Improved Rates for Differentially Private Stochastic Convex Optimization with Heavy-Tailed Data

Gautam Kamath [*1] **Xingtu Liu** [*1] **Huanyu Zhang** [*2]

Abstract

We study stochastic convex optimization with heavy-tailed data under the constraint of differential privacy (DP). Most prior work on this problem is restricted to the case where the loss function is Lipschitz. Instead, as introduced by Wang, Xiao, Devadas, and Xu (Wang et al., 2020), we study general convex loss functions with the assumption that the distribution of gradients has bounded k-th moments. We provide improved upper bounds on the excess population risk under concentrated DP for convex and strongly convex loss functions. Along the way, we derive new algorithms for private mean estimation of heavy-tailed distributions, under both pure and concentrated DP. Finally, we prove nearly-matching lower bounds for private stochastic convex optimization with strongly convex losses and mean estimation, showing new separations between pure and concentrated DP.

1. Introduction

Stochastic convex optimization (SCO) is a classic optimization problem in machine learning. The goal is, given a loss function ℓ and a dataset $x_1, \ldots, x_n$ drawn i.i.d. from some unknown distribution $\mathcal{D}$, to output a parameter vector w which minimizes the *population risk* $L_{\mathcal{D}}(w) = \mathbb{E}_{x \sim \mathcal{D}}[\ell(w; x)]$. The quality of a solution $\hat{w}$ is measured in terms of the excess risk over the minimizer in the parameter set $\mathcal{W}$, $L_{\mathcal{D}}(\hat{w}) - \min_{w \in \mathcal{W}} L_{\mathcal{D}}(w)$. We study SCO under the constraint of *differential privacy* (Dwork et al., 2006) (DP), a rigorous notion of privacy which guarantees that an algorithm's output distribution is insensitive to modification of a small number of datapoints.

The field of DP optimization has seen a significant amount of work. Early results focused on differentially private *empirical risk minimization* (ERM), a non-statistical problem in which the goal is to privately output a parameter vector w which minimizes the loss function ℓ over a fixed dataset $x_1, \ldots, x_n$: that is we would like to optimize $\min_w \frac{1}{n} \sum_{i=1}^n \ell(w, x_i)$. See, for example, Chaudhuri & Monteleoni (2008); Chaudhuri et al. (2011); Rubinstein et al. (2012); Kifer et al. (2012); Thakurta & Smith (2013); Song et al. (2013); Jain & Thakurta (2014); Bassily et al. (2014); Talwar et al. (2015); Kasiviswanathan & Jin (2016); Wu et al. (2017); Wang et al. (2017); Iyengar et al. (2019); Wang et al. (2019); Zhang et al. (2021); Wang et al. (2021). The first result to address the statistical problem of DP SCO was Bassily et al. (2014), using generalization properties of differential privacy and regularized ERM. However, the excess risk bounds were suboptimal. Bassily et al. (2019) addressed this and closed the gap by providing tight upper bounds on DP SCO. Following this result there has been renewed interest in DP SCO, with works reducing the gradient complexity and running time (Feldman et al., 2020; Kulkarni et al., 2021), and deriving results for different geometries (Asi et al., 2021; Bassily et al., 2021).

Despite the wealth of work in this area, a significant restriction in almost all results is that the loss function is assumed to be *Lipschitz*. This assumption bounds the magnitude of each datapoint's gradient, a very convenient property for restricting the sensitivity in the design of differentially private algorithms. While convenient, it is often an unrealistic assumption which does not hold in practice, and DP optimizers resort to heuristic clipping of gradients to enforce a bound on their magnitude (Abadi et al., 2016). One can remove the strong Lipschitz assumption by instead assuming that the distribution of gradients is somehow well-behaved. In this vein, Wang et al. (2020) and Hu et al. (2022) introduce and study the problem of DP SCO with heavy-tailed data.[1] Their work removes the requirement that the loss function is Lipschitz, and instead assumes that the distribu-

[*]Equal contribution [1]Cheriton School of Computer Science, University of Waterloo [2]Meta. Correspondence to: Huanyu Zhang <huanyuzhang@fb.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]Note that the phrase heavy-tailed *data* is a slight misnomer – they actually consider a setting with heavy-tailed *gradients*. Though the two are naturally related, they are not equivalent. Despite this unfortunate mismatch, we use the same nomenclature as Wang et al. (2020) to signify that we consider the same setting as they do.

tion of the gradient has bounded second moments. However, they leave open the question of whether the rates of their algorithms can be improved.

1.1. Results

We answer this question affirmatively, giving algorithms with better rates for DP SCO with heavy-tailed gradients. Our main upper bound is the following.

Theorem 1.1 (Informal, see Theorems 5.2, 5.4, and 5.6)**.** *Suppose we have a convex and smooth loss function $\ell : \mathcal{W} \times \mathbb{R}^d \to \mathbb{R}$ and there exists a distribution $\mathcal{D}$ over $\mathbb{R}^d$ such that for any parameter vector $w \in \mathcal{W}$, when $x \sim \mathcal{D}$, the k-th moment of $\nabla \ell(w, x)$ is bounded. Then there exists a computationally efficient ε^2-concentrated differentially private algorithm which, given $x_1, \ldots, x_n \sim \mathcal{D}$, outputs a parameter vector w^{priv} satisfying the following:*

$$\mathbb{E}[L_{\mathcal{D}}(w^{priv}) - L_{\mathcal{D}}(w^*)] \leq \tilde{O}\left(\min\left\{ \frac{d}{\sqrt{n}} + \frac{d^2}{\varepsilon n} \cdot \left(\frac{\varepsilon n}{d^{3/2}}\right)^{\frac{1}{k}}, \min_{0.5 \leq q \leq 2} \left(\frac{d^{\frac{3-q}{2}}}{\sqrt{n}} + \frac{d^{\frac{1+q}{2}}}{\varepsilon^{\frac{1}{2}}\sqrt{n}} \right) \right\} \right),$$

where $w^* = \arg\min_w L_{\mathcal{D}}(w)$. Furthermore, if ℓ is strongly convex, a similar algorithm guarantees the following:

$$\mathbb{E}[L_{\mathcal{D}}(w^{priv}) - L_{\mathcal{D}}(w^*)] \leq \tilde{O}\left(\frac{d}{n} + d \cdot \left(\frac{\sqrt{d}}{\varepsilon n} \right)^{\frac{2(k-1)}{k}} \right).$$

In the first bound, q plays the role of balancing the non-private and private cost for the second term . For example, consider the extreme case when $\varepsilon = \infty$ (non-private), $q = 2$ minimizes the non-private cost in that term.

This theorem is stated under the constraint of ε^2-concentrated differential privacy, which also implies the more common notion of $(O(\varepsilon\sqrt{\log(1/\delta)}), \delta)$-differential privacy for any $\delta > 0$ (see Lemma 2.3).[2] Thus, ignoring factors which are logarithmic in $1/\delta$, the same rates in Theorem 1.1 also hold under the weaker notion of (ε, δ)-differential privacy.

Prior work on DP SCO with heavy-tailed data is due to Wang et al. (2020). Their main results are algorithms for a case with bounded second moments ($k = 2$), guaranteeing excess risk bounds of $\tilde{O}\left(\left(\frac{d^3}{\varepsilon^2 n}\right)^{1/3}\right)$ and $\tilde{O}\left(\frac{d^3}{\varepsilon^2 n}\right)$ for the convex and strongly convex cases, respectively, which our results significantly improve on.[3] Furthermore, our results are applicable to distributions with bounded moment conditions of all orders k, while Wang et al. (2020) only applies to distributions with bounded second moments ($k = 2$). Finally, while it may appear that one advantage of our upper bounds is that they hold under the stronger notion of concentrated DP, the results of Wang et al. (2020) could easily be analyzed under concentrated DP as well.

We also provide lower bounds to complement our upper bounds.

Theorem 1.2 (Informal, see Theorems 6.1 and 6.4)**.** *Let $\ell : \mathcal{W} \times \mathbb{R}^d \to \mathbb{R}$ be a convex and smooth loss function and $\mathcal{D}$ be a distribution over $\mathbb{R}^d$, such that for any parameter vector $w \in \mathcal{W}$, when $x \sim \mathcal{D}$, the k-th moment of $\nabla \ell(w, x)$ is bounded. Suppose there exists an ε^2-concentrated differentially private algorithm which is given $x_1, \ldots, x_n \sim \mathcal{D}$ and outputs a parameter vector w^{priv}. Then there exists a choice of convex and smooth loss function ℓ and distribution $\mathcal{D}$ such that*

$$\mathbb{E}[L_{\mathcal{D}}(w^{priv}) - L_{\mathcal{D}}(w^*)] \geq \Omega\left(\sqrt{\frac{d}{n}} + \sqrt{d} \cdot \left(\frac{\sqrt{d}}{\varepsilon n} \right)^{\frac{k-1}{k}} \right),$$

where $w^* = \arg\min_w L_{\mathcal{D}}(w)$. Furthermore, there exists a choice of strongly convex and smooth loss function ℓ and distribution $\mathcal{D}$ such that

$$\mathbb{E}[L_{\mathcal{D}}(w^{priv}) - L_{\mathcal{D}}(w^*)] \geq \Omega\left(\frac{d}{n} + d \cdot \left(\frac{\sqrt{d}}{\varepsilon n} \right)^{\frac{2(k-1)}{k}} \right).$$

Observe that our upper and lower bounds nearly match for the strongly convex case. For the convex case and $k = 2$, the individual terms in our upper bound match the corresponding terms in our lower bound when q is chosen to be 0.5 and 2.

Those familiar with the literature on DP SCO under a Lipschitz condition may be puzzled by the apparent discrepancy on the dimension-dependence in our results. Specifically, results of Bassily et al. (2019) (which assume that the ℓ_2 norm of gradients are bound by a constant) give an optimal rate of $\Theta(1/\sqrt{n} + \sqrt{d}/\varepsilon n)$. In contrast, if one focuses on our convex lower bound in Theorem 1.2 with $k = \infty$, we show the rate can be no better than $\Omega(\sqrt{d/n} + d/\varepsilon n)$, which is worse

[2] While it may seem unusual to write ε^2-concentrated DP, this allows for direct comparison with $(\varepsilon, 0)$-DP and (ε, δ)-DP results, two privacy notions it is intermediate to.

[3] The quoted bounds are weaker than those alleged in Wang et al. (2020). After communicating with authors of Wang et al. (2020) and Holland (2019) (which Wang et al. (2020) depends on), we confirmed an issue in the analysis of Wang et al. (2020) which leads to an underestimate of the dependence on d. In brief, if the truncation parameter s is adopted as they suggest, the dependence on d in Lemma 6 (Equation 13) and Lemma 7 (Equation 14) in the supplement of Wang et al. (2020) should be $d^{\frac{3}{2}}$ instead of d, leading to an extra multiplicative factor of $d^{\frac{1}{3}}$ in the upper bound for convex functions and a factor of d for strongly convex functions.

by a factor of $\sqrt{d}$. This discrepancy can be explained by the fact that our moment condition (Definition 2.11) bounds each *coordinate* of the gradient by a constant, leading to an overall bound on the ℓ_2-norm of the gradient by $O(\sqrt{d})$, larger than the $O(1)$ bound in the Lipschitz case by precisely this $\sqrt{d}$ factor. Thus a scaling argument is required to provide the most direct comparison. As comparison between the two settings is nuanced for $k < \infty$ and not the focus of our investigation (since, in particular, our algorithms apply in settings where Lipschitz bounds required by other works may not hold), we omit further discussion.

As one of our key tools, we introduce new algorithms and lower bounds for differentially private mean estimation for distributions with heavy tails.

Theorem 1.3 (Informal, see Corollary 4.2 and Theorem 6.3)**.** *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with $\mathbb{E}[\mathcal{D}] = \mu$ and bounded k-th moment. There exists a computationally efficient ε^2-concentrated differentially private algorithm which, given $x_1, \ldots, x_n \sim \mathcal{D}$, outputs $\hat{\mu}$ which, with probability at least 0.9, satisfies:*

$$\|\hat{\mu} - \mu\|_2 \leq \tilde{O}\left(\sqrt{\frac{d}{n}} + \sqrt{d} \cdot \left(\frac{\sqrt{d}}{\varepsilon n}\right)^{\frac{k-1}{k}}\right).$$

Furthermore, if the algorithm is required to satisfy ε-differential privacy instead of ε^2-concentrated differential privacy, the guarantee instead becomes

$$\|\hat{\mu} - \mu\|_2 \leq \tilde{O}\left(\sqrt{\frac{d}{n}} + \sqrt{d} \cdot \left(\frac{d}{\varepsilon n}\right)^{\frac{k-1}{k}}\right).$$

Finally, considering instead the expected ℓ_2 error $\mathbb{E}[\|\hat{\mu} - \mu\|_2]$, these rates can not be improved by more than poly-logarithmic factors.

Some prior works have considered private heavy-tailed mean estimation (Barber & Duchi, 2014; Kamath et al., 2020), achieving different rates than what we report here. The distinction arises in the definition of bounded moments: letting $\mathcal{D}$ be the distribution of interest, $\mathbb{E}[\mathcal{D}] = \mu$, and $X \sim \mathcal{D}$, Barber & Duchi (2014) considers distributions $\mathcal{D}$ where $\mathbb{E}[\|X - \mu\|_2^k]$ is bounded, while Kamath et al. (2020) considers distributions $\mathcal{D}$ where, for all unit vectors v, $\mathbb{E}[\langle X - \mu, v\rangle^k]$ is bounded. In contrast, we consider distributions $\mathcal{D}$ where $\mathbb{E}[\langle X - \mu, v\rangle^k]$ is bounded for all standard basis vectors v, to match the definition employed in Wang et al. (2020). We observe an interesting separation which arises under this definition. For worst-case distributions over the hypercube, it is known that the sample complexity of private mean estimation is separated by a $\sqrt{d}$ factor under pure and approximate differential privacy (Bun et al., 2014; Steinke & Ullman, 2015; Dwork et al., 2015). On the other hand, under the strong direction-wise moment bound of Kamath et al. (2020), the best known algorithms and lower bounds indicate that no such separation exists between these two notions. However, under our weaker moment bound, our results show that this same $\sqrt{d}$ separation between the sample complexity of pure and approximate differential privacy arises once again.[4] Pinpointing the precise conditions under which such a separation exists remains an interesting direction for future investigation.

Furthermore, our results are the first to derive matching upper and lower bounds for heavy-tailed mean estimation under a privacy notion other than pure DP.[5] To prove our lower bounds we introduce a concentrated DP version of Fano's inequality, building upon the pure DP version from Acharya et al. (2021). Considering the wide applicability of pure DP Fano's inequality, we believe our CDP version can also be applied to establish tight lower bounds for other statistical problems. The proof appears in Section B.10.

Theorem 1.4 (ρ-CDP Fano's inequality)**.** *Let $\mathcal{V} = \{p_1, ..., p_M\} \subseteq \mathcal{P}$ be a set of probability distributions, $\theta : \mathcal{P} \to \mathbb{R}^d$ be a parameter of interest, and $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a loss function. Suppose for all $i \neq j$, it satisfies (a) $\ell(\theta(p_i), \theta(p_j)) \geq r$, (b) $d_{\mathrm{TV}}(p_i, p_j) \leq \alpha$, (c) $d_{\mathrm{KL}}(p_i, p_j) \leq \beta$. Then for any ρ-CDP estimator $\hat{\theta}$,*

$$\frac{1}{M}\sum_{i\in[M]} \mathbb{E}_{X\sim p_i^n}\left[\ell\Big(\hat{\theta}(X), \theta(p_i)\Big)\right] \geq$$

$$\frac{r}{2}\max\left\{1 - \frac{\beta + \log 2}{\log M}, 1 - \frac{\rho\big(n^2\alpha^2 + n\alpha(1-\alpha)\big) + \log 2}{\log M}\right\}.$$

1.2. Techniques

Our upper bounds operate using a gradient-descent-based method, relying upon algorithms for private mean estimation. In particular, we instantiate an oracle which outputs an estimate of the true gradient at a point. One oracle we adopt is based on an adaption of the algorithm in Kamath et al. (2020), which addresses the problem of private mean estimation of heavy-tailed distributions. That said, for several reasons, a naive black-box application of their results are insufficient to achieve the rates in Theorem 1.1. First, the accuracy guarantees in Kamath et al. (2020) give a prescribed ℓ_2-error with high probability. While such guarantees for an oracle allow one to achieve non-trivial rates, they are far from enough. Instead, we can get better results when

[4] We actually show a stronger separation, between pure and concentrated differential privacy.

