Proceedings of Machine Learning Research vol 201:1–69, 2023 34th International Conference on Algorithmic Learning Theory

Private Stochastic Optimization with Large Worst-Case Lipschitz Parameter: Optimal Rates for (Non-Smooth) Convex Losses and Extension to Non-Convex Losses

Andrew Lowy LOWYA@USC.EDU

Meisam Razaviyayn RAZAVIYA@USC.EDU
University of Southern California

Editors: Shipra Agrawal and Francesco Orabona

Abstract

We study differentially private (DP) stochastic optimization (SO) with loss functions whose worst-case Lipschitz parameter over all data points may be extremely large. To date, the vast majority of work on DP SO assumes that the loss is uniformly Lipschitz continuous over data (i.e. stochastic gradients are uniformly bounded over all data points). While this assumption is convenient, it often leads to pessimistic excess risk bounds. In many practical problems, the worst-case (uniform) Lipschitz parameter of the loss over all data points may be extremely large due to outliers. In such cases, the error bounds for DP SO, which scale with the worst-case Lipschitz parameter of the loss, are vacuous. To address these limitations, this work provides near-optimal excess risk bounds that do not depend on the uniform Lipschitz parameter of the loss. Building on a recent line of work (Wang et al., 2020; Kamath et al., 2022), we assume that stochastic gradients have bounded k-th order *moments* for some $k \geqslant 2$. Compared with works on uniformly Lipschitz DP SO, our excess risk scales with the k-th moment bound instead of the uniform Lipschitz parameter of the loss, allowing for significantly faster rates in the presence of outliers and/or heavy-tailed data. For *convex* and *strongly convex* loss functions, we provide the first asymptotically *optimal* excess risk bounds (up to a logarithmic factor). In contrast to (Wang et al., 2020; Kamath et al., 2022), our bounds do not require the loss function to be differentiable/smooth. We also devise an accelerated algorithm for smooth losses that runs in linear time and has excess risk that is tight in certain practical parameter regimes. Additionally, our work is the first to address *non-convex* non-uniformly Lipschitz loss functions satisfying the *Proximal-PL inequality*; this covers some practical machine learning models. Our Proximal-PL algorithm has near-optimal excess risk.

1. Introduction

As the use of machine learning (ML) models in industry and society has grown dramatically in recent years, so too have concerns about the privacy of personal data that is used in training such models. It is well-documented that ML models may leak training data, e.g., via model inversion attacks and membership-inference attacks (Fredrikson et al., 2015; Shokri et al., 2017; Faizullabhoy and Korolova, 2018; Nasr et al., 2019; Carlini et al., 2021). *Differential privacy* (DP) (Dwork et al., 2006) ensures that data cannot be leaked, and a plethora of work has been devoted to differentially private machine learning and optimization (Chaudhuri and Monteleoni, 2008; Duchi et al., 2013; Bassily et al., 2014; Ullman, 2015; Wang et al., 2017;

Bassily et al., 2019; Feldman et al., 2020a; Lowy and Razaviyayn, 2021b; Cheu et al., 2021; Asi et al., 2021b). Of particular importance is the fundamental problem of DP *stochastic (convex) optimization* (S(C)O): given n i.i.d. samples $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$ from an unknown distribution $\mathcal{D}$, we aim to privately solve

$$\min_{w \in \mathcal{W}} \left\{ F(w) := \mathbb{E}_{x \sim \mathcal{D}}[f(w, x)] \right\}, \tag{1}$$

where $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ is the loss function and $\mathcal{W} \subset \mathbb{R}^d$ is the parameter domain. Since finding the exact solution to (1) is not generally possible, we measure the quality of the obtained solution via *excess risk* (a.k.a. excess population loss): The excess risk of a (randomized) algorithm $\mathcal{A}$ for solving (1) is defined as $\mathbb{E}F(\mathcal{A}(X)) - \min_{w \in \mathcal{W}} F(w)$, where the expectation is taken over both the random draw of the data X and the algorithm $\mathcal{A}$.

A large body of literature is devoted to characterizing the optimal achievable differentially private excess risk of (1) when the function $f(\cdot, x)$ is uniformly L_f-*Lipschitz* for all $x \in \mathcal{X}$—see e.g., (Bassily et al., 2019; Feldman et al., 2020a; Asi et al., 2021b; Bassily et al., 2021b; Lowy and Razaviyayn, 2021b). In these works, the gradient of f is assumed to be uniformly bounded with $\sup_{w \in \mathcal{W}, x \in \mathcal{X}} \|\nabla_w f(w, x)\| \leqslant L_f$, and excess risk bounds scale with L_f. While this assumption is convenient for bounding the *sensitivity* (Dwork et al., 2006) of the steps of the algorithm, it is often unrealistic in practice or leads to pessimistic excess risk bounds. In many practical applications, data contains outliers, is unbounded or heavy-tailed (see e.g. (Crovella et al., 1998; Markovich, 2008; Woolson and Clarke, 2011) and references therein for such applications). Consequently, L_f may be prohibitively large. For example, even the linear regression loss $f(w, x) = \frac{1}{2}(\langle w, x^{(1)} \rangle - x^{(2)})^2$ with compact $\mathcal{W}$ and data from $\mathcal{X} = \mathcal{X}^{(1)} \times \mathcal{X}^{(2)}$, leads to $L_f \geqslant \text{diameter}(\mathcal{X}^{(1)})^2$, which could be huge or even infinite. Similar observations can be made for other useful ML models such as deep neural nets (Lei and Ying, 2021), and the situation becomes even grimmer in the presence of heavy-tailed data. In these cases, existing excess risk bounds, which scale with L_f, becomes vacuous.

While L_f can be very large in practice (due to outliers), the k-th *moment* of the stochastic gradients is often reasonably small for some $k \geqslant 2$ (see, e.g., Example 1). This is because the k-th moment $\widetilde{r}_k := \mathbb{E}\left[\sup_{w \in \mathcal{W}} \|\nabla_w f(w, x)\|_2^k\right]^{1/k}$ depends on the *average* behavior of the stochastic gradients, while L_f depends on the *worst-case* behavior over all data points. Motivated by this observation and building on the prior results (Wang et al., 2020; Kamath et al., 2022), this work characterizes the optimal differentially private excess risk bounds for the class of problems with a given parameter $\widetilde{r}_k$. Specifically, for the class of problems with parameter $\widetilde{r}_k$, we answer the following questions (up to a logarithmic factor):

- Question I: What are the minimax optimal rates for (strongly) convex DP SO?
- Question II: What utility guarantees are achievable for non-convex DP SO?

Prior works have made progress in addressing the first question above:[1] The work of (Wang et al., 2020) provided the first excess risk upper bounds for *smooth* DP (strongly)

1. (Wang et al., 2020; Kamath et al., 2022) consider a slightly different problem class than the class $\widetilde{r}_k$, which we consider: see Appendix B. However, our results imply asymptotically optimal rates for the problem class considered in (Wang et al., 2020; Kamath et al., 2022) under mild assumptions: see Appendix E.4.

convex SO. (Kamath et al., 2022) gave improved, yet suboptimal, upper bounds for *smooth* (strongly) convex $f(\cdot, x)$, and lower bounds for (strongly) convex SO. In this work, we provide *optimal algorithms for convex and strongly convex* losses, resolving Question I up to logarithmic factors. Our bounds hold even for *non-differentiable/non-smooth* F. Regarding Question II, we give the *first algorithm for DP SO with non-convex loss* functions satisfying the Proximal Polyak-Łojasiewicz (PL) condition (Polyak, 1963; Karimi et al., 2016). We provide a summary of our results for the case $k = 2$ in Figure 1, and a thorough discussion of related work in Appendix A.

1.1. Preliminaries

Let $\|\cdot\|$ be the ℓ_2 norm. Let $\mathcal{W}$ be a convex, compact set of ℓ_2 diameter D. Function $g : \mathcal{W} \to \mathbb{R}$ is *μ-strongly convex* if $g(\alpha w + (1-\alpha)w') \leqslant \alpha g(w) + (1-\alpha)g(w') - \frac{\alpha(1-\alpha)\mu}{2}\|w - w'\|^2$ for all $\alpha \in [0,1]$ and all $w, w' \in \mathcal{W}$. If $\mu = 0$, we say g is *convex.* For convex $f(\cdot, x)$, denote any *subgradient* of $f(w,x)$ w.r.t. w by $\nabla f(w,x) \in \partial_w f(w,x)$: i.e. $f(w',x) \geqslant f(w,x) + \langle \nabla f(w,x), w' - w\rangle$ for all $w' \in \mathcal{W}$. Function g is *β-smooth* if it is differentiable and its derivative ∇g is β-Lipschitz. For β-smooth, μ-strongly convex g, denote its *condition number* by $\kappa = \beta/\mu$. For functions a and b of input parameters, write $a \lesssim b$ if there is an absolute constant A such that $a \leqslant Ab$ for all feasible values of input parameters. Write $a = \widetilde{\mathcal{O}}(b)$ if $a \lesssim \ell b$ for a logarithmic function ℓ of input parameters. We assume that the stochastic gradient distributions have bounded k-th moment for some $k \geqslant 2$:

Assumption 1 *There exists $k \geqslant 2$ and $\widetilde{r}^{(k)} > 0$ such that $\mathbb{E}\left[\sup_{w\in\mathcal{W}} \|\nabla f(w,x)\|_2^k\right] \leqslant \widetilde{r}^{(k)}$ for all $\nabla f(w, x_i) \in \partial_w f(w, x_i)$. Denote $\widetilde{r}_k := (\widetilde{r}^{(k)})^{1/k}$.*

Clearly, $\widetilde{r}_k \leqslant L_f = \sup_{\{\nabla f(w,x) \in \partial_w f(w,x)\}} \sup_{w,x} \|\nabla f(w,x)\|$, but this inequality is often very loose:

Example 1 *For linear regression on a unit ball $\mathcal{W}$ with 1-dimensional data $x^{(1)}, x^{(2)} \in [-10^{10}, 10^{10}]$ having Truncated Normal distributions and $Var(x^{(1)}) = Var(x^{(2)}) \leqslant 1$, we have $L_f \geqslant 10^{20}$. On the other hand, $\widetilde{r}_k$ is much smaller than L_f for $k < \infty$: e.g., $\widetilde{r}_2 \leqslant 5$, $\widetilde{r}_4 \leqslant 8$, and $\widetilde{r}_8 \leqslant 14$.*

Differential Privacy: *Differential privacy* (Dwork et al., 2006) ensures that no adversary—even one with enormous resources—can infer much more about any person who contributes training data than if that person's data were absent. If two data sets X and X' differ in a single entry (i.e. $d_{\text{hamming}}(X, X') = 1$), then we say that X and X' are *adjacent.*

Definition 1 (Differential Privacy) *Let $\epsilon \geqslant 0$, $\delta \in [0, 1)$. A randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$ is (ϵ, δ)-differentially private (DP) if for all pairs of adjacent data sets $X, X' \in \mathcal{X}^n$ and all measurable subsets $S \subseteq \mathcal{W}$, we have $\mathbb{P}(\mathcal{A}(X) \in S) \leqslant e^{\epsilon}\mathbb{P}(\mathcal{A}(X') \in S) + \delta$.*

In this work, we focus on *zero-concentrated differential privacy* (Bun and Steinke, 2016):

Definition 2 (Zero-Concentrated Differential Privacy (zCDP)) *A randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$ satisfies ρ-zero-concentrated differential privacy (ρ-zCDP) if for all pairs of adjacent data sets $X, X' \in \mathcal{X}^n$ and all $\alpha \in (1, \infty)$, we have $D_\alpha(\mathcal{A}(X)||\mathcal{A}(X')) \leqslant \rho\alpha$, where $D_\alpha(\mathcal{A}(X)||\mathcal{A}(X'))$ is the α-Rényi divergence[2] between the distributions of $\mathcal{A}(X)$ and $\mathcal{A}(X')$.*

2. For distributions P and Q with probability density/mass functions p and q, $D_\alpha(P||Q) := \frac{1}{\alpha-1}\ln\left(\int p(x)^\alpha q(x)^{1-\alpha} dx\right)$ (Rényi, 1961, Eq. 3.3).

zCDP is weaker than $(\epsilon, 0)$-DP, but stronger than (ϵ, δ)-DP ($\delta > 0$) in the following sense:

Proposition 3 *(Bun and Steinke, 2016, Proposition 1.3) If $\mathcal{A}$ is ρ-zCDP, then $\mathcal{A}$ is $(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta)$ for any $\delta > 0$.*

Thus, if $\epsilon \leqslant \sqrt{\log(1/\delta)}$, then any $\frac{\epsilon^2}{2}$-zCDP algorithm is $(2\epsilon\sqrt{\log(1/\delta)}, \delta)$-DP. Appendix D contains more background on differential privacy.

1.2. Contributions and Related Work

We discuss our contributions in the context of related work. See Figure 1 for a summary of our results when $k = 2$, and Appendix A for a more thorough discussion of related work.

Optimal Rates for Non-Smooth (Strongly) Convex Losses (Section 3): We establish asymptotically optimal (up to logarithms) excess risk bounds for DP SCO under Assumption 1, without requiring differentiability of $f(\cdot, x)$:

Theorem 4 (Informal, see Theorem 6, Theorem 12, Theorem 13, Theorem 14) *Let $f(\cdot, x)$ be convex and $L_f < \infty$. Grant Assumption 1. Then, there is a polynomial-time $\frac{\epsilon^2}{2}$-zCDP algorithm $\mathcal{A}$ such that $\mathbb{E}F(\mathcal{A}(X)) - F^* = \widetilde{\mathcal{O}}\left(\widetilde{r}_{2k} D \left(\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{d}}{\epsilon n}\right)^{(k-1)/k}\right)\right)$. If $f(\cdot, x)$ is μ-strongly convex, then $\mathbb{E}F(\mathcal{A}(X)) - F^* = \widetilde{\mathcal{O}}\left(\frac{\widetilde{r}_{2k}^2}{\mu}\left(\frac{1}{n} + \left(\frac{\sqrt{d}}{\epsilon n}\right)^{(2k-2)/k}\right)\right)$. Further, these bounds are minimax optimal up to factors of $\widetilde{\mathcal{O}}(\widetilde{r}_{2k}/\widetilde{r}_k)$ and $\widetilde{\mathcal{O}}(\widetilde{r}_{2k}^2/\widetilde{r}_k^2)$ respectively.*

As $k \to \infty$, $\widetilde{r}_k, \widetilde{r}_{2k} \to L_f$ and Theorem 4 recovers the known rates for uniformly L_f-Lipschitz DP SCO (Bassily et al., 2019; Feldman et al., 2020a). However, when $k < \infty$ and $\widetilde{r}_{2k} \ll L_f$, the excess risk bounds in Theorem 4 may be much smaller than the uniformly Lipschitz excess risk bounds, which increase with L_f.

The works (Wang et al., 2020; Kamath et al., 2022) make a slightly different assumption than Assumption 1: they instead assume that the k-th order central moment of each coordinate $\nabla_j f(w, x)$ is bounded by $\gamma^{1/k}$ for all $j \in [d], w \in \mathcal{W}$. We also provide asymptotically optimal excess risk bounds for the class of problems satisfying the coordinate-wise moment assumption of (Wang et al., 2020; Kamath et al., 2022) and having *subexponential* stochastic subgradients: see Appendix E.4.

The previous state-of-the-art convex upper bound was suboptimal: $\mathcal{O}\left(\widetilde{r}_k D \sqrt{\frac{d}{n}}\right)$ for $\epsilon \approx 1$ (Kamath et al., 2022, Theorem 5.4).[3] Their result also required $f(\cdot, x)$ to be β_f-smooth for all $x \in \mathcal{X}$ with $\beta_f \leqslant 10$, which can be restrictive with outlier data: e.g. this implies that $f(\cdot, x)$ is uniformly L_f-Lipschitz with $L_f \leqslant \beta_f D \leqslant 10D$ if $\nabla f(w^*(x), x) = 0$ for some $w^*(x) \in \mathcal{W}$. By comparison, our near-optimal bounds hold even for non-differentiable f with $L_f \gg 1$.

Our optimal μ-strongly convex bound also improves over the best previous upper bound of (Kamath et al., 2022, Theorem 5.6), which required uniform β_f-smoothness of $f(\cdot, x)$.

3. We write the bound in (Kamath et al., 2022, Theorem 5.4) in terms of Assumption 1, replacing their $\gamma^{1/k} d$ by $\widetilde{r}\sqrt{d}$.

Function Class	Lower bound	Upper bound	Linear-Time Upper bound	Prior state-of-the-art
Nonsmooth Convex	$\underbrace{\sqrt{d}\left[\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{d}}{\epsilon n}\right)^{1/2}\right]}_{\zeta}$ (Theorem 13) + (Kamath et al., 2022, Thm. 6.4)	$\underbrace{\sqrt{d}\left[\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{d}}{\epsilon n}\right)^{1/2}\right]}_{\zeta}$ (Theorems 6 and 35)	None	None
Smooth Convex			$\zeta\left(1 + \left(\frac{\epsilon n}{d^{3/2}}\right)^{1/18}\right)$ (Thm. 15)	(Kamath et al., 2022, Thm. 5.4) **(not linear time)**: $\min\left\{\frac{d}{\sqrt{\epsilon n}}, \zeta + \frac{d^{5/4}}{\sqrt{n}}\right\}$ if f is uniformly smooth & $\nabla F(w^*) = 0$
Nonsmooth Strongly Convex	$d\left(\frac{1}{n} + \frac{\sqrt{d}}{\epsilon n}\right)$ (Theorem 14) + (Kamath et al., 2022, Theorem 6.1)	$d\left(\frac{1}{n} + \frac{\sqrt{d}}{\epsilon n}\right)$ (Theorems 12 and 35)	None	None
Smooth Strongly Convex			$d\left(\frac{1}{n} + \frac{\sqrt{d\kappa}}{\epsilon n}\right)$ (Theorem 17)	(Kamath et al., 2022, Thm. 5.6) + Appendix C: $d\left(\frac{\kappa_f^3}{n} + \frac{\kappa_f^{3/2}\sqrt{d}}{\epsilon n}\right)$ if f is uniformly smooth & $\nabla F(w^*) = 0$
Proximal PL			$d\kappa\left(\frac{1}{n} + \frac{\sqrt{d}}{\epsilon n}\right)$ (Theorem 19)	None

Figure 1: Excess risk for $k = 2$, $\tilde{r} = \sqrt{d}$; we omit logarithms. $\kappa = \beta/\mu$ is the condition number of F; $\kappa_f = \beta_f/\mu$ is the worst-case condition number of $f(\cdot, x)$.

In fact, (Kamath et al., 2022, Theorem 5.6) was incorrect, as we explain in Appendix C.[4] After communicating with the authors of (Kamath et al., 2022), they updated and corrected the result and proof in the arXiv version of their paper. The arXiv version of (Kamath et al., 2022, Theorem 5.6)—which we derive in Appendix C for completeness—is suboptimal by a factor of $\widetilde{\Omega}((\beta_f/\mu)^3)$. In practice, the worst-case condition number β_f/μ can be very large, especially in the presence of outliers or heavy-tailed data. Our near-optimal excess risk bound removes this dependence on β_f/μ and holds even for non-differentiable f.

Our Algorithm 3 combines the iterative localization technique of (Feldman et al., 2020a; Asi et al., 2021b) with a noisy *clipped* subgradient method. With clipped (hence *biased*) stochastic subgradients and non-Lipschitz/non-smooth $f(\cdot, x)$, the excess risk analysis of our algorithm is harder than in the uniformly Lipschitz setting. Instead of the uniform convergence analysis used in (Wang et al., 2020; Kamath et al., 2022), we derive new results about the *stability* (Kearns and Ron, 1999; Bousquet and Elisseeff, 2002) and generalization error of learning with loss functions that are not uniformly Lipschitz or differentiable; prior results (e.g. (Shalev-Shwartz et al., 2009; Lei and Ying, 2020)) were limited to β_f-smooth and/or L_f-Lipschitz $f(\cdot, x)$. Specifically, we show the following *for non-Lipschitz/non-smooth* $f(\cdot, x)$: a) *On-average model stability* (Lei and Ying, 2020) implies generalization (Proposition 9); and b) regularized empirical risk minimization is on-average model stable, hence it generalizes (Proposition 10). We combine these results with an empirical error bound for *biased*, noisy subgradient method to bound the excess risk of our algorithm (Theorem 6). We obtain our strongly convex bound (Theorem 12) by a reduction to the convex case, ala (Hazan and Kale, 2014; Feldman et al., 2020a).

We also refine (to describe the dependence on $\tilde{r}_k, D, \mu$), extend (to $k \gg 1$), and tighten (for $\mu = 0$) the lower bounds of (Kamath et al., 2022): see Theorems 13 and 14.

Linear-Time Algorithms for Smooth (Strongly) Convex Losses (Section 4): For convex, β-smooth F, we provide a novel accelerated DP algorithm (Algorithm 4), building

4. In short, the mistake is that Jensen's inequality is used in the wrong direction to claim that the T-th iterate of their algorithm w_T satisfies $\mathbb{E}[\|w_T - w^*\|^2] \leqslant (\mathbb{E}\|w_T - w^*\|)^2$, which is false.

on the work of (Ghadimi and Lan, 2012).[5] Our algorithm is *linear time* and attains excess risk that improves over the previous state-of-the-art (*not linear time*) algorithm (Kamath et al., 2022, Theorem 5.4) in practical parameter regimes (e.g. $d \gtrsim n^{1/6}$). The excess risk of our algorithm is tight in certain cases: e.g., $d \gtrsim (\epsilon n)^{2/3}$ or "sufficiently smooth" F (see Remark 16). Our excess risk bound holds even if $L_f = \infty$, which is the case, for instance, for linear regression with unbounded (e.g. Gaussian) data. To prove our bound, we give the first analysis of accelerated SGD with biased stochastic gradients.

For μ-strongly convex, β-smooth losses, acceleration results in excessive bias accumulation, so we propose a simple noisy clipped SGD. Our algorithm builds on (Kamath et al., 2022), but uses a lower-bias clipping mechanism from (Barber and Duchi, 2014) and a new, tighter analysis. We attain excess risk that is near-optimal up to a $\widetilde{\mathcal{O}}((\beta/\mu)^{(k-1)/k})$ factor: see Theorem 17. Our bound strictly improves over the best previous bound of (Kamath et al., 2022). Further, our bound does not require $L_f < \infty$.

First Algorithm for Non-Convex (Proximal-PL) Losses (Section 5): We consider losses satisfying the *Proximal Polyak-Łojasiewicz (PPL) inequality* (Polyak, 1963; Karimi et al., 2016) (Definition 18), an extension of the classical PL inequality to the proximal setting. This covers important models like (some) neural nets, linear/logistic regression, and LASSO (Karimi et al., 2016; Lei and Ying, 2021). We propose a DP proximal clipped SGD to attain near-optimal excess risk that almost matches the *strongly convex* rate: see Theorem 19.

We also provide (in Appendix I) the **first *shuffle differentially private (SDP)*** (Bittau et al., 2017; Cheu et al., 2019) **algorithms for SO with large worst-case Lipschitz parameter**. Our SDP algorithms achieve the same risk bounds as their zCDP counterparts *without requiring a trusted curator.*

2. Private Heavy-Tailed Mean Estimation Building Blocks

In each iteration of our SO algorithms, we need a way to privately estimate the mean $\nabla F(w_t) = \mathbb{E}_{x\sim\mathcal{D}}[\nabla f(w_t, x)]$. If $f(\cdot, x)$ is uniformly Lipschitz, then one can simply draw a random sample x^t from X and add noise to the stochastic gradient $\nabla f(w_t, x^t)$ to obtain a DP estimator of $\nabla F(w_t)$: the ℓ_2-sensitivity of the stochastic gradients is bounded by $\sup_{x,x'\in\mathcal{X}} \|\nabla f(w_t, x) - \nabla f(w_t, x')\| \leqslant 2L_f$, so the Gaussian mechanism guarantees DP (by Proposition 22). However, in the setting that we consider, L_f (and hence the sensitivity) may be huge, leading the privacy noise variance to also be huge. Thus, we *clip* the stochastic gradients (to force the sensitivity to be bounded) before adding noise. Specifically, we invoke Algorithm 1 on a batch of s stochastic gradients at each iteration of our algorithms. In Algorithm 1, $\Pi_C(z) := \operatorname{argmin}_{y\in B_2(0,C)} \|y - z\|^2$ denotes the projection onto the centered ℓ_2 ball of radius C in $\mathbb{R}^d$. Lemma 5 bounds the bias and variance of Algorithm 1.

Lemma 5 (Barber and Duchi (2014)) *Let $\{z_i\}_{i=1}^s \sim \mathcal{D}^s$ be $\mathbb{R}^d$-valued random vectors with $\mathbb{E}z_i = \nu$ and $\mathbb{E}\|z_i\|^k \leqslant r^{(k)}$ for some $k \geqslant 2$. Denote the noiseless average of clipped samples by $\widehat{\nu} := \frac{1}{s}\sum_{i=1}^s \Pi_C(z_i)$ and $\widetilde{\nu} := \widehat{\nu} + N$. Then, $\|\mathbb{E}\widetilde{\nu} - \nu\| = \|\mathbb{E}\widehat{\nu} - \nu\| \leqslant \mathbb{E}\|\widehat{\nu} - \nu\| \leqslant \frac{r^{(k)}}{(k-1)C^{k-1}}$, and $\mathbb{E}\|\widetilde{\nu} - \mathbb{E}\widetilde{\nu}\|^2 = \mathbb{E}\|\widetilde{\nu} - \mathbb{E}\widehat{\nu}\|^2 \leqslant d\sigma^2 + \frac{r^{(2)}}{s}$.*

5. In contrast to (Wang et al., 2020; Kamath et al., 2022), *we do not require* $f(\cdot, x)$ *to be* β_f*-smooth* for all x.

Algorithm 1 MeanOracle1($\{z_i\}_{i=1}^s; s; C; \frac{\epsilon^2}{2}$) (Barber and Duchi, 2014)

1: **Input:** $Z = \{z_i\}_{i=1}^s$, $C > 0$, $\epsilon > 0$. Set $\sigma^2 = \frac{4C^2}{s^2\epsilon^2}$ for $\frac{\epsilon^2}{2}$-zCDP.
2: Draw $N \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ and compute $\widetilde{\nu} := \frac{1}{s}\sum_{i=1}^s \Pi_C(z_i) + N$.
3: **Output:** $\widetilde{\nu}$.

3. Optimal Rates for Non-Smooth (Strongly) Convex Losses

In this section, we establish the optimal rates (up to logarithms) for the class of DP SCO problems satisfying Assumption 1. We present our result for convex losses in Section 3.1, and our result for strongly convex losses in Section 3.2. In Section 3.3, we provide lower bounds, which show that our upper bounds are tight (up to logarithms).

3.1. Localized Noisy Clipped Subgradient Method for Convex Losses

Our algorithm (Algorithm 3) uses iterative localization (Feldman et al., 2020a; Asi et al., 2021b) with clipping (in Algorithm 2) to handle stochastic subgradients with large norm.[6]

Algorithm 2 Noisy Clipped Regularized Subgradient Method for DP ERM

1: **Input:** Data $X \in \mathcal{X}^n$, $T \in \mathbb{N}$, stepsize η, clip thresh. C, regularization $\lambda \geqslant 0$, constraint set $\mathcal{W}$ and initialization $w_0 \in \mathcal{W}$.
2: **for** $t \in \{0, 1, \cdots, T-1\}$ **do**
3: $\quad\widetilde{\nabla} F_t(w_t) :=$ MeanOracle1($\{\nabla f(w_t, x_i)\}_{i=1}^n; n; C; \frac{\epsilon^2}{2T}$) for subgradients $\nabla f(w_t, x_i)$.
4: $\quad w_{t+1} = \Pi_{\mathcal{W}}\left[w_t - \eta\left(\widetilde{\nabla} F_t(w_t) + \lambda(w_t - w_0)\right)\right]$
5: **end for**
6: **Output:** w_T.

Algorithm 3 Localized Noisy Clipped Subgradient Method for DP SCO

1: **Input:** Data $X \in \mathcal{X}^n$, stepsize η, clip thresh. $\{C_i\}_{i=1}^{\log_2(n)}$, iteration num. $\{T_i\}_{i=1}^{\log_2(n)}$, hyperparameter $p \geqslant 1$.
2: Initialize $w_0 \in \mathcal{W}$. Let $l := \log_2(n)$.
3: **for** $i \in [l]$ **do**
4: $\quad$Set $n_i = 2^{-i} n, \eta_i = 4^{-i}\eta$, $\lambda_i = \frac{1}{\eta_i n_i^p}$, $T_i = \widetilde{\Theta}\left(\frac{1}{\lambda_i \eta_i}\right)$, and $D_i = \frac{2L_f}{\lambda_i}$.
5: $\quad$Draw new batch $\mathcal{B}_i$ of $n_i = |\mathcal{B}_i|$ samples from X without replacement.
6: $\quad$Let $\widehat{F}_i(w) := \frac{1}{n_i}\sum_{j \in \mathcal{B}_i} f(w, x_j) + \frac{\lambda_i}{2}\|w - w_{i-1}\|^2$.
7: $\quad$Use Algorithm 2 with initialization w_{i-1} to minimize $\widehat{F}_i$ over $\mathcal{W}_i := \{w \in \mathcal{W} : \|w - w_{i-1}\| \leqslant D_i\}$, for T_i iterations with clip threshold C_i and noise $\sigma_i^2 = \frac{4C_i^2 T_i}{n_i^2 \epsilon^2}$. Let w_i be the output of Algorithm 2.
8: **end for**
9: **Output:** w_l.

6. We assume WLOG that $n = 2^l$ for some $l \in \mathbb{N}$. If this is not the case, then throw out samples until it is; since the number of remaining samples is at least $n/2$, our bounds still hold up to a constant factor.

The main ideas of Algorithm 3 are:

1. *Clipping only the non-regularized component of the subgradient to control sensitivity and bias:* Notice that when we call Algorithm 2 in phase i of Algorithm 3, we only clip the subgradients of $f(w, x_j)$ and not the regularized loss $f(w, x_j) + \frac{\lambda}{2}\|w - w_{i-1}\|^2$. Compared to clipping the full gradient of the regularized loss, our selective clipping approach significantly reduces the bias of our subgradient estimator. This is essential for obtaining our near-optimal excess risk. Further, this reduction in bias comes at no cost to the variance of our subgradient estimator: the ℓ_2-sensitivity of our estimator is unaffected by the regularization term.

2. *Solve regularized ERM subproblem with a stable DP algorithm*: We run a *multi-pass* zCDP solver on a *regularized* empirical loss: Multiple passes let us reduce the noise variance in phase i by a factor of T_i (via strong composition for zCDP) and get a more accurate solution to the ERM subproblem. Regularization makes the empirical loss strongly convex, which improves *on-average model stability* and hence generalization of the obtained solution (see Proposition 9 and 29).

3. *Localization* (Feldman et al., 2020a; Asi et al., 2021a) (i.e. iteratively "zooming in" on a solution): In early phases (small i), when we are far away from the optimum w^*, we use more samples (larger n_i) and large learning rate η_i to make progress quickly. As i increases, w_i is closer to w^*, so fewer samples and slower learning rate suffice. Since step size η_i shrinks (geometrically) faster than n_i, the effective variance of the privacy noise $\eta_i^2\sigma_i^2$ decreases as i increases. This prevents w_{i+1} from moving too far away from w_i (and hence from w^*). We further enforce this localization behavior by increasing the regularization parameter λ_i and shrinking D_i over time. We choose D_i as small as possible subject to the constraint that $\text{argmin}_{w \in \mathcal{W}} \widehat{F}_i(w) \in \mathcal{W}_i$. This constraint ensures that Algorithm 2 can find w_i with small excess risk.

Next, we provide privacy and excess risk guarantees for Algorithm 3:

Theorem 6 *Grant Assumption 1. Let $\epsilon \leqslant \sqrt{d}$ and let $f(\cdot, x)$ be convex and L_f-Lipschitz for all x for some $L_f < \infty$. Then, there are algorithmic parameters such that Algorithm 3 is $\frac{\epsilon^2}{2}$-zCDP, and has excess risk*

$$\mathbb{E}F(w_l) - F^* \lesssim \widetilde{r}_{2k} D \left(\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{d \ln(n)}}{\epsilon n} \right)^{\frac{k-1}{k}} \right).$$

Moreover, this excess risk is attained in $\widetilde{\mathcal{O}}(n^2 L_f^2 / \widetilde{r}_{2k}^2)$ subgradient evaluations.

The excess risk bound in Theorem 6 is optimal up to a logarithmic factor.[7] A key feature of this bound is that it does not depend on L_f. The only cost of larger L_f is higher subgradient computational complexity. Further, Theorem 6 holds for any $L_f < \infty$ and even non-differentiable/non-smooth f. By contrast, prior works (Wang et al., 2020; Kamath et al.,

7. The lower bound construction in Theorem 13 is L_f-Lipschitz for $L_f \leqslant n^2 \widetilde{r}_k < \infty$.

2022) required uniform β_f-smoothness of $f(\cdot, x)$, which implies the restriction $L_f \leqslant \beta_f D$ for loss functions that have a vanishing gradient at some point.[8] In Section 4, we develop algorithms that can even handle the case $L_f = \infty$ and still give excess risk bounds that are tight in certain practical parameter regimes. The algorithms in Section 4 have runtime $O(nd)$ (independent of L_f), which may make them more practical than Algorithm 3.

