OPT2022: 14th Annual Workshop on Optimization for Machine Learning

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

Andrew Lowy LOWYA@USC.EDU
Meisam Razaviyayn RAZAVIYA@USC.EDU
University of Southern California

Abstract

We study differentially private (DP) stochastic optimization (SO) with loss functions whose worst-case Lipschitz parameter over all data points may be extremely large. To date, the vast majority of work on DP SO assumes that the loss is uniformly Lipschitz continuous over data (i.e. stochastic gradients are uniformly bounded over all data points). While this assumption is convenient, it often leads to pessimistic excess risk bounds. In many practical problems, the worst-case (uniform) Lipschitz parameter of the loss over all data points may be extremely large due to outliers. In such cases, the error bounds for DP SO, which scale with the worst-case Lipschitz parameter of the loss, are vacuous. To address these limitations, this work provides near-optimal excess risk bounds that do not depend on the uniform Lipschitz parameter of the loss. Building on a recent line of work [36, 62], we assume that stochastic gradients have bounded k-th order *moments* for some $k \geqslant 2$. Compared with works on uniformly Lipschitz DP SO, our excess risk scales with the k-th moment bound instead of the uniform Lipschitz parameter of the loss, allowing for significantly faster rates in the presence of outliers and/or heavy-tailed data. For *convex* and *strongly convex* loss functions, we provide the first asymptotically *optimal* excess risk bounds (up to a logarithmic factor). In contrast to [36, 62], our bounds do not require the loss function to be differentiable/smooth. We also devise an accelerated algorithm for smooth losses that runs in linear time and has excess risk that is tight in certain practical parameter regimes. Additionally, our work is the first to address *non-convex* non-uniformly Lipschitz loss functions satisfying the *Proximal-PL inequality*; this covers some practical machine learning models. Our Proximal-PL algorithm has near-optimal excess risk.

1. Introduction

As the use of machine learning (ML) models in industry and society has grown dramatically in recent years, so too have concerns about the privacy of personal data that is used in training such models. It is well-documented that ML models may leak training data, e.g., via model inversion attacks and membership-inference attacks [17, 28, 30, 49, 53]. *Differential privacy* (DP) [27] ensures that data cannot be leaked, and a plethora of work has been devoted to differentially private machine learning and optimization [7, 9, 10, 18, 21, 25, 29, 44, 57, 60]. Of particular importance is the fundamental problem of DP *stochastic (convex) optimization* (S(C)O): given n i.i.d. samples $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$ from an unknown distribution $\mathcal{D}$, we aim to privately solve

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

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

A large body of literature is devoted to characterizing the optimal achievable differentially private excess risk of (1) when the function $f(\cdot, x)$ is uniformly L_f*-Lipschitz* for all $x \in \mathcal{X}$—see e.g., [7, 10, 13, 29, 44]. In these works, the gradient of f is assumed to be uniformly bounded with $\sup_{w \in \mathcal{W}, x \in \mathcal{X}} \|\nabla_w f(w, x)\| \leqslant L_f$, and excess risk bounds scale with L_f. While this assumption is convenient for bounding the *sensitivity* [27] of the steps of the algorithm, it is often unrealistic in practice or leads to pessimistic excess risk bounds. In many practical applications, data contains outliers, is unbounded or heavy-tailed (see e.g. [22, 46, 63] and references therein for such applications). Consequently, L_f may be prohibitively large. For example, even the linear regression loss $f(w, x) = \frac{1}{2}(\langle w, x^{(1)} \rangle - x^{(2)})^2$ with compact $\mathcal{W}$ and data from $\mathcal{X} = \mathcal{X}^{(1)} \times \mathcal{X}^{(2)}$, leads to $L_f \geqslant \text{diameter}(\mathcal{X}^{(1)})^2$, which could be huge. Similar observations can be made for other useful ML models such as deep neural nets [42], and the situation becomes even grimmer in the presence of heavy-tailed data. In these cases, existing excess risk bounds, which scale with L_f, becomes vacuous.

While L_f can be very large in practice (due to outliers), the k-th *moment* of the stochastic gradients is often reasonably small for some $k \geqslant 2$ (see, e.g., Example 1). This is because the k-th moment $\widetilde{r}_k := \mathbb{E}\left[\sup_{w \in \mathcal{W}} \|\nabla_w f(w, x)\|_2^k\right]^{1/k}$ depends on the *average* behavior of the stochastic gradients, while L_f depends on the *worst-case* behavior over all data points. Motivated by this observation and building on the prior results [36, 62], this work characterizes the optimal differentially private excess risk bounds for the class of problems with a given parameter $\widetilde{r}_k$. Specifically, for the class of problems with parameter $\widetilde{r}_k$, we answer the following questions:

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

Prior works have made progress in addressing the first question above:[1] The work of [62] provided the first excess risk upper bounds for *smooth* DP (strongly) convex SO. [36] gave improved, yet suboptimal, upper bounds for *smooth* (strongly) convex $f(\cdot, x)$, and lower bounds for (strongly) convex SO. In this work, we provide *optimal algorithms for convex and strongly convex* losses, resolving Question I up to logarithmic factors. Our bounds hold even for *non-differentiable/non-smooth* F. Regarding Question II, we give the *first algorithm for DP SO with non-convex loss* functions satisfying the Proximal-Polyak-Łojasiewicz condition [37, 50]. We provide a summary of our results for the case $k = 2$ in Figure 1, and a thorough discussion of related work in Appendix B.

1.1. Preliminaries

Let $\|\cdot\|$ be the ℓ_2 norm. Let $\mathcal{W}$ be a convex, compact set of ℓ_2 diameter D. Function $g : \mathcal{W} \to \mathbb{R}$ is μ-*strongly convex* if $g(\alpha w + (1-\alpha)w') \leqslant \alpha g(w) + (1-\alpha)g(w') - \frac{\alpha(1-\alpha)\mu}{2}\|w - w'\|^2$ for all $\alpha \in [0, 1]$ and all $w, w' \in \mathcal{W}$. If $\mu = 0$, we say g is *convex*. For convex $f(\cdot, x)$, denote any *subgradient* of $f(w, x)$ w.r.t. w by $\nabla f(w, x) \in \partial_w f(w, x)$: i.e. $f(w', x) \geqslant f(w, x) + \langle \nabla f(w, x), w' - w \rangle$ for all $w' \in \mathcal{W}$. Function g is β-*smooth* if it is differentiable and its derivative ∇g is β-Lipschitz. For β-smooth, μ-strongly convex g, denote its *condition number* by $\kappa = \beta/\mu$. For functions a and b of

[1] [36, 62] consider a slightly different problem class than the class $\widetilde{r}^{(k)}$, which we consider: see Appendix A. However, our results imply asymptotically optimal rates for the problem class considered in [36, 62] under mild assumptions: see Appendix F.3.

Function Class	Lower bound	Upper bound	Linear-Time Upper bound	Prior state-of-the-art
Nonsmooth Convex	$\underbrace{\sqrt{d}\left(\frac{1}{\sqrt{n}}+\left(\frac{\sqrt{d}}{\epsilon n}\right)^{1/2}\right)}_{\zeta}$ (Theorem 21) + [35, Thm. 6.4]	$\underbrace{\sqrt{d}\left(\frac{1}{\sqrt{n}}+\left(\frac{\sqrt{d}}{\epsilon n}\right)^{1/2}\right)}_{\zeta}$ (Theorems 10 and 20)	None	None
Smooth Convex			$\zeta\left(1+\left(\frac{\epsilon n}{d^{3/2}}\right)^{1/18}\right)$ (Thm. 26)	[35, Thm. 5.4] **(not linear time)**: $\min\left\{\frac{d}{\sqrt{\epsilon n}}, \zeta+\frac{d^{5/4}}{\sqrt{n}}\right\}$ if f is uniformly smooth & $\nabla F(w^*)=0$
Nonsmooth Strongly Convex	$d\left(\frac{1}{n}+\frac{\sqrt{d}}{\epsilon n}\right)$ (Theorem 23) + [35, Theorem 6.1]	$d\left(\frac{1}{n}+\frac{\sqrt{d}}{\epsilon n}\right)$ (Theorems 18 and 20)	None	None
Smooth Strongly Convex			$d\left(\frac{1}{n}+\frac{\sqrt{d\kappa}}{\epsilon n}\right)$ (Theorem 30)	[35, Thm. 5.6] + Appendix B: $d\left(\frac{\kappa_f^3}{n}+\frac{\kappa_f^{3/2}\sqrt{d}}{\epsilon n}\right)$ if f is uniformly smooth & $\nabla F(w^*)=0$
Proximal PL			$d\kappa\left(\frac{1}{n}+\frac{\sqrt{d}}{\epsilon n}\right)$ (Theorem 35)	None

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

input parameters, write $a \lesssim b$ if there is an absolute constant A such that $a \leqslant Ab$ for all feasible values of input parameters. Write $a=\widetilde{O}(b)$ if $a \lesssim \ell b$ for a logarithmic function ℓ of input parameters. We assume that the stochastic gradient distributions have bounded k-th moment for some $k \geqslant 2$:

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

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

Example 1 *For linear regression on a unit ball $\mathcal{W}$ with 1-dimensional data $x^{(1)}, x^{(2)} \in [-10^6, 10^6]$ having truncated Normal distributions and Var$(x^{(1)})$ = Var$(x^{(2)}) \leqslant 1$, we have $L_f \geqslant 10^{12}$. On the other hand, $\tilde{r}_k$ is much smaller than L_f for small to moderate k: e.g., $\tilde{r}_2 \leqslant 5$, $\tilde{r}_4 \leqslant 8$, and $\tilde{r}_8 \leqslant 14$.*

For finite k, Assumption 1 is weaker than assuming that $f(\cdot,x)$ is uniformly Lipschitz for all x since it allows for the p-th moments of $\sup_{w \in \mathcal{W}} \|\nabla f(w,x)\|$ to be unbounded for $\infty \geqslant p > k$.

Differential Privacy [27] ensures that no adversary—even one with enormous resources—can infer much more about any person who contributes training data than if that person's data were absent.

Definition 1 (Zero-Concentrated Differential Privacy (zCDP) [16]) *A randomized algorithm $\mathcal{A}: \mathcal{X}^n \to \mathcal{W}$ satisfies ρ-zero-concentrated differential privacy (ρ-zCDP) if for all $X, X' \in \mathcal{X}^n$ differing in a single entry and all $\alpha \in (1,\infty)$, we have $D_\alpha(\mathcal{A}(X)||\mathcal{A}(X')) \leqslant \rho\alpha$, where $D_\alpha(\mathcal{A}(X)||\mathcal{A}(X'))$ is the α-Rényi divergence[2] between the distributions of $\mathcal{A}(X)$ and $\mathcal{A}(X')$.*

zCDP lives between *pure DP* and *approximate DP* [16]: see Appendix D for the precise statement.

2. Optimal Rates for Non-Smooth (Strongly) Convex Losses (see Appendix F)

We establish asymptotically optimal (up to logarithms) excess risk bounds for DP SCO under Assumption 1, without requiring differentiability of $f(\cdot,x)$:

Theorem 2 (Informal, see Theorem 10, Theorem 18, Theorem 21, Theorem 23) *Let $f(\cdot,x)$ be convex. Grant Assumption 1. Then, there is a polynomial-time $\frac{\epsilon^2}{2}$-zCDP algorithm $\mathcal{A}$ such*

[2] For distributions P and Q with probability density/mass functions p and q, $D_\alpha(P||Q) := \frac{1}{\alpha-1}\ln\left(\int p(x)^\alpha q(x)^{1-\alpha}dx\right)$ [51, Eq. 3.3].

that $\mathbb{E}F(\mathcal{A}(X)) - F^* = \widetilde{\mathcal{O}}\left(\widetilde{r}_{2k} D \left(\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{d}}{\epsilon n}\right)^{(k-1)/k}\right)\right)$. *If* $f(\cdot, x)$ *is* μ*-strongly convex, then* $\mathbb{E}F(\mathcal{A}(X)) - F^* = \widetilde{\mathcal{O}}\left(\frac{\widetilde{r}_{2k}^2}{\mu}\left(\frac{1}{n} + \left(\frac{\sqrt{d}}{\epsilon n}\right)^{(2k-2)/k}\right)\right)$. *Further, these bounds are minimax optimal up to a factor of* $\widetilde{\mathcal{O}}(\widetilde{r}_{2k}^2/\widetilde{r}_k^2)$.

The works [36, 62] make a slightly different assumption than Assumption 1: they instead assume that the k-th order central moment of each coordinate $\nabla_j f(w, x)$ is bounded by $\gamma^{1/k}$ for all $j \in [d], w \in \mathcal{W}$. We also provide asymptotically optimal excess risk bounds for the class of problems satisfying the coordinate-wise moment assumption of [36, 62] and having *subexponential* stochastic subgradients: see Appendix F.3.

The previous state-of-the-art convex upper bound was suboptimal: $\mathcal{O}\left(\widetilde{r}D\sqrt{\frac{d}{n}}\right)$ for $\epsilon \approx 1$ [36, Theorem 5.4].[3] Their result also required $f(\cdot, x)$ to be β_f-smooth for all $x \in \mathcal{X}$, which can be restrictive with outlier data: e.g. this implies that $f(\cdot, x)$ is L_f-Lipschitz with $L_f \leqslant 2\beta_f D$ if $\nabla f(w^*(x), x) = 0$ for some $w^*(x) \in \mathcal{W}$.

Our μ-strongly convex bound also improves over the best previous upper bound of [36, Theorem 5.6], which again required β_f-smoothness of $f(\cdot, x)$. In fact, [36, Theorem 5.6] was incorrect as stated in the ICML 2022 version of their paper, as we explain in Appendix C.[4] However, after communicating with the authors of [36], they updated their result and proof in the arXiv version of their paper. The corrected version of [36, Theorem 5.6]—which we derive in Appendix C for completeness—is suboptimal by a factor of $\widetilde{\mathcal{O}}((\beta_f/\mu)^3$. In practical applications, the condition number β_f/μ can be very large, especially in the presence of outliers or heavy-tailed data. Our near-optimal excess risk bounds remove this dependence on β_f/μ.

Our Algorithm 3 combines the iterative localization technique of [7, 29] with a *clipped* noisy subgradient method[5] run on a regularized empirical objective. In each iteration, we use a geometrically decreasing step size, which shrinks the effective variance of the privacy noise over time. We also geometrically decrease the number of samples used, and we increase the regularization parameter over time as our iterates get closer to the optimum to improve the *stability* [14, 38, 41] of our algorithm. With clipped (hence *biased*) stochastic gradients and non-smooth/non-uniformly Lipschitz loss, the excess risk analysis of our algorithm is harder than in the smooth and uniformly Lipschitz settings. Instead of the uniform convergence analysis used in [36, 62], we derive new results about the stability and generalization error of (regularized) ERM with non-smooth, non-Lipschitz losses. In a bit more detail, our upper bound proof consists of the following steps: i) We bound the excess empirical risk of our noisy clipped subgradient subroutine (Lemma 12). ii) We bound the *on-average model stability* [41] of our algorithm (Proposition 16). On-average model stability is a weaker notion than *uniform stability* [10, 14]; this is necessary to obtain our learnability guarantees for non-Lipschitz $f(\cdot, x)$. iii) We prove that on-average model stable algorithms generalize well (Proposition 14), strengthening and extending results from [41, 52] to non-smooth/non-Lipschitz $f(\cdot, x)$. We obtain our strongly convex bound (Theorem 18) by a reduction to the convex case, ala [29, 32].