[5] While the combined results of Kamath et al. (2020) and Barber & Duchi (2014) prove quantitatively matching upper and lower bounds for a related heavy-tailed setting, they are for qualitatively different notions of privacy. Specifically, the upper bound is under the easier constraint of concentrated DP, while the lower bound is under the harder constraint of pure DP, resulting in a qualitative gap. In contrast, we prove matching upper and lower bounds for pure DP, and separate matching upper and lower bounds for concentrated DP, the latter of which is often qualitatively harder.

the estimator is known to have low bias. This is where the intersection of privacy and heavy-tailed data gives rise to a new technical challenge: no unbiased mean estimation algorithm for this setting is known to exist. To deal with these issues, we explicitly derive bounds on the bias and variance of the estimator. We must additionally switch the order of various steps in their algorithm, to achieve sharper bounds on the variance while keeping the bias unchanged. Even with these changes in place, the bound would still be lossy – as a final modification, we find that a different bias-variance tradeoff is required in each iteration to achieve the best possible error. Namely, if we tolerate additional variance to reduce the bias of each step, this results in an improved final accuracy.

2. Preliminaries

2.1. Privacy Preliminaries

In our work we consider a few different variants of differential privacy. The first is the standard notion.

Definition 2.1 (Differential Privacy (DP) (Dwork et al., 2006))**.** A randomized algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies (ε, δ)-differential privacy if for every pair of neighbouring datasets $X, X' \in \mathcal{X}^n$ (i.e., datasets that differ in exactly one entry), $\forall Y \subseteq \mathcal{Y}$, $\mathbb{P}[M(X) \in Y] \leq e^{\varepsilon} \cdot \mathbb{P}[M(X') \in Y] + \delta$. When $\delta = 0$, we say that M satisfies ε-differential privacy or pure differential privacy.

The second is *concentrated differential privacy* (Dwork & Rothblum, 2016), and its refinement *zero-concentrated differential privacy* (Bun & Steinke, 2016). Since in this work we exclusively concern ourselves with the latter, in a slight overloading of nomenclature, we refer to it more concisely as concentrated differential privacy.

Definition 2.2 (Concentrated Differential Privacy (CDP) (Bun & Steinke, 2016))**.** A randomized algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies ρ-CDP if for every pair of neighboring datasets $X, X' \in \mathcal{X}^n$, $\forall \alpha \in (1, \infty)$ $D_\alpha(M(X) \| M(X')) \leq \rho\alpha$, where $D_\alpha(M(X) \| M(X'))$ is the α-Rényi divergence [6] between $M(X)$ and $M(X')$.

Roughly speaking, CDP lives between pure $(\varepsilon, 0)$-DP and approximate (ε, δ)-DP, formalized in the following lemma.

Lemma 2.3 (Bun & Steinke (2016))**.** *For every $\varepsilon > 0$, if M is ε-DP, then M is $\left(\frac{1}{2}\varepsilon^2\right)$-CDP. If M is $\left(\frac{1}{2}\varepsilon^2\right)$-CDP, then M is $\left(\frac{1}{2}\varepsilon^2 + \varepsilon\sqrt{2\log(1/\delta)}, \delta\right)$-DP for every $\delta > 0$.*

Differential privacy enjoys adaptive composition.

[6] Let P and Q be two probability distributions defined over some probability space, the α-Rényi divergence of order $\alpha > 1$ is defined as $D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim Q}\left[\left(\frac{P(x)}{Q(x)}\right)^\alpha\right]$.

Lemma 2.4 (Composition of DP (Dwork et al., 2006; 2010; Bun & Steinke, 2016))**.** *If M is an adaptive composition of differentially private algorithms $M_1, \ldots, M_T$, then: if $M_1, \ldots, M_T$ are $(\varepsilon_1, 0), \ldots, (\varepsilon_T, 0)$-DP then M is $(\varepsilon, 0)$-DP for $\varepsilon = \sum_t \varepsilon_t$. Furthermore: if $M_1, \ldots, M_T$ are $\rho_1, \ldots, \rho_T$-CDP then M is ρ-CDP for $\rho = \sum_t \rho_t$.*

Finally, we introduce two additive noise mechanisms, which transform non-private algorithms to private algorithms.

Lemma 2.5 (Additive noise mechanisms (Dwork et al., 2006; Bun & Steinke, 2016))**.** *Let $M : \mathcal{X}^n \to \mathcal{R}^d$ be a non-private algorithm. Let $\Delta_1(M) \triangleq \max_{X \sim X'} \|M(X) - M(X')\|_1$ denote the ℓ_1-sensitivity of M, which measures the maximum change of the output in ℓ_1-norm for two neighbouring datasets $X \sim X'$. Define the ℓ_2-sensitivity $\Delta_2(M)$ analogously in terms of the ℓ_2-norm. The Laplace mechanism is the output $M(X) + N$, where $N = (N_1, \ldots, N_d)$, and $\forall j \in [d]$, with $N_j \sim Lap\left(0, \frac{\Delta_1(M)}{\varepsilon}\right)$, and is $(\varepsilon, 0)$-DP. If instead $N \sim \mathcal{N}\left(0, \frac{\Delta_2^2(M)}{2\rho}\mathbb{I}_{d \times d}\right)$, this is the Gaussian mechanism, which satisfies ρ-CDP.*

2.2. Optimization Preliminaries

We require the following standard set of optimization preliminaries.

Definition 2.6. A function $f : \mathcal{W} \to \mathbb{R}$ is l-Lipschitz if for all $w_1, w_2 \in \mathcal{W}$ we have $|f(w_1) - f(w_2)| \leq l \cdot \|w_1 - w_2\|_2$.

Definition 2.7. A function f is λ-strongly convex on $\mathcal{W}$ if for all $w_1, w_2 \in \mathcal{W}$ we have $f(w_1) \geq f(w_2) + \langle \nabla f(w_2), w_1 - w_2 \rangle + \frac{\lambda}{2}\|w_1 - w_2\|_2^2$.

Definition 2.8. A function f is L-smooth on $\mathcal{W}$ if for all $w_1, w_2 \in \mathcal{W}$, $f(w_1) \leq f(w_2) + \langle \nabla f(w_2), w_1 - w_2 \rangle + \frac{L}{2}\|w_1 - w_2\|_2^2$.

Definition 2.9. Given a convex set $\mathcal{W}$, we denote the projection of any $\theta \in \mathbb{R}^d$ to the convex set $\mathcal{W}$ by $Proj_{\mathcal{W}}(\theta) = \arg\min_{w \in \mathcal{W}} \|\theta - w\|_2$.

2.3. Problem Setup

Definition 2.10 (Stochastic Convex Optimization (SCO))**.** Let $\mathcal{D}$ be some unknown distribution over $\mathcal{X}$ and $X = \{x_1, \ldots, x_n\}$ be i.i.d. samples from $\mathcal{D}$. Given a convex constraint set $\mathcal{W} \subseteq \mathbb{R}^d$ and a convex loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, the goal of *stochastic convex optimization (SCO)* is to find a minimizer w^{priv} for the population risk $L_{\mathcal{D}}(w^{priv}) = \mathbb{E}_{x \sim \mathcal{D}}[\ell(w^{priv}, x)]$. The utility of an algorithm $\mathcal{A}$ is measured by the *expected* excess population risk

$$\mathop{\mathbb{E}}_{X \sim \mathcal{D}^n, \mathcal{A}}\left[L_{\mathcal{D}}(w^{priv}) - \min_{w \in \mathcal{W}} L_{\mathcal{D}}(w)\right].$$

We use the following coordinate-wise definition of bounded moments, identical to that of Wang et al. (2020).

Definition 2.11. Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean μ. We say that for $k \geq 2$, the k-th moment of $\mathcal{D}$ is bounded by γ, if for every $j \in [d]$, $\mathbb{E}[|\langle X - \mu, e_j \rangle|^k] \leq \gamma$, where e_j is the j-th standard basis vector.

Let $B_R(\vec{c}) \subset \mathbb{R}^d$ be the ball of radius $R > 0$ centered at $\vec{c} \in \mathbb{R}^d$. All our theorems rely on the following set of assumptions.

Assumption 2.12. We assume the following:

1. The loss function $\ell(w, x)$ is non-negative, differentiable and convex for all $w \in \mathcal{W}$ and $x \in \mathcal{X}$.

2. For any $x \in \mathcal{X}$, $\ell(w, x)$ is L-smooth on $\mathcal{W}$.

3. The constraint set $\mathcal{W}$ is bounded with diameter M.

4. The gradient of the loss function at the optimum is zero.

5. For any $w \in \mathcal{W}$, the distribution of the gradient of the loss function has bounded k-th moments for some $k \geq 2$: $\nabla \ell(w, x) \sim \mathcal{P}$ satisfies Definition 2.11 with $\gamma = 1$.

6. For any $w \in \mathcal{W}$, the distribution of the gradient has bounded mean: $\mathbb{E}[\nabla \ell(w, x)] \in B_R(\vec{0})$, where $R = O(1)$.

The first four points in Assumption 2.12 are standard when studying convex learning problems. The fifth is a relaxation of the *Lipschitz* condition in non-heavy-tailed SCO problems, in which the gradient is assumed to be uniformly bounded by a constant. While the gradient in our setting is unbounded, it is realistic to assume that the *expected* gradient is inside a ball with some radius R. In fact, packing lower bounds for private mean estimation necessitate such an assumption under most notions of DP (Karwa & Vadhan, 2018). As a direct corollary, $\|\nabla L_{\mathcal{D}}(w)\|_2 \leq R$, so $L_{\mathcal{D}}(w)$ is R-Lipschitz. We note that all these assumptions are explicitly or implicitly assumed in Wang et al. (2020).

3. A Framework for Stochastic Convex Optimization

In this section, we present a general framework for private SCO. Before diving into the details, we first provide some intuition on how we approach this problem, via the classic optimization model.

Let $L_{\mathcal{D}}(\cdot)$ be the *expected* loss function we are trying to minimize. Although the data $x \sim \mathcal{X}$, the loss function $\ell(\cdot)$, and its gradient $\nabla \ell(\cdot)$ may be heavy-tailed, $L_{\mathcal{D}}(\cdot)$ is well-behaved: specifically, Assumption 2.12 implies that it is both convex and R-Lipschitz. Therefore, if $L_{\mathcal{D}}(\cdot)$ were known, the problem would reduce to a classical convex optimization problem, solvable by gradient descent (GD). Of course, $L_{\mathcal{D}}(\cdot)$ is not known to the optimizer, and we can not directly run gradient descent. Instead, we estimate $\nabla L_{\mathcal{D}}(\cdot)$ from the samples X, incurring an additional loss based on the quality of the approximation.

Algorithm 1 SCO algorithmic framework $\text{SCOF}_{\eta,T,\text{MeanOracle}}(X)$

Input: $X = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}^d$, algorithm MeanOracle, parameters η, T
Initialize $w^0 \in \mathcal{W}$
for $t = 1, 2, \ldots, T$ **do**
  if ℓ is *convex* **then**
    Select $S_t = X$
  else if ℓ is *strongly convex* **then**
    Select $S_t = \{x_{(t-1)n/T+1}, \ldots, x_{tn/T}\}$
  end if
  $\nabla \widetilde{L}_{\mathcal{D}}(w^{t-1}) = \text{MeanOracle}(\{\nabla \ell(w^{t-1}, x_i)\}_{x_i \in S_t})$
  $w^t = \text{Proj}_{\mathcal{W}}(w^{t-1} - \eta_{t-1} \nabla \widetilde{L}_{\mathcal{D}}(w^{t-1}))$
end for
Output: $\{w^1, w^2, \ldots, w^T\}$

There are two approaches to choosing samples used for this estimate in each iteration. The first strategy is to choose the entire dataset X. This breaks independence between the different iterations, so one must argue using uniform convergence to bound the estimation error for all $w \in \mathcal{W}$ simultaneously. The second strategy is to choose disjoint samples for each iteration, which maintains independence between iterations at the cost of less data and thus more error for each iteration. We adopt the first strategy in our analysis for convex functions, and the second strategy for strongly convex functions.

Our GD-based framework, SCOF, is presented in Algorithm 1. The true gradient $\nabla L_{\mathcal{D}}(w^{t-1})$ is replaced by an estimate $\nabla \widetilde{L}_{\mathcal{D}}(w^{t-1})$ obtained by a mean estimation algorithm.

Observe that Algorithm 1 is differentially private if the mean estimator MeanOracle is differentially private, a consequence of composition and post-processing of differential privacy. In Theorems 3.1 and 3.2, we quantify the population risk of Algorithm 1 based on the accuracy of MeanOracle. Theorem 3.1 considers convex loss functions, while Theorem 3.2 achieves better rates when the loss function is strongly convex. Although the proof techniques resemble those in previous work, e.g., Agarwal et al. (2018), we include the analysis for completeness, with proofs in Appendix B.1 and Appendix B.2.

Theorem 3.1 (Convex)**.** *Suppose that MeanOracle guarantees that for any $w \in \mathcal{W}$, $\|\mathbb{E}[\nabla \widetilde{L}_{\mathcal{D}}(w)] - \nabla L_{\mathcal{D}}(w)\|_2 \leq B$ and $\mathbb{E}[\|\nabla \widetilde{L}_{\mathcal{D}}(w) - \nabla L_{\mathcal{D}}(w)\|_2^2] \leq G^2$. Under Assumption 2.12, for any $\eta > 0$ the output $w^{priv} = \frac{1}{T}\sum_{t\in[T]} w^t$ produced by SCOF satisfies*

$$\mathop{\mathbb{E}}_{X \sim \mathcal{D}^n, \mathcal{A}} \left[L_{\mathcal{D}}(w^{priv}) - L_{\mathcal{D}}(w^*) \right] \leq \frac{M^2}{2\eta T} + \eta(R^2 + G^2) + MB,$$

where $w^* = \arg\min_w L_{\mathcal{D}}(w)$.

Theorem 3.2 (Strongly convex). *Suppose that MeanOracle guarantees that, for any $w^t, t \in [T]$, $\mathbb{E}[\|\nabla \widetilde{L}_{\mathcal{D}}(w^t) - \nabla L_{\mathcal{D}}(w^t)\|_2] \le G$. Under Assumption 2.12, and the further assumption that the population risk is λ-strongly convex, if $\eta = \frac{1}{\lambda+L}$, the output $w^{priv} = w^T$ produced by SCOF satisfies*

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}^n,\mathcal{A}}[L_{\mathcal{D}}(w^{priv}) - L_{\mathcal{D}}(w^*)] \le \left(1 - \frac{\lambda L}{(\lambda+L)^2}\right)^T M + \frac{(\lambda+L)G}{\lambda L}.$$

Specifically, if $T = \log\left(\frac{(\lambda+L)G}{\lambda L}\right) / \log\left(\frac{\lambda^2+L^2+\lambda L}{(\lambda+L)^2}\right)$, the output w^{priv} satisfies

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}^n,\mathcal{A}}\left[L_{\mathcal{D}}(w^{priv}) - L_{\mathcal{D}}(w^*)\right] \le \frac{(\lambda+L)^2(M+1)^2G^2}{2\lambda^2 L},$$

where $w^* = \arg\min_w L_{\mathcal{D}}(w)$.

4. Mean Estimation Oracle

We employ an adaption of Kamath et al. (2020)'s CDPHDME algorithm, which privately estimates the mean of a heavy-tailed distribution. In order to improve upon the bounds one would obtain via a black-box application, we provide a novel analysis for an adapted version of this algorithm, which differs from theirs in two crucial aspects.

First, their analysis only applies for a specific choice of the truncation parameter (τ in Theorem 4.1), which is selected to be optimal for their one-round algorithm. However, note that our problem is different from theirs, since GD requires multiple steps instead of only one round. If we naively follow the same parameter setting as they do, we will get a loose bound on the excess risk. Therefore, we generalize their analysis to accommodate a range of values for τ to fit our needs.

Second, while their analysis provides ℓ_2-error guarantees, Theorem 3.1 necessitates bounds on the bias and variance of the estimator. We thus modify steps in their algorithm to reduce the variance (while leaving the bias untouched).

Our new analysis can be summarized by Theorem 4.1, where we provide theoretical guarantees for our estimators in both the private and non-private settings. We defer our algorithm and the proof to Appendix B.3.

