On the Gradient Complexity of Private Optimization with Private Oracles

Michael Menart MICHAEL.MENART@UTORONTO.CA
Department of Computer Science, University of Toronto, Vector Institute

Aleksandar Nikolov ANIKOLOV@CS.TORONTO.EDU
Department of Computer Science, University of Toronto

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study the running time, in terms of first order oracle queries, of differentially private empirical/population risk minimization of Lipschitz convex losses. We primarily consider the setting where the loss is non-smooth and the optimizer interacts with a private proxy oracle, which sends only private messages about a minibatch of gradients. In this setting, we show that expected running time $\Omega(\min\{\frac{\sqrt{d}}{\alpha^2}, \frac{d}{\log(1/\alpha)}\})$ is necessary to achieve α excess risk on problems of dimension d when $d \geq 1/\alpha^2$. Upper bounds via DP-SGD show these results are tight when $d > \tilde{\Omega}(1/\alpha^4)$. In fact, the lower bound nearly matches the best known upper bound for general private optimizers in this regime. A consequence of our results is that, in high dimensions, the ubiquitous DP-SGD algorithm necessarily suffers a dimension dependent runtime slowdown and further that DP-SGD is optimal among the subclass of DP optimizers that use private oracles. We further show our lower bound can be strengthened to $\Omega(\min\{\frac{d}{\bar{m}\alpha^2}, \frac{d}{\log(1/\alpha)}\})$ for algorithms which use minibatches of size at most $\bar{m} < \sqrt{d}$. We next consider smooth losses, where we relax the private oracle assumption and give lower bounds under only the condition that the optimizer is private. Here, we lower bound the expected number of first order oracle calls by $\tilde{\Omega}\big(\frac{\sqrt{d}}{\alpha} + \min\{\frac{1}{\alpha^2}, n\}\big)$, where n is the size of the dataset. Modifications to existing algorithms show this bound is nearly tight. To our knowledge, ours are the first oracle complexity lower bounds to leverage differential privacy beyond the local privacy model. Compared to non-private lower bounds, our results show that differentially private optimizers pay a dimension dependent runtime penalty. Finally, as a natural extension of our proof technique, we show lower bounds in the non-smooth setting for optimizers interacting with information limited oracles. If the proxy oracle transmits at most Γ-bits of information about the gradients in the minibatch, then $\Omega\big(\min\{\frac{d}{\alpha^2\Gamma}, \frac{d}{\log(1/\alpha)}\}\big)$ oracle calls are needed. This result shows fundamental limitations of gradient quantization techniques in optimization.

Keywords: Optimization, Differential Privacy, Oracle Complexity

1. Introduction

Many fundamental problems in machine learning and statistics take the form of convex optimization problems. Many such problems can be formulated as empirical risk minimization (ERM), or stochastic convex optimization (SCO)[1]. For a dataset of n convex losses $\ell_1, \ldots, \ell_n$ and convex constraint set $\mathcal{W} \subseteq \mathbb{R}^d$ of diameter at most B, the former asks for an approximate minimizer of the empirical loss: $\min_{w\in\mathcal{W}}\{\mathcal{L}(w) := \frac{1}{n}\sum_{i=1}^n \ell_i(w)\}$. For an unknown distribution $\mathcal{D}$, the latter problem is solved by finding an approximate minimizer of population loss $\mathcal{L}_{\mathcal{D}}(w) := \mathbb{E}_{\ell\sim\mathcal{D}}[\ell(w)]$ given

1. Not to be confused with the alternative use of stochastic optimization, where one assumes access to a noisy gradient oracle. While related, these settings are distinct. We will refer to this other setting as the stochastic oracle setting.

independent samples from $\mathcal{D}$. Such problems have been studied for decades under a variety of regularity conditions on the loss, most commonly L-Lipschitzness and/or β-smoothness (i.e. Lipschitz continuous gradients). The runtime efficiency of such algorithms is often measured in the number of first-order oracle calls (i.e. gradient and loss evaluations) made during the optimization procedure. Such characterizations date as far back as the work of Nemirovski and Yudin (Nemirovsky and Yudin, 1985), which showed that the oracle complexity is $\Theta(\min\{d\log(1/\alpha), 1/\alpha^2\})$ for minimizing a single (i.e. $n = 1$) non-smooth function [2].

The power of this framework combined with modern privacy concerns has resulted in a rich literature studying differentially private (DP) analogs of these problems (Chaudhuri et al., 2011; Bassily et al., 2014, 2019; Feldman et al., 2020; Kulkarni et al., 2021; Asi et al., 2021; Choquette-Choo et al., 2025). For over a decade, it has been known that the best achievable asymptotic rate of the excess empirical risk for ERM under (ϵ, δ)-DP (DP-ERM) is $\alpha^*_{\epsilon,\delta} := BL\sqrt{d\log(1/\delta)}/(n\epsilon)$, which was first achieved in $O(n^2)$ running time using DP-SGD, (Bassily et al., 2014). The optimal rate for SCO under (ϵ, δ)-DP (DP-SCO), which is $\Theta(\alpha^*_{\epsilon,\delta} + \frac{BL}{\sqrt{n}})$, was established subsequently, albeit with even higher runtime overhead (Bassily et al., 2019). Various works have gradually improved the runtime of DP-SCO/DP-ERM since. Figure 1 summarizes the current best known upper bounds.

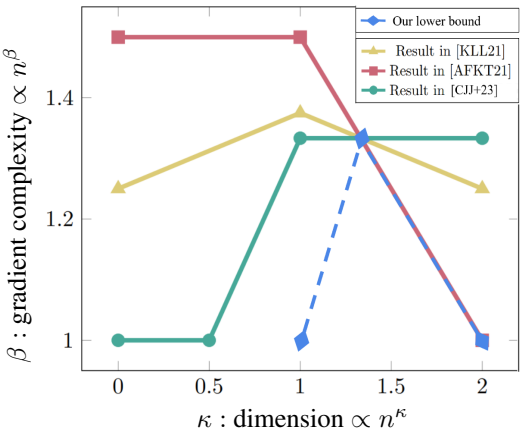

Figure 1: Summary of best known oracle complexity upper bounds and our lower bound for achieving optimal DP-SCO excess risk rate $O(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n})$, ignoring log factors and dependence on privacy parameters. The best upper bounds for private oracle methods is given by the minimum of Asi et al. (2021) and Kulkarni et al. (2021). One work, Carmon et al. (2023), improves this rate in low dimensions with an algorithm that does not satisfy oracle privacy.

Despite a large body of work on improving runtime upper bounds for DP optimization, and the importance of characterizing DP runtime repeatedly cited as an important open problem (Bassily et al., 2020; Kulkarni et al., 2021; Carmon et al., 2023), oracle complexity lower bounds that leverage DP are virtually non-existent. Our results complement past work, which has proven fundamental *utility* costs to ensuring privacy, by finally showing that a class of DP optimizers incurs significant running time cost as well. To our knowledge, ours are the first oracle complexity lower bounds to leverage differential privacy beyond the local privacy model.

