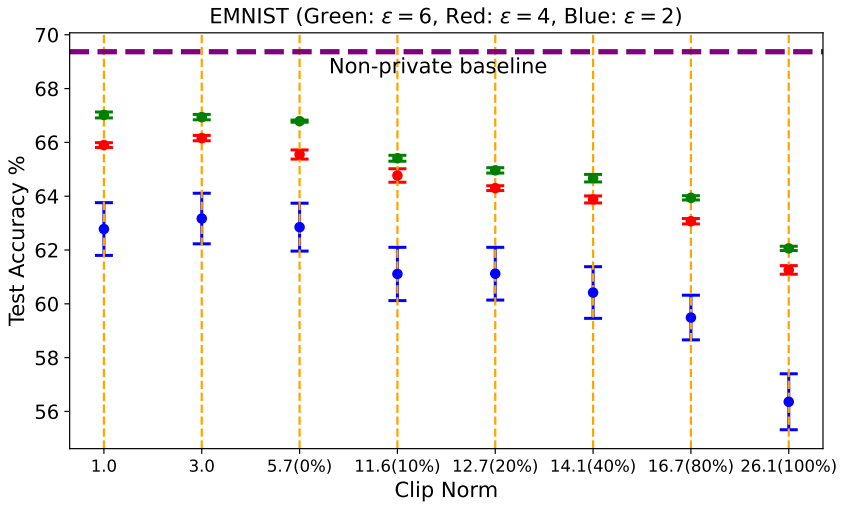
Contents

- Appendix A: Empirical Results on EMNIST and Fashion-MNIST
- Appendix B: Tables of Results for Section 5.2 and Appendix A
- Appendix C: Some Empirical Results in the Non-Convex Case
- Appendix D: Logistic Regression Satisfies Assumption 4.1
- Appendix E: Example for Assumption 6.4
- Appendix F: Some Useful Lemmas
- Appendix G: Proof of Theorem 3.4
- Appendix H: Full Version and Proof of Theorem 5.4
- Appendix I: Result for the Constrained Convex Case under Assumption 6.1
- Appendix J: Full Version and Proof of Theorem 6.2
- Appendix K: Full Version and Proof of Theorem 6.5
- Appendix L: Full Version and Proof of Theorem 6.6
- Appendix M: Proof of Theorem I.3
- Appendix N: Full Version and Proof of Theorem 6.8

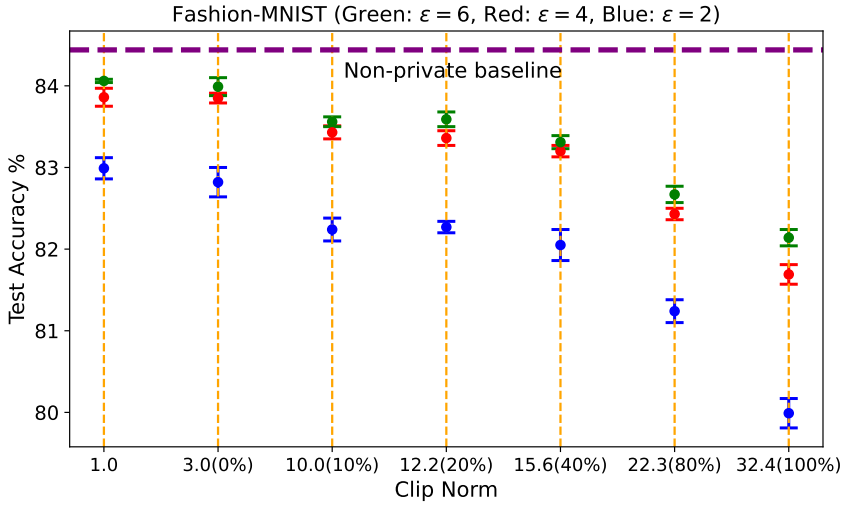
A. Empirical Results on EMNIST and Fashion-MNIST

Here, we show results for private multinomial logistic regression on two more datasets, viz., (balanced) EMNIST with 47 classes and Fashion-MNIST with 10 classes. For both these datasets, we just use the flattened images as features for logistic regression. Other experimental details are the same as in Section 5.2.

In Figure 2, we plot the best test accuracy obtained for different values of the clip norm τ (by tuning over the learning rate η) averaged over the last 5 epochs and across 3 independent runs (just like Figure 1). The figure caption discusses the results. The exact accuracy values for these two datasets are tabulated in Table 6 in Appendix B.



(a) EMNIST



(b) Fashion-MNIST

Figure 2. Average test accuracy (depicted by the blobs) $\pm$ 1 standard deviation (depicted by the bars around the blobs) in the last 5 epochs for different values of clip norm τ. "%" in the x-axis stands for percentile. Just as in Figure 1, *clip norms* $\leq G_1$ *(*0th *percentile of per-sample Lipschitz constants) perform better than clip norms* $> G_1$*, and the performance with* $\tau = G_1$ *is much better than that with* $\tau = G_n$ *(*100th *percentile).* These observations validate our theory and recommendation in Section 5.1.

B. Tables of Results for Section 5.2 and Appendix A

Table 4. **Caltech-256, Food-101 and CIFAR-100:** Average test accuracy $\pm$ 1 standard deviation in the last 5 epochs for different values of clip norm τ in the experiments of Section 5.2. Note that "pctl." stands for percentile. "Non-private baseline" is the accuracy of vanilla non-private SGD in the same setting.

Caltech-256	**(a)** $(2, 10^{-5})$-DP	**(b)** $(4, 10^{-5})$-DP	**(c)** $(6, 10^{-5})$-DP
$\tau = 0.1$	73.70 $\pm$ 0.28 %	79.00 $\pm$ 0.03 %	80.50 $\pm$ 0.04 %
$\tau = 1.0$	74.03 $\pm$ 0.47 %	79.00 $\pm$ 0.08 %	80.39 $\pm$ 0.07 %
$\tau = 5.0$	73.80 $\pm$ 0.40 %	79.00 $\pm$ 0.04 %	80.00 $\pm$ 0.04 %
$\tau = 10.0$	73.72 $\pm$ 0.28 %	78.79 $\pm$ 0.08 %	79.38 $\pm$ 0.05 %
$\tau = (0^{\text{th}}$ pctl.)	73.54 $\pm$ 0.32 %	78.42 $\pm$ 0.05 %	79.41 $\pm$ 0.06 %
$\tau = (10^{\text{th}}$ pctl.)	69.38 $\pm$ 0.45 %	75.32 $\pm$ 0.08 %	77.63 $\pm$ 0.05 %
$\tau = (20^{\text{th}}$ pctl.)	68.73 $\pm$ 0.30 %	74.81 $\pm$ 0.02 %	77.37 $\pm$ 0.08 %
$\tau = (40^{\text{th}}$ pctl.)	67.36 $\pm$ 0.29 %	73.43 $\pm$ 0.06 %	76.82 $\pm$ 0.09 %
$\tau = (80^{\text{th}}$ pctl.)	65.25 $\pm$ 0.30 %	71.35 $\pm$ 0.05 %	75.38 $\pm$ 0.08 %
$\tau = (100^{\text{th}}$ pctl.)	50.34 $\pm$ 0.57 %	63.98 $\pm$ 0.07 %	70.04 $\pm$ 0.10 %
Non-private baseline	83.97 $\pm$ 0.02 %		

Food-101	**(a)** $(2, 10^{-5})$-DP	**(b)** $(4, 10^{-5})$-DP	**(c)** $(6, 10^{-5})$-DP
$\tau = 0.1$	48.36 $\pm$ 1.08 %	54.30 $\pm$ 0.16 %	56.40 $\pm$ 0.12 %
$\tau = 1.0$	48.57 $\pm$ 1.22 %	54.28 $\pm$ 0.21 %	56.62 $\pm$ 0.09 %
$\tau = 5.0$	48.92 $\pm$ 1.14 %	54.85 $\pm$ 0.14 %	56.67 $\pm$ 0.10 %
$\tau = 10.0$	49.26 $\pm$ 1.00 %	54.73 $\pm$ 0.15 %	56.49 $\pm$ 0.05 %
$\tau = (0^{\text{th}}$ pctl.)	48.84 $\pm$ 0.89 %	54.31 $\pm$ 0.11 %	56.13 $\pm$ 0.08 %
$\tau = (10^{\text{th}}$ pctl.)	45.54 $\pm$ 0.88 %	52.22 $\pm$ 0.24 %	54.29 $\pm$ 0.04 %
$\tau = (20^{\text{th}}$ pctl.)	44.95 $\pm$ 0.94 %	51.91 $\pm$ 0.20 %	54.06 $\pm$ 0.09 %
$\tau = (40^{\text{th}}$ pctl.)	43.69 $\pm$ 0.87 %	51.43 $\pm$ 0.22 %	53.88 $\pm$ 0.02 %
$\tau = (80^{\text{th}}$ pctl.)	42.52 $\pm$ 0.96 %	50.60 $\pm$ 0.20 %	52.92 $\pm$ 0.09 %
$\tau = (100^{\text{th}}$ pctl.)	37.36 $\pm$ 1.26 %	46.71 $\pm$ 0.18 %	49.73 $\pm$ 0.11 %
Non-private baseline	63.20 $\pm$ 0.06 %		

CIFAR-100	**(a)** $(2, 10^{-5})$-DP	**(b)** $(4, 10^{-5})$-DP	**(c)** $(6, 10^{-5})$-DP
$\tau = 0.1$	54.32 $\pm$ 0.64 %	60.35 $\pm$ 0.13 %	61.09 $\pm$ 0.07 %
$\tau = 1.0$	54.49 $\pm$ 0.63 %	60.26 $\pm$ 0.10 %	61.15 $\pm$ 0.11 %
$\tau = 5.0$	54.33 $\pm$ 0.42 %	60.17 $\pm$ 0.08 %	61.16 $\pm$ 0.11 %
$\tau = 10.0$	55.10 $\pm$ 0.26 %	59.44 $\pm$ 0.12 %	60.36 $\pm$ 0.12 %
$\tau = 17.7(0^{\text{th}}$ pctl.)	54.40 $\pm$ 0.36 %	58.93 $\pm$ 0.10 %	59.96 $\pm$ 0.09 %
$\tau = 29.2(10^{\text{th}}$ pctl.)	52.16 $\pm$ 0.40 %	57.40 $\pm$ 0.17 %	59.89 $\pm$ 0.09 %
$\tau = 31.6(20^{\text{th}}$ pctl.)	51.61 $\pm$ 0.47 %	57.57 $\pm$ 0.10 %	58.72 $\pm$ 0.08 %
$\tau = 34.7(40^{\text{th}}$ pctl.)	49.98 $\pm$ 0.52 %	56.67 $\pm$ 0.16 %	58.15 $\pm$ 0.11 %
$\tau = 40.1(80^{\text{th}}$ pctl.)	47.23 $\pm$ 0.48 %	55.15 $\pm$ 0.09 %	56.87 $\pm$ 0.08 %
$\tau = 55.7(100^{\text{th}}$ pctl.)	45.40 $\pm$ 0.96 %	51.77 $\pm$ 0.09 %	54.46 $\pm$ 0.07 %
Non-private baseline	66.91 $\pm$ 0.05 %		

Table 5. **CIFAR-10, TweetEval Emoji and Emotion:** Average test accuracy $\pm$ 1 standard deviation in the last 5 epochs for different values of clip norm τ in the experiments of Section 5.2. Note that "pctl." stands for percentile. "Non-private baseline" is the accuracy of vanilla non-private SGD in the same setting.

CIFAR-10	**(a)** $(2, 10^{-5})$-DP	**(b)** $(4, 10^{-5})$-DP	**(c)** $(6, 10^{-5})$-DP
$\tau = 0.1$	83.61 $\pm$ 0.21 %	84.84 $\pm$ 0.04 %	85.41 $\pm$ 0.05 %
$\tau = 1.0$	83.87 $\pm$ 0.15 %	84.83 $\pm$ 0.04 %	85.58 $\pm$ 0.06 %
$\tau = 5.0$	83.90 $\pm$ 0.20 %	84.87 $\pm$ 0.10 %	85.45 $\pm$ 0.02 %
$\tau = 10.0$	83.97 $\pm$ 0.15 %	85.15 $\pm$ 0.07 %	85.43 $\pm$ 0.06 %
$\tau = 19.8$(0th pctl.)	83.79 $\pm$ 0.20 %	84.75 $\pm$ 0.06 %	85.31 $\pm$ 0.08 %
$\tau = 31.7$(10th pctl.)	83.14 $\pm$ 0.18 %	84.67 $\pm$ 0.15 %	84.98 $\pm$ 0.10 %
$\tau = 33.6$(20th pctl.)	82.75 $\pm$ 0.15 %	84.50 $\pm$ 0.09 %	85.05 $\pm$ 0.03 %
$\tau = 36.1$(40th pctl.)	82.94 $\pm$ 0.23 %	84.51 $\pm$ 0.11 %	84.90 $\pm$ 0.07 %
$\tau = 40.8$(80th pctl.)	82.56 $\pm$ 0.22 %	83.98 $\pm$ 0.10 %	84.73 $\pm$ 0.08 %
$\tau = 55.4$(100th pctl.)	81.78 $\pm$ 0.16 %	83.21 $\pm$ 0.11 %	84.25 $\pm$ 0.06 %
Non-private baseline	86.78 $\pm$ 0.05 %		

TweetEval Emoji	**(a)** $(2, 10^{-5})$-DP	**(b)** $(4, 10^{-5})$-DP	**(c)** $(6, 10^{-5})$-DP
$\tau = 0.1$	26.15 $\pm$ 0.26 %	28.13 $\pm$ 0.12 %	28.59 $\pm$ 0.21 %
$\tau = 1.0$	26.21 $\pm$ 0.30 %	28.21 $\pm$ 0.10 %	28.68 $\pm$ 0.17 %
$\tau = 5.0$	26.28 $\pm$ 0.33 %	28.14 $\pm$ 0.10 %	28.77 $\pm$ 0.23 %
$\tau = 10.0$	26.34 $\pm$ 0.37 %	28.00 $\pm$ 0.15 %	28.57 $\pm$ 0.18 %
$\tau = 13.8$(0th pctl.)	26.04 $\pm$ 0.29 %	27.86 $\pm$ 0.06 %	28.40 $\pm$ 0.13 %
$\tau = 15.9$(10th pctl.)	25.92 $\pm$ 0.29 %	27.67 $\pm$ 0.10 %	28.06 $\pm$ 0.06 %
$\tau = 16.2$(20th pctl.)	25.78 $\pm$ 0.32 %	27.63 $\pm$ 0.08 %	28.06 $\pm$ 0.05 %
$\tau = 16.7$(40th pctl.)	25.83 $\pm$ 0.30 %	27.50 $\pm$ 0.07 %	28.01 $\pm$ 0.07 %
$\tau = 17.5$(80th pctl.)	25.70 $\pm$ 0.31 %	27.46 $\pm$ 0.07 %	27.93 $\pm$ 0.07 %
$\tau = 20.8$(100th pctl.)	25.55 $\pm$ 0.26 %	27.06 $\pm$ 0.11 %	27.73 $\pm$ 0.05 %
Non-private baseline	30.10 $\pm$ 0.03 %		

Emotion	**(a)** $(2, 10^{-5})$-DP	**(b)** $(4, 10^{-5})$-DP	**(c)** $(6, 10^{-5})$-DP
$\tau = 0.1$	51.95 $\pm$ 0.42 %	54.88 $\pm$ 0.30 %	56.45 $\pm$ 0.25 %
$\tau = 1.0$	52.39 $\pm$ 0.61 %	55.05 $\pm$ 0.26 %	56.10 $\pm$ 0.16 %
$\tau = 5.0$	52.40 $\pm$ 0.57 %	55.25 $\pm$ 0.13 %	56.35 $\pm$ 0.20 %
$\tau = 10.0$	52.24 $\pm$ 0.64 %	54.95 $\pm$ 0.22 %	55.82 $\pm$ 0.33 %
$\tau = 14.6$(0th pctl.)	51.86 $\pm$ 0.70 %	54.52 $\pm$ 0.29 %	55.83 $\pm$ 0.16 %
$\tau = 16.4$(10th pctl.)	51.59 $\pm$ 0.46 %	54.25 $\pm$ 0.29 %	55.03 $\pm$ 0.20 %
$\tau = 16.8$(20th pctl.)	51.03 $\pm$ 0.42 %	54.16 $\pm$ 0.34 %	55.09 $\pm$ 0.19 %
$\tau = 17.2$(40th pctl.)	50.84 $\pm$ 0.48 %	54.21 $\pm$ 0.35 %	55.04 $\pm$ 0.22 %
$\tau = 17.8$(80th pctl.)	50.94 $\pm$ 0.50 %	53.78 $\pm$ 0.34 %	54.97 $\pm$ 0.09 %
$\tau = 20.2$(100th pctl.)	50.34 $\pm$ 0.46 %	53.64 $\pm$ 0.22 %	54.21 $\pm$ 0.10 %
Non-private baseline	60.43 $\pm$ 0.22 %		

Table 6. **EMNIST and Fashion-MNIST:** Average test accuracy $\pm$ 1 standard deviation in the last 5 epochs for different values of clip norm τ in the experiments of Appendix A. Note that "pctl." stands for percentile. "Non-private baseline" is the accuracy of vanilla non-private SGD in the same setting.