The proof of Theorem 6 consists of three main steps: i) We bound the empirical error of the noisy clipped subgradient subroutine (Lemma 7). ii) We prove that if an algorithm is *on-average model stable* (Definition 8), then it generalizes (Proposition 9). iii) We bound the on-average model stability of regularized ERM with non-smooth/non-Lipschitz $f(\cdot, x)$ (Proposition 29), leading to an excess population loss bound for Algorithm 2 run on the regularized empirical objective (c.f. line 7 of Algorithm 3). By using these results with the proof technique of (Feldman et al., 2020a), we can obtain Theorem 6.

First, we bound the empirical error of the step in line 7 of Algorithm 3, by extending the analysis of noisy subgradient method to *biased* subgradient oracles:

Lemma 7 *Fix $X \in \mathcal{X}^n$ and let $\widehat{F}_\lambda(w) = \frac{1}{n}\sum_{i=1}^n f(w, x_i) + \frac{\lambda}{2}\|w - w_0\|^2$ for $w_0 \in \mathcal{W}$, where $\mathcal{W}$ is a closed convex domain with diameter D. Assume $f(\cdot, x)$ is convex and $\widehat{r}_n(X)^{(k)} \geqslant \sup_{w \in \mathcal{W}} \left\{\frac{1}{n}\sum_{i=1}^n \|\nabla f(w, x_i)\|^k\right\}$ for all $\nabla f(w, x_i) \in \partial_w f(w, x_i)$. Denote $\widehat{r}_n(X) = \left[\widehat{r}_n(X)^{(k)}\right]^{1/k}$ and $\hat{w} = \operatorname{argmin}_{w \in \mathcal{W}} \widehat{F}_\lambda(w)$. Let $\eta \leqslant \frac{2}{\lambda}$. Then, the output of Algorithm 2 satisfies*

$$\mathbb{E}\|w_T - \hat{w}\|^2 \leqslant \exp\left(-\frac{\lambda\eta T}{2}\right)\|w_0 - \hat{w}\|^2 + \frac{8\eta}{\lambda}\left(\widehat{r}_n(X)^2 + \lambda^2 D^2 + d\sigma^2\right) + \frac{20}{\lambda^2}\left(\frac{\widehat{r}_n(X)^{(k)}}{(k-1)C^{k-1}}\right)^2,$$

where $\sigma^2 = \frac{4C^2 T}{n^2\epsilon^2}$.

Detailed proofs for this subsection are deferred to Appendix E.2.

Our next goal is to bound the generalization error of regularized ERM with convex loss functions that are not differentiable or uniformly Lipschitz. We will use a stability argument to obtain such a bound. Recall the notion of *on-average model stability* (Lei and Ying, 2020):

Definition 8 *Let $X = (x_1, \cdots, x_n)$ and $X' = (x_1', \cdots, x_n')$ be drawn independently from $\mathcal{D}$. For $i \in [n]$, let $X^i := (x_1, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots, x_n)$. We say randomized algorithm $\mathcal{A}$ has on-average model stability α (i.e. $\mathcal{A}$ is α-on-average model stable) if $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|\mathcal{A}(X) - \mathcal{A}(X^i)\|^2\right] \leqslant \alpha^2$. The expectation is over the randomness of $\mathcal{A}$ and the draws of X and X'.*

On-average model stability is weaker than the notion of *uniform stability* (Bousquet and Elisseeff, 2002), which has been used in DP Lipschitz SCO (e.g. by (Bassily et al., 2019)); this is necessary for obtaining our learnability guarantees without uniform Lipschitz continuity.

The main result in (Lei and Ying, 2020) showed that on-average model stable algorithms generalize well if $f(\cdot, x)$ is β_f-*smooth* for all x, which leads to a restriction on L_f. We show that neither smoothness nor Lipschitz continuity of f is needed to ensure generalizability:

8. Additionally, (Kamath et al., 2022) assumes $\beta_f \leqslant 10$.

Proposition 9 *Let $f(\cdot, x)$ be convex for all x. Suppose $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$ is α-on-average model stable. Let $\widehat{F}_X(w) := \frac{1}{n}\sum_{i=1}^n f(w, x_i)$ be an empirical loss. Then for any $\zeta > 0$,*

$$\mathbb{E}[F(\mathcal{A}(X)) - \widehat{F}_X(\mathcal{A}(X))] \leqslant \frac{\widetilde{r}^{(2)}}{2\zeta} + \frac{\zeta}{2}\alpha^2.$$

Using Proposition 9, we can bound the generalization error and excess (population) risk of regularized ERM:

Proposition 10 *Let $f(\cdot, x)$ be convex, $w_{i-1}, y \in \mathcal{W}$, and $\hat{w}_i := \text{argmin}_{w \in \mathcal{W}} \widehat{F}_i(w)$, where $\widehat{F}_i(w) := \frac{1}{n_i}\sum_{j \in \mathcal{B}_i} f(w, x_j) + \frac{\lambda_i}{2}\|w - w_{i-1}\|^2$ (c.f. line 6 of Algorithm 3). Then,*

$$\mathbb{E}[F(\hat{w}_i)] - F(y) \leqslant \frac{2\widetilde{r}^{(2)}}{\lambda_i n_i} + \frac{\lambda_i}{2}\|y - w_{i-1}\|^2,$$

where the expectation is over both the random draws of X from $\mathcal{D}$ and $\mathcal{B}_i$ from X.

With the pieces developed above, we can now sketch the proof of Theorem 6:

Proof [Sketch of the Proof of Theorem 6] **Privacy:** Since the batches $\{\mathcal{B}_i\}_{i=1}^l$ are disjoint, it suffices to show that w_i (produced by T_i iterations of Algorithm 2 in line 7 of Algorithm 3) is $\frac{\epsilon^2}{2}$-zCDP $\forall i \in [l]$. The ℓ_2 sensitivity of the clipped subgradient update is $\Delta = \sup_{w, X \sim X'} \|\frac{1}{n_i}\sum_{j=1}^{n_i} \Pi_{C_i}(\nabla f(w, x_j)) - \Pi_{C_i}(\nabla f(w, x_j'))\| \leqslant 2C_i/n_i$. (Note that the regularization term does not contribute to sensitivity.) Thus, the privacy guarantees of the Gaussian mechanism (Proposition 22) and the composition theorem for zCDP (Lemma 23) imply that Algorithm 3 is $\frac{\epsilon^2}{2}$-zCDP.

Excess risk: First, our choice of D_i ensures that $\hat{w}_i \in \mathcal{W}_i$. Second, since $L_f < \infty$, there exists $p \geqslant 1$ such that $L_f = \widetilde{\mathcal{O}}\left(n^{p/2}\widetilde{r}_{2k}\left(\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{d\ln(n)}}{\epsilon n}\right)^{\frac{k-1}{k}}\right)\right)$. Combining Lemma 7 with Lemma 5 and proper choices of η and T_i, we get:

$$\mathbb{E}\|w_i - \hat{w}_i\|^2 \lesssim \eta_i^2 n_i^p (L_f^2 + d\sigma_i^2) + \frac{\eta_i^2 n_i^{2p}\widetilde{r}^{(2k)}}{C_i^{2k-2}} \lesssim \left(\frac{\eta^2 n^p}{32^i}\left(L_f^2 + \frac{dC_i^2 T_i}{\epsilon^2 n_i^2} + \frac{n^p \widetilde{r}^{(2k)}}{C_i^{2k-2} 2^{pi}}\right)\right). \quad (2)$$

Now, following the strategy used in the proof of (Feldman et al., 2020a, Theorem 4.4), we write $\mathbb{E}F(w_l) - F(w^*) = \mathbb{E}[F(w_l) - F(\hat{w}_l)] + \sum_{i=1}^l \mathbb{E}[F(\hat{w}_i) - F(\hat{w}_{i-1})]$, where $\hat{w}_0 := w^*$. Using (2) and $\widetilde{r}_k$-Lipschitz continuity of F (which is implied by Assumption 1), we can bound $\mathbb{E}[F(w_l) - F(\hat{w}_l)]$ for the right η and C_l. To bound the sum (second term), we use Proposition 10 to obtain

$$\sum_{i=1}^l \mathbb{E}[F(\hat{w}_i) - F(\hat{w}_{i-1})] \lesssim r^2 \eta n^{p-1} + \frac{D^2}{\eta n^p} + \eta L_f^2 + \eta \sum_{i=2}^l 4^{-i} n_i^p \widetilde{r}_{2k}^2 \left(\frac{d\ln(n)}{\epsilon^2 n_i^2}\right)^{\frac{k-1}{k}}$$
$$\lesssim \eta\left[r^2 n^{p-1} + L_f^2 + \widetilde{r}_{2k}^2 n^p \left(\frac{d\ln(n)}{\epsilon^2 n^2}\right)^{\frac{k-1}{k}}\right] + \frac{D^2}{\eta n^p},$$

for the right choice of C_i. Then properly choosing η completes the excess risk proof.

Computational complexity: The choice $T_i = \widetilde{\mathcal{O}}(n_i^p)$ implies that the number of subgradient evaluations is bounded by $\sum_{i=1}^{l} n_i T_i = \widetilde{\mathcal{O}}(n^{p+1})$. Further, $n^p \leqslant n \left(\frac{L_f}{\widetilde{r}_{2k}}\right)^2$ ■

Remark 11 (Improved Computational Complexity for Approximate or Shuffle DP) *If one desires (ϵ, δ)-DP or (ϵ, δ)-SDP instead of zCDP, then the computational complexity of Algorithm 3 can be improved: see Appendix E.2.*

3.2. The Strongly Convex Case

Following (Feldman et al., 2020a), we use a folklore reduction to the convex case (detailed in Appendix E.3) in order to obtain the following upper bound via Theorem 6:

Theorem 12 *Grant Assumption 1. Let $\epsilon \leqslant \sqrt{d}$ and let $f(\cdot, x)$ be μ-strongly convex and L_f-Lipschitz for all $x \in \mathcal{X}$ for some $L_f < \infty$. Then, there is a polynomial-time $\frac{\epsilon^2}{2}$-zCDP algorithm $\mathcal{A}$ based on Algorithm 3 with excess risk*

$$\mathbb{E}F(\mathcal{A}(X)) - F^* \lesssim \frac{\widetilde{r}_{2k}^2}{\mu}\left(\frac{1}{n} + \left(\frac{\sqrt{d\ln(n)}}{\epsilon n}\right)^{\frac{2k-2}{k}}\right).$$

3.3. Lower Bounds

The work of (Kamath et al., 2022) proved lower bounds that are tight (by our upper bounds in Section 3) in most parameter regimes for $D = \mu = 1$, $\widetilde{r}_k = \sqrt{d}$, and $k = \mathcal{O}(1)$.[9] Our (relatively modest) contribution in this subsection is: refining these lower bounds to display the correct dependence on $\widetilde{r}_k, D, \mu$; tightening the convex lower bound (Kamath et al., 2022, Theorem 6.4) in the regime $d > n$; and extending (Kamath et al., 2022, Theorems 6.1 and 6.4) to $k \gg 1$. Our lower bound constructions satisfy the condition on L_f in the statements of Theorems 6 and 12. Our first lower bound holds even for affine functions:

Theorem 13 (Smooth Convex, Informal) *Let $\rho \leqslant d$. For any ρ-zCDP algorithm $\mathcal{A}$, there exist closed convex sets $\mathcal{W}, \mathcal{X} \subset \mathbb{R}^d$ such that $\|w - w'\| \leqslant 2D$ for all $w, w' \in \mathcal{W}$, a β_f-smooth, L_f-Lipschitz, linear, convex (in w) loss $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, and distributions $\mathcal{D}$ and $\mathcal{D}'$ on $\mathcal{X}$ such that Assumption 1 holds and if $X \sim \mathcal{D}^n$, then*

$$\mathbb{E}F(\mathcal{A}(X)) - F^* = \Omega\left(\widetilde{r}_k D\left(\frac{1}{\sqrt{n}} + \min\left\{1, \left(\frac{\sqrt{d}}{\sqrt{\rho}n}\right)^{\frac{k-1}{k}}\right\}\right)\right).$$

Remark 37 (in Appendix F) discusses parameter regimes in which Theorem 13 is strictly tighter than (Kamath et al., 2022, Theorem 6.4), as well as differences in our proof vs. theirs.

Next, we provide lower bounds for smooth, strongly convex loss functions:

Theorem 14 (Smooth Strongly Convex, Informal) *Let $\rho \leqslant d$. For any ρ-zCDP algorithm $\mathcal{A}$, there exist compact convex sets $\mathcal{W}, \mathcal{X} \subset \mathbb{R}^d$, a L_f-Lipschitz, μ-smooth, μ-strongly*

9. The lower bounds asserted in (Kamath et al., 2022) only hold if $k \lesssim 1$ since the moments of the Gaussian distribution that they construct grow exponentially/factorially with k.

convex (in w) loss $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, and distributions $\mathcal{D}$ and $\mathcal{D}'$ on $\mathcal{X}$ such that: Assumption 1 holds, and if $X \sim \mathcal{D}^n$, then

$$\mathbb{E}F(\mathcal{A}(X)) - F^* = \Omega\left(\frac{\tilde{r}_k^2}{\mu}\left(\frac{1}{n} + \min\left\{1, \left(\frac{\sqrt{d}}{\sqrt{\rho}n}\right)^{\frac{2k-2}{k}}\right\}\right)\right).$$

Thus, our upper bounds are indeed tight (up to logarithms). Having resolved *Question I*, next we will develop more computationally efficient, *linear-time* algorithms for smooth $F(\cdot)$.

4. Linear-Time Algorithms for Smooth (Strongly) Convex Losses

4.1. Noisy Clipped Accelerated SGD for Smooth Convex Losses

Algorithm 4 is a one-pass accelerated algorithm, which builds on (non-private) AC-SA of (Ghadimi and Lan, 2012); its privacy and excess risk guarantees are given in Theorem 15.

Algorithm 4 Noisy Clipped Accelerated SGD (AC-SA) for Heavy-Tailed DP SCO

Input: Data $X \in \mathcal{X}^n$, iteration number $T \leqslant n$, stepsize parameters $\{\eta_t\}_{t\in[T]}, \{\alpha_t\}_{t\in[T]}$ with $\alpha_1 = 1, \alpha_t \in (0,1)\ \forall t \geqslant 2$.
Initialize $w_0^{ag} = w_0 \in \mathcal{W}$ and $t = 1$.
for $t \in [T]$ **do**
  $w_t^{md} := (1-\alpha_t)w_{t-1}^{ag} + \alpha_t w_{t-1}$.
  Draw new batch $\mathcal{B}_t$ (without replacement) of n/T samples from X.
  $\tilde{\nabla}F_t(w_t^{md}) :=$ `MeanOracle1`$(\{\nabla f(w_t^{md}, x)\}_{x\in\mathcal{B}_t}; \frac{n}{T}; \frac{\epsilon^2}{2})$
  $w_t := \mathrm{argmin}_{w\in\mathcal{W}}\left\{\alpha_t\langle\tilde{\nabla}F_t(w_t^{md}), w\rangle + \frac{\eta_t}{2}\|w_{t-1} - w\|^2\right\}$.
  $w_t^{ag} := \alpha_t w_t + (1-\alpha_t)w_{t-1}^{ag}$.
end for
Output: w_T^{ag}.

Theorem 15 (Informal) *Grant Assumption 1. Let F be convex and β-smooth. Then, there are parameters such that Algorithm 4 is $\frac{\epsilon^2}{2}$-zCDP and:*

$$\mathbb{E}F(w_T^{ag}) - F^* \lesssim \tilde{r}_k D\left[\frac{1}{\sqrt{n}} + \max\left\{\left(\left(\frac{\beta D}{\tilde{r}_k}\right)^{1/4}\frac{\sqrt{d}}{\epsilon n}\right)^{\frac{4(k-1)}{5k-1}}, \left(\frac{\sqrt{d}}{\epsilon n}\right)^{\frac{k-1}{k}}\right\}\right]. \tag{3}$$

Moreover, Algorithm 4 uses at most n gradient evaluations.

Besides being linear-time, another advantage of Theorem 15 is that it holds even for problems with $L_f = \infty$. The key ingredient used to prove (3) is a novel convergence guarantee for AC-SA with *biased*, noisy stochastic gradients: see Proposition 40 in Appendix G.1. Combining Proposition 40 with Lemma 5 and a careful choice of stepsizes, clip threshold, and T yields Theorem 15.

Remark 16 (Optimal rate for "sufficiently smooth" functions) *Note that the upper bound (3) scales with the smoothness parameter β. Thus, for sufficiently small β, the optimal rates are attained. For example, if $k = 2$, the upper bound in (3) matches the lower bound in Theorem 13 when $\beta \lesssim \frac{\tilde{r}_k}{D}\left(\frac{d^5}{\epsilon n}\right)^{1/18}$; e.g. if β and D are constants and $d \geqslant (\epsilon n)^{1/5}$. In particular, for affine functions (which were not addressed in (Wang et al., 2020; Kamath et al., 2022) since these works assume $\nabla F(w^*) = 0$), $\beta = 0$ and Algorithm 4 is optimal.*

Having discussed the dependence on β, let us focus on understanding how the bound in Theorem 15 scales with n, d and ϵ. Thus, let us fix $\beta = D = 1$ and $\tilde{r}_k = \sqrt{d}$ for simplicity. If $k = 2$, then the bound in (3) simplifies to $\mathcal{O}\left(\sqrt{\frac{d}{n}} + \max\left\{\frac{d^{2/3}}{(\epsilon n)^{4/9}}, \frac{d^{3/4}}{\sqrt{\epsilon n}}\right\}\right)$, whereas the lower bound in Theorem 13 (part 2) is $\Omega\left(\sqrt{\frac{d}{n}} + \frac{d^{3/4}}{\sqrt{\epsilon n}}\right)$. Therefore, the bound in (3) is tight if $d^{3/2} \gtrsim \epsilon n$. For general n, d, ϵ, (3) is *nearly* tight up to a multiplicative factor of $\left(\frac{\epsilon n}{d^{3/2}}\right)^{1/18}$. By comparison, the previous state-of-the-art (*not linear-time*) bound for $\epsilon \approx 1$ was $\mathcal{O}\left(\frac{d}{\sqrt{n}}\right)$ (Kamath et al., 2022, Theorem 5.4). Our bound (3) improves over (Kamath et al., 2022, Theorem 5.4) if $d \gtrsim n^{1/6}$, which is typical in practical ML applications. As $k \to \infty$, (3) becomes $\mathcal{O}\left(\sqrt{\frac{d}{n}} + \left(\frac{d}{n}\right)^{4/5}\right)$ for $\epsilon \approx 1$, which is strictly better than the bound in (Kamath et al., 2022, Theorem 5.4).

4.2. Noisy Clipped SGD for Strongly Convex Losses

Our algorithm for strongly convex losses (Algorithm 6 in Appendix G.2) is a simple one-pass noisy clipped SGD. Compared to the algorithm of (Kamath et al., 2022), our approach differs in the choice of `MeanOracle`, step size, and iterate averaging weights, and in our *analysis*.

Theorem 17 (Informal) *Grant Assumption 1. Let F be μ-strongly convex, β-smooth with $\frac{\beta}{\mu} \leqslant n/\ln(n)$. Then, there are algorithmic parameters such that Algorithm 6 is $\frac{\epsilon^2}{2}$-zCDP and:*

$$\mathbb{E}F(\widehat{w}_T) - F^* \lesssim \frac{\tilde{r}_k^2}{\mu}\left(\frac{1}{n} + \left(\frac{\sqrt{d(\beta/\mu)\ln(n)}}{\epsilon n}\right)^{\frac{2k-2}{k}}\right). \quad (4)$$

Moreover, Algorithm 6 uses at most n gradient evaluations.

The bound (4) is optimal up to a $\widetilde{\mathcal{O}}((\beta/\mu)^{(k-1)/k})$ factor and improves over the best previous bound in (Kamath et al., 2022, Theorem 5.6) by removing the dependence on β_f (which can be much larger than β in the presence of outliers). The proof of Theorem 17 (in Appendix G.2) relies on a novel convergence guarantee for projected SGD with biased noisy stochastic gradients: Proposition 42. Compared to results in (Asi et al., 2021a) for convex ERM and (Ajalloeian and Stich, 2020) for PL SO, Proposition 42 is tighter, which is needed to obtain near-optimal excess risk: we leverage smoothness and strong convexity. Our new analysis also avoids the issue in the proofs of (the ICML versions of) (Wang et al., 2020; Kamath et al., 2022).

5. Algorithm for Non-Convex Proximal-PL Loss Functions

Assume: $f(w,x) = f^0(w,x) + f^1(w)$; $f^0(\cdot,x)$ is differentiable (maybe non-convex); f^1 is proper, closed, and convex (maybe non-differentiable) for all $x \in \mathcal{X}$; and $F(w) = F^0(w) + f^1(w) = \mathbb{E}_{x\sim\mathcal{D}}[f^0(w,x)] + f^1(w)$ satisfies the *Proximal-PL* condition (Karimi et al., 2016):

Definition 18 (μ-PPL) *Let $F(w) = F^0(w) + f^1(w)$ be bounded below; F^0 is β-smooth and f^1 is convex. F satisfies Proximal Polyak-Łojasiewicz inequality with parameter $\mu > 0$ if*

$$\mu[F(w) - \inf_{w'} F(w')] \leqslant -\beta \min_y \left[\langle \nabla F^0(w), y - w\rangle + \frac{\beta}{2}\|y-w\|^2 + f^1(y) - f^1(w)\right], \ \forall\ w \in \mathbb{R}^d.$$

Definition 18 generalizes the classical PL condition ($f^1 = 0$), allowing for constrained optimization or non-smooth regularizer depending on f^1 (Polyak, 1963; Karimi et al., 2016).

Recall that the *proximal operator* of a convex function g is defined as $\texttt{prox}_{\eta g}(z) := \text{argmin}_{y\in\mathbb{R}^d}\left(\eta g(y) + \frac{1}{2}\|y-z\|^2\right)$ for $\eta > 0$. We propose *Noisy Clipped Proximal SGD* (Algorithm 8 in Appendix H) for PPL losses. The algorithm runs as follows. For $t \in [T]$: first draw a new batch $\mathcal{B}_t$ (without replacement) of n/T samples from X; let $\widetilde{\nabla} F_t^0(w_t) := \texttt{MeanOracle1}(\{\nabla f^0(w_t,x)\}_{x\in\mathcal{B}_t}; \frac{n}{T}; \frac{\epsilon^2}{2})$; then update $w_{t+1} = \texttt{prox}_{\eta_t f^1}\left(w_t - \eta_t \widetilde{\nabla} F_t^0(w_t)\right)$. Finally, return the last iterate, w_T. Thus, the algorithm is linear time. Furthermore:

Theorem 19 (Informal) *Grant Assumption 1. Let F be μ-PPL for β-smooth F^0, with $\frac{\beta}{\mu} \leqslant n/\ln(n)$. Then, there are parameters such that Algorithm 8 is $\frac{\epsilon^2}{2}$-zCDP, and:*

$$\mathbb{E}F(w_T) - F^* \lesssim \frac{\widetilde{r}_k^2}{\mu}\left(\left(\frac{\sqrt{d}}{\epsilon n}(\beta/\mu)\ln(n)\right)^{\frac{2k-2}{k}} + \frac{(\beta/\mu)\ln(n)}{n}\right).$$

Moreover, Algorithm 8 uses at most n gradient evaluations.

The bound in Theorem 19 nearly matches the smooth *strongly convex* (hence PPL) lower bound in Theorem 14 up to $\widetilde{\mathcal{O}}((\beta/\mu)^{(2k-2)/2})$, and is attained without convexity.

To prove Theorem 19, we derive a convergence guarantee for proximal SGD with generic biased, noisy stochastic gradients in terms of the bias and variance of the oracle (see Proposition 45). We then apply this guarantee for `MeanOracle1` (Algorithm 1) with carefully chosen stepsizes, clip threshold, and T, using Lemma 5. Proposition 45 generalizes (Ajalloeian and Stich, 2020, Theorem 6)–which covered the unconstrained classical PL problem–to the proximal setting. However, the proof of Proposition 45 is very different from the proof of (Ajalloeian and Stich, 2020, Theorem 6), since `prox` makes it hard to bound excess risk without convexity when the stochastic gradients are biased/noisy. Instead, our proof builds on the proof of (Lowy et al., 2022, Theorem 3.1), using techniques from the analysis of *objective perturbation* (Chaudhuri et al., 2011; Kifer et al., 2012). See Appendix H for details.

6. Concluding Remarks and Open Questions

This paper was motivated by practical problems in which data contains outliers and potentially heavy tails, causing the worst-case Lipschitz parameter of the loss over all data points to be prohibitively large. In such cases, existing bounds for DP SO that scale with the worst-case Lipschitz parameter become vacuous. Thus, we operated under the more relaxed assumption of stochastic gradient distributions having bounded k-th moments. The k-th moment bound can be much smaller than the worst-case Lipschitz parameter in practice. For (strongly) convex loss functions, we established the asymptotically optimal rates (up to logarithms), even with non-differentiable losses. We also provided linear-time algorithms for smooth losses that are optimal in certain practical parameter regimes, but suboptimal in general. An interesting open question is: does there exist a linear-time algorithm with optimal excess risk? We also initiated the study of non-convex DP SO without uniform Lipschitz continuity, showing that the optimal strongly convex rates can nearly be attained without convexity, via the proximal-PL condition. We leave the treatment of general non-convex losses for future work.

Acknowledgements

We would like to thank John Duchi, Larry Goldstein, and Stas Minsker for very helpful conversations and pointers related to our lower bounds and the proof of Lemma 56. We also thank the authors of (Kamath et al., 2022) for clarifying some steps in the proof of their Theorem 4.1 and providing useful feedback on the first draft of this manuscript. We would also like to thank the USC-Meta Center for Research and Education in AI and Learning for supporting this research. Finally, we thank an anonymous ALT reviewer for identifying an issue in a previous version of this paper.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Oct 2016. doi: 10.1145/2976749.2978318. URL http://dx.doi.org/10.1145/2976749.2978318.

Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pages 48–78. PMLR, 2021.

Ahmad Ajalloeian and Sebastian U Stich. On the convergence of sgd with biased gradients. *arXiv preprint arXiv:2008.00051*, 2020.

Alex Andoni. COMS E6998-9: algorithmic techniques for massive data, 2015. URL https://www.mit.edu/~andoni/F15_AlgoTechMassiveData/files/scribe2.pdf.

Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. *Advances in Neural Information Processing Systems*, 34, 2021.

Raman Arora, Raef Bassily, Cristóbal Guzmán, Michael Menart, and Enayat Ullah. Differentially private generalized linear models revisited. *arXiv preprint arXiv:2205.03014*, 2022.

Hilal Asi, John Duchi, Alireza Fallah, Omid Javidbakht, and Kunal Talwar. Private adaptive gradient methods for convex optimization. In *International Conference on Machine Learning*, pages 383–392. PMLR, 2021a.

Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in ℓ_1 geometry. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 393–403. PMLR, 18–24 Jul 2021b. URL https://proceedings.mlr.press/v139/asi21b.html.

Rina Foygel Barber and John C Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*, 2014.

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

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

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020.

Raef Bassily, Cristóbal Guzmán, and Michael Menart. Differentially private stochastic optimization: New results in convex and non-convex settings. *arXiv preprint arXiv:2107.05585*, 2021a.

Raef Bassily, Cristóbal Guzmán, and Anupama Nandi. Non-euclidean differentially private stochastic convex optimization. In *Conference on Learning Theory*, pages 474–499. PMLR, 2021b.

Andrea Bittau, Ulfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnes, and Bernhard Seefeld. Prochlo: Strong privacy for analytics in the crowd. In *Proceedings of the Symposium on Operating Systems Principles (SOSP)*, pages 441–459, 2017.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

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

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Proceedings, Part I, of the 14th International Conference on Theory of Cryptography - Volume 9985*, page 635–658, Berlin, Heidelberg, 2016. Springer-Verlag. ISBN 9783662536407. doi: 10.1007/978-3-662-53641-4_24. URL https://doi.org/10.1007/978-3-662-53641-4_24.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. *Advances in neural information processing systems*, 21, 2008.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Xiangyi Chen, Steven Z Wu, and Mingyi Hong. Understanding gradient clipping in private sgd: A geometric perspective. *Advances in Neural Information Processing Systems*, 33: 13773–13782, 2020.

Albert Cheu, Adam Smith, Jonathan Ullman, David Zeber, and Maxim Zhilyaev. Distributed differential privacy via shuffling. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 375–403. Springer, 2019.

Albert Cheu, Matthew Joseph, Jieming Mao, and Binghui Peng. Shuffle private stochastic convex optimization. *arXiv preprint arXiv:2106.09805*, 2021.

Mark E Crovella, Murad S Taqqu, and Azer Bestavros. Heavy-tailed probability distributions. *A Practical Guide to Heavy Tails Statistical Techniques and Applications*, 1998.

Rudrajit Das, Satyen Kale, Zheng Xu, Tong Zhang, and Sujay Sanghavi. Beyond uniform lipschitz condition in differentially private optimization. *arXiv preprint arXiv:2206.10713*, 2022.

John Duchi. Lecture notes for statistics 311/electrical engineering 377. *URL: https://stanford.edu/class/stats311/lecture-notes.pdf*, 2021.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438, 2013. doi: 10.1109/FOCS.2013.53.

Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*. 2014.

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

Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Shuang Song, Kunal Talwar, and Abhradeep Thakurta. Encode, shuffle, analyze privacy revisited: Formalizations and empirical evaluation. *arXiv preprint arXiv:2001.03618*, 2020.

Irfan Faizullabhoy and Aleksandra Korolova. Facebook's advertising platform: New attack vectors and the need for interventions. *arXiv preprint arXiv:1803.10099*, 2018.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020a.

Vitaly Feldman, Audra McMillan, and Kunal Talwar. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling, 2020b.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012. doi: 10.1137/110848864. URL `https://doi.org/10.1137/110848864`.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

Matthew J Holland. Robust descent using smoothed multiplicative noise. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 703–711. PMLR, 2019.

Lijie Hu, Shuo Ni, Hanshen Xiao, and Di Wang. High dimensional differentially private stochastic optimization with heavy-tailed data. *arXiv preprint arXiv:2107.11136*, 2021.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *arXiv preprint:1912.04977*, 2019.

Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2204–2235. PMLR, 09–12 Jul 2020. URL `https://proceedings.mlr.press/v125/kamath20a.html`.

Gautam Kamath, Xingtu Liu, and Huanyu Zhang. Improved rates for differentially private stochastic convex optimization with heavy-tailed data. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10633–10660. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/kamath22a.html`.

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

Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural computation*, 11(6):1427–1453, 1999.

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

Andrei Nikolaevich Kolmogorov and Vladimir Mikhailovich Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.

Janardhan Kulkarni, Yin Tat Lee, and Daogao Liu. Private non-smooth empirical risk minimization and stochastic convex optimization in subquadratic steps. *arXiv preprint arXiv:2103.15352*, 2021.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020.

Yunwen Lei and Yiming Ying. Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations*, 2021.

Andrew Lowy and Meisam Razaviyayn. Output perturbation for differentially private convex optimization with improved population loss bounds, runtimes and applications to private adversarial training. *arXiv preprint:2102.04704*, 2021a.

Andrew Lowy and Meisam Razaviyayn. Private federated learning without a trusted server: Optimal algorithms for convex losses, 2021b.

Andrew Lowy, Ali Ghafelebashi, and Meisam Razaviyayn. Private non-convex federated learning without a trusted server. *arXiv preprint arXiv:2203.06735*, 2022.

Natalia Markovich. *Nonparametric analysis of univariate heavy-tailed data: research and practice*. John Wiley & Sons, 2008.

Cain Mckay. *Probability and Statistics*. Scientific e-Resources, 2019.

Frank D McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30, 2009.

Stanislav Minsker. U-statistics of growing order and sub-gaussian mean estimators with sharp constants. *arXiv preprint arXiv:2202.11842*, 2022.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE, 2019.

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

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pages 547–562. University of California Press, 1961.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, volume 2, page 5, 2009.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Evading the curse of dimensionality in unconstrained private glms. In *International Conference on Artificial Intelligence and Statistics*, pages 2638–2646. PMLR, 2021.

Sebastian U. Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint:1907.04232*, 2019.

Youming Tao, Yulian Wu, Peng Zhao, and Di Wang. Optimal rates of (locally) differentially private heavy-tailed multi-armed bandits. *arXiv preprint arXiv:2106.02575*, 2021.

Jonathan Ullman. Private multiplicative weights beyond linear queries. In *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 303–312, 2015.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

D Wang, M Ye, and J Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Proc. 31st Annual Conference on Advances in Neural Information Processing Systems (NIPS 2017)*, 2017.

Di Wang and Jinhui Xu. Differentially private ℓ_1-norm linear regression with heavy-tailed data. *arXiv preprint arXiv:2201.03204*, 2022.

Di Wang, Hanshen Xiao, Srinivas Devadas, and Jinhui Xu. On differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pages 10081–10091. PMLR, 2020.

Robert F Woolson and William R Clarke. *Statistical methods for the analysis of biomedical data*. John Wiley & Sons, 2011.

Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

Appendix

Appendix A. Additional Discussion of Related Work

DP SCO Without Uniform Lipschitz Continuity: The study of DP SCO without uniformly Lipschitz continuous loss functions was initiated by (Wang et al., 2020), who provided upper bounds for smooth convex/strongly convex loss. The work of (Kamath et al., 2022) provided lower bounds and improved, yet *suboptimal*, upper bounds for the convex case. Both of the works (Wang et al., 2020; Kamath et al., 2022) require f to be β_f-smooth. It is also worth mentioning that (Wang et al., 2020; Kamath et al., 2022) restricted attention to losses satisfying $\nabla F(w^*) = 0$ for $w^* \in \mathcal{W}$, i.e. $\mathcal{W}$ is a compact set containing the *unconstrained* optimum $w^* = \text{argmin}_{w \in \mathbb{R}^d} F(w) \in \mathcal{W}$. By comparison, we consider the more general *constrained* optimization problem $\min_{w \in \mathcal{W}} F(w)$, where $\mathcal{W}$ need not contain the global unconstrained optimum.

Here we provide a brief discussion of the techniques used in (Wang et al., 2020; Kamath et al., 2022). The work of (Wang et al., 2020) used a full batch (clipped, noisy) gradient descent based algorithm, building on the heavy-tailed mean estimator of (Holland, 2019). They bounded the excess risk of their algorithm by using a uniform convergence (Vapnik, 1999) argument, resulting in a suboptimal dependence on the dimension d. The work of (Kamath et al., 2022) used essentially the same approach as (Wang et al., 2020), but obtained an improved rate with a more careful analysis.[10] However, as discussed, the bound in (Kamath et al., 2022) is $\mathcal{O}\left(rD\sqrt{\frac{d}{n}}\right)$ when $\epsilon \approx 1$, which is still suboptimal.[11]

More recently, DP optimization with outliers was studied in special cases of sparse learning (Hu et al., 2021), multi-arm bandits (Tao et al., 2021), and ℓ_1-norm linear regression (Wang and Xu, 2022).

DP ERM and DP GLMs without Uniform Lipschitz continuity: The work of (Asi et al., 2021a) provides bounds for constrained DP *ERM* with arbitrary convex loss functions using a Noisy Clipped SGD algorithm that is similar to our Algorithm 6, except that their algorithm is multi-pass and ours is one-pass. In a concurrent work, (Das et al., 2022) considered DP *ERM* in the *unconstrained* setting with convex and non-convex loss functions. Their algorithm, noisy clipped SGD, is also similar to Algorithm 6 and the algorithm of (Asi et al., 2021a). The results in (Das et al., 2022) are not directly comparable to (Asi et al., 2021a) since (Das et al., 2022) consider the unconstrained setting while (Asi et al., 2021a) consider the constrained setting, but the rates in (Asi et al., 2021a) are faster. (Das et al., 2022) also analyzes the convergence of noisy clipped SGD with smooth non-convex loss functions.

The works of (Song et al., 2021; Arora et al., 2022) consider *generalized linear models (GLMs)*, a particular subclass of convex loss functions and provide empirical and population risk bounds for the *unconstrained* DP optimization problem. The unconstrained setting is not comparable to the constrained setting that we consider here: in the unconstrained case,

10. Additionally, (Kamath et al., 2022, Theorem 5.2) provided a bound via noisy gradient descent with the clipping mechanism of (Kamath et al., 2020), but this bound is inferior (in the practical privacy regime $\epsilon \approx 1$) to their bound in (Kamath et al., 2022, Theorem 5.4) that used the estimator of (Holland, 2019).
11. The bound in (Kamath et al., 2022, Theorem 5.4) for $k = 2$ is stated in the notation of Assumption 3 and thus has an extra factor of $\sqrt{d}$, compared to the bound written here. We write their bound in terms of our Assumption 1, replacing their γd term by $r\sqrt{d}$.

a dimension-independent upper bound is achievable, whereas our lower bounds (which apply to GLMs) imply that a dependence on the dimension d is necessary in the constrained case.

Other works on gradient clipping: The gradient clipping technique (and adaptive variants of it) has been studied empirically in works such as (Abadi et al., 2016; Chen et al., 2020; Andrew et al., 2021), to name a few. The work of (Chen et al., 2020) shows that gradient clipping can prevent SGD from converging, and describes the clipping bias with a disparity measure between the gradient distribution and a geometrically symmetric distribution.

Optimization with biased gradient oracles: The works (Ajalloeian and Stich, 2020; Asi et al., 2021a) analyze SGD with biased gradient oracles. Our work provides a tighter bound for smooth, strongly convex functions and analyzes accelerated SGD and proximal SGD with biased gradient oracles.

DP SO with Uniformly Lipschitz loss functions: In the absence of outlier data, there are a multitude of works studying uniformly Lipschitz DP SO, mostly in the convex/strongly convex case. We do not attempt to provide a comprehensive list of these here, but will name the most notable ones, which provide optimal or state-of-the-art utility guarantees. The first suboptimal bounds for DP SCO were provided in (Bassily et al., 2014). The work of (Bassily et al., 2019) established the optimal rate for non-strongly convex DP SCO, by bounding the uniform stability of Noisy DP SGD (without clipping). The strongly convex case was addressed by (Feldman et al., 2020a), who also provided optimal rates in linear times for sufficiently smooth, convex losses. Since then, other works have provided faster and simpler (optimal) algorithms for the non-smooth DP SCO problem (Bassily et al., 2020; Asi et al., 2021b; Kulkarni et al., 2021; Bassily et al., 2021a) and considered DP SCO with different geometries (Asi et al., 2021b; Bassily et al., 2021b). State-of-the-art rates for DP SO with the proximal PL condition are due to (Lowy et al., 2022).

Appendix B. Other Bounded Moment Conditions Besides Assumption 1

In this section, we give the alternate bounded moment assumption made in (Wang et al., 2020; Kamath et al., 2022) and a third bounded moment condition, and discuss the relationships between these assumptions. The notation presented here will be necessary in order to state the sharper versions of our linear-time excess risk bounds and the asymptotically optimal excess risk bounds under the coordinate-wise assumption of (Wang et al., 2020; Kamath et al., 2022) (which our Algorithm 3 also attains). First, we introduce a relaxation of Assumption 1:

Assumption 2 *There exists $k \geqslant 2$ and $r^{(k)} > 0$ such that $\sup_{w \in \mathcal{W}} \mathbb{E}\left[\|\nabla f(w,x)\|_2^k\right] \leqslant r^{(k)}$, for all subgradients $\nabla f(w, x_i) \in \partial_w f(w, x_i)$. Denote $r_k := (r^{(k)})^{1/k}$.*

Assumption 1 implies Assumption 2 for $r \leqslant \widetilde{r}$. Next, we precisely state the coordinate-wise moment bound assumption that is used in (Wang et al., 2020; Kamath et al., 2022) for differentiable f:

Assumption 3 (Used by (Wang et al., 2020; Kamath et al., 2022)[12] , but not in this work) *There exists $k \geqslant 2$ and $\gamma > 0$ such that $\sup_{w \in \mathcal{W}} \mathbb{E}|\langle \nabla f(w,x) - \nabla F(w), e_j \rangle|^k \leqslant \gamma$, for all $j \in [d]$, where e_j denotes the j-th standard basis vector in $\mathbb{R}^d$. Also, $L \triangleq \sup_{w \in \mathcal{W}} \|\nabla F(w)\| \leqslant \sqrt{d}\gamma^{1/k}$.*

Lemma 20 allows us compare our results in Section 4 obtained under Assumption 2 to the results in (Wang et al., 2020; Kamath et al., 2022), which require Assumption 3.

Lemma 20 *Suppose Assumption 3 holds. Then, Assumption 2 holds for $r_k \leqslant 4\sqrt{d}\gamma^{1/k}$.*

Proof We use the following inequality, which can easily be verified inductively, using Cauchy-Schwartz and Young's inequalities: for any vectors $u, v \in \mathbb{R}^d$, we have

$$\|u\|^k \leqslant 2^{k-1}\left(\|u - v\|^k + \|v\|^k\right). \tag{5}$$

Therefore,

$$\begin{aligned}
r^{(k)} &= \sup_{w \in \mathcal{W}} \mathbb{E}\|\nabla f(w,x)\|^k \\
&\leqslant 2^{k-1}\left(\sup_{w \in \mathcal{W}} \mathbb{E}\|\nabla f(w,x) - \nabla F(w)\|^k + L^k\right) \\
&= 2^{k-1}\left(\sup_{w \in \mathcal{W}} \mathbb{E}\left[\left\{\sum_{j=1}^d |\langle \nabla f(w,x) - \nabla F(w), e_j \rangle|^2\right\}^{k/2}\right] + L^k\right) \\
&\leqslant (2L)^k + 2^k d^{k/2} \sup_{w \in \mathcal{W}} \mathbb{E}\left[\frac{1}{d}\sum_{j=1}^d |\langle \nabla f(w,x) - \nabla F(w), e_j \rangle|^k\right],
\end{aligned}$$

where we used convexity of the function $\phi(y) = y^{k/2}$ for all $y \geqslant 0, k \geqslant 2$ and Jensen's inequality in the last inequality. Now using linearity of expectation and Assumption 3 gives us

$$r^{(k)} \leqslant 2^k\left(L^k + d^{k/2}\gamma\right) \leqslant 2^{k+1} d^{k/2}\gamma,$$

since $L^k \leqslant d^{k/2}\gamma$ by hypothesis. ■

Remark 21 *Since Assumption 2 is implied by Assumption 3, the upper bounds that we obtain under Assumption 2 also hold (up to constants) if we grant Assumption 3 instead, with $r \leftrightarrow \sqrt{d}\gamma^{1/k}$. Also, in Appendix E.4, we will use Lemma 20 to show that our optimal excess risk bounds under Assumption 1 imply asymptotically optimal excess risk bounds under Assumption 3.*

12. The work of (Kamath et al., 2022) assumes that $L \lesssim \gamma^{1/k} = 1$. On the other hand, (Wang et al., 2020) assumes that F is β-smooth and $\nabla F(w^*) = 0$ for some $w^* \in \mathcal{W}$, which implies $L \leqslant 2\beta D$.

Appendix C. Correcting the Errors in the Strongly Convex Upper Bounds Claimed in (Kamath et al., 2022; Wang et al., 2020)

While the ICML 2022 paper (Kamath et al., 2022, Theorem 5.6) claimed an upper bound for smooth strongly convex losses that is tight up to a factor of $\widetilde{\mathcal{O}}(\kappa_f^2)$—where $\kappa_f = \beta_f/\mu$ is the uniform condition number of $f(\cdot, x)$ over all $x \in \mathcal{X}$–we identify an issue with their proof that invalidates their result. A similar issue appears in the proof of (Wang et al., 2020, Theorems 5 and 7), which (Kamath et al., 2022) built upon. We then show how to salvage a correct upper bound within the framework of (Kamath et al., 2022), albeit at the cost of an additional factor of κ_f.[13]

The proof of (Kamath et al., 2022, Theorem 5.6) relies on (Kamath et al., 2022, Theorem 3.2). The proof of (Kamath et al., 2022, Theorem 3.2), in turn, bounds $\mathbb{E}\|w_T - w^*\| \leqslant \frac{(\lambda+L)(M+1)G}{\lambda L}$ in the notation of (Kamath et al., 2022), where L is the smoothness parameter, λ is the strong convexity parameter (so $L \geqslant \lambda$), and M is the diameter of $\mathcal{W}$. Then, it is *incorrectly* deduced that $\mathbb{E}[\|w_T - w^*\|^2] \leqslant \left(\frac{(\lambda+L)(M+1)G}{\lambda L}\right)^2$ (final line of the proof). Notice that $\mathbb{E}[\|w_T - w^*\|^2]$ can be much larger than $(\mathbb{E}\|w_T - w^*\|)^2$ in general: for example, if $\|w_T - w^*\|$ has the Pareto distribution with shape parameter $\alpha \in (1, 2]$ and scale parameter 1, then $(\mathbb{E}\|w_T - w^*\|)^2 = \left(\frac{\alpha}{\alpha-1}\right)^2 \ll \mathbb{E}(\|w_T - w^*\|^2) = \infty$.

As a first attempt to correct this issue, one could use Young's inequality to instead bound

$$
\begin{aligned}
\mathbb{E}[\|w_T - w^*\|^2] &\leqslant 2\left(1 - \frac{2\lambda L}{(\lambda+L)^2}\right)\mathbb{E}[\|w_{T-1} - w^*\|^2] + \frac{2G^2}{(\lambda+L)^2} \\
&\leqslant \left[2\left(1 - \frac{2\lambda L}{(\lambda+L)^2}\right)\right]^T \|w_0 - w^*\|^2 + \frac{2G^2}{(\lambda+L)^2}\sum_{t=0}^{T-1}\left[2\left(1 - \frac{2\lambda L}{(\lambda+L)^2}\right)\right]^t,
\end{aligned}
$$

but the geometric series above diverges to $+\infty$ as $T \to \infty$, since $2\left(1 - \frac{2\lambda L}{(\lambda+L)^2}\right) \geqslant 1 \iff (\lambda - L)^2 \geqslant 0$.

Next, we show how to modify the proof of (Kamath et al., 2022, Theorem 5.6) in order to obtain a correct excess risk upper bound of

$$
\widetilde{\mathcal{O}}\left(\frac{\gamma^{2/k}}{\mu} d\left[\frac{(\beta_f/\mu)^3}{n} + \left(\frac{\sqrt{d(\beta_f/\mu)^3}}{\epsilon n}\right)^{(2k-2)/k}\right]\right) \tag{6}
$$

(in our notation). This correction was derived in collaboration with the authors of (Kamath et al., 2022), who have also updated the arXiv version of their paper accordingly. By waiting until the very of the proof of (Kamath et al., 2022, Theorem 3.2) to take expectation, we can derive

$$
\|w_t - w^*\| \leqslant \left(1 - \frac{\lambda L}{(\lambda+L)^2}\right)\|w_{t-1} - w^*\| + \frac{\|\widetilde{\nabla} F(w_{t-1}) - \nabla F(w_{t-1})\|}{\lambda + L} \tag{7}
$$

13. The corrected result of (Kamath et al., 2022), derived here, is also included in the latest arXiv version of their paper. We communicated with the authors of (Kamath et al., 2022) to obtain this correct version.

for all t, where we use their $L = \beta_f$ and $\lambda = \mu$ notation but our notation F and $\widetilde{\nabla} F$ for the population loss and its biased noisy gradient estimate (instead of their $L_{\mathcal{D}}$ notation). By iterating (7), we can get

$$\|w_T - w^*\| \leqslant \left(1 - \frac{2\lambda L}{(\lambda+L)^2}\right)^T \|w_0 - w^*\| + \sum_{t=0}^{T-1}\left(1 - \frac{2\lambda L}{(\lambda+L)^2}\right)^t \left[\frac{\|\widetilde{\nabla} F(w_{T-t}) - \nabla F(w_{T-t})\|}{\lambda + L}\right].$$

Squaring both sides and using Cauchy-Schwartz, we get

$$\|w_T - w^*\|^2 \leqslant 2\left(1 - \frac{2\lambda L}{(\lambda+L)^2}\right)^{2T} \|w_0 - w^*\|^2 + T\sum_{t=0}^{T-1}\left(1 - \frac{2\lambda L}{(\lambda+L)^2}\right)^{2t} \left[\frac{\|\widetilde{\nabla} F(w_{T-t}) - \nabla F(w_{T-t})\|}{\lambda + L}\right]^2.$$

Using L-smoothness of F and the assumption made in (Kamath et al., 2022) that $\nabla F(w^*) = 0$, and *then* taking expectation yields

$$\mathbb{E}F(w_T) - F^* \lesssim L\|w_0 - w^*\|^2 \left(1 - \frac{2\lambda L}{(\lambda+L)^2}\right)^{2T} + TG^2\frac{L}{\lambda}, \tag{8}$$

where $G^2 \geqslant \mathbb{E}\left[\|\widetilde{\nabla} F(w_{T-t}) - \nabla F(w_{T-t})\|^2\right]$ for all t. It is necessary and sufficient to choose $T = \widetilde{\Omega}(L/\lambda)$ to make the first term in the right-hand side of (8) less than the second term (up to logarithms). With this choice of T, we get

$$\mathbb{E}F(w_T) - F^* = \widetilde{\mathcal{O}}\left(G^2\kappa_f^2\right), \tag{9}$$

where $\kappa_f = L/\lambda$. Next, we apply the bound on G^2 for the MeanOracle that is used in (Kamath et al., 2022); this bound is stated in the version of (Kamath et al., 2022, Lemma B.5) that appears in the updated (November 1, 2022) arXiv version of their paper. The bound (for general γ) is $G^2 = \widetilde{\mathcal{O}}\left(\gamma^{2/k}\left[\frac{Td}{n} + d\left(\frac{\sqrt{d}T^{3/2}}{\epsilon n}\right)^{(2k-2)/k}\right]\right)$. Plugging this bound on G^2 into (9) yields (6).

Appendix D. More Differential Privacy Preliminaries

We collect some basic facts about DP algorithms that will be useful in the proofs of our results. Our algorithms use the *Gaussian mechanism* to achieve zCDP:

Proposition 22 *(Bun and Steinke, 2016, Proposition 1.6) Let $q : \mathcal{X}^n \to \mathbb{R}$ be a query with ℓ_2-sensitivity $\Delta := \sup_{X \sim X'} \|q(X) - q(X')\|$. Then the Gaussian mechanism, defined by $\mathcal{M} : \mathcal{X}^n \to \mathbb{R}$, $M(X) := q(X) + u$ for $u \sim \mathcal{N}(0, \sigma^2)$, is ρ-zCDP if $\sigma^2 \geqslant \frac{\Delta^2}{2\rho}$.*

The (adaptive) composition of zCDP algorithms is zCDP, with privacy parameters adding:

Lemma 23 *(Bun and Steinke, 2016, Lemma 2.3) Suppose $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ satisfies ρ-zCDP and $\mathcal{A}' : \mathcal{X}^n \times \mathcal{Y} \to \mathcal{Z}$ satisfies ρ'-zCDP (as a function of its first argument). Define the composition of $\mathcal{A}$ and $\mathcal{A}'$, $\mathcal{A}'' : \mathcal{X}^n \to \mathcal{Z}$ by $\mathcal{A}''(X) = \mathcal{A}'(X, \mathcal{A}(X))$. Then $\mathcal{A}''$ satisfies $(\rho + \rho')$-zCDP. In particular, the composition of T ρ-zCDP mechanisms is a $T\rho$-zCDP mechanism.*

The definitions of DP and zCDP given above do not dictate *how* the algorithm $\mathcal{A}$ operates. In particular, they allow $\mathcal{A}$ to send sensitive data to a third party curator/analyst, who can then add noise to the data. However, in certain practical applications (e.g. federated learning (Kairouz et al., 2019)), there is no third party that can be trusted to handle sensitive user data. On the other hand, it is often more realistic to have a secure *shuffler* (a.k.a. mixnet): in each iteration of the algorithm, the shuffler receives encrypted noisy reports (e.g. noisy stochastic gradients) from each user and applies a uniformly random permutation to the n reports, thereby anonymizing them (and amplifying privacy) (Bittau et al., 2017; Cheu et al., 2019; Erlingsson et al., 2020; Feldman et al., 2020b). An algorithm is *shuffle private* if all of these "shuffled" reports are DP:

Definition 24 *(Shuffle Differential Privacy (Bittau et al., 2017; Cheu et al., 2019)) A randomized algorithm is (ϵ, δ)-shuffle DP (SDP) if the collection of reports output by the shuffler satisfies Definition 1.*

Appendix E. Details and Proofs for Section 3: Optimal Rates for (Strongly) Convex Losses

In order to precisely state (sharper forms of) Theorems 6 and 12, we will need to introduce some notation.

E.1. Notation

For a batch of data $X \in \mathcal{X}^m$, we define the k-th *empirical moment* of $f(w, \cdot)$ by

$$\widehat{r}_m(X)^{(k)} = \sup_{w \in \mathcal{W}} \sup_{\{\nabla f(w,x_i) \in \partial_w f(w,x_i)\}} \frac{1}{m} \sum_{i=1}^{m} \|\nabla f(w, x_i)\|^k,$$

where the supremum is also over all subgradients $\nabla f(w, x_i) \in \partial_w f(w, x_i)$ in case f is not differentiable. For $X \sim \mathcal{D}^m$, we denote the k-th *expected empirical moment* by

$$\widetilde{e}_m^{(k)} := \mathbb{E}[\widehat{r}_m(X)^{(k)}]$$

and let

$$\widetilde{r}_{k,m} := (\widetilde{e}_m^{(k)})^{1/k}.$$

Note that $\widetilde{r}_{k,1} = \widetilde{r}_k$. Our excess risk upper bounds will depend on a weighted average of the expected empirical moments for different batch sizes $m \in \{1, 2, 4, 8, \cdots, n\}$, with more weight being given to $\widetilde{r}_m$ for large m (which are smaller, by Lemma 25 below): for $n = 2^l$, define

$$\widetilde{R}_{k,n} := \sqrt{\sum_{i=1}^{l} 2^{-i} \widetilde{r}_{k,n_i}^2},$$

where $n_i = 2^{-i} n$.

Lemma 25 *Under Assumptions 1 and 2, we have: $\widetilde{r}^{(k)} = \widetilde{e}_1^{(k)} \geqslant \widetilde{e}_2^{(k)} \geqslant \widetilde{e}_4^{(k)} \geqslant \widetilde{e}_8^{(k)} \geqslant \cdots \geqslant r^{(k)}$. In particular, $\widetilde{R}_{k,n} \leqslant \widetilde{r}_k$.*

Proof Let $l \in \mathbb{N}$, $n = 2^l$ and consider

$$\begin{aligned}\widehat{r}_n(X)^{(k)} &= \frac{1}{n}\sup_w \left(\sum_{i=1}^{n/2}\|\nabla f(w,x_i)\|^k + \sum_{i=n/2+1}^{n}\|\nabla f(w,x_i)\|^k\right)\\ &\leqslant \frac{1}{n}\left(\sup_w \sum_{i=1}^{n/2}\|\nabla f(w,x_i)\|^k + \sup_w \sum_{i=n/2+1}^{n}\|\nabla f(w,x_i)\|^k\right).\end{aligned}$$

Taking expectations over the random draw of $X \sim \mathcal{D}^n$ yields $\widetilde{e}_n^{(k)} \leqslant \widetilde{e}_{n/2}^{(k)}$. Thus, $\widetilde{R}_{k,n} \leqslant \widetilde{r}_k$ by the definition of $\widetilde{R}_n$. ■

E.2. Localized Noisy Clipped Subgradient Method (Section 3.1)

We begin by proving the technical ingredients that will be used in the proof of Theorem 6. First, we will prove a variant of Lemma 5 that bounds the bias and variance of the subgradient estimator in Algorithm 2.

Lemma 26 *Let $\widehat{F}_\lambda(w) = \frac{1}{n}\sum_{i=1}^n f(w,x_i) + \frac{\lambda}{2}\|w-w_0\|^2$ be a regularized empirical loss on a closed convex domain $\mathcal{W}$ with ℓ_2-diameter D. Let $\widetilde{\nabla} F_\lambda(w_t) = \nabla \widehat{F}_\lambda(w_t) + b_t + N_t = \frac{1}{n}\sum_{i=1}^n \Pi_C(\nabla f(w,x_i)) + \lambda(w-w_0) + N_t$ be the biased, noisy subgradients of the regularized empirical loss in Algorithm 2, with $N_t \sim \mathcal{N}(0,\sigma^2\mathbf{I}_d)$ and $b_t = \frac{1}{n}\sum_{i=1}^n \Pi_C(\nabla f(w_t,x_i)) - \frac{1}{n}\sum_{i=1}^n \nabla f(w_t,x_i)$. Assume $\widehat{r}_n(X)^{(k)} \geqslant \sup_{w\in\mathcal{W}}\left\{\frac{1}{n}\sum_{i=1}^n\|\nabla f(w,x_i)\|^k\right\}$ for all $\nabla f(w,x_i) \in \partial_w f(w,x_i)$. Then, for any $T \geqslant 1$, we have:*

$$\hat{B} := \sup_{t\in[T]}\|b_t\| \leqslant \frac{\widehat{r}_n(X)^{(k)}}{(k-1)C^{k-1}}$$

and

$$\hat{\Sigma}^2 := \sup_{t\in[T]}\mathbb{E}\|N_t\|^2 = d\sigma^2.$$

Proof Fix any t. We have

$$\begin{aligned}\|b_t\| &= \left\|\frac{1}{n}\sum_{i=1}^n \Pi_C(\nabla f(w_t,x_i)) - \frac{1}{n}\sum_{i=1}^n \nabla f(w_t,x_i)\right\|\\ &\leqslant \frac{1}{(k-1)C^{k-1}}\left[\frac{1}{n}\sum_{i=1}^n\|\nabla f(w_t,x_i)\|^k\right], \end{aligned} \tag{10}$$

by Lemma 5 applied with $\mathcal{D}$ as the empirical distribution on X, and z_i in Lemma 5 corresponding to $\nabla f(w_t,x_i)$ in (10). Taking supremum over t of both sides of (10) and recalling the definition of $\widehat{r}_n(X)^{(k)}$ proves the bias bound. The noise variance bound is immediate from the distribution of N_t. ■

Using Lemma 26, we can obtain the following convergence guarantee for Algorithm 2:

Lemma 27 (Re-statement of Lemma 7) *Fix $X \in \mathcal{X}^n$ and let $\widehat{F}_\lambda(w) = \frac{1}{n}\sum_{i=1}^n f(w, x_i) + \frac{\lambda}{2}\|w - w_0\|^2$ for $w_0 \in \mathcal{W}$, where $\mathcal{W}$ is a closed convex domain with diameter D. Assume $f(\cdot, x)$ is convex and $\widehat{r}_n(X)^{(k)} \geqslant \sup_{w \in \mathcal{W}}\left\{\frac{1}{n}\sum_{i=1}^n \|\nabla f(w, x_i)\|^k\right\}$ for all $\nabla f(w, x_i) \in \partial_w f(w, x_i)$. Denote $\widehat{r}_n(X) = \left[\widehat{r}_n(X)^{(k)}\right]^{1/k}$ and $\hat{w} = \operatorname{argmin}_{w \in \mathcal{W}} \widehat{F}_\lambda(w)$. Let $\eta \leqslant \frac{2}{\lambda}$. Then, the output of Algorithm 2 satisfies*

$$\mathbb{E}\|w_T - \hat{w}\|^2 \leqslant \exp\left(-\frac{\lambda \eta T}{2}\right)\|w_0 - \hat{w}\|^2 + \frac{8\eta}{\lambda}\left(\widehat{r}_n(X)^2 + \lambda^2 D^2 + d\sigma^2\right) + \frac{20}{\lambda^2}\left(\frac{\widehat{r}_n(X)^{(k)}}{(k-1)C^{k-1}}\right)^2,$$

where $\sigma^2 = \frac{4C^2 T}{n^2 \epsilon^2}$.

Proof We use the notation of Lemma 26 and write $\widetilde{\nabla} F_\lambda(w_t) = \nabla \widehat{F}_\lambda(w_t) + b_t + N_t = \frac{1}{n}\sum_{i=1}^n \Pi_C(\nabla f(w, x_i)) + \lambda(w - w_0) + N_t$ as the biased, noisy subgradients of the regularized empirical loss in Algorithm 2, with $N_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ and $b_t = \frac{1}{n}\sum_{i=1}^n \Pi_C(\nabla f(w_t, x_i)) - \frac{1}{n}\sum_{i=1}^n \nabla f(w_t, x_i)$. Denote $y_{t+1} = w_t - \eta \widetilde{\nabla} F_\lambda(w_t)$, so that $w_{t+1} = \Pi_{\mathcal{W}}(y_{t+1})$. For now, condition on the randomness of the algorithm (noise). By strong convexity, we have

$$\begin{aligned}
\widehat{F}_\lambda(w_t) - \widehat{F}_\lambda(\hat{w}) &\leqslant \langle \nabla \widehat{F}_\lambda(w_t), w_t - \hat{w}\rangle - \frac{\lambda}{2}\|w_t - \hat{w}\|^2 \\
&= \langle \widetilde{\nabla} F_\lambda(w_t), w_t - \hat{w}\rangle - \frac{\lambda}{2}\|w_t - \hat{w}\|^2 + \langle \nabla \widehat{F}_\lambda(w_t) - \widetilde{\nabla} F_\lambda(w_t), w_t - \hat{w}\rangle \\
&= \frac{1}{2\eta}\left(\|w_t - \hat{w}\|^2 + \|w_t - y_{t+1}\|^2 - \|y_{t+1} - \hat{w}\|^2\right) - \frac{\lambda}{2}\|w_t - \hat{w}\|^2 \\
&\quad + \langle \nabla \widehat{F}_\lambda(w_t) - \widetilde{\nabla} F_\lambda(w_t), w_t - \hat{w}\rangle \\
&= \frac{1}{2\eta}\left(\|w_t - \hat{w}\|^2(1 - \lambda\eta) - \|y_{t+1} - \hat{w}\|^2\right) + \frac{\eta}{2}\|\widetilde{\nabla} F_\lambda(w_t)\|^2 \\
&\quad + \langle \nabla \widehat{F}_\lambda(w_t) - \widetilde{\nabla} F_\lambda(w_t), w_t - \hat{w}\rangle \\
&\leqslant \frac{1}{2\eta}\left(\|w_t - \hat{w}\|^2(1 - \lambda\eta) - \|w_{t+1} - \hat{w}\|^2\right) + \frac{\eta}{2}\|\widetilde{\nabla} F_\lambda(w_t)\|^2 - \langle b_t + N_t, w_t - \hat{w}\rangle,
\end{aligned}$$

where we used non-expansiveness of projection and the definition of $\widetilde{\nabla} F_\lambda(w_t)$ in the last line. Now, re-arranging this inequality and taking expectation, we get

$$\begin{aligned}
\mathbb{E}[\|w_{t+1} - \hat{w}\|^2] &\leqslant -2\eta \mathbb{E}[\widehat{F}_\lambda(w_t) - \widehat{F}_\lambda(\hat{w})] + \mathbb{E}\|w_t - \hat{w}\|^2(1 - \lambda\eta) + \eta^2 \mathbb{E}\|\widetilde{\nabla} F_\lambda(w_t)\|^2 \\
&\quad - 2\eta \mathbb{E}\langle b_t + N_t, w_t - \hat{w}\rangle \\
&\leqslant \mathbb{E}\|w_t - \hat{w}\|^2(1 - \lambda\eta) + \eta^2 \mathbb{E}\|\widetilde{\nabla} F_\lambda(w_t)\|^2 - 2\eta \mathbb{E}\langle b_t, w_t - \hat{w}\rangle,
\end{aligned}$$

by optimality of $\hat{w}$ and the assumption that the noise N_t is independent of $w_t - \hat{w}$ and zero mean. Also,

$$\begin{aligned}
\mathbb{E}\|\widetilde{\nabla} F_\lambda(w_t)\|^2 &\leqslant 2\left(\mathbb{E}\|\nabla \widehat{F}_\lambda(w_t)\|^2 + \|b_t\|^2 + \mathbb{E}\|N_t\|^2\right) \\
&\leqslant 2\left(2\widehat{r}_n(X)^2 + 2\lambda^2 D^2 + \hat{B}^2 + \hat{\Sigma}^2\right),
\end{aligned}$$

where $\hat{B} := \sup_{t \in [T]} \|b_t\| \leqslant \frac{\widehat{r}_n(X)^{(k)}}{(k-1)C^{k-1}}$ and $\hat{\Sigma}^2 := \sup_{t \in [T]} \mathbb{E}\|N_t\|^2 = d\sigma^2$. by Lemma 26. We also used Young's and Jensen's inequalities and the fact that $\mathbb{E}N_t = 0$. Further,

$$|\mathbb{E}\langle b_t, w_t - \hat{w}\rangle| \leqslant \frac{\hat{B}^2}{\lambda} + \frac{\lambda}{4}\mathbb{E}\|w_t - \hat{w}\|^2,$$

by Young's inequality. Combining these pieces yields

$$\mathbb{E}\|w_{t+1} - \hat{w}\|^2 \leqslant \left(1 - \frac{\lambda\eta}{2}\right)\mathbb{E}\|w_t - \hat{w}\|^2 + 4\eta^2\left(\widehat{r}_n(X)^2 + \lambda^2 D^2 + \hat{B}^2 + \hat{\Sigma}^2\right) + \frac{2\eta\hat{B}^2}{\lambda}. \quad (11)$$

Iterating (11) gives us

$$\begin{aligned}
\mathbb{E}\|w_T - \hat{w}\|^2 &\leqslant \left(1 - \frac{\lambda\eta}{2}\right)^T \|w_0 - \hat{w}\|^2 + \left[4\eta^2\left(\widehat{r}_n(X)^2 + \lambda^2 D^2 + \hat{B}^2 + \hat{\Sigma}^2\right) + \frac{2\eta\hat{B}^2}{\lambda}\right]\sum_{t=0}^{T-1}\left(1 - \frac{\lambda\eta}{2}\right)^t \\
&\leqslant \left(1 - \frac{\lambda\eta}{2}\right)^T \|w_0 - \hat{w}\|^2 + \left[4\eta^2\left(\widehat{r}_n(X)^2 + \lambda^2 D^2 + \hat{B}^2 + \hat{\Sigma}^2\right) + \frac{2\eta\hat{B}^2}{\lambda}\right]\left(\frac{2}{\lambda\eta}\right) \\
&= \left(1 - \frac{\lambda\eta}{2}\right)^T \|w_0 - \hat{w}\|^2 + \frac{8\eta}{\lambda}\left(\widehat{r}_n(X)^2 + \lambda^2 D^2 + \hat{B}^2 + \hat{\Sigma}^2\right) + \frac{4\hat{B}^2}{\lambda^2} \\
&\leqslant \exp\left(-\frac{\lambda\eta T}{2}\right)\|w_0 - \hat{w}\|^2 + \frac{8\eta}{\lambda}\left(\widehat{r}_n(X)^2 + \lambda^2 D^2 + \hat{B}^2 + \hat{\Sigma}^2\right) + \frac{4\hat{B}^2}{\lambda^2} \\
&\leqslant \exp\left(-\frac{\lambda\eta T}{2}\right)\|w_0 - \hat{w}\|^2 + \frac{8\eta}{\lambda}\left(\widehat{r}_n(X)^2 + \lambda^2 D^2 + \hat{\Sigma}^2\right) + \frac{20\hat{B}^2}{\lambda^2},
\end{aligned}$$

since $\eta \leqslant \frac{2}{\lambda}$. Plugging in the bounds on $\hat{B}$ and $\hat{\Sigma}$ from Lemma 26 completes the proof. ■