[3] We write the bound in [36, Theorem 5.4] in terms of Assumption 1, replacing their $\gamma^{1/k}d$ by $\widetilde{r}\sqrt{d}$.

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

[5] Clipped noisy subgradient method (Algorithm 2) works as follows: In each round t, we: draw a random minibatch of samples; compute and then *clip* (i.e. project onto an ℓ_2 ball centered around 0) their stochastic subgradients; average the clipped stochastic subgradients and add noise; and then take a noisy minibatch stochastic subgradient step.

We also refine (to describe the dependence on $\widetilde{r}_k D, \mu$), extend (to $k \gg 1$), and tighten (for $\mu = 0$) the lower bounds of [36]: see Theorems 21 and 23.

3. Linear-Time Algorithms for Smooth (Strongly) Convex Losses (see Appendix G)

For convex, β-smooth F, we provide a novel accelerated DP algorithm (Algorithm 4), building on the work of Ghadimi and Lan [31].[6] Our algorithm is *linear time* and attains excess risk that improves over the previous state-of-the-art (*not linear time*) algorithm [36, Theorem 5.4] in practical parameter regimes (e.g. $d \gtrsim n^{1/6}$). The excess risk of our algorithm is tight in certain cases: e.g., $d \gtrsim (\epsilon n)^{2/3}$ or "sufficiently smooth" F (see Remark 27). To prove our upper bound, we give the first analysis of accelerated SGD with biased stochastic gradients (Proposition 25).

For μ-strongly convex, β-smooth losses, acceleration results in excessive bias accumulation, so we propose a simple noisy clipped SGD. Our algorithm builds on [36], but uses a lower-bias clipping mechanism from [8] and a tighter analysis. We attain excess risk that is near-optimal up to a $\widetilde{O}((\beta/\mu)^{(k-1)/k})$ factor: see Theorem 30. Our linear-time bound strictly improves over the best previous bound of [36, Theorem 5.6]. Our analysis leverages a novel bound (Proposition 31) for SGD with biased stochastic gradients; this bound is tighter than existing results of [3, 6, 36], which is needed to obtain near-optimal excess risk.

4. First Algorithm for Non-Convex (Proximal-PL) Losses (see Appendix H)

We consider losses satisfying the *Proximal Polyak-Łojasiewicz (PPL) inequality* [37, 50] (Definition 34), an extension of the classical PL inequality to the proximal setting. Many practical ML models satisfy the PPL condition, such as some classes of neural nets, linear/logistic regression, and LASSO [37, 42]. We propose a DP proximal clipped SGD to attain near-optimal excess risk that almost matches the *strongly convex* rate: see Theorem 35. The proof of this result is difficult because it is unclear how to separate the privacy noise from the non-private terms in the proximal/non-convex setting. We prove Proposition 36 by building on [45], which analyzed *uniformly Lipschitz* PPL losses with *unbiased* stochastic gradients using techniques from the analysis of *objective perturbation* [19].

5. Concluding Remarks and Open Questions

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

[6] In contrast to [36, 62], *we do not require* $f(\cdot, x)$ *to be* β*-smooth* for all x.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix

Appendix A. Other Bounded Moment Conditions Besides Assumption 1

In this section, we give the alternate bounded moment assumption made in [36, 62] and a third bounded moment condition, and discuss the relationships between these assumptions. The notation presented here will be necessary in order to state the sharper versions of our linear-time excess risk bounds and the asymptotically optimal excess risk bounds under the coordinate-wise assumption of [36, 62] (which our Algorithm 3 also attains). First, we introduce a relaxation of Assumption 1:

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

Assumption 1 implies Assumption 2 for $r \leqslant \widetilde{r}$. Next, we precisely state the coordinate-wise moment bound assumption that is used in [36, 62] for differentiable f:

Assumption 3 (Used by [36, 62][7] , but not in this work) *There exists $k \geqslant 2$ and $\gamma > 0$ such that $\sup_{w \in \mathcal{W}} \mathbb{E}|\langle \nabla f(w,x) - \nabla F(w), e_j\rangle|^k \leqslant \gamma$, for all $j \in [d]$, where e_j denotes the j-th standard basis vector in $\mathbb{R}^d$. Also, $L \triangleq \sup_{w \in \mathcal{W}} \|\nabla F(w)\| \leqslant \sqrt{d}\gamma^{1/k}$.*

Lemma 3 allows us compare our results in Section 3 obtained under Assumption 2 to the results in [36, 62], which require Assumption 3.

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

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

Proof [Proof of Lemma 3] We use the following inequality, which can easily be verified inductively, using Cauchy-Schwartz and Young's inequalities: for any vectors $u, v \in \mathbb{R}^d$, we have

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

Therefore,

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

[7] The work of [36] assumes that $L \lesssim \gamma^{1/k} = 1$. On the other hand, [62] assumes that F is β-smooth and $\nabla F(w^*) = 0$ for some $w^* \in \mathcal{W}$, which implies $L \leqslant 2\beta D$.

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

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

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

Appendix B. Additional Discussion of Related Work

DP SCO Without Uniform Lipschitz Continuity: The study of DP SCO without uniformly Lipschitz continuous loss functions was initiated by [62], who provided upper bounds for smooth convex/strongly convex loss. The work of [36] provided lower bounds and improved, yet *suboptimal*, upper bounds for the convex case. Both of the works [36, 62] require f to be β_f-smooth. It is also worth mentioning that [36, 62] restricted attention to losses satisfying $\nabla F(w^*) = 0$ for $w^* \in \mathcal{W}$, i.e. $\mathcal{W}$ is a compact set containing the *unconstrained* optimum $w^* = \text{argmin}_{w \in \mathbb{R}^d} F(w) \in \mathcal{W}$. By comparison, we consider the more general *constrained* optimization problem $\min_{w \in \mathcal{W}} F(w)$, where $\mathcal{W}$ need not contain the global unconstrained optimum.

Here we provide a brief discussion of the techniques used in [36, 62]. The work of [62] used a full batch (clipped, noisy) gradient descent based algorithm, building on the heavy-tailed mean estimator of [33]. They bounded the excess risk of their algorithm by using a uniform convergence [58] argument, resulting in a suboptimal dependence on the dimension d. The work of [36] used essentially the same approach as [62], but obtained an improved rate with a more careful analysis.[8] However, as discussed, the bound in [36] is $\mathcal{O}\left(rD\sqrt{\frac{d}{n}}\right)$ when $\epsilon \approx 1$, which is still suboptimal.[9]

More recently, DP optimization with outliers was studied in special cases of sparse learning [34], multi-arm bandits [56], and ℓ_1-norm linear regression [61].

DP ERM and DP GLMs without Uniform Lipschitz continuity: The work of [6] provides bounds for constrained DP *ERM* with arbitrary convex loss functions using a Noisy Clipped SGD algorithm that is similar to our Algorithm 5, except that their algorithm is multi-pass and ours is one pass. In a concurrent work, [23] considered DP *ERM* in the *unconstrained* setting with convex and non-convex loss functions. Their algorithm, noisy clipped SGD, is also similar to Algorithm 5 and the algorithm of [6]. The results in [23] are not directly comparable to [6] since [23] consider the unconstrained setting while [6] consider the constrained setting, but the rates in [6] are faster. [23] also analyzes the convergence of noisy clipped SGD with smooth non-convex loss functions.

The works of [5, 54] consider *generalized linear models (GLMs)*, a particular subclass of convex loss functions and provide empirical and population risk bounds for the *unconstrained* DP optimization problem. The unconstrained setting is not comparable to the constrained setting that we consider here: in the unconstrained case, a dimension-independent upper bound is achievable,

[8] Additionally, [36, Theorem 5.2] provided a bound via noisy gradient descent with the clipping mechanism of [35], but this bound is inferior (in the practical privacy regime $\epsilon \approx 1$) to their bound in [36, Theorem 5.4] that used the estimator of [33].

[9] The bound in [36, Theorem 5.4] for $k = 2$ is stated in the notation of Assumption 3 and thus has an extra factor of $\sqrt{d}$, compared to the bound written here. We write their bound in terms of our Assumption 1, replacing their γd term by $r\sqrt{d}$.

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

Other works on gradient clipping: The gradient clipping technique (and adaptive variants of it) has been studied empirically in works such as [1, 4, 20], to name a few. The work of [20] shows that gradient clipping can prevent SGD from converging, and describes the clipping bias with a disparity measure between the gradient distribution and a geometrically symmetric distribution.

Optimization with biased gradient oracles: The works [3, 6] analyze SGD with biased gradient oracles. Our work provides a tighter bound for smooth, strongly convex functions and analyzes accelerated SGD and proximal SGD with biased gradient oracles.

DP SO with Uniformly Lipschitz loss functions: In the absence of outlier data, there are a multitude of works studying Lipschitz DP SO, mostly in the convex/strongly convex case. We do not attempt to provide a comprehensive list of these here, but will name the most notable ones, which provide optimal or state-of-the-art utility guarantees. The first suboptimal bounds for DP SCO were provided in [9]. The work of [10] established the optimal rate for non-strongly convex DP SCO, by bounding the uniform stability of Noisy DP SGD (without clipping). The strongly convex case was addressed by [29], who also provided optimal rates in linear times for sufficiently smooth, convex losses. Since then, other works have provided faster and simpler (optimal) algorithms for the non-smooth DP SCO problem [7, 11, 12, 40] and considered DP SCO with different geometries [7, 13]. State-of-the-art rates for DP SO with the proximal PL condition are due to [45].

Appendix C. Correcting the Errors in the Strongly Convex Upper Bounds Claimed in [36, 62]

While [36, Theorem 5.6] claims an upper bound for smooth strongly convex losses that is tight up to a factor of $\widetilde{\mathcal{O}}(\kappa_f^2)$, we identify an issue with their proof that invalidates their result. A similar issue appears in the proof of [62, Theorems 5 and 7], which [36] built upon. We then show how to salvage a correct upper bound within the framework of [36], albeit at the cost of an additional factor of κ_f.

The proof of [36, Theorem 5.6] relies on [36, Theorem 3.2]. The proof of [36, Theorem 3.2], in turn, bounds $\mathbb{E}\|w_T - w^*\| \leqslant \frac{(\lambda+L)(M+1)G}{\lambda L}$ in the notation of [36], where L is the smoothness parameter, λ is the strong convexity parameter (so $L \geqslant \lambda$), and M is the diameter of $\mathcal{W}$. Then, it is *incorrectly* deduced that $\mathbb{E}[\|w_T - w^*\|^2] \leqslant \left(\frac{(\lambda+L)(M+1)G}{\lambda L}\right)^2$ (final line of the proof). Notice that $\mathbb{E}[\|w_T - w^*\|^2]$ can be much larger than $(\mathbb{E}\|w_T - w^*\|)^2$ in general: for example, if $\|w_T - w^*\|$ has the Pareto distribution with shape parameter $\alpha \in (1, 2]$ and scale parameter 1, then $(\mathbb{E}\|w_T - w^*\|)^2 = \left(\frac{\alpha}{\alpha-1}\right)^2 \ll \mathbb{E}(\|w_T - w^*\|^2) = \infty$. To try to correct this issue, one could use Young's inequality to instead bound

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

but the geometric series above diverges to $+\infty$ as $T \to \infty$, since $2\left(1 - \frac{2\lambda L}{(\lambda+L)^2}\right) \geqslant 1 \iff (\lambda - L)^2 \geqslant 0$. Evidently, there is no "easy fix" for the issue in the proofs of [36, Theorem 3.2] and

[62, Theorem 5] (at least without imposing severe restrictions on λ and L and hence dramatically shrinking the function class).

Next, we show how to modify the proof of [36, Theorem 5.6] in order to obtain a correct excess risk upper bound of

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

(in our notation). This correction was derived in collaboration with the authors of [36], who have also updated the arXiv version of their paper accordingly. By waiting until the very of the proof of [36, Theorem 3.2] to take expectation, we can derive

$$\|w_t-w^*\|\leqslant\left(1-\frac{\lambda L}{(\lambda+L)^2}\right)\|w_{t-1}-w^*\|+\frac{\|\widetilde{\nabla}F(w_{t-1})-\nabla F(w_{t-1})\|}{\lambda+L} \tag{4}$$

for all t, where we use their $L=\beta_f$ and $\lambda=\mu$ notation but our notation F and $\widetilde{\nabla}F$ for the population loss and its biased noisy gradient estimate (instead of their $L_{\mathcal{D}}$ notation). By iterating (4), we can get

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

Squaring both sides and using Cauchy-Schwartz, we get

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

Using L-smoothness of F and the assumption made in [36] that $\nabla F(w^*)=0$, and *then* taking expectation yields

$$\mathbb{E}F(w_T)-F^*\lesssim L\|w_0-w^*\|^2\left(1-\frac{2\lambda L}{(\lambda+L)^2}\right)^{2T}+TG^2\frac{L}{\lambda}, \tag{5}$$

where $G^2\geqslant\mathbb{E}\left[\|\widetilde{\nabla}F(w_{T-t})-\nabla F(w_{T-t})\|^2\right]$ for all t. It is necessary and sufficient to choose $T=\widetilde{\Omega}(L/\lambda)$ to make the first term in (5) less than the second term (up to logarithms). With this choice of T, we get

$$\mathbb{E}F(w_T)-F^*=\widetilde{\mathcal{O}}\left(G^2\kappa_f^2\right), \tag{6}$$

where $\kappa_f=L/\lambda$. Next, we apply the bound on G^2 for the MeanOracle that is used in [36]; this bound is stated in the version of [36, Lemma B.5] that appears in the updated (November 1, 2022) arXiv version of their paper. The bound (for general γ) is $G^2=\widetilde{\mathcal{O}}\left(\gamma^{2/k}\left[\frac{Td}{n}+d\left(\frac{\sqrt{d}T^{3/2}}{\epsilon n}\right)^{(2k-2)/k}\right]\right)$. Plugging this bound on G^2 into (6) yields (3).

Appendix D. More Preliminaries

Remark on Terminology: Following [36, 62], we sometimes use the terminology "heavy-tailed" to refer to distributions satisfying Assumption 1, 2, or 3. However, as discussed, our results are useful for a broader class of distributions than those typically considered "heavy-tailed": we are also interested in bounded data that contains outliers, or data that is unbounded but light-tailed (e.g. Gaussian), since we may have $\tilde{r} \ll L_f$ in such cases.

Notation: For functions a and b of input parameters, write $a \lesssim b$ if there is an absolute constant A such that $a \lesssim Ab$ for all feasible values of input parameters.

More Differential Privacy Preliminaries: Recall the classical definition of differential privacy [27]:

Definition 4 (Differential Privacy) *Let $\epsilon \geqslant 0,\ \delta \in [0, 1)$. A randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$ is (ϵ, δ)-differentially private (DP) if for all pairs of adjacent data sets $X, X' \in \mathcal{X}^n$ (i.e. $d_{hamming}(X, X') = 1$) and all measurable subsets $S \subseteq \mathcal{W}$, we have $\mathbb{P}(\mathcal{A}(X) \in S) \leqslant e^{\epsilon}\mathbb{P}(\mathcal{A}(X') \in S) + \delta$.*

If $\delta = 0$, we say $\mathcal{A}$ satisfies *pure DP*. If $\delta > 0$, then $\mathcal{A}$ satisfies *approximate DP*. zCDP is weaker than pure DP, but stronger than approximate DP in the following sense:

Proposition 5 *[16, Proposition 1.3] If $\mathcal{A}$ is ρ-zCDP, then $\mathcal{A}$ is $(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta)$ for any $\delta > 0$.*

Thus, if $\epsilon \leqslant \sqrt{\log(1/\delta)}$, then any $\frac{\epsilon^2}{2}$-zCDP algorithm is $(2\epsilon\sqrt{\log(1/\delta)}, \delta)$-DP.