EMNIST	**(a)** $(2, 10^{-5})$-DP	**(b)** $(4, 10^{-5})$-DP	**(c)** $(6, 10^{-5})$-DP
$\tau = 1.0$	62.78 ± 0.98 %	65.90 ± 0.09 %	67.02 ± 0.11 %
$\tau = 3.0$	63.17 ± 0.94 %	66.16 ± 0.10 %	66.94 ± 0.10 %
$\tau = 5.7(0^{\text{th}}$ pctl.)	62.85 ± 0.89 %	65.55 ± 0.17 %	66.79 ± 0.04 %
$\tau = 11.6(10^{\text{th}}$ pctl.)	61.11 ± 0.99 %	64.77 ± 0.25 %	65.41 ± 0.11 %
$\tau = 12.7(20^{\text{th}}$ pctl.)	61.12 ± 0.98 %	64.30 ± 0.09 %	64.96 ± 0.10 %
$\tau = 14.1(40^{\text{th}}$ pctl.)	60.42 ± 0.96 %	63.88 ± 0.13 %	64.67 ± 0.14 %
$\tau = 16.7(80^{\text{th}}$ pctl.)	59.49 ± 0.83 %	63.07 ± 0.10 %	63.94 ± 0.08 %
$\tau = 26.1(100^{\text{th}}$ pctl.)	56.36 ± 1.04 %	61.26 ± 0.16 %	62.06 ± 0.08 %
Non-private baseline	69.37 ± 0.04 %		

Fashion-MNIST	**(a)** $(2, 10^{-5})$-DP	**(b)** $(4, 10^{-5})$-DP	**(c)** $(6, 10^{-5})$-DP
$\tau = 1.0$	82.99 ± 0.13 %	83.86 ± 0.11 %	84.06 ± 0.02 %
$\tau = 3.0(0^{\text{th}}$ pctl.)	82.82 ± 0.18 %	83.85 ± 0.06 %	83.99 ± 0.11 %
$\tau = 10.0(10^{\text{th}}$ pctl.)	82.24 ± 0.14 %	83.43 ± 0.08 %	83.56 ± 0.06 %
$\tau = 12.2(20^{\text{th}}$ pctl.)	82.27 ± 0.07 %	83.36 ± 0.09 %	83.59 ± 0.09 %
$\tau = 15.6(40^{\text{th}}$ pctl.)	82.05 ± 0.19 %	83.20 ± 0.07 %	83.31 ± 0.08 %
$\tau = 22.3(80^{\text{th}}$ pctl.)	81.24 ± 0.14 %	82.43 ± 0.07 %	82.67 ± 0.10 %
$\tau = 32.4(100^{\text{th}}$ pctl.)	79.99 ± 0.18 %	81.69 ± 0.12 %	82.14 ± 0.10 %
Non-private baseline	84.44 ± 0.05 %		

C. Some Empirical Results in the Non-Convex Case

Here we show some empirical results on a nonconvex neural network (NN) problem. Specifically, we consider a two-layer feedforward NN having one hidden layer with tanh activation. We use tanh instead of ReLU activation due to two reasons: (i) Papernot et al. (2020) show that tanh performs better than ReLU in private training of NNs (which we also observed), and (ii) we expect Lipschitz constants to be smaller with tanh than ReLU. We run our experiments on CIFAR-100 and EMNIST; we use pre-trained features for the former, while for the latter, we use the raw images. Specifically, for CIFAR-100, we use 512-dimensional features obtained from the pre-softmax layer of a pre-trained ResNet-18 model on ImageNet (same as in Section 5.2). For EMNIST, we use the flattened images as features (just as mentioned in Appendix A). For both datasets, we set the dimension of the hidden layer of the NN to be 256. Computing the per-sample Lipschitz constants is much harder here so we just test several values of the clip norm τ, viz., $\{1, 3, 6, 12, 18, 24, 30, 36\}$, and show the performance trend as a function of τ. All other experimental details are the same as in Section 5.2.

In Figure 3, we plot the best test accuracy obtained for different values of τ (by tuning over η) averaged over the last 5 epochs and across 3 independent runs. The figure caption discusses the results. The exact values are tabulated in Table 7.

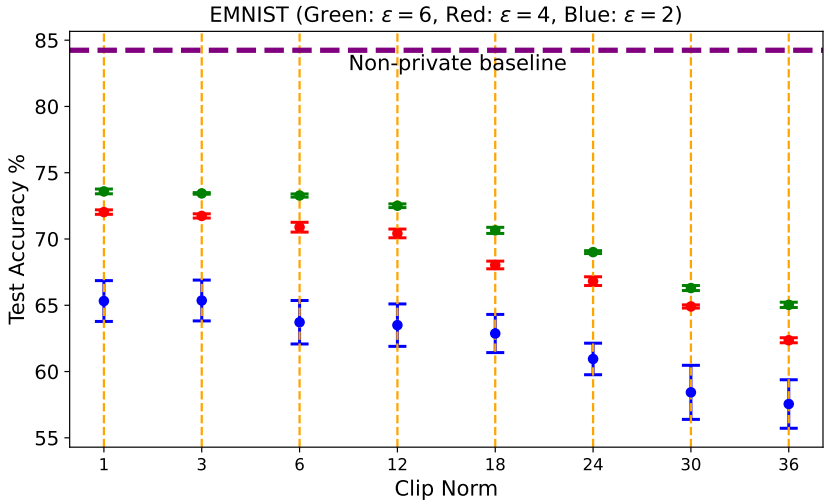
So empirically, smaller clip norms perform better in two-layer nonconvex NNs, similar to the convex case.

A single NVIDIA TITAN Xp GPU was used for all the experiments in this paper.

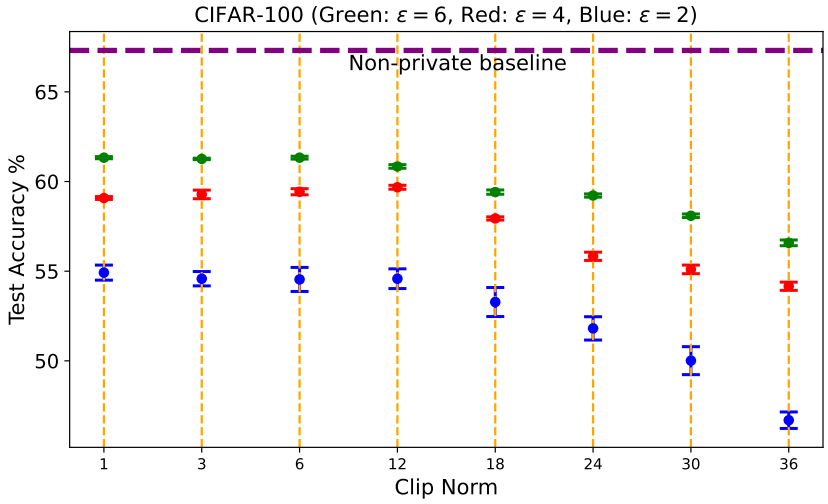
Table 7. **EMNIST and CIFAR-100 with two-layer NN:** Average test accuracy $\pm$ 1 standard deviation in the last 5 epochs for different values of clip norm τ in the experiments above. "Non-private baseline" is the accuracy of vanilla non-private SGD in the same setting.

EMNIST	**(a)** $(2, 10^{-5})$-DP	**(b)** $(4, 10^{-5})$-DP	**(c)** $(6, 10^{-5})$-DP
$\tau = 1$	65.32 $\pm$ 1.54 %	72.03 $\pm$ 0.17 %	73.59 $\pm$ 0.18 %
$\tau = 3$	65.36 $\pm$ 1.54 %	71.74 $\pm$ 0.16 %	73.44 $\pm$ 0.06 %
$\tau = 6$	63.72 $\pm$ 1.64 %	70.89 $\pm$ 0.37 %	73.28 $\pm$ 0.12 %
$\tau = 12$	63.50 $\pm$ 1.60 %	70.42 $\pm$ 0.33 %	72.51 $\pm$ 0.14 %
$\tau = 18$	62.87 $\pm$ 1.44 %	68.04 $\pm$ 0.29 %	70.65 $\pm$ 0.23 %
$\tau = 24$	60.95 $\pm$ 1.19 %	66.82 $\pm$ 0.33 %	69.01 $\pm$ 0.11 %
$\tau = 30$	58.43 $\pm$ 2.04 %	64.91 $\pm$ 0.12 %	66.30 $\pm$ 0.19 %
$\tau = 36$	57.55 $\pm$ 1.83 %	62.36 $\pm$ 0.19 %	65.03 $\pm$ 0.20 %
Non-private baseline	84.24 $\pm$ 0.05 %		

CIFAR-100	**(a)** $(2, 10^{-5})$-DP	**(b)** $(4, 10^{-5})$-DP	**(c)** $(6, 10^{-5})$-DP
$\tau = 1$	54.92 $\pm$ 0.42 %	59.08 $\pm$ 0.08 %	61.33 $\pm$ 0.06 %
$\tau = 3$	54.58 $\pm$ 0.40 %	59.28 $\pm$ 0.24 %	61.26 $\pm$ 0.04 %
$\tau = 6$	54.54 $\pm$ 0.67 %	59.43 $\pm$ 0.17 %	61.33 $\pm$ 0.08 %
$\tau = 12$	54.58 $\pm$ 0.55 %	59.48 $\pm$ 0.11 %	60.84 $\pm$ 0.10 %
$\tau = 18$	53.28 $\pm$ 0.81 %	57.94 $\pm$ 0.09 %	59.41 $\pm$ 0.12 %
$\tau = 24$	51.81 $\pm$ 0.65 %	55.83 $\pm$ 0.23 %	59.22 $\pm$ 0.09 %
$\tau = 30$	50.01 $\pm$ 0.78 %	55.10 $\pm$ 0.24 %	58.09 $\pm$ 0.10 %
$\tau = 36$	46.69 $\pm$ 0.46 %	54.16 $\pm$ 0.23 %	56.58 $\pm$ 0.16 %
Non-private baseline	67.31 $\pm$ 0.04 %		



(a) EMNIST



(b) CIFAR-100

Figure 3. **EMNIST and CIFAR-100 with two-layer NN:** Average test accuracy (depicted by the blobs) $\pm$ 1 standard deviation (depicted by the bars above and below the blobs) in the last 5 epochs for different values of clip norm τ. The general trend above is that the accuracy drops as the clip norm increases; this is similar to what we saw in the experiments on convex problems, and consistent with the main message of Theorem 5.4 (even though it is for the convex case), viz., smaller clip norms should perform better as they attain a lower risk bound.

D. Logistic Regression Satisfies Assumption 4.1

Consider doing logistic regression for multi-class classification with the cross-entropy loss, where m is the number of classes. Suppose $\boldsymbol{x} \sim \mathcal{F}$ (with a '1' appended to account for the bias term) is the feature vector and $y \in [m]$ is the corresponding class number. Let the model parameter $\boldsymbol{w}$ be split as $\boldsymbol{w} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m]$, where each $\{\boldsymbol{w}_j\}_{j=1}^m \in \mathbb{R}^d$, d being the dimension of $\boldsymbol{x}$; so, $\boldsymbol{w}_j$ denotes the parameter vector corresponding to class j. Then, our predicted probability of $\boldsymbol{x}$ belonging to class j with the softmax predictor is:

$$p_j = \frac{\exp(\boldsymbol{w}_j^T \boldsymbol{x})}{\sum_{k=1}^m \exp(\boldsymbol{w}_k^T \boldsymbol{x})}.$$

We use the standard cross-entropy loss for logistic regression which gives us:

$$\ell(\boldsymbol{w}, \boldsymbol{x}, y) = -\log(p_y). \tag{7}$$

Now, with some differentiation, it can be checked that:

$$\|\nabla \ell_{\boldsymbol{w}}(\boldsymbol{w}, \boldsymbol{x}, y)\| = \left(\sqrt{\sum_{j \neq y} p_j^2 + (1 - p_y)^2}\right) \|\boldsymbol{x}\| \leq \sqrt{2}\|\boldsymbol{x}\|. \tag{8}$$

Thus, logistic regression satisfies Assumption 4.1 with $G(\boldsymbol{x}, y) = \sqrt{2}\|\boldsymbol{x}\|$ in any parameter domain $\mathcal{W}$.

E. Example for Assumption 6.4

Suppose $f_i(\boldsymbol{w}) = \|\boldsymbol{w} - \boldsymbol{w}_i^*\|$. Then, $f(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{w} - \boldsymbol{w}_i^*\|$ and $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w})$.

Let $\overline{\boldsymbol{w}}^* = \frac{1}{n}\sum_{i=1}^n \boldsymbol{w}_i^*$. For $\|\boldsymbol{w} - \boldsymbol{w}^*\| \geq D := \max(2\|\overline{\boldsymbol{w}}^* - \boldsymbol{w}^*\|, \frac{4}{n}\sum_{i=1}^n \|\overline{\boldsymbol{w}}^* - \boldsymbol{w}_i^*\|)$, it can be shown that $f(\boldsymbol{w}) - f(\boldsymbol{w}^*) \geq \frac{1}{4}\|\boldsymbol{w} - \boldsymbol{w}^*\|$. Noting that $G = \mathcal{O}(1)$ here (as the sub-gradient of f_i is bounded by 1 in magnitude), Assumption 6.4 easily holds here for small φ.

Proof. Using the triangle inequality, we have for any $\boldsymbol{w}$ satisfying $\|\boldsymbol{w} - \boldsymbol{w}^*\| \geq D$:

$$\begin{aligned}
f(\boldsymbol{w}) &= \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{w} - \boldsymbol{w}_i^*\| && (9)\\
&\geq \left\|\boldsymbol{w} - \frac{1}{n}\sum_{i=1}^n \boldsymbol{w}_i^*\right\| && (10)\\
&= \|\boldsymbol{w} - \overline{\boldsymbol{w}}^*\| && (11)\\
&= \|\boldsymbol{w} - \boldsymbol{w}^* - (\overline{\boldsymbol{w}}^* - \boldsymbol{w}^*)\| && (12)\\
&\geq \left|\|\boldsymbol{w} - \boldsymbol{w}^*\| - \|\overline{\boldsymbol{w}}^* - \boldsymbol{w}^*\|\right| && (13)\\
&= \left|\frac{\|\boldsymbol{w} - \boldsymbol{w}^*\|}{2} + \frac{\|\boldsymbol{w} - \boldsymbol{w}^*\|}{2} - \|\overline{\boldsymbol{w}}^* - \boldsymbol{w}^*\|\right| && (14)\\
&\geq \frac{\|\boldsymbol{w} - \boldsymbol{w}^*\|}{2}. && (15)
\end{aligned}$$

Equation (15) follows because $\frac{\|\boldsymbol{w} - \boldsymbol{w}^*\|}{2} \geq \frac{D}{2} \geq \|\overline{\boldsymbol{w}}^* - \boldsymbol{w}^*\|$ (from the definition of D). Next, since $\boldsymbol{w}^*$ is the minimizer of f and using the definition of D, we have:

$$f(\boldsymbol{w}^*) \leq f(\overline{\boldsymbol{w}}^*) = \frac{1}{n}\sum_{i=1}^n \|\overline{\boldsymbol{w}}^* - \boldsymbol{w}_i^*\| \leq \frac{D}{4} \leq \frac{\|\boldsymbol{w} - \boldsymbol{w}^*\|}{4}. \tag{16}$$

Finally, subtracting Equation (16) from Equation (15) gives us the desired result. ■

F. Some Useful Lemmas

Lemma F.1 (Clipping Bias). *Suppose $\boldsymbol{v}(\boldsymbol{\zeta})$ (where $\boldsymbol{\zeta}$ denotes the source of randomness) is an unbiased estimator of $\boldsymbol{v}$, i.e. $\mathbb{E}_{\boldsymbol{\zeta}}[\boldsymbol{v}(\boldsymbol{\zeta})] = \boldsymbol{v}$. Let $b(\tau)$ denote the clipping bias of clip$(\boldsymbol{v}(\boldsymbol{\zeta}), \tau)$, i.e.*

$$b(\tau) = \Big\| \boldsymbol{v} - \mathbb{E}_{\boldsymbol{\zeta}}\Big[\text{clip}(\boldsymbol{v}(\boldsymbol{\zeta}), \tau)\Big]\Big\|.$$

Then for any $p > 1$,

$$b(\tau) \leq \Big(\mathbb{E}[\|\boldsymbol{v}(\boldsymbol{\zeta})\|^p]\Big)^{\frac{1}{p}} \Big(\mathbb{P}(\|\boldsymbol{v}(\boldsymbol{\zeta})\| \geq \tau)\Big)^{1-\frac{1}{p}} - \tau\mathbb{P}(\|\boldsymbol{v}(\boldsymbol{\zeta})\| \geq \tau).$$