Proposition 28 (Precise statement of Proposition 9) *Let $f(\cdot, x)$ be convex for all x and grant Assumption 2 for $k = 2$. Suppose $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$ is α-on-average model stable. Then for any $\zeta > 0$, we have*

$$\mathbb{E}[F(\mathcal{A}(X)) - \widehat{F}_X(\mathcal{A}(X))] \leqslant \frac{r^{(2)}}{2\zeta} + \frac{\zeta}{2}\alpha^2.$$

Proof Let X, X', X^i be constructed as in Definition 8. We may write $\mathbb{E}[F(\mathcal{A}(X)) - \widehat{F}_X(\mathcal{A}(X))] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^n f(\mathcal{A}(X^i), x_i) - f(\mathcal{A}(X), x_i)]$, by symmetry and independence of x_i and $\mathcal{A}(X^i)$ (c.f. (Lei and Ying, 2020, Equation B.2)). Then by convexity, we have

$$\begin{aligned}
\mathbb{E}[F(\mathcal{A}(X)) - \widehat{F}_X(\mathcal{A}(X))] &\leqslant \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\langle \mathcal{A}(X^i) - \mathcal{A}(X), \nabla f(\mathcal{A}(X^i), x_i)\rangle] \\
&\leqslant \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{\zeta}{2}\|\mathcal{A}(X^i) - \mathcal{A}(X)\|^2 + \frac{1}{2\zeta}\|\nabla f(\mathcal{A}(X^i), x_i)\|^2\right].
\end{aligned}$$

Now, since $\mathcal{A}(X^i)$ is independent of x_i, we have:

$$\begin{aligned}\mathbb{E}\|\nabla f(\mathcal{A}(X^i), x_i)\|^2 &= \mathbb{E}[\mathbb{E}[\|\nabla f(\mathcal{A}(X^i), x_i)\|^2 | \mathcal{A}(X^i)]] \\ &\leqslant \sup_{w \in \mathcal{W}} \mathbb{E}[\|\nabla f(\mathcal{A}(X^i), x_i)\|^2 | \mathcal{A}(X^i) = w] \\ &= \sup_{w \in \mathcal{W}} \mathbb{E}[\|\nabla f(w, x_i)\|^2] \\ &\leqslant r^{(2)}.\end{aligned}$$

Combining the above inequalities and recalling Definition 8 yields the result. ■

To prove our excess risk bound for regularized ERM (i.e. Proposition 10), we require the following bound on the generalization error of ERM with strongly convex loss:

Proposition 29 *Let $f(\cdot, x)$ be λ-strongly convex, and grant Assumption 2. Let $\mathcal{A}(X) := \operatorname{argmin}_{w \in \mathcal{W}} \widehat{F}_X(w)$ be the ERM algorithm. Then,*

$$\mathbb{E}[F(\mathcal{A}(X)) - \widehat{F}_X(\mathcal{A}(X))] \leqslant \frac{2r^{(2)}}{\lambda n}.$$

Proof We first bound the stability of ERM and then use Proposition 9 to get a bound on the generalization error. The beginning of the proof is similar to the proof of (Lei and Ying, 2020, Proposition D.6): Let X, X', X^i be constructed as in Definition 8. By strong convexity of $\widehat{F}_{X^i}$ and optimality of $\mathcal{A}(X^i)$, we have

$$\frac{\lambda}{2}\|\mathcal{A}(X) - \mathcal{A}(X^i)\|^2 \leqslant \widehat{F}_{X^i}(\mathcal{A}(X)) - \widehat{F}_{X^i}(\mathcal{A}(X^i)),$$

which implies

$$\frac{1}{n}\sum_{i=1}^n \|\mathcal{A}(X) - \mathcal{A}(X^i)\|^2 \leqslant \frac{2}{\lambda n}\sum_{i=1}^n \left[\widehat{F}_{X^i}(\mathcal{A}(X)) - \widehat{F}_{X^i}(\mathcal{A}(X^i))\right]. \tag{12}$$

Now, for any $w \in \mathcal{W}$,

$$\begin{aligned}n\sum_{i=1}^n \widehat{F}_{X^i}(w) &= \sum_{i=1}^n [f(w, x_i') + \sum_{j \neq i} f(w, x_j)] \\ &= (n-1)n\widehat{F}_X(w) + n\widehat{F}_{X'}(w).\end{aligned}$$

Hence

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n \widehat{F}_{X^i}(\mathcal{A}(X))\right] &= \left(\frac{n-1}{n}\right)\mathbb{E}\widehat{F}_X(\mathcal{A}(X)) + \frac{1}{n}\mathbb{E}\widehat{F}_{X'}(\mathcal{A}(X)) \\ &= \left(\frac{n-1}{n}\right)\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n \widehat{F}_{X^i}(\mathcal{A}(X^i))\right] + \frac{1}{n}\mathbb{E}F(\mathcal{A}(X)),\end{aligned}$$

by symmetry and independence of $\mathcal{A}(X)$ and X'. Re-arranging the above equality and using symmetry yields

$$\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}\widehat{F}_{X^i}(\mathcal{A}(X))-\widehat{F}_{X^i}(\mathcal{A}(X^i))\right]=\frac{1}{n}\mathbb{E}\left[F(\mathcal{A}(X))-\widehat{F}_X(\mathcal{A}(X))\right]. \tag{13}$$

Combining (12) with (13) shows that ERM is α-on-average model stable for

$$\alpha^2=\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\|\mathcal{A}(X)-\mathcal{A}(X^i)\|^2\right]\leqslant\frac{2}{\lambda n}\mathbb{E}\left[F(\mathcal{A}(X))-\widehat{F}_X(\mathcal{A}(X))\right]. \tag{14}$$

The rest of the proof is where we depart from the analysis of (Lei and Ying, 2020) (which required smoothness of $f(\cdot,x)$): Bounding the right-hand side of (14) by Proposition 9 yields

$$\alpha^2\leqslant\frac{2}{\lambda n}\left(\frac{r^{(2)}}{2\zeta}+\frac{\zeta}{2}\alpha^2\right)$$

for any $\zeta>0$. Choosing $\zeta=\frac{\lambda n}{2}$, we obtain

$$\frac{\alpha^2}{2}\leqslant\frac{r^{(2)}}{\lambda n\zeta}=\frac{2r^{(2)}}{\lambda^2n^2},$$

and $\alpha^2\leqslant\frac{4r^{(2)}}{\lambda^2n^2}$. Applying Proposition 9 again yields (for any $\zeta'>0$)

$$\begin{aligned}\mathbb{E}[F(\mathcal{A}(X))-\widehat{F}_X(\mathcal{A}(X))]&\leqslant\frac{r^{(2)}}{2\zeta'}+\frac{\zeta'}{2}\left(\frac{4r^{(2)}}{\lambda^2n^2}\right)\\&\leqslant\frac{2r^{(2)}}{\lambda n},\end{aligned}$$

by the choice $\zeta'=\frac{\lambda n}{2}$. ■

Proposition 30 (Precise statement of Proposition 10) *Let $f(\cdot,x)$ be convex, $w_{i-1},y\in\mathcal{W}$, and $\hat{w}_i:=\operatorname{argmin}_{w\in\mathcal{W}}\widehat{F}_i(w)$, where $\widehat{F}_i(w):=\frac{1}{n_i}\sum_{j\in\mathcal{B}_i}f(w,x_j)+\frac{\lambda_i}{2}\|w-w_{i-1}\|^2$ (c.f. line 6 of Algorithm 3). Then,*

$$\mathbb{E}[F(\hat{w}_i)]-F(y)\leqslant\frac{2r^{(2)}}{\lambda_i n_i}+\frac{\lambda_i}{2}\|y-w_{i-1}\|^2,$$

where the expectation is over both the random draws of X from $\mathcal{D}$ and $\mathcal{B}_i$ from X.

Proof Denote the regularized population loss by $G_i(w):=\mathbb{E}[\widehat{F}_i(w)]=F(w)+\frac{\lambda_i}{2}\|w-w_{i-1}\|^2$. By Proposition 29, we have

$$\mathbb{E}[G_i(\hat{w}_i)-\widehat{F}_i(\hat{w}_i)]\leqslant\frac{2r^{(2)}}{\lambda_i n_i}.$$

Thus,

$$\begin{aligned}\frac{\lambda_i}{2}\mathbb{E}\|\hat{w}_i - w_{i-1}\|^2 + \mathbb{E}F(\hat{w}_i) &= \mathbb{E}G_i(\hat{w}_i) \\ &\leqslant \frac{2r^{(2)}}{\lambda_i n_i} + \mathbb{E}[\widehat{F}_i(\hat{w}_i)] \\ &\leqslant \frac{2r^{(2)}}{\lambda_i n_i} + \frac{\lambda_i}{2}\|y - w_{i-1}\|^2 + F(y), \end{aligned} \tag{15}$$

since $\mathbb{E}[\widehat{F}_i(\hat{w}_i)] = \mathbb{E}[\min_{w\in\mathcal{W}} \widehat{F}_i(w)] \leqslant \min_{w\in\mathcal{W}} \mathbb{E}[\widehat{F}_i(w)] = \min_{w\in\mathcal{W}} G_i(w) \leqslant \frac{\lambda_i}{2}\|y - w_{i-1}\|^2 + F(y)$. Subtracting $F(y)$ from both sides of (15) completes the proof. ■

We are ready to state and prove the precise form of Theorem 6, using the notation of Appendix E.1:

Theorem 31 (Precise Statement of Theorem 6) *Grant Assumption 1. Let $f(\cdot, x)$ be convex and L_f-Lipschitz for all x, with $L_f < \infty$ and let $\epsilon \leqslant \sqrt{d}$. Then, there are algorithmic parameters such that Algorithm 3 is $\frac{\epsilon^2}{2}$-zCDP, and has excess risk*

$$\mathbb{E}F(w_l) - F^* \lesssim \widetilde{R}_{2k,n} D\left(\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{d\ln(n)}}{\epsilon n}\right)^{\frac{k-1}{k}}\right).$$

Moreover, this excess risk is attained in $\widetilde{\mathcal{O}}(n^2 L_f^2/\widetilde{R}_{2k,n})$ subgradient evaluations.

Proof We choose $\sigma_i^2 = \frac{4C_i^2 T_i}{n_i^2\epsilon^2}$ for C_i and T_i to be determined exactly later. Note that for λ_i and η_i defined in Algorithm 3, we have $\eta_i \leqslant \frac{2}{\lambda_i}$ for all $i \in [l]$.

Privacy: Since the batches $\{\mathcal{B}_i\}_{i=1}^{l}$ are disjoint, it suffices (by parallel composition (McSherry, 2009)) to show that w_i (produced by T_i iterations of Algorithm 2 in line 7 of Algorithm 3) is $\frac{\epsilon^2}{2}$-zCDP for all $i \in [l]$. With clip threshold C_i and batch size n_i, the ℓ_2 sensitivity of the clipped subgradient update is bounded by $\Delta = \sup_{w, X\sim X'} \frac{1}{n_i}\|\sum_{j=1}^{n_i} \Pi_{C_i}(\nabla f(w, x_j)) - \Pi_{C_i}(\nabla f(w, x'_j))\| = \frac{1}{n_i}\sup_{w,x,x'}\|\Pi_{C_i}(\nabla f(w,x)) - \Pi_{C_i}(\nabla f(w, x'))\| \leqslant \frac{2C_i}{n_i}$. (Note that the terms arising from regularization cancel out.) Thus, by Proposition 22, conditional on the previous updates $w_{1:i}$, the $(i+1)$-st update in line 5 of Algorithm 2 satisfies $\frac{\epsilon^2}{2T_i}$-zCDP. Hence, Lemma 23 implies that w_i (in line 7 of Algorithm 3) is $\frac{\epsilon^2}{2}$-zCDP.

Excess risk: First, our choice of D_i ensures that $\hat{w}_i \in \mathcal{W}_i$, since

$$\begin{aligned}\widehat{F}_i(\hat{w}_i) = \frac{1}{n_i}\sum_{j\in\mathcal{B}_i} f(\hat{w}_i, x_j) + \frac{\lambda_i}{2}\|\hat{w}_i - w_{i-1}\|^2 \leqslant \widehat{F}_i(w_{i-1}) = \frac{1}{n_i}\sum_{j\in\mathcal{B}_i} f(w_{i-1}, x_j) \\ \implies \frac{\lambda_i}{2}\|\hat{w}_i - w_{i-1}\|^2 \leqslant L_f\|\hat{w}_i - w_{i-1}\| \\ \implies \|\hat{w}_i - w_{i-1}\| \leqslant \frac{2L_f}{\lambda_i} = D_i,\end{aligned}$$

by definition of $\hat{w}_i$ and L_f-Lipschitz continuity of $f(\cdot, x_j)$ for all j. Then by Lemma 7, we have

$$\mathbb{E}\|w_i - \hat{w}_i\|^2 \leqslant \exp\left(-\frac{\lambda_i \eta_i T_i}{2}\right)\|w_{i-1} - \hat{w}_i\|^2 + \frac{8\eta_i}{\lambda_i}\left(\widehat{r}_{n_i}(\mathcal{B}_i)^{(2)} + \lambda_i^2 D_i^2 + d\sigma_i^2\right) + \frac{20}{\lambda_i^2}\left(\frac{\widehat{r}_{n_i}(\mathcal{B}_i)^{(k)}}{(k-1)C_i^{k-1}}\right)^2,$$

conditional on w_{i-1} and the draws of $X \sim \mathcal{D}^n$ and $\mathcal{B}_i \sim X^{n_i}$. Taking expectation over the random sampling yields

$$\mathbb{E}\|w_i - \hat{w}_i\|^2 \leqslant \exp\left(-\frac{\lambda_i \eta_i T_i}{2}\right)\|w_{i-1} - \hat{w}_i\|^2 + \frac{8\eta_i}{\lambda_i}\left(\widetilde{e}_{n_i}^{(2)} + \lambda_i^2 D_i^2 + d\sigma_i^2\right) + \frac{20}{\lambda_i^2}\frac{\widetilde{e}_{n_i}^{(2k)}}{C_i^{2k-2}(k-1)^2},$$

where $d\sigma_i^2 \leqslant \frac{4dC_i^2 T_i}{n_i^2 \epsilon^2}$. Now, since $L_f < \infty$, there exists $p \geqslant 1$ such that $L_f \lesssim n^{p/2}\widetilde{R}_{2k,n}\left(\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{d\ln(n)}}{\epsilon n}\right)^{\frac{k-1}{k}}\right)$. Choosing $T_i = \frac{1}{\lambda_i \eta_i}\ln\left(\frac{D^2\lambda_i}{d\sigma_i^2 \eta_i}\right) \lesssim n_i^p \ln(n)$ and η to be determined later (polynomial in n), we get

$$\begin{aligned}
\mathbb{E}\|w_i - \hat{w}_i\|^2 &\lesssim \frac{\eta_i}{\lambda_i}\left(L_f^2 + d\sigma_i^2\right) + \frac{\widetilde{e}_{n_i}^{(2k)}}{\lambda_i^2 C_i^{2k-2}} \\
&\lesssim \eta_i^2 n_i^p (L_f^2 + d\sigma_i^2) + \frac{\eta_i^2 n_i^{2p}\widetilde{e}_{n_i}^{(2k)}}{C_i^{2k-2}} \\
&\lesssim \left(\frac{\eta^2 n^p}{16^i 2^{ip}}\left(L_f^2 + \frac{dC_i^2 T_i}{\epsilon^2 n_i^2} + \frac{n^p \widetilde{e}_{n_i}^{(2k)}}{C_i^{2k-2}2^{pi}}\right)\right). \qquad (16)
\end{aligned}$$

Note that under Assumption 1, F is L-Lipshitz, where $L = \sup_{w\in\mathcal{W}}\|\nabla F(w)\| \leqslant r$ by Jensen's inequality. Now, following the strategy used in the proof of (Feldman et al., 2020a, Theorem 4.4), we write

$$\mathbb{E}F(w_l) - F(w^*) = \mathbb{E}[F(w_l) - F(\hat{w}_l)] + \sum_{i=1}^{l}\mathbb{E}[F(\hat{w}_i) - F(\hat{w}_{i-1})],$$

where $\hat{w}_0 := w^*$. Using (16), the first term can be bounded as follows:

$$\begin{aligned}
\mathbb{E}[F(w_l) - F(\hat{w}_l)] &\leqslant L\sqrt{\mathbb{E}\|w_l - \hat{w}_l\|^2} \\
&\lesssim L\sqrt{\eta_l^2\left(L_f^2 + \frac{C_l^2 d}{\epsilon^2} + \frac{\widetilde{e}_{n_l}^{(2k)}}{C_l^{2k-2}}\right)} \\
&\lesssim L\left[\frac{\eta}{n^2}\left(L_f + \frac{\sqrt{d}C_l}{\epsilon} + \frac{\widetilde{r}_{2k}^k}{C_l^{k-1}}\right)\right] \\
&\lesssim L\left[\frac{\eta}{n^2}\left(L_f + \widetilde{r}_{2k}\left(\frac{\sqrt{d}}{\epsilon}\right)^{(k-1)/k}\right)\right]
\end{aligned}$$

if we choose $C_l = \widetilde{r}_{2k}\left(\frac{\epsilon}{\sqrt{d}}\right)^{1/k}$. Therefore,

$$\mathbb{E}[F(w_l) - F(\hat{w}_l)] \lesssim \widetilde{R}_{2k,n} D \left(\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{d\ln(n)}}{\epsilon n}\right)^{\frac{k-1}{k}}\right), \tag{17}$$

if we choose

$$\eta \lesssim \frac{\widetilde{R}_{2k,n} D n^2}{L} \min\left(\frac{1}{L_f}, \frac{1}{\widetilde{r}_{2k}}\left(\frac{\epsilon}{\sqrt{d}}\right)^{(k-1)/k}\right)\left(\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{d\ln(n)}}{\epsilon n}\right)^{\frac{k-1}{k}}\right) =: \eta_A.$$

Next, Proposition 10 implies

$$\mathbb{E}[F(\hat{w}_i) - F(\hat{w}_{i-1})] \leqslant \frac{2r^2}{\lambda_i n_i} + \frac{\lambda_i}{2}\mathbb{E}\|\hat{w}_{i-1} - w_{i-1}\|^2$$

for all $i \in [l]$. Hence

$$\begin{aligned}
\sum_{i=1}^{l} \mathbb{E}[F(\hat{w}_i) - F(\hat{w}_{i-1})] &\lesssim \frac{r^2}{\lambda_1 n_1} + \lambda_1 D^2 + \sum_{i=2}^{l}\left[\frac{r^2}{\lambda_i n_i} + \lambda_i \eta_i^2 \left(n_i^p (L_f^2 + d\sigma_i^2) + \frac{n_i^{2p}\widetilde{e}_{n_i}^{(2k)}}{C_i^{2k-2}}\right)\right] \\
&\lesssim r^2 \eta n^{p-1} + \frac{D^2}{\eta n^p} + \sum_{i=2}^{l} r^2 \eta_i n_i^{p-1} + \sum_{i=2}^{l} \frac{\eta_i}{n_i^p}\left(n_i^p (L_f^2 + \frac{dC_i^2 T_i}{\epsilon^2 n_i^2}) + \frac{n_i^{2p}\widetilde{e}_{n_i}^{(2k)}}{C_i^{2k-2}}\right) \\
&\lesssim r^2 \eta n^{p-1} + \frac{D^2}{\eta n^p} + \sum_{i=2}^{l} \eta_i \left(L_f^2 + \frac{dC_i^2 n_i^p \ln(n)}{\epsilon^2 n_i^2} + \frac{n_i^p \widetilde{e}_{n_i}^{(2k)}}{C_i^{2k-2}}\right).
\end{aligned}$$

Choosing $C_i = \widetilde{r}_{2k,n_i}\left(\frac{\epsilon n_i}{\sqrt{d\ln(n)}}\right)^{1/k}$ approximately equalizes the two terms above involving C_i and we get

$$\begin{aligned}
\sum_{i=1}^{l} \mathbb{E}[F(\hat{w}_i) - F(\hat{w}_{i-1})] &\lesssim r^2 \eta n^{p-1} + \frac{D^2}{\eta n^p} + \eta L_f^2 + \eta \sum_{i=2}^{l} 4^{-i} n_i^p \widetilde{r}_{2k,n_i}^2 \left(\frac{d\ln(n)}{\epsilon^2 n_i^2}\right)^{\frac{k-1}{k}} \\
&\lesssim \eta\left[r^2 n^{p-1} + L_f^2 + \widetilde{R}_{2k,n}^2 n^p \left(\frac{d\ln(n)}{\epsilon^2 n^2}\right)^{\frac{k-1}{k}}\right] + \frac{D^2}{\eta n^p}.
\end{aligned}$$

Now, choosing

$$\eta = \min\left(\eta_A, \frac{D}{n^{p/2}} \min\left\{\frac{1}{r n^{(p-1)/2}}, \frac{1}{L_f}, \frac{1}{\widetilde{R}_{2k,n} n^{p/2}}\left(\frac{\epsilon n}{\sqrt{d\ln(n)}}\right)^{(k-1)/k}\right\}\right)$$

yields

$$\begin{aligned}
\sum_{i=1}^{l} \mathbb{E}[F(\hat{w}_i) - F(\hat{w}_{i-1})] &\lesssim \widetilde{R}_{2k,n} D \left(\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{d\ln(n)}}{\epsilon n}\right)^{\frac{k-1}{k}}\right) + \frac{L_f D}{n^{p/2}} + \frac{D^2}{\eta_A n^p} \\
&\lesssim \widetilde{R}_{2k,n} D \left(\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{d\ln(n)}}{\epsilon n}\right)^{\frac{k-1}{k}}\right),
\end{aligned}$$

since $L_f \lesssim n^{p/2}\widetilde{R}_{2k,n}\left(\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{d\ln(n)}}{\epsilon n}\right)^{\frac{k-1}{k}}\right)$. Combining the above pieces completes the excess risk proof.

Subgradient complexity: Our choice of $T_i = \widetilde{\Theta}\left(\frac{1}{\lambda_i \eta_i}\right) \lesssim n_i^p \ln(n)$ implies that Algorithm 3 uses $\sum_{i=1}^{l} n_i T_i \lesssim \ln(n) n^{p+1}$ subgradient evaluations. Further, $n^p \leqslant n\left(\frac{L_f}{\widetilde{R}_{2k,n}}\right)^2$. ■

Remark 32 (Details of Remark 11) *If one desires (ϵ, δ)-DP or (ϵ, δ)-SDP instead of zCDP, then the gradient complexity of Algorithm 3 can be improved to $\mathcal{O}(n^{p+\frac{1}{2}}\ln(n))$ by using Clipped Noisy Stochastic Subgradient Method instead of Algorithm 2 as the subroutine in line 7 of Algorithm 3. Choosing batch sizes $m_i \approx \sqrt{n_i} < n_i$ in this subroutine (and increasing σ_i^2 by a factor of $\mathcal{O}(\log(1/\delta))$) ensures (ϵ, δ)-DP by (Abadi et al., 2016, Theorem 1) via privacy amplification by subsampling. The same excess risk bounds hold for any minibatch size $m_i \in [n_i]$, as the proof of Theorem 6 shows.*

E.3. The Strongly Convex Case (Section 3.2)

Our algorithm is an instantiation of the meta-algorithm described in (Feldman et al., 2020a): Initialize $w_0 \in \mathcal{W}$. For $j \in [M] := \lceil \log_2(\log_2(n)) \rceil$, let $N_j = 2^{j-2}n/\log_2(n)$, $\mathcal{C}_j = \left\{\sum_{h<j} N_h + 1, \ldots, \sum_{h \leqslant j} N_h\right\}$, and let w_j be the output of Algorithm 3 run with input data $X_j = (x_s)_{s \in \mathcal{C}_j}$ initialized at w_{j-1}. Output w_M. Assume without loss of generality that $N_j = 2^p$ for some $p \in \mathbb{N}$. Then, with the notation of Appendix E.1, we have the following guarantees:

Theorem 33 (Precise Statement of Theorem 12) *Grant Assumption 1. Let $\epsilon \leqslant \sqrt{d}$ and $f(\cdot, x)$ be μ-strongly convex and L_f-Lipschitz for all $x \in \mathcal{X}$, with $L_f < \infty$. Then, there is a polynomial-time $\frac{\epsilon^2}{2}$-zCDP algorithm $\mathcal{A}$ based on Algorithm 3 with excess risk*

$$\mathbb{E}F(\mathcal{A}(X)) - F^* \lesssim \frac{\widetilde{R}^2_{2k,n/4}}{\mu}\left(\frac{1}{n} + \left(\frac{\sqrt{d\ln(n)}}{\epsilon n}\right)^{\frac{2k-2}{k}}\right).$$

Proof **Privacy**: Since the batches X_j used in each phase of the algorithm are disjoint and Algorithm 3 is $\frac{\epsilon^2}{2}$-zCDP, privacy of the algorithm follows from parallel composition of DP (McSherry, 2009).

Excess risk: Note that N_j samples are used in phase j of the algorithm. For $j \geqslant 0$, let $D_j^2 = \mathbb{E}[\|w_j - w^*\|^2]$ and $\Delta_j = \mathbb{E}[F(w_j) - F^*]$. By strong convexity, we have $D_j^2 \leqslant \frac{2\Delta_j}{\mu}$. Also,

$$\begin{aligned} \Delta_{j+1} &\leqslant a\widetilde{R}_{2k,N_j} D_j \left(\frac{1}{\sqrt{N_j}} + \left(\frac{\sqrt{d\ln(N_j)}}{\epsilon N_j}\right)^{\frac{k-1}{k}}\right) \\ &\leqslant a\widetilde{R}_{2k,N_j}\sqrt{\frac{2\Delta_j}{\mu}}\left(\frac{1}{\sqrt{N_j}} + \left(\frac{\sqrt{d\ln(N_j)}}{\epsilon N_j}\right)^{\frac{k-1}{k}}\right) \qquad (18) \end{aligned}$$

for an absolute constant $a \geqslant 1$, by Theorem 6. Denote $E_j = \left[a\widetilde{R}_{2k,N_j}\sqrt{\frac{2}{\mu}}\left(\frac{1}{\sqrt{N_j}} + \left(\frac{\sqrt{d\ln(N_j)}}{\epsilon N_j}\right)^{\frac{k-1}{k}}\right)\right]^2$.
Then since $N_j = 2N_{j+1}$, we have

$$\frac{E_j}{E_{j+1}} \leqslant 4\left(\frac{\widetilde{R}_{2k,N_j}}{\widetilde{R}_{2k,N_{j+1}}}\right)^2 \leqslant 8, \tag{19}$$

where the second inequality holds because for any $m = 2^q$, we have:

$$\widetilde{R}^2_{2k,m/2} = \sum_{i=1}^{\log_2(m)-1} 2^{-i}\widetilde{r}^2_{2k,2^{-(i+1)}m} = \sum_{i=2}^{\log_2(m)} 2^{-(i-1)}\widetilde{r}^2_{2k,2^{-i}m} = 2\sum_{i=2}^{\log_2(m)} 2^{-i}\widetilde{r}^2_{2k,2^{-i}m} \leqslant 2\widetilde{R}^2_{2k,m}.$$

Now, (19) implies that (18) can be re-arranged as

$$\frac{\Delta_{j+1}}{64E_{j+1}} \leqslant \sqrt{\frac{\Delta_j}{64E_j}} \leqslant \left(\frac{\Delta_0}{64E_0}\right)^{1/2^{j+1}}. \tag{20}$$

Further, if $M \geqslant \log\log\left(\frac{\Delta_0}{E_0}\right)$, then

$$\frac{\Delta_M}{64E_M} \leqslant \left(\frac{\Delta_0}{64E_0}\right)^{1/2^M} \leqslant \left(\frac{\Delta_0}{64E_0}\right)^{1/\log(\Delta_0/E_0)} \leqslant 2^A\left(\frac{1}{64}\right)^{1/\log(\Delta_0/E_0)} \leqslant 2^A,$$

for an absolute constant $A > 0$, since $\Delta_0 \leqslant \frac{2L^2}{\mu}$ and $E_0 \geqslant \frac{2L^2}{\mu n}$ implies $\Delta_0/E_0 = \frac{n}{a^2} \leqslant n$ and $\frac{1}{\log(\Delta_0/E_0)} = \frac{1}{\log(n) - 2\log(a)} \leqslant \frac{A}{\log(n)}$ for some $A > 0$, so that $\left(\frac{\Delta_0}{E_0}\right)^{1/\log(\Delta_0/E_0)} \leqslant n^{A/\log(n)} \leqslant 2^A$.
Therefore,

$$\Delta_M \leqslant 2^A 64E_M = \mathcal{O}\left(\frac{\widetilde{R}^2_{2k,n/4}}{\mu}\left(\frac{1}{n} + \left(\frac{\sqrt{d\ln(n)}}{\epsilon n}\right)^{\frac{2k-2}{k}}\right)\right),$$

since $N_M = n/4$. ∎

E.4. Asymptotic Upper Bounds Under Assumptions 2 and 3

We first recall the notion of *subexponential* distribution:

Definition 34 (Subexponential Distribution) *A random variable Y is subexponential if there is an absolute constant $s > 0$ such that $\mathbb{P}(|Y| \geqslant t) \leqslant 2\exp\left(-\frac{t}{s}\right)$ for all $t \geqslant 0$. For subexponential Y, we define $\|Y\|_{\psi_1} := \inf\left\{s > 0 : \mathbb{P}(|Y| \geqslant t) \leqslant 2\exp\left(-\frac{t}{s}\right)\ \forall\ t \geqslant 0\right\}$.*

Essentially all (heavy-tailed) distributions that arise in practice are subexponential (Mckay, 2019).

Now, we establish asymptotically optimal upper bounds for a broad subclass of the problem class considered in (Wang et al., 2020; Kamath et al., 2022): namely, *subexponential* stochastic subgradient distributions satisfying Assumption 2 or Assumption 3. In Theorem 35 below (which uses the notation of Appendix E.1), we give upper bounds under Assumption 2:

Theorem 35 *Let $f(\cdot, x)$ be convex. Assume $\widetilde{r}_{2k} < \infty$ and $Y_i = \|\nabla f(w, x_i)\|^{2k}$ is subexponential with $E_n \geqslant \max_{i \in [n]} (\|Y_i\|_{\psi_1})$ $\forall w \in \mathcal{W}$, $\nabla f(w, x_i) \in \partial_w f(w, x_i)$. Assume that for sufficiently large n, we have $\sup_{w,x} \|\nabla f(w, x)\|^{2k} \leqslant n^q r^{(2k)}$ for some $q \geqslant 1$ and $\max\left(\frac{E_n}{r^{(2k)}}, \frac{E_n^2}{(r^{(2k)})^2}\right) \ln\left(\frac{3nD\beta}{4r_{2k}}\right) \leqslant \frac{n}{dq}$, where $\|\nabla f(w, x) - \nabla f(w', x)\| \leqslant \beta \|w - w'\|$ for all $w, w' \in \mathcal{W}, x \in \mathcal{X}$, and subgradients $\nabla f(w, x) \in \partial_w f(w, x)$. Then, $\lim_{n \to \infty} \widetilde{R}_{2k,n} \leqslant 4r_{2k}$. Further, there exists $N \in \mathbb{N}$ such that for all $n \geqslant N$, the output of Algorithm 3 satisfies*

$$\mathbb{E}F(w_l) - F^* = \mathcal{O}\left(r_{2k} D \left(\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{d \ln(n)}}{\epsilon n}\right)^{\frac{k-1}{k}}\right)\right).$$

If $f(\cdot, x)$ is μ-strongly convex, then the output of algorithm $\mathcal{A}$ (in Section 3.2) satisfies

$$\mathbb{E}F(\mathcal{A}(X)) - F^* = \mathcal{O}\left(\frac{r_{2k}^2}{\mu}\left(\frac{1}{n} + \left(\frac{\sqrt{d \ln(n)}}{\epsilon n}\right)^{\frac{2k-2}{k}}\right)\right).$$

While a bound on $\sup_{w,x} \|\nabla f(w, x)\|$ is needed in Theorem 35, it can grow as fast as any polynomial in n and only needs to hold for sufficiently large n. As $n \to \infty$, this assumption is easily satisfied. Likewise, Theorem 35 depends only logarithmically on the Lipschitz parameter of the subgradients β, so the result still holds up to constant factors if, say, $\beta \leqslant n^p (r/D)$ as $n \to \infty$ for some $p \geqslant 1$. Crucially, our excess risk bounds do not depend on L_f or β.