Proxy oracles. Our main result is a lower bound on the oracle complexity of non-smooth ERM/SCO with access to a private "proxy" oracle. In our proxy oracle model, at round t, an optimizer sends M_t gradient queries (M_t may be adaptively chosen) to the proxy oracle. The proxy oracle then computes a minibatch of M_t gradients using the true gradient oracle, and sends a response to the optimizer. By way of example, the commonly studied stochastic oracle can be considered a type of proxy oracle which responds with noisy estimates of the true gradients. Ideologically, however,

2. The lower bounds in Nemirovsky and Yudin (1985) are loose by log factors for randomized algorithms. This gap has since been closed (Braun et al., 2017).

stochastic oracles have generally been studied from an adversarial perspective (i.e. worst case noise) whereas we are interested in *cooperative* proxy oracles, i.e. best case proxies satisfying some constraint. Private proxy oracles are those whose messages are a ρ-zCDP (zero concentrated differential privacy) processing of the true oracle responses. We remark that even when providing approximate DP guarantees of the overall process, DP optimizers generally use zCDP oracles due to favorable "moment's accounting" composition guarantees (Abadi et al., 2016; Bun et al., 2018a). A common case for private oracle methods is for the proxy oracle response to be an estimate of the gradient, often perturbed by Gaussian noise, as in the canonical DP-SGD algorithm. However, our framework allows arbitrary messages. To make the most direct comparison to traditional oracle lower bounds, we still measure the overall oracle complexity of the algorithm in the number of calls to the *true* gradient oracle (i.e. the sum of the minibatch sizes). The formal definition of the proxy oracle model is given in Section 2.

As common examples, methods which apply zCDP mechanisms to disjoint minibatches, thus obtaining privacy via parallel composition, and also methods which apply zCDP mechanisms to randomly sampled minibatches, obtaining privacy via advanced composition and amplification via subsampling, both fall into the private oracle framework. However, the private oracle model places no restrictions on data reuse or how minibatches are sampled. Because of this, it is possible to construct private oracle methods which do not lead to a DP optimizer. It is perhaps surprising that our lower bound still holds for such methods, and arises merely from leveraging the fact that information about the loss passes through a differentially private channel.

Our oracle model is motivated by several factors. Most obviously, it provides insight into design limitations of private optimizers more generally. Notably, when $d \geq (\alpha^*_{\epsilon,\delta})^{-4}$, the fastest known method for DP-SCO is a private oracle method (Asi et al., 2021). Techniques aside, a myriad of important scenarios naturally lend themselves to private oracle settings. One example is when the optimization procedure is being performed by an untrusted server communicating with nodes holding data, potentially from multiple individuals, as occurs in federated learning (Lowy and Razaviyayn, 2023). A common scenario is for the server to make gradient queries to the nodes, and thus ensuring privacy in this setting means the nodes send only privacy preserving messages about the gradients back to server. Another example is when intermediate models obtained during optimization need to themselves be used or released. Private oracle methods provide a versatile way to guarantee privacy of the entire collection of models generated during training. As such, the study of private oracle methods not only serves to provide insight into private optimization more generally, but is a meaningful algorithm class in its own right. Further, while there have been theoretical algorithmic advances in DP optimization in more structured settings, such as smooth or low dimensional problems (Feldman et al., 2020; Carmon et al., 2023), private oracle methods remain the dominant method in practice, for both convex and non-convex settings (Yu et al., 2021; Ponomareva et al., 2023; Cummings et al., 2024). Finally, by studying private oracle methods, we are, as our lower bound shows, able to characterize the effect of *batch sizes* on private training dynamics. Our results formally show the negative impact of small batch sizes on private learning, a phenomenon which has been explored in previous work, but without a minimax characterization (Räisä et al., 2024; Ponomareva et al., 2023). This contrasts sharply with the non-private setting, where lower bounds show that $\omega(1)$ batch sizes worsen runtime, at least for $d = \tilde{\Omega}(1/\alpha^4)$ (Bubeck et al., 2019).

1.1. Results

Non-smooth loss, private oracle. In the non-smooth setting when $d > 1/\alpha^2$, we show any optimizer interacting with a private oracle has expected running time $\Omega\big(\min\{\frac{1}{\alpha^2}(\sqrt{\frac{d}{\rho}}+\frac{d}{\bar{m}\rho}), \frac{d}{\log(1/\alpha)}\}\big)$, where ρ is the zCDP privacy parameter of the proxy oracle and $\bar{m}$ is some (possibly infinite) upper bound on the batch sizes. We further show this lower bound is tight for private oracle methods when $d \geq 1/\alpha^4$ by providing a more general analysis of the DP-SGD algorithm. Even in the regime $d \in [1/\alpha^2, 1/\alpha^4]$, where our lower bound is $\tilde{\Omega}(d)$, it is still stronger than the non-private lower bound of $\Omega(1/\alpha^2)$. Of particular note is the case $\rho = 1$, $\alpha = \alpha^*_{\epsilon,\delta}$, and $\bar{m} = \infty$, in which our lower bound is $\Omega(\frac{n^2\epsilon^2}{\sqrt{d}\log(1/\delta)})$, nearly matching the best known upper bound for general (ϵ, δ)-DP optimizers in the regime $d \geq (\alpha^*_{\epsilon,\delta})^{-4}$. The $\tilde{O}(d)$ limit on our lower bound can be removed for algorithms that make at most d *unique* queries to the private oracle. As such, one particularly relevant consequence of our result pertains to the ubiquitous DP-SGD algorithm. Given a minibatch estimate of the gradient, g_t, at parameters w_t, DP-SGD updates parameters via $w_{t+1} = \Pi_{\mathcal{W}}[w_t - \eta(g_t + \mathcal{N}(0, \mathbb{I}_d\sigma^2))]$.

Corollary 1 *(Informal corollary of Theorem 5) Let $\mathcal{A}$ be an α-accurate (for non-smooth losses) implementation of DP-SGD with batch size m. Then its running time is* $\Omega\big(\min\big\{\frac{\sqrt{d}+d/m}{\alpha^2}, \frac{dm}{\log(1/\alpha)}\big\}\big)$.

To our knowledge, outside of the local privacy model, ours are the first results which formally show that a class of private optimizers suffer worse runtime compared to their non-private counterparts. Even for the DP-SGD algorithm specifically, which is the backbone of DP optimization in practice and the subject of intense study, we are unaware of prior work formally showing runtime costs due to privacy.

Non-smooth loss, information limited oracle. Our proof technique uses information theoretic tools, and thus naturally extends to proxy oracles which transmit at most Γ bits of information to the optimizer. For such oracles, we show that $\Omega\big(\min\{\frac{d}{\alpha^2\Gamma}, \frac{d}{\log(1/\alpha)}\}\big)$ expected running time is necessary. Compared to classic lower bound constructions in non-smooth optimization, which require the optimizer to discover $1/\alpha^2$ vectors/gradients embedded in the loss function, our lower bound shows that, ultimately, the optimizer must indeed use "the entirety" of each of the gradients. This establishes fundamental limits for gradient quantization strategies in machine learning, which have received substantial interest due to the prominence of distributed gradient methods (Alistarh et al., 2017; Stich et al., 2018; Mayekar and Tyagi, 2020b; Faghri et al., 2020; Wang et al., 2023).