Theorem 4.1. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean $\mu \in B_R\left(\vec{0}\right)$ with $R \le 10$ and k-th moment bounded by 1. For any $\tau \ge 10$ and a universal constant $C \ge 14$, there exists a polynomial-time (non-private) algorithm (Algorithm 3) that takes n samples from $\mathcal{D}$, and outputs $\widehat{\mu} \in \mathbb{R}^d$, such that with probability at least $1-\beta$,*

$$\|\hat{\mu} - \mu\|_2 = O\left(\sqrt{d}\cdot\left(\sqrt{\frac{\log\left(\frac{d}{\beta}\right)}{n}} + \left(\frac{C}{\tau}\right)^{k-1}\right)\right).$$

A ρ-CDP adaption, CDPCWME(ρ,τ), takes n samples from $\mathcal{D}$, and outputs $\widetilde{\mu} = \hat{\mu} + \mathcal{N}\left(0, \frac{1152\tau^2 d\log^2\left(\frac{2d}{\beta}\right)}{\rho n^2}\mathbb{I}_{d\times d}\right)$, where $\hat{\mu}$ is the non-private output of Algorithm 3, such that with probability at least $1-\beta$,

$$\|\widetilde{\mu} - \mu\|_2 = O\left(\sqrt{d}\cdot\left(\sqrt{\frac{\log\left(\frac{d}{\beta}\right)}{n}} + \left(\frac{C}{\tau}\right)^{k-1}\right) + \frac{\tau\log\left(\frac{d}{\beta}\right)\sqrt{d}}{\sqrt{\rho}n}\left(\sqrt{d} + \sqrt{\log\frac{1}{\beta}}\right)\right).$$

Finally, an $(\varepsilon,0)$-DP adaption, DPCWME(ε,τ), takes n samples from $\mathcal{D}$, and outputs $\widetilde{\mu}$ with $\widetilde{\mu}_j = \hat{\mu}_j + Lap\left(0, \frac{48\tau d\log\left(\frac{2d}{\beta}\right)}{\varepsilon n}\right)$, where $\hat{\mu}$ is the non-private output of Algorithm 3, such that with probability at least $1-\beta$,

$$\|\widetilde{\mu} - \mu\|_2 = O\left(\sqrt{d}\cdot\left(\sqrt{\frac{\log\left(\frac{d}{\beta}\right)}{n}} + \left(\frac{C}{\tau}\right)^{k-1}\right) + \frac{\tau d^{\frac{3}{2}}\log^2\left(\frac{d}{\beta}\right)}{\varepsilon n}\right).$$

Setting $\beta = \frac{1}{10}$, $\tau = \left(\frac{\sqrt{\rho}n}{\sqrt{d}}\right)^{\frac{1}{k}}$ for CDPCWME(ρ,τ), and $\tau = \left(\frac{\varepsilon n}{d}\right)^{\frac{1}{k}}$ for DPCWME(ε,τ), gives the following corollary.

Corollary 4.2. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean $\mu \in B_R\left(\vec{0}\right)$ with $R \le 10$ and k-th moment bounded by 1. There exists a polynomial-time ρ-CDP algorithm CDPCWME$\left(\rho, \left(\frac{\sqrt{\rho}n}{\sqrt{d}}\right)^{\frac{1}{k}}\right)$ that takes n samples from $\mathcal{D}$, and outputs $\widetilde{\mu} \in \mathbb{R}^d$, such that with probability at least 0.9,*

$$\|\widetilde{\mu} - \mu\|_2 = \widetilde{O}\left(\sqrt{\frac{d}{n}} + \sqrt{d}\cdot\left(\frac{\sqrt{d}}{\sqrt{\rho}n}\right)^{\frac{k-1}{k}}\right).$$

Furthermore, there exists a polynomial-time ε-DP algorithm DPCWME$\left(\varepsilon, \left(\frac{\varepsilon n}{d}\right)^{\frac{1}{k}}\right)$ that takes n samples from $\mathcal{D}$, and outputs $\widetilde{\mu} \in \mathbb{R}^d$, such that with probability at least 0.9,

$$\|\widetilde{\mu} - \mu\|_2 = \widetilde{O}\left(\sqrt{\frac{d}{n}} + \sqrt{d}\cdot\left(\frac{d}{\varepsilon n}\right)^{\frac{k-1}{k}}\right).$$

5. Algorithms for SCO with Heavy-Tailed Data

In this section we introduce our main algorithm for ρ-CDP SCO, Algorithm 2. We analyze utility for convex loss functions in Section 5.1, while the results for strongly convex loss functions are in Section 5.2.

Algorithm 2 CDP-SCO algorithm with heavy-tailed data

1: **Input:** $X = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}^d$, parameters η, ρ, T
2: $\{w^t\}_{t=1}^T = SCOF_{\eta,T,MeanOracle(\rho/T)}(X)$
3: **Output:** w^{priv}

Privacy is straightforward: since each of the T steps of the algorithm is ρ/T-CDP, composition of CDP (Lemma 2.4) gives the following privacy guarantee.

Lemma 5.1. *Algorithm 2 is ρ-CDP.*

5.1. Convex Setting

In this section, we consider convex and smooth loss functions. We provide accuracy guarantees for Algorithm 2 in the following two theorems, each of which instantiates our framework with a different mean estimation oracle. The proofs follow by appropriately selecting the truncation parameter τ, and balancing the bias and variance in SCOF. We defer the proofs to Appendix B.4 and Appendix B.5, respectively.

Theorem 5.2 (Convex)**.** *Suppose we have a stochastic convex optimization problem which satisfies Assumption 2.12. Assuming $R \leq 10$, $L \leq 10$, Algorithm 2, instantiated with CDPCWME with parameters $T = \frac{R^2\rho n^2}{\tau^2 d^4}$, $\eta = \frac{M}{R\sqrt{T}}$, and $\tau = \left(\frac{\sqrt{\rho}n}{Md^{\frac{3}{2}}}\right)^{\frac{1}{k}}$, outputs $w^{priv} = \frac{1}{T}\sum_{t\in[T]} w^t$, such that*

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}^n,\mathcal{A}}[L_{\mathcal{D}}(w^{priv}) - L_{\mathcal{D}}(w^*)] \leq O\left(\frac{Md}{\sqrt{n}} + \frac{Md^2}{n\sqrt{\rho}}\cdot\left(\frac{\sqrt{\rho}n}{Md^{\frac{3}{2}}}\right)^{\frac{1}{k}}\right),$$

where $w^* = \arg\min_w L_{\mathcal{D}}(w)$, and M is the diameter of the constraint set $\mathcal{W}$. The running time is $O(ndT)$[7] *.*

Remark 5.3. Our non-standard choice of the truncation parameter τ in Theorem 5.2 is crucial to obtaining our results. If one were to naïvely adopt $\tau = \left(\frac{\sqrt{\rho}n}{d^{\frac{3}{2}}\sqrt{T}}\right)^{\frac{1}{k}}$ to balance the bias and standard deviation for each iteration, we would achieve much worse bounds. Instead, in order to reduce bias we truncate far less aggressively than done in Kamath et al. (2020), which comes at the cost of increased variance. For example, considering the case when $d = 1$, if we were to use the choice of $\tau = \left(\frac{\sqrt{\rho}n}{\sqrt{T}}\right)^{\frac{1}{k}}$ for the convex case, the error would be $O\left(\frac{1}{\sqrt{T}} + \left(\frac{1}{\tau}\right)^{k-1}\right) = O\left(\frac{1}{\sqrt{T}} + \left(\frac{\sqrt{T}}{\sqrt{\rho}n}\right)^{\frac{k-1}{k}}\right)$. Fixing $T = \left(\sqrt{\rho}n\right)^{\frac{2k-2}{2k-1}}$, we obtain the bound of $O\left(\left(\frac{1}{\sqrt{\rho}n}\right)^{\frac{k-1}{2k-1}}\right)$ instead of our bound of $O\left(\left(\frac{1}{\sqrt{\rho}n}\right)^{\frac{k-1}{k}}\right)$ in Theorem 5.2. In the limit as $k \to \infty$, our bound is quadratically better.

Alternatively, one can adopt the mean estimation oracle of Holland (2019), as done by Wang et al. (2020). However, we provide a more careful analysis, resulting in a significantly improved error rate. We provide further details of the mean estimation oracle and a proof of the following theorem in Appendix B.5.

Theorem 5.4 (Convex)**.** *Suppose we have a stochastic convex optimization problem which satisfies Assumption 2.12. Assuming $R \leq 10$ and $L \leq 10$, for any $0.5 \leq q \leq 2$, Algorithm 2, instantiated with CDPNSME (Algorithm 4) with parameters $T = \frac{R^2\rho n^2}{\tau^2 d^2}$, $\eta = \frac{M}{R\sqrt{T}}$, and $\tau = \left(\frac{\sqrt{\rho}n}{Md^q}\right)^{\frac{1}{2}}$, outputs $w^{priv} = \frac{1}{T}\sum_{t\in[T]} w^t$, such that*

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}^n,\mathcal{A}}[L_{\mathcal{D}}(w^{priv}) - L_{\mathcal{D}}(w^*)] \leq O\left(\frac{\sqrt{M}d^{\frac{3-q}{2}}}{\sqrt{n}} + \frac{\sqrt{M}d^{\frac{1+q}{2}}}{\rho^{\frac{1}{4}}\sqrt{n}}\right),$$

where $w^* = \arg\min_w L_{\mathcal{D}}(w)$, and M is the diameter of the constraint set $\mathcal{W}$. The running time is $O(ndT)$.

Remark 5.5. By varying q from 0.5 to 2, Theorem 5.4 is able to achieve different balances between the non-private (first) and private (second) error terms. In practice, one should choose q which minimizes their sum, which depends on the instance parameters. When $q = 0.5$, our bound is $O\left(\frac{\sqrt{M}d^{\frac{5}{4}}}{\sqrt{n}} + \frac{\sqrt{M}d^{\frac{3}{4}}}{\rho^{\frac{1}{4}}\sqrt{n}}\right)$, where the private term matches the $k = 2$ lower bound in Theorem 6.4. Additionally, when $q = 1$ and ρ and M are constants, our bound is $O\left(\frac{d}{\sqrt{n}}\right)$, strictly improving upon $O\left(\frac{d}{n^{\frac{1}{3}}}\right)$ in Wang et al. (2020).

Combining Theorems 5.2 and 5.4 gives the convex part of Theorem 1.1.

5.2. Strongly Convex Setting

In this section, we consider strongly convex and smooth loss functions. The analysis is somewhat simpler than the convex case, due to number of iterations being only logarithmic. The proof of the following theorem is in Appendix B.6.

[7] Suppose M and R are constants, this bound is vacuous unless the loss is $O(1)$, i.e., we implicitly require n is large enough such that the denominator is larger (in order) than the numerator. The same constraint applies to the other theorems.

Theorem 5.6 (Strongly Convex)**.** *Suppose we have a stochastic convex optimization problem which satisfies Assumption 2.12, and additionally, the loss function ℓ is λ-strongly convex. Algorithm 2, instantiated with CDPCWME with parameters $T = \log\left(\frac{(\lambda+L)G}{\lambda L}\right) / \log\left(\frac{\lambda^2+L^2+\lambda L}{(\lambda+L)^2}\right)$ with $G = \widetilde{O}\left(\sqrt{\frac{d}{n}} + \sqrt{d}\cdot\left(\frac{\sqrt{d}}{\sqrt{\rho}n}\right)^{\frac{k-1}{k}}\right)$, $\eta = \frac{1}{\lambda+L}$ and $\tau = \left(\frac{\sqrt{\rho}n}{\sqrt{d}T^{\frac{3}{2}}}\right)^{1/k}$, outputs $w^{priv} = w^T$, such that*

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}^n,\mathcal{A}}[L_{\mathcal{D}}(w^{priv}) - L_{\mathcal{D}}(w^*)] \leq \frac{(M+1)^2(\lambda+L)^2}{\lambda^2 L}\cdot\tilde{O}\left(\frac{d}{n} + d\cdot\left(\frac{\sqrt{d}}{\sqrt{\rho}n}\right)^{\frac{2k-2}{k}}\right),$$

where $w^* = \arg\min_w L_{\mathcal{D}}(w)$. The running time is $O(ndT)$.

Remark 5.7. Although in Theorem 5.2 and 5.6, we provide our utility guarantees in terms of the expectation, they can be easily generalized to the high-probability setting. In Appendix B.11, we present the high-probability version of Theorem 5.2 as an example.

6. Lower Bounds for DP SCO with Heavy-Tailed Data

In this section, we present our lower bounds for ρ-CDP SCO. Our results are generally attained by reducing from mean estimation to SCO, where similar connections have been explored when proving lower bounds for DP empirical risk minimization (Bassily et al., 2014). In order to prove some of our lower bounds, we introduce a new technical tool, a CDP version of Fano's inequality (Theorem 1.4), which is of independent interest.

6.1. Strongly-Convex Loss Functions

Theorem 6.1 (Strongly convex case)**.** *Let $n, d \in \mathbb{N}$ and $\rho > 0$. There exists a strongly convex and smooth loss function $\ell : \mathcal{W}\times\mathbb{R}^d$, such that for every ρ-CDP algorithm $\mathcal{A}$ (whose output on input X is denoted by $w^{priv} = \mathcal{A}(X)$), there exists a distribution $\mathcal{D}$ on $\mathbb{R}^d$ such that $\forall w \in \mathcal{W}$, $\sup_{j\in[d]}\mathbb{E}_{x\sim\mathcal{D}}\left[\left|\langle\nabla\ell(w,x) - \mathbb{E}\left[\nabla\ell(w,x)\right], e_j\rangle\right|^k\right] \leq 1$ (e_j is the j-th standard basis vector), which satisfies*

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

where $w^* = \arg\min_w L_{\mathcal{D}}(w)$.

Proof. The following lemma shows a reduction from mean estimation to SCO. The proof is deferred to Appendix B.7.

Lemma 6.2. *Let $n, d \in \mathbb{N}$, and $\rho > 0$. There exists a strongly convex and smooth loss function $\ell : \mathcal{W}\times\mathbb{R}^d$, such that for every ρ-CDP algorithm $\mathcal{A}$ (whose output on input X is denoted by $w^{priv} = \mathcal{A}(X)$), and every distribution $\mathcal{D}$ on $\mathbb{R}^d$ with $\mathbb{E}[\mathcal{D}] = \mu$,*

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}^n,\mathcal{A}}[L_{\mathcal{D}}(w^{priv}) - L_{\mathcal{D}}(w^*)] = \mathop{\mathbb{E}}_{X\sim\mathcal{D}^n,\mathcal{A}}\left[\frac{1}{2}\|w^{priv} - \mu\|_2^2\right],$$

where $w^* = \arg\min_w L_{\mathcal{D}}(w)$.

The following lemma provides lower bounds for mean estimation, under both DP and CDP. The first term is the non-private sample complexity, and is folklore for Gaussian mean estimation. To prove the second term, we leverage our CDP version of Fano's inequality (Theorem 1.4), based on the packing of distributions employed by Barber & Duchi (2014). Detail are in Appendix B.8.

Lemma 6.3. *Let $n, d \in \mathbb{N}$ and $\rho > 0$. For every ρ-CDP algorithm $\mathcal{A}$, there exists a distribution $\mathcal{D}$ on $\mathbb{R}^d$ with $\mathbb{E}[\mathcal{D}] = \mu$ and $\sup_{j\in[d]}\mathbb{E}_{x\sim\mathcal{D}}[|\langle e_j, x-\mu\rangle|^k] \leq 1$ (e_j is the j-th standard basis vector), such that*

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}^n,\mathcal{A}}[\|\mathcal{A}(X) - \mu\|_2] \geq \Omega\left(\sqrt{\frac{d}{n}} + \sqrt{d}\cdot\min\left(1, \left(\frac{\sqrt{d}}{\sqrt{\rho}n}\right)^{\frac{k-1}{k}}\right)\right).$$

Additionally, for every $(\varepsilon, 0)$-DP algorithm $\mathcal{A}$, there exists a distribution $\mathcal{D}$ on $\mathbb{R}^d$ with $\mathbb{E}[\mathcal{D}] = \mu$ and $\sup_{j\in[d]}\mathbb{E}_{x\sim\mathcal{D}}[|\langle e_j, x-\mu\rangle|^k] \leq 1$ (e_j is the j-th standard basis vector), such that

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}^n,\mathcal{A}}[\|\mathcal{A}(X) - \mu\|_2] \geq \Omega\left(\sqrt{\frac{d}{n}} + \sqrt{d}\cdot\min\left(1, \left(\frac{d}{\varepsilon n}\right)^{\frac{k-1}{k}}\right)\right).$$

Observe that by Jensen's inequality, for ρ-CDP algorithms,

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}^n,\mathcal{A}}[\|\mathcal{A}(X) - \mu\|_2^2] \geq \Omega\left(\frac{d}{n} + d\cdot\min\left(1, \left(\frac{\sqrt{d}}{\sqrt{\rho}n}\right)^{\frac{2k-2}{k}}\right)\right).$$

Combining Lemma 6.2 and Lemma 6.3 yields Theorem 6.1. □

6.2. Convex Loss Functions

The convex case is slightly different from the strongly convex case, as it can not be reduced to mean estimation in a black-box fashion. As before, we apply our CDP version of Fano's inequality (Theorem 1.4), based on the packing of distributions employed by Barber & Duchi (2014). The proof appears in Appendix B.9.