Asymptotically optimal upper bounds for Assumption 3 are an immediate consequence of Lemma 20 combined with Theorem 35. Namely, under Assumption 3, the upper bounds in Theorem 35 hold with r replaced by $\sqrt{d}\gamma^{1/k}$ (by Lemma 20). These upper bounds, and the ones in Theorem 35, are tight up to logarithms for their respective problem classes, by the lower bounds in Appendix F.

Proof [Proof of Theorem 35] **Step One:** *There exists $N \in \mathbb{N}$ such that $\widetilde{r}_{2k,n}^2 \leqslant 16 r_{2k}^2$ for all $n \geqslant N$.*

We will first use a covering argument to show that $\widehat{r}_n(X)^{(2k)}$ is upper bounded by $2^{2k+1} r^{(2k)}$ with high probability. For any $\alpha > 0$, we may choose an α-net with $N_\alpha \leqslant \left(\frac{3D}{2\alpha}\right)^d$ balls centered around points in $\mathcal{W}_\alpha = \{w_1, w_2, \cdots, w_{N_\alpha}\} \subset \mathcal{W}$ such that for any $w \in \mathcal{W}$ there exists $i \in [N_\alpha]$ with $\|w - w_i\| \leqslant \alpha$ (see e.g. (Kolmogorov and Tikhomirov, 1959) for the existence of such $\mathcal{W}_\alpha$). For $w \in \mathcal{W}$, let $\widetilde{w}$ denote the element of $\mathcal{W}_\alpha$ that is closest to w, so that $\|w - \widetilde{w}\| \leqslant \alpha$. Now, for any $X \in \mathcal{X}^n$, we have

$$\begin{aligned}
\widehat{r}_n(X)^{(2k)} &= \sup_w \left\{\frac{1}{n}\sum_{i=1}^n \|\nabla f(w, x_i)) - \nabla f(\widetilde{w}, x_i) + \nabla f(\widetilde{w}, x_i)\|^{2k}\right\} \\
&\leqslant 2^{2k} \sup_w \left\{\frac{1}{n}\sum_{i=1}^n \|\nabla f(w, x_i) - \nabla f(\widetilde{w}, x_i)\|^{2k} + \|\nabla f(\widetilde{w}, x_i)\|^{2k}\right\} \\
&\leqslant 2^{2k} \left[\beta^{2k}\alpha^{2k} + \frac{1}{n}\max_{j \in [N_\alpha]} \sum_{i=1}^n \|\nabla f(w_j, x_i)\|^{2k}\right],
\end{aligned}$$

where we used Cauchy-Schwartz and Young's inequality for the first inequality, and the assumption of β-Lipschitz subgradients plus the definition of $\mathcal{W}_\alpha$ for the second inequality. Further,

$$\mathbb{P}\left(\frac{2^{2k}}{n}\max_{j\in[N_\alpha]}\sum_{i=1}^{n}\|\nabla f(w_j,x_i)\|^{2k}\geqslant 2^{2k+1}r^{(2k)}\right)\leqslant N_\alpha\max_{j\in[N_\alpha]}\mathbb{P}\left(\sum_{i=1}^{n}\|\nabla f(w_j,x_i)\|^{2k}\geqslant 2^{2k+1}r^{(2k)}\right)$$
$$\leqslant N_\alpha\exp\left(-n\min\left(\frac{r^{(2k)}}{E_n},\frac{(r^{(2k)})^2}{E_n^2}\right)\right),$$

by a union bound and Bernstein's inequality (see e.g. (Vershynin, 2018, Corollary 2.8.3)). Choosing $\alpha=\frac{2r_{2k}}{\beta}$ ensures that $\mathbb{P}(2^{2k}\beta^{2k}\alpha^{2k}>2^{2k+1}r^{(2k)})=0$ and hence (by union bound)

$$\mathbb{P}\left(\widehat{r}_n(X)^{(2k)}\geqslant 2^{2k+1}r^{(2k)}\right)\leqslant N_\alpha\exp\left(-n\min\left(\frac{r^{(2k)}}{E_n},\frac{(r^{(2k)})^2}{E_n^2}\right)\right)$$
$$\leqslant\left(\frac{3D\beta}{4r_{2k}}\right)^d\exp\left(-n\min\left(\frac{r^{(2k)}}{E_n},\frac{(r^{(2k)})^2}{E_n^2}\right)\right)$$
$$\leqslant\frac{1}{n^q},$$

by the assumption on n. Next, we use this concentration inequality to derive a bound on $\widetilde{e}_n^{(2k)}$:

$$\widetilde{e}_n^{(2k)}=\mathbb{E}\left[\widehat{r}_n(X)^{(2k)}\right]\leqslant\mathbb{E}\left[\widehat{r}_n(X)^{(2k)}|\widehat{r}_n(X)^{(2k)}\geqslant 2^{2k+1}r^{(2k)}\right]\frac{1}{n^q}+2^{2k+1}r^{(2k)}$$
$$\leqslant\frac{\sup_{w,x}\|\nabla f(w,x)\|^{2k}}{n^q}+2^{2k+1}r^{(2k)}$$
$$\leqslant(1+2^{2k+1})r^{(2k)},$$

for sufficiently large n. Thus, $\widetilde{r}^2_{2k,n}\leqslant 16r^2_{2k}$ for all sufficiently large n. This establishes Step One.

Step Two: $\lim_{n\to\infty}\widetilde{R}_{2k,n}\leqslant 4r_{2k}$.
For all $n=2^l, l,i\in\mathbb{N}$, define $h_n(i)=2^{-i}\widetilde{r}^2_{2k,2^{-i}n}\mathbb{1}_{\{i\in[\log_2(n)]\}}$. Note that $0\leqslant h_n(i)\leqslant g(i):=2^{-i}\widetilde{r}^2_{2k}$ for all n,i, and that $\sum_{i=1}^{\infty}g(i)=\widetilde{r}^2_{2k}<\infty$ (i.e. g is integrable with respect to the counting measure). Furthermore, the limit $\lim_{n\to\infty}h_n(i)=2^{-i}\lim_{n\to\infty}\widetilde{r}^2_{2k,2^{-i}n}$ exists since Lemma 25 implies that the sequence $\{\widetilde{r}^2_{2k,2^{-i}n}\}_{n=1}^{\infty}$ is monotonic and bounded for every

$i \in \mathbb{N}$. Thus, by Lebesgue's dominated convergence theorem, we have

$$\begin{aligned}\lim_{n\to\infty} \widetilde{R}_{2k,n}^2 &= \lim_{n\to\infty} \sum_{i=1}^{\infty} h_n(i) \\ &= \sum_{i=1}^{\infty} \lim_{n\to\infty} h_n(i) \\ &\leqslant \sum_{i=1}^{\infty} 2^{-i} \lim_{n\to\infty} \widetilde{r}_{2k,2^{-i}n}^2 \\ &\leqslant 16 \sum_{i=1}^{\infty} 2^{-i} r_{2k}^2 \\ &= 16 r_{2k}^2,\end{aligned}$$

where the last inequality follows from **Step One**. Therefore, $\lim_{n\to\infty} \widetilde{R}_{2k,n} \leqslant 4r_{2k}$. By Theorem 6 and Theorem 12, this also implies the last two claims in Theorem 35. ∎

Appendix F. Lower Bounds (Section 3.3)

In this section, we prove the lower bounds stated in Section 3.3, and also provide tight lower bounds under Assumptions 2 and 3.

Theorem 36 (Precise Statement of Theorem 13) *Let $k \geqslant 2$, $D, \gamma, r^{(k)}, \widetilde{r}^{(k)} > 0$, $\beta_f \geqslant 0$, $d \geqslant 40, n > 7202$, and $\rho \leqslant d$. Then, for any ρ-zCDP algorithm $\mathcal{A}$, there exist $\mathcal{W}, \mathcal{X} \subset \mathbb{R}^d$ such that $\|w - w'\| \leqslant 2D$ for all $w, w' \in \mathcal{W}$, a β_f-smooth, linear, convex (in w) loss $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, and distributions $\mathcal{D}$ and $\mathcal{D}'$ on $\mathcal{X}$ such that:*
1. Assumption 1 holds and if $X' \sim \mathcal{D}'^n$, then

$$\mathbb{E}F(\mathcal{A}(X')) - F^* = \Omega\left(\widetilde{r}_k D \left(\frac{1}{\sqrt{n}} + \min\left\{1, \left(\frac{\sqrt{d}}{\sqrt{\rho n}}\right)^{\frac{k-1}{k}}\right\}\right)\right). \tag{21}$$

2. Assumption 2 holds and if $X' \sim \mathcal{D}'^n$, then

$$\mathbb{E}F(\mathcal{A}(X')) - F^* = \Omega\left(r_k D \left(\frac{1}{\sqrt{n}} + \min\left\{1, \left(\frac{\sqrt{d}}{\sqrt{\rho n}}\right)^{\frac{k-1}{k}}\right\}\right)\right). \tag{22}$$

3. Assumption 3 holds and if $X \sim \mathcal{D}^n$, then

$$\mathbb{E}F(\mathcal{A}(X)) - F^* = \Omega\left(\gamma^{1/k} D \left(\sqrt{\frac{d}{n}} + \sqrt{d}\min\left\{1, \left(\frac{\sqrt{d}}{\sqrt{\rho n}}\right)^{\frac{k-1}{k}}\right\}\right)\right).$$

Proof We will prove part 3 first.
3. We begin by proving the result for $\gamma = D = 1$. In this case, it is proved in (Kamath et al.,

2022) that

$$\mathbb{E}F(\mathcal{A}(X)) - F^* = \Omega\left(\sqrt{d}\min\left\{1, \left(\frac{\sqrt{d}}{\sqrt{\rho}n}\right)^{\frac{k-1}{k}}\right\}\right)$$

for $f(w,x) = -\langle w, x\rangle$ with $\mathcal{W} = B_2^d(0,1)$ and $\mathcal{X} = \{\pm 1\}^d$, and a distribution satisfying Assumption 3 with $\gamma = 1$. Then $f(\cdot, x)$ is linear, convex, and β-smooth for all $\beta \geqslant 0$. We prove the first (non-private) term in the lower bound. By the Gilbert-Varshamov bound (see e.g. (Acharya et al., 2021, Lemma 6)) and the assumption $d \geqslant 40$, there exists a set $\mathcal{V} \subseteq \{\pm 1\}^d$ with $|\mathcal{V}| \geqslant 2^{d/20}$, $d_{\text{Ham}}(\nu, \nu') \geqslant \frac{d}{8}$ for all $\nu, \nu' \in \mathcal{V}$, $\nu \neq \nu'$. For $\nu \in \mathcal{V}$, define the product distribution $Q_\nu = (Q_{\nu_1}, \cdots Q_{\nu_d})$, where for all $j \in [d]$,

$$Q_{\nu_j} = \begin{cases} 1 & \text{with probability } \frac{1+\delta_{\nu_j}}{2} \\ -1 & \text{with probability } \frac{1-\delta_{\nu_j}}{2} \end{cases}$$

for $\delta_{\nu_j} \in (0,1)$ to be chosen later. Then $\mathbb{E}Q_{\nu_j} := \mu_{\nu_j} = \delta_{\nu_j}$ and for any $w \in \mathcal{W}$, $x \sim Q_\nu$, we have

$$\mathbb{E}|\langle \nabla f(w,x) - \nabla F(w), e_j\rangle|^k = \mathbb{E}|\langle -x + \mathbb{E}x, e_j\rangle|^k \tag{23}$$
$$= \mathbb{E}|x - \mu_{\nu_j}|^k \tag{24}$$
$$= \frac{1+\delta_{\nu_j}}{2}|1 - \delta_{\nu_j}|^k + \frac{1-\delta_{\nu_j}}{2}|1 + \delta_{\nu_j}|^k \tag{25}$$
$$\leqslant 1 - \delta_{\nu_j}^2 \leqslant 1 \tag{26}$$

for $\delta_{\nu_j} \in (0,1)$. Now, let $p := \sqrt{d/n}$ and $\delta_{\nu_j} := \frac{p\nu_j}{\sqrt{d}}$. Note that $\mathbb{E}Q_\nu := \mu_\nu = \frac{p\nu}{\sqrt{d}}$ and $w_\nu := \frac{\mu_\nu}{\|\mu_\nu\|} = \frac{\nu}{\|\nu\|}$. Also, $\|\mu_\nu\| = p := \|\mu\|$ for all $\nu \in \mathcal{V}$. Now, denoting $F_{Q_\nu}(w) := \mathbb{E}_{x\sim Q_\nu} f(w,x)$, we have for any $w \in \mathcal{W}$ (possibly depending on $X \sim Q_\nu^n$) that

$$\max_{\nu\in\mathcal{V}} \mathbb{E}\left[F_{Q_\nu}(w) - \min_{w'\in\mathcal{W}} F_{Q_\nu}(w')\right] = \max_{\nu\in\mathcal{V}} \mathbb{E}\left[\left\langle \frac{\mu_\nu}{\|\mu\|}, \mu_\nu\right\rangle - \langle w, \mu_\nu\rangle\right] \tag{27}$$
$$= \max_{\nu\in\mathcal{V}} \mathbb{E}\left[\|\mu\| - \langle w, \mu_\nu\rangle\right] \tag{28}$$
$$= \max_{\nu\in\mathcal{V}} \mathbb{E}\left(\|\mu\|[1 - \langle w, w_\nu\rangle]\right) \tag{29}$$
$$\geqslant \max_{\nu\in\mathcal{V}} \mathbb{E}\left[\frac{1}{2}\|\mu\|\|w - w_\nu\|^2\right], \tag{30}$$

since $\|w\|, \|w_\nu\| \leqslant 1$. Further, denoting $\hat{w} := \text{argmin}_{\nu\in\mathcal{V}} \|w_\nu - w\|$, we have $\|\hat{w} - w_\nu\|^2 \leqslant 4\|w_\nu - w\|^2$ for all $\nu \in \mathcal{V}$ (via Young's inequality). Hence

$$\max_{\nu\in\mathcal{V}} \mathbb{E}\left[F_{Q_\nu}(w) - \min_{w'\in\mathcal{W}} F_{Q_\nu}(w')\right] \geqslant \frac{\|\mu\|}{8} \max_{\nu\in\mathcal{V}} \mathbb{E}\|\hat{w} - w_\nu\|^2. \tag{31}$$

Now we apply Fano's method (see e.g. (Yu, 1997, Lemma 3)) to lower bound $\max_{\nu\in\mathcal{V}} \mathbb{E}\|\hat{w} - w_\nu\|^2$. For all $\nu \neq \nu'$, we have $\|w_\nu - w_{\nu'}\|^2 \geqslant \frac{\|\nu-\nu'\|^2}{\|\nu\|^2} \geqslant 1$ since $d_{\text{Ham}}(\nu, \nu') \geqslant \frac{d}{2}$ and $\nu \in \{\pm 1\}^d$

implies $\|\nu - \nu'\|^2 \geqslant \frac{d}{2}$ and $\|\nu\|^2 = d$. Also, a straightforward computation shows that for any $j \in [d]$ and $\nu, \nu' \in \mathcal{V}$,

$$D_{KL}(Q_{\nu_j} || Q_{\nu_j'}) \leqslant \frac{1 + \frac{p}{\sqrt{d}}}{2}\left[\log\left(\frac{\sqrt{d}+p}{\sqrt{d}}\right) + \log\left(\frac{\sqrt{d}}{\sqrt{d}-p}\right)\right] \tag{32}$$

$$\leqslant \log\left(\frac{1+\frac{p}{\sqrt{d}}}{1-\frac{p}{\sqrt{d}}}\right) \tag{33}$$

$$\leqslant \frac{3p}{\sqrt{d}}, \tag{34}$$

for our choice of p, provided $\frac{p}{\sqrt{d}} = \frac{1}{\sqrt{n}} \in (0, \frac{1}{2})$, which holds if $n > 4$. Hence by the chain rule for KL-divergence,

$$D_{KL}(Q_\nu || Q_{\nu'}) \leqslant 3p\sqrt{d} = 3\frac{d}{\sqrt{n}}$$

for all $\nu, \nu' \in \mathcal{V}$. Thus, for any $w \in \mathcal{W}$, Fano's method yields

$$\max_{\nu \in \mathcal{V}} \mathbb{E}\|w - w_\nu\|^2 \geqslant \frac{1}{2}\left(1 - \frac{3p\sqrt{d} + \log(2)}{(d/20)}\right) = \frac{1}{2}\left(1 - \frac{60\frac{d}{\sqrt{n}} - 20\log(2)}{d}\right),$$

which is $\Omega(1)$ for $d \geqslant 40 > 20\log(2)$ and $n > 7202$. Combining this with (31) and plugging in $\|\mu\| = \sqrt{\frac{d}{n}}$ shows that

$$\mathbb{E}F_{Q_\nu}(\mathcal{A}(X)) - F^*_{Q_\nu} = \Omega\left(\sqrt{\frac{d}{n}}\right)$$

for some $\nu \in \mathcal{V}$ (for any algorithm $\mathcal{A}$), where $X \sim Q_\nu^n$.

Next, we scale our hard instance for arbitrary $\gamma, D > 0$. First, we scale the distribution $Q_\nu \to \tilde{Q}_\nu = \gamma^{1/k} Q_\nu$, which is supported on $\tilde{\mathcal{X}} = \{\pm\gamma^{1/k}\}^d$. Denote its mean by $\mathbb{E}Q_\nu := \widetilde{\mu}_\nu = \gamma^{1/k}\mu_\nu$. Also we scale $\mathcal{W} \to \widetilde{\mathcal{W}} = D\mathcal{W} = B_2^d(0, D)$. So our final (linear, convex, smooth) hard instance is $f : \widetilde{\mathcal{W}} \times \widetilde{\mathcal{X}} \to \mathbb{R}$, $f(\widetilde{w}, \widetilde{x}) = -\langle \widetilde{w}, \widetilde{x}\rangle$, $\widetilde{F}(\widetilde{w}) := \mathbb{E}_{\widetilde{x} \sim \tilde{Q}_\nu} f(\widetilde{w}, \widetilde{x})$. Denote $F(w) := \mathbb{E}_{x \sim Q_\nu} f(w, x)$. Note that

$$\begin{aligned}
\mathbb{E}|\langle \nabla f(\widetilde{w}, \widetilde{x}) - \nabla \widetilde{F}(\widetilde{w}), e_j\rangle|^k &= \mathbb{E}|\langle -\widetilde{x} + \mathbb{E}\widetilde{x}, e_j\rangle|^k \\
&= \mathbb{E}|\widetilde{x} - \widetilde{\mu}_{\nu_j}|^k \\
&= \mathbb{E}|\gamma^{1/k}(x - \mu_{\nu_j})|^k \leqslant \gamma.
\end{aligned}$$

Further, we have $w^* := \operatorname{argmin}_{w \in \mathcal{W}} F(w) = \frac{\mu_\nu}{\|\mu_\nu\|}$ and $\widetilde{w}^* := \operatorname{argmin}_{\widetilde{w} \in \widetilde{\mathcal{W}}} \widetilde{F}(\widetilde{w}) = Dw^*$. Therefore, for any $w \in \mathcal{W}, \widetilde{w} = Dw \in \widetilde{\mathcal{W}}$, we have

$$\widetilde{F}(\widetilde{w}) - \widetilde{F}(\widetilde{w}^*) = -\langle \widetilde{w}, \widetilde{\mu}_\nu\rangle + \langle \widetilde{w}^*, \widetilde{\mu}_\nu\rangle \tag{35}$$

$$= \langle D(w^* - w), \gamma^{1/k}\mu_\nu\rangle \tag{36}$$

$$= D\gamma^{1/k}[F(w) - F(w^*)]. \tag{37}$$

Thus,

$$\mathbb{E}\widetilde{F}(\mathcal{A}(\widetilde{X})) - \widetilde{F}^* = \gamma^{1/k} D[\mathbb{E}F(\mathcal{A}(X)) - F^*],$$

so applying the lower bound for the case $D = \gamma = 1$ (i.e. for the unscaled F) yields the desired lower bound via $\widetilde{F}$.

1. We will use nearly the same unscaled hard instances used to prove the private and non-private terms of the lower bound in part 3, but the scaling will differ. Starting with the *non-private* term, we scale the distribution $Q_\nu \to \widetilde{Q}_\nu = \frac{\widetilde{r}_k}{\sqrt{d}} Q_\nu$ and $\mathcal{X} \to \widetilde{\mathcal{X}} = \frac{\widetilde{r}_k}{\sqrt{d}} \mathcal{X}$. Also, scale $\mathcal{W} \to \widetilde{\mathcal{W}} = D\mathcal{W} = B_2^d(0, D)$. Let $f(\widetilde{w}, \widetilde{x}) := -\langle \widetilde{w}, \widetilde{x} \rangle$, which satisfies all the hypotheses of the theorem. Also,

$$\mathbb{E}_{\widetilde{x} \sim \widetilde{Q}_\nu} \left[\sup_{\widetilde{w}} \|\nabla f(\widetilde{w}, \widetilde{x})\|^k \right] = \left(\frac{\widetilde{r}_k}{\sqrt{d}} \right)^k \mathbb{E}_{x \sim Q_\nu} \|x\|^k \leqslant \left(\frac{\widetilde{r}_k}{\sqrt{d}} \right)^k d^{k/2} = \widetilde{r}^{(k)}.$$

Now $\widetilde{w}^* = Dw^*$ as before and letting $\widetilde{F}(\cdot) := \mathbb{E}_{\widetilde{x} \sim \widetilde{Q}_\nu} f(\cdot, \widetilde{x})$, we have

$$\widetilde{F}(\widetilde{w}) - \widetilde{F}(\widetilde{w}^*) = \frac{\widetilde{r}_k D}{\sqrt{d}} [F(w) - F^*]$$

for any $\widetilde{w} = Dw$. Thus, applying the unscaled non-private lower bound established above yields a lower bound of $\Omega\left(\frac{\widetilde{r}D}{\sqrt{n}}\right)$ on the non-private excess risk of our scaled instance.

Next, we turn to the scaled *private* lower bound. The unscaled hard distribution Q'_ν given by

$$Q'_\nu = \begin{cases} 0 & \text{with probability } 1 - p \\ p^{-1/k}\nu & \text{with probability } p \end{cases}$$

(with the same linear f and same $\mathcal{W}$) provides the unscaled lower bound

$$\mathbb{E}F(\mathcal{A}(X)) - F^* = \Omega\left(\sqrt{d} \min\left\{ 1, \left(\frac{\sqrt{d}}{\sqrt{\rho} n} \right)^{\frac{k-1}{k}} \right\} \right),$$

by the proof of (Kamath et al., 2022, Theorem 6.4). We scale $Q'_\nu \to \widetilde{Q}'_\nu = \frac{\widetilde{r}}{\sqrt{d}} Q'_\nu$, $\mathcal{X} \to \widetilde{\mathcal{X}} = \frac{\widetilde{r}}{\sqrt{d}} \mathcal{X}$, and $\mathcal{W} \to \widetilde{\mathcal{W}} = D\mathcal{W}$. Then for any $\widetilde{w} \in \widetilde{\mathcal{W}}$,

$$\mathbb{E}_{\widetilde{x} \sim \widetilde{Q'}} \left[\sup_{\widetilde{w}} \|\nabla f(\widetilde{w}, \widetilde{x})\|^k \right] = \left(\frac{\widetilde{r}_k}{\sqrt{d}} \right)^k \mathbb{E}_{x \sim Q'} \|x\|^k = p\|p^{-1/k}\nu\|^k = \widetilde{r}^{(k)}.$$

Moreover, excess risk scales by a factor of $\frac{\widetilde{r}D}{\sqrt{d}}$, as we saw above. Thus, applying the unscaled lower bound completes the proof of part 1.

2. We use an identical construction to that used above in part 1 except that the scaling factor $\widetilde{r}_k$ gets replaced by r_k. It is easy to see that $\mathbb{E}\left[\sup_{w \in \mathcal{W}} \|\nabla f(w, x)\|^k\right] = \sup_{w \in \mathcal{W}} \mathbb{E}\left[\|\nabla f(w, x)\|^k\right]$ for our construction, hence the result follows. ■

Remark 37 *The main differences in our proof of part 3 of Theorem 13 from the proof of (Kamath et al., 2022, Theorem 6.4) (for $\gamma = D = 1$) are: 1) we construct a Bernoulli product distribution (built on (Duchi, 2021, Example 7.7)) instead of a Gaussian, which establishes a lower bound that holds for all $k \geqslant 2$ instead of just $k = \mathcal{O}(1)$; and 2) we choose a different parameter value (larger p in the notation of the proof) in our application of Fano's method, which results in a tighter lower bound: the term $\min\{1, \sqrt{d/n}\}$ in (Kamath et al., 2022, Theorem 6.4) gets replaced with $\sqrt{d/n}$.*[14] *Also, there exist parameter settings for which our lower bound is indeed strictly greater than the lower bound in (Kamath et al., 2022, Theorem 6.4): for instance, if $d > n > d/\rho$ and $k \to \infty$, then our lower bound simplifies to $\Omega(\sqrt{\frac{d}{n}})$. On the other hand, the lower bound in (Kamath et al., 2022, Theorem 6.4) breaks as $k \to \infty$ (since the k-th moment of their Gaussian goes to infinity); however, even if were extended to $k \to \infty$ (e.g. by replacing their Gaussian with our Bernoulli distribution), then the resulting lower bound $\Omega(1 + \frac{d}{\sqrt{\rho n}})$ would still be smaller than the one we prove above.*[15]

Theorem 38 (Precise Statement of Theorem 14) *Let $k \geqslant 2$, $\mu, \gamma, \widetilde{r}_k, r_k > 0$, $n \in \mathbb{N}$, $d \geqslant 40$, and $\rho \leqslant d$. Then, for any ρ-zCDP algorithm $\mathcal{A}$, there exist convex, compact sets $\mathcal{W}, \mathcal{X} \subset \mathbb{R}^d$ of diameter D, a μ-smooth, μ-strongly convex (in w) loss $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, and distributions $\mathcal{D}$ and $\mathcal{D}'$ on $\mathcal{X}$ such that:*

1. Assumption 1 holds with $D \approx \frac{\widetilde{r}_k}{\mu}$, and if $X' \sim \mathcal{D}'^n$, then

$$\mathbb{E}F(\mathcal{A}(X')) - F^* = \Omega\left(\frac{\widetilde{r}_k^2}{\mu}\left(\frac{1}{n} + \min\left\{1, \left(\frac{\sqrt{d}}{\sqrt{\rho} n}\right)^{\frac{2k-2}{k}}\right\}\right)\right).$$

2. Assumption 2 holds with $D \approx \frac{r_k}{\mu}$, and if $X' \sim \mathcal{D}'^n$, then

$$\mathbb{E}F(\mathcal{A}(X')) - F^* = \Omega\left(\frac{r_k^2}{\mu}\left(\frac{1}{n} + \min\left\{1, \left(\frac{\sqrt{d}}{\sqrt{\rho} n}\right)^{\frac{2k-2}{k}}\right\}\right)\right).$$

3. Assumption 3 holds, $D \approx \frac{\gamma^{1/k}\sqrt{d}}{\mu}$, and if $X \sim \mathcal{D}^n$, then

$$\mathbb{E}F(\mathcal{A}(X)) - F^* = \Omega\left(\frac{\gamma^{2/k}}{\mu}\left(\frac{d}{n} + d\min\left\{1, \left(\frac{\sqrt{d}}{\sqrt{\rho} n}\right)^{\frac{2k-2}{k}}\right\}\right)\right).$$

Proof We will prove part 3 first. 3. We first consider $\gamma = \mu = 1$ and then scale our hard instance. For $f(w, x) := \frac{1}{2}\|w - x\|^2$, (Kamath et al., 2022) construct a convex/compact domain $\mathcal{W} \times \mathcal{X} \subset \mathbb{R}^d \times \mathbb{R}^d$ and distribution $\mathcal{D}$ on $\mathcal{X}$ such that

$$\mathbb{E}F(\mathcal{A}(X)) - F^* = \Omega\left(d\min\left\{1, \left(\frac{\sqrt{d}}{\sqrt{\rho} n}\right)^{\frac{2k-2}{k}}\right\}\right)$$

14. Note that (Kamath et al., 2022, Theorem 6.4) writes $\sqrt{d/n}$ for the first term. However, the proof (see Equation 16 in their paper) only establishes the bound $\min\{1, \sqrt{d/n}\}$.

15. By Lemma 20, lower bounds under Assumption 3 imply lower bounds under Assumption 2 with $\gamma^{1/k}$ replaced by $r/\sqrt{d}$. Nevertheless, we provide direct proofs under both assumptions for additional clarity.

for any k and any ρ-zCDP algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$ if $X \sim \mathcal{D}^n$.[16] So, it remains to a) prove the first term (d/n) in the lower bound, and then b) show that the scaled instance satisfies the exact hypotheses in the theorem and has excess loss that scales by a factor of $\gamma^{2/k}/\mu$. We start with task a). Observe that for f defined above and any distribution $\mathcal{D}$ such that $\mathbb{E}\mathcal{D} \in \mathcal{W}$, we have

$$\mathbb{E}F(\mathcal{A}(X)) - F^* = \frac{1}{2}\mathbb{E}\|\mathcal{A}(X) - \mathbb{E}\mathcal{D}\|^2 \tag{38}$$

(see (Kamath et al., 2022, Lemma 6.2)), and

$$\mathbb{E}|\langle \nabla f(w, x) - \nabla F(w), e_j\rangle|^k = \mathbb{E}|\langle x - \mathbb{E}x, e_j\rangle|^k.$$

Thus, it suffices to prove that $\mathbb{E}\|\mathcal{A}(X) - \mathbb{E}\mathcal{D}\|^2 \gtrsim \frac{d}{n}$ for some $\mathcal{D}$ such that $\mathbb{E}|\langle x - \mathbb{E}x, e_j\rangle|^k \leqslant 1$. This is a known result for products of Bernoulli distributions; nevertheless, we provide a detailed proof below. First consider the case $d = 1$. Then the proof follows along the lines of (Duchi, 2021, Example 7.7). Define the following pair of distributions on $\{\pm 1\}$:

$$P_0 := \begin{cases} 1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

and

$$P_1 := \begin{cases} 1 & \text{with probability } \frac{1+\delta}{2} \\ -1 & \text{with probability } \frac{1-\delta}{2} \end{cases}$$

for $\delta \in (0, 1)$ to be chosen later. Notice that if X is a random variable with distribution P_ν ($\nu \in \{0, 1\}$), then $\mathbb{E}|X - \mu|^k \leqslant \mathbb{E}|X|^k \leqslant 1$. Also, $\mathbb{E}P_\nu = \delta\nu$ for $\nu \in \{0, 1\}$ and $|\mathbb{E}P_1 - \mathbb{E}P_0| = \delta$ (i.e. the two distributions are δ-separated with respect to the metric $\rho(a, b) = |a - b|$). Then by LeCam's method (see e.g. (Duchi, 2021, Eq. 7.33) and take $\Phi(\cdot) = (\cdot)^2$),

$$\max_{\nu \in \{0,1\}} \mathbb{E}_{X \sim P_\nu^n}|\mathcal{A}(X) - \mathbb{E}P_\nu|^2 \geqslant \frac{\delta^2}{8}\left[1 - \|P_0^n - P_1^n\|_{TV}\right].$$

Now, by Pinsker's inequality and the chain rule for KL-divergence, we have

$$\|P_0^n - P_1^n\|_{TV}^2 \leqslant \frac{1}{2}D_{KL}(P_0^n||P_1^n) = \frac{n}{2}D_{KL}(P_0||P_1) = \frac{n}{2}\log\left(\frac{1}{1-\delta^2}\right).$$

Choosing $\delta = \frac{1}{\sqrt{2n}} < \frac{1}{\sqrt{2}}$ implies $\|P_0^n - P_1^n\|_{TV}^2 \leqslant n\delta^2 = \frac{1}{2}$. Hence there exists a distribution $\hat{\mathcal{D}} \in \{P_0, P_1\}$ on $\mathbb{R}$ such that

$$\mathbb{E}_{X \sim \hat{\mathcal{D}}^n}|\mathcal{A}(X) - \mathbb{E}\hat{\mathcal{D}}|^2 \geqslant \frac{\delta^2}{8}\left[1 - \frac{1}{\sqrt{2}}\right] \geqslant \frac{1}{64n}$$

For general $d \geqslant 1$, we take the product distribution $\mathcal{D} := \hat{\mathcal{D}}^d$ on $\mathcal{X} = \{\pm 1\}^d$ and choose $\mathcal{W} = B_2^d(0, \sqrt{d})$ to ensure $\mathbb{E}\mathcal{D} \in \mathcal{W}$ (so that (38) holds). Clearly, $\mathbb{E}|\langle \mathcal{D} - \mathbb{E}\mathcal{D}, e_j\rangle|^k \leqslant 1$ for all $j \in [d]$. Further, the mean squared error of any algorithm for estimating the mean of $\mathcal{D}$ is

$$\mathbb{E}_{X \sim \mathcal{D}^n}\|\mathcal{A}(X) - \mathbb{E}\mathcal{D}\|^2 = \sum_{j=1}^d \mathbb{E}|\mathcal{A}(X)_j - \mathbb{E}\mathcal{D}_j|^2 \geqslant \frac{d}{64n}, \tag{39}$$

16. In fact, $\mathcal{W}$ and $\mathcal{X}$ can be chosen to be Euclidean balls of radius $\sqrt{d}p^{-1/k}$ for p defined in the proof of (Kamath et al., 2022, Lemma 6.3), which ensures that $\mathbb{E}\mathcal{D} \in \mathcal{W} = \mathcal{X}$.

by applying the $d = 1$ result to each coordinate.