Smooth loss, private optimizer. In the smooth case, for (ϵ, δ)-DP ERM algorithms (with access to a true oracle), we show a lower bound on expected running time of $\Omega\big(\frac{\sqrt{d}}{\alpha\sqrt{\log(1/\delta)}} + \min\{\frac{1}{\alpha^2}, n\}\big)$, and provide upper bounds which show this is nearly tight. We note previous work has provided tight upper bounds for $\alpha = \alpha^*_{\epsilon,\delta}$ (Feldman et al., 2020), but by generalizing these methods, we show the lower bound is tight up to log factors for all α. Recent upper bounds show our lower bound cannot be improved via the private oracle model (Choquette-Choo et al., 2025), at least for optimal error.

Between DP-ERM and DP-SCO. In Appendix F.1, we give a reduction showing that for any α, DP-SCO is no harder than DP-ERM, up to log factors. This result follows fairly directly from combining existing results. The reverse reduction was shown in (Bassily et al., 2019). For this reason, we focus on DP-ERM in this work.

1.2. Techniques

Techniques for lower bounding oracle complexity generally live in one of two disjoint classes. The first class is the "vector discovery" framework, were the loss function is randomly instantiated via a collection of $K > 0$ random vectors from $\mathbb{R}^d$. One then tries to argue that, in order to minimize the loss function, the optimizer must observe each vector by having it returned from the oracle. As an example, a staple of such techniques is the so-called Nemirovski function, defined as $\mathcal{N}(w) = \max_{j\in[K]} \{\langle w, X_j\rangle - j\gamma\}$, for random "problem vectors" $X_1, ..., X_K \in \mathbb{R}^d$ and offset $\gamma \geq 0$. The second class of lower bounds use information theoretic methods. Here, the loss is randomly sampled from a distribution with high entropy. One then argues that loss minimization requires obtaining high mutual information with the loss (i.e. identifying the loss instance).

Our technique for non-smooth losses uses elements from both these approaches. We use a Nemirovski-like loss function of the form, $\mathcal{L}(w) = \max\left\{ \max_{k\in[K]} \{|\langle w, X_k\rangle - \alpha|\}, \|\Pi_V w\|\right\}$, where $X_1, ..., X_K$ are problem vectors in $\mathbb{R}^d$ and Π_V is the orthogonal projection onto some subspace V, which is orthogonal to $\mathsf{Span}(X_1, \ldots, X_K)$. When $n > 1$, our construction simply copies this function n times. In our analysis we show that the optimizer must discover each problem vector. But in contrast to previous analysis in the vector discovery framework, "discovery" requires obtaining high mutual information with each X_k. This is in part possible due to the presence of the regularization term, $\|\Pi_V w\|$, which will become more apparent in our proof. A crucial step in our proof is showing that the optimizer is unable to adequately learn V unless it makes $\Omega(d)$ queries.

On the other hand, we also diverge from existing information theoretic techniques in how we track mutual information. Previous techniques essentially upper bound $I(W; \mathcal{L})$, where W is the candidate solution. Clearly, we cannot hope to show more than $I(W; \mathcal{L}) \geq d\log(d/\alpha)$ is necessary, since $w^* = \min_{w\in\mathcal{W}} \{\mathcal{L}(w)\}$ can be communicated in $\tilde{O}(d)$ bits. Our solution to this bottleneck is to instead track the sum $\sum_{k=1}^{K} I(X_k; W|X_{\neq k})$. Under the correct distribution, estimating X_k is still a difficult/high-dimensional problem, even under knowledge of $X_{\neq k}$, allowing us to show that this sum must be $\Omega(dK)$. Crucial to this analysis is bounding the information leaked about X_k when the oracle *does not* return X_k, as each oracle response depends on all of $\mathcal{L}$. Previous information theoretic oracle lower bounds often circumvent this issue by considering linear/simple losses, where the query point is largely irrelevant to the gradient. In our case, because we are tracking the conditional information $I(X_k; W|X_{\neq k})$, when $\mathcal{O}(w)$ does not return X_k, we can show the oracle response is just the answer to a $\log(K)$ bit question about X_k, and thus leaks little information. This information is small enough that we can then argue the only efficient way to obtain information about X_k is by making oracle queries that return X_k.

For proxy oracles with bounded information capacity, even oracle responses which return X_k may not reveal sufficient information to estimate it. We can then use the above ideas to lower bound the runtime by lower bounding the number of times the optimizer must get the oracle to return X_k. In the private case, controlling the amount of information leaked is more technical. While it is true that ρ-zCDP mechanisms run on a dataset of (random) size M leak at most $\rho\mathbb{E}[M^2]$ bits, such a bound is too weak to achieve our lower bounds, and a more careful accounting of the information gain must be used.

1.3. Related work

To our knowledge, all existing work which leverages privacy for oracle complexity lower bounds considers the *local* model of privacy, of which the most relevant is Acharya et al. (2021). In the

language of our framework, they study the case where the private oracle always has batch size 1, satisfies ϵ-pure differential privacy, and itself only has access to a stochastic oracle, rather than a true oracle. In this setting, they show $\Omega(\frac{d}{\alpha^2\epsilon^2})$ queries are necessary. In the same setting but with an Γ-information limited oracle they show a lower bound of $\Omega(\frac{d}{\alpha^2\Gamma})$. Their assumption that the private/information limited oracle only has access to a stochastic oracle is significant, and without it their lower bound would lose its polynomial dependence on α, as the loss they consider is linear, and so only $\tilde{O}(d)$ bits need to be transmitted before the optimizer can obtain the solution.

Outside of privacy, there is a substantial literature on oracle complexity bounds. For finite sums (i.e. $n > 1$), and sufficiently large dimension, Woodworth and Srebro (2016) proved a complexity lower bound in the non-smooth case of $\tilde{\Omega}\big(\min\{\frac{1}{\alpha^2}, n + \frac{\sqrt{n}}{\alpha}\}\big)$ and, in the smooth case, $\tilde{\Omega}(\min\{\frac{1}{\alpha^2}, n + \sqrt{\frac{n}{\alpha}}\})$, which are nearly tight (Johnson and Zhang, 2013). Their loss construction is very distinct from ours. Closer constructions to ours can be found in Bubeck et al. (2019) and Marsden et al. (2022), which study the oracle complexity of highly parallelized and memory limited optimization respectively. Their loss constructions resembles ours in the sense that they use a Nermivoski-like function combined with a "regularizer", but in both cases the form and analysis of their regularizer differs substantially from ours. Examples of works using information theoretic techniques include (Agarwal et al., 2012; Braun et al., 2017; Gopi et al., 2022). However, the source or hardness in all these methods come from the difficulty of estimating (at most) $\tilde{O}(d)$ bits, making them fundamentally distinct from our method. We note also Braun et al. (2017) considered algorithms with randomized running time, as we do.

The technique for our lower bound in the smooth case is more related to work on lower bounds for DP mean estimation (Bun et al., 2018b; Dwork et al., 2015). A reduction between (smooth) optimization and mean estimation was shown in Bassily et al. (2014).

Finally, outside of (Acharya et al., 2021), other works have studied the complexity implications of communication limitations in optimization (Mayekar and Tyagi, 2020a,b; Huang et al., 2022; Salgia et al., 2025), but these works still assume the proxy oracle only has access to a stochastic oracle and quantizes only a single gradient. The works (Arjevani and Shamir, 2015; Woodworth et al., 2021; Scaman et al., 2019) studied distributed optimization lower bound under different communication constraints, which are not directly comparable to information capacity limitations.