Proof. We shall omit the subscript $\boldsymbol{\zeta}$ in expectations henceforth, and it should be inferred from context. We can bound the clipping bias $b(\tau)$ with a clip norm τ as:

$$\begin{aligned}
b(\tau) &= \Big\| \boldsymbol{v} - \mathbb{E}\Big[\text{clip}(\boldsymbol{v}(\boldsymbol{\zeta}), \tau)\Big]\Big\| && (17)\\
&= \Big\| \boldsymbol{v} - \mathbb{E}\Big[\boldsymbol{v}(\boldsymbol{\zeta}) \min\Big(1, \frac{\tau}{\|\boldsymbol{v}(\boldsymbol{\zeta})\|}\Big)\Big]\Big\| && (18)\\
&= \Big\| \mathbb{E}\Big[\boldsymbol{v}(\boldsymbol{\zeta})\Big(1 - \frac{\tau}{\|\boldsymbol{v}(\boldsymbol{\zeta})\|}\Big)\mathbb{1}(\|\boldsymbol{v}(\boldsymbol{\zeta})\| \geq \tau)\Big]\Big\| && (19)\\
&\leq \mathbb{E}\Big[\|\boldsymbol{v}(\boldsymbol{\zeta})\|\Big(1 - \frac{\tau}{\|\boldsymbol{v}(\boldsymbol{\zeta})\|}\Big)\mathbb{1}(\|\boldsymbol{v}(\boldsymbol{\zeta})\| \geq \tau)\Big] && (20)\\
&= \mathbb{E}\Big[\|\boldsymbol{v}(\boldsymbol{\zeta})\|\mathbb{1}(\|\boldsymbol{v}(\boldsymbol{\zeta})\| \geq \tau)\Big] - \tau\mathbb{E}[\mathbb{1}(\|\boldsymbol{v}(\boldsymbol{\zeta})\| \geq \tau)] && (21)\\
&\leq \Big(\mathbb{E}[\|\boldsymbol{v}(\boldsymbol{\zeta})\|^p]\Big)^{\frac{1}{p}} \Big(\mathbb{E}\Big[\Big(\mathbb{1}(\|\boldsymbol{v}(\boldsymbol{\zeta})\| \geq \tau)\Big)^q\Big]\Big)^{\frac{1}{q}} - \tau\mathbb{P}(\|\boldsymbol{v}(\boldsymbol{\zeta})\| \geq \tau), && (22)
\end{aligned}$$

for $p, q \in (1, \infty)$ such that $\frac{1}{p} + \frac{1}{q} = 1$; this follows from Hölder's inequality. Now

$$\mathbb{E}\Big[\Big(\mathbb{1}(\|\boldsymbol{v}(\boldsymbol{\zeta})\| \geq \tau)\Big)^q\Big] = \mathbb{E}[\mathbb{1}(\|\boldsymbol{v}(\boldsymbol{\zeta})\| \geq \tau)] = \mathbb{P}(\|\boldsymbol{v}(\boldsymbol{\zeta})\| \geq \tau). \tag{23}$$

Plugging this back in Equation (22) and substituting $\frac{1}{q} = 1 - \frac{1}{p}$, we get the desired result for $b(\tau)$. ■

Corollary F.2 (Clipping Bias). *In the setting of Lemma F.1, we have the following simpler upper bound for any $p > 1$:*

$$b(\tau) \leq \frac{\mathbb{E}[\|\boldsymbol{v}(\boldsymbol{\zeta})\|^p]}{\tau^{p-1}}.$$

Proof. From Lemma F.1, we have that:

$$b(\tau) \leq \Big(\mathbb{E}[\|\boldsymbol{v}(\boldsymbol{\zeta})\|^p]\Big)^{\frac{1}{p}} \Big(\mathbb{P}(\|\boldsymbol{v}(\boldsymbol{\zeta})\| \geq \tau)\Big)^{1-\frac{1}{p}}, \tag{24}$$

for any $p > 1$. Using Markov's inequality, we have:

$$\mathbb{P}(\|\boldsymbol{v}(\boldsymbol{\zeta})\| \geq \tau) \leq \frac{\mathbb{E}[\|\boldsymbol{v}(\boldsymbol{\zeta})\|^p]}{\tau^p}. \tag{25}$$

Plugging this in Equation (24), we get the desired result. ■

It is worth mentioning here that a result similar to Corollary F.2 has been established in Lemma 10 of Zhang et al. (2020).

G. Proof of Theorem 3.4

Using the result of Abadi et al. (2016), we know that any $\boldsymbol{w}_t$, where $t \in \{0, \ldots, T-1\}$, will be (ε, δ)-DP if we set $\sigma_n^2 = \nu \frac{T \log(\frac{1}{\delta})}{n^2 \varepsilon^2} \tau^2$ for some absolute constant ν. Thus, $\boldsymbol{w}_{\widehat{t}}$ (where $\widehat{t}$ is chosen uniformly at random from $\{0, \ldots, T-1\}$ as defined in Algorithm 1) will also be (ε, δ)-DP.

H. Full Version and Proof of Theorem 5.4

Theorem H.1 (Convex Case). *Suppose each f_i is convex and $\mathcal{W}$ is a convex set (which can be $\mathbb{R}^d$). In Algorithm 1, for all $t < T$, set $\eta_t = \eta = \frac{C}{T\tau}\left(\frac{1}{T} + \varphi^2\right)^{-1/2}$ for clip norm τ, where $C > 0$ is a parameter of our choice. Recall $\boldsymbol{w}^* \in \text{argmin}_{\boldsymbol{w}\in\mathcal{W}} f(\boldsymbol{w})$ and $\widehat{t} \sim \text{Unif}[0, T-1]$. Then, DP-SGD (Algorithm 1) has the following convergence guarantee:*

$$\frac{1}{n}\sum_{i\in[n]} \mathbb{E}\Big[\min\Big(1, \frac{\tau}{\|\nabla f_i(\boldsymbol{w}_{\widehat{t}})\|}\Big)(f_i(\boldsymbol{w}_{\widehat{t}}) - f_i(\boldsymbol{w}^*))\Big] \leq \left(\frac{\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2}{2C} + C\right)\tau\sqrt{\frac{1}{T} + \varphi^2}.$$

Now suppose Assumption 5.1 holds. Then, DP-SGD has the following upper bound on the optimization risk as a function of the clip norm $\tau \in (0, G_n]$:

$$\text{OR}(T) \leq \frac{1}{\alpha(\tau)}\left(\left(\frac{\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2}{2C} + C\right)G_n\sqrt{\frac{1}{T} + \varphi^2}\right) + \left(\frac{G_n}{\tau\alpha(\tau)} - 1\right)\Delta(\boldsymbol{w}^*),$$

where $\Delta(\boldsymbol{w}^*) \geq 0$ and $\alpha(\tau) \geq 1$ are as defined in Definitions 5.2 and 5.3, respectively.

Theorem 5.4 can be obtained from the above theorem by plugging in $T = \frac{1}{3\varphi^2}$. The proof of Theorem H.1 is below.

Proof:

Proof. Suppose we use a constant clip norm τ and constant learning rate η. For any $\boldsymbol{w}^* \in \text{argmin}_{\boldsymbol{w}\in\mathcal{W}} f(\boldsymbol{w})$, $\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\| \leq \|\boldsymbol{z}_{t+1} - \boldsymbol{w}^*\|$ as $\boldsymbol{w}_{t+1}$ is the projection of $\boldsymbol{z}_{t+1}$ onto the convex set $\mathcal{W}$. Thus:

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2] &\leq \mathbb{E}[\|\boldsymbol{z}_{t+1} - \boldsymbol{w}^*\|^2] && (26)\\
&= \mathbb{E}[\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2] - 2\eta\mathbb{E}[\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}^*\rangle] + \eta^2\mathbb{E}[\|\boldsymbol{g}_t\|^2] && (27)\\
&= \mathbb{E}[\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2] - 2\eta\mathbb{E}\Big[\Big\langle \frac{1}{b}\sum_{i\in\mathcal{S}_t} \text{clip}(\nabla f_i(\boldsymbol{w}_t), \tau), \boldsymbol{w}_t - \boldsymbol{w}^*\Big\rangle\Big] + \eta^2\mathbb{E}[\|\boldsymbol{g}_t\|^2] && (28)\\
&= \mathbb{E}[\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2] - \frac{2\eta}{n}\sum_{i\in[n]} \mathbb{E}\Big[\min\Big(1, \frac{\tau}{\|\nabla f_i(\boldsymbol{w}_t)\|}\Big)\langle \nabla f_i(\boldsymbol{w}_t), \boldsymbol{w}_t - \boldsymbol{w}^*\rangle\Big] + \eta^2\mathbb{E}[\|\boldsymbol{g}_t\|^2] && (29)\\
&\leq \mathbb{E}[\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2] - \frac{2\eta}{n}\sum_{i\in[n]} \mathbb{E}\Big[\min\Big(1, \frac{\tau}{\|\nabla f_i(\boldsymbol{w}_t)\|}\Big)(f_i(\boldsymbol{w}_t) - f_i(\boldsymbol{w}^*))\Big] + \eta^2\mathbb{E}[\|\boldsymbol{g}_t\|^2]. && (30)
\end{aligned}$$

Equation (30) follows from the convexity of f_i. Next, rearranging the above a bit, followed by summing for $t = 0$ through to $T-1$, and then dividing by $2\eta T$ throughout, we get:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\Big\{\frac{1}{n}\sum_{i\in[n]} \mathbb{E}\Big[\min\Big(1, \frac{\tau}{\|\nabla f_i(\boldsymbol{w}_t)\|}\Big)(f_i(\boldsymbol{w}_t) - f_i(\boldsymbol{w}^*))\Big]\Big\} &\leq \frac{\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2 - \mathbb{E}[\|\boldsymbol{w}_T - \boldsymbol{w}^*\|^2]}{2\eta T} + \frac{\eta}{2T}\sum_{t=0}^{T-1}\mathbb{E}[\|\boldsymbol{g}_t\|^2] && (31)\\
&\leq \frac{\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2}{2\eta T} + \frac{\eta}{2T}\sum_{t=0}^{T-1}\mathbb{E}[\|\boldsymbol{g}_t\|^2]. && (32)
\end{aligned}$$

Henceforth, we shall denote $\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|$ by D_0 for brevity.

Next:

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{g}_t\|^2] &= \mathbb{E}\Big[\Big\|\frac{1}{b}\sum_{i\in\mathcal{S}_t} \text{clip}(\nabla f_i(\boldsymbol{w}_t), \tau) + \boldsymbol{\zeta_t}\Big\|^2\Big] && (33)\\
&= \underbrace{\mathbb{E}\Big[\Big\|\frac{1}{b}\sum_{i\in\mathcal{S}_t} \text{clip}(\nabla f_i(\boldsymbol{w}_t), \tau)\Big\|^2\Big]}_{\text{(I)}} + d\sigma_n^2. && (34)
\end{aligned}$$

Let $z_{t,i} = 1$ if sample $i \in \mathcal{S}_t$ and 0 otherwise; note that $\mathbb{P}(z_{t,i} = 1) = \frac{b}{n}$. With this, we can rewrite (I) as:

$$\text{(I)} = \mathbb{E}\Big[\Big\|\frac{1}{b}\sum_{i\in[n]} \text{clip}(\nabla f_i(\boldsymbol{w}_t), \tau) z_{t,i}\Big\|^2\Big] \tag{35}$$

$$= \frac{1}{b^2}\sum_{i\in[n]} \|\text{clip}(\nabla f_i(\boldsymbol{w}_t), \tau)\|^2 \mathrm{E}[z_{t,i}^2] + \frac{1}{b^2}\sum_{i\neq j}\langle \text{clip}(\nabla f_i(\boldsymbol{w}_t), \tau), \text{clip}(\nabla f_j(\boldsymbol{w}_t), \tau)\rangle \mathrm{E}[z_{t,i} z_{t,j}]. \tag{36}$$

Note that $\mathrm{E}[z_{t,i}^2] = \frac{b}{n}\ \forall\ i \in [n]$, $\mathrm{E}[z_{t,i}z_{t,j}] = \frac{b^2}{n^2}\ \forall\ i \neq j$, $\|\text{clip}(\nabla f_i(\boldsymbol{w}_t), \tau)\|^2 \leq \tau^2\ \forall\ i \in [n]$ and $\langle \text{clip}(\nabla f_i(\boldsymbol{w}_t), \tau), \text{clip}(\nabla f_j(\boldsymbol{w}_t), \tau)\rangle \leq \|\text{clip}(\nabla f_i(\boldsymbol{w}_t), \tau)\|\|\text{clip}(\nabla f_j(\boldsymbol{w}_t), \tau)\| \leq \tau^2\ \forall\ i \neq j$. Using all this above, we get:

$$\text{(I)} \leq \tau^2\Big(1 + \frac{1}{b} - \frac{1}{n}\Big). \tag{37}$$

Plugging the above as well as the value of σ_n^2 back in Equation (34), we get:

$$\mathbb{E}[\|\boldsymbol{g}_t\|^2] \leq \tau^2\Big(1 + \frac{1}{b} - \frac{1}{n} + \frac{\nu d T \log(1/\delta)}{n^2\varepsilon^2}\Big) \tag{38}$$

$$\leq \tau^2\Big(2 + \frac{\nu d T \log(1/\delta)}{n^2\varepsilon^2}\Big) \tag{39}$$

$$\leq 2\tau^2\Big(1 + \frac{\nu d T \log(1/\delta)}{n^2\varepsilon^2}\Big). \tag{40}$$

Equation (39) follows because $b \geq 1$.

Plugging Equation (40) in Equation (32), we get:

$$\frac{1}{T}\sum_{t=0}^{T-1}\Big\{\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\Big[\min\Big(1, \frac{\tau}{\|\nabla f_i(\boldsymbol{w}_t)\|}\Big)(f_i(\boldsymbol{w}_t) - f_i(\boldsymbol{w}^*))\Big]\Big\} \leq \frac{D_0^2}{2\eta T} + \eta T \tau^2\Big(\frac{1}{T} + \frac{\nu d \log(\frac{1}{\delta})}{n^2\varepsilon^2}\Big). \tag{41}$$

Plugging in $\eta = \frac{C}{T\tau\sqrt{\frac{1}{T} + \frac{\nu d \log(1/\delta)}{n^2\varepsilon^2}}}$ in the above equation, where $C > 0$ is a constant of our choice, we get:

$$\frac{1}{T}\sum_{t=0}^{T-1}\Big\{\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\Big[\min\Big(1, \frac{\tau}{\|\nabla f_i(\boldsymbol{w}_t)\|}\Big)(f_i(\boldsymbol{w}_t) - f_i(\boldsymbol{w}^*))\Big]\Big\} \leq \Big(\frac{D_0^2}{2C} + C\Big)\tau\sqrt{\frac{1}{T} + \frac{\nu d \log(1/\delta)}{n^2\varepsilon^2}}. \tag{42}$$

Recalling that $\widehat{t} \sim \text{Unif}[0, T-1]$, we can rewrite the above as:

$$\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\Big[\min\Big(1, \frac{\tau}{\|\nabla f_i(\boldsymbol{w}_{\widehat{t}})\|}\Big)(f_i(\boldsymbol{w}_{\widehat{t}}) - f_i(\boldsymbol{w}^*))\Big] \leq \Big(\frac{D_0^2}{2C} + C\Big)\tau\sqrt{\frac{1}{T} + \frac{\nu d \log(1/\delta)}{n^2\varepsilon^2}}. \tag{43}$$

Next, Equation (43) can be further rewritten as:

$$\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\Big[\min\Big(1, \frac{\tau}{\|\nabla f_i(\boldsymbol{w}_{\widehat{t}})\|}\Big)(f_i(\boldsymbol{w}_{\widehat{t}}) - f_i^*)\Big] \leq \Big(\frac{D_0^2}{2C} + C\Big)\tau\sqrt{\frac{1}{T} + \frac{\nu d \log(1/\delta)}{n^2\varepsilon^2}} + \frac{1}{n}\sum_{i\in[n]}\mathbb{E}\Big[\min\Big(1, \frac{\tau}{\|\nabla f_i(\boldsymbol{w}_{\widehat{t}})\|}\Big)\Big](f_i(\boldsymbol{w}^*) - f_i^*), \tag{44}$$

where $f_i^* = \min_{\boldsymbol{w}\in\mathcal{W}} f_i(\boldsymbol{w})$.