Theorem 6.4 (Convex case)**.** *Let $n, d \in \mathbb{N}$ and $\rho > 0$. There exists a convex and smooth loss function $\ell : \mathcal{W} \times \mathbb{R}^d$, such that for every ρ-CDP algorithm $\mathcal{A}$ (whose output on input X is denoted by $w^{priv} = \mathcal{A}(X)$), there exists a distribution $\mathcal{D}$ on $\mathbb{R}^d$ such that $\forall w \in \mathcal{W}$, $\sup_{j\in[d]} \mathbb{E}_{x\sim\mathcal{D}} \left[\left| \langle \nabla \ell(w, x) - \mathbb{E}\left[\nabla \ell(w, x)\right], e_j \rangle \right|^k \right] \leq 1$ (e_j is the j-th standard basis), which satisfies*

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}^n, \mathcal{A}}[L_{\mathcal{D}}(w^{priv}) - L_{\mathcal{D}}(w^*))] \geq \Omega\left(\sqrt{\frac{d}{n}} + \sqrt{d} \cdot \min\left(1, \left(\frac{\sqrt{d}}{\sqrt{\rho} n}\right)^{\frac{k-1}{k}}\right)\right),$$

where $w^* = \arg\min_w L_{\mathcal{D}}(w)$.

7. Acknowledgements

We thank Daniel Levy, Ziteng Sun, and an anonymous reviewer for pointing out an issue in a previous version of our paper.

Gautam Kamath is supported by an NSERC Discovery Grant, an unrestricted gift from Google, and a University of Waterloo startup grant.

Xingtu Liu was supported by an NSERC Discovery Grant.

This work was partially done while Huanyu Zhang was a graduate student at Cornell University, supported by NSF #1815893.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM Conference on Computer and Communications Security*, CCS '16, pp. 308–318, New York, NY, USA, 2016. ACM.

Acharya, J., Sun, Z., and Zhang, H. Differentially private assouad, fano, and le cam. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, ALT '21, pp. 48–78. JMLR, Inc., 2021.

Agarwal, N., Suresh, A. T., Yu, F. X. X., Kumar, S., and McMahan, B. cpSGD: Communication-efficient and differentially-private distributed SGD. In *Advances in Neural Information Processing Systems 31*, NeurIPS '18, pp. 7575–7586. Curran Associates, Inc., 2018.

Asi, H., Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: Optimal rates in ℓ_1 geometry. In *International Conference on Machine Learning*, pp. 393–403. PMLR, 2021.

Barber, R. F. and Duchi, J. C. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*, 2014.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '14, pp. 464–473, Washington, DC, USA, 2014. IEEE Computer Society.

Bassily, R., Feldman, V., Talwar, K., and Thakurta, A. G. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pp. 11282–11291. Curran Associates, Inc., 2019.

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

Bubeck, S. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8 (2–3):231–357, 2015.

Bun, M. and Steinke, T. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Proceedings of the 14th Conference on Theory of Cryptography*, TCC '16-B, pp. 635–658, Berlin, Heidelberg, 2016. Springer.

Bun, M., Ullman, J., and Vadhan, S. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the 46th Annual ACM Symposium on the Theory of Computing*, STOC '14, pp. 1–10, New York, NY, USA, 2014. ACM.

Chaudhuri, K. and Monteleoni, C. Privacy-preserving logistic regression. In *Advances in Neural Information Processing Systems 21*, NIPS '08, pp. 289–296. Curran Associates, Inc., 2008.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(29):1069–1109, 2011.

Dwork, C. and Rothblum, G. N. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In

Proceedings of the 3rd Conference on Theory of Cryptography, TCC '06, pp. 265–284, Berlin, Heidelberg, 2006. Springer.

Dwork, C., Rothblum, G. N., and Vadhan, S. Boosting and differential privacy. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, FOCS '10, pp. 51–60, Washington, DC, USA, 2010. IEEE Computer Society.

Dwork, C., Smith, A., Steinke, T., Ullman, J., and Vadhan, S. Robust traceability from trace amounts. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pp. 650–669, Washington, DC, USA, 2015. IEEE Computer Society.

Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: Optimal rates in linear time. In *Proceedings of the 52nd Annual ACM Symposium on the Theory of Computing*, STOC '20, New York, NY, USA, 2020. ACM.

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

Hu, L., Ni, S., Xiao, H., and Wang, D. High dimensional differentially private stochastic optimization with heavy-tailed data. In *Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, SIGMOD/PODS '22, pp. 227–236, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392600. doi: 10.1145/3517804.3524144.

Iyengar, R., Near, J. P., Song, D., Thakkar, O., Thakurta, A., and Wang, L. Towards practical differentially private convex optimization. In *Proceedings of the 40th IEEE Symposium on Security and Privacy*, SP '19, pp. 299–316, Washington, DC, USA, 2019. IEEE Computer Society.

Jain, P. and Thakurta, A. G. (near) dimension independent risk bounds for differentially private learning. In *Proceedings of the 31st International Conference on Machine Learning*, ICML '14, pp. 476–484. JMLR, Inc., 2014.

Kamath, G., Singhal, V., and Ullman, J. Private mean estimation of heavy-tailed distributions. In *Proceedings of the 33rd Annual Conference on Learning Theory*, COLT '20, pp. 2204–2235, 2020.

Karwa, V. and Vadhan, S. Finite sample differentially private confidence intervals. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science*, ITCS '18, pp. 44:1–44:9, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Kasiviswanathan, S. P. and Jin, H. Efficient private empirical risk minimization for high-dimensional learning. In *Proceedings of the 33rd International Conference on Machine Learning*, ICML '16, pp. 488–497. JMLR, Inc., 2016.

Kifer, D., Smith, A., and Thakurta, A. Private convex empirical risk minimization and high-dimensional regression. In *Proceedings of the 25th Annual Conference on Learning Theory*, COLT '12, pp. 25.1–25.40, 2012.

Kulkarni, J., Lee, Y. T., and Liu, D. Private non-smooth erm and sco in subquadratic steps. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 4053–4064. Curran Associates, Inc., 2021.

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pp. 1302–1338, 2000.

Rubinstein, B., Bartlett, P., Huang, L., and Taft, N. Learning in a large function space: Privacy-preserving mechanisms for SVM learning. *The Journal of Privacy and Confidentiality*, 4(1):65–100, 2012.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning - From Theory to Algorithms.* Cambridge University Press, 2014.

Shamir, O. A variant of azuma's inequality for martingales with subgaussian tails. *CoRR*, abs/1110.2392, 2011.

Song, S., Chaudhuri, K., and Sarwate, A. D. Stochastic gradient descent with differentially private updates. In *Proceedings of the 2013 IEEE Global Conference on Signal and Information Processing*, GlobalSIP '13, pp. 245–248, Washington, DC, USA, 2013. IEEE Computer Society.

Steinke, T. and Ullman, J. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Proceedings of the 28th Annual Conference on Learning Theory*, COLT '15, pp. 1588–1628, 2015.

Talwar, K., Thakurta, A., and Zhang, L. Nearly-optimal private LASSO. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pp. 3025–3033. Curran Associates, Inc., 2015.

Thakurta, A. G. and Smith, A. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Proceedings of the 26th Annual Conference on Learning Theory*, COLT '13, pp. 819–850, 2013.

Wang, D., Ye, M., and Xu, J. Differentially private empirical risk minimization revisited: Faster and more general. In

Advances in Neural Information Processing Systems 30, NIPS '17, pp. 2722–2731. Curran Associates, Inc., 2017.

Wang, D., Xiao, H., Devadas, S., and Xu, J. On differentially private stochastic convex optimization with heavy-tailed data. In *Proceedings of the 37th International Conference on Machine Learning*, ICML '20, pp. 10081–10091. JMLR, Inc., 2020.

Wang, D., Zhang, H., Gaboardi, M., and Xu, J. Estimating smooth glm in non-interactive local differential privacy model with public unlabeled data. In *Algorithmic Learning Theory*, pp. 1207–1213. PMLR, 2021.

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

Wu, X., Li, F., Kumar, A., Chaudhuri, K., Jha, S., and Naughton, J. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proceedings of the 2017 ACM SIGMOD International Conference on Management of Data*, SIGMOD '17, pp. 1307–1322, New York, NY, USA, 2017. ACM.

Zhang, H., Mironov, I., and Hejazinia, M. Wide network learning with differential privacy. *arXiv preprint arXiv:2103.01294*, 2021.

A. Useful Inequalities

Lemma A.1. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ with mean μ, and k-th moment bounded by γ. Then the following holds for any $a > 1$.*

$$\mathbb{P}_{\mathbb{X}\sim\mathcal{D}}[|X-\mu| > a\gamma^{\frac{1}{k}}] \le \frac{1}{a^k}.$$

The following lemma comes from Kamath et al. (2020). We prove it here for completeness.

Lemma A.2. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ with mean μ, and k-th moment bounded by 1. Suppose $X_1, \dots, X_n$ are generated from $\mathcal{D}$, then with probability at least 0.99,*

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \le \frac{10}{\sqrt{n}}.$$

Proof. By Jensen's inequality,

$$\mathbb{E}\left[(X-\mu)^2\right] \le \mathbb{E}\left[|X-\mu|^k\right]^{\frac{2}{k}} \le 1.$$

Then

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2\right] &= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n} X_i - n\mu\right)^2\right] \\ &= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}(X_i-\mu)^2\right] \le \frac{1}{n}.\end{aligned}$$

By Chebyshev's inequality,

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \ge \frac{10}{\sqrt{n}}\right) \le \frac{1}{100}.$$

□

B. Omitted Proofs

B.1. Proof of Theorem 3.1

We let $L_{\mathcal{D}}(w^t) = \mathbb{E}_{x\sim\mathcal{D}}[\ell(w^t, x)]$. By Assumption 2.12, for all t,

$$\|\nabla L_{\mathcal{D}}(w^t)\|_2 = \left\|\nabla \mathbb{E}_{x\sim\mathcal{D}}[\ell(w^t,x)]\right\|_2 = \left\|\mathbb{E}_{x\sim\mathcal{D}}[\nabla\ell(w^t,x)]\right\|_2 \le R.$$

Let $w'^t = w^{t-1} - \eta\nabla\widetilde{L}_{\mathcal{D}}(w^{t-1})$, and w^t denotes its projection to $\mathcal{W}$. By the convexity of $L_{\mathcal{D}}(\cdot)$ (see, e.g., Section 14.1.1 in Shalev-Shwartz & Ben-David (2014)), we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{A},X\sim\mathcal{D}^n}\left[L_{\mathcal{D}}(w^{priv}) - L_{\mathcal{D}}(w^*)\right] \\ =&\mathbb{E}_{\mathcal{A},X\sim\mathcal{D}^n}\left[L_{\mathcal{D}}\left(\frac{1}{T}\sum_{t=1}^{T} w^t\right) - L_{\mathcal{D}}(w^*)\right] \\ \le&\mathbb{E}_{\mathcal{A},X\sim\mathcal{D}^n}\left[\frac{1}{T}\sum_{t=1}^{T}\left(L_{\mathcal{D}}\left(w^t\right)\right) - L_{\mathcal{D}}(w^*)\right] \qquad (1) \\ =&\mathbb{E}_{\mathcal{A},X\sim\mathcal{D}^n}\left[\frac{1}{T}\sum_{t=1}^{T}\left(L_{\mathcal{D}}\left(w^t\right) - L_{\mathcal{D}}(w^*)\right)\right] \\ \le&\mathbb{E}_{\mathcal{A},X\sim\mathcal{D}^n}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\eta}\left\langle\eta\nabla L_{\mathcal{D}}(w^t), w^t - w^*\right\rangle\right] \qquad (2)\end{aligned}$$

where (1) is by the Jensen's inequality and (2) is by the convexity of $L_{\mathcal{D}}$. Continuing the proof,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{A},X\sim\mathcal{D}^n}\left[L_{\mathcal{D}}(w^{priv})-L_{\mathcal{D}}(w^*)\right]\\
\leq&\mathop{\mathbb{E}}_{\mathcal{A},X\sim\mathcal{D}^n}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\eta}\left\langle\eta\nabla L_{\mathcal{D}}(w^t)+\eta\nabla\widetilde{L}_{\mathcal{D}}(w^t)-\eta\nabla\widetilde{L}_{\mathcal{D}}(w^t),w^t-w^*\right\rangle\right]\\
=&\mathop{\mathbb{E}}_{\mathcal{A},X\sim\mathcal{D}^n}\left[\frac{1}{T}\sum_{t=1}^{T}\left\langle\nabla L_{\mathcal{D}}(w^t)-\nabla\widetilde{L}_{\mathcal{D}}(w^t),w^t-w^*\right\rangle\right]+\mathop{\mathbb{E}}_{\mathcal{A},X\sim\mathcal{D}^n}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\eta}\left\langle\eta\nabla\widetilde{L}_{\mathcal{D}}(w^t),w^t-w^*\right\rangle\right].
\end{aligned}$$

We bound the first term, note that $\|w^t-w^*\|_2\leq M$, and $\|\mathbb{E}[\nabla\widetilde{L}_{\mathcal{D}}(w)]-\nabla L_{\mathcal{D}}(w)\|_2\leq B$,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{A},X\sim\mathcal{D}^n}\left[\frac{1}{T}\sum_{t=1}^{T}\left\langle\nabla L_{\mathcal{D}}(w^t)-\nabla\widetilde{L}_{\mathcal{D}}(w^t),w^t-w^*\right\rangle\right]\\
=&\frac{1}{T}\sum_{t=1}^{T}\left\langle\nabla L_{\mathcal{D}}(w^t)-\mathop{\mathbb{E}}_{\mathcal{A},X\sim\mathcal{D}^n}\left[\nabla\widetilde{L}_{\mathcal{D}}(w^t)\right],w^t-w^*\right\rangle\leq BM. \qquad (3)
\end{aligned}$$

Then we move to the second term.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{A},X\sim\mathcal{D}^n}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\eta}\left\langle\eta\nabla\widetilde{L}_{\mathcal{D}}(w^t),w^t-w^*\right\rangle\right]\\
=&\mathop{\mathbb{E}}_{\mathcal{A},X\sim\mathcal{D}^n}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2\eta}\left(-\left\|w^t-w^*-\eta\nabla\widetilde{L}_{\mathcal{D}}(w^t)\right\|^2+\left\|w^t-w^*\right\|^2\right)+\frac{\eta}{2}\left\|\nabla\widetilde{L}_{\mathcal{D}}(w^t)\right\|^2\right)\right] \qquad (4)\\
=&\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2\eta}\left(-\mathbb{E}\left[\left\|w'^{t+1}-w^*\right\|^2\right]+\mathbb{E}\left[\left\|w^t-w^*\right\|^2\right]\right)+\frac{\eta}{2}\cdot\mathbb{E}\left[\left\|\nabla\widetilde{L}_{\mathcal{D}}(w^t)\right\|^2\right]\right) \qquad (5)\\
\leq&\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2\eta}\left(-\mathbb{E}\left[\left\|w^{t+1}-w^*\right\|^2\right]+\mathbb{E}\left[\left\|w^t-w^*\right\|^2\right]\right)+\frac{\eta}{2}\cdot\mathbb{E}\left[\left\|\nabla\widetilde{L}_{\mathcal{D}}(w^t)\right\|^2\right]\right) \qquad (6)\\
=&\frac{1}{2\eta T}\left(-\mathbb{E}\left[\left\|w^T-w^*\right\|^2\right]+\mathbb{E}\left[\left\|w^1-w^*\right\|^2\right]\right)+\frac{\eta}{2T}\cdot\mathbb{E}\left[\sum_{t=1}^{T}\left\|\nabla\widetilde{L}_{\mathcal{D}}(w^t)\right\|^2\right] \qquad (7)\\
\leq&\frac{M^2}{2\eta T}+\frac{\eta}{2T}\cdot\mathbb{E}\left[\sum_{t=1}^{T}\left\|\nabla\widetilde{L}_{\mathcal{D}}(w^t)\right\|^2\right]. \qquad (8)
\end{aligned}$$

where (4) comes from the fact that $\forall a,b\in\mathbb{R}^d$, $\langle a,b\rangle=\frac{1}{2}\Big(\|a\|_2^2+\|b\|_2^2-\|a-b\|_2^2\Big)$, and (5) is by the updating rule, (6) comes from the fact that $\left\|w'^{t+1}-w^*\right\|_2\geq\left\|w^{t+1}-w^*\right\|_2$, and (7) is by the telescopic sum.

Finally, for all $t\in[T]$,

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla\widetilde{L}_{\mathcal{D}}(w^t)\right\|^2\right]&=\mathbb{E}\left[\left\|\nabla\widetilde{L}_{\mathcal{D}}(w^t)-\nabla L_{\mathcal{D}}(w^t)+\nabla L_{\mathcal{D}}(w^t)\right\|^2\right]\\
&\leq 2\mathbb{E}\left[\left\|\nabla\widetilde{L}_{\mathcal{D}}(w^t)-\nabla L_{\mathcal{D}}(w^t)\right\|^2\right]+2\left\|\nabla L_{\mathcal{D}}(w^t)\right\|^2\\
&\leq 2G^2+2R^2, \qquad (9)
\end{aligned}$$

where we note that $\mathbb{E}[\|\nabla\widetilde{L}_{\mathcal{D}}(w)-\nabla L_{\mathcal{D}}(w)\|_2^2]\leq G^2$, and $\|\nabla L_{\mathcal{D}}(w^t)\|^2\leq R^2$.