2. Preliminaries

Notation. We let $\mathcal{B}(r)$ denotes the d-dimensional Euclidean ball with radius r centered at zero. For a set of vectors S, Π_S denotes the orthogonal projection onto $\mathsf{Span}(S)$ and $\Pi_S^{\perp}$ denotes the projection onto the orthogonal complement; $\Pi_{S,S'}$ is the projection onto $\mathsf{Span}(S) \cup \mathsf{Span}(S')$. When $\mathcal{W}$ is a compact set, we use $\Pi_{\mathcal{W}}$ as the projection onto $\mathcal{W}$ itself. We let ρ_X denote the law of X and use $\rho(x)$ when disambiguation is obvious by context. For a collection of random variables, $A_1, ..., A_T$, we denote $A = [A_1, ..., A_T]$, $A_{\leq t} = [A_1, \ldots, A_t]$ and similarly for $A_{<t}$ and $A_{\neq t}$. We define $\alpha^*_{\epsilon,\delta} := BL\sqrt{d\log(1/\delta)}/(n\epsilon)$.

Information theory. For a discrete random variables X and Y, the entropy and mutual information is defined as $H(X) = \sum_x x\log(1/\rho(x))$ and $I(X;Y) = \sum_x \sum_y \rho(x,y)\log\left(\frac{\rho(x,y)}{\rho(x)\rho(y)}\right)$ respectively. We take log to be the natural logarithm, such that entropy/information is measure in nats. For arbitrary random variables, the more general definition can be used; $I(X;Y) = \sup_{\mathcal{P},\mathcal{Q}}\{I([X]_{\mathcal{P}};[Y]_{\mathcal{Q}})\}$, where the supremum is over all finite partitions of the support, and $[X]_{\mathcal{P}}$

denotes the quantization of X via the partition of its support, $\mathcal{P}$ (Cover and Thomas, 2006). We define the information capacity of a function as follows.

Definition 2 (Information Capacity, Cover and Thomas (2006)) *The information capacity, Γ, of a randomized function, $f : \mathcal{X} \mapsto \mathcal{Y}$, is $\Gamma = \max_{\rho_X} \{I(f(X); X)\}$, where the maximum is over all distributions supported on $\mathcal{X}$.*

The most basic example of a function with information capacity at most Γ is one whose range is $\{0,1\}^\Gamma$. We will also frequently use the fact that the α-packing number of $\mathcal{B}(r)$, i.e. the size of the largest set of vectors $\mathcal{V} \subseteq \mathcal{B}(r)$ such that $\forall v, v' \in \mathcal{V} : \|v - v'\| \geq \alpha$, lies in $[(r/\alpha)^d, (3r/\alpha)^d]$ (Vershynin, 2018).

Differential privacy. An algorithm $\mathcal{A}$ is (ϵ, δ)-*differentially private* if for all datasets S and S' differing in one data point and all events $\mathcal{E}$ in the range of the $\mathcal{A}$, we have, $\mathbb{P}[\mathcal{A}(S) \in \mathcal{E}] \leq e^\epsilon \mathbb{P}[\mathcal{A}(S') \in \mathcal{E}] + \delta$ (Dwork et al., 2006). An algorithm $\mathcal{A}$ is *ρ-zero concentrated differentially private (zCDP)* if for all datasets S and S' differing in one data point and all $\alpha \in (1, \infty)$, it holds that $D_\alpha(\mathcal{A}(S)||\mathcal{A}(S')) \leq \rho\alpha$, where $D_\alpha(X||Y) = \frac{1}{\alpha - 1} \int \rho_X(x)^\alpha \rho_Y(x)^{1-\alpha} dx$ denotes the α-Rényi divergence (Bun and Steinke, 2016).

First order optimization. We consider the problem of minimizing finite sum losses. For a set of $n > 0$ losses $\mathcal{L} = \{\ell_1, ..., \ell_n\}$, with some abuse of notation we let $\mathcal{L}(w) = \frac{1}{n} \sum_{\ell \in \mathcal{L}} \ell(w)$. We consider the case where each ℓ_i is L-Lipschitz and possibly also smooth, over a compact convex set $\mathcal{W} \subset \mathbb{R}^d$ of diameter at most B. Denoting $w^* = \arg\min_{w \in \mathcal{W}} \{\mathcal{L}(w)\}$, we define the suboptimality/excess empirical risk of w as $\mathcal{L}(w) - \mathcal{L}(w^*)$. In the DP-SCO problem, we assume $\{\ell_1, ..., \ell_n\} \sim \mathcal{D}^n$ for some distribution $\mathcal{D}$, and call $\mathbb{E}_{\ell \sim \mathcal{D}}[\ell(w)] - \arg\min_{w' \in \mathcal{W}} \{\mathbb{E}_{\ell \sim \mathcal{D}}[\ell(w')]\}$ the excess population risk. First order oracles are a common way to model the interaction between the optimizer and loss function. In our work, we also consider optimizers of the form of Algorithm 1 interacting with a proxy oracle.

Definition 3 (First Order Oracle) *For losses $\mathcal{L} = \{\ell_1, ..., \ell_n\}$, a first order oracle $\mathcal{O} = \mathcal{O}_\mathcal{L}$ is a function satisfying $\mathcal{O}(r, i) \in \{(\ell_i(r), g) : g \in \nabla \ell_i(r)\}$; here ∇ denotes the subgradient.*

Definition 4 (Proxy Oracle) *Given an oracle $\mathcal{O}$, a (first order) proxy oracle $\tilde{\mathcal{O}}$ is an algorithm of the form given by Algorithm 2. For some range $\mathcal{Y}$ and any side information $\perp$, it is uniquely defined by the set of possibly randomized response functions $\tilde{\mathcal{O}}_\perp : (\mathbb{R} \times \mathbb{R}^d)^* \mapsto \mathcal{Y}$.*

We emphasize that the response of the proxy oracle need not be an estimate of the gradient, or even a vector in $\mathbb{R}^d$. As examples, the oracle could return an estimate of gradient variation (see e.g. Arora et al. (2023)), a sketch of the entire gradient minibatch, or even updated model parameters. When every $\tilde{\mathcal{O}}_\perp$ has information capacity at most Γ, we call the oracle Γ-information limited. When every $\tilde{\mathcal{O}}_\perp$ is a ρ-zCDP mechanism (with respect to its dataset of gradients), we call $\tilde{\mathcal{O}}$ a *ρ-private oracle*. We also consider a relaxation of this requirement to truncated CDP in Appendix C.2.

In the protocol defined by Algorithms 1 and 2 we refer to a *batch* as the set of non-adaptive queries made at some iteration $t \in [T]$. The non-adaptivity assumption is necessary both for this framework to be meaningful and for our lower bounds to hold; otherwise, one could push the entirety of any optimization algorithm into one call of the proxy oracle. In this case, $\mathcal{A}$ only needs to be a differentially private aggregator of some dataset of T gradients, which is much weaker than even assuming $\mathcal{A}$ is a private optimizer. Consequently, it is possible to achieve faster algorithms; see remark in Appendix D.2.