From Assumption 5.1, $\|\nabla f_i(\boldsymbol{w}_{\widehat{t}})\| \leq G_i$; thus, $\min\Big(1, \frac{\tau}{\|\nabla f_i(\boldsymbol{w}_{\widehat{t}})\|}\Big) \geq \min\big(1, \frac{\tau}{G_i}\big)$. Using this and the fact that $f_i(\boldsymbol{w}_{\widehat{t}}) - f_i^* \geq 0$, we get:

$$\min\Big(1, \frac{\tau}{\|\nabla f_i(\boldsymbol{w}_{\widehat{t}})\|}\Big)(f_i(\boldsymbol{w}_{\widehat{t}}) - f_i^*) \geq \min\Big(1, \frac{\tau}{G_i}\Big)(f_i(\boldsymbol{w}_{\widehat{t}}) - f_i^*).$$

Using this in Equation (44) together with the fact that $\min\left(1, \frac{\tau}{\|\nabla f_i(\boldsymbol{w}_{\hat{t}})\|}\right) \leq 1$ and then dividing by τ throughout, we get:

$$\frac{1}{n}\sum_{i\in[n]} \min\left(\frac{1}{\tau}, \frac{1}{G_i}\right)\mathbb{E}[f_i(\boldsymbol{w}_{\hat{t}}) - f_i^*] \leq \left(\frac{D_0^2}{2C} + C\right)\sqrt{\frac{1}{T} + \frac{\nu d \log(1/\delta)}{n^2\varepsilon^2}} + \frac{1}{n}\sum_{i\in[n]} \frac{(f_i(\boldsymbol{w}^*) - f_i^*)}{\tau}. \tag{45}$$

Next, from the definition of $\alpha(\tau)$ in Definition 5.3, we have that:

$$\frac{1}{n}\sum_{i\in[n]} \min\left(\frac{1}{\tau}, \frac{1}{G_i}\right)\mathbb{E}[f_i(\boldsymbol{w}_{\hat{t}}) - f_i^*] \geq \frac{\alpha(\tau)}{n}\sum_{i\in[n]} \frac{\mathbb{E}[f_i(\boldsymbol{w}_{\hat{t}}) - f_i^*]}{G_n} \tag{46}$$

$$= \left(\frac{\alpha(\tau)}{G_n}\right)\underbrace{\left(\mathbb{E}[f(\boldsymbol{w}_{\hat{t}})] - f(\boldsymbol{w}^*)\right)}_{=\mathrm{OR}(T)} + \left(\frac{\alpha(\tau)}{G_n}\right)\left(\frac{1}{n}\sum_{i\in[n]}(f_i(\boldsymbol{w}^*) - f_i^*)\right). \tag{47}$$

Next, using Equation (47) in Equation (45) and the definition of $\mathrm{OR}(T)$, we get (after some rearrangement):

$$\mathrm{OR}(T) \leq \frac{1}{\alpha(\tau)}\left(\left(\frac{D_0^2}{2C} + C\right)G_n\sqrt{\frac{1}{T} + \frac{\nu d \log(1/\delta)}{n^2\varepsilon^2}}\right) + \left(\frac{G_n}{\tau\alpha(\tau)} - 1\right)\underbrace{\left(\frac{1}{n}\sum_{i\in[n]}(f_i(\boldsymbol{w}^*) - f_i^*)\right)}_{\Delta(\boldsymbol{w}^*)}. \tag{48}$$

Recalling the definition of $\Delta(\boldsymbol{w}^*)$, we get the desired result. ■

I. Result for the Constrained Convex Case under Assumption 6.1

As mentioned in the main paper, Kamath et al. (2021) consider stochastic convex optimization (SCO) and derive upper and lower bounds of $\mathcal{O}(\varphi^{1-\frac{1}{k}})$ and $\Omega(\varphi^{1-\frac{1}{k}})$ for the generalization error (i.e., $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\ell(\boldsymbol{w}_{\text{priv}}, \boldsymbol{x}) - \ell(\boldsymbol{w}^{**}, \boldsymbol{x})]$, where $\boldsymbol{w}^{**} := \arg\min_{\boldsymbol{w}\in\mathbb{R}^d}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\ell(\boldsymbol{w}, \boldsymbol{x})]$) in the constrained convex case under an assumption similar to Assumption 6.1[7]. We shall now show that the same bounds hold for the training error (i.e., $f(\boldsymbol{w}_{\text{priv}}) - f(\boldsymbol{w}^*)$) in empirical risk minimization (which is what we consider in this work) in the constrained convex case under Assumption 6.1.

Theorem I.1 (**Constrained Convex Case**). *Suppose Assumption 6.1 holds, f is convex and $\mathcal{W}$ is a bounded convex set with diameter $D_{\mathcal{W}} < \infty$. Fix some $\gamma \in (0, 1)$. In Algorithm 1, set $\tau = \frac{G}{\gamma^{1/k}}\left(\frac{1}{T} + \varphi^2\right)^{-\frac{1}{2k}}$ and $\eta_t = \eta = \frac{D_{\mathcal{W}}}{T\tau}\left(\frac{1}{T} + \varphi^2\right)^{-\frac{1}{2}}$ for all $t < T$. Then with a probability of at least $(1-\gamma)$ which is w.r.t. the random dataset $\mathcal{Z}$ that we obtain, DP-SGD (Algorithm 1) has the following guarantee:*

$$\mathrm{OR}(T) \leq \frac{5D_{\mathcal{W}}G}{2\gamma^{1/k}}\left(\frac{1}{T} + \varphi^2\right)^{\frac{1}{2}(1-\frac{1}{k})}.$$

So if we set $T = \frac{1}{\varphi^2}$ above, we get the following bound for the risk:

$$\mathrm{OR}(T) \leq \frac{5D_{\mathcal{W}}G}{2^{\frac{1}{2}(1+\frac{1}{k})}\gamma^{1/k}}\varphi^{(1-\frac{1}{k})}.$$

Remark I.2 (**Comparison with Lipschitz Case**). As per the above theorem, the optimization risk is $\mathcal{O}(\varphi^{1-\frac{1}{k}})$ in the bounded k^{th} moment **constrained** convex case. In comparison, the risk is $\mathcal{O}(\varphi)$ in the **Lipschitz** convex case (equivalent to $k = \infty$); see Bassily et al. (2014).

We also have a matching lower bound in this case.

Theorem I.3 (**Lower Bound for Constrained Convex Case**). *Suppose $\varphi < o(1)$ and $\delta < \exp(-\varepsilon^2)$. There exists a convex loss function ℓ and a bounded convex set $\mathcal{W}$, such that for every (ε, δ)-DP algorithm $\mathcal{A}$ which tries to solve for $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}\in\mathcal{W}} f(\boldsymbol{w})$ where f is the average loss for a dataset $\mathcal{Z}$ of n samples drawn from the data distribution $\mathcal{D}$, there exists a choice of $\mathcal{D}$ such that:*

[7] Specifically, Kamath et al. (2021) assume *coordinate-wise* bounded *centered* moments.

- *f satisfies Assumption 6.1.*
- $\mathbb{E}_{\mathcal{Z}\sim\mathcal{D}^n,\mathcal{A}}\Big[f(\boldsymbol{w}_{\mathcal{Z}}^{(\mathcal{A})}) - f(\boldsymbol{w}^*)\Big] \geq \Omega\big(\varphi^{1-\frac{1}{k}}\big)$*, where $\boldsymbol{w}_{\mathcal{Z}}^{(\mathcal{A})}$ is the output of algorithm $\mathcal{A}$ on the dataset $\mathcal{Z}$.*

Remark I.4 (**Tightness of Theorem I.1**). The $\mathcal{O}(\varphi^{1-\frac{1}{k}})$ bound on the risk in Theorem I.1 is tight (for $\delta < \exp(-\varepsilon^2)$).

We now prove Theorem I.1; the proof of Theorem I.3 is deferred to Appendix M.

Proof of Theorem I.1:

Proof. First, using Lemma I.5, we have that:

$$\Big\|\frac{1}{n}\sum_{i\in[n]}\text{clip}(\nabla f_i(\boldsymbol{w}),\tau) - \nabla f(\boldsymbol{w})\Big\| \leq \frac{G_{\mathcal{Z}}^k}{\tau^{k-1}}, \tag{49}$$

where $G_{\mathcal{Z}}^k = \frac{1}{n}\sum_{i=1}^n (G(\boldsymbol{x}_i, y_i))^k$.

For any $\boldsymbol{w}^* \in \text{argmin}_{\boldsymbol{w}\in\mathcal{W}} f(\boldsymbol{w})$, $\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\| \leq \|\boldsymbol{z}_{t+1} - \boldsymbol{w}^*\|$ as $\boldsymbol{w}_{t+1} = \Pi_{\mathcal{W}}(\boldsymbol{z}_{t+1})$. So:

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2] &\leq \mathbb{E}[\|\boldsymbol{z}_{t+1} - \boldsymbol{w}^*\|^2] && (50)\\
&= \mathbb{E}[\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2] - 2\eta\mathbb{E}[\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}^*\rangle] + \eta^2\mathbb{E}[\|\boldsymbol{g}_t\|^2] && (51)\\
&= \mathbb{E}[\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2] - 2\eta\mathbb{E}\Big[\Big\langle \frac{1}{b}\sum_{i\in\mathcal{S}_t}\text{clip}(\nabla f_i(\boldsymbol{w}_t),\tau), \boldsymbol{w}_t - \boldsymbol{w}^*\Big\rangle\Big] + \eta^2\mathbb{E}[\|\boldsymbol{g}_t\|^2] && (52)\\
&= \mathbb{E}[\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2] - 2\eta\mathbb{E}\Big[\Big\langle \frac{1}{n}\sum_{i\in[n]}\text{clip}(\nabla f_i(\boldsymbol{w}_t),\tau), \boldsymbol{w}_t - \boldsymbol{w}^*\Big\rangle\Big] + \eta^2\mathbb{E}[\|\boldsymbol{g}_t\|^2] && (53)\\
&\leq \mathbb{E}[\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2] - 2\eta\mathbb{E}[\langle \nabla f(\boldsymbol{w}_t), \boldsymbol{w}_t - \boldsymbol{w}^*\rangle] && (54)\\
&\quad + 2\eta\mathbb{E}\Big[\Big\|\frac{1}{n}\sum_{i\in[n]}\text{clip}(\nabla f_i(\boldsymbol{w}_t),\tau) - \nabla f(\boldsymbol{w}_t)\Big\|\|\boldsymbol{w}_t - \boldsymbol{w}^*\|\Big] + \eta^2\mathbb{E}[\|\boldsymbol{g}_t\|^2]\\
&\leq \mathbb{E}[\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2] - 2\eta\mathbb{E}[f(\boldsymbol{w}_t) - f(\boldsymbol{w}^*)] && (55)\\
&\quad + 2\eta D_{\mathcal{W}}\mathbb{E}\Big[\Big\|\frac{1}{n}\sum_{i\in[n]}\text{clip}(\nabla f_i(\boldsymbol{w}_t),\tau) - \nabla f(\boldsymbol{w}_t)\Big\|\Big] + \eta^2\mathbb{E}[\|\boldsymbol{g}_t\|^2].
\end{aligned}$$

Equation (55) follows from the convexity of f together with the fact that $\|\boldsymbol{w}_t - \boldsymbol{w}^*\| \leq D_{\mathcal{W}}$. Next, rearranging the above a bit, followed by summing for $t = 0$ through to $T-1$, and then dividing by $2\eta T$ throughout, we get:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\Big(\mathbb{E}[f(\boldsymbol{w}_t)] - f(\boldsymbol{w}^*)\Big) &\leq \frac{D_{\mathcal{W}}^2}{2\eta T} + \frac{\eta}{2T}\sum_{t=0}^{T-1}\mathbb{E}[\|\boldsymbol{g}_t\|^2] + \frac{D_{\mathcal{W}}}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\Big\|\frac{1}{n}\sum_{i\in[n]}\text{clip}(\nabla f_i(\boldsymbol{w}_t),\tau) - \nabla f(\boldsymbol{w}_t)\Big\|\Big] && (56)\\
&\leq \frac{D_{\mathcal{W}}^2}{2\eta T} + \eta T\tau^2\left(\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2} + \frac{1}{T}\right) + \frac{D_{\mathcal{W}}G_{\mathcal{Z}}^k}{\tau^{k-1}}, && (57)
\end{aligned}$$

where the last step follows by using Equation (40) and Equation (49).

Plugging in $\eta = \frac{D_{\mathcal{W}}}{T\tau\sqrt{\frac{1}{T} + \frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}}}$ above, we get:

$$\frac{1}{T}\sum_{t=0}^{T-1}\Big(\mathbb{E}[f(\boldsymbol{w}_t)] - f(\boldsymbol{w}^*)\Big) \leq \frac{3D_{\mathcal{W}}\tau}{2}\sqrt{\frac{1}{T} + \frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}} + \frac{D_{\mathcal{W}}G_{\mathcal{Z}}^k}{\tau^{k-1}}. \tag{58}$$

Let us choose $\tau = \frac{G}{\gamma^{1/k}}\left(\frac{1}{T} + \frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\right)^{-\frac{1}{2k}}$ above, where $\gamma \in (0,1)$. With that, we get:

$$\frac{1}{T}\sum_{t=0}^{T-1}\Big(\mathbb{E}[f(\boldsymbol{w}_t)] - f(\boldsymbol{w}^*)\Big) \leq D_{\mathcal{W}}\left(\frac{3G}{2\gamma^{\frac{1}{k}}} + \frac{G_{\mathcal{Z}}^k\gamma^{1-\frac{1}{k}}}{G^{k-1}}\right)\left(\frac{1}{T} + \frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\right)^{\frac{1}{2}(1-\frac{1}{k})}. \tag{59}$$

Now, using Markov's inequality, $G_{\mathcal{Z}}^k \leq \frac{G^k}{\gamma}$ with a probability of at least $1-\gamma$ w.r.t. the random dataset $\mathcal{Z}$. Plugging this above, we get:

$$\frac{1}{T}\sum_{t=0}^{T-1}\Big(\mathbb{E}[f(\boldsymbol{w}_t)] - f(\boldsymbol{w}^*)\Big) \leq \frac{5D_{\mathcal{W}}G}{2\gamma^{1/k}}\left(\frac{1}{T} + \frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\right)^{\frac{1}{2}(1-\frac{1}{k})}, \tag{60}$$

with a probability of at least $1-\gamma$ w.r.t. the random dataset $\mathcal{Z}$.

Lastly, plugging in $\varphi = \frac{\sqrt{\nu d\log(1/\delta)}}{n\varepsilon}$, noting that $\mathbb{E}[f(\boldsymbol{w}_{\widehat{t}})] - f(\boldsymbol{w}^*) = \frac{1}{T}\sum_{t=0}^{T-1}\Big(\mathbb{E}[f(\boldsymbol{w}_t)] - f(\boldsymbol{w}^*)\Big)$ and using the definition of $\mathrm{OR}(T)$, we get the final result. ■

Lemma I.5 (**Clipping Bias under Assumption 4.1**). *Under Assumption 4.1, we have for any $\boldsymbol{w} \in \mathcal{W}$:*

$$\Big\|\frac{1}{n}\sum_{i\in[n]}\mathrm{clip}(\nabla f_i(\boldsymbol{w}),\tau) - \nabla f(\boldsymbol{w})\Big\| \leq \frac{G_{\mathcal{Z}}^k}{\tau^{k-1}},$$

where $G_{\mathcal{Z}}^k := \frac{1}{n}\sum_{i=1}^n (G(\boldsymbol{x}_i, y_i))^k$.

Proof. Using Corollary F.2 and Assumption 4.1, we have for any $\boldsymbol{w} \in \mathcal{W}$:

$$\Big\|\frac{1}{n}\sum_{i\in[n]}\mathrm{clip}(\nabla f_i(\boldsymbol{w}),\tau) - \nabla f(\boldsymbol{w})\Big\| = \|\mathbb{E}_i[\mathrm{clip}(\nabla f_i(\boldsymbol{w}),\tau)] - \nabla f(\boldsymbol{w})\| \tag{61}$$

$$\leq \frac{\mathbb{E}_i[\|\nabla f_i(\boldsymbol{w})\|^k]}{\tau^{k-1}} \tag{62}$$

$$\leq \frac{G_{\mathcal{Z}}^k}{\tau^{k-1}}, \tag{63}$$

where $G_{\mathcal{Z}}^k = \frac{1}{n}\sum_{i=1}^n (G(\boldsymbol{x}_i, y_i))^k$. ■

J. Full Version and Proof of Theorem 6.2

Theorem J.1 (**Unconstrained Convex Case**). *Suppose Assumption 6.1 holds, f is convex and $\mathcal{W} = \mathbb{R}^d$. Fix some $\gamma \in (0,1)$ and $C > 0$. In Algorithm 1, set $\tau = \frac{G}{\gamma^{1/k}}\left(\frac{1}{T} + \varphi^2\right)^{-\frac{1}{k+1}}$ and $\eta_t = \eta = \frac{C}{T\tau}\left(\frac{1}{T} + \varphi^2\right)^{-\frac{1}{2}}$ for all $t < T$. Then with a probability of at least $(1-\gamma)$ which is w.r.t. the random dataset $\mathcal{Z}$ that we obtain, DP-SGD (Algorithm 1) has the following guarantee:*

$$\mathrm{OR}(T) \leq \frac{G}{\gamma^{1/k}}\left\{\frac{1}{2}\left(\frac{\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2}{C} + 4C\right)\left(\frac{1}{T} + \varphi^2\right)^{\frac{1}{2}(1-\frac{2}{k+1})} + (\|\boldsymbol{w}_0 - \boldsymbol{w}^*\| + C)\left(\frac{1}{T} + \varphi^2\right)^{(1-\frac{2}{k+1})}\right\}.$$

So if we set $T = \frac{1}{\varphi^2}$ above (which is what we do in Theorem 6.2), we get the following bound for the risk:

$$\mathrm{OR}(T) \leq \frac{G}{\gamma^{1/k}}\left\{\frac{1}{2^{\frac{1}{2}+\frac{1}{k+1}}}\left(\frac{\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2}{C} + 4C\right)\varphi^{(1-\frac{2}{k+1})} + 2^{1-\frac{2}{k+1}}(\|\boldsymbol{w}_0 - \boldsymbol{w}^*\| + C)\varphi^{2(1-\frac{2}{k+1})}\right\}.$$

We prove this result now.