Our algorithms use the *Gaussian mechanism* to achieve zCDP:

Proposition 6 *[16, Proposition 1.6] Let $q : \mathcal{X}^n \to \mathbb{R}$ be a query with ℓ_2-sensitivity $\Delta := \sup_{X \sim X'} \|q(X) - q(X')\|$. Then the Gaussian mechanism, defined by $\mathcal{M} : \mathcal{X}^n \to \mathbb{R}$, $M(X) := q(X) + u$ for $u \sim \mathcal{N}(0, \sigma^2)$, is ρ-zCDP if $\sigma^2 \geqslant \frac{\Delta^2}{2\rho}$.*

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

Lemma 7 *[16, Lemma 2.3] Suppose $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ satisfies ρ-zCDP and $\mathcal{A}' : \mathcal{X}^n \times \mathcal{Y} \to \mathcal{Z}$ satisfies ρ'-zCDP (as a function of its first argument). Define the composition of $\mathcal{A}$ and $\mathcal{A}'$, $\mathcal{A}'' : \mathcal{X}^n \to \mathcal{Z}$ by $\mathcal{A}''(X) = \mathcal{A}'(X, \mathcal{A}(X))$. Then $\mathcal{A}''$ satisfies $(\rho + \rho')$-zCDP. In particular, the composition of T ρ-zCDP mechanisms is a $T\rho$-zCDP mechanism.*

Appendix E. Private Heavy-Tailed Mean Estimation Building Blocks

In each iteration of our SO algorithms, we need a way to privately estimate the mean $\nabla F(w_t) = \mathbb{E}_{x \sim \mathcal{D}}[\nabla f(w_t, x)]$. If $f(\cdot, x)$ is Lipschitz, then one can simply draw a random sample x^t from X and add noise to the stochastic gradient $\nabla f(w_t, x^t)$ to obtain a DP estimator of $\nabla F(w_t)$: the ℓ_2-sensitivity of the stochastic gradients is bounded by $\sup_{x, x' \in \mathcal{X}} \|\nabla f(w_t, x) - \nabla f(w_t, x')\| \leqslant 2L_f$, so the Gaussian mechanism guarantees DP (by Proposition 6). However, in the setting that we consider, L_f (and hence the sensitivity) may be unbounded, so noisy stochastic gradients are not DP. Thus, we *clip* the stochastic gradients (to force the sensitivity to be bounded) before adding noise. Specifically, we invoke Algorithm 1 on a batch of s stochastic gradients at each iteration of our algorithms. In Algorithm 1, $\Pi_C(z) := \text{argmin}_{y \in B_2(0, C)} \|y - z\|^2$ denotes the projection onto the centered ℓ_2 ball of radius C in $\mathbb{R}^d$. Lemma 8 bounds the bias and variance of Algorithm 1.

Algorithm 1 ℓ_2 Clip `MeanOracle1`$(\{x_i\}_{i=1}^s; s; C; \frac{\epsilon^2}{2})$ [8]

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

Lemma 8 ([8]) *Let $\{x_i\}_{i=1}^s \sim \mathcal{D}^s$ be $\mathbb{R}^d$-valued random vectors with $\mathbb{E}x_i = \nu$ and $\mathbb{E}\|x_i\|^k \leqslant r^{(k)}$ for some $k \geqslant 2$. Denote the noiseless average of clipped samples by $\hat{\nu} := \frac{1}{s}\sum_{i=1}^s \Pi_C(x_i)$. Then, $\|\mathbb{E}\tilde{\nu} - \nu\| = \|\mathbb{E}\hat{\nu} - \nu\| \leqslant \mathbb{E}\|\hat{\nu} - \nu\| \leqslant \frac{r^{(k)}}{(k-1)C^{k-1}}$, and $\mathbb{E}\|\tilde{\nu} - \mathbb{E}\tilde{\nu}\|^2 = \mathbb{E}\|\tilde{\nu} - \mathbb{E}\hat{\nu}\|^2 \leqslant d\sigma^2 + \frac{r^{(2)}}{s}$.*

Appendix F. Details and Proofs of Results in Section 2

In this section, we establish the optimal rates (up to logarithms) for the class of DP SCO problems satisfying Assumption 1. We present our result for convex losses in Section 2, and our result for strongly convex losses in Appendix F.2. In Appendix F.3, we show how our (tight) upper bounds under Assumption 1 lead to asymptotically tight upper bounds under Assumptions 2 and 3. In Appendix F.4, we provide lower bounds, which show that our upper bounds are tight.

F.1. Localized Noisy Clipped Subgradient Method for Convex Losses

Our algorithm (Algorithm 3) uses iterative localization [7, 29] with clipping (in Algorithm 2) to handle heavy-tailed data.[10]

Algorithm 2 Noisy Clipped Regularized Subgradient Method for DP ERM

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

The main ideas of Algorithm 3 are:

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

2. *Solve regularized ERM subproblem with a stable DP algorithm*: We run a *multi-pass* zCDP solver on a *regularized* empirical loss: Multiple passes let us reduce the noise variance in

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

Algorithm 3 Localized Noisy Clipped Subgradient Method for DP SCO

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

phase i by a factor of T_i (via strong composition for zCDP) and get a more accurate solution to the ERM subproblem. Regularization makes the empirical loss strongly convex, which improves *on-average model stability* and hence generalization of the obtained solution (see Proposition 14 and 15).

3. *Localization* [6, 29] (i.e. iteratively "zooming in" on a solution): In early phases (small i), when we are far away from the optimum w^*, we use more samples (larger n_i) and large learning rate η_i to make progress quickly. As i increases, w_i is closer to w^*, so fewer samples and slower learning rate suffice. Since step size η_i shrinks (geometrically) faster than n_i, the effective variance of the privacy noise $\eta_i^2\sigma_i^2$ decreases as i increases. This prevents w_{i+1} from moving too far away from w_i (and hence from w^*). We further enforce this localization behavior by increasing the regularization parameter λ_i and shrinking D_i over time. We choose D_i as small as possible subject to the constraint that $\operatorname{argmin}_{w \in \mathcal{W}} \widehat{F}_i(w) \in \mathcal{W}_i$. This constraint ensures that Algorithm 2 can find w_i with small excess risk.

Next, we will provide privacy and excess risk guarantees for Algorithm 3. In order to precisely state our result, we will need to introduce some notation:

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

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

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

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

and let

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

Note that $\widetilde{r}_{k,1} = \widetilde{r}_k$. Our excess risk upper bounds will depend on a weighted average of the expected empirical moments for different batch sizes $m \in \{1, 2, 4, 8, \cdots, n\}$, with more weight being given to $\widetilde{r}_m$ for large m (which are smaller, by Lemma 9 below): for $n = 2^l$, define

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

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

Lemma 9 *Under Assumptions 1 and 2, we have:* $\widetilde{r}^{(k)} = \widetilde{e}_1^{(k)} \geqslant \widetilde{e}_2^{(k)} \geqslant \widetilde{e}_4^{(k)} \geqslant \widetilde{e}_8^{(k)} \geqslant \cdots \geqslant r^{(k)}$. *In particular,* $\widetilde{R}_{k,n} \leqslant \widetilde{r}_k$.

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

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

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

With this notation, we can give the precise statement of our upper bound:

Theorem 10 *Grant Assumption 1. Let $f(\cdot, x)$ be convex and L_f-Lipschitz for all x, with $L_f \lesssim n^{p/2} \widetilde{R}_{2k,n} \left(\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{d \ln(n)}}{\epsilon n} \right)^{\frac{k-1}{k}} \right)$ for some $p \geqslant 2$ and $\epsilon \leqslant \sqrt{d}$. Then, there are algorithmic parameters such that Algorithm 3 is $\frac{\epsilon^2}{2}$-zCDP, and has excess risk*

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

Moreover, this excess risk is attained in $\widetilde{\mathcal{O}}(n^{p+1})$ subgradient evaluations. If $p \in [1, 2)$, then the same excess risk bound holds up to logarithmic factors.

The proof of Theorem 10 consists of three main steps: i) We bound the empirical error of the noisy clipped subgradient subroutine (Lemma 12). ii) We prove that if an algorithm is *on-average model stable* (Definition 13), then it generalizes (Proposition 14). iii) We bound the on-average model stability of regularized ERM with non-smooth/non-Lipschitz loss (Proposition 15), leading to an excess population loss bound for Algorithm 2 run on the regularized empirical objective (c.f. line 7 of Algorithm 3). By using these results with the proof technique of [29], we can obtain Theorem 10.

We proceed to prove the technical ingredients that will be used in the proof of Theorem 10. First, we will prove a variant of Lemma 8 that bounds the bias and variance of the subgradient estimator in Algorithm 2.

Lemma 11 *Let $\widehat{F}_\lambda(w) = \frac{1}{n}\sum_{i=1}^n f(w, x_i) + \frac{\lambda}{2}\|w - w_0\|^2$ be a regularized empirical loss on a closed convex domain $\mathcal{W}$ with ℓ_2-diameter D. Let $\widetilde{\nabla} F_\lambda(w_t) = \nabla \widehat{F}_\lambda(w_t) + b_t + N_t = \frac{1}{n}\sum_{i=1}^n \Pi_C(\nabla f(w, x_i)) + \lambda(w - w_0) + N_t$ be the biased, noisy subgradients of the regularized empirical loss in Algorithm 2, with $N_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ and $b_t = \frac{1}{n}\sum_{i=1}^n \Pi_C(\nabla f(w_t, x_i)) - \frac{1}{n}\sum_{i=1}^n \nabla f(w_t, x_i)$. Assume $\widehat{r}_n(X)^{(k)} \geqslant \sup_{w \in \mathcal{W}} \left\{\frac{1}{n}\sum_{i=1}^n \|\nabla f(w, x_i)\|^k\right\}$ for all $\nabla f(w, x_i) \in \partial_w f(w, x_i)$. Then, for any $T \geqslant 1$, we have:*

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

and

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

Proof Fix any t. We have

$$\begin{aligned} \|b_t\| &= \left\| \frac{1}{n}\sum_{i=1}^n \Pi_C(\nabla f(w_t, x_i)) - \frac{1}{n}\sum_{i=1}^n \nabla f(w_t, x_i) \right\| \\ &\leqslant \frac{1}{(k-1)C^{k-1}} \left[\frac{1}{n}\sum_{i=1}^n \|\nabla f(w_t, x_i)\|^k \right], \end{aligned} \tag{7}$$

by Lemma 8 applied with $\mathcal{D}$ as the empirical distribution on X, and x_i in Lemma 8 corresponding to $\nabla f(w_t, x_i)$ in (7). Taking supremum over t of both sides of (7) and recalling the definition of $\widehat{r}_n(X)^{(k)}$ proves the bias bound. The noise variance bound is immediate from the distribution of N_t. ∎

Using Lemma 11, we can obtain the following convergence guarantee for Algorithm 2:

Lemma 12 *Fix $X \in \mathcal{X}^n$ and let $\widehat{F}_\lambda(w) = \frac{1}{n}\sum_{i=1}^n f(w, x_i) + \frac{\lambda}{2}\|w - w_0\|^2$ for $w_0 \in \mathcal{W}$, where $\mathcal{W}$ is a closed convex domain with diameter D. Assume $f(\cdot, x)$ is convex and $\widehat{r}_n(X)^{(k)} \geqslant \sup_{w \in \mathcal{W}} \left\{\frac{1}{n}\sum_{i=1}^n \|\nabla f(w, x_i)\|^k\right\}$ for all $\nabla f(w, x_i) \in \partial_w f(w, x_i)$. Denote $\widehat{r}_n(X) = \left[\widehat{r}_n(X)^{(k)}\right]^{1/k}$ and $\hat{w} = \operatorname{argmin}_{w \in \mathcal{W}} \widehat{F}_\lambda(w)$. Let $\eta \leqslant \frac{2}{\lambda}$. Then, the output of Algorithm 2 satisfies*

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

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

Proof We use the notation of Lemma 11 and write $\widetilde{\nabla} F_\lambda(w_t) = \nabla \widehat{F}_\lambda(w_t) + b_t + N_t = \frac{1}{n}\sum_{i=1}^n \Pi_C(\nabla f(w, x_i)) + \lambda(w - w_0) + N_t$ as the biased, noisy subgradients of the regularized empirical loss in Algorithm 2, with $N_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ and $b_t = \frac{1}{n}\sum_{i=1}^n \Pi_C(\nabla f(w_t, x_i)) - \frac{1}{n}\sum_{i=1}^n \nabla f(w_t, x_i)$. Denote $y_{t+1} = w_t - \widetilde{\nabla} F_\lambda(w_t)$, so that $w_{t+1} = \Pi_{\mathcal{W}}(y_{t+1})$. For now, condition on the randomness of the algorithm (noise). By strong convexity, we have

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

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

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

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

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

where $\hat{B} := \sup_{t \in [T]} \|b_t\| \leqslant \frac{\widehat{r}_n(X)^{(k)}}{(k-1)C^{k-1}}$ and $\hat{\Sigma}^2 := \sup_{t \in [T]} \mathbb{E}\|N_t\|^2 = d\sigma^2$. by Lemma 11. We also used Young's and Jensen's inequalities and the fact that $\mathbb{E}N_t = 0$. Further,

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

by Young's inequality. Combining these pieces yields

$$\mathbb{E}\|w_{t+1} - \hat{w}\|^2 \leqslant \left(1 - \frac{\lambda\eta}{2}\right)\mathbb{E}\|w_t - \hat{w}\|^2 + 4\eta^2\left(\widehat{r}_n(X)^2 + \lambda^2 D^2 + \hat{B}^2 + \hat{\Sigma}^2\right) + \frac{2\eta\hat{B}^2}{\lambda}. \tag{8}$$

Iterating (8) gives us

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

since $\eta \leqslant \frac{2}{\lambda}$. Plugging in the bounds on $\hat{B}$ and $\hat{\Sigma}$ from Lemma 11 completes the proof. ■

Our next goal is to bound the generalization error of regularized ERM with non-Lipschitz/non-smooth convex loss functions. We will use a stability argument to obtain such a bound. Recall the notion of *on-average model stability* [41]:

Definition 13 *Let $X = (x_1, \cdots, x_n)$ and $X' = (x'_1, \cdots, x'_n)$ be drawn independently from $\mathcal{D}$. For $i \in [n]$, let $X^i := (x_1, \cdots, x_{i-1}, x'_i, x_{i+1}, \cdots, x_n)$. We say randomized algorithm $\mathcal{A}$ has on-average model stability α (i.e. $\mathcal{A}$ is α-on-average model stable) if $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|\mathcal{A}(X) - \mathcal{A}(X^i)\|^2\right] \leqslant \alpha^2$. The expectation is over the randomness of $\mathcal{A}$ and the draws of X and X'.*

On-average model stability is weaker than the notion of *uniform stability* [14], which has been used in DP Lipschitz SCO (e.g. by [10]); this is necessary for obtaining our learnability guarantees without Lipschitz continuity.