Algorithm 1 *Interaction protocol for optimizer, $\mathcal{A}$, and proxy oracle, $\tilde{\mathcal{O}}$.*

Require: Lipschitz parameter L, Constraint set $\mathcal{W}$ of diameter at most B

Set $t = 1$
while $\mathcal{A}$ chooses to continue **do**
  For $M_t \geq 0$, choose M_t queries, $(R_{t,1}, I_{t,1}), ..., (R_{t,M_t}, I_{t,M_t}) \in \mathbb{R}^d \times [n]$ to send to $\tilde{\mathcal{O}}$
  Receive Y_t from $\tilde{\mathcal{O}}$
  $t = t + 1$
end while
$T = t - 1$, and let $\bar{T}$ be the number of unique vectors in $\{R_{t,l}\}_{t\in[T], l\in M_t}$ *(for analysis only)*

Output: $\mathcal{A}$ releases solution: $W \in \mathcal{W}$

Algorithm 2 *Proxy oracle, $\tilde{\mathcal{O}}$.*

Require: Batch size $m > 0$, Queries $\{(R_{t,l}, I_{t,l})\}_{l=1}^m$, Iteration $t > 0$, Side information $\perp$

Compute first order information $G_t = \{G_{t,1}, ..., G_{t,m}\}$ where $G_{t,l} = \mathcal{O}(R_{t,l}, I_{t,l})$

Output: $Y_t = \tilde{\mathcal{O}}_\perp(G_{t,1}, ..., G_{t,m})$

Runtime characterization. For any $B, L, \beta \in \mathbb{R}^+ \cup \{\infty\}$, let $\mathcal{F}_{L,\beta}$ denote the set of all L-Lipschitz β-smooth loss functions over $\mathbb{R}^d$ and $\mathcal{K}_B$ denote the collection of all convex sets inside $\mathcal{B}(B)$. Let $\mathsf{ExTime}(\mathcal{A}, \tilde{\mathcal{O}}, \mathcal{O}_\mathcal{L}, \mathcal{W}, \alpha)$ denote the expected running time of $\mathcal{A}$ (measured in the number of evaluations of the true oracle $\mathcal{O}_\mathcal{L}$) needed to achieve expected suboptimality α on $\mathcal{L}$, when run with (proxy) oracle $\tilde{\mathcal{O}}$ on constraint set $\mathcal{W}$. We then define the worst case running time, $\overline{\mathsf{Time}}(\mathcal{A}, \tilde{\mathcal{O}}, \alpha, B, L, \beta) = \sup_{\mathcal{W}\in\mathcal{K}_B} \sup_{\mathcal{L}\in\mathcal{F}_{L,\beta}^n} \sup_{\mathcal{O}_\mathcal{L}} \{\mathsf{ExTime}(\mathcal{A}, \tilde{\mathcal{O}}, \mathcal{O}_\mathcal{L}, \mathcal{W}, \alpha)\}$, where the supremum over $\mathcal{O}_\mathcal{L}$ is taken over valid true oracles for $\mathcal{L}$; note the only flexibility is in how the oracle resolves the subgradient. We will consider different classes of algorithms throughout, and so it is useful for notation to omit for now any quantifiers over $\mathcal{A}, \tilde{\mathcal{O}}$ that would specify minimax complexity. We say an algorithm is α-accurate for $(\mathcal{F}, \mathcal{K})$ if for any $\mathcal{L} \in \mathcal{F}$ and $\mathcal{W} \in \mathcal{K}$ it yields a solution with expected suboptimality at most α.

3. Non-smooth optimization with private oracles

Our main result is a lower bound on the oracle complexity of optimization via private oracles in the large scale regime (i.e. $d \geq 1/\alpha^2$). Compared to optimization with access to a true oracle, whose complexity is $\Theta(1/\alpha^2)$ in this regime, our lower bound shows that optimization via a best-case private oracle incurs a dimension dependent runtime penalty.

Theorem 5 *Let C_2 be a universal constant. Let $\mathcal{A}$ be any optimizer satisfying the form of Algorithm 1 and $\mathbb{E}[\bar{T}] \leq \frac{d}{640\log(BL/\alpha)}$. Let $\tilde{\mathcal{O}}$ be any proxy oracle such that each $\tilde{\mathcal{O}}_\perp$ is ρ-zCDP. If $d \geq \frac{C_2 B^2 L^2}{\alpha^2}$ then, $\overline{\mathsf{Time}}(\mathcal{A}, \tilde{\mathcal{O}}, \alpha, B, L, \infty) = \Omega\big(\frac{L^2 B^2 \sqrt{d}}{\alpha^2 \sqrt{\rho}}\big)$. If for some $\bar{m} > 0$, $\|M\|_\infty \leq \bar{m}$ w.p. 1, then additionally, $\overline{\mathsf{Time}}(\mathcal{A}, \tilde{\mathcal{O}}, \alpha, B, L, \infty) = \Omega\big(\frac{L^2 B^2 d}{\alpha^2 \bar{m} \rho}\big)$.*

The full proof is in Appendix B. We also provide a detailed sketch after the following discussion. The unique query assumption expands the applicability of the lower bound when considering algorithms such as DP-SGD, which query the oracle many times at a single point each iteration. Observe

that if we drop the assumed bound on $\mathbb{E}\left[\bar{T}\right]$ we obtain a lower bound of $\Omega\big(\min\{\frac{\sqrt{d}}{\alpha^2\sqrt{\rho}}, \frac{d}{\log(1/\alpha)}\}\big)$, and similarly in the case where $\bar{m}$ is bounded [3]. Also, the same lower bound holds for DP-SCO via a reduction; see Appendix F.2. While it is not clear whether the $\Omega(d/\log(1/\alpha))$ term in this bound is tight, upper bounds from Kulkarni et al. (2021) show that for some regime of α and ρ, the bound must be weaker than $\frac{\sqrt{d}}{\alpha^2\sqrt{\rho}}$ when d is roughly less than $1/\alpha^4$. Specifically, (Kulkarni et al., 2021, Theorem 4.11) provides an algorithm which, for any $\epsilon, \delta \in [0, 0.5]$, is $\alpha^*_{\epsilon,\delta}$-accurate, uses a ρ'-zCDP oracle with $\rho' = (\frac{\epsilon}{\log(1/\delta)})$, and runs in time $O\big(\frac{n^{3/2}\epsilon}{d^{1/8}\log^{1/4}(1/\delta)} + \frac{n^2\epsilon^2}{d\log(1/\delta)}\big)$. This is faster than $\frac{\sqrt{d}}{(\alpha^*_{\epsilon,\delta})^2\rho'}$ when $d < \log(1/\delta)/(\alpha^*_{\epsilon,\delta})^4$. We also know the lower bound is tight when $d \geq \frac{\log^2(1/\alpha)}{\alpha^4\rho}$ and $\bar{m} \geq 1/(\alpha^2\rho)$ due to the following.