Next, we move to task b). For this, we re-scale each of our hard distributions (non-private given above, and private given in the proof of (Kamath et al., 2022, Lemma 6.3) and below in our proof of part 2 of the theorem–see (43)): $\mathcal{D} \to \frac{\gamma^{1/k}}{\mu}\mathcal{D} = \widetilde{\mathcal{D}}$, $\mathcal{X} \to \frac{\gamma^{1/k}}{\mu}\mathcal{X} = \widetilde{\mathcal{X}}$, $\mathcal{W} \to \frac{\gamma^{1/k}}{\mu}\mathcal{W} = \widetilde{\mathcal{W}}$ and $f : \mathcal{W} \times \mathcal{X} \to \mu f = \widetilde{f} : \widetilde{\mathcal{W}} \times \widetilde{\mathcal{X}}$. Then $\widetilde{f}(\cdot, \widetilde{x})$ is μ-strongly convex and μ-smooth for all $\widetilde{x} \in \widetilde{\mathcal{X}}$ and

$$\mathbb{E}|\langle \nabla \widetilde{f}(\widetilde{w}, \widetilde{x}) - \nabla \widetilde{F}(\widetilde{w}), e_j\rangle|^k = \mu^k \mathbb{E}|\langle \widetilde{x} - \mathbb{E}\widetilde{x}, e_j\rangle|^k = \mu^k \mathbb{E}\left|\left(\frac{\gamma^{1/k}}{\mu}\right)\langle x - \mathbb{E}x, e_j\rangle\right|^k = \gamma \mathbb{E}|\langle x - \mathbb{E}x, e_j\rangle|^k \leqslant \gamma$$

for any $j \in [d]$, $x \sim \mathcal{D}$, $\widetilde{x} \sim \widetilde{\mathcal{D}}$, $\widetilde{w} \in \widetilde{\mathcal{W}}$. Thus, the scaled hard instance is in the required class of functions/distributions. Further, denote $\widetilde{F}(w) = \mathbb{E}\widetilde{f}(w, x)$, $\widetilde{w}^* := \operatorname{argmin}_{\widetilde{w} \in \widetilde{\mathcal{W}}} \widetilde{F}(\widetilde{w}) = \mathbb{E}\widetilde{\mathcal{D}} = \frac{\gamma^{1/k}}{\mu}\mathbb{E}\mathcal{D}$. Then, for any $w \in \mathcal{W}$, $\widetilde{w} := \frac{\gamma^{1/k}}{\mu} w$, we have:

$$\widetilde{F}(\widetilde{w}) - \widetilde{F}(\widetilde{w}^*) = \frac{\mu}{2}\mathbb{E}\left[\|\widetilde{w} - \widetilde{x}\|^2 - \|\widetilde{w}^* - \widetilde{x}\|^2\right] \tag{40}$$

$$= \frac{\mu}{2}\left(\frac{\gamma^{2/k}}{\mu^2}\right)\mathbb{E}\left[\|w - x\|^2 - \|w^* - x\|^2\right] \tag{41}$$

$$= \frac{\gamma^{2/k}}{\mu}[F(w) - F(w^*)]. \tag{42}$$

In particular, for $w := \mathcal{A}(X)$ and $\widetilde{w} := \frac{\gamma^{1/k}}{\mu}\mathcal{A}(X)$, we get

$$\mathbb{E}_{\mathcal{A}, X \sim \mathcal{D}^n}\left[\widetilde{F}\left(\frac{\gamma^{1/k}}{\mu}\mathcal{A}(X)\right) - \widetilde{F}^*\right] = \frac{\gamma^{2/k}}{\mu}\mathbb{E}_{\mathcal{A}, X \sim \mathcal{D}^n}\left[F(\mathcal{A}(X)) - F^*\right]$$

for any algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$. Writing $\tilde{\mathcal{A}}(\tilde{X}) := \frac{\gamma^{1/k}}{\mu}\mathcal{A}(X)$ and $\tilde{X} := \frac{\gamma^{1/k}}{\mu}X$ for $X \in \mathcal{X}^n$, we conclude

$$\mathbb{E}_{\widetilde{\mathcal{A}}, \tilde{X} \sim \tilde{\mathcal{D}}^n}\left[\widetilde{F}\left(\tilde{\mathcal{A}}(\tilde{X})\right) - \widetilde{F}^*\right] = \frac{\gamma^{2/k}}{\mu}\mathbb{E}_{\mathcal{A}, X \sim \mathcal{D}^n}\left[F(\mathcal{A}(X)) - F^*\right]$$

for any $\widetilde{\mathcal{A}} : \widetilde{\mathcal{X}}^n \to \widetilde{\mathcal{W}}$. Therefore, an application of the unscaled lower bound

$$\mathbb{E}_{\mathcal{A}, X \sim \mathcal{D}^n}\left[F(\mathcal{A}(X)) - F^*\right] = \Omega\left(\frac{d}{n} + d\min\left\{1, \left(\frac{\sqrt{d}}{n\sqrt{\rho}}\right)^{\frac{2k-2}{k}}\right\}\right),$$

which follows by combining part 3a) above with (Kamath et al., 2022, Lemma 6.3), completes the proof of part 3.

1. We begin by proving the first (non-private) term in the lower bound: For our *unscaled* hard instance, we will take the same distribution $\mathcal{D} = P_\nu^d$ (for some $\nu \in \{0, 1\}$) on $\mathcal{X} = \{\pm 1\}^d$

and quadratic f described above in part 1a with $\mathcal{W} := B_2^d(0, \sqrt{d})$. The choice of $\mathcal{W}$ ensures $\mathbb{E}\mathcal{D} \in \mathcal{W}$ so that (38) holds. Further,

$$\mathbb{E}\left[\sup_{w\in\mathcal{W}} \|\nabla f(w,x)\|^k\right] = \mathbb{E}\left[\sup_{w\in\mathcal{W}} \|w-x\|^k\right] \leqslant \mathbb{E}[\|3x\|^k] \leqslant (9d)^{k/2}.$$

Thus, if we scale $f \to \widetilde{f} = \mu f$, $\mathcal{W} \to \widetilde{\mathcal{W}} := \frac{\widetilde{r}_k}{\mu\sqrt{9d}}\mathcal{W}$, $\mathcal{X} \to \widetilde{\mathcal{X}} := \frac{\widetilde{r}_k}{\mu\sqrt{9d}}\mathcal{X}$ and $\mathcal{D} \to \widetilde{\mathcal{D}} = \frac{\widetilde{r}_k}{\mu\sqrt{9d}}\mathcal{D}$, then $\widetilde{f}(\cdot, \widetilde{x})$ is μ-strongly convex and μ-smooth, and

$$\mathbb{E}\left[\sup_{\widetilde{w}\in\widetilde{\mathcal{W}}} \left\|\nabla \widetilde{f}(\widetilde{w},\widetilde{x})\right\|^k\right] = \mathbb{E}\left[\sup_{\widetilde{w}\in\widetilde{\mathcal{W}}} \|\widetilde{w}-\widetilde{x}\|^k\right] = \mu^k \widetilde{r}^{(k)}\left(\frac{1}{\mu\sqrt{9d}}\right)^k \mathbb{E}\left[\sup_{w\in\mathcal{W}} \|\nabla f(w,x)\|^k\right] \leqslant \widetilde{r}^{(k)}.$$

Moreover, if $\left(\frac{\widetilde{r}_k}{3\mu\sqrt{d}}\right)\mathcal{A} = \widetilde{\mathcal{A}} : \widetilde{X}^n \to \widetilde{\mathcal{W}}$ is any algorithm and $\widetilde{X} \sim \widetilde{\mathcal{D}}^n$, then by (39) and (38), we have

$$\mathbb{E}\widetilde{F}(\widetilde{\mathcal{A}}(\widetilde{X})) - \widetilde{F}^* = \frac{\mu}{2}\mathbb{E}\|\widetilde{\mathcal{A}}(\widetilde{X}) - \mathbb{E}\widetilde{\mathcal{D}}\|^2 = \frac{\mu}{2}\left(\frac{\widetilde{r}_k}{\mu\sqrt{9d}}\right)^2 \mathbb{E}\|\mathcal{A}(X) - \mathbb{E}\mathcal{D}\|^2 \gtrsim \frac{\widetilde{r}_k^2}{\mu n}.$$

Next, we prove the second (private) term in the lower bound. Let f be as defined above. For our unscaled hard distribution, we follow (Barber and Duchi, 2014; Kamath et al., 2022) and define a family of distributions $\{Q_\nu\}_{\nu\in\mathcal{V}}$ on $\mathbb{R}^d$, where $\mathcal{V} \subset \{\pm 1\}^d$ will be defined later. For any given $\nu \in \mathcal{V}$, we define the distribution Q_ν as follows: $X_\nu \sim Q_\nu$ iff

$$X_\nu = \begin{cases} 0 & \text{with probability } 1-p \\ p^{-1/k}\nu & \text{with probability } p \end{cases} \tag{43}$$

where $p := \min\left(1, \frac{\sqrt{d}}{n\sqrt{\rho}}\right)$. Now, we select a set $\mathcal{V} \subset \{\pm 1\}^d$ such that $|\mathcal{V}| \geqslant 2^{d/20}$ and $d_{\text{Ham}}(\nu, \nu') \geqslant \frac{d}{8}$ for all $\nu, \nu' \in \mathcal{V}, \nu \neq \nu'$: such $\mathcal{V}$ exists by standard Gilbert-Varshamov bound (see e.g. (Acharya et al., 2021, Lemma 6)). For any $\nu \in \mathcal{V}$, if $x \sim Q_\nu$ and $w \in \mathcal{W} := B_2^d(0, \sqrt{d}p^{-1/k})$, then

$$\mathbb{E}\left[\sup_{w\in\mathcal{W}} \|\nabla f(w,x)\|^k\right] = \mathbb{E}\left[\sup_{w\in\mathcal{W}} \|w-x\|^k\right] \leqslant \mathbb{E}[\|2x\|^k] = 2^k(p\|p^{-1/k}\nu\|^k) = 2^k\|\nu\|^k = 2^k d^{k/2}.$$

Note also that our choice of $\mathcal{W}$ and $p \leqslant 1$ ensures that $\mathbb{E}[Q_\nu] \in \mathcal{W}$. Moreover, as in the proof of (Kamath et al., 2022, Lemma 6.3), zCDP Fano's inequality (see (Kamath et al., 2022, Theorem 1.4)) implies that for any ρ-zCDP algorithm $\mathcal{A}$,

$$\sup_{\nu\in\mathcal{V}} \mathbb{E}_{X\sim Q_\nu^n, \mathcal{A}}\|\mathcal{A}(X) - \mathbb{E}Q_\nu\|^2 = \Omega\left(d\min\left\{1, \left(\frac{\sqrt{d}}{n\sqrt{\rho}}\right)^{\frac{2k-2}{k}}\right\}\right). \tag{44}$$

Thus,

$$\mathbb{E}_{X\sim Q_\nu^n, \mathcal{A}}F(\mathcal{A}(X)) - F^* = \Omega\left(d\min\left\{1, \left(\frac{\sqrt{d}}{n\sqrt{\rho}}\right)^{\frac{2k-2}{k}}\right\}\right)$$

for some $\nu \in \mathcal{V}$, by (38). Now we scale our hard instance: $f \to \widetilde{f} = \mu f$, $\mathcal{W} \to \widetilde{\mathcal{W}} := \frac{\widetilde{r}_k}{2\mu\sqrt{d}}\mathcal{W}$, $\mathcal{X} \to \widetilde{\mathcal{X}} := \frac{\widetilde{r}_k}{2\mu\sqrt{d}}\mathcal{X}$ and $\mathcal{D} \to \widetilde{\mathcal{D}} = \frac{\widetilde{r}_k}{2\mu\sqrt{d}}\mathcal{D}$. Then $\widetilde{f}(\cdot, \widetilde{x})$ is μ-strongly convex and μ-smooth, and

$$\mathbb{E}\left[\sup_{\widetilde{w}\in\widetilde{\mathcal{W}}} \|\nabla \widetilde{f}(\widetilde{w}, \widetilde{x})\|^k\right] = \mathbb{E}\left[\sup_{\widetilde{w}\in\widetilde{\mathcal{W}}} \|\widetilde{w} - \widetilde{x}\|^k\right] = \mu^k \left(\frac{\widetilde{r}_k}{2\mu\sqrt{d}}\right)^k \mathbb{E}\left[\sup_{w\in\mathcal{W}} \|\nabla f(w, x)\|^k\right] \leqslant \widetilde{r}^{(k)}.$$

Moreover, if $\left(\frac{\widetilde{r}_k}{2\mu\sqrt{d}}\right)\mathcal{A} = \widetilde{\mathcal{A}} : \widetilde{X}^n \to \widetilde{\mathcal{W}}$ is any ρ-zCDP algorithm and $\widetilde{X} \sim \widetilde{\mathcal{D}}^n$, then

$$\begin{aligned}
\mathbb{E}\widetilde{F}(\widetilde{\mathcal{A}}(\widetilde{X})) - \widetilde{F}^* &= \frac{\mu}{2}\mathbb{E}\|\widetilde{\mathcal{A}}(\widetilde{X}) - \mathbb{E}\widetilde{\mathcal{D}}\|^2 \\
&= \frac{\mu}{2}\left(\frac{\widetilde{r}_k}{2\mu\sqrt{d}}\right)^2 \mathbb{E}\|\mathcal{A}(X) - \mathbb{E}\mathcal{D}\|^2 \\
&\geqslant \frac{\widetilde{r}_k^2}{16\mu d}\Omega\left(d \min\left\{1, \left(\frac{\sqrt{d}}{n\sqrt{\rho}}\right)^{\frac{2k-2}{k}}\right\}\right),
\end{aligned}$$

by (44).

2. We use an identical construction to that used above in part 1 except that the scaling factor $\widetilde{r}_k$ gets replaced by r_k. It is easy to see that $\mathbb{E}\left[\sup_{w\in\mathcal{W}} \|\nabla f(w, x)\|^k\right] \approx \sup_{w\in\mathcal{W}} \mathbb{E}\left[\|\nabla f(w, x)\|^k\right]$ for our construction, and the lower bound in part 2 follows just as it did in part 1. This completes the proof. ■

Remark 39 *Note that the lower bound proofs construct bounded (hence subexponential) distributions and uniformly L_f-Lipschitz, β_f-smooth losses that easily satisfy the conditions in our upper bound theorems.*

Appendix G. Details and Proofs for Section 4: Linear Time Algorithms

G.1. Noisy Clipped Accelerated SGD for Smooth Convex Losses (Section 4.1)

We first present Algorithm 5, which is a generalized version of Algorithm 4 that allows for any `MeanOracle`. This will be useful for our analysis. Proposition 40 provides excess risk guarantees for Algorithm 5 in terms of the bias and variance of the `MeanOracle`.

Proposition 40 *Consider Algorithm 5 run with a* `MeanOracle` *satisfying $\widetilde{\nabla} F_t(w_t^{md}) = \nabla F(w_t^{md}) + b_t + N_t$, where $\|b_t\| \leqslant B$ (with probability 1), $\mathbb{E}N_t = 0$, $\mathbb{E}\|N_t\|^2 \leqslant \Sigma^2$ for all $t \in [T-1]$, and $\{N_t\}_{t=1}^T$ are independent. Assume that $F : \mathcal{W} \to \mathbb{R}$ is convex and β-smooth, $F(w_0) - F^* \leqslant \Delta$, and $\|w_0 - w^*\| \leqslant D$. Suppose parameters are chosen in Algorithm 5 so that for all $t \in [T]$, $\eta_t > \beta\alpha_t^2$ and $\eta_t/\Gamma_t = \eta_1/\Gamma_1$, where*

$$\Gamma_t := \begin{cases} 1, & t = 1 \\ (1-\alpha_t)\Gamma_t, & t \geqslant 2. \end{cases}$$

Then,

$$\mathbb{E}F(w_T^{ag}) - F^* \leqslant \frac{\Gamma_T \eta_1 D^2}{2} + \Gamma_T \sum_{t=1}^T \left[\frac{2\alpha_t^2(\Sigma^2 + B^2)}{\Gamma_t(\eta_t - \beta\alpha_t^2)} + \frac{\alpha_t}{\Gamma_t}BD\right].$$

Algorithm 5 Generic Framework for DP Accelerated Stochastic Approximation (AC-SA)

Input: Data $X \in \mathcal{X}^n$, iteration number $T \leqslant n$, stepsize parameters $\{\eta_t\}_{t\in[T]}, \{\alpha_t\}_{t\in[T]}$ with $\alpha_1 = 1, \alpha_t \in (0,1)\ \forall t \geqslant 2$, DP `MeanOracle`.
Initialize $w_0^{ag} = w_0 \in \mathcal{W}$ and $t = 1$.
for $t \in [T]$ **do**
  $w_t^{md} := (1-\alpha_t)w_{t-1}^{ag} + \alpha_t w_{t-1}$.
  Draw new batch $\mathcal{B}_t$ (without replacement) of n/T samples from X.
  $\widetilde{\nabla} F_t(w_t^{md}) :=$ `MeanOracle`$(\{\nabla f(w_t^{md}, x)\}_{x\in\mathcal{B}_t}; \frac{n}{T}; \frac{\epsilon^2}{2})$
  $w_t := \text{argmin}_{w\in\mathcal{W}} \left\{\alpha_t\langle\widetilde{\nabla} F_t(w_t^{md}), w\rangle + \frac{\eta_t}{2}\|w_{t-1} - w\|^2\right\}$.
  $w_t^{ag} := \alpha_t w_t + (1-\alpha_t)w_{t-1}^{ag}$.
end for
Output: w_T^{ag}.

In particular, choosing $\alpha_t = \frac{2}{t+1}$ *and* $\eta_t = \frac{4\eta}{t(t+1)}$, $\forall t \geqslant 1$, *where* $\eta \geqslant 2\beta$ *implies*

$$\mathbb{E}F(w_T^{ag}) - F^* \leqslant \frac{4\eta D^2}{T(T+1)} + \frac{4(\Sigma^2 + B^2)(T+2)}{3\eta} + BD.$$

Further, setting $\eta = \max\left\{2\beta, \frac{T^{3/2}\sqrt{\Sigma^2+B^2}}{D}\right\}$ *implies*

$$\mathbb{E}F(w_T^{ag}) - F^* \lesssim \frac{\beta D^2}{T^2} + \frac{D(\Sigma + B)}{\sqrt{T}} + BD. \tag{45}$$

Proof We begin by extending (Ghadimi and Lan, 2012, Proposition 4) to biased/noisy stochastic gradients. Fix any $w_{t-1}, w_{t-1}^{ag} \in \mathcal{W}$. By (Ghadimi and Lan, 2012, Lemma 3), we have

$$F(w_t^{ag}) \leqslant (1-\alpha_t)F(w_{t-1}^{ag}) + \alpha[F(z) + \langle\nabla F(z), w_t - z\rangle] + \frac{\beta}{2}\|w_t^{ag} - z\|^2, \tag{46}$$

for any $z \in \mathcal{W}$. Denote

$$\Upsilon_t(w) := \alpha_t\langle N_t + b_t, w - w_{t-1}\rangle + \frac{\alpha_t^2\|N_t + b_t\|^2}{\eta_t - \beta\alpha_t^2}$$

and $d_t := w_t^{ag} - w_t^{md} = \alpha_t(w_t - w_{t-1})$. Then using (46) with $z = w_t^{md}$, we have

$$\begin{aligned} F(w_t^{ag}) &\leqslant (1-\alpha_t)F(w_{t-1}^{ag}) + \alpha_t[F(w_t^{md}) + \langle\nabla F(w_t^{md}), w_t - w_t^{md}\rangle] + \frac{\beta}{2}\|d_t\|^2 \\ &= (1-\alpha_t)F(w_{t-1}^{ag}) + \alpha_t[F(w_t^{md}) + \langle\nabla F(w_t^{md}), w_t - w_t^{md}\rangle] \\ &\quad + \frac{\eta_t}{2}\|w_{t-1} - w_t\|^2 - \frac{\eta_t - \beta\alpha_t^2}{2\alpha_t^2}\|d_t\|^2, \end{aligned} \tag{47}$$

by the expression for d_t. Now we apply (Ghadimi and Lan, 2012, Lemma 2) with $p(u) = \alpha_t[\langle\widetilde{\nabla} F_t(w_t^{md}, u\rangle]$, $\mu_1 = 0$, $\mu_2 = \eta_t$, $\tilde{x} = w_t^{md}$, and $\tilde{y} = w_{t-1}$ to obtain (conditional on all randomness) for any $w \in \mathcal{W}$:

$$\begin{aligned}
&\alpha_t[F(w_t^{md}) + \langle \tilde{\nabla} F_t(w_t^{md}), w_t - w_t^{md}\rangle] + \frac{\eta_t}{2}\|w_{t-1} - w_t\|^2 \\
&\leqslant \alpha_t[F(w_t^{md}) + \langle \nabla F(w_t^{md}), w - w_t^{md}\rangle] \\
&\quad + \alpha_t\langle N_t + b_t, w - w_t^{md}\rangle + \frac{\eta_t}{2}\|w_{t-1} - w\|^2 - \frac{\eta_t}{2}\|w_t - w\|^2.
\end{aligned}$$

Next, we combine the above inequality with (47) to get

$$\begin{aligned}
F(w_t^{ag}) &\leqslant (1-\alpha_t)F(w_{t-1}^{ag}) + \alpha_t[F(w_t^{md}) + \langle \nabla F(w_t^{md}), w - w_t^{md}\rangle] \\
&\quad + \frac{\eta_t}{2}\left[\|w_{t-1} - w\|^2 - \|w_t - w\|^2\right] + \underbrace{-\frac{\eta_t - \beta\alpha_t^2}{2\alpha_t^2}\|d_t\|^2 + \alpha_t\langle N_t + b_t, w - w_t\rangle}_{U_t}, \quad (48)
\end{aligned}$$

for all $w \in \mathcal{W}$. By Cauchy-Schwartz, we can bound

$$\begin{aligned}
U_t &\leqslant -\frac{\eta_t - \beta\alpha_t^2}{2\alpha_t^2}\|d_t\|^2 + \|N_t + b_t\|\|d_t\| + \alpha_t\langle N_t + b_t, w - w_{t-1}\rangle \\
&\leqslant \Upsilon_t(w), \quad (49)
\end{aligned}$$

where the last inequality follows from maximizing the concave quadratic function $q(\|d_t\|) := -\left[\frac{\eta_t - \beta\alpha_t^2}{2\alpha_t^2}\right]\|d_t\|^2 + \|N_t + b_t\|\|d_t\|$ with respect to $\|d_t\|$. Plugging the bound (49) back into (48) shows that

$$\begin{aligned}
F(w_t^{ag}) &\leqslant (1-\alpha_t)F(w_{t-1}^{ag}) + \alpha_t[F(w_t^{md}) + \langle \nabla F(w_t^{md}), w - w_t^{md}\rangle] + \frac{\eta_t}{2}\left[\|w_{t-1} - w\|^2 - \|w_t - w\|^2\right] \\
&\quad + \Upsilon_t(w). \quad (50)
\end{aligned}$$

Then it can be shown (see (Ghadimi and Lan, 2012, Proposition 5)) that the assumptions on η_t and α_t imply that

$$F(w_T^{ag}) - \Gamma_T \sum_{t=1}^T \left[\frac{\alpha_t}{\Gamma_t}\left(F(w_t^{md}) + \langle \nabla F(w_t^{md}), w - w_t^{md}\rangle\right)\right] \leqslant \Gamma_T \sum_{t=1}^T \frac{\eta_t}{2\Gamma_t}[\|w_{t-1} - w\|^2 - \|w_t - w\|^2] \quad (51)$$

$$+ \Gamma_T \sum_{t=1}^T \frac{\Upsilon_t(w)}{\Gamma_t}, \quad (52)$$

for any $w \in \mathcal{W}$ and any $T \geqslant 1$. Now,

$$\sum_{t=1}^T \frac{\alpha_t}{\Gamma_t} = \frac{1}{\Gamma_T}$$

by definition. Hence by convexity of F,

$$\sum_{t=1}^T \left[\frac{\alpha_t}{\Gamma_t}\left(F(w_t^{md}) + \langle \nabla F(w_t^{md}), w - w_t^{md}\rangle\right)\right] \leqslant F(w), \ \forall w \in \mathcal{W}.$$

Also, since $\Gamma_t/\eta_t = \Gamma_1/\eta_1$ for all $t \geqslant 1$, we have

$$\Gamma_T \sum_{t=1}^{T} \frac{\eta_t}{2\Gamma_t}[\|w_{t-1} - w\|^2 - \|w_t - w\|^2] = \Gamma_t \frac{\eta_1}{2\Gamma_1}[\|w_0 - w\|^2 - \|w_T - w\|^2] \leqslant \Gamma_T \eta_1 \frac{1}{2}\|w_0 - w\|^2,$$

since $\Gamma_1 = 1$. Substituting the above bounds into (51), we get

$$F(w_T^{ag}) - F(w) \leqslant \Gamma_T \eta_1 \frac{1}{2}\|w_0 - w\|^2 + \Gamma_T \sum_{t=1}^{T} \frac{\Upsilon_t(w)}{\Gamma_t}, \ \forall w \in \mathcal{W}. \tag{53}$$

Now, setting $w = w^*$ and taking expectation yields

$$\begin{aligned} \mathbb{E}[F(w_T^{ag}) - F^*] &\leqslant \frac{\Gamma_T \eta_1 D^2}{2} + \Gamma_T \sum_{t=1}^{T} \frac{\mathbb{E}\Upsilon_t(w^*)}{\Gamma_t} & (54)\\ &\leqslant \frac{\Gamma_T \eta_1 D^2}{2} + \Gamma_T \sum_{t=1}^{T} \left[\frac{1}{\Gamma_t}\left(\alpha_t \mathbb{E}\langle b_t, w^* - w_{t-1}\rangle + \frac{2\alpha_t^2(\Sigma^2 + B^2)}{\eta_t - \beta\alpha_t^2}\right)\right] & (55)\\ &\leqslant \frac{\Gamma_T \eta_1 D^2}{2} + \Gamma_T \sum_{t=1}^{T} \left[\frac{1}{\Gamma_t}\left(\alpha_t B D + \frac{2\alpha_t^2(\Sigma^2 + B^2)}{\eta_t - \beta\alpha_t^2}\right)\right], & (56) \end{aligned}$$

where we used conditional independence of N_t and $w^* - w_{t-1}$ given w_{t-1}, Young's inequality, Cauchy-Schwartz, and the definitions of B^2 and Σ^2. This establishes the first claim of the theorem. The second and third claims are simple corollaries, which can be verified as in (Ghadimi and Lan, 2012, Proposition 7) and the ensuing discussion. ■

Theorem 41 (Complete Version of Theorem 15) *Grant Assumption 2. Let $\epsilon > 0$ and assume F is convex and β-smooth. Then, there are algorithmic parameters such that Algorithm 4 is $\frac{\epsilon^2}{2}$-zCDP. Further, if*

$$n \geqslant T := \left\lceil \min\left\{\left(\frac{\beta D}{r}\right)^{2k/(5k-1)}\left(\frac{\epsilon n}{\sqrt{d}}\right)^{(2k-2)/(5k-1)}, \sqrt{\frac{\beta D}{r}} n^{1/4}\right\}\right\rceil,$$

then,

$$\mathbb{E}F(w_T^{ag}) - F^* \lesssim r_k D\left[\frac{1}{\sqrt{n}} + \max\left\{\left(\left(\frac{\beta D}{r_k}\right)^{1/4}\frac{\sqrt{d}}{\epsilon n}\right)^{\frac{4(k-1)}{5k-1}}, \left(\frac{\sqrt{d}}{\epsilon n}\right)^{\frac{k-1}{k}}\right\}\right].$$

Proof **Privacy:** Choose $\sigma^2 = \frac{4C^2T^2}{\epsilon^2 n^2}$. First, the collection of all $\widetilde{\nabla} F_t(w_t^{md}),\ t \in [T]$ is $\frac{\epsilon^2}{2}$-zCDP: since the batches of data drawn in each iteration are disjoint, it suffices (by parallel composition (McSherry, 2009)) to show that $\widetilde{\nabla} F_t(w_t^{md})$ is $\frac{\epsilon^2}{2}$-zCDP for all t. Now, the ℓ_2 sensitivity of each clipped gradient update is bounded by $\Delta = \sup_{w, X \sim X'} \|\frac{T}{n}\sum_{x \in \mathcal{B}_t} \Pi_C(\nabla f(w,x)) - \sum_{x' \in \mathcal{B}'_t} \Pi_C(\nabla f(w,x'))\| = \sup_{w,x,x'} \|\frac{T}{n}\Pi_C(\nabla f(w,x)) - \Pi_C(\nabla f(w,x'))\| \leqslant \frac{2CT}{n}$. Thus, $\widetilde{\nabla} F_t(w_t^{md})$ is $\frac{\epsilon^2}{2}$-zCDP by Proposition 22. Second, the iterates w_t^{ag} are deterministic functions of

$\tilde{\nabla} F_t(w_t^{md})$, so the post-processing property of differential privacy (Dwork and Roth, 2014; Bun and Steinke, 2016) ensures that Algorithm 4 is $\frac{\epsilon^2}{2}$-zCDP.

Excess risk: Consider round $t \in [T]$ of Algorithm 4, where Algorithm 1 is run on input data $\{\nabla f(w_t, x_i^t)\}_{i=1}^{n/T}$. Denote the bias of Algorithm 1 by $b_t := \mathbb{E}\tilde{\nabla} F_t(w_t) - \nabla F(w_t)$, where $\tilde{\nabla} F_t(w_t) = \tilde{\nu}$ in the notation of Algorithm 1. Also let $\hat{\nabla} F_t(w_t) := \hat{\mu}$ (in the notation of Lemma 5) and denote the noise by $N_t = \tilde{\nabla} F_t(w_t) - \nabla F(w_t) - b_t = \tilde{\nabla} F_t(w_t) - \mathbb{E}\tilde{\nabla} F_t(w_t)$. Then we have $B := \sup_{t \in [T]} \|b_t\| \leqslant \frac{r^{(k)}}{(k-1)C^{k-1}}$ and $\Sigma^2 := \sup_{t \in [T]} \mathbb{E}[\|N_t\|^2] \leqslant d\sigma^2 + \frac{r^2 T}{n} \lesssim \frac{dC^2T^2}{\epsilon^2 n^2} + \frac{r^2 T}{n}$, by Lemma 5. Plugging these estimates for B and Σ^2 into Proposition 40 and setting $C = r(\frac{\epsilon n}{\sqrt{dT}})^{1/k}$, we get

$$\begin{aligned}
\mathbb{E}F(w_T^{ag}) - F^* &\lesssim \frac{\beta D^2}{T^2} + \frac{D(\Sigma + B)}{\sqrt{T}} + BD \\
&\lesssim \frac{\beta D^2}{T^2} + \frac{CD\sqrt{dT}}{\epsilon n} + \frac{rD}{\sqrt{n}} + \frac{r^{(k)}D}{C^{k-1}} \\
&\lesssim \frac{\beta D^2}{T^2} + rD\left[\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{dT}}{\epsilon n}\right)^{(k-1)/k}\right]. \qquad (57)
\end{aligned}$$

Now, our choice of T implies that $\frac{\beta D^2}{T^2} \leqslant rD\left[\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{dT}}{\epsilon n}\right)^{(k-1)/k}\right]$ and we get the result upon plugging in T. ∎

G.2. Noisy Clipped SGD for Strongly Convex Losses (Section 4.2)

We begin by presenting the pseudocode for our noisy clipped SGD in Algorithm 6.

Algorithm 6 Noisy Clipped SGD for Heavy-Tailed DP SCO

```
Input: Data X ∈ 𝒳^n, T ⩽ n, stepsizes {η_t}_{t=0}^T, averaging weights {ζ_t}_{t=0}^T, w_0 ∈ 𝒲.
for t ∈ {0, 1, ··· , T} do
    Draw new batch ℬ_t (without replacement) of n/T samples from X.
    ∇̃F_t(w_t) := MeanOracle1({∇f(w_t, x)}_{x∈ℬ_t}; n/T; ε²/2)
    w_{t+1} = Π_𝒲 [w_t − η_t ∇̃F_t(w_t)]
end for
Output: ŵ_T := (1/Z_T) Σ_{t=0}^T ζ_t w_{t+1}, where Z_T = Σ_{t=0}^T ζ_t.
```

Algorithm 7 is a generalized version of Algorithm 6 that allows for any `MeanOracle` and will be useful in our analysis. In Proposition 42, we provide the convergence guarantees for Algorithm 7 in terms of the bias and variance of the `MeanOracle`.