The main result in [41] showed that on-average model stable algorithms generalize well if $f(\cdot, x)$ is β_f-*smooth* for all x, which leads to a restriction on L_f. We show that neither smoothness nor Lipschitz continuity of f is needed to ensure generalizability:

Proposition 14 *Let $f(\cdot, x)$ be convex for all x and grant Assumption 2 for $k = 2$. Suppose $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$ is α-on-average model stable. Then for any $\zeta > 0$, we have*

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

Proof Let X, X', X^i be constructed as in Definition 13. We may write $\mathbb{E}[F(\mathcal{A}(X)) - \widehat{F}_X(\mathcal{A}(X))] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^n f(\mathcal{A}(X^i), x_i) - f(\mathcal{A}(X), x_i)]$, by symmetry and independence of x_i and $\mathcal{A}(X^i)$ (c.f. [41, Equation B.2]). Then by convexity, we have

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

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

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

Combining the above inequalities and recalling Definition 13 yields the result. ■

Using Proposition 14, we will bound the generalization error and excess (population) risk of regularized ERM in Proposition 16. To prove Proposition 16, we first require the following bound on the generalization error of ERM with strongly convex loss:

Proposition 15 *Let $f(\cdot, x)$ be λ-strongly convex, and grant Assumption 2. Let $\mathcal{A}(X) := \operatorname{argmin}_{w\in\mathcal{W}} \widehat{F}_X(w)$ be the ERM algorithm. Then,*

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

Proof We first bound the stability of ERM and then use Proposition 14 to get a bound on the generalization error. The beginning of the proof is similar to the proof of [41, Proposition D.6]: Let X, X', X^i be constructed as in Definition 13. By strong convexity of $\widehat{F}_{X^i}$ and optimality of $\mathcal{A}(X^i)$, we have

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

which implies

$$\frac{1}{n}\sum_{i=1}^n \|\mathcal{A}(X) - \mathcal{A}(X^i)\|^2 \leqslant \frac{2}{\lambda n}\sum_{i=1}^n \left[\widehat{F}_{X^i}(\mathcal{A}(X)) - \widehat{F}_{X^i}(\mathcal{A}(X^i))\right]. \tag{9}$$

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

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

Hence

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

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

$$\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n \widehat{F}_{X^i}(\mathcal{A}(X)) - \widehat{F}_{X^i}(\mathcal{A}(X^i))\right] = \frac{1}{n}\mathbb{E}\left[F(\mathcal{A}(X)) - \widehat{F}_X(\mathcal{A}(X))\right]. \tag{10}$$

Combining (9) with (10) shows that ERM is α-on-average model stable for

$$\alpha^2 = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\|\mathcal{A}(X)-\mathcal{A}(X^i)\|^2\right] \leqslant \frac{2}{\lambda n}\mathbb{E}\left[F(\mathcal{A}(X)) - \widehat{F}_X(\mathcal{A}(X))\right]. \tag{11}$$

The rest of the proof is where we depart from the analysis of [41] (which required smoothness of $f(\cdot, x)$): Bounding the right-hand side of (11) by Proposition 14 yields

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

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

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

and $\alpha^2 \leqslant \frac{4r^2}{\lambda^2 n^2}$. Applying Proposition 14 again yields (for any $\zeta' > 0$)

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

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

Proposition 16 *Let $f(\cdot, x)$ be convex, $w_{i-1}, y \in \mathcal{W}$, and $\hat{w}_i := \mathrm{argmin}_{w\in\mathcal{W}} \widehat{F}_i(w)$, where $\widehat{F}_i(w) := \frac{1}{n_i}\sum_{j\in\mathcal{B}_i} f(w, x_j) + \frac{\lambda_i}{2}\|w - w_{i-1}\|^2$ (c.f. line 6 of Algorithm 3). Then,*

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

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

Proof Denote the regularized population loss by $G_i(w) := \mathbb{E}[\widehat{F}_i(w)] = F(w) + \frac{\lambda_i}{2}\|w - w_{i-1}\|^2$. By Proposition 15, we have

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

Thus,

$$\begin{aligned}\frac{\lambda_i}{2}\mathbb{E}\|\hat{w}_i - w_{i-1}\|^2 + \mathbb{E}F(\hat{w}_i) &= \mathbb{E}G_i(\hat{w}_i) \\ &\leqslant \frac{2r^2}{\lambda_i n_i} + \mathbb{E}[\widehat{F}_i(\hat{w}_i)] \\ &\leqslant \frac{2r^2}{\lambda_i n_i} + \frac{\lambda_i}{2}\|y - w_{i-1}\|^2 + F(y),\end{aligned} \tag{12}$$

since $\mathbb{E}[\widehat{F}_i(\hat{w}_i)] = \mathbb{E}[\min_{w\in\mathcal{W}} \widehat{F}_i(w)] \leqslant \min_{w\in\mathcal{W}} \mathbb{E}[\widehat{F}_i(w)] = \min_{w\in\mathcal{W}} G_i(w) \leqslant \frac{\lambda_i}{2}\|y - w_{i-1}\|^2 + F(y)$. Subtracting $F(y)$ from both sides of (12) completes the proof. ■

We are ready to prove Theorem 10:

Proof [Proof of Theorem 10] We choose $\sigma_i^2 = \frac{4C_i^2 T_i}{n_i^2 \epsilon^2}$ for C_i and T_i to be determined exactly later. Note that for λ_i and η_i defined in Algorithm 3, we have $\eta_i \leqslant \frac{2}{\lambda_i}$ for all $i \in [l]$.

Privacy: Since the batches $\{\mathcal{B}_i\}_{i=1}^l$ are disjoint, it suffices (by parallel composition [48]) to show that w_i (produced by T_i iterations of Algorithm 2 in line 7 of Algorithm 3) is $\frac{\epsilon^2}{2}$-zCDP for all $i \in [l]$. With clip threshold C_i and batch size n_i, the ℓ_2 sensitivity of the clipped subgradient update is bounded by $\Delta = \sup_{w, X \sim X'} \frac{1}{n_i} \| \sum_{j=1}^{n_i} \Pi_{C_i}(\nabla f(w, x_j)) - \Pi_{C_i}(\nabla f(w, x_j'))\| = \frac{1}{n_i} \sup_{w,x,x'} \|\Pi_{C_i}(\nabla f(w,x)) - \Pi_{C_i}(\nabla f(w,x'))\| \leqslant \frac{2C_i}{n_i}$. (Note that the terms arising from regularization cancel out.) Thus, by Proposition 6, conditional on the previous updates $w_{1:t}$, the $(t+1)$-st update in line 5 of Algorithm 2 satisfies $\frac{\epsilon^2}{2T_i}$-zCDP. Hence, Lemma 7 implies that w_i (in line 7 of Algorithm 3) is $\frac{\epsilon^2}{2}$-zCDP.

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

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

by definition of $\hat{w}_i$ and L_f-Lipschitz continuity of $f(\cdot, x_j)$ for all j. Then by Lemma 12, we have

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

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

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

where $d\sigma_i^2 \leqslant \frac{4dC_i^2 T_i}{n_i^2 \epsilon^2}$. Choosing $T_i = \frac{1}{\lambda_i \eta_i}\ln\left(\frac{D^2 \lambda_i}{d\sigma_i^2 \eta_i}\right) \lesssim n_i^p \ln(n)$ and η to be determined later (polynomial in n), we get

$$\begin{aligned}\mathbb{E}\|w_i - \hat{w}_i\|^2 &\lesssim \frac{\eta_i}{\lambda_i}\left(L_f^2 + d\sigma_i^2\right) + \frac{\widetilde{e}_{n_i}^{(2k)}}{\lambda_i^2 C_i^{2k-2}} \\ &\lesssim \eta_i^2 n_i^p (L_f^2 + d\sigma_i^2) + \frac{\eta_i^2 n_i^{2p} \widetilde{e}_{n_i}^{(2k)}}{C_i^{2k-2}} \\ &\lesssim \left(\frac{\eta^2 n^p}{32^i}\left(L_f^2 + \frac{dC_i^2 T_i}{\epsilon^2 n_i^2} + \frac{n^p \widetilde{e}_{n_i}^{(2k)}}{C_i^{2k-2} 2^{pi}}\right)\right).\end{aligned} \tag{13}$$

Note that under Assumption 1, F is L-Lipshitz, where $L = \sup_{w \in \mathcal{W}} \|\nabla F(w)\| \leqslant r$ by Jensen's inequality. Now, following the strategy used in the proof of [29, Theorem 4.4], we write

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

where $\hat{w}_0 := w^*$. Using (13), the first term can be bounded as follows:

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

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

$$\mathbb{E}[F(w_l) - F(\hat{w}_l)] \lesssim \widetilde{R}_{2k,n} D \left(\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{d \ln(n)}}{\epsilon n} \right)^{\frac{k-1}{k}} \right), \tag{14}$$

if we choose

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

Next, Proposition 16 implies

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

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

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

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

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

where the last line holds verbatim if $p \geqslant 2$ and holds up to an additional factor of $\ln(n)$ otherwise. Assume $p \geqslant 2$. Now, choosing

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

yields

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

by the upper bound that we assumed on L_f. Combining the above pieces completes the excess risk proof.

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

Remark 17 (Reduced Computational Cost for Approximate DP) *If one desires (ϵ,δ)-DP or (ϵ,δ)-SDP instead of zCDP, then the gradient complexity of Algorithm 3 can be improved to $\mathcal{O}(n^{p+\frac{1}{2}}\ln(n))$ by using Clipped Noisy Stochastic Subgradient Method instead of Algorithm 2 as the subroutine in line 7 of Algorithm 3. Choosing batch sizes $m_i \approx \sqrt{n_i} < n_i$ in this subroutine (and increasing σ_i^2 by a factor of $\mathcal{O}(\log(1/\delta))$) ensures (ϵ,δ)-DP by [1, Theorem 1] via privacy amplification by subsampling. The same excess risk bounds hold for any minibatch size $m_i \in [n_i]$, as the proof of Theorem 10 shows.*

F.2. The Strongly Convex Case

Our algorithm is an instantiation of the meta-algorithm described in [29]: Initialize $w_0 \in \mathcal{W}$. For $j \in [M] := \lceil\log_2(\log_2(n))\rceil$, let $N_j = 2^{j-2}n/\log_2(n)$, $\mathcal{C}_j = \left\{\sum_{h<j} N_h + 1, \ldots, \sum_{h\leqslant j} N_h\right\}$, and let w_j be the output of Algorithm 3 run with input data $X_j = (x_s)_{s\in\mathcal{C}_j}$ initialized at w_{j-1}. Output w_M. Assume without loss of generality that $N_j = 2^p$ for some $p \in \mathbb{N}$. Then, with the notation defined in Appendix F, we have the following guarantees:

Theorem 18 *Grant Assumption 1. Let $\epsilon \leqslant \sqrt{d}$ and $f(\cdot, x)$ be μ-strongly convex and L_f-Lipschitz for all $x \in \mathcal{X}$, with $L_f \lesssim n^{p/2}\widetilde{R}_{2k,n}\left(\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{d\ln(n)}}{\epsilon n}\right)^{\frac{k-1}{k}}\right)$ for some $p \geqslant 2$. Then, there is a polynomial-time $\frac{\epsilon^2}{2}$-zCDP algorithm $\mathcal{A}$ based on Algorithm 3 with excess risk*

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

Proof **Privacy**: Since the batches X_j used in each phase of the algorithm are disjoint and Algorithm 3 is $\frac{\epsilon^2}{2}$-zCDP, privacy of the algorithm follows from parallel composition of DP [48].
Excess risk: Note that N_j samples are used in phase j of the algorithm. For $j \geqslant 0$, let $D_j^2 = \mathbb{E}[\|w_j - w^*\|^2]$ and $\Delta_j = \mathbb{E}[F(w_j) - F^*]$. By strong convexity, we have $D_j^2 \leqslant \frac{2\Delta_j}{\mu}$. Also,

$$\begin{aligned}\Delta_{j+1} &\leqslant a\widetilde{R}_{2k,N_j}D_j\left(\frac{1}{\sqrt{N_j}} + \left(\frac{\sqrt{d\ln(N_j)}}{\epsilon N_j}\right)^{\frac{k-1}{k}}\right)\\ &\leqslant a\widetilde{R}_{2k,N_j}\sqrt{\frac{2\Delta_j}{\mu}}\left(\frac{1}{\sqrt{N_j}} + \left(\frac{\sqrt{d\ln(N_j)}}{\epsilon N_j}\right)^{\frac{k-1}{k}}\right) \qquad (15)\end{aligned}$$

for an absolute constant $a \geqslant 1$, by Theorem 10. Denote $E_j = \left[a\widetilde{R}_{2k,N_j}\sqrt{\frac{2}{\mu}}\left(\frac{1}{\sqrt{N_j}} + \left(\frac{\sqrt{d\ln(N_j)}}{\epsilon N_j}\right)^{\frac{k-1}{k}}\right)\right]^2$. Then since $N_j = 2N_{j+1}$, we have

$$\begin{aligned}\frac{E_j}{E_{j+1}} &\leqslant 4\left(\frac{\widetilde{R}_{2k,N_j}}{\widetilde{R}_{2k,N_{j+1}}}\right)^2\\ &\leqslant 8, \qquad (16)\end{aligned}$$

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

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

Now, (16) implies that (15) can be re-arranged as

$$\frac{\Delta_{j+1}}{64E_{j+1}} \leqslant \sqrt{\frac{\Delta_j}{64E_j}} \leqslant \left(\frac{\Delta_0}{64E_0}\right)^{1/2^{j+1}}. \qquad (17)$$

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

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

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

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

since $N_M = n/4$. ■

F.3. Asymptotic Upper Bounds Under Assumption 3

We first recall the notion of *subexponential* distribution:

Definition 19 (Subexponential Distribution) *A random variable Y is subexponential if there is an absolute constant $s > 0$ such that $\mathbb{P}(|Y| \geqslant t) \leqslant 2\exp\left(-\frac{t}{s}\right)$ for all $t \geqslant 0$. For subexponential Y, we define $\|Y\|_{\psi_1} := \inf\left\{s > 0 : \mathbb{P}(|Y| \geqslant t) \leqslant 2\exp\left(-\frac{t}{s}\right) \ \forall\, t \geqslant 0\right\}$.*

Essentially all (heavy-tailed) distributions that arise in practice are subexponential [47].

Now, we establish asymptotically optimal upper bounds for a broad subclass of the problem class considered in [36, 62]: namely, *subexponential* stochastic subgradient distributions satisfying Assumption 3. These upper bounds are a consequence of Lemma 3 combined with the following theorem (which uses the notation introduced in Section 2):

Theorem 20 *Let $f(\cdot, x)$ be convex. Assume $\widetilde{r}_{2k} < \infty$ and $Y_i = \|\nabla f(w, x_i)\|^{2k}$ is subexponential with $E_n \geqslant \max_{i\in[n]}\left(\|Y_i\|_{\psi_1}\right) \forall w \in \mathcal{W}$, $\nabla f(w, x_i) \in \partial_w f(w, x_i)$. Assume that for sufficiently large n, we have $\sup_{w,x}\|\nabla f(w,x)\|^{2k} \leqslant n^q r^{(2k)}$ for some $q \geqslant 1$ and $\max\left(\frac{E_n}{r^{(2k)}}, \frac{E_n^2}{(r^{(2k)})^2}\right)\ln\left(\frac{3nD\beta}{4r_{2k}}\right) \leqslant \frac{n}{dq}$, where $\|\nabla f(w,x) - \nabla f(w',x)\| \leqslant \beta\|w - w'\|$ for all $w, w' \in \mathcal{W}, x \in \mathcal{X}$, and subgradients $\nabla f(w,x) \in \partial_w f(w,x)$. Then, $\lim_{n\to\infty}\widetilde{R}_{2k,n} \leqslant 4r_{2k}$. Further, there exists $N \in \mathbb{N}$ such that for all $n \geqslant N$, the output of Algorithm 3 satisfies*

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

If $f(\cdot, x)$ is μ-strongly convex, then the output of algorithm $\mathcal{A}$ (in Appendix F.2) satisfies

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

While a bound on $\sup_{w,x}\|\nabla f(w,x)\|$ is needed in Theorem 20, it can grow as fast as any polynomial in n and only needs to hold for sufficiently large n. As $n \to \infty$, this assumption is easily satisfied. Likewise, Theorem 20 depends only logarithmically on the Lipschitz parameter of the subgradients β, so the result still holds up to constant factors if, say, $\beta \leqslant n^p(r/D)$ as $n \to \infty$ for some $p \geqslant 1$. Crucially, our excess risk bounds do not depend on L_f or β.