Theorem 6 *Let $\alpha, \bar{m}, \rho > 0$. There exists an algorithm of the form given by Algorithm 1 which, using a ρ-zCDP proxy oracle, is $O(\alpha)$-accurate for $(\mathcal{F}^n_{L,\infty}, \mathcal{K}_B)$, and runs in $O\big(\frac{B^2L^2}{\alpha^2}\big(\frac{\sqrt{d}}{\sqrt{\rho}} + \frac{d}{\bar{m}\rho}\big)\big)$ gradient computations. Further, for $\epsilon, \delta \in [0, 1]$, the algorithm is (ϵ, δ)-DP when run with parameters $\alpha \geq 26\alpha^*_{\epsilon,\delta}$ and $\rho = \frac{1}{\log(1/\delta)}$.*

The algorithm in question is simply DP-SGD with a careful tuning of the hyperparameters; see Appendix C.1 for a description and proof. Further, using our reduction in Appendix F.1, this result implies essentially the same upper bound for DP-SCO for any $\alpha \geq 1/\sqrt{n}$. Note the running time does not depend on the final desired choice of ϵ. Rather, the running time only indirectly depends on ϵ in that ϵ affects the minimum achievable error. Furthermore, at $\alpha = \Theta(\alpha^*_{\epsilon,\delta})$, the running time is $O\big(\frac{n^2\epsilon^2}{\sqrt{d\log(1/\delta)}}\big)$, which we note improves upon the previous best ERM rates (implicit in Bassily et al. (2019); Asi et al. (2021)) by a $\sqrt{\epsilon}$ factor.

Remark 7 *The discrepancy between the privacy notion used for the proxy oracle, zCDP, and the notion used for the final guarantee of DP-SGD, approximate DP, stems from zCDP's inability to be amplified via subsampling and the poor group privacy properties of approximate DP. Regardless, most algorithms in the literature that fit in the proxy oracle model use a zCDP oracle, even when providing approximate DP guarantees for the overall algorithm. In part, this is because zCDP enables composition guarantees which are tighter than what one would obtain with an approximate DP oracle (Abadi et al., 2016; Bun et al., 2018a). Both our upper and lower bounds can be rephrased using the notion of truncated CDP, which is weaker than zCDP and stronger than approximate DP. We provide these details to Appendix C.2. Whether the relaxation provided by an approximate DP oracle is meaningful enough to provide stronger running time upper bounds is a possible direction for further research.*

3.1. Proof sketch of Theorem 5 (Full proof in Appendix B)

Hard problem instance. For this sketch we will assume $L = B = 1$. To prove our result, we construct a hard distribution over loss functions. Let $\mathcal{W} = \mathcal{B}(1)$. Let $K = \frac{1}{C_1\alpha^2}$ for some constant C_1, We sample V as a uniformly random $d/2$-dimensional subspace and $X = \{X_1, \ldots, X_K\}$ as a uniformly random set of orthonormal vectors sampled orthogonal to V. We then define $\ell_i(w) = \ell(w) := \max\left\{\max_{k\in[K]}\{f_k(w)\}, h(w)\right\}, \forall i \in [n]$, where $\forall k \in [K]$, $f_k(w) = |\langle w, X_k\rangle - C_1\alpha|$

3. We will omit factors of B and L in our discussions for simplicity. They can be obtained by replacing α with $\alpha/(BL)$

and $h(w) = 2\|\Pi_V w\|$. As the loss construction of interest is the same for each $i \in [n]$, from here on we will ignore the query indices, $\{I_{t,l}\}_{t\in[T],l\in[M_t]}$, in the queries made the oracle.

Overview. Roughly, we will establish Theorem 5 by proving upper and lower bounds on the sum $\sum_{j=1}^{K} I(W; X_k|X_{\neq k}, V)$. The information upper bound will leverage properties of both the private oracle, $\tilde{\mathcal{O}}$, and true oracle, $\mathcal{O}$, to control the information leaked about X_k in terms of the number of times X_k is returned by the true oracle. The information lower bound will leverage the structure of the loss function, and the difficulty of learning the regularizing subspace V. It will be helpful to consider the mutual information with respect to an appropriately $O(\alpha)$-accurate discretization of $C_1\alpha X_k$, which we will denote as $\hat{X}_k$ respectively.

Bounding Information Obtained. To describe our information upper bound we introduce the random variables $\{Q_{t,l}\}$, which track the oracle response indices. Specifically, for $t \in [T]$ and $l \in [M_t]$, $Q_{t,l} = k$ if $\mathcal{O}(R_{t,l}) = (f_k(R_{t,l}), \nabla f_k(R_{t,l}))$ and equals $K + 1$ otherwise. Further, for each $k \in [K]$, define $\mathsf{Cnt}_k : \{0, \ldots, K + 1\}^* \mapsto \mathbb{Z}$ as, $\mathsf{Cnt}_k(q) = |\{l \in [\mathsf{length}(q)] : q_l = k\}|$. In words, $\mathsf{Cnt}_k(Q_t)$ is the number of times $\mathcal{O}$ evaluates via f_k in the minibatch used at iteration t. We will also extend the random variables defined in Algorithms 1 and 2 by defining $Y_{t,l} = 0$ for $t > T$ or $l \geq M_t$, and similarly for Q,R, G, and M. Recall $\bar{T}$ is a random variable corresponding to the number of unique vectors in $\{R_{t,l}\}_{t\in[T],l\in[M_t]}$. We now have the following.

Lemma 8 *Under the assumptions of Theorem 5, for any $k \in [K]$ it holds that, $I(W; \hat{X}_k|X_{\neq k}, V) = O\big(\mathbb{E}\Big[\sum_{t=1}^{\infty} \min\{\rho \cdot \mathsf{Cnt}_k^2(Q_t), d\}\Big] + \mathbb{E}\left[\bar{T}\right] \log(K)\big)$; expectation is taken w.r.t. $\mathcal{A}$, $\tilde{\mathcal{O}}$, X, V.*

Proof sketch of Lemma 8 (Full proof in Appendix B.1). Throughout the following we condition on $V = v$ and $X_{\neq k} = x_{\neq k}$ for some v, $x_{\neq k}$ in their support. To simplify expressions, we will let P_t denote the random variable containing $Y_{<t}$ and all content sent by $\mathcal{A}$ to $\tilde{\mathcal{O}}$ up through t.

First observe that the information any Q_t contains about $\hat{X}_k$ is bounded by its entropy, which is at most $\log(K + 1)$. Thus, via applications of the chain rule, the total information obtained can be decomposed into a bound on the information obtained from the Q random variables and the information obtained from the DP mechanisms,

$$I(W; \hat{X}_k) \leq \mathbb{E}\left[\bar{T}\right] \log(K + 1) + \sum_{t=1}^{\infty} I(Y_t; \hat{X}_k|Q_{\leq t}, P_t). \tag{1}$$