Proposition 42 *Let $F : \mathcal{W} \to \mathbb{R}$ be μ-strongly convex and β-smooth with condition number $\kappa := \frac{\beta}{\mu}$. Let $w_{t+1} := \Pi_{\mathcal{W}}[w_t - \eta_t \tilde{\nabla} F_t(w_t)]$, where $\tilde{\nabla} F_t(w_t) = \nabla F(w_t) + b_t + N_t$, such that the bias and noise (which can depend on w_t and the samples drawn) satisfy $\|b_t\| \leqslant B$ (with probability 1), $\mathbb{E}N_t = 0$, $\mathbb{E}\|N_t\|^2 \leqslant \Sigma^2$ for all $t \in [T-1]$, and that $\{N_t\}_{t=1}^T$ are*

Algorithm 7 Generic Noisy Clipped SGD Framework for Heavy-Tailed SCO

Input: Data $X \in \mathcal{X}^n$, $T \leqslant n$, `MeanOracle`, stepsizes $\{\eta_t\}_{t=0}^T$, averaging weights $\{\zeta_t\}_{t=0}^T$.
Initialize $w_0 \in \mathcal{W}$.
for $t \in \{0, 1, \cdots, T\}$ **do**
  Draw new batch $\mathcal{B}_t$ (without replacement) of n/T samples from X.
  $\widetilde{\nabla} F_t(w_t) :=$ `MeanOracle`$(\{\nabla f(w_t, x)\}_{x \in \mathcal{B}_t}; \frac{n}{T}; \frac{\epsilon^2}{2})$
  $w_{t+1} = \Pi_{\mathcal{W}}\left[w_t - \eta_t \widetilde{\nabla} F_t(w_t)\right]$
end for
Output: $\widehat{w}_T := \frac{1}{Z_T}\sum_{t=0}^T \zeta_t w_{t+1}$, where $Z_T = \sum_{t=0}^T \zeta_t$.

independent. Then, there exist stepsizes $\{\eta_t\}_{t=1}^T$ *and weights* $\{\zeta_t\}_{t=0}^T$ *such that the average iterate* $\widehat{w}_T := \frac{1}{\sum_{t=0}^T \zeta_t}\sum_{t=0}^T \zeta_t w_{t+1}$ *satisfies*

$$\mathbb{E}F(\widehat{w}_T) - F^* \leqslant 32\beta D^2 \exp\left(-\frac{T}{4\kappa}\right) + \frac{72\Sigma^2}{\mu T} + \frac{2B^2}{\mu}.$$

Proof Define $g(w_t) = -\frac{1}{\eta_t}(w_{t+1} - w_t)$. Then

$$\begin{aligned}\mathbb{E}\|w_{t+1} - w^*\|^2 &= \mathbb{E}\|w_t - \eta_t g(w_t) - w^*\|^2 \\ &= \mathbb{E}\|w_t - w^*\|^2 - 2\eta_t \mathbb{E}\langle g(w_t), w_t - w^*\rangle + \eta_t^2 \mathbb{E}\|g(w_t)\|^2. \end{aligned} \tag{58}$$

Now, conditional on all randomness, we use smoothness and strong convexity to write:

$$\begin{aligned}F(w_{t+1}) - F(w^*) &= F(w_{t+1}) - F(w_t) + F(w_t) - F(w^*) \\ &\leqslant \langle F(w_t), w_{t+1} - w_t\rangle + \frac{\beta}{2}\|w_{t+1} - w_t\|^2 + \langle \nabla F(w_t), w_t - w^*\rangle - \frac{\mu}{2}\|w_t - w^*\|^2 \\ &= \langle \widetilde{\nabla} F_t(w_t), w_{t+1} - w^*\rangle + \langle \nabla F(w_t) - \widetilde{\nabla} F_t(w_t), w_{t+1} - w^*\rangle + \frac{\beta\eta_t^2}{2}\|g(w_t)\|^2 \\ &\quad - \frac{\mu}{2}\|w_t - w^*\|^2 \\ &\leqslant \langle g(w_t), w_{t+1} - w^*\rangle + \langle \nabla F(w_t) - \widetilde{\nabla} F_t(w_t), w_{t+1} - w^*\rangle + \frac{\beta\eta_t^2}{2}\|g(w_t)\|^2 \\ &\quad - \frac{\mu}{2}\|w_t - w^*\|^2 \\ &= \langle g(w_t), w_{t+1} - w_t\rangle + \langle g(w_t), w_t - w^*\rangle - \langle b_t + N_t, w_{t+1} - w^*\rangle + \frac{\beta\eta_t^2}{2}\|g(w_t)\|^2 \\ &\quad - \frac{\mu}{2}\|w_t - w^*\|^2 \\ &= \langle g(w_t), w_t - w^*\rangle - \langle b_t + N_t, w_{t+1} - w^*\rangle + \left(\frac{\beta\eta_t^2}{2} - \eta_t\right)\|g(w_t)\|^2 - \frac{\mu}{2}\|w_t - w^*\|^2,\end{aligned}$$

where we used the fact that $\langle \Pi_{\mathcal{W}}(y) - x, \Pi_{\mathcal{W}}(y) - y\rangle \leqslant 0$ for all $x \in \mathcal{W}, y \in \mathbb{R}^d$ (c.f. (Bubeck, 2015, Lemma 3.1)) to obtain the last inequality. Thus,

$$\begin{aligned}-2\eta_t \mathbb{E}\langle g(w_t), w_t - w^*\rangle &\leqslant -2\eta_t \mathbb{E}[F(w_{t+1}) - F^*] \\ &\quad + 2\eta_t \mathbb{E}\left[-\langle b_t + N_t, w_{t+1} - w^*\rangle + \left(\frac{\beta\eta_t^2}{2} - \eta_t\right)\|g(w_t)\|^2 - \frac{\mu}{2}\|w_t - w^*\|^2\right].\end{aligned}$$

Combining the above inequality with (58), we get

$$\mathbb{E}\|w_{t+1} - w^*\|^2 \leqslant (1 - \mu\eta_t)\mathbb{E}\|w_t - w^*\|^2 - 2\eta_t\mathbb{E}[F(w_{t+1}) - F^*] - 2\eta_t\mathbb{E}\langle b_t + N_t, w_{t+1} - w^*\rangle + 2\eta_t\left(\frac{\eta_t^2\beta}{2} - \eta_t\right)\mathbb{E}\|g(w_t)\|^2. \tag{59}$$

Next, consider

$$\begin{aligned}
|\mathbb{E}\langle b_t + N_t, w_{t+1} - w^*\rangle| &\leqslant |\mathbb{E}\langle b_t + N_t, w_{t+1} - w_t\rangle| + |\mathbb{E}\langle b_t + N_t, w_t - w^*\rangle| \\
&= |\mathbb{E}\langle b_t + N_t, w_{t+1} - w_t\rangle| + |\mathbb{E}\langle b_t, w_t - w^*\rangle| \\
&\leqslant |\mathbb{E}\langle b_t + N_t, w_{t+1} - w_t\rangle| + \frac{B^2}{\mu} + \frac{\mu}{4}\mathbb{E}\|w_t - w^*\|^2
\end{aligned}$$

by independence of N_t (which has zero mean) and $w_t - w^*$, and Young's inequality. Next, note that $v := w_t - \eta_t(\nabla F(w_t) + b_t)$ is independent of N_t, so $\mathbb{E}\langle N_t, \Pi_{\mathcal{W}}(v)\rangle = 0$. Thus,

$$\begin{aligned}
|\mathbb{E}\langle N_t, w_{t+1} - w_t\rangle| &= |\mathbb{E}\langle N_t, w_{t+1}\rangle| \\
&= |\mathbb{E}\langle N_t, \Pi_{\mathcal{W}}\left[w_t - \eta_t\left(\nabla F(w_t) + b_t + N_t\right)\right]\rangle| \\
&= |\mathbb{E}\langle N_t, \Pi_{\mathcal{W}}\left[v - \eta_t N_t\right]\rangle| \\
&= |\mathbb{E}\langle N_t, \Pi_{\mathcal{W}}\left[v\right] - \Pi_{\mathcal{W}}[v - \eta_t N_t]\rangle| \\
&\leqslant \mathbb{E}\left[\|N_t\|\|\Pi_{\mathcal{W}}\left[v\right] - \Pi_{\mathcal{W}}[v - \eta_t N_t]\,\|\right] \\
&\leqslant \mathbb{E}\left[\|N_t\|\|\eta_t N_t\|\right] \\
&\leqslant \eta_t\Sigma^2,
\end{aligned}$$

by Cauchy-Schwartz and non-expansiveness of projection. Further,

$$\begin{aligned}
|\mathbb{E}\langle b_t, w_{t+1} - w_t\rangle| &= |\mathbb{E}\langle b_t, -\eta_t g(w_t)\rangle| \\
&\leqslant \frac{B^2}{\mu} + \frac{\eta_t^2\mu}{4}\mathbb{E}\|g(w_t)\|^2,
\end{aligned}$$

by Young's inequality. Therefore,

$$-2\eta_t\mathbb{E}\langle b_t + N_t, w_{t+1} - w^*\rangle \leqslant 2\eta_t\left[\frac{2B^2}{\mu} + \frac{\eta_t^2\mu}{4}\mathbb{E}\|g(w_t)\|^2 + \eta_t\Sigma^2 + \frac{\mu}{4}\mathbb{E}\|w_t - w^*\|^2\right].$$

Plugging this bound back into (59) and choosing $\eta_t \leqslant \frac{1}{\beta} \leqslant \frac{1}{\mu}$ yields:

$$\begin{aligned}
\mathbb{E}\|w_{t+1} - w^*\|^2 &\leqslant \left(1 - \frac{\mu\eta_t}{2}\right)\mathbb{E}\|w_t - w^*\|^2 - 2\eta_t\mathbb{E}[F(w_{t+1}) - F^*] + \frac{4\eta_t B^2}{\mu} + 2\eta_t^2\Sigma^2 \\
&\quad + 2\eta_t\left(\frac{\eta_t^2\beta}{2} - \eta_t + \frac{\eta_t^2\mu}{4}\right)\mathbb{E}\|g(w_t)\|^2 \\
&\leqslant \left(1 - \frac{\mu\eta_t}{2}\right)\mathbb{E}\|w_t - w^*\|^2 - 2\eta_t\mathbb{E}[F(w_{t+1}) - F^*] + \frac{4\eta_t B^2}{\mu} + 2\eta_t^2\Sigma^2.
\end{aligned}$$

Next, we apply Lemma 43 (see below) with $r_t := \mathbb{E}\|w_t - w^*\|^2$, $s_t := \mathbb{E}F(w_{t+1}) - F^* - \frac{2B^2}{\mu}$, $a := \frac{\mu}{2}$, $b := 2$, $c = 2\Sigma^2$, and $g = \beta$. We may assume $s_t \geqslant 0$ for all t: if this inequality breaks for some t, then simply return w_{t+1} instead of $\widehat{w}_T$ to obtain $\mathbb{E}F(w_t) - F^* < \frac{2B^2}{\mu}$. Thus,

$$\frac{1}{\Gamma_T}\sum_{t=0}^{T}\gamma_t \mathbb{E}[F(w_{t+1}) - F^*] \leqslant \frac{1}{2}\left[32\beta D^2 \exp\left(\frac{-\mu T}{4\beta}\right) + \frac{144\Sigma^2}{\mu T} + \frac{2B^2}{\mu}\right]$$

Finally, Jensen's inequality yields the theorem. ■

Lemma 43 *(Stich, 2019, Lemma 3) Let $b > 0$, let $a, c \geqslant 0$, and $\{\eta_t\}_{t \geqslant 0}$ be non-negative step-sizes such that $\eta_t \leqslant \frac{1}{g}$ for all $t \geqslant 0$ for some parameter $g \geqslant a$. Let $\{r_t\}_{t\geqslant 0}$ and $\{s_t\}_{t\geqslant 0}$ be two non-negative sequences of real numbers which satisfy*

$$r_{t+1} \leqslant (1 - a\eta_t)r_t - b\eta_t s_t + c\eta_t^2$$

for all $t \geqslant 0$. Then there exist particular choices of step-sizes $\eta_t \leqslant \frac{1}{g}$ and averaging weights $\zeta_t \geqslant 0$ such that

$$\frac{b}{\Gamma_T}\sum_{t=0}^{T} s_t\zeta_t + a r_{T+1} \leqslant 32 g r_0 \exp\left(\frac{-aT}{2g}\right) + \frac{36c}{aT},$$

where $\Gamma_T := \sum_{t=0}^{T}\gamma_t$.

We are now prepared to prove Theorem 17.

Theorem 44 (Precise statement of Theorem 17) *Grant Assumption 2. Let $\epsilon > 0$, and assume F is μ-strongly convex and β-smooth with $\kappa = \frac{\beta}{\mu} \leqslant n/\ln(n)$. Then, there are parameters such that Algorithm 6 is $\frac{\epsilon^2}{2}$-zCDP, and*

$$\mathbb{E}F(\widehat{w}_T) - F^* \lesssim \frac{r_k^2}{\mu}\left(\frac{1}{n} + \left(\frac{\sqrt{d\kappa\ln(n)}}{\epsilon n}\right)^{\frac{2k-2}{k}}\right). \tag{60}$$

Proof **Privacy:** Choose $\sigma^2 = \frac{4C^2T^2}{\epsilon^2 n^2}$. Since the batches of data drawn in each iteration are disjoint, it suffices (by parallel composition (McSherry, 2009)) to show that $\widetilde{\nabla}F_t(w_t)$ is $\frac{\epsilon^2}{2}$-zCDP for all t. Now, the ℓ_2 sensitivity of each clipped gradient update is bounded by $\Delta = \sup_{w, X\sim X'}\|\frac{T}{n}\sum_{x\in\mathcal{B}_t}\Pi_C(\nabla f(w,x)) - \sum_{x'\in\mathcal{B}_t'}\Pi_C(\nabla f(w,x'))\| = \sup_{w,x,x'}\|\frac{T}{n}\Pi_C(\nabla f(w,x)) - \Pi_C(\nabla f(w,x'))\| \leqslant \frac{2CT}{n}$. Hence Proposition 22 implies that the algorithm is $\frac{\epsilon^2}{2}$-zCDP.
Excess risk: For any iteration $t \in [T]$, denote the bias of Algorithm 1 by $b_t := \mathbb{E}\widetilde{\nabla}F_t(w_t) - \nabla F(w_t)$, where $\widetilde{\nabla}F_t(w_t) = \widetilde{\nu}$ in the notation of Algorithm 1. Also let $\widehat{\nabla}F_t(w_t) := \hat{\nu}$ (in the notation of Lemma 5) and denote the noise by $N_t = \widetilde{\nabla}F_t(w_t) - \nabla F(w_t) - b_t = \widetilde{\nabla}F_t(w_t) - \mathbb{E}\widetilde{\nabla}F_t(w_t)$. Then we have $B := \sup_{t\in[T]}\|b_t\| \leqslant \frac{r^{(k)}}{(k-1)C^{k-1}}$ and $\Sigma^2 := \sup_{t\in[T]}\mathbb{E}[\|N_t\|^2] \leqslant d\sigma^2 + \frac{r^2T}{n} \lesssim \frac{dC^2T^2}{\epsilon^2n^2} + \frac{r^2T}{n}$, by Lemma 5. Plugging these bias and variance estimates into Proposition 42, we get

$$\mathbb{E}F(\widehat{w}_T) - F^* \lesssim \beta D^2 \exp\left(-\frac{T}{4\kappa}\right) + \frac{1}{\mu T}\left(\frac{dC^2T^2}{\epsilon^2n^2} + \frac{r^2T}{n}\right) + \frac{(r^{(k)})^2}{C^{2k-2}\mu}.$$

Choosing $C = r\left(\frac{\epsilon^2 n^2}{dT}\right)^{1/2k}$ implies

$$\mathbb{E}F(\widehat{w}_T) - F^* \lesssim \beta D^2 \exp\left(-\frac{T}{4\kappa}\right) + \frac{r^2}{\mu}\left(\frac{1}{n} + \left(\frac{dT}{\epsilon^2 n^2}\right)^{(k-1)/k}\right).$$

Finally, choosing $T = \left\lceil 4\kappa \ln\left(\frac{\mu\beta D^2}{r^2}\left(n + \left(\frac{\epsilon^2 n^2}{d}\right)^{(k-1)/k}\right)\right)\right\rceil \lesssim \kappa \ln(n)$ yields the result. ■

Appendix H. Details and Proofs for Section 5: Non-Convex Proximal PL Losses

Pseudocode for our algorithm for PPL losses is given in Algorithm 8. Algorithm 9 is a generalization of Algorithm 8 which allows for arbitrary `MeanOracle`. This will be useful for our analysis. Proposition 45 provides a convergence guarantee for Algorithm 9 in terms of the bias and variance of the `MeanOracle`.

Algorithm 8 Noisy Clipped Proximal SGD for Heavy-Tailed DP SO

1: **Input:** Data $X \in \mathcal{X}^n$, $T \leqslant n$, stepsizes $\{\eta_t\}_{t=0}^{T-1}$.
2: Initialize $w_0 \in \mathcal{W}$.
3: **for** $t \in \{0, 1, \cdots, T-1\}$ **do**
4: Draw new batch $\mathcal{B}_t$ (without replacement) of n/T samples from X.
5: $\widetilde{\nabla} F_t^0(w_t) := \texttt{MeanOracle1}(\{\nabla f^0(w_t, x)\}_{x \in \mathcal{B}_t}; \frac{n}{T}; \frac{\epsilon^2}{2})$
6: $w_{t+1} = \texttt{prox}_{\eta_t f^1}\left(w_t - \eta_t \widetilde{\nabla} F_t^0(w_t)\right)$
7: **end for**
8: **Output:** w_T.

Algorithm 9 Generic Noisy Proximal SGD Framework for Heavy-Tailed DP SO

1: **Input:** Data $X \in \mathcal{X}^n$, $T \leqslant n$, `MeanOracle` (and truncation/minibatch parameters), privacy parameter $\rho = \epsilon^2/2$, stepsizes $\{\eta_t\}_{t=0}^{T-1}$.
2: Initialize $w_0 \in \mathcal{W}$.
3: **for** $t \in \{0, 1, \cdots, T-1\}$ **do**
4: Draw new batch $\mathcal{B}_t$ (without replacement) of n/T samples from X.
5: $\widetilde{\nabla} F_t^0(w_t) := \texttt{MeanOracle}(\{\nabla f^0(w_t, x)\}_{x \in \mathcal{B}_t}; \frac{n}{T}; \frac{\epsilon^2}{2})$
6: $w_{t+1} = \texttt{prox}_{\eta_t f^1}\left(w_t - \eta_t \widetilde{\nabla} F_t^0(w_t)\right)$
7: **end for**
8: **Output:** w_T.

Proposition 45 *Consider Algorithm 9 with biased, noisy stochastic gradients: $\widetilde{\nabla} F_t^0(w_t) = \nabla F^0(w_t) + b_t + N_t$, and stepsize $\eta = \frac{1}{2\beta}$. Assume that the bias and noise (which can depend on w_t and the samples drawn) satisfy $\|b_t\| \leqslant B$ (with probability 1), $\mathbb{E}N_t = 0$, $\mathbb{E}\|N_t\|^2 \leqslant \Sigma^2$ for* *all* $t \in [T-1]$*, and that* $\{N_t\}_{t=1}^T$ *are independent. Assume further that* F *is* μ*-Proximal-PL,* F^0 *is* β*-smooth, and* $F(w_0) - F^* \leqslant \Delta$*. Then,*

$$\mathbb{E}F(w_T) - F^* \leqslant \left(1 - \frac{\mu}{2\beta}\right)^T \Delta + \frac{4(B^2 + \Sigma^2)}{\mu}.$$

Proof Our proof extends the ideas in (Lowy et al., 2022) to generic *biased* and noisy gradients without using Lipschitz continuity of f. By β-smoothness, for any $r \in [T-1]$, we have

$$\begin{aligned}
\mathbb{E}F(w_{r+1}) &= \mathbb{E}[F^0(w_{r+1}) + f^1(w_r) + f^1(w_{r+1}) - f^1(w_r)] \\
&\leqslant \mathbb{E}\left\{F(w_r) + \left[\langle \widetilde{\nabla} F_r^0(w_r), w_{r+1} - w_r \rangle + \frac{\beta}{2}\|w_{r+1} - w_r\|^2 + f^1(w_{r+1}) - f^1(w_r)\right]\right\} \\
&\quad + \mathbb{E}\langle \nabla F^0(w_r) - \widetilde{\nabla} F_r^0(w_r), w_{r+1} - w_r \rangle \\
&= \mathbb{E}F(w_r) + \mathbb{E}\Big[\langle \nabla F^0(w_r), w_{r+1} - w_r \rangle + \frac{\beta}{2}\|w_{r+1} - w_r\|^2 + f^1(w_{r+1}) - f^1(w_r) \\
&\quad + \langle b_r + N_r, w_{r+1} - w_r \rangle\Big] - \mathbb{E}\langle b_r + N_r, w_{r+1} - w_r \rangle \\
&\leqslant \mathbb{E}F(w_r) + \mathbb{E}\Big[\langle \nabla F^0(w_r), w_{r+1} - w_r \rangle + \beta\|w_{r+1} - w_r\|^2 + f^1(w_{r+1}) - f^1(w_r) \\
&\quad + \langle b_r + N_r, w_{r+1} - w_r \rangle\Big] + \frac{B^2 + \Sigma^2}{\beta},
\end{aligned} \tag{61}$$

where we used Young's inequality to bound

$$-\mathbb{E}\langle b_r + N_r, w_{r+1} - w_r \rangle \leqslant \frac{B^2 + \Sigma^2}{\beta} + \frac{\beta}{2}\|w_{r+1} - w_r\|^2. \tag{62}$$

Next, we will bound the quantity

$$\mathbb{E}\left[\langle \nabla F^0(w_r), w_{r+1} - w_r \rangle + \beta\|w_{r+1} - w_r\|^2 + f^1(w_{r+1}) - f^1(w_r) + \langle b_r + N_r, w_{r+1} - w_r \rangle\right].$$

Denote $H_r^{\text{priv}}(y) := \langle \nabla F^0(w_r), y - w_r \rangle + \beta\|y - w_r\|^2 + f^1(y) - f^1(w_r) + \langle b_r + N_r, y - w_r \rangle$ and $H_r(y) := \langle \nabla F^0(w_r), y - w_r \rangle + \beta\|y - w_r\|^2 + f^1(y) - f^1(w_r)$. Note that H_r and H_r^{priv} are 2β-strongly convex. Denote the minimizers of these two functions by y_* and y_*^{priv} respectively. Now, conditional on w_r and $N_r + b_r$, we claim that

$$H_r(y_*^{\text{priv}}) - H_r(y_*) \leqslant \frac{\|N_r + b_r\|^2}{2\beta}. \tag{63}$$

To prove (63), we will need the following lemma:

Lemma 46 *(Lowy and Razaviyayn, 2021a, Lemma B.2) Let* $H(y), h(y)$ *be convex functions on some convex closed set* $\mathcal{Y} \subseteq \mathbb{R}^d$ *and suppose that* H *is* 2β*-strongly convex. Assume further that* h *is* L_h*-Lipschitz. Define* $y_1 = \arg\min_{y \in \mathcal{Y}} H(y)$ *and* $y_2 = \arg\min_{y \in \mathcal{Y}}[H(y) + h(y)]$*. Then* $\|y_1 - y_2\|_2 \leqslant \frac{L_h}{2\beta}$*.*

We apply Theorem 46 with $H(y) := H_r(y)$, $h(y) := \langle N_r + b_r, y\rangle$, $L_h = \|N_r + b_r\|$, $y_1 = y_*$, and $y_2 = y_*^{\text{priv}}$ to get

$$\|y_* - y_*^{\text{priv}}\| \leqslant \frac{\|N_r + b_r\|}{2\beta}.$$

On the other hand,

$$H_r^{\text{priv}}(y_*^{\text{priv}}) = H_r(y_*^{\text{priv}}) + \langle N_r + b_r, y_*^{\text{priv}}\rangle \leqslant H_r^{\text{priv}}(y_*) = H_r(y_*) + \langle N_r + b_r, y_*\rangle.$$

Combining these two inequalities yields

$$\begin{aligned} H_r(y_*^{\text{priv}}) - H_r(y_*) &\leqslant \langle N_r + b_r, y_* - y_*^{\text{priv}}\rangle \\ &\leqslant \|N_r + b_r\|\|y_* - y_*^{\text{priv}}\| \\ &\leqslant \frac{\|N_r + b_r\|^2}{2\beta}, \end{aligned} \tag{64}$$

as claimed. Also, note that $w_{r+1} = y_*^{\text{priv}}$. Hence

$$\begin{aligned} &\mathbb{E}\left[\langle \nabla F^0(w_r), w_{r+1} - w_r\rangle + \beta\|w_{r+1} - w_r\|^2 + f^1(w_{r+1}) - f^1(w_r) + \langle b_r + N_r, w_{r+1} - w_r\rangle\right] \\ &\quad = \mathbb{E}\left[\min_{y \in \mathbb{R}^d} H_r^{\text{priv}}(y)\right] \end{aligned}$$

satisfies

$$\mathbb{E}\left[\min_{y \in \mathbb{R}^d} H_r^{\text{priv}}(y)\right] \leqslant \mathbb{E}\left[\min_y \left\{\langle \nabla F^0(w_r), y - w_r\rangle + \beta\|y - w_r\|^2 + f^1(y) - f^1(w_r)\right\}\right] + \frac{\Sigma^2 + B^2}{\beta} \tag{65}$$

$$\leqslant -\frac{\mu}{2\beta}\mathbb{E}\left[F(w_r) - F^*\right] + \frac{\Sigma^2 + B^2}{\beta}, \tag{66}$$

where we used the assumptions that F is μ-PPL and F^0 is 2β-smooth in the last inequality. Plugging the above bounds back into (61), we obtain

$$\mathbb{E}F(w_{r+1}) \leqslant \mathbb{E}F(w_r) - \frac{\mu}{2\beta}[F(w_r) - F^*] + \frac{2(\Sigma^2 + B^2)}{\beta}, \tag{67}$$

whence

$$\mathbb{E}[F(w_{r+1}) - F^*] \leqslant \mathbb{E}[F(w_r) - F^*](1 - \frac{\mu}{2\beta}) + \frac{2(\Sigma^2 + B^2)}{\beta}. \tag{68}$$

Using (68) recursively and summing the geometric series, we get

$$\mathbb{E}[F(w_T) - F^*] \leqslant \Delta\left(1 - \frac{\mu}{2\beta}\right)^T + \frac{4(\Sigma^2 + B^2)}{\mu}. \tag{69}$$

■

Theorem 47 (Precise statement of Theorem 19) *Grant Assumption 2. Let $\epsilon > 0$ and assume $F(w) = F^0(w) + f^1(w)$ is μ-PPL for β-smooth F^0, with $\kappa = \frac{\beta}{\mu} \leqslant n/\ln(n)$. Then, there are parameters such that Algorithm 8 is $\frac{\epsilon^2}{2}$-zCDP, and*

$$\mathbb{E}F(w_T) - F^* \lesssim \frac{r_k^2}{\mu}\left(\left(\frac{\sqrt{d}}{\epsilon n}\kappa\ln(n)\right)^{\frac{2k-2}{k}} + \frac{\kappa\ln(n)}{n}\right).$$

Proof We choose $\sigma^2 = \frac{4C^2T^2}{\epsilon^2 n^2}$.
Privacy: By parallel composition (since each sample is used only once) and the post-processing property of DP (since the iterates are deterministic functions of the output of `MeanOracle1`), it suffices to show that $\widetilde{\nabla} F_t(w_t)$ is $\frac{\epsilon^2}{2}$-zCDP for all $t \geqslant 0$. By our choice of σ^2 and Proposition 22, $\widetilde{\nabla} F_t(w_t)$ is $\frac{\epsilon^2}{2}$-zCDP, since it's sensitivity is bounded by $\sup_{X\sim X',w}\frac{T}{n}\left\|\sum_{x\in\mathcal{B}_t}\Pi_C[\nabla f^0(w,x)] - \sum_{x'\in\mathcal{B}'_t}\Pi_C[\nabla f^0(w,x')]\right\| \leqslant \frac{T}{n}\sup_{x,x',w}\|\Pi_C[\nabla f^0(w,x)] - \Pi_C[\nabla f^0(w,x')]\| \leqslant \frac{2CT}{n}$.
Excess risk: For any iteration $t \in [T]$, denote the bias of `MeanOracle1` (Algorithm 1) by $b_t := \mathbb{E}\widetilde{\nabla} F_t(w_t) - \nabla F(w_t)$, where $\widetilde{\nabla} F_t(w_t) = \tilde{\nu}$ in the notation of Algorithm 1. Also let $\widehat{\nabla} F_t(w_t) := \hat{\nu}$ (in the notation of Lemma 5) and denote the noise by $N_t = \widetilde{\nabla} F_t(w_t) - \nabla F(w_t) - b_t = \widetilde{\nabla} F_t(w_t) - \mathbb{E}\widetilde{\nabla} F_t(w_t)$. Then we have $B := \sup_{t\in[T]}\|b_t\| \leqslant \frac{r^{(k)}}{(k-1)C^{k-1}}$ and $\Sigma^2 := \sup_{t\in[T]}\mathbb{E}[\|N_t\|^2] \leqslant d\sigma^2 + \frac{r^2 T}{n} \leqslant \frac{4dC^2T^2}{\epsilon^2 n^2} + \frac{r^2 T}{n}$, by Lemma 5. Plugging these bounds on B^2 and Σ^2 into Proposition 45, and choosing $T = 2\left\lceil \kappa \ln\left(\frac{\Delta\mu}{B^2+\Sigma^2}\right)\right\rceil \lesssim \kappa\ln(n)$ where $\Delta \geqslant F(w_0) - F^*$, we have:

$$\mathbb{E}F(w_T) - F^* \leqslant \frac{5(B^2+\Sigma^2)}{\mu} \leqslant \frac{5}{\mu}\left(\frac{2r^2T}{n} + \frac{2(r^{(k)})^2}{(k-1)^2C^{2k-2}} + \frac{2dC^2T^2}{\epsilon^2 n^2}\right),$$

for any $C > 0$. Choosing $C = r\left(\frac{\epsilon^2 n^2}{dT^2}\right)^{1/2k}$ makes the last two terms in the above display equal, and we get

$$\mathbb{E}F(w_T) - F^* \lesssim \frac{r^2}{\mu}\left(\left(\frac{\sqrt{d}}{\epsilon n}\kappa\ln(n)\right)^{\frac{2k-2}{k}} + \frac{\kappa\ln(n)}{n}\right),$$

as desired.
■

Appendix I. Shuffle Differentially Private Algorithms

In the next two subsections, we present two SDP algorithms for DP heavy-tailed mean estimation. The first is an SDP version of Algorithm 1 and the second is an SDP version of the coordinate-wise protocol of (Kamath et al., 2020, 2022). Both of our algorithms offer the same utility guarantees as their zCDP counterparts (up to logarithms). In particular, this implies that the upper bounds obtained in the main body of this paper can also be attained via SDP protocols that do not require individuals to trust any third party curator with their sensitive data (assuming the existence of a secure shuffler).

I.1. ℓ_2 Clip Shuffle Private Mean Estimator

For heavy-tailed SO problems satisfying Assumption 2, we propose using the SDP mean estimation protocol described in Algorithm 10. Algorithm 10 relies on the shuffle private vector summation protocol of (Cheu et al., 2021), which is given in Algorithm 11. The useful properties of Algorithm 11 are contained in Lemma 48.

Algorithm 10 ℓ_2 Clip `ShuffleMeanOracle1`$(\{x_i\}_{i=1}^s; C; (\epsilon, \delta))$

Input: $X = \{x_i\}_{i=1}^s$, $x_i = (x_{i,1}, \cdots, x_{i,d}) \in \mathbb{R}^d$, $C > 0$, $(\epsilon, \delta) \in (\mathbb{R}_+ \times (0, 1/2))$.
for $i \in [s]$ **do**
  $z_i := \Pi_C(x_i)$.
end for
$\tilde{\nu} := \mathcal{P}_{\text{vec}}(\{z_i\}_{i=1}^s; s; C; (\epsilon, \delta))$.
Output: $\tilde{\nu}$.