Proof:

Proof. Everything remains the same till Equation (54) in the proof of Theorem I.1. That is, we have:

$$\mathbb{E}[\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2] \leq \mathbb{E}[\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2] - 2\eta\mathbb{E}[\langle\nabla f(\boldsymbol{w}_t), \boldsymbol{w}_t - \boldsymbol{w}^*\rangle] \tag{64}$$
$$+ 2\eta\mathbb{E}\Big[\Big\|\frac{1}{n}\sum_{i\in[n]}\mathrm{clip}(\nabla f_i(\boldsymbol{w}_t),\tau) - \nabla f(\boldsymbol{w}_t)\Big\|\|\boldsymbol{w}_t - \boldsymbol{w}^*\|\Big] + \eta^2\mathbb{E}[\|\boldsymbol{g}_t\|^2],$$

where $\boldsymbol{w}^* = \text{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w})$.

Using Lemma I.5, we have:

$$\mathbb{E}\Big[\Big\|\frac{1}{n}\sum_{i\in[n]}\text{clip}(\nabla f_i(\boldsymbol{w}_t),\tau)-\nabla f(\boldsymbol{w}_t)\Big\|\|\boldsymbol{w}_t-\boldsymbol{w}^*\|\Big] \le \frac{G_{\mathcal{Z}}^k}{\tau^{k-1}}\mathbb{E}[\|\boldsymbol{w}_t-\boldsymbol{w}^*\|], \tag{65}$$

where $G_{\mathcal{Z}}^k = \frac{1}{n}\sum_{i=1}^n (G(\boldsymbol{x}_i,y_i))^k$.

Now using Lemma J.2 in Equation (65), we get:

$$\mathbb{E}\Big[\Big\|\frac{1}{n}\sum_{i\in[n]}\text{clip}(\nabla f_i(\boldsymbol{w}_t),\tau)-\nabla f(\boldsymbol{w}_t)\Big\|\|\boldsymbol{w}_t-\boldsymbol{w}^*\|\Big] \le \frac{G_{\mathcal{Z}}^k}{\tau^{k-1}}\left(\|\boldsymbol{w}_0-\boldsymbol{w}^*\|+\eta T\left(G_{\mathcal{Z}}+\tau\sqrt{\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}}\right)\right). \tag{66}$$

Using the above equation and Equation (40) in Equation (64) as well as the convexity of f, we get:

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{w}_{t+1}-\boldsymbol{w}^*\|^2] &\le \mathbb{E}[\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2]-2\eta\mathbb{E}[f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)] \\ &+\frac{2\eta G_{\mathcal{Z}}^k}{\tau^{k-1}}\left(\|\boldsymbol{w}_0-\boldsymbol{w}^*\|+\eta T G_{\mathcal{Z}}+\eta T\tau\sqrt{\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}}\right)+2\eta^2\tau^2\Big(1+\frac{\nu d T\log(1/\delta)}{n^2\varepsilon^2}\Big).\end{aligned} \tag{67}$$

Next, summing the above for $t=0$ through to $T-1$, rearranging a bit and then dividing by $2\eta T$ throughout, we get the following:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\Big(\mathbb{E}[f(\boldsymbol{w}_t)]-f(\boldsymbol{w}^*)\Big) &\le \frac{\|\boldsymbol{w}_0-\boldsymbol{w}^*\|^2}{2\eta T}+\eta T\tau^2\left(\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}+\frac{1}{T}\right) \\ &+\frac{G_{\mathcal{Z}}^k}{\tau^{k-1}}\Big(\|\boldsymbol{w}_0-\boldsymbol{w}^*\|+\eta T G_{\mathcal{Z}}+\eta T\tau\sqrt{\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}}\Big).\end{aligned} \tag{68}$$

Let us choose $\eta = \frac{C}{T\tau\sqrt{\frac{1}{T}+\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}}}$, where $C>0$ is some constant of our choice. With that, we get after simplifying a bit:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\Big(\mathbb{E}[f(\boldsymbol{w}_t)]-f(\boldsymbol{w}^*)\Big) &\le \Big(\frac{\|\boldsymbol{w}_0-\boldsymbol{w}^*\|^2}{2C}+C\Big)\tau\sqrt{\frac{1}{T}+\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}}+\frac{(\|\boldsymbol{w}_0-\boldsymbol{w}^*\|+C)G_{\mathcal{Z}}^k}{\tau^{k-1}} \\ &+\frac{CG_{\mathcal{Z}}^{k+1}}{\tau^k}\frac{1}{\sqrt{\frac{1}{T}+\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}}}.\end{aligned}$$

Let us choose $\tau = \frac{G}{\gamma^{1/k}}\left(\frac{1}{T}+\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\right)^{-\frac{1}{k+1}}$ above, where $\gamma\in(0,1)$. With that, we get:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\Big(\mathbb{E}[f(\boldsymbol{w}_t)]-f(\boldsymbol{w}^*)\Big) &\le \frac{G}{\gamma^{1/k}}\Bigg\{\left(\frac{\|\boldsymbol{w}_0-\boldsymbol{w}^*\|^2}{2C}+C+\frac{CG_{\mathcal{Z}}^{k+1}}{\big(G/\gamma^{1/k}\big)^{k+1}}\right)\left(\frac{1}{T}+\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\right)^{\frac{1}{2}(1-\frac{2}{k+1})} \\ &+(\|\boldsymbol{w}_0-\boldsymbol{w}^*\|+C)\left(\frac{G_{\mathcal{Z}}^k}{\big(G/\gamma^{1/k}\big)^k}\right)\left(\frac{1}{T}+\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\right)^{(1-\frac{2}{k+1})}\Bigg\}.\end{aligned} \tag{69}$$

Let:

$$\text{(I)} := \left(\frac{CG_{\mathcal{Z}}^{k+1}}{\big(G/\gamma^{1/k}\big)^{k+1}}\right)\left(\frac{1}{T}+\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\right)^{\frac{1}{2}(1-\frac{2}{k+1})}+(\|\boldsymbol{w}_0-\boldsymbol{w}^*\|+C)\left(\frac{G_{\mathcal{Z}}^k}{\big(G/\gamma^{1/k}\big)^k}\right)\left(\frac{1}{T}+\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\right)^{(1-\frac{2}{k+1})}.$$

Now note that $G_{\mathcal{Z}} \le \frac{G}{\gamma^{1/k}}$ implies:

$$\text{(I)} \le \underbrace{C\left(\frac{1}{T}+\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\right)^{\frac{1}{2}(1-\frac{2}{k+1})}+(\|\boldsymbol{w}_0-\boldsymbol{w}^*\|+C)\left(\frac{1}{T}+\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\right)^{(1-\frac{2}{k+1})}}_{:=\text{(II)}}.$$

Thus, $\mathbb{P}_{\mathcal{Z}}\big((\text{I}) \leq (\text{II})\big) \geq \mathbb{P}_{\mathcal{Z}}(G_{\mathcal{Z}} \leq G/\gamma^{1/k})$. But using Markov's inequality, $G_{\mathcal{Z}}^k \leq \frac{G^k}{\gamma}$ with a probability of at least $1-\gamma$ w.r.t. the random dataset $\mathcal{Z}$. Thus, (I) $\leq$ (II) with a probability of at least $1-\gamma$. Using this in Equation (69), we get:

$$\frac{1}{T}\sum_{t=0}^{T-1}\Big(\mathbb{E}[f(\boldsymbol{w}_t)] - f(\boldsymbol{w}^*)\Big) \leq \frac{G}{\gamma^{1/k}}\Bigg\{\left(\frac{\|\boldsymbol{w}_0-\boldsymbol{w}^*\|^2}{2C}+2C\right)\left(\frac{1}{T}+\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\right)^{\frac{1}{2}(1-\frac{2}{k+1})} + (\|\boldsymbol{w}_0-\boldsymbol{w}^*\|+C)\left(\frac{1}{T}+\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\right)^{(1-\frac{2}{k+1})}\Bigg\}, \quad (70)$$

with a probability of at least $1-\gamma$ w.r.t. the random dataset $\mathcal{Z}$.

Lastly, plugging in $\varphi = \frac{\sqrt{\nu d\log(1/\delta)}}{n\varepsilon}$, noting that $\mathbb{E}[f(\boldsymbol{w}_{\hat{t}})] - f(\boldsymbol{w}^*) = \frac{1}{T}\sum_{t=0}^{T-1}\Big(\mathbb{E}[f(\boldsymbol{w}_t)] - f(\boldsymbol{w}^*)\Big)$ and using the definition of OR(T), we get the final result. ■

Lemma J.2. *In the setting of the proof of Theorem J.1, for any $0 < t < T$, we have:*

$$\mathbb{E}[\|\boldsymbol{w}_t - \boldsymbol{w}^*\|] \leq \|\boldsymbol{w}_0-\boldsymbol{w}^*\| + \eta T\left(G_{\mathcal{Z}} + \tau\sqrt{\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}}\right), \quad (71)$$

where $G_{\mathcal{Z}}^k := \frac{1}{n}\sum_{i=1}^n (G(\boldsymbol{x}_i, y_i))^k$.

Proof. Let us denote $\frac{1}{b}\sum_{i\in\mathcal{S}_t}\text{clip}(\nabla f_i(\boldsymbol{w}_t),\tau)$ by $\boldsymbol{u}_t$. Now:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}_t}[\|\boldsymbol{u}_t\|] &\leq \mathbb{E}_{\mathcal{S}_t}\Big[\frac{1}{b}\sum_{i\in\mathcal{S}_t}\|\text{clip}(\nabla f_i(\boldsymbol{w}_t),\tau)\|\Big] && (72)\\
&= \frac{1}{n}\sum_{i\in[n]}\|\text{clip}(\nabla f_i(\boldsymbol{w}_t),\tau)\| && (73)\\
&\leq \frac{1}{n}\sum_{i\in[n]}\|\nabla f_i(\boldsymbol{w}_t)\| && (74)\\
&\leq \Big(\frac{1}{n}\sum_{i\in[n]}\|\nabla f_i(\boldsymbol{w}_t)\|^k\Big)^{1/k} \qquad \text{(using Jensen's inequality)} && (75)\\
&\leq G_{\mathcal{Z}}, && (76)
\end{aligned}$$

where $G_{\mathcal{Z}}^k := \frac{1}{n}\sum_{i=1}^n (G(\boldsymbol{x}_i, y_i))^k$.

Now for any $t > 0$:

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{w}_t - \boldsymbol{w}^*\|] &\leq \mathbb{E}[\|(\boldsymbol{w}_t-\boldsymbol{w}_0)+(\boldsymbol{w}_0-\boldsymbol{w}^*)\|] && (77)\\
&\leq \mathbb{E}[\|\boldsymbol{w}_t-\boldsymbol{w}_0\|] + \|\boldsymbol{w}_0-\boldsymbol{w}^*\| && (78)\\
&\leq \eta\mathbb{E}\Big[\Big\|\sum_{t'=0}^{t-1}(\boldsymbol{u}_{t'}+\boldsymbol{\zeta}_{t'})\Big\|\Big] + \|\boldsymbol{w}_0-\boldsymbol{w}^*\| && (79)\\
&\leq \eta\mathbb{E}\Big[\Big\|\sum_{t'=0}^{t-1}\boldsymbol{u}_{t'}\Big\|\Big] + \eta\mathbb{E}\Big[\Big\|\sum_{t'=0}^{t-1}\boldsymbol{\zeta}_{t'}\Big\|\Big] + \|\boldsymbol{w}_0-\boldsymbol{w}^*\| && (80)\\
&\leq \eta\sum_{t'=0}^{t-1}\mathbb{E}[\|\boldsymbol{u}_{t'}\|] + \eta\sqrt{\mathbb{E}\Big[\Big\|\sum_{t'=0}^{t-1}\boldsymbol{\zeta}_{t'}\Big\|^2\Big]} + \|\boldsymbol{w}_0-\boldsymbol{w}^*\| && (81)\\
&\leq \eta t G_{\mathcal{Z}} + \eta\sqrt{t\sigma_n^2 d} + \|\boldsymbol{w}_0-\boldsymbol{w}^*\|, && (82)
\end{aligned}$$

where Equation (82) follows by using Equation (76) and because $\sum_{t'=0}^{t-1}\boldsymbol{\zeta}_{t'}$ is $\mathcal{N}(\vec{0}, t\sigma_n^2\mathrm{I}_d)$. Plugging in the value of σ_n^2 and using the fact that $t < T$, we get the desired result. ■

K. Full Version and Proof of Theorem 6.5

Theorem K.1 (**Unconstrained Convex Case Under Assumption 6.4**). *Suppose Assumptions 6.1 and 6.4 hold, f is convex and $\mathcal{W} = \mathbb{R}^d$. Fix some $C > 0$. In Algorithm 1, set $T \geq \frac{1}{\varphi^2}$, $\tau = G\Big(\frac{1}{T} + \varphi^2\Big)^{-\frac{1}{2k}}$ and $\eta_t = \eta = \frac{C}{T\tau}\Big(\frac{1}{T} + \varphi^2\Big)^{-\frac{1}{2}}$ for all $t < T$. Then with a probability of at least $3/4$ which is w.r.t. the random dataset $\mathcal{Z}$ that we obtain, DP-SGD (Algorithm 1) has the following **improved** guarantee:*

$$\mathrm{OR}(T) \leq G\left\{\Big(\frac{\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2}{C} + 2C\Big)\Big(\frac{1}{T} + \varphi^2\Big)^{\frac{1}{2}(1-\frac{1}{k})} + 16\varphi^{(1-\frac{1}{k})}D\right\}.$$

So if we set $T = \frac{1}{\varphi^2}$ above (which is what we do in Theorem 6.5), we get the following bound for the risk:

$$\mathrm{OR}(T) \leq G\left\{2^{\frac{1}{2}(1-\frac{1}{k})}\Big(\frac{\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2}{C} + 2C\Big) + 16D\right\}\varphi^{(1-\frac{1}{k})}.$$

We now prove this result.

Proof:

Proof. Similar to Equation (64) in the proof of Theorem J.1, taking expectation only w.r.t. the randomness in the current iteration t, we have:

$$\begin{aligned}\mathbb{E}_t[\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2] \leq \|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 - 2\eta\langle\nabla f(\boldsymbol{w}_t), \boldsymbol{w}_t - \boldsymbol{w}^*\rangle \\ + 2\eta\Big\|\frac{1}{n}\sum_{i\in[n]}\mathrm{clip}(\nabla f_i(\boldsymbol{w}_t), \tau) - \nabla f(\boldsymbol{w}_t)\Big\|\|\boldsymbol{w}_t - \boldsymbol{w}^*\| + \eta^2\mathbb{E}_t[\|\boldsymbol{g}_t\|^2],\end{aligned} \tag{83}$$

where $\boldsymbol{w}^* = \mathrm{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w})$.

Equation (65) in the proof of Theorem J.1 also holds here, i.e., we have:

$$\Big\|\frac{1}{n}\sum_{i\in[n]}\mathrm{clip}(\nabla f_i(\boldsymbol{w}_t), \tau) - \nabla f(\boldsymbol{w}_t)\Big\|\|\boldsymbol{w}_t - \boldsymbol{w}^*\| \leq \frac{G_{\mathcal{Z}}^k}{\tau^{k-1}}\|\boldsymbol{w}_t - \boldsymbol{w}^*\|, \tag{84}$$

where $G_{\mathcal{Z}}^k = \frac{1}{n}\sum_{i=1}^n (G(\boldsymbol{x}_i, y_i))^k$.