Under Assumption 3, the upper bounds in Theorem 20 hold with r replaced by $\sqrt{d}\gamma^{1/k}$ (by Lemma 3). These upper bounds, and the ones in Theorem 20, are tight up to logarithms for their respective problem classes, by the lower bounds in Appendix F.4.

Proof [Proof of Theorem 20] **Step One:** *There exists $N \in \mathbb{N}$ such that $\widetilde{r}_{2k,n}^2 \leqslant 16 r_{2k}^2$ for all $n \geqslant N$.* We will first use a covering argument to show that $\widehat{r}_n(X)^{(2k)}$ is upper bounded by $2^{2k+1}r^{(2k)}$ with high probability. For any $\alpha > 0$, we may choose an α-net with $N_\alpha \leqslant \left(\frac{3D}{2\alpha}\right)^d$ balls centered around points in $\mathcal{W}_\alpha = \{w_1, w_2, \cdots, w_{N_\alpha}\} \subset \mathcal{W}$ such that for any $w \in \mathcal{W}$ there exists $i \in [N_\alpha]$ with $\|w - w_i\| \leqslant \alpha$ (see e.g. [39] for the existence of such $\mathcal{W}_\alpha$). For $w \in \mathcal{W}$, let $\widetilde{w}$ denote the element of $\mathcal{W}_\alpha$ that is closest to w, so that $\|w - \widetilde{w}\| \leqslant \alpha$. Now, for any $X \in \mathcal{X}^n$, we have

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

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

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

by a union bound and Bernstein's inequality (see e.g. [59, Corollary 2.8.3]). Choosing $\alpha = \frac{2r_{2k}}{\beta}$ ensures that $\mathbb{P}(2^{2k}\beta^{2k}\alpha^{2k} > 2^{2k+1}r^{(2k)}) = 0$ and hence (by union bound)

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

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

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

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

Step Two: $\lim_{n \to \infty} \widetilde{R}_{2k,n} \leqslant 4 r_{2k}$.
For all $n = 2^l, l, i \in \mathbb{N}$, define $h_n(i) = 2^{-i} \widetilde{r}^2_{2k, 2^{-i}n} \mathbb{1}_{\{i \in [\log_2(n)]\}}$. Note that $0 \leqslant h_n(i) \leqslant g(i) := 2^{-i} \widetilde{r}^2_{2k}$ for all n, i, and that $\sum_{i=1}^{\infty} g(i) = \widetilde{r}^2_{2k} < \infty$ (i.e. g is integrable with respect to the counting measure). Furthermore, the limit $\lim_{n \to \infty} h_n(i) = 2^{-i} \lim_{n \to \infty} \widetilde{r}^2_{2k, 2^{-i}n}$ exists since Lemma 9 implies that the sequence $\{\widetilde{r}^2_{2k, 2^{-i}n}\}_{n=1}^{\infty}$ is monotonic and bounded for every $i \in \mathbb{N}$. Thus, by Lebesgue's dominated convergence theorem, we have

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

where the last inequality follows from **Step One**. Therefore, $\lim_{n \to \infty} \widetilde{R}_{2k,n} \leqslant 4 r_{2k}$. By Theorem 10 and Theorem 18, this also implies the last two claims in Theorem 20. ∎

F.4. Lower Bounds

The work of [36] proved lower bounds under Assumption 3 that are tight (by our upper bounds) in most parameter regimes for $\gamma = D = \mu = 1$ and $k = \mathcal{O}(1)$.[11] Our (relatively modest) contribution in this subsection is: refining these lower bounds to display the correct dependence on $\gamma, r, \widetilde{r}, D, \mu$; tightening the convex lower bound [36, Theorem 6.4] in the regime $d > n$; and extending [36, Theorems 6.1 and 6.4] to $k \gg 1$. Our first lower bounds hold even for affine functions:

Theorem 21 *Let $k \geqslant 2$, $D, \gamma, r^{(k)}, \widetilde{r}^{(k)} > 0$, $\beta_f \geqslant 0$, $d \geqslant 40, n > 7202$, and $\rho \leqslant d$. Then, for any ρ-zCDP algorithm $\mathcal{A}$, there exist $\mathcal{W}, \mathcal{X} \subset \mathbb{R}^d$ such that $\|w - w'\| \leqslant 2D$ for all $w, w' \in \mathcal{W}$, a β_f-smooth, linear, convex (in w) loss $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, and distributions $\mathcal{D}$ and $\mathcal{D}'$ on $\mathcal{X}$ such that:*
1. Assumption 1 holds and if $X' \sim \mathcal{D}'^n$, then

$$\mathbb{E}F(\mathcal{A}(X')) - F^* = \Omega\left(\widetilde{r}_k D \left(\frac{1}{\sqrt{n}} + \min\left\{1, \left(\frac{\sqrt{d}}{\sqrt{\rho} n}\right)^{\frac{k-1}{k}}\right\}\right)\right). \tag{18}$$

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

$$\mathbb{E}F(\mathcal{A}(X')) - F^* = \Omega\left(r_k D \left(\frac{1}{\sqrt{n}} + \min\left\{1, \left(\frac{\sqrt{d}}{\sqrt{\rho} n}\right)^{\frac{k-1}{k}}\right\}\right)\right). \tag{19}$$

[11] The lower bounds asserted in [36] only hold if $k \lesssim 1$ since the moments of the Gaussian distribution that they construct grow exponentially/factorially with k.

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

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

Proof We will prove part 3 first.
3. We begin by proving the result for $\gamma = D = 1$. In this case, it is proved in [36] that

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

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

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

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

$$\begin{aligned} \mathbb{E}|\langle \nabla f(w,x) - \nabla F(w), e_j\rangle|^k &= \mathbb{E}|\langle -x + \mathbb{E}x, e_j\rangle|^k && (20)\\ &= \mathbb{E}|x - \mu_{\nu_j}|^k && (21)\\ &= \frac{1+\delta_{\nu_j}}{2}|1 - \delta_{\nu_j}|^k + \frac{1-\delta_{\nu_j}}{2}|1 + \delta_{\nu_j}|^k && (22)\\ &\leqslant 1 - \delta_{\nu_j}^2 \leqslant 1 && (23)\end{aligned}$$

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

$$\begin{aligned} \max_{\nu\in\mathcal{V}} \mathbb{E}\left[F_{Q_\nu}(w) - \min_{w'\in\mathcal{W}} F_{Q_\nu}(w')\right] &= \max_{\nu\in\mathcal{V}} \mathbb{E}\left[\left\langle \frac{\mu_\nu}{\|\mu\|}, \mu_\nu \right\rangle - \langle w, \mu_\nu\rangle\right] && (24)\\ &= \max_{\nu\in\mathcal{V}} \mathbb{E}\left[\|\mu\| - \langle w, \mu_\nu\rangle\right] && (25)\\ &= \max_{\nu\in\mathcal{V}} \mathbb{E}\left(\|\mu\|[1 - \langle w, w_\nu\rangle]\right) && (26)\\ &\geqslant \max_{\nu\in\mathcal{V}} \mathbb{E}\left[\frac{1}{2}\|\mu\|\|w - w_\nu\|^2\right], && (27)\end{aligned}$$

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

$$\max_{\nu\in\mathcal{V}} \mathbb{E}\left[F_{Q_\nu}(w) - \min_{w'\in\mathcal{W}} F_{Q_\nu}(w')\right] \geqslant \frac{\|\mu\|}{8} \max_{\nu\in\mathcal{V}} \mathbb{E}\|\hat{w} - w_\nu\|^2. \qquad (28)$$

Now we apply Fano's method (see e.g. [64, Lemma 3]) to lower bound $\max_{\nu \in \mathcal{V}} \mathbb{E}\|\hat{w} - w_\nu\|^2$. For all $\nu \neq \nu'$, we have $\|w_\nu - w_{\nu'}\|^2 \geqslant \frac{\|\nu - \nu'\|^2}{\|\nu\|^2} \geqslant 1$ since $d_{\text{Ham}}(\nu, \nu') \geqslant \frac{d}{2}$ and $\nu \in \{\pm 1\}^d$ implies $\|\nu - \nu'\|^2 \geqslant \frac{d}{2}$ and $\|\nu\|^2 = d$. Also, a straightforward computation shows that for any $j \in [d]$ and $\nu, \nu' \in \mathcal{V}$,

$$D_{KL}(Q_{\nu_j}||Q_{\nu'_j}) \leqslant \frac{1 + \frac{p}{\sqrt{d}}}{2} \left[\log\left(\frac{\sqrt{d} + p}{\sqrt{d}}\right) + \log\left(\frac{\sqrt{d}}{\sqrt{d} - p}\right)\right] \tag{29}$$

$$\leqslant \log\left(\frac{1 + \frac{p}{\sqrt{d}}}{1 - \frac{p}{\sqrt{d}}}\right) \tag{30}$$

$$\leqslant \frac{3p}{\sqrt{d}}, \tag{31}$$

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

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

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

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

which is $\Omega(1)$ for $d \geqslant 40 > 20\log(2)$ and $n > 7202$. Combining this with (28) and plugging in $\|\mu\| = \sqrt{\frac{d}{n}}$ shows that

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

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

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

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

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

$$\widetilde{F}(\tilde{w}) - \widetilde{F}(\tilde{w}^*) = -\langle \tilde{w}, \tilde{\mu}_\nu \rangle + \langle \tilde{w}^*, \tilde{\mu}_\nu \rangle \tag{32}$$

$$= \langle D(w^* - w), \gamma^{1/k}\mu_\nu \rangle \tag{33}$$

$$= D\gamma^{1/k}[F(w) - F(w^*)]. \tag{34}$$

Thus,

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

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

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

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

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

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

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

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

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

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

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

by the proof of [36, Theorem 6.4]. We scale $Q'_\nu \to \widetilde{Q}'_\nu = \frac{\widetilde{r}}{\sqrt{d}} Q'_\nu$, $\mathcal{X} \to \widetilde{\mathcal{X}} = \frac{\widetilde{r}}{\sqrt{d}}\mathcal{X}$, and $\mathcal{W} \to \widetilde{\mathcal{W}} = D\mathcal{W}$. Then for any $\widetilde{w} \in \widetilde{\mathcal{W}}$,

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

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

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

Remark 22 *The main differences in our proof of part 3 of Theorem 21 from the proof of [36, Theorem 6.4] (for $\gamma = D = 1$) are: 1) we construct a Bernoulli product distribution (built on [24, Example 7.7]) instead of a Gaussian, which establishes a lower bound that holds for all $k \geqslant 2$*

instead of just $k = \mathcal{O}(1)$; and 2) we choose a different parameter value (larger p in the notation of the proof) in our application of Fano's method, which results in a tighter lower bound: the term $\min\{1, \sqrt{d/n}\}$ in [36, Theorem 6.4] gets replaced with $\sqrt{d/n}$.[12] *Also, there exist parameter settings for which our lower bound is indeed strictly greater than the lower bound in [36, Theorem 6.4]: for instance, if $d > n > d/\rho$ and $k \to \infty$, then our lower bound simplifies to $\Omega(\sqrt{\frac{d}{n}})$. On the other hand, the lower bound in [36, Theorem 6.4] breaks as $k \to \infty$ (since the k-th moment of their Gaussian goes to infinity); however, even if were extended to $k \to \infty$ (e.g. by replacing their Gaussian with our Bernoulli distribution), then the resulting lower bound $\Omega(1 + \frac{d}{\sqrt{\rho n}})$ would still be smaller than the one we prove above.*[13]

Theorem 23 *Let $k \geqslant 2$, $\mu, \gamma, \tilde{r}_k, r_k > 0$, $n \in \mathbb{N}$, $d \geqslant 40$, and $\rho \leqslant d$. Then, for any ρ-zCDP algorithm $\mathcal{A}$, there exist convex, compact sets $\mathcal{W}, \mathcal{X} \subset \mathbb{R}^d$ of diameter D, a μ-smooth, μ-strongly convex (in w) loss $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, and distributions $\mathcal{D}$ and $\mathcal{D}'$ on $\mathcal{X}$ such that:*
1. Assumption 1 holds with $D \approx \frac{\tilde{r}_k}{\mu}$, and if $X' \sim \mathcal{D}'^n$, then

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

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

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

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

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

Proof We will prove part 3 first. 3. We first consider $\gamma = \mu = 1$ and then scale our hard instance. For $f(w, x) := \frac{1}{2}\|w - x\|^2$, [36] construct a convex/compact domain $\mathcal{W} \times \mathcal{X} \subset \mathbb{R}^d \times \mathbb{R}^d$ and distribution $\mathcal{D}$ on $\mathcal{X}$ such that

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

for any k and any ρ-zCDP algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$ if $X \sim \mathcal{D}^n$.[14] So, it remains to a) prove the first term (d/n) in the lower bound, and then b) show that the scaled instance satisfies the exact

[12] Note that [36, Theorem 6.4] writes $\sqrt{d/n}$ for the first term. However, the proof (see Equation 16 in their paper) only establishes the bound $\min\{1, \sqrt{d/n}\}$.