Algorithm 11 $\mathcal{P}_{\text{vec}}$, a shuffle private protocol for vector summation

Input: database of d-dimensional vectors $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_s$ with maximum norm bounded by $C > 0$; privacy parameters (ϵ, δ).
procedure: Local Randomizer $\mathcal{R}_{\text{vec}}(\mathbf{x}_i)$
  for $j \in [d]$ **do**
    Shift component to enforce non-negativity: $\mathbf{w}_{i,j} \leftarrow \mathbf{x}_{i,j} + C$
    $\mathbf{m}_j \leftarrow \mathcal{R}_{1D}(\mathbf{w}_{i,j})$
  end for
  Output labeled messages $\{(j, \mathbf{m}_j)\}_{j \in [d]}$
end procedure
procedure: Analyzer $\mathcal{A}_{\text{vec}}(\mathbf{y})$
  for $j \in [d]$ **do**
    Run analyzer on coordinate j's messages $z_j \leftarrow \mathcal{A}_{\text{1D}}(\mathbf{y}_j)$
    Re-center: $o_j \leftarrow z_j - L$
  end for
  Output the vector of estimates $\mathbf{o} = (o_1, \cdots o_d)$
end procedure

Lemma 48 *(Cheu et al., 2021, Theorem 3.2) Let $\epsilon \leqslant 15, \delta \in (0, 1/2)$, $d, s \in \mathbb{N}$ and $C > 0$. There are choices of parameters b, g, p for $\mathcal{P}_{1D}$ such that for an input data set $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_s)$ of vectors with maximum norm $\|\mathbf{x}_i\| \leqslant C$, the following holds:*
1) Algorithm 11 is (ϵ, δ)-SDP.
2) $\mathcal{P}_{vec}(\mathbf{X})$ is an unbiased estimate of $\sum_{i=1}^s \mathbf{x}_i$ with bounded variance

$$\mathbb{E}\left[\left\|\mathcal{P}_{vec}(\mathbf{X}) - \sum_{i=1}^s \mathbf{x}_i\right\|^2\right] = \mathcal{O}\left(\frac{dC^2}{\epsilon^2} \ln^2(d/\delta)\right).$$

By the post-processing property of DP, we immediately obtain:

Lemma 49 (Privacy, Bias, and Variance of Algorithm 10) *Let $\{z_i\}_{i=1}^s \sim \mathcal{D}^s$ have mean $\mathbb{E}z_i = \nu$ and $\mathbb{E}\|z_i\|^k \leqslant r^{(k)}$ for some $k \geqslant 2$. Denote the noiseless average of clipped samples in Algorithm 10 by $\widehat{\nu} := \frac{1}{n}\sum_{i=1}^s \Pi_C(z_i)$. Then, there exist algorithmic parameters such that Algorithm 10 is (ϵ, δ)-SDP and such that the following bias and variance bounds hold:*

$$\|\mathbb{E}\widetilde{\nu} - \nu\| = \|\mathbb{E}\widehat{\nu} - \nu\| \leqslant \mathbb{E}\|\widehat{\nu} - \nu\| \leqslant \frac{r^{(k)}}{(k-1)C^{k-1}}, \tag{70}$$

and

$$\mathbb{E}\|\widetilde{\nu} - \mathbb{E}\widetilde{\nu}\|^2 = \mathbb{E}\|\widetilde{\nu} - \mathbb{E}\widehat{\nu}\|^2 = \mathcal{O}\left(\frac{dC^2\ln^2(d/\delta)}{\epsilon^2 s^2} + \frac{r^2}{s}\right). \tag{71}$$

Proof **Privacy:** The privacy claim is immediate from Lemma 48 and the post-processing property of DP (Dwork and Roth, 2014, Proposition 2.1).
Bias: The bias bound follows as in Lemma 5, since $\mathcal{P}_{\text{vec}}$ is an unbiased estimator (by Lemma 48).
Variance: We have

$$\begin{aligned}
\mathbb{E}\|\widetilde{\nu} - \mathbb{E}\widetilde{\nu}\|^2 &= \mathbb{E}\|\widetilde{\nu} - \mathbb{E}\widehat{\nu}\|^2 \\
&= \mathbb{E}\|\widetilde{\nu} - \widehat{\nu}\|^2 + \mathbb{E}\|\widehat{\nu} - \mathbb{E}\widehat{\nu}\|^2 \\
&\lesssim \frac{dC^2\ln^2(d/\delta)}{\epsilon^2 s^2} + \frac{1}{s}\mathbb{E}\|\Pi_C(z_1) - \mathbb{E}\Pi_C(z_1)\|^2 \\
&\leqslant \frac{dC^2\ln^2(d/\delta)}{\epsilon^2 s^2} + \frac{1}{s}\mathbb{E}\|z_1 - \mathbb{E}z_1\|^2 \\
&\leqslant \frac{dC^2\ln^2(d/\delta)}{\epsilon^2 s^2} + \frac{r^{(2)}}{s},
\end{aligned}$$

where we used that the samples $\{z_i\}_{i=1}^s$ are i.i.d., the variance bound in Lemma 48, and (Barber and Duchi, 2014, Lemma 4), which states that $\mathbb{E}\|\Pi_C(X) - \mathbb{E}\Pi_C(X)\|^2 \leqslant \mathbb{E}\|X - \mathbb{E}X\|^2$ for any random vector X. ■

Remark 50 *Comparing Lemma 49 to Lemma 5, we see that the bias and variance of the two `MeanOracles` are the same up to logarithmic factors. Therefore, replacing Algorithm 1 by Algorithm 10 in our stochastic optimization algorithms yields SDP algorithms with excess risk that matches the bounds provided in this paper (via Algorithm 1) up to logarithmic factors.*

I.2. Coordinate-wise Shuffle Private Mean Estimation Oracle

For SO problems satisfying Assumption 3, we propose Algorithm 13 as a shuffle private mean estimation oracle. Algorithm 13 is a shuffle private variation of Algorithm 12, which was employed by (Kamath et al., 2020, 2022).

The bias/variance and privacy properties of Algorithm 12 are summarized in Lemma 51.

Lemma 51 (Privacy, Bias, and Variance of Algorithm 12, (Kamath et al., 2020, 2022)) *Let $\{x_i\}_{i=1}^s \sim \mathcal{D}^s$ have mean $\mathbb{E}x_i = \nu$, $\|\nu\| \leqslant L$, and $\mathbb{E}|x_{i,j} - \nu_j|^k \leqslant \gamma$ for some $k \geqslant 2$. Denote by $\widehat{\nu}$ the output of the non-private algorithm that is identical to Algorithm 12 except without*

Algorithm 12 Coordinate-wise Private $\texttt{MeanOracle2}(\{x_i\}_{i=1}^s; s; \tau; \frac{\epsilon^2}{2}; m)$ (Kamath et al., 2020, 2022)

Input: $X = \{x_i\}_{i=1}^s$, $x_i = (x_{i,1}, \cdots, x_{i,d}) \in \mathbb{R}^d$, $\epsilon > 0, \tau > 0$, $m \in [s]$ such that m divides s.
for $j \in [d]$ **do**
  Partition j-th coordinates of data into m disjoint groups of size s/m.
  for $i \in [m]$ **do**
    Clip data in i-th group: $Z_j^i := \left\{\Pi_{[-\tau,\tau]}(x_{(i-1)\frac{s}{m}+1,j}), \cdots, \Pi_{[-\tau,\tau]}(x_{i\frac{s}{m},j})\right\}$.
    Compute average of Z_j^i: $\widehat{\nu}_j^i := \frac{m}{s}\sum_{z \in Z_j^i} z$.
  end for
  Compute median of group means: $\widehat{\nu}_j := \text{median}(\widehat{\nu}_j^1, \cdots, \widehat{\nu}_j^m)$.
  Draw $u \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$, with $\sigma^2 = \frac{4\tau^2 m^2 d}{s^2\epsilon^2}$.
end for
Output: $\widetilde{\nu} = (\widehat{\nu}_1, \cdots, \widehat{\nu}_d) + u$.

the added Gaussian noise. Then, Algorithm 12 with $\sigma^2 = \frac{72\tau^2 m^2 d}{\epsilon^2 s^2}$ is $\frac{\epsilon^2}{2}$-zCDP. Further, if $\tau \geqslant 2L$, then there is $m = \widetilde{\mathcal{O}}(1) \in [s]$ such that:

$$\|\mathbb{E}\widetilde{\nu} - \nu\| = \|\mathbb{E}\widehat{\nu} - \mu\| \leqslant \mathbb{E}\|\widehat{\nu} - \nu\| = \widetilde{\mathcal{O}}\left(\sqrt{d}\left(\frac{\gamma^{1/k}}{\sqrt{s}} + \left(\frac{2}{\tau}\right)^{k-1}\gamma\right)\right) =: B, \tag{72}$$

and

$$\mathbb{E}\|N\|^2 = \mathbb{E}\|\widetilde{\nu} - \mathbb{E}\widetilde{\nu}\|^2 = \mathbb{E}\|\widetilde{\nu} - \mathbb{E}\widehat{\nu}\|^2 = \widetilde{\mathcal{O}}\left(B^2 + d\sigma^2\right). \tag{73}$$

Algorithm 13 Coordinate-wise Shuffle Private $\texttt{ShuffleMeanOracle2}(\{x_i\}_{i=1}^s; s; \tau; \epsilon, \delta; m)$

Input: $X = \{x_i\}_{i=1}^s$, $x_i = (x_{i,1}, \cdots, x_{i,d}) \in \mathbb{R}^d$, $\epsilon > 0, \delta \in (0,1)$, $m \in [s]$.
for $j \in [d]$ **do**
  $\epsilon_j := \frac{\epsilon}{4\sqrt{2d\ln(1/\delta_j)}}, \delta_j := \frac{\delta}{2d}$.
  Partition j-th coordinates of data into m disjoint groups of size s/m.
  for $i \in [m]$ **do**
    Clip j-th coordinate of data in i-th group: $Z_j^i := \left\{\Pi_{[-\tau,\tau]}(x_{(i-1)\frac{s}{m}+1,j}), \cdots, \Pi_{[-\tau,\tau]}(x_{i\frac{s}{m},j})\right\}$.
    Shift to enforce non-negativity: $Z_j^i \leftarrow Z_j^i + (\tau, \cdots, \tau)$.
    Compute noisy average of s/m scalars in Z_j^i: $\widetilde{\nu}_j^i := \frac{m}{s}\mathcal{P}_{1D}(Z_j^i; \frac{s}{m}; 2\tau; \epsilon_j, \delta_j)$.
    Re-center: $\widetilde{\nu}_j^i \leftarrow \widetilde{\nu}_j^i - \tau$.
  end for
  Compute median of noisy means: $\widetilde{\nu}_j := \text{median}(\widetilde{\nu}_j^1, \cdots, \widetilde{\nu}_j^m)$.
end for
Output: $\widetilde{\nu} = (\widetilde{\nu}_1, \cdots, \widetilde{\nu}_d)$.

The $\mathcal{P}_{1D}$ subroutine used in Algorithm 13 is an SDP protocol for summing scalars that we borrow from (Cheu et al., 2021). It is outlined in Algorithm 14. Algorithm 14 decomposes into a local randomizer $\mathcal{R}$ that individuals execute on their own data, and an analyzer component $\mathcal{A}$ that the shuffler executes. $\mathcal{S}(\mathbf{y})$ denotes the shuffled vector $\mathbf{y}$: i.e. the vector whose components are random permutations of the components of $\mathbf{y}$. We describe the privacy guarantee, bias, and variance of Algorithm 13 in Proposition 52.

Algorithm 14 $\mathcal{P}_{1\text{D}}$, a shuffle private protocol for summing scalars (Cheu et al., 2021)

Input: Scalars $Z = (z_1, \cdots z_s) \in [0, \tau]^s$; design parameters $g, b \in \mathbb{N}; p \in (0, \frac{1}{2})$.
procedure: Local Randomizer $\mathcal{R}_{1D}(z_i)$
  for $i \in [s]$ **do**
    $\bar{z}_i \leftarrow \lfloor z_i g/\tau \rfloor$.
    Sample rounding value $\eta_1 \sim \mathbf{Ber}(z_i g/\tau - \bar{z}_i)$.
    Set $\hat{z}_i \leftarrow \bar{z}_i + \eta_1$.
    Sample privacy noise value $\eta_2 \sim \mathbf{Bin}(b, p)$.
    Report $y_i = (y_{i,1}, \cdots, y_{i,g+b}) \in \{0,1\}^{g+b}$ containing $\hat{z}_i + \eta_2$ copies of 1 and $g + b - (\hat{z}_i + \eta_2)$ copies of 0.
  end for
end procedure
procedure: Shuffler $\mathcal{S}(\mathbf{y})$
  Shuffler receives $\mathbf{y} := (y_1, \cdots, y_s)$, draws a uniformly random permutation π of $[g+b] \times [s]$, and sends $\mathcal{S}(\mathbf{y}) := (y_{\pi(1,1)}, \cdots, y_{\pi(s,g+b)})$ to analyzer.
end procedure
procedure: Analyzer $\mathcal{A}_{1\text{D}}(\mathcal{S}(\mathbf{y}))$
  Output: $\frac{\tau}{g}[(\sum_{i=1}^{s}\sum_{l=1}^{b+g} y_{\pi(i,l)}) - pbs]$.

Proposition 52 (Privacy, Bias, and Variance of Algorithm 13) *Let $\{x_i\}_{i=1}^s \sim \mathcal{D}^s$ have mean $\mathbb{E}x_i = \nu$, $\|\nu\| \leqslant L$, and $\mathbb{E}\|x_{i,j} - \nu_j\|^k \leqslant \gamma$ for some $k \geqslant 2$. Let $\epsilon \leqslant 8\ln(2d/\delta)$, $\delta \in (0, 1/2)$, and choose $\tau \geqslant 2L$. Then, there exist choices of parameters (g, b, p, m) such that Algorithm 8 is (ϵ, δ)-SDP, and has bias and variance bounded as follows:*

$$\|\mathbb{E}\tilde{\nu} - \nu\| = \tilde{\mathcal{O}}\left(\sqrt{d}\left(\frac{\gamma^{1/k}}{\sqrt{s}} + \left(\frac{2}{\tau}\right)^{k-1}\gamma\right)\right) \tag{74}$$

and

$$\mathbb{E}\|N\|^2 = \mathbb{E}\|\tilde{\nu} - \mathbb{E}\tilde{\nu}\|^2 = \tilde{\mathcal{O}}\left(\frac{\tau^2 d \ln^2(d/\delta)}{s^2\epsilon^2} + d\left(\frac{1}{s}\gamma^{2/k} + \gamma^2\left(\frac{2}{\tau}\right)^{2k-2}\right)\right). \tag{75}$$

Remark 53 *Comparing Proposition 52 with Lemma 51, we see that the bias and variance of the two `MeanOracle`s Algorithm 13 and Algorithm 12 are the same up to logarithmic factors. Therefore, using Algorithm 13 as `MeanOracle` for stochastic optimization results in the same excess risk bounds as one would get by using Algorithm 12, up to logarithmic factors.*

The proof of Proposition 52 will require Lemma 55, which is due to (Cheu et al., 2021). First, we need the following notation:

Definition 54 (δ-Approximate Max Divergence) *For random variables X and Y, define*

$$D_{\infty}^{\delta}(X||Y) = \sup_{S \subseteq supp(X):\mathbb{P}(X \in S) \geqslant \delta} \ln \left[\frac{\mathbb{P}(X \in S) - \delta}{\mathbb{P}(Y \in S)}\right].$$

An important fact is that a randomized algorithm $\mathcal{A}$ is (ϵ, δ)-DP if and only if $D_{\infty}^{\delta}(\mathcal{A}(X)||\mathcal{A}(X')) \leqslant \epsilon$ for all adjacent data sets $X \sim X'$ (Dwork and Roth, 2014).

Lemma 55 *(Cheu et al., 2021, Lemma 3.1) Let $s \in \mathbb{N}, \epsilon \leqslant 15$. Let $g \geqslant \tau\sqrt{s}$, $b > \frac{180g^2 \ln(2/\delta)}{\epsilon^2 s}$ and $p = \frac{90g^2 \ln(2/\delta)}{b\epsilon^2 s}$. Then for any adjacent scalar databases $Z, Z' \in [0, \tau]^r$ differing on user u ($z_u \neq z'_u$), we have:*
a) $D_{\infty}^{\delta}(\mathcal{S} \circ \mathcal{R}_{1D}^{r}(Z)||\mathcal{S} \circ \mathcal{R}_{1D}^{r}(Z')) \leqslant \epsilon\left(\frac{2}{g} + \frac{|z_u - z'_u|}{\tau}\right)$.
b) Unbiasedness: $\mathbb{E}[\mathcal{P}_{1D}(Z)] = \sum_{i=1}^{s} z_i$.
c) Variance bound: $\mathbb{E}[(\mathcal{P}_{1D}(Z) - \sum_{i=1}^{s} z_i)^2] = \mathcal{O}\left(\frac{\tau^2}{\epsilon^2}\ln(1/\delta)\right)$.

To prove the utility guarantees in Proposition 52, we begin by providing (Lemma 56) high probability bounds on the bias and noise induced by the *non-private* version of Algorithm 13, in which $\mathcal{P}_{1D}$ in line 8 is replaced by the sum of $z \in Z_j^i$. Lemma 56 is a refinement of (Kamath et al., 2022, Theorem 4.1), with correct scaling for arbitrary $\gamma > 0$ and exact constants:

Lemma 56 *Let $\zeta \in (0,1)$ and $X \sim \mathcal{D}$ be a random d-dimensional vector with mean ν such that $\|\nu\| \leqslant L$ and $\mathbb{E}|\langle X - \nu, e_j\rangle|^k \leqslant \gamma$ for some $k \geqslant 2$ for all $j \in [d]$. Consider the non-private version of Algorithm 13 run with $m = \lceil 20\log(4d/\zeta)\rceil$ and $\tau \geqslant 2L$, where $\mathcal{P}_{1D}$ in line 8 is replaced by the sum of $z \in Z_j^i$. Denote the output of this algorithm by $\hat{\nu}$. Then with probability at least $1 - \zeta$, we have*

$$\|\hat{\nu} - \nu\| \leqslant 10\sqrt{d}\left(\sqrt{\frac{m}{s}}\gamma^{1/k} + \gamma\left(\frac{2}{\tau}\right)^{k-1}\right).$$

Proof Denote $X = (x_1, \cdots, x_d)$, $\nu = (\nu_1, \cdots, \nu_d)$, and $z_j := \Pi_{[-\tau,\tau]}(x_j)$ for $j \in [d]$. By Lemma 57 (stated and proved below),

$$|\mathbb{E}z_j - \nu_j| \leqslant 10\gamma\left(\frac{2}{\tau}\right)^{k-1}. \tag{76}$$

Now an application of (Minsker, 2022, Lemma 3) with $\rho(t) := \begin{cases} \frac{t^2}{2} & \text{if } t \in [-\tau, \tau] \\ \tau t & \text{if } t > \tau \\ -\tau t & \text{if } t < -\tau \end{cases}$ shows that $\mathbb{E}|z_j - \mathbb{E}z_j|^k \leqslant \gamma$. Hence by Lemma 58 (stated and proved below),

$$\mathbb{P}\left(\left|\hat{\nu}_j^i - \mathbb{E}z_j\right| \leqslant 10\sqrt{\frac{m}{s}}\gamma^{1/k}\right) = \mathbb{P}\left(\left|\frac{m}{s}\sum_{z \in Z_j^i} z - \mathbb{E}z_j\right| \leqslant 10\sqrt{\frac{m}{s}}\gamma^{1/k}\right) \geqslant 0.99, \ \forall i \in [m], \ j \in [d]. \tag{77}$$

Next, using the "median trick" (via Chernoff/Hoeffding bound, see e.g. (Andoni, 2015)), we get

$$\mathbb{P}\left(\left|\hat{\nu}_j - \mathbb{E}z_j\right| \leqslant 10\sqrt{\frac{m}{s}}\gamma^{1/k}\right) \leqslant 2e^{-m/20} \leqslant \frac{\zeta}{2d}, \tag{78}$$

where the last inequality follows from our choice of m. Now, by union bound, we have for any $a > 0$ that

$$\mathbb{P}(\|\hat{\nu} - \nu\| \geqslant \sqrt{d}a) \leqslant \sum_{j=1}^{d} \mathbb{P}(|\hat{\nu}_j - \nu_j| \geqslant a) \tag{79}$$

$$\leqslant \sum_{j=1}^{d} [\mathbb{P}(|\hat{\nu}_j - \mathbb{E}z_j| \geqslant a/2) + \mathbb{P}(|\mathbb{E}z_j - \nu_j| \geqslant a/2)]. \tag{80}$$

Plugging in $a = 2\left[10\gamma^{1/k}\sqrt{\frac{m}{s}} + 10\gamma\left(\frac{2}{\tau}\right)^{k-1}\right]$ implies

$$\mathbb{P}\left(\|\hat{\nu} - \nu\| \geqslant \sqrt{d}\left(10\gamma^{1/k}\sqrt{\frac{m}{s}} + 10\gamma\left(\frac{2}{\tau}\right)^{k-1}\right)\right) \leqslant d\left(\frac{\zeta}{2d} + 0\right) = \frac{\zeta}{2}, \tag{81}$$

by (76) and (78). ■

Below we give the lemmas that we used in the proof of Lemma 56. The following is a refinement of (Kamath et al., 2022, Lemma B.1) with the proper scaling in γ:

Lemma 57 *Let $x \sim \mathcal{D}$ be a random variable with mean ν and $\mathbb{E}|x - \nu|^k \leqslant \gamma$ for some $k \geqslant 2$. Let $z = \Pi_{[-\tau,\tau]}(x)$ for $\tau \geqslant 2|\nu|$. Then,*

$$|\nu - \mathbb{E}z| \leqslant 10\gamma\left(\frac{2}{\tau}\right)^{k-1}.$$

Proof We begin by recalling the following form of Chebyshev's inequality:

$$\mathbb{P}(|x - \nu| \geqslant c) \leqslant \frac{\mathbb{E}|x - \nu|^k}{c^k} \leqslant \frac{\gamma}{c^k} \tag{82}$$

for any $c > 0$, via Markov's inequality. By symmetry, we may assume without loss of generality that $\nu \leqslant 0$. Note that

$$\nu - \mathbb{E}z = \mathbb{E}\left[x\mathbb{1}_{x<-\tau} + x\mathbb{1}_{x>\tau} + x\mathbb{1}_{x\in[-\tau,\tau]} - (-\tau\mathbb{1}_{x<-\tau} + \tau\mathbb{1}_{x>\tau} + x\mathbb{1}_{x\in[-\tau,\tau]})\right] \tag{83}$$

$$= \mathbb{E}\left[(x+\tau)\mathbb{1}_{x<-\tau} + (x-\tau)\mathbb{1}_{x>\tau}\right]. \tag{84}$$

So,

$$|\nu - \mathbb{E}z| \leqslant \underbrace{|\mathbb{E}(x+\tau)\mathbb{1}_{x<-\tau}|}_{\text{ⓐ}} + \underbrace{|\mathbb{E}(x-\tau)\mathbb{1}_{x>\tau}|}_{\text{ⓑ}}, \tag{85}$$

by the triangle inequality. Now,

$$\begin{aligned}
ⓐ &= |\mathbb{E}[x-\nu-(-\tau-\nu)\mathbb{1}_{x<-\tau}| \\
&\leqslant \mathbb{E}[|x-\nu|\mathbb{1}_{x<-\tau}] + |-\tau-\nu|\mathbb{E}\mathbb{1}_{x<-\tau} \\
&\leqslant \left(\mathbb{E}|x-\nu|^k\right)^{1/k} (\mathbb{P}(x<-\tau))^{(k-1)/k} + |-\tau-\nu|\mathbb{P}(x<-\tau), \qquad (86)
\end{aligned}$$

by Holder's inequality. Also, since $-\frac{\tau}{2} \leqslant \nu \leqslant 0$, we have

$$\mathbb{P}(x<-\tau) \leqslant \mathbb{P}\left(x<\nu-\frac{\tau}{2}\right) \leqslant \mathbb{P}\left(|x-\nu|>\frac{\tau}{2}\right) \leqslant \frac{\gamma 2^k}{\tau^k}, \qquad (87)$$

via (82). Plugging (87) into (86) and using the bounded moment assumption, we get

$$\begin{aligned}
ⓐ &\leqslant \gamma^{1/k}\left(\frac{\gamma 2^k}{\tau^k}\right)^{(k-1)/k} + (\tau+|\nu|)\frac{\gamma 2^k}{\tau^k} \qquad (88)\\
&\leqslant \gamma\left(\frac{2}{\tau}\right)^{k-1} + 4\gamma\left(\frac{2}{\tau}\right)^{k-1} \qquad (89)\\
&= 5\gamma\left(\frac{2}{\tau}\right)^{k-1}. \qquad (90)
\end{aligned}$$

Likewise, a symmetric argument shows that ⓑ $\leqslant 5\gamma\left(\frac{2}{\tau}\right)^{k-1}$. Hence the lemma follows. ■

Below is a is a more precise and general version of (Kamath et al., 2022, Lemma A.2) (scaling with γ):

Lemma 58 *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ with mean ν and $\mathbb{E}|\mathcal{D}-\nu|^k \leqslant \gamma$ for some $k \geqslant 2$. Let $x_1, \cdots, x_n$ be i.i.d. samples from $\mathcal{D}$. Then, with probability at least* 0.99*,*

$$\left|\frac{1}{n}\sum_{i=1}^n x_i - \nu\right| \leqslant \frac{10\gamma^{1/k}}{\sqrt{n}}.$$

Proof First, by Jensen's inequality, we have

$$\mathbb{E}[(x-\nu)^2] \leqslant \mathbb{E}\left[|x-\nu|^k\right]^{2/k} \leqslant \gamma^{2/k}. \qquad (91)$$

Hence,

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n x_i - \nu\right)^2\right] &= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n (x_i-\nu)^2\right] \qquad (92)\\
&\leqslant \frac{\gamma^{2/k}}{n}, \qquad (93)
\end{aligned}$$

where we used the assumption that $\{x_i\}_{i=1}^n$ are i.i.d. and (91). Thus, by Chebyshev's inequality,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n x_i - \nu\right| \geqslant \frac{10\gamma^{1/k}}{\sqrt{n}}\right) \leqslant \frac{1}{100}.$$

■

Now, we provide the proof of Proposition 52:

Proof [Proof of Proposition 52] **Privacy:** Let X and X' be adjacent data sets in $\mathcal{X}^s$. Assume without loss of generality that $x_1 \neq x'_1$ and $x_l = x'_l$ for $l > 1$. By the post-processing property of DP and the fact that each sample is only processed once (due to disjoint batches) during the algorithm, it suffices to fix $i \in [m]$ and show that the composition of all d invocations of $\mathcal{P}_{1D}$ (line 8) for $j \in [d]$ is (ϵ, δ)-DP. Assume without loss of generality that $Z_j := Z_j^1$ and $Z'_j := (Z_j^1)'$ contain the truncated $z_{1,j} := \Pi_{[0,2\tau]}(x_1, j)$ and $z'_{1,j} := \Pi_{[0,2\tau]}(x'_1, j)$, respectively. Then, by the first part of Lemma 55, there are choices of b and p such that

$$D_\infty^{\delta_j}(\mathcal{S} \circ \mathcal{R}_{1D}^r(Z_j) || \mathcal{S} \circ \mathcal{R}_{1D}^r(Z'_j)) \leqslant \epsilon_j \left(\frac{2}{g} + \frac{|z_{1,j} - z'_{1,j}|}{2\tau} \right) \leqslant 2\epsilon_j,$$

for all j by the triangle inequality, provided $g \geqslant \max\{2, 2\tau\sqrt{s}\}$. Hence each invocation of $\mathcal{P}_{1D}$ is $(2\epsilon_j, \delta_j)$-DP. Thus, by the advanced composition theorem (Dwork and Roth, 2014, Theorem 3.20), Algorithm 13 is (ϵ', δ)-SDP, where

$$\epsilon' = \sum_{j=1}^d (2\epsilon_j) \left(e^{2\epsilon_j} - 1\right) + 2\sqrt{2 \sum_{j=1}^d \epsilon_j^2 \ln(1/\delta_j)} \tag{94}$$

$$\leqslant 8d\epsilon_1^2 + \frac{\epsilon}{2} \tag{95}$$

$$\leqslant \frac{\epsilon^2}{4\ln(1/\delta_1)} + \frac{\epsilon}{2} \tag{96}$$

$$\leqslant \epsilon, \tag{97}$$

by our choices of ϵ_j, δ_j and the assumption that $\epsilon \leqslant 8\ln(1/\delta_1) = 8\ln(2d/\delta)$.

Bias: Let $\hat{\nu} = (\hat{\nu}_1, \cdots, \hat{\nu}_d)$ denote the output of the non-private version of Algorithm 13 where $\mathcal{P}_{1D}$ in line 8 is replaced by the (noiseless) sum of $z \in Z_j^i$. Then, Lemma 56 tells us that

$$\|\hat{\nu} - \nu\|^2 \leqslant 200d \left(\frac{m}{s}\gamma^{2/k} + \gamma^2 \left(\frac{2}{\tau}\right)^{2k-2} \right)$$

with probability at least $1 - \zeta$, if $m = \lceil 20\log(4d/\zeta) \rceil$. Thus,

$$\begin{aligned} b^2 = \|\mathbb{E}\tilde{\nu} - \nu\|^2 = \|\mathbb{E}\hat{\nu} - \nu\|^2 \leqslant \mathbb{E}\|\hat{\nu} - \nu\|^2 \\ \leqslant 200d \left(\frac{m}{s}\gamma^{2/k} + \gamma^2 \left(\frac{2}{\tau}\right)^{2k-2} \right)(1 - \zeta) + 2\sup\left(\|\hat{\nu}\|^2 + \|\nu\|^2\right)\zeta \\ \leqslant 200d \left(\frac{m}{s}\gamma^{2/k} + \gamma^2 \left(\frac{2}{\tau}\right)^{2k-2} \right) + 4\tau^2 d\zeta, \end{aligned}$$

since $\tau \geqslant L \geqslant \|\nu\|$ by assumption. Then choosing

$$\zeta = \frac{1}{\tau^2}\left[\left(\frac{1}{s}\gamma^{2/k} + \gamma^2 \left(\frac{2}{\tau}\right)^{2k-2} \right) \right]$$

and noting that ζ is polynomial in all parameters (so that $m = \widetilde{\mathcal{O}}(1)$) implies

$$b^2 \leqslant \mathbb{E}\|\hat{\nu} - \nu\|^2 = \widetilde{\mathcal{O}}\left(d\left(\frac{1}{s}\gamma^{2/k} + \gamma^2\left(\frac{2}{\tau}\right)^{2k-2}\right)\right). \tag{98}$$

Variance: We have

$$\mathbb{E}\|N\|^2 \leqslant 2\left[\underbrace{\mathbb{E}\|\widetilde{\nu} - \hat{\nu}\|^2}_{\text{ⓐ}} + \underbrace{\mathbb{E}\|\hat{\nu} - \mathbb{E}\hat{\nu}\|^2}_{\text{ⓑ}}\right]. \tag{99}$$

We bound ⓐ as follows. For any $j \in [d], i \in [m]$, denote $\hat{\nu}_j^i := \frac{m}{s}\sum_{z \in Z_j^i} z$ (c.f. line 8 of Algorithm 13), the mean of the (noiseless) s/m clipped j-th coordinates of data in group i. Denote $\widetilde{\nu}_j^i$ as in Algorithm 13, which is the same as $\hat{\nu}_j^i$ except that the summation over Z_j^i is replaced by $\mathcal{P}_{1D}(Z_j^i)$. Also, denote $\hat{\nu}_j = \text{median}(\hat{\nu}_j^1, \cdots, \hat{\nu}_j^m)$ and $\widetilde{\nu}_j = \text{median}(\widetilde{\nu}_j^1, \cdots, \widetilde{\nu}_j^m)$. Then for any $j \in [d]$, we have

$$\begin{aligned}
|\hat{\nu}_j - \widetilde{\nu}_j|^2 &= |\text{median}(\hat{\nu}_j^1, \cdots, \hat{\nu}_j^m) - \text{median}(\widetilde{\nu}_j^1, \cdots, \widetilde{\nu}_j^m)|^2 \\
&\leqslant \left(\sum_{i=1}^m |\hat{\nu}_j^i - \widetilde{\nu}_j^i|\right)^2 \\
&= \left(\sum_{i=1}^m \left|\frac{m}{s}\left[\sum_{z \in Z_j^i} z - \mathcal{P}_{1D}(Z_j^i)\right]\right|\right)^2 \\
&\leqslant \left(\frac{m}{s}\sum_{i=1}^m \left|\mathcal{P}_{1D}(Z_j^i) - \sum_{z \in Z_j^i} z\right|\right)^2 \\
&\leqslant \frac{m^2}{s^2} m \sum_{i=1}^m \left|\mathcal{P}_{1D}(Z_j^i) - \sum_{z \in Z_j^i} z\right|^2.
\end{aligned}$$

Now using Lemma 55 (part c), we get

$$\begin{aligned}
\mathbb{E}[|\hat{\nu}_j - \widetilde{\nu}_j|^2] &\leqslant \frac{m^4}{s^2}\mathbb{E}\left[\left|\mathcal{P}_{1D}(Z_j^i) - \sum_{z \in Z_j^i} z\right|^2\right] \\
&\lesssim \frac{m^4}{s^2}\frac{\tau^2}{\epsilon_j^2}\ln(1/\delta_j) \\
&\lesssim \frac{m^4 \tau^2 d \ln^2(d/\delta)}{s^2 \epsilon^2}.
\end{aligned}$$

Thus, summing over $j \in [d]$, we get

$$\mathbb{E}[\|\hat{\nu} - \widetilde{\nu}\|^2] \lesssim \frac{m^4 \tau^2 d \ln^2(d/\delta)}{s^2 \epsilon^2}. \tag{100}$$

Next, we bound ⓑ:

$$\begin{aligned}
ⓑ &= \mathbb{E}\|\hat{\nu} - \mathbb{E}\hat{\nu}\|^2 \\
&\leqslant 2\left[\mathbb{E}\|\hat{\nu} - \nu\|^2 + \|\nu - \mathbb{E}\hat{\nu}\|^2\right] \\
&\leqslant 2\left[\mathbb{E}\|\hat{\nu} - \nu\|^2 + b^2\right] \\
&= \widetilde{\mathcal{O}}\left(d\left(\frac{1}{s}\gamma^{2/k} + \gamma^2\left(\frac{2}{\tau}\right)^{2k-2}\right)\right),
\end{aligned}$$

by the bias bound (98). Therefore,

$$\mathbb{E}\|N\|^2 = \widetilde{\mathcal{O}}\left(\frac{\tau^2 d \ln^2(d/\delta)}{s^2\epsilon^2} + d\left(\frac{1}{s}\gamma^{2/k} + \gamma^2\left(\frac{2}{\tau}\right)^{2k-2}\right)\right).$$

■