We conclude the proof by combining (3), (8), and (9).

B.2. Proof of Theorem 3.2

The argument is broadly similar to the proof of Theorem 5 in Wang et al. (2020), albeit with some minor modifications. Let $w'^t = w^{t-1} - \eta \nabla \widetilde{L}_{\mathcal{D}}(w^{t-1})$. Now we have

$$\begin{aligned}\|w'^t - w^*\|_2 &= \|w^{t-1} - \eta \nabla \widetilde{L}_{\mathcal{D}}(w^{t-1}) - w^*\|_2 \\ &\leq \|w^{t-1} - \eta \nabla L_{\mathcal{D}}(w^{t-1}) - w^*\|_2 + \eta \|\nabla \widetilde{L}_{\mathcal{D}}(w^{t-1}) - \nabla L_{\mathcal{D}}(w^{t-1})\|_2.\end{aligned}$$

It should be noticed that, the second term is bounded by ηG in expectation, since $\mathbb{E}[\|\nabla \widetilde{L}_{\mathcal{D}}(w^{t-1}) - \nabla L_{\mathcal{D}}(w^{t-1})\|_2] \leq G$. For the first term, by the coercivity of strongly convex functions (Lemma 3.11 in Bubeck (2015))

$$\langle w^{t-1} - w^*, \nabla L_{\mathcal{D}}(w^{t-1}) \rangle \geq \frac{\lambda L}{\lambda + L} \|w^{t-1} - w^*\|_2^2 + \frac{1}{\lambda + L} \|\nabla L_{\mathcal{D}}(w^{t-1})\|_2^2$$

and by taking $\eta = \frac{1}{\lambda + L}$ we have

$$\begin{aligned}\|w^{t-1} - \eta \nabla L_{\mathcal{D}}(w^{t-1}) - w^*\|_2^2 &= \|w^{t-1} - w^*\|_2^2 + \|\eta \nabla L_{\mathcal{D}}(w^{t-1})\|_2^2 - 2\langle w^{t-1} - w^*, \eta \nabla L_{\mathcal{D}}(w^{t-1}) \rangle \\ &\leq \left(1 - \frac{2\lambda L}{(\lambda + L)^2}\right) \|w^{t-1} - w^*\|_2^2 - \frac{1}{(\lambda + L)^2} \|\nabla L_{\mathcal{D}}(w^{t-1})\|_2^2 \\ &\leq \left(1 - \frac{2\lambda L}{(\lambda + L)^2}\right) \|w^{t-1} - w^*\|_2^2.\end{aligned}$$

Now using the inequality $\sqrt{1 - x} \leq 1 - \frac{x}{2}$ we combine two terms together to have

$$\mathbb{E}[\|w'^t - w^*\|_2] \leq \left(1 - \frac{\lambda L}{(\lambda + L)^2}\right) \mathbb{E}[\|w^{t-1} - w^*\|_2] + \frac{G}{\lambda + L}.$$

Recall that w^t is the projection of w'^t on $\mathcal{W}$, which implies $\|w^t - w^*\|_2 \leq \|w'^t - w^*\|_2$. Therefore,

$$\mathbb{E}[\|w^t - w^*\|_2] \leq \left(1 - \frac{\lambda L}{(\lambda + L)^2}\right) \mathbb{E}[\|w^{t-1} - w^*\|_2] + \frac{G}{\lambda + L}.$$

After T multiplications and simplifying the geometric series,

$$\mathbb{E}[\|w^T - w^*\|_2] \leq \left(1 - \frac{\lambda L}{(\lambda + L)^2}\right)^T M + \frac{(\lambda + L)^2}{\lambda L} \frac{G}{\lambda + L}.$$

Letting $T = \log\left(\frac{(\lambda + L)G}{\lambda L}\right) / \log\left(\frac{\lambda^2 + L^2 + \lambda L}{(\lambda + L)^2}\right)$,

$$\mathbb{E}[\|w^T - w^*\|_2] \leq \frac{(\lambda + L)(M + 1)G}{\lambda L}.$$

Since $L_{\mathcal{D}}(w)$ is L-smooth, we have

$$\mathop{\mathbb{E}}_{\mathcal{A}, X \sim \mathcal{D}^n} \left[L_{\mathcal{D}}(w^T)\right] - L_{\mathcal{D}}(w^*) \leq \frac{L}{2} \mathbb{E}[\|w^T - w^*\|_2^2] \leq \frac{(\lambda + L)^2 (M + 1)^2 G^2}{2\lambda^2 L}.$$

which concludes the proof.

B.3. Proof of Theorem 4.1

Algorithm 3 High-Dimensional Mean Estimator

Input: Samples $X = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}^d$. Parameters $0 < R < 10, \tau \geq 10$
Set parameters: $m \leftarrow 4\log(2d/\beta)$
$I = [-3\tau, 3\tau]$
for $j \leftarrow 1, ..., d$ **do**
  for $i \leftarrow 1, ..., m$ **do**
    $Z_j^i \leftarrow \left(\text{clip}(x, I) \text{ for } x \in \left(X_{(i-1)\cdot\frac{n}{m}+1}(j), \ldots, X_{i\cdot\frac{n}{m}}(j)\right)\right)$
    $\hat{\mu}_j^i \leftarrow \frac{m}{n}\sum_{x\in Z_j^i} x$
  end for
  $\hat{\mu}_j = \text{median}\left(\hat{\mu}_j^1, \ldots, \hat{\mu}_j^m\right)$
end for
Let $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_d)$
Output: $\hat{\mu}$

We adopt a coordinate analysis for the algorithm. For each coordinate, Algorithm 3 truncates each point to an interval. We first recall a lemma from Kamath et al. (2020), which quantifies the bias induced.

Lemma B.1. *[Lemma 3.1 in Kamath et al. (2020)] Let $\tau \geq 10$, and $\mathcal{D}$ be a distribution over $\mathbb{R}$ with mean μ and k-th moment bounded by 1. Suppose $x \sim \mathcal{D}$, $ce \in \mathbb{R}$ and Z is the following random variable,*

$$Z = \begin{cases} a \text{ if } x < ce - 3\tau, \\ b \text{ if } x > ce + 3\tau, \\ x \text{ if } x \in [ce - 3\tau, ce + 3\tau]. \end{cases}$$

If $\mu - ce \leq 3\tau$, then $|\mu - \mathbb{E}[Z]| \leq 3 \cdot \left(\frac{C}{\tau}\right)^{k-1}$, where $C \geq 14$ is a universal constant.

Intuitively, this lemma tells that if the heavy-tailed random variable is truncated to an interval with length 6τ and its center ce close enough to the true mean μ, the induced bias is small. With this in mind, we proceed to prove Theorem 4.1.

Proof of Theorem 4.1. We firstly show the accuracy guarantees of the non-private algorithms.

We analyze the algorithm coordinatewisely. For a fixed dimension j, let $Z_j = \text{clip}(x(j), I)$ with $x \sim \mathcal{D}$, we note that the k-th moment of Z_j is bounded by 1. Since $R \leq \tau$, by Lemma B.1,

$$|\mathbb{E}[Z_j] - \mu_j| \leq 3 \cdot \left(\frac{C}{\tau}\right)^{k-1}, \tag{10}$$

where $\mu_j = \mathbb{E}[X(j)]$.

Let $m = 4\log(2d/\beta)$. For a fixed i, Z_j^i is a combination of $\frac{n}{m}$ i.i.d. realizations of Z_j. By Lemma A.2, we have

$$\Pr\left(\left|\hat{\mu}_j^i - \mathbb{E}[Z_j]\right| \leq 10 \cdot \sqrt{\frac{m}{n}}\right) \geq 0.9.$$

Note that $\hat{\mu}_j = \text{median}\left(\hat{\mu}_j^1, \ldots, \hat{\mu}_j^m\right)$. By Hoeffding's inequality,

$$\Pr\left(|\hat{\mu}_j - \mathbb{E}[Z_j]| \geq 10 \cdot \sqrt{\frac{m}{n}}\right) \leq e^{-\frac{m}{4}}.$$

We apply the union bound to all the dimensions. Combined with (10), we get

$$\Pr\left(\|\hat{\mu} - \mu\|_2 \geq \sqrt{d} \cdot \left(10 \cdot \sqrt{\frac{m}{n}} + 3 \cdot \left(\frac{C}{\tau}\right)^{k-1}\right)\right) \leq d \cdot e^{-\frac{m}{4}} \leq \frac{\beta}{2}.$$

Next we move to private adaptions, where the key step is to bound the sensitivity of the non-private algorithm.

Fixing one dimension $j \in [d]$, for two neighboring datasets X and X', we want to show that $|\hat{\mu}_j(X) - \hat{\mu}_j(X')| \leq \frac{12\tau m}{n}$ for each j. With this in mind, ℓ_1 sensitivity of $\hat{\mu}$ is upper bounded by $\frac{12\tau md}{n}$, and the ℓ_2 sensitivity is upper bounded by $\frac{12\tau m\sqrt{d}}{n}$.

Now it suffices to bound the ℓ_∞ sensitivity. Let $\hat{\mu}_j = median(\hat{\mu}_j^1, \ldots, \hat{\mu}_j^i, \ldots, \hat{\mu}_j^m)$. Let X and X' be the two neighboring datasets which differ at one sample. Suppose m is odd, there are two cases:

1. $\hat{\mu}_j^{i^*}(X)$ is the median for dataset X, and $\hat{\mu}_j^{i^*}(X')$ is the median for dataset X'.
2. $\hat{\mu}_j^{i^*}(X)$ is the median for dataset X, while $\hat{\mu}_j^{i'}(X')$ is the median for dataset X'.

For the first case, $\left|\hat{\mu}_j^{i^*}(X) - \hat{\mu}_j^{i^*}(X')\right| \leq \frac{12\tau m}{n}$, since X and X' differ at one sample. For the second case, note that it can only happen when $\left|\hat{\mu}_j^{i^*}(X) - \hat{\mu}_j^{i'}(X)\right| \leq \frac{6\tau m}{n}$. Furthermore, we have $\left|\hat{\mu}_j^{i'}(X) - \hat{\mu}_j^{i'}(X')\right| \leq \frac{6\tau m}{n}$. By triangle inequality, $\left|\hat{\mu}_j^{i^*}(X) - \hat{\mu}_j^{i'}(X')\right| \leq \frac{12\tau m}{n}$, which provides an upper bound of the ℓ_∞ sensitivity. The case when m is even is similar and omitted.[8]

For CDP adaption, by the guarantee of Gaussian mechanism (Lemma 2.5), the algorithm satisfies ρ-CDP when the noise added is $\mathcal{N}\left(0, \frac{72\tau^2 m^2 d}{\rho n^2}\mathbb{I}_{d\times d}\right)$.

Besides, since $N \sim \mathcal{N}\left(0, \sigma^2\mathbb{I}_{d\times d}\right)$, where $\sigma^2 = \frac{72\tau^2 m^2 d}{\rho n^2}$. By the tail property of chi-squared distribution (Laurent & Massart, 2000),

$$\Pr\left(\|N\|_2 \geq 2\sigma\left(\sqrt{d} + \sqrt{\log\left(\frac{1}{\beta}\right)}\right)\right) \leq \frac{\beta}{2}.$$

Note that $\|\widetilde{\mu} - \mu\|_2 \leq \|N\|_2 + \|\hat{\mu} - \mu\|_2$, we conclude the proof by the union bound.

With respect to DP adaption, by the guarantee of Laplace mechanism (Lemma 2.5), the algorithm satisfies ε-DP when the noise added is $\mathrm{Lap}\left(0, \frac{12\tau md}{\varepsilon n}\right)$ for each dimension.

Besides, let $N_j \sim \mathrm{Lap}\left(0, \frac{12\tau md}{\varepsilon n}\right)$, by the tail property of Laplace distribution,

$$\Pr\left(|N_j| \geq \frac{48\tau d}{\varepsilon n} \cdot \log^2(2d/\beta)\right) \leq \frac{\beta}{2d}.$$

By union bound, with probability at least $1 - \frac{\beta}{2}$,

$$\Pr\left(\|N\|_2 \geq \frac{48\tau d^{\frac{3}{2}}}{\varepsilon n} \cdot \log^2(2d/\beta)\right) \leq \frac{\beta}{2}.$$

Note that $\|\widetilde{\mu} - \mu\|_2 \leq \|N\|_2 + \|\hat{\mu} - \mu\|_2$, we conclude the proof by applying the union bound.

□

B.4. Proof of Theorem 5.2

Lemma B.2. *Consider Algorithm 1 instantiated with CDPCWME$\left(\frac{\rho}{T}, \tau\right)$ as MeanOracle (Algorithm 3). Under Assumption 2.12 and further assuming $R \leq 10$, $L \leq 10$, when $\tau \geq 10$, the following holds for all $w \in \mathcal{W}$ simultaneously:*

$$\|\mathbb{E}[\nabla\widetilde{L}_{\mathcal{D}}(w)] - \nabla L_{\mathcal{D}}(w)\|_2 \leq \widetilde{O}\left(\frac{d}{\sqrt{n}} + \sqrt{d} \cdot \left(\frac{C}{\tau}\right)^{k-1}\right).$$

[8] To facilitate the sensitivity analysis, let $\{x_i'\}_{i=1}^n$ be the ordered set of $\{x_i\}_{i=1}^n$. If n is even, we define the median to be $\frac{1}{2}\left(x'_{\frac{n}{2}} + x'_{\frac{n}{2}+1}\right)$ rather than an arbitrary value ranging from $x'_{\frac{n}{2}}$ to $x'_{\frac{n}{2}+1}$.

and

$$\mathbb{E}[\|\nabla\widetilde{L}_{\mathcal{D}}(w)-\nabla L_{\mathcal{D}}(w)\|_2^2]\le\widetilde{O}\left(\frac{\tau^2d^4T}{\rho n^2}+\frac{d^2}{n}+d\cdot\left(\frac{C}{\tau}\right)^{2k-2}\right).$$

where $\nabla\widetilde{L}_{\mathcal{D}}(w)$ is the estimated gradient in Algorithm 1.

Proof. We start with bounding the bias. First we note that $\mathbb{E}[\nabla\widetilde{L}_{\mathcal{D}}(w)|\nabla\hat{L}_{\mathcal{D}}(w)]=\nabla\hat{L}_{\mathcal{D}}(w)$, which denotes the output of the non-private algorithm.

In order to obtain bounds that hold uniformly over the choice of w, we follow a standard strategy of covering. Note that the number of balls of radius α required to cover $\mathcal{W}$ is bounded as $N_\alpha\le\left(\frac{M}{\alpha}\right)^d$. Let $\mathcal{W}_\alpha=\{\widetilde{w}_1,\dots,\widetilde{w}_{N_\alpha}\}$ denote the centers of this covering. For an arbitrary $w\in W$, and $\widetilde{w}\in\mathcal{W}_\alpha$,

$$\left\|\nabla\hat{L}_{\mathcal{D}}(w)-\nabla L_{\mathcal{D}}(w)\right\|_2\le\left\|\nabla\hat{L}_{\mathcal{D}}(w)-\nabla\hat{L}_{\mathcal{D}}(\widetilde{w})\right\|_2+\left\|\nabla\hat{L}_{\mathcal{D}}(\widetilde{w})-\nabla L_{\mathcal{D}}(\widetilde{w})\right\|_2+\left\|\nabla L_{\mathcal{D}}(\widetilde{w})-\nabla L_{\mathcal{D}}(w)\right\|_2.$$

We bound each term separately.

For the first term, we need to analyze how much the output of the non-private estimator $\nabla\hat{L}_{\mathcal{D}}(\cdot)$ changes, when the input switches from w to $\widetilde{w}$.

Let $\beta=\left(\frac{\alpha}{M}\right)^{2d}$, and $m=4\log(2d/\beta)=8d\log\left(\frac{3M}{\alpha}\right)$. According to the smoothness assumption, for each dimension j, and batch $i\in[m]$, the average of each batch differs by no more than $L\alpha$. Therefore, for each dimension j, the median differs by no more than $Lm\alpha$. Summing over all the dimensions,

$$\left\|\nabla\hat{L}_{\mathcal{D}}(w)-\nabla\hat{L}_{\mathcal{D}}(\widetilde{w})\right\|_2\le Lm\alpha\cdot\sqrt{d}.$$

For the second term, let $\beta=\left(\frac{\alpha}{M}\right)^{2d}$. According to Theorem 4.1, with probability at least $1-\beta$,

$$\left\|\nabla\hat{L}_{\mathcal{D}}(\widetilde{w})-\nabla L_{\mathcal{D}}(\widetilde{w})\right\|_2\le C'\left(\sqrt{d}\cdot\left(\sqrt{\frac{\log\left(\frac{d}{\beta}\right)}{n}}+\left(\frac{C}{\tau}\right)^{k-1}\right)\right),$$

where C' is a universal constant.