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

[14] In fact, $\mathcal{W}$ and $\mathcal{X}$ can be chosen to be Euclidean balls of radius $\sqrt{d}p^{-1/k}$ for p defined in the proof of [36, Lemma 6.3], which ensures that $\mathbb{E}\mathcal{D} \in \mathcal{W} = \mathcal{X}$.

hypotheses in the theorem and has excess loss that scales by a factor of $\gamma^{2/k}/\mu$. We start with task a). Observe that for f defined above and any distribution $\mathcal{D}$ such that $\mathbb{E}\mathcal{D} \in \mathcal{W}$, we have

$$\mathbb{E}F(\mathcal{A}(X)) - F^* = \frac{1}{2}\mathbb{E}\|\mathcal{A}(X) - \mathbb{E}\mathcal{D}\|^2 \tag{35}$$

(see [36, Lemma 6.2]), and

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

Thus, it suffices to prove that $\mathbb{E}\|\mathcal{A}(X) - \mathbb{E}\mathcal{D}\|^2 \gtrsim \frac{d}{n}$ for some $\mathcal{D}$ such that $\mathbb{E}|\langle x - \mathbb{E}x, e_j\rangle|^k \leqslant 1$. This is a known result for products of Bernoulli distributions; nevertheless, we provide a detailed proof below. First consider the case $d = 1$. Then the proof follows along the lines of [24, Example 7.7]. Define the following pair of distributions on $\{\pm 1\}$:

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

and

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

for $\delta \in (0,1)$ to be chosen later. Notice that if X is a random variable with distribution P_ν ($\nu \in \{0,1\}$), then $\mathbb{E}|X - \mu|^k \leqslant \mathbb{E}|X|^k \leqslant 1$. Also, $\mathbb{E}P_\nu = \delta\nu$ for $\nu \in \{0,1\}$ and $|\mathbb{E}P_1 - \mathbb{E}P_0| = \delta$ (i.e. the two distributions are δ-separated with respect to the metric $\rho(a,b) = |a-b|$). Then by LeCam's method (see e.g. [24, Eq. 7.33] and take $\Phi(\cdot) = (\cdot)^2$),

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

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

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

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

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

For general $d \geqslant 1$, we take the product distribution $\mathcal{D} := \hat{\mathcal{D}}^d$ on $\mathcal{X} = \{\pm 1\}^d$ and choose $\mathcal{W} = B_2^d(0, \sqrt{d})$ to ensure $\mathbb{E}\mathcal{D} \in \mathcal{W}$ (so that (35) holds). Clearly, $\mathbb{E}|\langle \mathcal{D} - \mathbb{E}\mathcal{D}, e_j\rangle|^k \leqslant 1$ for all $j \in [d]$. Further, the mean squared error of any algorithm for estimating the mean of $\mathcal{D}$ is

$$\mathbb{E}_{X \sim \mathcal{D}^n}\|\mathcal{A}(X) - \mathbb{E}\mathcal{D}\|^2 = \sum_{j=1}^{d}\mathbb{E}|\mathcal{A}(X)_j - \mathbb{E}\mathcal{D}_j|^2 \geqslant \frac{d}{64n}, \tag{36}$$

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

Next, we move to task b). For this, we re-scale each of our hard distributions (non-private given above, and private given in the proof of [36, Lemma 6.3] and below in our proof of part 2 of the theorem–see (40)): $\mathcal{D} \to \frac{\gamma^{1/k}}{\mu}\mathcal{D} = \widetilde{\mathcal{D}}$, $\mathcal{X} \to \frac{\gamma^{1/k}}{\mu}\mathcal{X} = \widetilde{\mathcal{X}}$, $\mathcal{W} \to \frac{\gamma^{1/k}}{\mu}\mathcal{W} = \widetilde{\mathcal{W}}$ and $f : \mathcal{W} \times \mathcal{X} \to \mu f = \widetilde{f} : \widetilde{\mathcal{W}} \times \widetilde{\mathcal{X}}$. Then $\widetilde{f}(\cdot, \widetilde{x})$ is μ-strongly convex and μ-smooth for all $\widetilde{x} \in \widetilde{\mathcal{X}}$ and

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

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

$$\begin{aligned} \widetilde{F}(\widetilde{w}) - \widetilde{F}(\widetilde{w}^*) &= \frac{\mu}{2}\mathbb{E}\left[\|\widetilde{w} - \widetilde{x}\|^2 - \|\widetilde{w}^* - \widetilde{x}\|^2\right] && (37)\\ &= \frac{\mu}{2}\left(\frac{\gamma^{2/k}}{\mu^2}\right)\mathbb{E}\left[\|w - x\|^2 - \|w^* - x\|^2\right] && (38)\\ &= \frac{\gamma^{2/k}}{\mu}[F(w) - F(w^*)]. && (39) \end{aligned}$$

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

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

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

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

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

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

which follows by combining part 3a) above with [36, Lemma 6.3], completes the proof of part 3.

1. We begin by proving the first (non-private) term in the lower bound: For our *unscaled* hard instance, we will take the same distribution $\mathcal{D} = P_\nu^d$ (for some $\nu \in \{0, 1\}$) on $\mathcal{X} = \{\pm 1\}^d$ and quadratic f described above in part 1a with $\mathcal{W} := B_2^d(0, \sqrt{d})$. The choice of $\mathcal{W}$ ensures $\mathbb{E}\mathcal{D} \in \mathcal{W}$ so that (35) holds. Further,

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

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

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

Moreover, if $\left(\frac{\widetilde{r}_k}{3\mu\sqrt{d}}\right)\mathcal{A} = \widetilde{\mathcal{A}} : \widetilde{X}^n \to \widetilde{\mathcal{W}}$ is any algorithm and $\widetilde{X} \sim \widetilde{\mathcal{D}}^n$, then by (36) and (35), we have

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

Next, we prove the second (private) term in the lower bound. Let f be as defined above. For our unscaled hard distribution, we follow [8, 36] and define a family of distributions $\{Q_\nu\}_{\nu\in\mathcal{V}}$ on $\mathbb{R}^d$, where $\mathcal{V} \subset \{\pm 1\}^d$ will be defined later. For any given $\nu \in \mathcal{V}$, we define the distribution Q_ν as follows: $X_\nu \sim Q_\nu$ iff

$$X_\nu = \begin{cases} 0 & \text{with probability } 1-p \\ p^{-1/k}\nu & \text{with probability } p \end{cases} \tag{40}$$

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

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

Note also that our choice of $\mathcal{W}$ and $p \leqslant 1$ ensures that $\mathbb{E}[Q_\nu] \in \mathcal{W}$. Moreover, as in the proof of [36, Lemma 6.3], zCDP Fano's inequality (see [36, Theorem 1.4]) implies that for any ρ-zCDP algorithm $\mathcal{A}$,

$$\sup_{\nu\in\mathcal{V}} \mathbb{E}_{X\sim Q_\nu^n, \mathcal{A}}\|\mathcal{A}(X) - \mathbb{E}Q_\nu\|^2 = \Omega\left(d\min\left\{1, \left(\frac{\sqrt{d}}{n\sqrt{\rho}}\right)^{\frac{2k-2}{k}}\right\}\right). \tag{41}$$

Thus,

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

for some $\nu \in \mathcal{V}$, by (35). Now we scale our hard instance: $f \to \widetilde{f} = \mu f$, $\mathcal{W} \to \widetilde{\mathcal{W}} := \frac{\widetilde{r}_k}{2\mu\sqrt{d}}\mathcal{W}$, $\mathcal{X} \to \widetilde{\mathcal{X}} := \frac{\widetilde{r}_k}{2\mu\sqrt{d}}\mathcal{X}$ and $\mathcal{D} \to \widetilde{\mathcal{D}} = \frac{\widetilde{r}_k}{2\mu\sqrt{d}}\mathcal{D}$. Then $\widetilde{f}(\cdot, \widetilde{x})$ is μ-strongly convex and μ-smooth, and

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

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

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

by (41).

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

Remark 24 *Note that the lower bound proofs construct bounded (hence subexponential) distributions and uniformly L_f-Lipschitz, β_f-smooth losses that easily satisfy the conditions in Theorem 20.*

Appendix G. Details and Proofs of Results in Section 3

G.1. Noisy Clipped Accelerated SGD for Smooth Convex Losses

We present our generic framework for accelerated SO with biased/noisy stochastic graadients in Algorithm 4. It is built on the AC-SA of [31], and is the first accelerated DP algorithm for heavy-tailed SO.

Algorithm 4 Generic Framework for DP Accelerated Stochastic Approximation (AC-SA)

```
Input: Data samples $X \in \mathcal{X}^n$, number of iterations $T \leqslant n$, convex loss $f(\cdot, x)$, step size
    parameters $\{\eta_t\}_{t\in[T]}, \{\alpha_t\}_{t\in[T]}$ such that $\alpha_1 = 1, \alpha_t \in (0,1)$ for all $t \geqslant 2$, DP mean estimation
    oracle MeanOracle (and its hyperparameters), privacy parameter $\rho = \epsilon^2/2$.
Initialize $w_0^{ag} = w_0 \in \mathcal{W}$ and $t = 1$.
for $t \in [T]$ do
    $w_t^{md} := (1-\alpha_t)w_{t-1}^{ag} + \alpha_t w_{t-1}$.
    Draw new batch $\mathcal{B}_t$ (without replacement) of $n/T$ samples from $X$.
    $\widetilde{\nabla} F_t(w_t^{md}) :=$ MeanOracle($\{\nabla f(w_t^{md}, x)\}_{x\in\mathcal{B}_t}; \frac{n}{T}; \frac{\epsilon^2}{2}$)
    $w_t := \mathrm{argmin}_{w\in\mathcal{W}}\left\{\alpha_t\langle\widetilde{\nabla} F_t(w_t^{md}), w\rangle + \frac{\eta_t}{2}\|w_{t-1} - w\|^2\right\}$.
    $w_t^{ag} := \alpha_t w_t + (1-\alpha_t)w_{t-1}^{ag}$.
end for
Output: $w_T^{ag}$.
```

As a first step towards bounding the excess risk of our algorithm, we provide excess risk guarantees for Algorithm 4 in terms of the bias and variance of the MeanOracle:

Proposition 25 (Informal) *Consider Algorithm 4 run with a* `MeanOracle` *satisfying $\widetilde{\nabla} F_t(w_t^{md}) = \nabla F(w_t^{md}) + b_t + N_t$, where $\|b_t\| \leqslant B$ (with probability 1), $\mathbb{E}N_t = 0$, $\mathbb{E}\|N_t\|^2 \leqslant \Sigma^2$ for all $t \in [T-1]$, and $\{N_t\}_{t=1}^T$ are independent. Assume that $F : \mathcal{W} \to \mathbb{R}$ is convex and β-smooth. Then there are choices of algorithmic parameters in Algorithm 4 such that:*

$$\mathbb{E}F(w_T^{ag}) - F^* \lesssim \frac{\beta D^2}{T^2} + \frac{D(\Sigma + B)}{\sqrt{T}} + BD. \tag{42}$$

The proof of Proposition 25, given towards the end of this subsection, involves a careful analysis of the bias and noise that propogates throughout the algorithm, and uses tools from [31].[15] We shall apply Proposition 25 with $b_t := \mathbb{E}\widetilde{\nabla} F_t(w_t) - \nabla F(w_t)$ and $N_t := \widetilde{\nabla} F_t(w_t) - \mathbb{E}\widetilde{\nabla} F_t(w_t)$ for $\widetilde{\nabla} F_t(w_t)$ given by Algorithm 1. Then, combining Proposition 25 with Lemma 8 yields the main result of this subsection:

Theorem 26 (Smooth, Convex - Informal) *Let $\epsilon > 0$ and assume F is convex and β-smooth. Then, there are parameters such that Algorithm 4 instantiated with* `MeanOracle` *Algorithm 1 is $\frac{\epsilon^2}{2}$-zCDP and*

$$\mathbb{E}F(w_T^{ag}) - F^* \lesssim r_k D \left[\frac{1}{\sqrt{n}} + \max\left\{ \left(\left(\frac{\beta D}{r_k}\right)^{1/4} \frac{\sqrt{d}}{\epsilon n} \right)^{\frac{4(k-1)}{5k-1}}, \left(\frac{\sqrt{d}}{\epsilon n}\right)^{\frac{k-1}{k}} \right\} \right]. \tag{43}$$

The full statement of Theorem 26 and its proof is given at the end of this subsection.

Remark 27 (Optimal rate for "sufficiently smooth" convex functions) *Notice that the upper bound in Theorem 26 scales with the smoothness parameter β. Thus, for sufficiently small β, the optimal rates (see Theorem 21) are attained. For example, when $k = 2$, the upper bound in (43) matches the respective lower bound in Theorem 21 when $\beta \lesssim \frac{r_k}{D}\left(\frac{d^5}{\epsilon n}\right)^{1/18}$; this would be the case for example if β and D are constants and $d \geqslant (\epsilon n)^{1/5}$. In particular, for affine functions–which were not addressed in prior works [36, 62] since these works assumed $\nabla F(w^*) = 0$–we have $\beta = 0$, so that Algorithm 4 is optimal (up to constant factors) for all $k \geqslant 2$.*[16]

Having discussed the dependence on β, let us focus on understanding how the bound in Theorem 26 scales with n, d and ϵ. Thus, let us fix $\beta = D = \gamma = 1$ and $r = \sqrt{d}$ for simplicity. If $k = 2$, then the bound in (43) simplifies to $\mathcal{O}\left(\sqrt{\frac{d}{n}} + \max\left\{\frac{d^{2/3}}{(\epsilon n)^{4/9}}, \frac{d^{3/4}}{\sqrt{\epsilon n}}\right\}\right)$, whereas the lower bound in Theorem 21 (part 2) is $\Omega\left(\sqrt{\frac{d}{n}} + \frac{d^{3/4}}{\sqrt{\epsilon n}}\right)$. Therefore, the bound in (43) is tight if $d^{3/2} \gtrsim \epsilon n$. For general n, d, ϵ, (43) is *nearly* tight up to a multiplicative factor of $\left(\frac{\epsilon n}{d^{3/2}}\right)^{1/18}$. By comparison, the previous state-of-the-art (*not linear time*) bound for $\epsilon \approx 1$ was $\mathcal{O}\left(\frac{d}{\sqrt{n}}\right)$ [36, Theorem 5.4]. Our bound (43) improves over [36, Theorem 5.4] if $d \gtrsim n^{1/6}$, which is typical in practical ML applications. As $k \to \infty$, (43) becomes

[15] Our analysis can be extended to the strongly convex setting, resulting in a bound that is essentially the same as the strongly convex bounds in [31, Theorem 1 and Proposition 7], plus a term scaling with BD. However, the bias term BD is too large to yield near-optimal strongly convex rates, which is why we only use acceleration in the convex case.

[16] An affine function is a function that is linear in w: i.e. $\nabla f(w, x)$ does not depend on w, hence $\nabla^2_{ww} f(w, x) = 0$. The assumption made in [36, 62] that $\nabla F(w^*) = 0$ is needed for the mean oracle of [33], which is used in [36, 62]. Also, note that the lower bound construction in Theorem 21 uses an affine function.

$\mathcal{O}\left(\sqrt{\frac{d}{n}} + \left(\frac{d}{n}\right)^{4/5}\right)$ for $\epsilon \approx 1$, which is strictly better than the bound in [36, Theorem 5.4]. Additionally, the gradient complexity of our algorithm is n, which is superior to the $\mathcal{O}(n^2/d)$ complexity of the algorithm in [36].

Next, we provide the formal versions and proofs of results presented above.