What remains is to bound $\sum_{t=1}^{\infty} I(Y_t; \hat{X}_k|Q_{\leq t}, P_t)$, the information obtained from the DP mechanisms. Recall G_t is the first order information returned by $\tilde{\mathcal{O}}$ during round t. Observe that if we condition on $Q_t = q_t$ for some q_t (recalling we have already conditioned on $\hat{X}_{\neq k} = \hat{x}_{\neq k}$), the only randomness left in G_t is in X_k. Let $G_t(x_k)$ denote the induced realization of G_t when $X_k = x_k$ under such a conditioning. We now have,

$$\begin{aligned} I(Y_t; \hat{X}_k|Q_t = q_t, P_t = p_t) &\leq I(Y_t; X_k|Q_t = q_t, P_t = p_t) \\ &= \mathop{\mathbb{E}}_{x_k \leftarrow X_k}\left[\mathsf{KL}\Big(\tilde{\mathcal{O}}_{\perp}(G_t(x_k)) \,\Big|\Big|\, \tilde{\mathcal{O}}_{\perp}(G_t)\Big)\right] \\ &\leq \mathop{\mathbb{E}}_{x_k, x_k' \leftarrow X_k}\left[\mathsf{KL}\Big(\tilde{\mathcal{O}}_{\perp}(G_t(x_k)) \,\Big|\Big|\, \tilde{\mathcal{O}}_{\perp}(G_t(x_k'))\Big)\right] \\ &\leq \mathsf{Cnt}_k^2(q_t)\rho. \end{aligned} \tag{2}$$

The KL divergence is between the induced conditional distributions given $Q_t = q_t$ and $P_t = p_t$ (as well as $X_{\neq k} = x_{\neq k}$, and $V = v$). The first three steps use standard properties of mutual information and KL divergence: discretization does not increase information, $I(A;B) = \mathbb{E}_B[\mathsf{KL}(A|B \mid\mid A)]$, and convexity of KL divergence. The last inequality uses the fact that we have conditioned on $Q_t = q_t$, as well as the definition of zCDP and its group privacy properties (i.e., that ρ-zCDP implies $s^2\rho$-group-zCDP for groups of size s). We have an additional upper bound on the mutual information via entropy,

$$I(Y_t; \hat{X}_k | Q_t = q_t, P_t = p_t) \leq H(\hat{X}_k | Q_t = q_t, P_t = p_t) = O(d). \tag{3}$$

The last equality uses the fact that our chosen discretization of X_k is of size $2^{O(d)}$. Thus after consolidating Eqns. (2) and (3) and taking appropriate expectations we have $I(Y_t; \hat{X}_k | Q_{\leq t}, P_t) \leq O\big(\mathbb{E}\left[\min\{\rho \cdot \mathsf{Cnt}_k^2(Q_t), d\}\right]\big)$. Plugging this into Eqn. (1) completes the proof.

Bounding Information Needed. We now bound the information needed in the following lemma.

Lemma 9 *Let $d > \frac{C_2}{\alpha^2}$. Under the problem distribution given at the start of Section 3.1, if $\mathcal{A}$ is of the form given by Algorithm 1 with $\mathbb{E}\left[\bar{T}\right] \leq \frac{d}{160\log(1/\alpha)}$, then $\min_k\{I(W; \hat{X}_k | X_{\neq k}, V)\} = \Omega(d)$.*

Proof sketch of Lemma 16 (Full proof in Appendix B.2) Fix some $k \in [K]$. Our information lower bound starts by leveraging a variant of Fano's method, from which we obtain that any estimator $\hat{W}$ satisfies,

$$\mathbb{P}\left[\|\hat{W} - \hat{X}_k\| \geq 40\alpha\right] \geq \frac{1}{2} - 8 \cdot \frac{I(\hat{W}; \hat{X}_k | X_{\neq k}, V)) + 1}{d}. \tag{4}$$

With this in hand, we would like to construct an accurate estimator from W using a bounded amount of additional information. In this regard, first observe that for every possible instantiation of X and V, there exists a minimizer, $w^* = C_1\alpha\sum_{k=1}^K X_k$, with 0 loss. Thus the accuracy condition of the optimization problem alone guarantees that $\mathbb{E}[|\langle W, X_k\rangle - C_1\alpha|] \leq \alpha$ and $\mathbb{E}[\|\Pi_V W\|] \leq \alpha$. The problematic piece is the component of W in the "unpenalized" subspace, which is $\Pi^{\perp}_{X,V}W$. This component may be large even if W has small loss. Further, while it would be easy enough to project out the components of W in $\mathsf{Span}(X_{\neq k})$ and V because we are considering the conditional mutual information, projecting out the component in the orthogonal complement of $\mathsf{Span}(X) \cup V$ would add too much information, as it localizes X_k to a small dimensional subspace. A key step will be to show that unless $\mathcal{A}$ makes enough (i.e. $\Omega(d)$) oracle queries to learn V, it cannot leverage the unpenalized subspace.

With this in mind, we now sketch how to construct the modified estimator, $\hat{W}$. Let $Z = \{Z_{t,l}\}_{t\in[T],l\in[M_t]}$ be defined as $Z_t = \nabla h(R_{t,l})$ and let $O = \{O_{t,l}\}_{t\in[T],l\in[M_t]}$ be such that $O_{t,l}$ is the unit vector orthogonal to $Z_{t,l}$ in the plane spanned by $Z_{t,l}$ and $R_{t,l}$ (regardless of whether or not the oracle response at query $R_{t,l}$ is the gradient of h). The important point for this sketch is that each $Z_{t,l} \in V$ and each $O_{t,l}$ is in the orthogonal complement of V. We then take $\hat{W}$ roughly equal to $\Pi^{\perp}_{X,\mathcal{S}}\Pi^{\perp}_{Z}W$, where $\mathcal{S}$ is the orthogonal complement of X_k inside $\mathsf{Span}(O)$. The actual construction $\hat{W}$ uses a modified version of $\mathcal{S}$ to ensure no more than $\frac{d}{160}$ bits of information about $\hat{X}_k$ are added. Importantly, we can show that so long as the $\mathcal{A}$ does not take advantage of the unpenalized subspace, $\hat{W}$ accurately estimates $\hat{X}_k$ and thus Eqn. (4) yields an information lower bound. We finish with the the following lemma, which indeed shows $\mathcal{A}$ cannot leverage the orthogonal complement of $\mathsf{Span}(X) \cup V$ (the unpenalized subspace).

Lemma 10 *Let $\mathcal{S}$ be the orthogonal complement of X_k inside $\mathsf{Span}(O)$. Then under the conditions of Lemma 16, $\mathbb{P}[\|\Pi^{\perp}_{X,\mathcal{S}}\Pi^{\perp}_{Z}W\| \geq 2\|\Pi_{V}\Pi^{\perp}_{X,\mathcal{S}}\Pi^{\perp}_{Z}W\| + 2\alpha] \leq \frac{1}{10}$.*

The full proof can be found in Appendix B.2; we summarize it here. First, when conditioning on $X = x$, we see that the chain $V \to (O, Z, R) \to W$ is Markovian. Thus it suffices to reason about the conditional distribution of V given O, Z and R. Put another way, we can characterize what $\mathcal{A}$ would learn about V even if it received ∇h at every query. We show that at best $\mathcal{A}$ learns V is a subspace that contains $\mathsf{Span}(Z)$ and is orthogonal to $\mathsf{Span}(O)$. Thus, provided the number of queries is not too large, we can show there is a "leftover" subspace of V, $\mathsf{Span}(V) \setminus \mathsf{Span}(Z)$, which is of dimension $\Omega(d)$ and has a conditional distribution that is uniform over an $\Omega(d)$ dimensional space. Now the properties of random projection ensure that the component of $\hat{W}$ in the unpenalized subspace is not much smaller than its component in the leftover space. As a result, we can establish that $\mathbb{E}[\|\hat{W} - \hat{X}_k\|] = O(\alpha)$ and obtain the desired information lower bound from Eqn. (4).