Note that $\beta\cdot N_\alpha\le\left(\frac{\alpha}{M}\right)^d$. By union bound, with probability at least $1-\left(\frac{\alpha}{M}\right)^d$, for all $\widetilde{w}\in\mathcal{W}_\alpha$,

$$\left\|\nabla\hat{L}_{\mathcal{D}}(\widetilde{w})-\nabla L_{\mathcal{D}}(\widetilde{w})\right\|_2\le C'\left(\sqrt{d}\cdot\left(\sqrt{\frac{\log\left(\frac{d}{\beta}\right)}{n}}+\left(\frac{C}{\tau}\right)^{k-1}\right)\right),$$

Note that $\left\|\nabla\hat{L}_{\mathcal{D}}(\widetilde{w})-\nabla L_{\mathcal{D}}(\widetilde{w})\right\|_2\le 2R$ for sure. Taking expectation, we have

$$\mathbb{E}\left[\left\|\nabla\hat{L}_{\mathcal{D}}(\widetilde{w})-\nabla L_{\mathcal{D}}(\widetilde{w})\right\|_2\right]\le C'\left(\sqrt{d}\cdot\left(\sqrt{\frac{\log\left(\frac{d}{\beta}\right)}{n}}+\left(\frac{C}{\tau}\right)^{k-1}\right)\right)+2R\cdot\left(\frac{\alpha}{M}\right)^d.$$

For the third term, according to the smoothness assumption,

$$\left\|\nabla L_{\mathcal{D}}(\hat{w})-\nabla L_{\mathcal{D}}(w)\right\|_2\le L\alpha.$$

Summing up all three terms, we have

$$\mathbb{E}\left[\left\|\nabla \hat{L}_{\mathcal{D}}(\widetilde{w}) - \nabla L_{\mathcal{D}}(\widetilde{w})\right\|_2\right] \leq L\alpha(m\sqrt{d}+1) + C'\left(\sqrt{d}\cdot\left(\sqrt{\frac{\log\left(\frac{d}{\beta}\right)}{n}} + \left(\frac{C}{\tau}\right)^{k-1}\right)\right) + 2R\cdot\left(\frac{\alpha}{M}\right)^d.$$

Finally taking $\alpha = \frac{1}{n^3}$, and note that $\mathbb{E}[\nabla \widetilde{L}_{\mathcal{D}}(w)] = \mathbb{E}\left[\mathbb{E}[\nabla \widetilde{L}_{\mathcal{D}}(w)]|\nabla \hat{L}_{\mathcal{D}}(w)\right] = \mathbb{E}[\nabla \hat{L}_{\mathcal{D}}(w)]$,

$$\begin{aligned}\|\mathbb{E}[\nabla \widetilde{L}_{\mathcal{D}}(w)] - \nabla L_{\mathcal{D}}(w)\|_2 &= \|\mathbb{E}[\nabla \hat{L}_{\mathcal{D}}(w)] - \nabla L_{\mathcal{D}}(w)\|_2 \\ &\leq \mathbb{E}\left[\left\|\nabla \hat{L}_{\mathcal{D}}(\widetilde{w}) - \nabla L_{\mathcal{D}}(\widetilde{w})\right\|_2\right] \\ &\leq \widetilde{O}\left(\frac{Ld^{1.5}}{n^3} + \frac{d}{\sqrt{n}} + \sqrt{d}\cdot\left(\frac{C}{\tau}\right)^{k-1} + \frac{R}{n^{3d}}\right). \end{aligned} \tag{11}$$

Assuming L, R are constants,

$$\|\mathbb{E}[\nabla \widetilde{L}_{\mathcal{D}}(w)] - \nabla L_{\mathcal{D}}(w)\|_2 \leq \widetilde{O}\left(\frac{d}{\sqrt{n}} + \sqrt{d}\cdot\left(\frac{C}{\tau}\right)^{k-1}\right).$$

Next we move to the variance. Note that

$$\mathbb{E}[\|\nabla \widetilde{L}_{\mathcal{D}}(w) - \nabla L_{\mathcal{D}}(w)\|_2^2] \leq 2\mathbb{E}[\|\nabla \widetilde{L}_{\mathcal{D}}(w) - \nabla \hat{L}_{\mathcal{D}}(w)\|_2^2] + 2\mathbb{E}[\|\nabla \hat{L}_{\mathcal{D}}(w) - \nabla L_{\mathcal{D}}(w)\|_2^2].$$

As designed in Theorem 4.1, we notice that

$$\nabla \hat{L}_{\mathcal{D}}(w^t) - \nabla \widetilde{L}_{\mathcal{D}}(w^t) = N_t \sim \mathcal{N}(0, \sigma \mathbb{I}_{d\times d}),$$

where $\sigma^2 = \frac{72\tau^2 m^2 dT}{\rho n^2}$.

Thus, $\mathbb{E}[\|\nabla \widetilde{L}_{\mathcal{D}}(w) - \nabla \hat{L}_{\mathcal{D}}(w)\|_2^2] = \frac{72\tau^2 m^2 d^2 T}{\rho n^2}$, and by (11),

$$\mathbb{E}[\|\nabla \widetilde{L}_{\mathcal{D}}(w) - \nabla L_{\mathcal{D}}(w)\|_2^2] \leq \widetilde{O}\left(\frac{\tau^2 d^4 T}{\rho n^2} + \frac{L^2 d^3}{n^6} + \frac{d^2}{n} + d\cdot\left(\frac{C}{\tau}\right)^{2k-2} + \left(\frac{R}{n^{3d}}\right)^2\right).$$

Assuming L, R are constants,

$$\mathbb{E}[\|\nabla \widetilde{L}_{\mathcal{D}}(w) - \nabla L_{\mathcal{D}}(w)\|_2^2] \leq \widetilde{O}\left(\frac{\tau^2 d^4 T}{\rho n^2} + \frac{d^2}{n} + d\cdot\left(\frac{C}{\tau}\right)^{2k-2}\right).$$

□

The proof now follows by choosing the right η and T in Theorem 3.1. To balance the first two terms, $\frac{M^2}{\eta T} + \frac{\eta}{2}R^2$, in Theorem 3.1, we let $\eta = \frac{M}{R\sqrt{T}}$.

Suppose $T = \frac{R^2\rho n^2}{\tau^2 d^4}$, with $\tau = \left(\frac{\sqrt{\rho}n}{Md^{\frac{3}{2}}}\right)^{\frac{1}{k}}$ we have

$$\frac{M^2}{\eta T} = \frac{\eta}{2}R^2 = O\left(\sqrt{d}\cdot\left(\frac{Md^{\frac{3}{2}}}{\sqrt{\rho}n}\right)^{\frac{k-1}{k}}\right).$$

Besides,

$$BM = \widetilde{O}\left(\frac{Md}{\sqrt{n}} + M\sqrt{d}\cdot\left(\frac{Md^{\frac{3}{2}}}{\sqrt{\rho}n}\right)^{\frac{k-1}{k}}\right).$$

Finally,

$$\eta G^2 = \widetilde{O}\left(\frac{d^2}{n} + d\cdot\left(\frac{Md^{\frac{3}{2}}}{\sqrt{\rho}n}\right)^{\frac{2k-2}{k}} + M\sqrt{d}\cdot\left(\frac{Md^{\frac{3}{2}}}{\sqrt{\rho}n}\right)^{\frac{k-1}{k}}\right).$$

Putting the various terms together completes the proof.

B.5. Proof of Theorem 5.4

We first introduced the mean estimation oracle in Holland (2019). For $x \in \mathbb{R}$, let

$$\phi(x) = \begin{cases} x - \frac{x^3}{6}, & -\sqrt{2} \le x \le \sqrt{2}, \\ \frac{2\sqrt{2}}{3}, & x > \sqrt{2}, \\ -\frac{2\sqrt{2}}{3}, & x < -\sqrt{2}. \end{cases}$$

Algorithm 4 CDP Noise Smoothing Mean Estimator (Holland, 2019; Wang et al., 2020)

Input: Samples $X = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}^d$. Parameters $\rho, \tau \ge 10$
Let $p = \mathcal{N}(0, c)$, and $N \sim p$, where c is a constant
for $j \leftarrow 1, ..., d$ **do**
  $\hat{\mu}_j = \frac{\tau}{n}\sum_{i=1}^n \int_{-\infty}^{\infty} \phi\Big(\frac{x_i(j)(1+N)}{\tau}\Big) dp(N)$
  Let $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_d)$
end for
Output: $\hat{\mu} + \mathcal{N}\Big(0, \frac{\tau^2 d}{\rho n^2}\cdot \mathbb{I}_{d\times d}\Big)$

Remark B.3. This estimator can be efficiently computed. Please refer to Holland (2019); Wang et al. (2020) for more detail.

It is not hard to see this algorithm satisfies ρ-CDP. We note that $\forall x, |\phi(x)| \le 1$, so the ℓ_2 sensitivity $\Delta_2(\hat{\mu}) \le \sqrt{d}$. We conclude the proof by applying Lemma 2.5.

We provide the accuracy guarantee of this algorithm in the following lemma.

Lemma B.4. *Consider Algorithm 1 instantiated with CDPNSME$\big(\frac{\rho}{T}, \tau\big)$ as MeanOracle (Algorithm 4). Under Assumption 2.12 and further assuming $R \le 10$, $L \le 10$, when $\tau \ge 10$, the following holds for all $w \in \mathcal{W}$ simulatenously:*

$$\|\mathbb{E}[\nabla\widetilde{L}_{\mathcal{D}}(w)] - \nabla L_{\mathcal{D}}(w)\|_2 \le \widetilde{O}\left(\frac{d^{\frac{3}{2}}\tau}{n} + \frac{\sqrt{d}}{\tau}\right).$$

and

$$\mathbb{E}[\|\nabla\widetilde{L}_{\mathcal{D}}(w) - \nabla L_{\mathcal{D}}(w)\|_2^2] \le \widetilde{O}\left(\frac{d^3\tau^2}{n^2} + \frac{d}{\tau^2} + \frac{\tau^2 d^2 T}{\rho n^2}\right).$$

where $\nabla\widetilde{L}_{\mathcal{D}}(w)$ is the estimated gradient in Algorithm 1.

Proof. This bias analysis directly comes from combining Remark 3 and Lemma 4 in Holland (2019). In fact, this analysis can be viewed as Lemma 5 of Holland (2019) with an explicit analysis on τ. [9]

[9] One may wonder why our result is different with Lemma 5 in Holland (2019) when setting $\tau = \sqrt{\frac{n}{d}}$. After communicating with authors of Holland (2019), we confirmed there was an issue in their Lemma 5, where their s_j (equivalent with our τ) should be $\sqrt{\frac{n}{d}}$ instead of $\sqrt{n}$.

With respect to the variance analysis,

$$\mathbb{E}[\|\nabla\widetilde{L}_{\mathcal{D}}(w)-\nabla L_{\mathcal{D}}(w)\|_2^2]\leq 2\mathbb{E}[\|\nabla\widetilde{L}_{\mathcal{D}}(w)-\nabla\hat{L}_{\mathcal{D}}(w)\|_2^2]+2\mathbb{E}[\|\nabla\hat{L}_{\mathcal{D}}(w)-\nabla L_{\mathcal{D}}(w)\|_2^2].$$

Note that the noise added is generated from $N\sim\mathcal{N}\Big(0,\frac{\tau^2 dT}{\rho n^2}\mathbb{I}_{d\times d}\Big)$. We conclude the proof by summing over all the dimensions.

□

The proof now follows by choosing the right η and T in Theorem 3.1. To balance the first two terms, $\frac{M^2}{\eta T}+\frac{\eta}{2}R^2$, in Theorem 3.1, we let $\eta=\frac{M}{R\sqrt{T}}$.

Suppose $T=\frac{R^2\rho n^2}{\tau^2 d^2}$, with $\tau=\left(\frac{\sqrt{\rho}n}{Md^q}\right)^{\frac{1}{2}}$ we have

$$\frac{M^2}{\eta T}=\frac{\eta}{2}R^2=O\left(\frac{\sqrt{M}d^{1-\frac{q}{2}}}{\rho^{\frac{1}{4}}\sqrt{n}}\right).$$

Besides,

$$BM=\widetilde{O}\left(\frac{Md^{\frac{3}{2}}}{n}\cdot\left(\frac{\sqrt{\rho}n}{Md^q}\right)^{\frac{1}{2}}+M\sqrt{d}\cdot\left(\frac{\sqrt{\rho}n}{Md^q}\right)^{-\frac{1}{2}}\right)=\widetilde{O}\left(\frac{\sqrt{M}d^{\frac{3-q}{2}}}{\sqrt{n}}+\frac{\sqrt{M}d^{\frac{1+q}{2}}}{\rho^{\frac{1}{4}}\sqrt{n}}\right).$$

Finally,

$$\eta G^2=\widetilde{O}\left(\frac{d^{3-q}}{n}+\frac{Md^{1+q}}{\sqrt{\rho}n}+\frac{\sqrt{M}d^{1-\frac{q}{2}}}{\rho^{\frac{1}{4}}\sqrt{n}}\right).$$

Note that when $0.5\leq q\leq 2$, $\frac{1+q}{2}\geq 1-\frac{q}{2}$, putting the various terms together completes the proof.

B.6. Proof of Theorem 5.6

In the strongly convex setting, for each iteration, the input of the MeanOracle is disjoint and independent, with size $\frac{n}{T}$. Therefore, there is no need to adopt the strategy of covering.

Note that Corollary 4.2 immediately guarantees the following accuracy when CDPCWME is instantiated as MeanOracle in SCOF:

Lemma B.5. *Consider Algorithm 1 instantiated with CDPCWME*$\left(\frac{\rho}{T},\left(\frac{\sqrt{\rho}n}{\sqrt{d}T^{\frac{3}{2}}}\right)^{1/k}\right)$ *as MeanOracle. Under Assumption 2.12, the following holds for all* $w^t, t\in[T]$*:*

$$\mathbb{E}[\|\nabla\widetilde{L}_{\mathcal{D}}(w^t)-\nabla L_{\mathcal{D}}(w^t)\|_2]\leq\widetilde{O}\left(\sqrt{\frac{Td}{n}}+\sqrt{d}\cdot\left(\frac{\sqrt{d}T^{\frac{3}{2}}}{\sqrt{\rho}n}\right)^{\frac{k-1}{k}}\right),$$

where $\nabla\widetilde{L}_{\mathcal{D}}(w^t)$ *is the estimated gradient in Algorithm 1.*

Note that T is poly-logarithmic on n and d. The proof follows by Theorem 3.2 immediately.

B.7. Proof of Lemma 6.2

Let $x \sim \mathcal{D}$, and $\ell(w;x) = \frac{1}{2}\|w-x\|_2^2$. Note that $w^* = \arg\min L_{\mathcal{D}}(w) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[x] = \mu$. Further using the expansion $\|a-b\|_2^2 = \|a\|_2^2 - 2\langle a,b\rangle + \|b\|_2^2$,

$$\begin{aligned}
L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) &= \frac{1}{2}\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\|w-x\|_2^2 - \|w^*-x\|_2^2] \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\|w\|_2^2 - 2\langle w,x\rangle + \|x\|_2^2 - \|w^*\|_2^2 + 2\langle w^*,x\rangle - \|x\|_2^2\right] \\
&= \frac{1}{2}\left(\|w\|_2^2 - 2\langle w,w^*\rangle - \|w^*\|_2^2 + 2\langle w^*,w^*\rangle\right) \\
&= \frac{1}{2}\left(\|w\|_2^2 - 2\langle w,w^*\rangle + \|w^*\|_2^2\right) \\
&= \frac{1}{2}\|w-w^*\|_2^2
\end{aligned}$$

Notice that ℓ is both strongly convex and smooth and the expected risk of w^{priv} is

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}^n,\mathcal{A}}[L_{\mathcal{D}}(w^{priv})] - L_{\mathcal{D}}(w^*) = \mathop{\mathbb{E}}_{X\sim\mathcal{D}^n,\mathcal{A}}\left[\frac{1}{2}\|w^{priv}-\mu\|_2^2\right],$$

which implies the result.

Now we prove the second half, note that $\nabla\ell(w,x) = w-x$, and $\mathbb{E}\left[\nabla\ell(w,x)\right] = w-\mu$,

$$\begin{aligned}
&\sup_{j\in[d]} \mathbb{E}_{x\sim\mathcal{D}}\left[\left|\langle e_j, \nabla\ell(w,x) - \mathbb{E}\left[\nabla\ell(w,x)\right]\rangle\right|^k\right] \\
&= \sup_{j\in[d]} \mathbb{E}_{x\sim\mathcal{D}}\left[\left|\langle e_j, w-x-(w-\mu)\rangle\right|^k\right] \\
&= \sup_{j\in[d]} \mathbb{E}_{x\sim\mathcal{D}}\left[\left|\langle e_j, x-\mu\rangle\right|^k\right] \le 1.
\end{aligned}$$

B.8. Proof of Lemma 6.3

We first prove the private term (the second term) in Lemma 6.3 for $(\varepsilon,0)$-DP.