Proposition 28 (Complete Version of Proposition 25) *Consider Algorithm 4 run with a* `MeanOracle` *satisfying $\tilde{\nabla} F_t(w_t^{md}) = \nabla F(w_t^{md}) + b_t + N_t$, where $\|b_t\| \leqslant B$ (with probability 1), $\mathbb{E}N_t = 0$, $\mathbb{E}\|N_t\|^2 \leqslant \Sigma^2$ for all $t \in [T-1]$, and $\{N_t\}_{t=1}^T$ are independent. Assume that $F : \mathcal{W} \to \mathbb{R}$ is convex and β-smooth, $F(w_0) - F^* \leqslant \Delta$, and $\|w_0 - w^*\| \leqslant D$. Suppose parameters are chosen in Algorithm 4 so that for all $t \in [T]$, $\eta_t > \beta\alpha_t^2$ and $\eta_t/\Gamma_t = \eta_1/\Gamma_1$, where*

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

Then,

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

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

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

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

$$\mathbb{E}F(w_T^{ag}) - F^* \lesssim \frac{\beta D^2}{T^2} + \frac{D(\Sigma + B)}{\sqrt{T}} + BD. \tag{44}$$

Proof We begin by extending [31, Proposition 4] to biased/noisy stochastic gradients. Fix any $w_{t-1}, w_{t-1}^{ag} \in \mathcal{W}$. By [31, Lemma 3], we have

$$F(w_t^{ag}) \leqslant (1-\alpha_t)F(w_{t-1}^{ag}) + \alpha[F(z) + \langle \nabla F(z), w_t - z\rangle] + \frac{\beta}{2}\|w_t^{ag} - z\|^2, \tag{45}$$

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

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

and $d_t := w_t^{ag} - w_t^{md} = \alpha_t(w_t - w_{t-1})$. Then using (45) with $z = w_t^{md}$, we have

$$\begin{aligned} F(w_t^{ag}) &\leqslant (1-\alpha_t)F(w_{t-1}^{ag}) + \alpha_t[F(w_t^{md}) + \langle \nabla F(w_t^{md}), w_t - w_t^{md}\rangle] + \frac{\beta}{2}\|d_t\|^2 \\ &= (1-\alpha_t)F(w_{t-1}^{ag}) + \alpha_t[F(w_t^{md}) + \langle \nabla F(w_t^{md}), w_t - w_t^{md}\rangle] + \frac{\eta_t}{2}\|w_{t-1} - w_t\|^2 - \frac{\eta_t - \beta\alpha_t^2}{2\alpha_t^2}\|d_t\|^2, \end{aligned} \tag{46}$$

by the expression for d_t. Now we apply [31, Lemma 2] with $p(u) = \alpha_t[\langle\widetilde{\nabla}F_t(w_t^{md}, u\rangle], \mu_1 = 0, \mu_2 = \eta_t, \tilde{x} = w_t^{md}$, and $\tilde{y} = w_{t-1}$ to obtain (conditional on all randomness) for any $w \in \mathcal{W}$:

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

Next, we combine the above inequality with (46) to get

$$\begin{aligned}F(w_t^{ag}) \leqslant (1-\alpha_t)F(w_{t-1}^{ag}) + \alpha_t[F(w_t^{md}) + \langle\nabla F(w_t^{md}), w - w_t^{md}\rangle] + \frac{\eta_t}{2}\left[\|w_{t-1} - w\|^2 - \|w_t - w\|^2\right] \\ + \underbrace{-\frac{\eta_t - \beta\alpha_t^2}{2\alpha_t^2}\|d_t\|^2 + \alpha_t\langle N_t + b_t, w - w_t\rangle}_{U_t}, \qquad (47)\end{aligned}$$

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

$$\begin{aligned}U_t &\leqslant -\frac{\eta_t - \beta\alpha_t^2}{2\alpha_t^2}\|d_t\|^2 + \|N_t + b_t\|\|d_t\| + \alpha_t\langle N_t + b_t, w - w_{t-1}\rangle \\ &\leqslant \Upsilon_t(w), \qquad (48)\end{aligned}$$

where the last inequality follows from maximizing the concave quadratic function $q(\|d_t\|) := -\left[\frac{\eta_t - \beta\alpha_t^2}{2\alpha_t^2}\right]\|d_t\|^2 + \|N_t + b_t\|\|d_t\|$ with respect to $\|d_t\|$. Plugging the bound (48) back into (47) shows that

$$\begin{aligned}F(w_t^{ag}) \leqslant (1-\alpha_t)F(w_{t-1}^{ag}) + \alpha_t[F(w_t^{md}) + \langle\nabla F(w_t^{md}), w - w_t^{md}\rangle] + \frac{\eta_t}{2}\left[\|w_{t-1} - w\|^2 - \|w_t - w\|^2\right] \\ + \Upsilon_t(w). \qquad (49)\end{aligned}$$

Then it can be shown (see [31, Proposition 5]) that the assumptions on η_t and α_t imply that

$$F(w_T^{ag}) - \Gamma_T\sum_{t=1}^T\left[\frac{\alpha_t}{\Gamma_t}\left(F(w_t^{md}) + \langle\nabla F(w_t^{md}), w - w_t^{md}\rangle\right)\right] \leqslant \Gamma_T\sum_{t=1}^T\frac{\eta_t}{2\Gamma_t}[\|w_{t-1} - w\|^2 - \|w_t - w\|^2] \qquad (50)$$

$$+ \Gamma_T\sum_{t=1}^T\frac{\Upsilon_t(w)}{\Gamma_t}, \qquad (51)$$

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

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

by definition. Hence by convexity of F,

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

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

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

since $\Gamma_1 = 1$. Substituting the above bounds into (50), we get

$$F(w_T^{ag}) - F(w) \leqslant \Gamma_T \eta_1 \frac{1}{2}\|w_0 - w\|^2 + \Gamma_T \sum_{t=1}^{T} \frac{\Upsilon_t(w)}{\Gamma_t}, \ \forall w \in \mathcal{W}. \tag{52}$$

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

$$\begin{aligned}
\mathbb{E}[F(w_T^{ag}) - F^*] &\leqslant \frac{\Gamma_T \eta_1 D^2}{2} + \Gamma_T \sum_{t=1}^{T} \frac{\mathbb{E}\Upsilon_t(w^*)}{\Gamma_t} & (53)\\
&\leqslant \frac{\Gamma_T \eta_1 D^2}{2} + \Gamma_T \sum_{t=1}^{T} \left[\frac{1}{\Gamma_t}\left(\alpha_t \mathbb{E}\langle b_t, w^* - w_{t-1}\rangle + \frac{2\alpha_t^2(\Sigma^2 + B^2)}{\eta_t - \beta\alpha_t^2}\right)\right] & (54)\\
&\leqslant \frac{\Gamma_T \eta_1 D^2}{2} + \Gamma_T \sum_{t=1}^{T} \left[\frac{1}{\Gamma_t}\left(\alpha_t BD + \frac{2\alpha_t^2(\Sigma^2 + B^2)}{\eta_t - \beta\alpha_t^2}\right)\right], & (55)
\end{aligned}$$

where we used conditional independence of N_t and $w^* - w_{t-1}$ given w_{t-1}, Young's inequality, Cauchy-Schwartz, and the definitions of B^2 and Σ^2. This establishes the first claim of the theorem. The second and third claims are simple corollaries, which can be verified as in [31, Proposition 7] and the ensuing discussion. ■

Theorem 29 (Complete Version of Theorem 26) *Let $\epsilon > 0$ and assume F is convex and β-smooth. Then, there are parameters such that Algorithm 4 instantiated with* `MeanOracle` *Algorithm 1 is $\frac{\epsilon^2}{2}$-zCDP. Further, if $n \geqslant T := \left\lceil \min\left\{\left(\frac{\beta D}{r_k}\right)^{2k/(5k-1)}\left(\frac{\epsilon n}{\sqrt{d}}\right)^{(2k-2)/(5k-1)}, \sqrt{\frac{\beta D}{r}} n^{1/4}\right\}\right\rceil$, then,*

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

Proof

Privacy: Choose $\sigma^2 = \frac{4C^2T^2}{\epsilon^2 n^2}$. First, the collection of all $\widetilde{\nabla} F_t(w_t^{md}),\ t \in [T]$ is $\frac{\epsilon^2}{2}$-zCDP: since the batches of data drawn in each iteration are disjoint, it suffices (by parallel composition [48]) to show that $\widetilde{\nabla} F_t(w_t^{md})$ is $\frac{\epsilon^2}{2}$-zCDP for all t. Now, the ℓ_2 sensitivity of each clipped gradient update is bounded by $\Delta = \sup_{w, X\sim X'} \|\frac{T}{n}\sum_{x\in\mathcal{B}_t} \Pi_C(\nabla f(w,x)) - \sum_{x'\in\mathcal{B}_t'} \Pi_C(\nabla f(w,x'))\| = \sup_{w,x,x'} \|\frac{T}{n}\Pi_C(\nabla f(w,x)) - \Pi_C(\nabla f(w,x'))\| \leqslant \frac{2CT}{n}$. Thus, $\widetilde{\nabla} F_t(w_t^{md})$ is $\frac{\epsilon^2}{2}$-zCDP by Proposition 6. Second, the iterates w_t^{ag} are deterministic functions of $\widetilde{\nabla} F_t(w_t^{md})$, so the post-processing property of differential privacy [16, 26] ensures that Algorithm 4 is $\frac{\epsilon^2}{2}$-zCDP.

Excess risk: Consider round $t \in [T]$ of Algorithm 4, where Algorithm 1 is run on input data $\{\nabla f(w_t, x_i^t)\}_{i=1}^{n/T}$. Denote the bias of Algorithm 1 by $b_t := \mathbb{E}\widetilde{\nabla} F_t(w_t) - \nabla F(w_t)$, where $\widetilde{\nabla} F_t(w_t) =$

$\tilde{\nu}$ in the notation of Algorithm 1. Also let $\widehat{\nabla}F_t(w_t) := \hat{\mu}$ (in the notation of Lemma 8) and denote the noise by $N_t = \widetilde{\nabla}F_t(w_t) - \nabla F(w_t) - b_t = \widetilde{\nabla}F_t(w_t) - \mathbb{E}\widetilde{\nabla}F_t(w_t)$. Then we have $B := \sup_{t \in [T]} \|b_t\| \leqslant \frac{r^k}{(k-1)C^{k-1}}$ and $\Sigma^2 := \sup_{t \in [T]} \mathbb{E}[\|N_t\|^2] \leqslant d\sigma^2 + \frac{r^2 T}{n} \lesssim \frac{dC^2T^2}{\epsilon^2 n^2} + \frac{r^2 T}{n}$, by Lemma 8. Plugging these estimates for B and Σ^2 into Proposition 25 and setting $C = r(\frac{\epsilon n}{\sqrt{dT}})^{1/k}$, we get

$$\begin{aligned}
\mathbb{E}F(w_T^{ag}) - F^* &\lesssim \frac{\beta D^2}{T^2} + \frac{D(\Sigma + B)}{\sqrt{T}} + BD \\
&\lesssim \frac{\beta D^2}{T^2} + \frac{CD\sqrt{dT}}{\epsilon n} + \frac{rD}{\sqrt{n}} + \frac{r^k D}{C^{k-1}} \\
&\lesssim \frac{\beta D^2}{T^2} + rD\left[\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{dT}}{\epsilon n}\right)^{(k-1)/k}\right]. \qquad (56)
\end{aligned}$$

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

G.2. Noisy Clipped SGD for Strongly Convex Losses

Our framework for strongly convex losses, given in Algorithm 5, is to run SGD with biased/noisy stochastic gradients given by some `MeanOracle`. This general framework is similar to the frameworks of [36, 62], but with some key differences. The main differences in our approach lie in the choice of algorithmic parameters (including `MeanOracle`, step size, and iterate averaging weights), as well as in our *analysis* of the algorithm.

Algorithm 5 Noisy SGD Framework for Heavy-Tailed SCO

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

We have the following privacy and utility guarantees for Algorithm 5:

Theorem 30 (Smooth, Strongly Convex) *Let $\epsilon > 0$, and assume F is μ-strongly convex and β-smooth with $\kappa = \frac{\beta}{\mu} \leqslant n/\ln(n)$. Then, there are parameters such that Algorithm 5 instantiated with `MeanOracle` Algorithm 1 is $\frac{\epsilon^2}{2}$-zCDP, and*

$$\mathbb{E}F(\widehat{w}_T) - F^* \lesssim \frac{r_k^2}{\mu}\left(\frac{1}{n} + \left(\frac{\sqrt{d\kappa\ln(n)}}{\epsilon n}\right)^{\frac{2k-2}{k}}\right). \qquad (57)$$

To obtain Theorem 30, we will apply Proposition 31 to the biased noisy stochastic gradients provided by the private mean oracle Algorithm 1. Combining Proposition 31 with Lemma 8 and a suitable choice of algorithmic parameters yields the excess risk bounds in Theorem 30. The full proof is deferred to the end of this subsection. As we saw in Theorem 23, the bound (57) is optimal up to a factor of $\widetilde{\mathcal{O}}(\kappa^{(k-1)/k})$.

To prove the excess risk bounds in Theorem 30, we first derive a novel convergence guarantee for projected SGD with generic noisy biased stochastic gradients:

Proposition 31 *Let $F : \mathcal{W} \to \mathbb{R}$ be μ-strongly convex and β-smooth with condition number $\kappa := \frac{\beta}{\mu}$. Let $w_{t+1} := \Pi_{\mathcal{W}}[w_t - \eta_t \widetilde{\nabla} F_t(w_t)]$, where $\widetilde{\nabla} F_t(w_t) = \nabla F(w_t) + b_t + N_t$, such that the bias and noise (which can depend on w_t and the samples drawn) satisfy $\|b_t\| \leqslant B$ (with probability 1), $\mathbb{E}N_t = 0$, $\mathbb{E}\|N_t\|^2 \leqslant \Sigma^2$ for all $t \in [T-1]$, and that $\{N_t\}_{t=1}^T$ are independent. Then, there exist stepsizes $\{\eta_t\}_{t=1}^T$ and weights $\{\zeta_t\}_{t=0}^T$ such that the average iterate $\widehat{w}_T := \frac{1}{\sum_{t=0}^T \zeta_t} \sum_{t=0}^T \zeta_t w_{t+1}$ satisfies*

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

Compared to the results in [6] for (non-strongly) convex DP ERM and [3] for non-private unconstrained PL losses, Proposition 31 is tighter, since we leverage smoothness and strong convexity. Our analysis also corrects the issue in [36, 62].

Proof [Proof of Proposition 31] Define $g(w_t) = -\frac{1}{\eta_t}(w_{t+1} - w_t)$. Then

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

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

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

where we used the fact that $\langle \Pi_{\mathcal{W}}(y) - x, \Pi_{\mathcal{W}}(y) - y \rangle \leqslant 0$ for all $x \in \mathcal{W}, y \in \mathbb{R}^d$ (c.f. [15, Lemma 3.1]) to obtain the last inequality. Thus,

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

Combining the above inequality with (58), we get

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

$$+ 2\eta_t \left(\frac{\eta_t^2 \beta}{2} - \eta_t \right) \mathbb{E}\|g(w_t)\|^2. \tag{60}$$

Next, consider

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

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

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

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

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

by Young's inequality. Therefore,

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

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

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

Next, we apply Lemma 32 (see below) with $r_t := \mathbb{E}\|w_t - w^*\|^2$, $s_t := \mathbb{E}F(w_{t+1}) - F^* - \frac{2B^2}{\mu}$, $a := \frac{\mu}{2}$, $b := 2$, $c = 2\Sigma^2$, and $g = \beta$. We may assume $s_t \geqslant 0$ for all t: if this inequality breaks for some t, then simply return w_{t+1} instead of $\widehat{w}_T$ to obtain $\mathbb{E}F(w_t) - F^* < \frac{2B^2}{\mu}$. Thus,

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

Finally, Jensen's inequality yields the theorem. ■

Lemma 32 *[55, Lemma 3] Let $b > 0$, let $a, c \geqslant 0$, and $\{\eta_t\}_{t\geqslant 0}$ be non-negative step-sizes such that $\eta_t \leqslant \frac{1}{g}$ for all $t \geqslant 0$ for some parameter $g \geqslant a$. Let $\{r_t\}_{t\geqslant 0}$ and $\{s_t\}_{t\geqslant 0}$ be two non-negative sequences of real numbers which satisfy*

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

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

$$\frac{b}{\Gamma_T}\sum_{t=0}^{T}s_t\zeta_t + ar_{T+1} \leqslant 32gr_0\exp\left(\frac{-aT}{2g}\right) + \frac{36c}{aT},$$

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

Theorem 33 (Re-statement of Theorem 30) *Let $\epsilon > 0$, and assume F is μ-strongly convex and β-smooth with $\kappa = \frac{\beta}{\mu} \leqslant n/\ln(n)$. Then, there are parameters such that Algorithm 5 instantiated with `MeanOracle` Algorithm 1 is $\frac{\epsilon^2}{2}$-zCDP, and*

$$\mathbb{E}F(\widehat{w}_T) - F^* \lesssim \frac{r_k^2}{\mu}\left(\frac{1}{n} + \left(\frac{\sqrt{d\kappa\ln(n)}}{\epsilon n}\right)^{\frac{2k-2}{k}}\right). \tag{61}$$

Proof **Privacy:** Choose $\sigma^2 = \frac{4C^2T^2}{\epsilon^2 n^2}$. Since the batches of data drawn in each iteration are disjoint, it suffices (by parallel composition [48]) to show that $\widetilde{\nabla}F_t(w_t)$ is $\frac{\epsilon^2}{2}$-zCDP for all t. Now, the ℓ_2 sensitivity of each clipped gradient update is bounded by $\Delta = \sup_{w,X\sim X'}\|\frac{T}{n}\sum_{x\in\mathcal{B}_t}\Pi_C(\nabla f(w,x)) - \sum_{x'\in\mathcal{B}'_t}\Pi_C(\nabla f(w,x'))\| = \sup_{w,x,x'}\|\frac{T}{n}\Pi_C(\nabla f(w,x)) - \Pi_C(\nabla f(w,x'))\| \leqslant \frac{2CT}{n}$. Hence Proposition 6 implies that the algorithm is $\frac{\epsilon^2}{2}$-zCDP.