Plugging in Equation (84) into Equation (83), we get:

$$\mathbb{E}_t[\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2] \leq \|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 - 2\eta\langle\nabla f(\boldsymbol{w}_t), \boldsymbol{w}_t - \boldsymbol{w}^*\rangle + 2\eta\Big(\frac{G_{\mathcal{Z}}^k}{\tau^{k-1}}\|\boldsymbol{w}_t - \boldsymbol{w}^*\|\Big) + \eta^2\mathbb{E}_t[\|\boldsymbol{g}_t\|^2]. \tag{85}$$

Further, using Equation (40) to bound $\mathbb{E}_t[\|\boldsymbol{g}_t\|^2]$ as well as the convexity of f above, we get:

$$\mathbb{E}_t[\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2] \leq \|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 \underbrace{-2\eta(f(\boldsymbol{w}_t) - f(\boldsymbol{w}^*)) + 2\eta\Big(\frac{G_{\mathcal{Z}}^k}{\tau^{k-1}}\|\boldsymbol{w}_t - \boldsymbol{w}^*\|\Big)}_{\text{(I)}} + 2\eta^2\tau^2\Big(1 + \frac{\nu dT\log(1/\delta)}{n^2\varepsilon^2}\Big). \tag{86}$$

Now, plugging in our choice of $\tau = G\Big(\frac{1}{T} + \frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\Big)^{-\frac{1}{2k}}$ in (I) and using the fact that $T \geq \frac{n^2\varepsilon^2}{\nu d\log(1/\delta)}$, we get:

$$\text{(I)} \leq -2\eta(f(\boldsymbol{w}_t) - f(\boldsymbol{w}^*)) + 4\eta\Big(\frac{G_{\mathcal{Z}}^k}{G^{k-1}}\Big)\Big(\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\Big)^{\frac{1}{2}(1-\frac{1}{k})}\|\boldsymbol{w}_t - \boldsymbol{w}^*\|. \tag{87}$$

Now note that $G_{\mathcal{Z}}^k\|\boldsymbol{w}_t - \boldsymbol{w}^*\|$ depends on the random dataset $\mathcal{Z}$. But $G_{\mathcal{Z}}^k \leq 4G^k$ implies $G_{\mathcal{Z}}^k\|\boldsymbol{w}_t - \boldsymbol{w}^*\| \leq 4G^k\|\boldsymbol{w}_t - \boldsymbol{w}^*\|$ for all t. Thus, $\mathbb{P}_{\mathcal{Z}}\Big(G_{\mathcal{Z}}^k\|\boldsymbol{w}_t - \boldsymbol{w}^*\| \leq 4G^k\|\boldsymbol{w}_t - \boldsymbol{w}^*\|,\ \forall t\Big) \geq \mathcal{P}_{\mathcal{Z}}\Big(G_{\mathcal{Z}}^k \leq 4G^k\Big) \geq \frac{3}{4}$, where the last step follows from Markov's inequality. Thus, for all t, we have:

$$\text{(I)} \leq -2\eta(f(\boldsymbol{w}_t) - f(\boldsymbol{w}^*)) + 16\eta G\Big(\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\Big)^{\frac{1}{2}(1-\frac{1}{k})}\|\boldsymbol{w}_t - \boldsymbol{w}^*\|, \tag{88}$$

with a probability of at least $\frac{3}{4}$ w.r.t. the random dataset $\mathcal{Z}$. Henceforth, we will not mention this and it should be inferred directly.

Case 1: $\|\boldsymbol{w}_t - \boldsymbol{w}^*\| \leq D$.
In this case, we simply have:

$$\text{(I)} \leq -2\eta(f(\boldsymbol{w}_t) - f(\boldsymbol{w}^*)) + 16\eta G\Big(\frac{\nu d \log(1/\delta)}{n^2\varepsilon^2}\Big)^{\frac{1}{2}(1-\frac{1}{k})} D. \tag{89}$$

Case 2: $\|\boldsymbol{w}_t - \boldsymbol{w}^*\| > D$.
In this case, we have:

$$\text{(I)} \leq -\eta(f(\boldsymbol{w}_t) - f(\boldsymbol{w}^*)) - \eta \underbrace{\Big\{(f(\boldsymbol{w}_t) - f(\boldsymbol{w}^*)) - 16G\Big(\frac{\nu d \log(1/\delta)}{n^2\varepsilon^2}\Big)^{\frac{1}{2}(1-\frac{1}{k})}\|\boldsymbol{w}_t - \boldsymbol{w}^*\|\Big\}}_{\geq 0 \text{ using Assumption 6.4}} \tag{90}$$

$$\leq -\eta(f(\boldsymbol{w}_t) - f(\boldsymbol{w}^*)). \tag{91}$$

Combining equations (89) and (90), we have:

$$\text{(I)} \leq -\eta(f(\boldsymbol{w}_t) - f(\boldsymbol{w}^*)) + 16\eta G\Big(\frac{\nu d \log(1/\delta)}{n^2\varepsilon^2}\Big)^{\frac{1}{2}(1-\frac{1}{k})} D. \tag{92}$$

Now plugging Equation (92) in Equation (86), we get:

$$\mathbb{E}_t[\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2] \leq \|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 - \eta(f(\boldsymbol{w}_t) - f(\boldsymbol{w}^*)) + 16\eta G\Big(\frac{\nu d \log(1/\delta)}{n^2\varepsilon^2}\Big)^{\frac{1}{2}(1-\frac{1}{k})} D + 2\eta^2\tau^2\Big(1 + \frac{\nu d T \log(1/\delta)}{n^2\varepsilon^2}\Big). \tag{93}$$

Next, summing the above for $t = 0$ through to $T-1$ after taking expectation throughout, rearranging a bit and then dividing by ηT throughout, we get the following:

$$\frac{1}{T}\sum_{t=0}^{T-1}\Big(\mathbb{E}[f(\boldsymbol{w}_t)] - f(\boldsymbol{w}^*)\Big) \leq \frac{\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2}{\eta T} + 2\eta T\tau^2\left(\frac{\nu d \log(1/\delta)}{n^2\varepsilon^2} + \frac{1}{T}\right) + 16G\Big(\frac{\nu d \log(1/\delta)}{n^2\varepsilon^2}\Big)^{\frac{1}{2}(1-\frac{1}{k})} D. \tag{94}$$

Plugging in our choice of $\eta = \frac{C}{T\tau\sqrt{\frac{1}{T} + \frac{\nu d \log(1/\delta)}{n^2\varepsilon^2}}}$, where $C > 0$ is some constant of our choice, and $\tau = G\Big(\frac{1}{T} + \frac{\nu d \log(1/\delta)}{n^2\varepsilon^2}\Big)^{-\frac{1}{2k}}$, we get:

$$\frac{1}{T}\sum_{t=0}^{T-1}\Big(\mathbb{E}[f(\boldsymbol{w}_t)] - f(\boldsymbol{w}^*)\Big) \leq \Big(\frac{\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2}{C} + 2C\Big)G\Big(\frac{1}{T} + \frac{\nu d \log(1/\delta)}{n^2\varepsilon^2}\Big)^{\frac{1}{2}(1-\frac{1}{k})} + 16G\Big(\frac{\nu d \log(1/\delta)}{n^2\varepsilon^2}\Big)^{\frac{1}{2}(1-\frac{1}{k})} D. \tag{95}$$

Lastly, plugging in $\varphi = \frac{\sqrt{\nu d \log(1/\delta)}}{n\varepsilon}$, noting that $\mathbb{E}[f(\boldsymbol{w}_{\widehat{t}})] - f(\boldsymbol{w}^*) = \frac{1}{T}\sum_{t=0}^{T-1}\Big(\mathbb{E}[f(\boldsymbol{w}_t)] - f(\boldsymbol{w}^*)\Big)$ and using the definition of $\text{OR}(T)$, we get the final result. ■

L. Full Version and Proof of Theorem 6.6

Theorem L.1 (Lower Bound for Unconstrained Convex Case Under Assumption 6.4). *Suppose $\varphi < o(1)$ and $\delta < \exp(-\varepsilon^2)$. There exists a convex loss function ℓ, such that for every (ε, δ)-DP algorithm $\mathcal{A}$ which tries to solve for $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w})$ where f is the average loss for a dataset $\mathcal{Z}$ of n samples drawn from the data distribution $\mathcal{D}$, there exists a choice of $\mathcal{D}$ such that:*

- *f satisfies Assumptions 6.1 and 6.4 (the latter up to constant terms and with a probability of at least $1 - \exp\big(-\mathcal{O}\big(\sqrt{d\log(1/\delta)}/\varepsilon\big)\big)$ w.r.t. $\mathcal{Z}$).*

- $\mathbb{E}_{\mathcal{Z}\sim\mathcal{D}^n,\mathcal{A}}\Big[f(\boldsymbol{w}_{\mathcal{Z}}^{(\mathcal{A})}) - f(\boldsymbol{w}^*)\Big] \geq \Omega\big(\varphi^{1-\frac{1}{k}}\big)$, *where* $\boldsymbol{w}_{\mathcal{Z}}^{(\mathcal{A})}$ *is the output of algorithm* $\mathcal{A}$ *on the dataset* $\mathcal{Z}$.

As mentioned in the main text, even though we follow the proof outline of Theorem 6.4 of Kamath et al. (2021), our proof is more involved. First, we had to use a *non-obvious* loss function ℓ (to obtain the lower bound on) for the *unconstrained* case. Second, since we are in the ERM setting, we have to lower bound the expected training error which is harder than lower bounding the expected generalization error in the SCO setting of Kamath et al. (2021); see the footnote in the line after Equation (124) (i.e., footnote 10) for details. Finally, note that Kamath et al. (2021) derive a bound for $(0,\rho)$-zCDP while our bound is for (ε,δ)-DP. To our knowledge, (ε,δ)-DP does not imply $(0,\tilde{\rho})$-zCDP for some $\tilde{\rho}$ due to which we had to re-derive an important tool in their analysis (specifically, Theorem 1.4 of Kamath et al. (2021)); see Lemma L.3. That is also the reason we had to impose the constraint of $\delta < \exp(-\varepsilon^2)$.

Proof. We shall borrow some ideas from Acharya et al. (2021) and Kamath et al. (2021).

Let $\mathcal{V}$ be a set of d-dimensional points satisfying:

- For all $\boldsymbol{v}\in\mathcal{V}$, $\boldsymbol{v}\in\{0,1\}^d$ and the number of 1's in $\boldsymbol{v}$ is $\frac{d}{2}$.
- For all $\boldsymbol{v},\boldsymbol{v}'\in\mathcal{V}$, $d_{\text{Ham}}(\boldsymbol{v},\boldsymbol{v}')\geq\frac{d}{8}$.

By Lemma 6 of Acharya et al. (2021), there must exist such a $\mathcal{V}$ with cardinality at least $2^{\frac{7d}{128}}$. Note that $\|\boldsymbol{v}\|=\sqrt{d/2}$ for all $\boldsymbol{v}\in\mathcal{V}$.

Next, similar to Kamath et al. (2021), let $Q_{\boldsymbol{v}}$ be a distribution whose support includes $\vec{0}_d$ and $p^{-1/k}\boldsymbol{v}$ for $p=\frac{2\sqrt{d\log(1/\delta)}}{n\varepsilon}<\frac{1}{2}$, such that $\mathbb{P}_{\boldsymbol{x}\sim Q_{\boldsymbol{v}}}(\boldsymbol{x}=\vec{0}_d)=1-p$ and $\mathbb{P}_{\boldsymbol{x}\sim Q_{\boldsymbol{v}}}(\boldsymbol{x}=p^{-1/k}\boldsymbol{v})=p$. Note that:

$$\mathbb{E}_{\boldsymbol{x}\sim Q_{\boldsymbol{v}}}[\boldsymbol{x}]=p^{1-\frac{1}{k}}\boldsymbol{v} \text{ and } \mathbb{E}_{\boldsymbol{x}\sim Q_{\boldsymbol{v}}}[\|\boldsymbol{x}\|^k]=\|\boldsymbol{v}\|^k. \tag{96}$$

Now, let $\ell:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ be defined as:

$$\ell(\boldsymbol{w},\boldsymbol{x})=-\langle\boldsymbol{w},\boldsymbol{x}\rangle+2\|\boldsymbol{x}\|\max\big(\|\boldsymbol{w}\|-1,0\big). \tag{97}$$

We have a dataset $\mathcal{Z}_{\boldsymbol{v}}$ of n i.i.d. samples $\{\boldsymbol{x}_i\}_{i=1}^n$ drawn from $Q_{\boldsymbol{v}}$. Then, we denote the ith sample's loss in $\mathcal{Z}_{\boldsymbol{v}}$ by:

$$f_{\boldsymbol{v},i}(\boldsymbol{w})=\ell(\boldsymbol{w},\boldsymbol{x}_i)=-\langle\boldsymbol{w},\boldsymbol{x}_i\rangle+2\|\boldsymbol{x}_i\|\max\big(\|\boldsymbol{w}\|-1,0\big). \tag{98}$$

Here, the subscript $\boldsymbol{v}$ denotes that $\mathcal{Z}_{\boldsymbol{v}}$ is drawn from $Q_{\boldsymbol{v}}$. Next, the average loss is:

$$f_{\boldsymbol{v}}(\boldsymbol{w})=\frac{1}{n}\sum_{i=1}^n f_{\boldsymbol{v},i}(\boldsymbol{w})=-\langle\boldsymbol{w},\overline{\boldsymbol{x}}\rangle+2\|\overline{\boldsymbol{x}}\|\max\big(\|\boldsymbol{w}\|-1,0\big), \tag{99}$$

where $\overline{\boldsymbol{x}}=\frac{1}{n}\sum_{i=1}^n\boldsymbol{x}_i$. Equation (99) follows because each $\boldsymbol{x}_i=c_i\boldsymbol{v}$, where $c_i=0$ with a probability of $1-p$ and $c_i=p^{-1/k}$ otherwise, due to which $\frac{1}{n}\sum_{i=1}^n\|\boldsymbol{x}_i\|=\Big\|\frac{1}{n}\sum_{i=1}^n\boldsymbol{x}_i\Big\|=\|\overline{\boldsymbol{x}}\|$.

Note that $q(\boldsymbol{w})=\max\big(\|\boldsymbol{w}\|-1,0\big)$ is convex. This is because both $q_1(\boldsymbol{w})=\|\boldsymbol{w}\|-1$ and $q_2(\boldsymbol{w})=0$ are convex and the point-wise maximum of convex functions is also convex. Thus, $f_{\boldsymbol{v}}(\boldsymbol{w})$ is convex.

Now, we shall show that Assumption 6.1 holds. It can be checked that $\|\nabla\ell(\boldsymbol{w},\boldsymbol{x})\|\leq 3\|\boldsymbol{x}\|$ for all $\boldsymbol{w}\in\mathbb{R}^d$. Thus,

$$\mathbb{E}_{\boldsymbol{x}\sim Q_{\boldsymbol{v}}}[\|\nabla\ell(\boldsymbol{w},\boldsymbol{x})\|^k]\leq 3^k\mathbb{E}_{\boldsymbol{x}\sim Q_{\boldsymbol{v}}}[\|\boldsymbol{x}\|^k]=(3\|\boldsymbol{v}\|)^k, \tag{100}$$

where the last step follows from Equation (96). Thus, Assumption 6.1 holds here with $G=3\|\boldsymbol{v}\|=3\sqrt{\frac{d}{2}}$.

Let us first obtain $\arg\min_{\boldsymbol{w}\in\mathbb{R}^d}f_{\boldsymbol{v}}(\boldsymbol{w})$. Note that if $\|\overline{\boldsymbol{x}}\|=0$ (which happens when all the $\boldsymbol{x}_i$'s are $\vec{0}_d$), $f_{\boldsymbol{v}}$ is identically 0 and so any point is a minimizer. So, let us focus on the case of $\|\overline{\boldsymbol{x}}\|>0$; we claim that $\arg\min_{\boldsymbol{w}\in\mathbb{R}^d}f_{\boldsymbol{v}}(\boldsymbol{w})=\hat{\boldsymbol{x}}:=\overline{\boldsymbol{x}}/\|\overline{\boldsymbol{x}}\|$.

Clearly, $\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{w}:\|\boldsymbol{w}\|\leq 1} f_{\boldsymbol{v}}(\boldsymbol{w})$. Also, for any $\boldsymbol{w}$ such that $\|\boldsymbol{w}\| > 1$, we have:

$$f_{\boldsymbol{v}}(\boldsymbol{w}) - f_{\boldsymbol{v}}(\hat{\boldsymbol{x}}) = -\langle \boldsymbol{w}, \overline{\boldsymbol{x}}\rangle + 2\|\overline{\boldsymbol{x}}\|\big(\|\boldsymbol{w}\| - 1\big) + \|\overline{\boldsymbol{x}}\| \tag{101}$$

$$= \Big(- \langle \boldsymbol{w}, \overline{\boldsymbol{x}}\rangle + \|\overline{\boldsymbol{x}}\|\|\boldsymbol{w}\|\Big) + \|\overline{\boldsymbol{x}}\|\|\boldsymbol{w}\| - \|\overline{\boldsymbol{x}}\| \tag{102}$$

$$\geq \|\overline{\boldsymbol{x}}\|(\|\boldsymbol{w}\| - 1) \tag{103}$$

$$> 0. \tag{104}$$

Equation (106) follows from the Cauchy-Schwarz inequality while Equation (104) follows because $\|\boldsymbol{w}\| > 1$. Thus, $f_{\boldsymbol{v}}(\boldsymbol{w}) > f_{\boldsymbol{v}}(\hat{\boldsymbol{x}})$ for all $\boldsymbol{w}$ such that $\|\boldsymbol{w}\| > 1$. Hence, we must have $\arg\min_{\boldsymbol{w}\in\mathbb{R}^d} f_{\boldsymbol{v}}(\boldsymbol{w}) = \hat{\boldsymbol{x}}$. Also, note that since each $\boldsymbol{x}_i = c_i\boldsymbol{v}$ where $c_i \in \{0, p^{-1/k}\}$, $\overline{\boldsymbol{x}} = \Big(\frac{1}{n}\sum_{i=1}^n c_i\Big)\boldsymbol{v}$ with $\frac{1}{n}\sum_{i=1}^n c_i > 0$ and so,

$$\hat{\boldsymbol{x}} = \boldsymbol{v}/\|\boldsymbol{v}\| := \boldsymbol{w}^*_{\boldsymbol{v}}, \tag{105}$$

which is independent of the dataset $\mathcal{Z}_{\boldsymbol{v}}$ that we receive; in fact, $\boldsymbol{w}^*_{\boldsymbol{v}} = \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \mathbb{E}_{\boldsymbol{x}\sim Q_{\boldsymbol{v}}}[\ell(\boldsymbol{w}, \boldsymbol{x})]$. We shall be using these facts later.