We adopt the packing set defined in the proof of Proposition 4 in Barber & Duchi (2014). Given $\nu\in\mathcal{V}$, with $\|\nu\|_1 = \frac{d}{2}$, and $\nu\in\{\pm1\}^d$, let $Q_\nu = (1-p)P_0 + pP_\nu$ for some $p\in[0,1]$, where P_0 is a point mass on $\{D=0\}$ and P_ν is a point mass on $\{D = p^{-1/k}\nu\}$.

Given Q_ν, we define $\mu_\nu\in\mathbb{R}^d$ to be the mean of Q_ν, i.e., $\mu_\nu = \mathbb{E}_{x\sim Q_\nu}[x]$.

As a corollary of standard Gilbert-Varshamov bound for constant-weight codes (e.g., see Lemma 6 in Acharya et al. (2021)), there exists a set $\mathcal{V}$ such that

- The cardinality of $\mathcal{V}$ satisfies $|\mathcal{V}| \ge 2^{\frac{d}{8}}$.
- For all $\nu\in\mathcal{V}$, $\nu\in\{\pm1\}^d$ with $\|\nu\|_1 = \frac{d}{2}$.
- For all $\nu_1,\nu_2\in\mathcal{V}$, $d_{\mathrm{Ham}}(\nu_1,\nu_2) \ge \frac{d}{8}$.

We first compute the norm of μ_ν. Note that $\forall\nu\in\mathcal{V}$, $\|\mu_\nu\|_2$ is the same, which is denoted by $\|\mu\|_2$.

$$\|\mu_\nu\|_2 = \|\mathbb{E}_{x\sim Q_\nu}[x]\|_2 = p^{\frac{k-1}{k}}\cdot\sqrt{\frac{d}{2}} \coloneqq \|\mu\|_2 .$$

Let $x \sim Q_\nu$, and e_j denote the j-th standard basis.

$$\sup_{j\in[d]} \mathbb{E}_{x\sim Q_\nu}\left[|\langle (x-\mu_\nu), e_j\rangle|^k\right] \le p\cdot\left(p^{-1/k}\right)^k = 1.$$

Now we are able to bound the error.

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}^n,\mathcal{A}}[\|\mathcal{A}(X)-\mu\|_2] \ge \frac{1}{|\mathcal{V}|}\sum_{\nu\in\mathcal{V}} \mathbb{E}_{X\sim Q_\nu^n}\left[\|\mu^{priv}(X)-\mu_\nu\|_2\right],$$

which comes from the fact that the worst case loss is no smaller than the average loss.

Note that $|\mathcal{V}| \ge 2^{\frac{d}{8}}$. Furthermore, $\forall \nu \ne \nu'$, $\|\mu_\nu - \mu_{\nu'}\|_2 \ge \frac{1}{2}\|\mu\|_2$; $d_{\mathrm{TV}}(Q_\nu, Q_{\nu'}) = p$, indicating that there exists a coupling between Q_ν and $Q_{\nu'}$ with a coupling distance np. Suppose $p = \min\left(1, \frac{d}{n\varepsilon}\right)$, by DP Fano's inequality (Theorem 2 in Acharya et al. (2021)), it can be shown that

$$\frac{1}{|\mathcal{V}|}\sum_{\nu\in\mathcal{V}} \mathbb{E}_{X\sim Q_\nu^n}\left[\|\mu^{priv}(X)-\mu_\nu\|_2\right] = \Omega\left(\min\left(1, \left(\frac{d}{\varepsilon n}\right)^{\frac{k-1}{k}}\right)\cdot\sqrt{d}\right).$$

With respect to ρ-CDP algorithms, we just take $p = \min\left(1, \frac{\sqrt{d}}{n\sqrt{\rho}}\right)$, by CDP Fano's inequality (Theorem 1.4), it can be shown that

$$\frac{1}{|\mathcal{V}|}\sum_{\nu\in\mathcal{V}} \mathbb{E}_{X\sim Q_\nu^n}\left[\|\mu^{priv}(X)-\mu_\nu\|_2\right] = \Omega\left(\min\left(1, \left(\frac{\sqrt{d}}{\sqrt{\rho}n}\right)^{\frac{k-1}{k}}\right)\cdot\sqrt{d}\right).$$

We conclude the proof by noting that the non-private term (the first term) in Lemma 6.3 comes from classical Gaussian mean estimation, and $\forall a, b_1, b_2$, $a \ge 0.5(b_1+b_2)$ if $a \ge \max(b_1, b_2)$.

Remark B.6. The previous analysis implicitly assumes the strongly convex parameter $\lambda = 1$. To see the dependency on λ, we let the loss function $\ell(w;x) = \frac{\lambda}{2}\|w-x\|_2^2$ instead. Meanwhile, to keep the k-th moment bounded by 1, we have to shrink the parameter space of x by λ. Therefore, the $\|w^{priv}-\mu\|_2$ gets scaled by $\frac{1}{\lambda}$, and the final loss gets scaled by $\lambda\cdot\frac{1}{\lambda^2} = \frac{1}{\lambda}$. We note that this dependency matches with our upper bound when $\lambda = L$, which is the smoothness parameter.

B.9. Proof of Theorem 6.4

Theorem B.7 (Convex case)**.** *Let $n, d \in \mathbb{N}$. There exists a convex and smooth loss function $\ell : \mathcal{W}\times\mathbb{R}^d$, such that for every $(\varepsilon, 0)$-DP algorithm (whose output on input X is denoted by $w^{priv} = \mathcal{A}(X)$), there exists a distribution $\mathcal{D}$ on $\mathbb{R}^d$ with $\forall w$, $\sup_{j\in[d]}\mathbb{E}_{x\sim\mathcal{D}}\left[|\langle\nabla\ell(w,x) - \mathbb{E}\left[\nabla\ell(w,x)\right], e_j\rangle|^k\right] \le 1$ (e_j is the j-th standard basis), which satisfies*

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}^n,\mathcal{A}}[L_\mathcal{D}(w^{priv}) - L_\mathcal{D}(w^*))] \ge \sqrt{\frac{d}{n}} + \Omega\left(\min\left(1, \left(\frac{d}{\varepsilon n}\right)^{\frac{k-1}{k}}\right)\cdot\sqrt{d}\right),$$

where $w^* = \arg\min_w L_\mathcal{D}(w)$.

With respect to ρ-CDP algorithms, the lower bound turns to

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}^n,\mathcal{A}}[L_\mathcal{D}(w^{priv}) - L_\mathcal{D}(w^*))] \ge \sqrt{\frac{d}{n}} + \Omega\left(\min\left(1, \left(\frac{\sqrt{d}}{\sqrt{\rho}n}\right)^{\frac{k-1}{k}}\right)\cdot\sqrt{d}\right),$$

We first prove the private term (the second term) in Theorem 6.4.

Similarly, we adopt the packing set defined in the proof of Proposition 4 in Barber & Duchi (2014). Given $\nu \in \mathcal{V}$, with $\|\nu\|_1 = \frac{d}{2}$, and $\nu \in \{\pm 1\}^d$, let $Q_\nu = (1-p)P_0 + pP_\nu$ for some $p \in [0,1]$, where P_0 is a point mass on $\{D = 0\}$ and P_ν is a point mass on $\{D = p^{-1/k}\nu\}$.

Given Q_ν, we define $\mu_\nu \in \mathbb{R}^d$ to be the mean of Q_ν, i.e., $\mu_\nu = \mathbb{E}_{x\sim Q_\nu}[x]$. Additionally, we define w_ν to be its normalization, i.e., $w_\nu = \frac{\mu_\nu}{\|\mu_\nu\|_2}$. Note that w_ν is in the same direction as μ_ν, with $\|w_\nu\|_2 = 1$.

As a corollary of standard Gilbert-Varshamov bound for constant-weight codes (e.g., see Lemma 6 in Acharya et al. (2021)), there exists a set $\mathcal{V}$ such that

- The cardinality of $\mathcal{V}$ satisfies $|\mathcal{V}| \geq 2^{\frac{d}{8}}$.
- For all $\nu \in \mathcal{V}$, $\nu \in \{\pm 1\}^d$ with $|\nu| = \frac{d}{2}$.
- For all $\nu_1, \nu_2 \in \mathcal{V}$, $d_{\mathrm{Ham}}(\nu_1, \nu_2) \geq \frac{d}{8}$.

We first compute the norm of μ_ν. Note that $\forall \nu \in \mathcal{V}$, $\|\mu_\nu\|_2$ is the same, which is denoted by $\|\mu\|_2$.

$$\|\mu_\nu\|_2 = \|\mathbb{E}_{x\sim Q_\nu}[x]\|_2 = p^{\frac{k-1}{k}} \cdot \sqrt{\frac{d}{2}} \coloneqq \|\mu\|_2 .$$

Without loss of generality, we assume the parameter space $\|W\|_2 = 1$, which is a unit ball. Then we define the loss function $\ell(w; x)$. Given $\nu \in \mathcal{V}$, and $x \sim Q_\nu$, we let

$$\ell(w; x) = -\langle w, x\rangle,$$

and

$$L_{Q_\nu}(w) = \mathop{\mathbb{E}}_{x\sim Q_\nu}[\ell(w; x)] = -\langle w, \mu_\nu\rangle.$$

Note that ℓ is both convex and smooth. Let $x \sim Q_\nu$. Note that $\nabla\ell(w, x) = -x$, and $\mathbb{E}\left[\nabla\ell(w, x)\right] = -\mu_\nu$,

$$\begin{aligned}
&\sup_{j\in[d]} \mathbb{E}_{x\sim Q_\nu}\left[\left|\nabla_j\ell(w, x) - \mathbb{E}\left[\nabla_j\ell(w, x)\right]\right|^k\right] \\
&= \sup_{j\in[d]} \mathbb{E}_{x\sim Q_\nu}\left[\left|-x_j + \mu_{\nu,j}\right|^k\right] \\
&\leq p \cdot \left(p^{-1/k}\right)^k = 1.
\end{aligned}$$

Now we are able to bound the error of SCO.

$$\begin{aligned}
\mathbb{E}\left[L_{\mathcal{D}}(w^{priv}) - \min_{\hat{w}\in\mathcal{W}} L_{\mathcal{D}}(\hat{w})\right] &\geq \frac{1}{|\mathcal{V}|}\sum_{\nu\in\mathcal{V}} \mathbb{E}\left[L_{Q_\nu}(w^{priv}) - \min_{\hat{w}\in\mathcal{W}} L_{Q_\nu}(\hat{w})\right] &(12)\\
&\geq \frac{1}{|\mathcal{V}|}\sum_{\nu\in\mathcal{V}} \mathbb{E}\left[\left\langle \frac{\mu_\nu}{\|\mu\|_2}, \mu_\nu\right\rangle - \left\langle w^{priv}, \mu_\nu\right\rangle\right] &(13)\\
&= \frac{1}{|\mathcal{V}|}\sum_{\nu\in\mathcal{V}} \mathbb{E}\left[\|\mu\|_2 - \left\langle w^{priv}, \mu_\nu\right\rangle\right] \\
&\geq \frac{1}{|\mathcal{V}|}\sum_{\nu\in\mathcal{V}} \mathbb{E}\left[\frac{1}{2}\cdot\|\mu\|_2 \cdot \|w^{priv} - w_\nu\|_2^2\right], &(14)
\end{aligned}$$

where (12) comes from the fact that the worst case loss is no smaller than the average loss, (13) comes from $w_\nu = \mathop{\mathrm{argmin}}_{\hat{w}\in\mathcal{W}} L_{Q_\nu}(\hat{w})$, and (14) comes from the fact that $\left\|w^{priv}\right\|_2 \leq 1$, and $\|w_\nu\|_2 \leq 1$.

Note that $|\mathcal{V}| \geq 2^{\frac{d}{8}}$. Furthermore, $\forall \nu \neq \nu'$, $\|w_\nu - w_{\nu'}\|_2 = \Omega(1)$; $d_{\mathrm{TV}}(w_\nu, w_{\nu'}) = p$, indicating that there exists a coupling between w_ν and $w_{\nu'}$ with a coupling distance np. Suppose $p = \min\left(1, \frac{d}{n\varepsilon}\right)$, by DP Fano's inequality (Theorem 2 in Acharya et al. (2021)), it can be shown that

$$\frac{1}{|\mathcal{V}|}\sum_{\nu\in\mathcal{V}} \mathbb{E}\left[\|w^{priv} - w_\nu\|_2^2\right] = \Omega(1).$$

Thus,

$$\mathbb{E}\left[L_{\mathcal{D}}(w^{priv}) - \min_{\hat{w}\in\mathcal{W}} L_{\mathcal{D}}(\hat{w})\right] \geq \Omega(1)\cdot\|\mu\|_2 = \Omega\left(\min\left(1, \left(\frac{d}{\varepsilon n}\right)^{\frac{k-1}{k}}\right)\cdot\sqrt{d}\right). \tag{15}$$

With respect to ρ-CDP algorithms, we just take $p = \min\left(1, \frac{\sqrt{d}}{n\sqrt{\rho}}\right)$, and replace DP Fano's inequality by CDP Fano's inequality, then all the proof follows.

Now we prove the first term. We generally follow the lower bound proof of estimating Gaussians (Acharya et al., 2021). Given $\nu \in \{0,1\}^d$, we define $Q_\nu = \mathcal{N}(\mu_\nu, \mathbb{I}_d)$, where $\mu_\nu = \frac{p}{\sqrt{d}}\cdot\nu$, for some $p \in [0,1]$. Similarly, we define $w_\nu = \frac{\mu_\nu}{\|\mu_\nu\|_2}$.

As a standard Gilbert-Varshamov bound for constant-weight codes (e.g., see Lemma 6 in Acharya et al. (2021)), there exists a set $\mathcal{V}$ with cardinality at least $\|\mathcal{V}\|_2 \geq 2^{\frac{d}{8}}$, with $\|\nu\|_1 = \frac{d}{2}$ for all $\nu \in \mathcal{V}$, and with $d_{\text{Ham}}(\nu,\nu') \geq \frac{d}{2}$ for all $\nu \neq \nu' \in \mathcal{V}$.

Suppose $p = \min\left(1, \sqrt{\frac{d}{n}}\right)$, we can compute the norm of the distribution mean. Note that $\|\nu\|_1 = \frac{d}{2}$,

$$\|\mu_\nu\|_2 = \frac{\sqrt{2}}{2}\min\left(1, \sqrt{\frac{d}{n}}\right) \coloneqq \|\mu\|_2 .$$

By a similar argument with the private case, it can be shown that

$$\mathbb{E}\left[L_{\mathcal{D}}(w^{priv}) - \min_{\hat{w}\in\mathcal{W}} L_{\mathcal{D}}(\hat{w})\right] \geq \frac{\|\mu\|_2}{8}\cdot\frac{1}{|\mathcal{V}|}\cdot\sum_{\nu\in\mathcal{V}}\mathbb{E}\left[\|\hat{w}^{priv} - w_\nu\|_2^2\right],$$

where $\hat{w}^{priv} \coloneqq \arg\min_{\nu\in\mathcal{V}} \left\|w_\nu - w^{priv}\right\|_2$.

Note that this is indeed a multi-way classification problem, where w_ν's are well-separated. By classical Fano's inequality,

$$\frac{1}{|\mathcal{V}|}\sum_{\nu\in\mathcal{V}}\mathbb{E}\left[\|\hat{w}^{priv} - w_\nu\|_2^2\right] = \Omega(1).$$

Thus,

$$\mathbb{E}\left[L_{\mathcal{D}}(w^{priv}) - \min_{\hat{w}\in\mathcal{W}} L_{\mathcal{D}}(\hat{w})\right] \geq \Omega(1)\cdot\|\mu\|_2 = \Omega\left(\min\left(1, \sqrt{\frac{d}{n}}\right)\right). \tag{16}$$

Combining (15) and (16), and note that $\forall a, b_1, b_2$, $a \geq 0.5(b_1+b_2)$ if $a \geq \max(b_1, b_2)$, we conclude the proof.

B.10. Proof of Theorem 1.4

We note that the first term comes from classical Fano's inequality. So it is enough to prove the second term.