Excess risk: For any iteration $t \in [T]$, denote the bias of Algorithm 1 by $b_t := \mathbb{E}\widetilde{\nabla} F_t(w_t) - \nabla F(w_t)$, where $\widetilde{\nabla} F_t(w_t) = \tilde{\nu}$ in the notation of Algorithm 1. Also let $\widehat{\nabla} F_t(w_t) := \hat{\nu}$ (in the notation of Lemma 8) and denote the noise by $N_t = \widetilde{\nabla} F_t(w_t) - \nabla F(w_t) - b_t = \widetilde{\nabla} F_t(w_t) - \mathbb{E}\widetilde{\nabla} F_t(w_t)$. Then we have $B := \sup_{t\in[T]} \|b_t\| \leqslant \frac{r^{(k)}}{(k-1)C^{k-1}}$ and $\Sigma^2 := \sup_{t\in[T]} \mathbb{E}[\|N_t\|^2] \leqslant d\sigma^2 + \frac{r_k^2 T}{n} \lesssim \frac{dC^2T^2}{\epsilon^2 n^2} + \frac{r_k^2 T}{n}$, by Lemma 8. Plugging these bias and variance estimates into Proposition 31, we get

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

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

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

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

■

Appendix H. Details and Proofs of Results in Section 4

Assume: $f(w, x) = f^0(w, x) + f^1(w)$; $f^0(\cdot, x)$ is differentiable (maybe non-convex), f^1 is proper, closed, and convex (maybe non-differentiable) for all $x \in \mathcal{X}$; and $F(w) = F^0(w) + f^1(w) = \mathbb{E}_{x\sim\mathcal{D}}[f^0(w, x)] + f^1(w)$ satisfies the *Proximal-PL* condition [37]:

Definition 34 (μ-PPL) *Let $F(w) = F^0(w) + f^1(w)$ be bounded below; F^0 is β-smooth and f^1 is convex. F satisfies Proximal Polyak-Łojasiewicz inequality with parameter $\mu > 0$ if*

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

Theorem 34 is an extension of the classical PL inequality [50], allowing for constrained optimization and/or non-smooth regularizer, depending on the choice of f^1. For PPL losses, we propose Algorithm 6, which is a linear time algorithm. Recall that the *proximal operator* of a convex function g is defined as

$$\texttt{prox}_{\eta g}(z) := \underset{y\in\mathbb{R}^d}{\operatorname{argmin}}\left(\eta g(y) + \frac{1}{2}\|y - z\|^2\right), \text{ for } \eta > 0.$$

Proximal operators generalize projections: if $g = \iota_{\mathcal{W}}$, then $\texttt{prox}_g(z) = \Pi_{\mathcal{W}}(z) := \operatorname{argmin}_{y\in\mathcal{W}} \|y - z\|^2$. The privacy and excess risk guarantees of Algorithm 6 are provided in Theorem 35:

Theorem 35 (Proximal-PL) *Let $\epsilon > 0$ and assume $F(w) = F^0(w) + f^1(w)$ is μ-PPL for β-smooth F^0, with $\kappa = \frac{\beta}{\mu} \leqslant n/\ln(n)$. Then, there are parameters such that Algorithm 6 is $\frac{\epsilon^2}{2}$-zCDP, and*

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

Algorithm 6 Noisy Proximal SGD Framework for Heavy-Tailed SO

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

The bound in Theorem 35 nearly matches the *smooth, strongly convex* lower bound in Theorem 23 up to the $\widetilde{\mathcal{O}}(\kappa^{(2k-2)/2})$ factor, and is attained without convexity. In particular, Algorithm 6 is nearly optimal.[17] To prove Theorem 35, we first derive Proposition 36, a generic convergence guarantee for Algorithm 6 with biased, noisy stochastic gradients in terms of the bias and variance of the oracle:

Proposition 36 *Consider Algorithm 6 with biased, noisy stochastic gradients:* $\widetilde{\nabla} F_t^0(w_t) = \nabla F^0(w_t) + b_t + N_t$, *and stepsize* $\eta = \frac{1}{2\beta}$. *Assume that the bias and noise satisfy* $\|b_t\| \leqslant B$ *(with probability* 1*),* $\mathbb{E}N_t = 0$, $\mathbb{E}\|N_t\|^2 \leqslant \Sigma^2$ *for all* $t \in [T-1]$, *and that* $\{N_t\}_{t=1}^T$ *are independent. Assume further that* F *is* μ*-PPL,* F^0 *is* β*-smooth, and* $F(w_0) - F^* \leqslant \Delta$. *Then,*

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

Proposition 36 generalizes [3, Theorem 6]–which provides a similar bound for the unconstrained, classical PL problem–to the proximal setting. However, the proof of Proposition 36 is very different from the proof of [3, Theorem 6], since the proximal operator makes it difficult to bound the excess loss without convexity when the stochastic gradients are biased/noisy. Our proof draws inspiration from the proof of [45, Theorem 3.1] (for the case of *unbiased* stochastic gradients and *Lipschitz* loss function). Specifically, we view each biased/noisy proximal evaluation as an *objective perturbation* [19] problem. Then, using techniques from the analysis of objective perturbation, we bound the difference between the errors of the biased, noisy stochastic proximal gradient steps and the unbiased noiseless proximal gradient steps, the latter of which can be bounded via the PPL inequality. Compared to the proof of [45, Theorem 3.1], here we need to carefully handle the bias term and bound the error without appealing to Lipschitz continuity of f.

[17] Since any smooth, strongly function satisfies the PPL condition [37], the lower bounds in Theorem 23 also apply to the PPL function class considered here.

Proof [Proof of Proposition 36] Our proof extends the ideas in [45] to generic *biased* and noisy gradients without using Lipschitzness of the loss. By β-smoothness, for any $r \in [T-1]$, we have

$$\begin{aligned}\mathbb{E}F(w_{r+1}) &= \mathbb{E}[F^0(w_{r+1}) + f^1(w_r) + f^1(w_{r+1}) - f^1(w_r)] \\ &\leqslant \mathbb{E}\left\{F(w_r) + \left[\langle \widetilde{\nabla} F_r^0(w_r), w_{r+1} - w_r\rangle + \frac{\beta}{2}\|w_{r+1} - w_r\|^2 + f^1(w_{r+1}) - f^1(w_r)\right]\right\} \\ &\quad + \mathbb{E}\langle \nabla F^0(w_r) - \widetilde{\nabla} F_r^0(w_r), w_{r+1} - w_r\rangle \\ &= \mathbb{E}F(w_r) + \mathbb{E}\Bigg[\langle \nabla F^0(w_r), w_{r+1} - w_r\rangle + \frac{\beta}{2}\|w_{r+1} - w_r\|^2 + f^1(w_{r+1}) - f^1(w_r) \qquad (62) \\ &\quad + \langle b_r + N_r, w_{r+1} - w_r\rangle\Bigg] \\ &\quad - \mathbb{E}\langle b_r + N_r, w_{r+1} - w_r\rangle \\ &\leqslant \mathbb{E}F(w_r) + \mathbb{E}\Big[\langle \nabla F^0(w_r), w_{r+1} - w_r\rangle + \beta\|w_{r+1} - w_r\|^2 + f^1(w_{r+1}) - f^1(w_r) \qquad (63) \\ &\quad + \langle b_r + N_r, w_{r+1} - w_r\rangle\Big] + \frac{B^2 + \Sigma^2}{\beta}, \qquad (64)\end{aligned}$$

where we used Young's inequality to bound

$$-\mathbb{E}\langle b_r + N_r, w_{r+1} - w_r\rangle \leqslant \frac{B^2 + \Sigma^2}{\beta} + \frac{\beta}{2}\|w_{r+1} - w_r\|^2. \qquad (65)$$

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

$$H_r(y_*^{\text{priv}}) - H_r(y_*) \leqslant \frac{\|N_r + b_r\|^2}{2\beta}. \qquad (66)$$

To prove (66), we will need the following lemma:

Lemma 37 *[43, Lemma B.2] Let $H(y), h(y)$ be convex functions on some convex closed set $\mathcal{Y} \subseteq \mathbb{R}^d$ and suppose that H is 2β-strongly convex. Assume further that h is L_h-Lipschitz. Define $y_1 = \arg\min_{y \in \mathcal{Y}} H(y)$ and $y_2 = \arg\min_{y \in \mathcal{Y}}[H(y) + h(y)]$. Then $\|y_1 - y_2\|_2 \leqslant \frac{L_h}{2\beta}$.*

We apply Lemma 37 with $H(y) := H_r(y),\ h(y) := \langle N_r + b_r, y\rangle$, $L_h = \|N_r + b_r\|$, $y_1 = y_*$, and $y_2 = y_*^{\text{priv}}$ to get

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

On the other hand,

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

Combining these two inequalities yields

$$\begin{aligned} H_r(y_*^{\text{priv}}) - H_r(y_*) &\leqslant \langle N_r + b_r, y_* - y_*^{\text{priv}} \rangle \\ &\leqslant \|N_r + b_r\| \|y_* - y_*^{\text{priv}}\| \\ &\leqslant \frac{\|N_r + b_r\|^2}{2\beta}, \end{aligned} \tag{67}$$

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

$$\mathbb{E}\left[\langle \nabla F^0(w_r), w_{r+1} - w_r \rangle + \beta \|w_{r+1} - w_r\|^2 + f^1(w_{r+1}) - f^1(w_r) + \langle b_r + N_r, w_{r+1} - w_r \rangle\right] = \mathbb{E}\left[\min_{y \in \mathbb{R}^d} H_r^{\text{priv}}(y)\right] \tag{68}$$

satisfies

$$\mathbb{E}\left[\min_{y \in \mathbb{R}^d} H_r^{\text{priv}}(y)\right] \leqslant \mathbb{E}\left[\min_y \left\{\langle \nabla F^0(w_r), y - w_r \rangle + \beta \|y - w_r\|^2 + f^1(y) - f^1(w_r)\right\}\right] + \frac{\Sigma^2 + B^2}{\beta} \tag{69}$$

$$\leqslant -\frac{\mu}{2\beta}\mathbb{E}\left[F(w_r) - F^*\right] + \frac{\Sigma^2 + B^2}{\beta}, \tag{70}$$

where we used the assumptions that F is μ-PPL and F^0 is 2β-smooth in the last inequality. Plugging the above bounds back into (64), we obtain

$$\mathbb{E}F(w_{r+1}) \leqslant \mathbb{E}F(w_r) - \frac{\mu}{2\beta}[F(w_r) - F^*] + \frac{2(\Sigma^2 + B^2)}{\beta}, \tag{71}$$

whence

$$\mathbb{E}[F(w_{r+1}) - F^*] \leqslant \mathbb{E}[F(w_r) - F^*](1 - \frac{\mu}{2\beta}) + \frac{2(\Sigma^2 + B^2)}{\beta}. \tag{72}$$

Using (72) recursively and summing the geometric series, we get

$$\mathbb{E}[F(w_T) - F^*] \leqslant \Delta\left(1 - \frac{\mu}{2\beta}\right)^T + \frac{4(\Sigma^2 + B^2)}{\mu}. \tag{73}$$

■

With Proposition 36 in hand, Theorem 35 then follows from substituting the bias and variance bounds of our `MeanOracle` (given in Lemma 8) into Proposition 36, and optimizing for clip threshold and batch size n/T. We provide the detailed proof below:

Theorem 38 (Re-statement of Theorem 35) *Let $\epsilon > 0$ and assume $F(w) = F^0(w) + f^1(w)$ is μ-PPL for β-smooth F^0, with $\kappa = \frac{\beta}{\mu} \leqslant n/\ln(n)$. Then, there are parameters such that Algorithm 6 is $\frac{\epsilon^2}{2}$-zCDP, and*

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

Proof We choose $\sigma^2 = \frac{4C^2T^2}{\epsilon^2 n^2}$.
Privacy: By parallel composition (since each sample is used only once) and the post-processing property of DP (since the iterates are deterministic functions of the output of `MeanOracle`), it suffices to show that $\widetilde{\nabla} F_t(w_t)$ is $\frac{\epsilon^2}{2}$-zCDP for all $t \geqslant 0$. By our choice of σ^2 and Proposition 6, $\widetilde{\nabla} F_t(w_t)$ is $\frac{\epsilon^2}{2}$-zCDP, since it's sensitivity is bounded by $\sup_{X \sim X', w} \frac{T}{n} \left\| \sum_{x \in \mathcal{B}_t} \Pi_C[\nabla f^0(w,x)] - \sum_{x' \in \mathcal{B'}_t} \Pi_C[\nabla f^0(w,x')] \right\| \leqslant \frac{T}{n} \sup_{x,x',w} \|\Pi_C[\nabla f^0(w,x)] - \Pi_C[\nabla f^0(w,x')]\| \leqslant \frac{2CT}{n}$.
Excess risk: For any iteration $t \in [T]$, denote the bias of Algorithm 1 by $b_t := \mathbb{E}\widetilde{\nabla} F_t(w_t) - \nabla F(w_t)$, where $\widetilde{\nabla} F_t(w_t) = \tilde{\nu}$ in the notation of Algorithm 1. Also let $\widehat{\nabla} F_t(w_t) := \hat{\nu}$ (in the notation of Lemma 8) and denote the noise by $N_t = \widetilde{\nabla} F_t(w_t) - \nabla F(w_t) - b_t = \widetilde{\nabla} F_t(w_t) - \mathbb{E}\widetilde{\nabla} F_t(w_t)$. Then we have $B := \sup_{t \in [T]} \|b_t\| \leqslant \frac{r^k}{(k-1)C^{k-1}}$ and $\Sigma^2 := \sup_{t \in [T]} \mathbb{E}[\|N_t\|^2] \leqslant d\sigma^2 + \frac{r_k^2 T}{n} \leqslant \frac{4dC^2T^2}{\epsilon^2 n^2} + \frac{r_k^2 T}{n}$, by Lemma 8. Plugging these bounds on B^2 and Σ^2 into Proposition 36, and choosing $T = 2\left\lceil \kappa \ln\left(\frac{\Delta\mu}{B^2+\Sigma^2}\right)\right\rceil \lesssim \kappa \ln(n)$ where $\Delta \geqslant F(w_0) - F^*$, we have:

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

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

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

as claimed.
■