Rewriting Equation (103) by replacing $\hat{\boldsymbol{x}}$ with $\boldsymbol{w}^*_{\boldsymbol{v}}$ in the LHS, we get:

$$f_{\boldsymbol{v}}(\boldsymbol{w}) - f_{\boldsymbol{v}}(\boldsymbol{w}^*_{\boldsymbol{v}}) \geq \|\overline{\boldsymbol{x}}\|(\|\boldsymbol{w}\| - 1)\ \forall\ \boldsymbol{w} \in \mathbb{R}^d. \tag{106}$$

Also note that Equation (106) holds trivially for the case of $\|\overline{\boldsymbol{x}}\| = 0$.

Finally, let us show that Assumption 6.4 also holds up to constant factors with high-probability over the dataset $\mathcal{Z}_{\boldsymbol{v}}$. Specifically, we shall restrict our attention to $\|\overline{\boldsymbol{x}}\| \geq \frac{p^{1-\frac{1}{k}}}{2}\|\boldsymbol{v}\|$. Since $\mathbb{E}[\|\overline{\boldsymbol{x}}\|] = p^{1-\frac{1}{k}}\|\boldsymbol{v}\|$, using the Chernoff bound for Binomial random variables[8], we have $\|\overline{\boldsymbol{x}}\| \geq \frac{p^{1-\frac{1}{k}}}{2}\|\boldsymbol{v}\|$ with a probability of at least

$$1 - \exp(-\mathcal{O}(np)) = 1 - \exp\Big(- \mathcal{O}\Big(\frac{\sqrt{d\log(1/\delta)}}{\varepsilon}\Big)\Big), \tag{107}$$

where the last step follows by plugging in $p = \frac{2\sqrt{d\log(1/\delta)}}{n\varepsilon}$.

Now, let us consider all $\boldsymbol{w}$ such that $\|\boldsymbol{w} - \boldsymbol{w}^*_{\boldsymbol{v}}\| \geq 4$ – in this case, since $\|\boldsymbol{w}^*_{\boldsymbol{v}}\| = 1$, we must have $\|\boldsymbol{w}\| \geq 3$. Further

$$\|\boldsymbol{w} - \boldsymbol{w}^*_{\boldsymbol{v}}\| \leq \|\boldsymbol{w}\| + 1 \tag{108}$$

$$\leq 2(\|\boldsymbol{w}\| - 1), \tag{109}$$

where the last step holds for $\|\boldsymbol{w}\| \geq 3$. Equation (108) follows by using the triangle inequality and the fact that $\|\boldsymbol{w}^*_{\boldsymbol{v}}\| = 1$. Using Equation (109) in Equation (106), we get for $\|\boldsymbol{w} - \boldsymbol{w}^*_{\boldsymbol{v}}\| \geq 4$:

$$f_{\boldsymbol{v}}(\boldsymbol{w}) - f_{\boldsymbol{v}}(\boldsymbol{w}^*_{\boldsymbol{v}}) \geq \Omega\Big(\|\overline{\boldsymbol{x}}\|\|\boldsymbol{w} - \boldsymbol{w}^*_{\boldsymbol{v}}\|\Big). \tag{110}$$

Recall that we are restricting our attention to $\|\overline{\boldsymbol{x}}\| \geq \frac{p^{1-\frac{1}{k}}}{2}\|\boldsymbol{v}\|$ which occurs with probability $\geq 1 - \exp\Big(-\mathcal{O}\Big(\frac{\sqrt{d\log(1/\delta)}}{\varepsilon}\Big)\Big)$ (w.r.t. the random dataset $\mathcal{Z}_{\boldsymbol{v}}$). Using this as well as the fact that $\|\boldsymbol{w} - \boldsymbol{w}^*_{\boldsymbol{v}}\|$ is independent of $\mathcal{Z}_{\boldsymbol{v}}$ in Equation (110), and then plugging in the value of $p = \frac{2\sqrt{d\log(1/\delta)}}{n\varepsilon}$ and $G = 3\|v\|$ (as defined after Equation (100)), we get:

$$f_{\boldsymbol{v}}(\boldsymbol{w}) - f_{\boldsymbol{v}}(\boldsymbol{w}^*_{\boldsymbol{v}}) \geq \Omega\Big(p^{1-\frac{1}{k}}\|\boldsymbol{v}\|\|\boldsymbol{w} - \boldsymbol{w}^*_{\boldsymbol{v}}\|\Big) = \Omega\Big(\big(\sqrt{d\log(1/\delta)}/n\varepsilon\big)^{1-\frac{1}{k}} G\|\boldsymbol{w} - \boldsymbol{w}^*_{\boldsymbol{v}}\|\Big), \tag{111}$$

with probability $\geq 1 - \exp\Big(-\mathcal{O}\Big(\frac{\sqrt{d\log(1/\delta)}}{\varepsilon}\Big)\Big)$. Hence, Assumption 6.4 holds up to constant factors with high-probability over the dataset $\mathcal{Z}_{\boldsymbol{v}}$.

[8] Specifically, for n i.i.d. binomial ($\in \{0, 1\}$) random variables $\{Z_i\}_{i=1}^n$ with $\mathbb{P}(Z_i = 1) = \mu\ \forall\ i \in [n]$, we use $\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n Z_i \leq \mu(1-\beta)\Big) \leq \exp(-\frac{n\mu\beta^2}{3})$ for $\beta \in (0, 1)$.

Now that we have shown that $f_{\boldsymbol{v}}$ is convex and satisfies Assumptions 6.1 and 6.4, let us move onto the lower bound. Let us denote the output of the (ε, δ)-DP algorithm $\mathcal{A}$ (on the dataset $\mathcal{Z}_{\boldsymbol{v}}$) by $\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}$[9]. We shall consider two cases.

Case 1, $\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}\| \geq 3$: Using Equation (106), we get

$$f_{\boldsymbol{v}}(\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}) - f_{\boldsymbol{v}}(\boldsymbol{w}_{\boldsymbol{v}}^{*}) \geq 2\|\overline{\boldsymbol{x}}\|. \tag{112}$$

Case 2, $\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}\| < 3$: First, let us consider $\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}\| \in (1, 3)$. Rewriting Equation (101) for $\|\overline{\boldsymbol{x}}\| > 0$ by replacing $\hat{\boldsymbol{x}}$ with $\boldsymbol{w}_{\boldsymbol{v}}^{*}$ in the LHS, we get:

$$f_{\boldsymbol{v}}(\boldsymbol{w}) - f_{\boldsymbol{v}}(\boldsymbol{w}_{\boldsymbol{v}}^{*}) \geq -\langle \boldsymbol{w}, \overline{\boldsymbol{x}} \rangle + 2\|\overline{\boldsymbol{x}}\|\big(\|\boldsymbol{w}\| - 1\big) + \|\overline{\boldsymbol{x}}\| \;\forall\, \boldsymbol{w} \in \mathbb{R}^d. \tag{113}$$

Also note that Equation (113) holds trivially for the case of $\|\overline{\boldsymbol{x}}\| = 0$. Using Equation (113), we get

$$\begin{aligned}
f_{\boldsymbol{v}}(\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}) - f_{\boldsymbol{v}}(\boldsymbol{w}_{\boldsymbol{v}}^{*}) &\geq -\langle \boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}, \overline{\boldsymbol{x}} \rangle + 2\|\overline{\boldsymbol{x}}\|\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}\| - \|\overline{\boldsymbol{x}}\| && (114)\\
&= \|\overline{\boldsymbol{x}}\|\Bigg(-\Big\langle \boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}, \underbrace{\frac{\overline{\boldsymbol{x}}}{\|\overline{\boldsymbol{x}}\|}}_{\boldsymbol{w}_{\boldsymbol{v}}^{*}} \Big\rangle + 2\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}\| - 1\Bigg) && (115)\\
&= \frac{\|\overline{\boldsymbol{x}}\|}{2}\Big(1 - 2\langle \boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}, \boldsymbol{w}_{\boldsymbol{v}}^{*} \rangle + 4\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}\| - 3\Big). && (116)
\end{aligned}$$

Next, using the fact that $\|\boldsymbol{w}_{\boldsymbol{v}}^{*}\|^2 = 1$ and $4\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}\| - 3 \geq \|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}\|^2$ for $\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}\| \in (1, 3)$ above, we get:

$$f_{\boldsymbol{v}}(\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}) - f_{\boldsymbol{v}}(\boldsymbol{w}_{\boldsymbol{v}}^{*}) \geq \frac{\|\overline{\boldsymbol{x}}\|}{2}\Big(\|\boldsymbol{w}_{\boldsymbol{v}}^{*}\|^2 - 2\langle \boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}, \boldsymbol{w}_{\boldsymbol{v}}^{*} \rangle + \|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}\|^2\Big) = \frac{\|\overline{\boldsymbol{x}}\|}{2}\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})} - \boldsymbol{w}_{\boldsymbol{v}}^{*}\|^2. \tag{117}$$

Let us now consider $\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}\| \leq 1$. In this case:

$$\begin{aligned}
f_{\boldsymbol{v}}(\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}) - f_{\boldsymbol{v}}(\boldsymbol{w}_{\boldsymbol{v}}^{*}) &\geq -\langle \boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}, \overline{\boldsymbol{x}} \rangle + \|\overline{\boldsymbol{x}}\| && (118)\\
&= \|\overline{\boldsymbol{x}}\|\Bigg(-\Big\langle \boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}, \underbrace{\frac{\overline{\boldsymbol{x}}}{\|\overline{\boldsymbol{x}}\|}}_{\boldsymbol{w}_{\boldsymbol{v}}^{*}} \Big\rangle + 1\Bigg) && (119)\\
&= \frac{\|\overline{\boldsymbol{x}}\|}{2}\Big(1 + 1 - 2\langle \boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}, \boldsymbol{w}_{\boldsymbol{v}}^{*} \rangle\Big) && (120)\\
&\geq \frac{\|\overline{\boldsymbol{x}}\|}{2}\Big(\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}\|^2 + \|\boldsymbol{w}_{\boldsymbol{v}}^{*}\|^2 - 2\langle \boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}, \boldsymbol{w}_{\boldsymbol{v}}^{*} \rangle\Big) && (121)\\
&= \frac{\|\overline{\boldsymbol{x}}\|}{2}\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})} - \boldsymbol{w}_{\boldsymbol{v}}^{*}\|^2. && (122)
\end{aligned}$$

Equation (121) follows because $\|\boldsymbol{w}_{\boldsymbol{v}}^{*}\| = 1$ and $\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}\| \leq 1$.

So for both $\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}\| \in (1, 3)$ and $\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}\| \leq 1$, we have:

$$f_{\boldsymbol{v}}(\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}) - f_{\boldsymbol{v}}(\boldsymbol{w}_{\boldsymbol{v}}^{*}) \geq \frac{\|\overline{\boldsymbol{x}}\|}{2}\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})} - \boldsymbol{w}_{\boldsymbol{v}}^{*}\|^2. \tag{123}$$

Combining the results of both cases, i.e., Equation (112) and Equation (123), we get:

$$f_{\boldsymbol{v}}(\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}) - f_{\boldsymbol{v}}(\boldsymbol{w}_{\boldsymbol{v}}^{*}) \geq 2\|\overline{\boldsymbol{x}}\| \min\Big(1, \frac{1}{4}\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})} - \boldsymbol{w}_{\boldsymbol{v}}^{*}\|^2\Big). \tag{124}$$

[9] A better choice of notation would have been to put $\mathcal{Z}_{\boldsymbol{v}}$ instead of just $\boldsymbol{v}$ in the subscript; however, we do not do so to avoid overloading notation going further.

Now taking expectation w.r.t. $Q_{\boldsymbol{v}}$ and $\mathcal{A}$ (henceforth, we shall omit these in the subscript of expectation for brevity), we get[10]:

$$\mathbb{E}\Big[f_{\boldsymbol{v}}(\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}) - f_{\boldsymbol{v}}(\boldsymbol{w}_{\boldsymbol{v}}^{*})\Big] \geq 2\mathbb{E}\Big[\|\overline{\boldsymbol{x}}\| \min\Big(1, \frac{1}{4}\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})} - \boldsymbol{w}_{\boldsymbol{v}}^{*}\|^2\Big)\Big] \tag{125}$$

$$\geq 2\mathbb{E}\Bigg[\|\overline{\boldsymbol{x}}\| \min\Big(1, \frac{1}{4}\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})} - \boldsymbol{w}_{\boldsymbol{v}}^{*}\|^2\Big)\Big|\|\overline{\boldsymbol{x}}\| \geq \frac{p^{1-\frac{1}{k}}}{2}\|\boldsymbol{v}\|\Bigg] \underbrace{\mathbb{P}\Big(\|\overline{\boldsymbol{x}}\| \geq \frac{p^{1-\frac{1}{k}}}{2}\|\boldsymbol{v}\|\Big)}_{=\Theta(1)} \tag{126}$$

$$\geq \Omega\big(p^{1-\frac{1}{k}}\|\boldsymbol{v}\|\big)\mathbb{E}\Bigg[\min\Big(1, \frac{1}{4}\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})} - \boldsymbol{w}_{\boldsymbol{v}}^{*}\|^2\Big)\Big|\|\overline{\boldsymbol{x}}\| \geq \frac{p^{1-\frac{1}{k}}}{2}\|\boldsymbol{v}\|\Bigg]. \tag{127}$$

To obtain Equation (127), we have used the fact that $\mathbb{P}\Big(\|\overline{\boldsymbol{x}}\| \geq \frac{p^{1-\frac{1}{k}}}{2}\|\boldsymbol{v}\|\Big) = \Theta(1)$ from Equation (107). Next, letting $M_{\boldsymbol{v},\mathcal{A}} = \min\Big(1, \frac{1}{4}\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})} - \boldsymbol{w}_{\boldsymbol{v}}^{*}\|^2\Big)$, we have:

$$\mathbb{E}[M_{\boldsymbol{v},\mathcal{A}}] = \mathbb{E}\Bigg[M_{\boldsymbol{v},\mathcal{A}}\Big|\|\overline{\boldsymbol{x}}\| \geq \frac{p^{1-\frac{1}{k}}}{2}\|\boldsymbol{v}\|\Bigg]\mathbb{P}\Big(\|\overline{\boldsymbol{x}}\| \geq \frac{p^{1-\frac{1}{k}}}{2}\|\boldsymbol{v}\|\Big) + \mathbb{E}\Bigg[M_{\boldsymbol{v},\mathcal{A}}\Big|\|\overline{\boldsymbol{x}}\| < \frac{p^{1-\frac{1}{k}}}{2}\|\boldsymbol{v}\|\Bigg]\mathbb{P}\Big(\|\overline{\boldsymbol{x}}\| < \frac{p^{1-\frac{1}{k}}}{2}\|\boldsymbol{v}\|\Big). \tag{128}$$

Now note that $\mathbb{E}\Bigg[M_{\boldsymbol{v},\mathcal{A}}\Big|\|\overline{\boldsymbol{x}}\| < \frac{p^{1-\frac{1}{k}}}{2}\|\boldsymbol{v}\|\Bigg] \leq 1$ (as $M_{\boldsymbol{v},\mathcal{A}} \leq 1$ by definition) and from Equation (107), $\mathbb{P}\Big(\|\overline{\boldsymbol{x}}\| < \frac{p^{1-\frac{1}{k}}}{2}\|\boldsymbol{v}\|\Big) \leq \exp\Big(-\mathcal{O}\Big(\frac{\sqrt{d\log(1/\delta)}}{\varepsilon}\Big)\Big)$. Also, from Lemma L.2, there exists a $\boldsymbol{v}$, say $\hat{\boldsymbol{v}}$, for which $\mathbb{E}[M_{\hat{\boldsymbol{v}},\mathcal{A}}] = \Omega(1)$ for $\delta < \exp(-\varepsilon^2)$. Thus, for $\boldsymbol{v} = \hat{\boldsymbol{v}}$:

$$\mathbb{E}\Bigg[M_{\hat{\boldsymbol{v}},\mathcal{A}}\Big|\|\overline{\boldsymbol{x}}\| \geq \frac{p^{1-\frac{1}{k}}}{2}\|\hat{\boldsymbol{v}}\|\Bigg] = \Theta\Big(\mathbb{E}[M_{\hat{\boldsymbol{v}},\mathcal{A}}]\Big) = \Omega(1). \tag{129}$$

Putting this back in Equation (127), we get for $\boldsymbol{v} = \hat{\boldsymbol{v}}$:

$$\mathbb{E}\Big[f_{\hat{\boldsymbol{v}}}(\boldsymbol{w}_{\hat{\boldsymbol{v}}}^{(\mathcal{A})}) - f_{\hat{\boldsymbol{v}}}(\boldsymbol{w}_{\hat{\boldsymbol{v}}}^{*})\Big] \geq \Omega\Big(p^{1-\frac{1}{k}}\|\hat{\boldsymbol{v}}\|\Big) = \Omega\Big(\Big(\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon}\Big)^{1-\frac{1}{k}}G\Big), \tag{130}$$

where the last step follows by plugging in the values of p and G.