Let i^* be a random variable uniformly sampled over $[M]$. Given i^*, we generate n i.i.d. samples $X \sim p_{i^*}$. Note that the distribution of X is a mixture of M distributions. Specifically, for any event S,

$$\Pr(X \in S) = \frac{1}{M}\sum_{i\in M}\Pr_{X\sim p_i^n}(X \in S).$$

Letting $X^i \sim p_i^n$, and $\hat{p}(X)$ be a classifier mapping from samples to the underlying distribution. For the mutual information $I(X, \hat{p}(X))$ between X and $\hat{p}(X)$,

$$\begin{aligned} I(X,\hat{p}(X)) &= \mathbb{E}_{x\sim X}\left[d_{\text{KL}}(\hat{p}(x), \hat{p}(X))\right] \\ &= \frac{1}{M}\sum_{i\in M}\mathbb{E}_{x\sim p_i^n}\left[d_{\text{KL}}(\hat{p}(x), \hat{p}(X))\right], \end{aligned}$$

where the first equation comes from the definition of the mutual information:

$$I(X,Y) = \mathbb{E}_X\left[d_{\mathrm{KL}}(Y|X,Y)\right].$$

By convexity of the KL divergence,

$$\begin{aligned} d_{\mathrm{KL}}(\hat{p}(x),\hat{p}(X)) &\le \frac{1}{M}\sum_{j\in M} d_{\mathrm{KL}}\big(\hat{p}(x),\hat{p}(X^j)\big) \\ &\le \frac{1}{M}\sum_{j\in M} \mathbb{E}_{x'\sim p_j^n}\left[d_{\mathrm{KL}}(\hat{p}(x),\hat{p}(x'))\right]. \end{aligned}$$

Therefore,

$$I(X,\hat{p}(X)) \le \frac{1}{M^2}\sum_{i\in M}\sum_{j\in M} \mathbb{E}_{x\sim p_i^n}\left[\mathbb{E}_{x'\sim p_j^n}\left[d_{\mathrm{KL}}(\hat{p}(x),\hat{p}(x'))\right]\right].$$

By the group privacy property of CDP (Proposition 1.9 in Bun & Steinke (2016)), which says that if $\hat{p}$ is ρ-CDP, $d_{\mathrm{KL}}(\hat{p}(x),\hat{p}(x')) \le \rho\cdot d_{\mathrm{Ham}}(x,x')^2$. Therefore, we have

$$I(X,\hat{p}(X)) \le \frac{\rho}{M^2}\sum_{i\in M}\sum_{j\in M} \mathbb{E}_{x\sim p_i^n}\left[\mathbb{E}_{x'\sim p_j^n}\left[d_{\mathrm{Ham}}(x,x')^2\right]\right].$$

Note that the TV distance between each pair of distributions is upper bounded by α. By the property of optimal coupling, there exists a coupling such that $\Pr_{z\sim p_i, z'\sim p_j}(z\neq z') = \alpha$. Therefore, $d_{\mathrm{Ham}}(x,x') \sim Bin(n,\alpha)$, and

$$\mathbb{E}_{x\leftarrow p_i^n}\left[\mathbb{E}_{x'\leftarrow p_j^n}\left[d_{\mathrm{Ham}}(x,x')^2\right]\right] \le n^2\alpha^2 + n\alpha(1-\alpha).$$

By Fano's inequality, let $p_e = \frac{1}{M}\sum_{i\in[M]}\Pr_{X\sim p_i^n}(\hat{p}(X)\neq p_i)$,

$$I(i^*,\hat{p}(X)) \ge (1-p_e)\log M - \log 2.$$

Noting that $I(i^*,\hat{p}(X)) \le I(X,\hat{p}(X))$, combining inequalities shows that

$$p_e \ge 1 - \frac{\rho\big(n^2\alpha^2 + n\alpha(1-\alpha)\big) + \log 2}{\log M}. \tag{17}$$

Finally, let $\hat{p}(X) \coloneqq \arg\min_{i\in M}\ell\Big(\hat{\theta}(X),\theta(p_i)\Big)$. By triangle inequality,

$$\ell(\theta(p_{i^*}),\theta(\hat{p}(X))) \le \ell\Big(\hat{\theta}(X),\theta(p_{i^*}))\Big) + \ell\Big(\hat{\theta}(X),\theta(\hat{p}(X))\Big) \le 2\ell\Big(\hat{\theta}(X),\theta(p_{i^*})\Big).$$

Therefore,

$$\begin{aligned} \frac{1}{M}\sum_{i\in[M]}\mathbb{E}_{X\sim p_i^n}\left[\ell\Big(\hat{\theta}(X),\theta(p_i)\Big)\right] &\ge \frac{1}{2M}\sum_{i\in[M]}\mathbb{E}_{X\sim p_i^n}\left[\ell(\theta(\hat{p}(X)),\theta(p_i))\right] \\ &\ge \frac{r}{2M}\sum_{i\in[M]}\Pr_{X\sim p_i^n}(\hat{p}(X)\neq p_i) \\ &= \frac{rp_e}{2}. \end{aligned}$$

Combined with (17), we conclude the proof.

B.11. Theorem 5.2 with High-probability Guarantees

In this paper, we provide all our utility guarantees in terms of the expectation over the randomness of samples and algorithms. However, they can be easily generalized to the high-probability setting. In this section, we present the high-probability version of Theorem 5.2 as an example.

Theorem B.8 (Theorem 5.2 in high probability)**.** *Suppose we have a stochastic convex optimization problem which satisfies Assumption 2.12. Assuming $R \le 10$, $L \le 10$, Algorithm 2, instantiated with CDPCWME with parameters $T = \frac{R^2\rho n^2}{\tau^2 d^4}$, $\eta = \frac{M}{R\sqrt{T}}$, and $\tau = \left(\frac{\sqrt{\rho}n}{Md^{\frac{3}{2}}}\right)^{\frac{1}{k}}$, outputs $w^{priv} = \frac{1}{T}\sum_{t\in[T]} w^t$, such that with probability at least $1-\beta$,*

$$L_{\mathcal{D}}(w^{priv}) - L_{\mathcal{D}}(w^*) \le O\left(\frac{Md\sqrt{\log\frac{Mdn}{\beta}}}{\sqrt{n}} + \log\left(\frac{Mdn}{\beta}\right)\cdot\left(\frac{Md^2}{n\sqrt{\rho}}\cdot\left(\frac{\sqrt{\rho}n}{Md^{\frac{3}{2}}}\right)^{\frac{1}{k}} + \frac{MLd^{1.5}}{n^3}\right)\right),$$

where $w^* = \arg\min_w L_{\mathcal{D}}(w)$, and M is the diameter of the constraint set $\mathcal{W}$.

Proof. Let $w'^t = w^{t-1} - \eta\nabla\widetilde{L}_{\mathcal{D}}(w^{t-1})$, and w^t denotes its projection to $\mathcal{W}$. Similar with the proof of Theorem 3.1,

$$\begin{aligned}
&L_{\mathcal{D}}(w^{priv}) - L_{\mathcal{D}}(w^*)\\
\le&\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\eta}\left\langle \eta\nabla L_{\mathcal{D}}(w^t), w^t - w^*\right\rangle\\
\le&\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\eta}\left\langle \eta\nabla L_{\mathcal{D}}(w^t) + \eta\nabla\widetilde{L}_{\mathcal{D}}(w^t) - \eta\nabla\widetilde{L}_{\mathcal{D}}(w^t), w^t - w^*\right\rangle\\
=&\frac{1}{T}\sum_{t=1}^{T}\left\langle \nabla L_{\mathcal{D}}(w^t) - \nabla\widetilde{L}_{\mathcal{D}}(w^t), w^t - w^*\right\rangle + \frac{1}{T}\sum_{t=1}^{T}\frac{1}{\eta}\left\langle \eta\nabla\widetilde{L}_{\mathcal{D}}(w^t), w^t - w^*\right\rangle\\
=&\text{LHS} + \text{RHS}.
\end{aligned}$$

We first bound the RHS, which corresponds to analyzing the variance. Let $w'^t = w^{t-1} - \eta\nabla\widetilde{L}_{\mathcal{D}}(w^{t-1})$, and w^t denotes its projection to $\mathcal{W}$. Similar with the proof of Theorem 3.1,

$$\begin{aligned}
\text{RHS} =&\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\eta}\left\langle \eta\nabla\widetilde{L}_{\mathcal{D}}(w^t), w^t - w^*\right\rangle\\
=&\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2\eta}\left(-\left\|w^t - w^* - \eta\nabla\widetilde{L}_{\mathcal{D}}(w^t)\right\|^2 + \left\|w^t - w^*\right\|^2\right) + \frac{\eta}{2}\left\|\nabla\widetilde{L}_{\mathcal{D}}(w^t)\right\|^2\right)\\
=&\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2\eta}\left(-\left\|w'^{t+1} - w^*\right\|^2 + \left\|w^t - w^*\right\|^2\right) + \frac{\eta}{2}\cdot\left\|\nabla\widetilde{L}_{\mathcal{D}}(w^t)\right\|^2\right)\\
\le&\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2\eta}\left(-\left\|w^{t+1} - w^*\right\|^2 + \left\|w^t - w^*\right\|^2\right) + \frac{\eta}{2}\cdot\left\|\nabla\widetilde{L}_{\mathcal{D}}(w^t)\right\|^2\right)\\
=&\frac{1}{2\eta T}\left(-\left\|w^T - w^*\right\|^2 + \left\|w^1 - w^*\right\|^2\right) + \frac{\eta}{2T}\cdot\sum_{t=1}^{T}\left\|\nabla\widetilde{L}_{\mathcal{D}}(w^t)\right\|^2\\
=&\frac{1}{2\eta T}\left(-\left\|w^T - w^*\right\|^2 + \left\|w^1 - w^*\right\|^2\right) + \frac{\eta}{2T}\cdot\sum_{t=1}^{T}\left\|\nabla\widetilde{L}_{\mathcal{D}}(w^t) - \nabla L_{\mathcal{D}}(w^t) + \nabla L_{\mathcal{D}}(w^t)\right\|^2\\
\le&\frac{1}{2\eta T}\left(-\left\|w^T - w^*\right\|^2 + \left\|w^1 - w^*\right\|^2\right) + \frac{\eta}{T}\cdot\sum_{t=1}^{T}\left(\left\|\nabla\widetilde{L}_{\mathcal{D}}(w^t) - \nabla L_{\mathcal{D}}(w^t)\right\|^2 + \left\|\nabla L_{\mathcal{D}}(w^t)\right\|^2\right).
\end{aligned}$$

By Assumption 2.12, we have

$$\|\nabla L_{\mathcal{D}}(w^t)\|^2 = \left\|\nabla \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\ell(w^t,x)]\right\|^2 = \left\|\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\nabla\ell(w^t,x)]\right\|^2 \le R^2$$

for all t, and $\|w' - w^*\|^2 \le M^2$ for any $w' \in \mathcal{W}$.

Thus,

$$\begin{aligned}\text{RHS} &\le \frac{M^2}{2\eta T} + \eta R^2 + \frac{\eta}{T}\sum_{t=1}^{T}\left(\left\|\nabla\widetilde{L}_{\mathcal{D}}(w^t) - \nabla L_{\mathcal{D}}(w^t)\right\|^2\right).\\ &\le \frac{M^2}{2\eta T} + \eta R^2 + \frac{2\eta}{T}\sum_{t=1}^{T}\left(\left\|\nabla\widetilde{L}_{\mathcal{D}}(w^t) - \nabla\hat{L}_{\mathcal{D}}(w^t)\right\|^2 + \left\|\nabla\hat{L}_{\mathcal{D}}(w^t) - \nabla L_{\mathcal{D}}(w^t)\right\|^2\right).\end{aligned}$$

Following a similar proof with the covering argument in the proof of Theorem 5.2, we can bound $\left\|\nabla\hat{L}_{\mathcal{D}}(w^t) - \nabla L_{\mathcal{D}}(w^t)\right\|^2$ for all $t \in [T]$ simulatenously. Specifically, replacing $\alpha = \frac{1}{n^3}$ and $m = d\log\frac{dMn}{\beta}$, we can show that with probability $1 - \frac{\beta}{10}$, for all $w \in \mathcal{W}$,

$$\left\|\nabla\hat{L}_{\mathcal{D}}(w) - \nabla L_{\mathcal{D}}(w)\right\| \le O\left(\sqrt{d}\cdot\left(\frac{C}{\tau}\right)^{k-1} + d\cdot\sqrt{\frac{\log\frac{Mdn}{\beta}}{n}} + \frac{Ld^{1.5}\log\frac{Mdn}{\beta}}{n^3}\right). \tag{18}$$

Then by Gaussian tail bound and the union bound, with probability $1 - \frac{\beta}{10}$, for all w_t with $t \in [T]$,

$$\left\|\nabla\hat{L}_{\mathcal{D}}(w) - \nabla\widetilde{L}_{\mathcal{D}}(w)\right\| \le O\left(\frac{\tau d^2\sqrt{T\log\frac{MndT}{\beta}}}{n\sqrt{\rho}}\right),$$

Combining the previous two equations, with probability at least $1 - \frac{\beta}{5}$,

$$\begin{aligned}\frac{\eta}{T}\sum_{t=1}^{T}\left\|\nabla L_{\mathcal{D}}(w^t) - \nabla\hat{L}_{\mathcal{D}}(w^t)\right\|^2 &\le O\left(\eta\cdot\left(d\cdot\left(\frac{C}{\tau}\right)^{2k-2} + \frac{d^2\log\frac{Mdn}{\beta}}{n} + \frac{L^2d^3\log^2\frac{Mdn}{\beta}}{n^6}\right)\right),\\ \frac{\eta}{T}\sum_{t=1}^{T}\left\|\nabla\widetilde{L}_{\mathcal{D}}(w^t) - \nabla\hat{L}_{\mathcal{D}}(w^t)\right\|^2 &\le O\left(\eta\cdot\frac{\tau^2d^4T\cdot\log\frac{MndT}{\beta}}{n^2\rho}\right).\end{aligned} \tag{19}$$

Next we bound the LHS, which corresponds to analyzing the bias. By the triangle inequality,

$$\begin{aligned}\text{LHS} &= \frac{1}{T}\sum_{t=1}^{T}\left\langle\nabla L_{\mathcal{D}}(w^t) - \nabla\widetilde{L}_{\mathcal{D}}(w^t), w^t - w^*\right\rangle\\ &\le \frac{1}{T}\sum_{t=1}^{T}\left\langle\nabla L_{\mathcal{D}}(w^t) - \nabla\hat{L}_{\mathcal{D}}(w^t), w^t - w^*\right\rangle + \frac{1}{T}\sum_{t=1}^{T}\left\langle\nabla\hat{L}_{\mathcal{D}}(w^t) - \nabla\widetilde{L}_{\mathcal{D}}(w^t), w^t - w^*\right\rangle.\end{aligned}$$

By (18), we have

$$\left\|\nabla\hat{L}_{\mathcal{D}}(w) - \nabla L_{\mathcal{D}}(w)\right\| \le O\left(\sqrt{d}\cdot\left(\frac{C}{\tau}\right)^{k-1} + d\cdot\sqrt{\frac{\log\frac{Mdn}{\beta}}{n}} + \frac{Ld^{1.5}\log\frac{Mdn}{\beta}}{n^3}\right).$$

Note that $\|w^t - w^*\|_2 \leq M$,

$$\frac{1}{T}\sum_{t=1}^{T}\left\langle \nabla L_{\mathcal{D}}(w^t) - \nabla \hat{L}_{\mathcal{D}}(w^t), w^t - w^* \right\rangle \leq M \cdot O\left(\sqrt{d}\cdot\left(\frac{C}{\tau}\right)^{k-1} + d\cdot\sqrt{\frac{\log\frac{Mdn}{\beta}}{n}} + \frac{Ld^{1.5}\log\frac{Mdn}{\beta}}{n^3}\right).$$

Until now, the proof is almost the same with the case under expectation. Lastly, we analyze the term of $\nabla \hat{L}_{\mathcal{D}}(w^t) - \nabla \widetilde{L}_{\mathcal{D}}(w^t)$. Note that this term is new, since it is zero when taking expectation. As designed in Theorem 4.1, we notice that

$$\nabla \hat{L}_{\mathcal{D}}(w^t) - \nabla \widetilde{L}_{\mathcal{D}}(w^t) = N_t \sim \mathcal{N}(0, \sigma \mathbb{I}_{d\times d}),$$

where $\sigma^2 = \frac{72\tau^2 d^3 T \log^2\frac{Mnd}{\beta}}{\rho n^2}$. Note that N_t is independent of $w^t - w^*$, with $\mathbb{E}[\langle N_t, w^t - w^*\rangle | w^t] = 0$. Therefore, $\{\langle N_t, w^t - w^*\rangle, w^t\}_{t=0}^{T}$ is a martingale difference sequence. By Azuma's inequality for sub-Gaussian distributions (see Theorem 2 in (Shamir, 2011)),

$$\frac{1}{T}\sum_{t=1}^{T}\left\langle \nabla \hat{L}_{\mathcal{D}}(w^t) - \nabla \widetilde{L}_{\mathcal{D}}(w^t), w^t - w^* \right\rangle \leq O\left(\frac{M\sqrt{d}\sigma\log(1/\beta)}{\sqrt{T}}\right) \tag{20}$$

with probability at least $1 - \frac{\beta}{10}$. Taking $\eta = \frac{M}{R\sqrt{T}}$, and $T = \frac{R^2\rho n^2}{\tau^2 d^4}$, this term is strictly dominated by (19).

Finally, putting everything together with the same η and T chosen in Theorem 5.2, and by the union bound, we conclude the result in Theorem B.8.

□