This finishes the proof. ■

Lemma L.2. *In the setting of the proof of Theorem L.1 and for $\delta < \exp(-\varepsilon^2)$, there exists some $\hat{\boldsymbol{v}} \in \mathcal{V}$ such that:*

$$\mathbb{E}_{Q_{\hat{\boldsymbol{v}}},\mathcal{A}}\Bigg[\min\Big(1, \frac{1}{4}\|\boldsymbol{w}_{\hat{\boldsymbol{v}}}^{(\mathcal{A})} - \boldsymbol{w}_{\hat{\boldsymbol{v}}}^{*}\|^2\Big)\Bigg] \geq \Omega(1). \tag{131}$$

Proof. From Lemma 22 of Bun & Steinke (2016), note that (ε, δ)-DP implies $\big(5(\frac{\varepsilon^2}{\log(1/\delta)})^{1/4} - \frac{\varepsilon^2}{4\log(1/\delta)}, \frac{\varepsilon^2}{4\log(1/\delta)}\big)$-zCDP; this follows by setting $\hat{\xi} = 0$ & $\hat{\rho} = \frac{\varepsilon^2}{\log(1/\delta)}$ in that lemma. Since $\hat{\rho} < 1$ in that lemma, we impose the constraint of $\delta < \exp(-\varepsilon^2)$.

[10] Lower bounding $\mathbb{E}\big[f_{\boldsymbol{v}}(\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}) - f_{\boldsymbol{v}}(\boldsymbol{w}_{\boldsymbol{v}}^{*})\big]$, i.e. the training error, is harder than lower bounding $\mathbb{E}\big[\ell(\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}) - \ell(\boldsymbol{w}_{\boldsymbol{v}}^{*})\big]$, i.e. the generalization error. This is because the lower bound for the training error includes $\|\overline{\boldsymbol{x}}\|$ and $\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}$, both of which depend on the dataset $\mathcal{Z}$, making the expectation of the lower bound challenging to compute. In contrast, in the lower bound for the generalization error, $\|\overline{\boldsymbol{x}}\|$ is replaced by $\|\mathbb{E}_{\boldsymbol{x}\sim Q_{\boldsymbol{v}}}[\boldsymbol{x}]\|$ and so, $\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}$ is the only quantity depending on $\mathcal{Z}$, making the expectation of the corresponding lower bound much simpler to compute.

Now, we shall use Lemma L.3 (stated and proved after this proof). In Lemma L.3, let us use the loss function $\widetilde{l}(\boldsymbol{w}, \boldsymbol{w}') = \min\left(1, \frac{1}{4}\|\boldsymbol{w} - \boldsymbol{w}'\|^2\right)$. As mentioned after Equation (105), $\boldsymbol{w}_{\boldsymbol{v}}^* = \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}$ is the minimizer of the expected loss ℓ over the distribution $Q_{\boldsymbol{v}}$. Now:

$$\widetilde{l}(\boldsymbol{w}_{\boldsymbol{v}}^*, \boldsymbol{w}_{\boldsymbol{v}'}^*) = \min\left(1, \frac{1}{4}\|\boldsymbol{w}_{\boldsymbol{v}}^* - \boldsymbol{w}_{\boldsymbol{v}'}^*\|^2\right) = \min\left(1, \frac{1}{4}\left\|\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} - \frac{\boldsymbol{v}'}{\|\boldsymbol{v}'\|}\right\|^2\right). \tag{132}$$

Using the fact that $\|\boldsymbol{v}\| = \|\boldsymbol{v}'\| = \sqrt{\frac{d}{2}}$ and $d_{\text{Ham}}(\boldsymbol{v}, \boldsymbol{v}') \geq \frac{d}{8}$ for all $\boldsymbol{v} \neq \boldsymbol{v}'$, we have that $\widetilde{l}(\boldsymbol{w}_{\boldsymbol{v}}^*, \boldsymbol{w}_{\boldsymbol{v}'}^*) \geq \Omega(1)$ for all $\boldsymbol{v} \neq \boldsymbol{v}'$. Also, $d_{TV}(Q_{\boldsymbol{v}}, Q_{\boldsymbol{v}'}) = p = \frac{2\sqrt{d\log(1/\delta)}}{n\varepsilon}$ by definition and $\log|\mathcal{V}| = \mathcal{O}(d)$. Using all of this in Lemma L.3, we get:

$$\frac{1}{|\mathcal{V}|}\sum_{\boldsymbol{v}\in\mathcal{V}} \mathbb{E}\Big[\widetilde{l}(\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}, \boldsymbol{w}_{\boldsymbol{v}}^*)\Big] \geq \Omega(1). \tag{133}$$

Since the average over all $\boldsymbol{v}$ is $\Omega(1)$, there must exist some $\boldsymbol{v}$, say $\hat{\boldsymbol{v}}$, for which:

$$\mathbb{E}\Big[\widetilde{l}(\boldsymbol{w}_{\hat{\boldsymbol{v}}}^{(\mathcal{A})}, \boldsymbol{w}_{\hat{\boldsymbol{v}}}^*)\Big] \geq \Omega(1). \tag{134}$$

This completes the proof. ■

Lemma L.3 (**(ξ, ρ)-zCDP Fano's inequality**: based on Thm. 1.4 of Kamath et al. (2021))**.** *Let $\{p_1, \ldots, p_M\} \subseteq \mathcal{P}$ be a set of probability distributions, $\boldsymbol{\theta} : \mathcal{P} \to \mathbb{R}^d$ be a parameter of interest, and $\tilde{l} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a loss function. Suppose for all $i \neq j$, it satisfies (a) $\tilde{l}(\boldsymbol{\theta}(p_i), \boldsymbol{\theta}(p_j)) \geq r$, (b) $d_{TV}(p_i, p_j) \leq \alpha$, (c) $d_{KL}(p_i, p_j) \leq \beta$. Then for any (ξ, ρ)-zCDP estimator $\hat{\boldsymbol{\theta}}$:*

$$\frac{1}{M}\sum_{i\in M} \mathbb{E}_{\boldsymbol{X}\sim p_i^n}\left[\tilde{l}\big(\hat{\boldsymbol{\theta}}(X), \boldsymbol{\theta}(p_i)\big)\right] \geq$$
$$\frac{r}{2}\max\Big\{1 - \frac{\beta + \log 2}{\log M}, 1 - \frac{\rho(n^2\alpha^2 + n\alpha(1-\alpha)) + \xi(un\alpha\log(eun\alpha) + n\log(en)\exp(-\frac{n\alpha(u-1)^2}{3})) + \log 2}{\log M}\Big\},$$
$$\textit{for any } u > 1. \tag{135}$$

Compared to Theorem 1.4 of Kamath et al. (2021) (who only consider the case of $\xi = 0$), note the extra term in red.

Proof. The proof is largely based on the proof of Theorem 1.4 of Kamath et al. (2021); we mention the changes required (while following their notation).

The first change is that instead of $d_{\text{KL}}(\hat{p}(x), \hat{p}(x')) \leq \rho d_{\text{Ham}}(x, x')^2$ (in Kamath et al. (2021)), here we have $d_{\text{KL}}(\hat{p}(x), \hat{p}(x')) \leq \xi d_{\text{Ham}}(x, x')\log(e.d_{\text{Ham}}(x, x')) + \rho d_{\text{Ham}}(x, x')^2$; this follows using Proposition 27 of Bun & Steinke (2016) and the fact that $\sum_{i=1}^k \frac{1}{i} \leq \log(e.k)$. Secondly, due to the extra term in blue, we need to bound $\mathbb{E}_{z\sim\text{Bin}(n,\alpha)}[z\log(e.z)]$ (this is because $d_{\text{Ham}}(x, x') \sim \text{Bin}(n, \alpha)$ as discussed in the proof of Theorem 1.4 of Kamath et al. (2021)). For any $u > 1$, we have:

$$\mathbb{E}[z\log(e.z)] \leq \mathbb{E}[z\log(e.z)|z < un\alpha] + \mathbb{E}[z\log(e.z)|z \geq un\alpha]\mathbb{P}(z \geq un\alpha) \tag{136}$$
$$\leq un\alpha\log(eun\alpha) + n\log(en)\exp\Big(-\frac{n\alpha(u-1)^2}{3}\Big), \tag{137}$$

where the last step uses the Chernoff bound. Combining all of this and following the rest of the steps in the proof of Theorem 1.4 of Kamath et al. (2021), we get the desired result. ■

M. Proof of Theorem I.3

The proof of Theorem I.3 follows from that of Theorem L.1 by just restricting ourselves to $\|\boldsymbol{w}\| \leq 1$ (as well as $\|\boldsymbol{w}_{\boldsymbol{v}}^{(\mathcal{A})}\| \leq 1$).

N. Full Version and Proof of Theorem 6.8

Theorem N.1 (**Unconstrained Nonconvex Case**). *Suppose Assumption 6.1 holds, f is L-smooth and $\mathcal{W} = \mathbb{R}^d$. Fix some $\gamma \in (0,1)$ and $C > 0$. In Algorithm 1, set $\tau = G\Big(\frac{G}{\gamma^2 C\sqrt{L}}\Big)^{\frac{1}{2k-1}}\Big(\frac{1}{T}+\varphi^2\Big)^{-\frac{1}{2(2k-1)}}$ and $\eta_t = \eta = \frac{C}{T\tau\sqrt{L}}\Big(\frac{1}{T}+\varphi^2\Big)^{-\frac{1}{2}}$ for all $t < T$. Then with a probability of at least $(1-\gamma)$ which is w.r.t. the random dataset $\mathcal{Z}$ that we obtain, DP-SGD (Algorithm 1) has the following guarantee:*

$$\mathrm{OR}(T) \leq \frac{(\sqrt{L})^{1-\frac{1}{2k-1}} G^{1+\frac{1}{2k-1}}}{\gamma^{\frac{2}{2k-1}} C^{\frac{1}{2k-1}}}\left(3C + \frac{2(f(\boldsymbol{w}_0) - \min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w}))}{C}\right)\Big(\frac{1}{T}+\varphi^2\Big)^{\frac{1}{2}(1-\frac{1}{2k-1})}.$$

So if we set $T = \frac{1}{\varphi^2}$ above (which is what we do in Theorem 6.8), we get the following bound for the risk:

$$\mathrm{OR}(T) \leq \frac{(\sqrt{2L})^{1-\frac{1}{2k-1}} G^{1+\frac{1}{2k-1}}}{\gamma^{\frac{2}{2k-1}} C^{\frac{1}{2k-1}}}\left(3C + \frac{2(f(\boldsymbol{w}_0) - \min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w}))}{C}\right)\varphi^{(1-\frac{1}{2k-1})}.$$

We prove this below.

Proof:

Proof. From Lemma I.5, recall that

$$\Big\|\frac{1}{n}\sum_{i\in[n]} \mathrm{clip}(\nabla f_i(\boldsymbol{w}),\tau) - \nabla f(\boldsymbol{w})\Big\| \leq \frac{G_{\mathcal{Z}}^k}{\tau^{k-1}}, \tag{138}$$

where $G_{\mathcal{Z}}^k = \frac{1}{n}\sum_{i=1}^n (G(\boldsymbol{x}_i, y_i))^k$.

Using the L-smoothness of f and taking expectation only with respect to the randomness in the current iteration, we have:

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{w}_{t+1})] &\leq f(\boldsymbol{w}_t) - \eta\mathbb{E}[\langle\nabla f(\boldsymbol{w}_t), \boldsymbol{g}_t\rangle] + \frac{\eta^2 L}{2}\mathbb{E}[\|\boldsymbol{g}_t\|^2] && (139)\\
&= f(\boldsymbol{w}_t) - \eta\Big[\Big\langle\nabla f(\boldsymbol{w}_t), \frac{1}{n}\sum_{i=1}^n \mathrm{clip}(\nabla f_i(\boldsymbol{w}_t),\tau)\Big\rangle\Big] + \eta^2 L\tau^2\Big(1 + \frac{\nu d T\log(\frac{1}{\delta})}{n^2\varepsilon^2}\Big) && (140)\\
&= f(\boldsymbol{w}_t) - \frac{\eta}{2}\left\{\Big\|\frac{1}{n}\sum_{i=1}^n \mathrm{clip}(\nabla f_i(\boldsymbol{w}_t),\tau)\Big\|^2 + \|\nabla f(\boldsymbol{w}_t)\|^2 - \Big\|\frac{1}{n}\sum_{i=1}^n \mathrm{clip}(\nabla f_i(\boldsymbol{w}_t),\tau) - \nabla f(\boldsymbol{w}_t)\Big\|^2\right\} && (141)\\
&\quad + \eta^2 L\tau^2\Big(1 + \frac{\nu d T\log(\frac{1}{\delta})}{n^2\varepsilon^2}\Big)\\
&\leq f(\boldsymbol{w}_t) - \frac{\eta}{2}\|\nabla f(\boldsymbol{w}_t)\|^2 + \frac{\eta}{2}\Big(\frac{G_{\mathcal{Z}}^{2k}}{\tau^{2(k-1)}}\Big) + \eta^2 L\tau^2\Big(1 + \frac{\nu d T\log(\frac{1}{\delta})}{n^2\varepsilon^2}\Big). && (142)
\end{aligned}$$

In Equation (140), we have used Equation (40). Equation (141) follows by using the fact for any two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, $\langle\boldsymbol{a},\boldsymbol{b}\rangle = \frac{1}{2}\Big(\|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 - \|\boldsymbol{a}-\boldsymbol{b}\|^2\Big)$. Equation (142) is obtained by using Equation (138).

Next, summing up the above for $t = 0$ through to $T-1$, taking expectation throughout and then after rearranging a bit and using the fact that $\mathbb{E}[f(\boldsymbol{w}_T)] \geq f^* = \min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w})$, we get:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\boldsymbol{w}_t)\|^2] \leq \frac{2(f(\boldsymbol{w}_0) - f^*)}{\eta T} + 2\eta T L\tau^2\Big(\frac{1}{T} + \frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\Big) + \frac{G_{\mathcal{Z}}^{2k}}{\tau^{2(k-1)}}. \tag{143}$$

Let us plug in $\eta = \frac{C}{T\tau\sqrt{L}\sqrt{\frac{1}{T}+\frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}}}$ above, where $C > 0$ is a constant of our choice. With that, we get:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\boldsymbol{w}_t)\|^2] \leq \underbrace{\Big(\frac{2(f(\boldsymbol{w}_0) - f^*)}{C} + 2C\Big)}_{:=C'}\sqrt{L}\tau\sqrt{\frac{1}{T} + \frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}} + \frac{G_{\mathcal{Z}}^{2k}}{\tau^{2(k-1)}}. \tag{144}$$

Let us now choose $\tau = G\Big(\frac{G}{\gamma^2 C\sqrt{L}}\Big)^{\frac{1}{2k-1}}\Big(\frac{1}{T} + \frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\Big)^{-\frac{1}{2(2k-1)}}$, where $\gamma \in (0,1)$. That gives us:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\boldsymbol{w}_t)\|^2] \leq \frac{(\sqrt{L})^{1-\frac{1}{2k-1}}G^{1+\frac{1}{2k-1}}}{\gamma^{\frac{2}{2k-1}}}\left(\frac{C'}{C^{\frac{1}{2k-1}}} + C^{1-\frac{1}{2k-1}}\Big(\frac{G_{\mathcal{Z}}^k}{G^k/\gamma}\Big)^2\right)\Big(\frac{1}{T} + \frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\Big)^{\frac{1}{2}(1-\frac{1}{2k-1})}. \quad (145)$$

Now, using Markov's inequality, $G_{\mathcal{Z}}^k \leq \frac{G^k}{\gamma}$ with a probability of at least $1-\gamma$ w.r.t. the random dataset $\mathcal{Z}$. Plugging this above, we get:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\boldsymbol{w}_t)\|^2] \leq \frac{(\sqrt{L})^{1-\frac{1}{2k-1}}G^{1+\frac{1}{2k-1}}}{\gamma^{\frac{2}{2k-1}}}\left(\frac{C'}{C^{\frac{1}{2k-1}}} + C^{1-\frac{1}{2k-1}}\right)\Big(\frac{1}{T} + \frac{\nu d\log(1/\delta)}{n^2\varepsilon^2}\Big)^{\frac{1}{2}(1-\frac{1}{2k-1})}, \quad (146)$$

with a probability of at least $1-\gamma$ w.r.t. the random dataset $\mathcal{Z}$.

Lastly, plugging in $\varphi = \frac{\sqrt{\nu d\log(1/\delta)}}{n\varepsilon}$ and the value of C', noting that $\mathbb{E}[\|\nabla f(\boldsymbol{w}_{\widehat{t}})\|^2] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\boldsymbol{w}_t)\|^2]$ and using the definition of $\mathrm{OR}(T)$, we get the final result. ■