Concluding the proof sketch. (Full details in Appendix B.3). Using the previously described information upper and lower bounds and summing over $k \in [K]$ we obtain,

$$\Omega(dK) = \sum_{k=1}^{K} I(W; X_k | X_{\neq k}, V) = O\Big(\mathbb{E}\Big[\sum_{k=1}^{K}\sum_{t=1}^{\infty} \min\{\rho \cdot \mathsf{Cnt}_k^2(Q_t), d\}\Big] + K\log(K)\mathbb{E}\left[\bar{T}\right]\Big).$$

Using $\min\{\rho \cdot \mathsf{Cnt}_k^2(Q_t), d\} \leq \mathsf{Cnt}_k(Q_t)\min\{\sqrt{\rho d}, \rho\bar{m}\}$ and $\mathbb{E}\left[\bar{T}\right] \leq \frac{d}{320\log(K)}$ and rearranging gives the desired lower bound on $\mathbb{E}\left[\sum_{t=1}^{\infty}\sum_{k=1}^{K}\mathsf{Cnt}_k(Q_t)\right]$ (i.e. the expected runtime).

4. Smooth optimization with private optimizers

We now turn our attention towards the oracle complexity of DP-ERM for smooth functions. For such functions, we are able to relax the private oracle assumption and only assume that the entire optimization procedure is differentially private.

Theorem 11 *Let $\delta \leq \frac{1}{16nd}$, $\epsilon \leq \log(1/\delta)$, and d be larger than some constant. Assume $\mathcal{A}$ is (ϵ, δ)-DP. Then, $\overline{\mathsf{Time}}(\mathcal{A}, \mathcal{O}, \alpha, B, L, \alpha/B^2) = \Omega\big(\frac{BL\sqrt{d}}{\alpha\sqrt{\log(1/\delta)}} + \min\big\{\frac{B^2L^2}{\alpha^2}, n\big\}\big)$.*

We give the proof in Appendix D.1. Like past lower bounds for *excess risk* (e.g. Bassily et al. (2014)), we leverage the difficulty of private mean estimation and the fact that optimizers can solve mean estimation problems. Concretely, even $O(1)$-accurate (ϵ, δ)-DP mean estimation requires the dataset contain $n \geq \sqrt{d}$ samples. Our dataset construction uses $n\alpha$ non-trivial datapoints and $n - n\alpha$ zero vectors. Our lower bound then stems from the fact that after T oracle queries, the expected number of non-trivial vectors observed is $T\alpha$. Then by showing $T\alpha = \tilde{\omega}(\sqrt{d})$ is necessary, we obtain the lower bound. Some care must be taken here, as it is not necessarily true that the estimator must be (ϵ, δ)-DP with respect to the observed samples; consider for example, privacy amplification via subsampling. Nonetheless, we can still show the observed samples cannot be "traceable", allowing us to provide similar guarantees. For this reason however, the lower bound does not scale with ϵ. Our upper bound, presented subsequently, shows this is necessary. The $\min\{\frac{1}{\alpha^2}, n\}$ term in the lower bound holds even for non-private algorithms and can be proved via standard techniques. A matching upper bound can be obtained by an combing modifications of existing algorithms, which we provide in Appendix D.2

5. Non-smooth optimization with information limited oracles

Our proof techniques from Section 3 can easily be adapted to provide lower bounds for information limited oracles. In this section, we consider an individual loss $\mathcal{L}$, i.e. $n = 1$, such that the objective is to approximate $\arg\min_{w \in \mathcal{W}} \{\mathcal{L}(w)\}$. We are interested in algorithms of the form of Algorithm 1 (ignoring the query indices) interacting with a proxy oracle of bounded information capacity.

Theorem 12 *Let $\tilde{\mathcal{O}}$ be a Γ-information limited proxy oracle and $\mathcal{A}$ an algorithm of the form given by Alg. 1 with $\mathbb{E}\left[\bar{T}\right] \leq \frac{d}{640 \log(BL/\alpha)}$. If $d \geq \frac{C_2 B^2 L^2}{\alpha^2}$ then $\overline{\mathsf{Time}}(\mathcal{A}, \tilde{\mathcal{O}}, \alpha, B, L, \infty) = \Omega\big(\frac{B^2 L^2 d}{\alpha^2 \Gamma}\big)$.*

Once again, we can drop the unique query limit and more simply lower bound the runtime as $\Omega\big(\min\{\frac{d}{\alpha^2\Gamma}, \frac{d}{\log(1/\alpha)}\}\big)$. Note also that it is possible to have $\Gamma = \omega(d)$, which can be reasonable when the batch sizes are $\omega(1)$. The proof of this result follows from a simplified version of the proof of Theorem 5. We defer this proof and further discussion on this result to Appendix E.

Acknowledgments

Michael Menart would like to thank Raef Bassily and Cristóbal Guzmán for the insights gained while working with them on earlier attempts at this problem. This research was supported by an NSERC Discovery Grant (RGPIN-2021-03206), and the Canada Research Chairs program (CRC-2020-00004).

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

Jayadev Acharya, Clement Canonne, Prathamesh Mayekar, and Himanshu Tyagi. Information-constrained optimization: can adaptive processing of gradients help? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 7126–7138. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/398475c83b47075e8897a083e97eb9f0-Paper.pdf.

Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Trans. Inf. Theory*, 58(5):3235–3249, 2012. doi: 10.1109/TIT.2011.2182178. URL https://doi.org/10.1109/TIT.2011.2182178.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6c340f25839e6acdc73414517203f5f0-Paper.pdf.

Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1756–1764, Cambridge, MA, USA, 2015. MIT Press.

Raman Arora, Raef Bassily, Cristóbal Guzmán, Michael Menart, and Enayat Ullah. Differentially private generalized linear models revisited. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022.

Raman Arora, Raef Bassily, Tomás González, Cristóbal A Guzmán, Michael Menart, and Enayat Ullah. Faster rates of convergence to stationary points in differentially private optimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 1060–1092. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/arora23a.html.

Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in l1 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.

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

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11279–11288, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3bd8fdb090f1f5eb66a00c84dbc5ad51-Abstract.html.

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, 2020.

Raef Bassily, Cristóbal Guzmán, and Michael Menart. Differentially private algorithms for the stochastic saddle point problem with optimal rates for the strong gap. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2482–2508. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/bassily23a.html.

Raef Bassily, Cristóbal Guzmán, and Michael Menart. Private algorithms for stochastic saddle points and variational inequalities: Beyond euclidean geometry. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 128603–128635. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e88476b0ce9d445037422fe68ca097e4-Paper-Conference.pdf.

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

Gábor Braun, Cristóbal Guzmán, and Sebastian Pokutta. Lower bounds on the oracle complexity of nonsmooth convex optimization via information theory. *IEEE Trans. Inf. Theor.*, 63(7): 4709–4724, July 2017. ISSN 0018-9448. doi: 10.1109/TIT.2017.2701343. URL https://doi.org/10.1109/TIT.2017.2701343.

Sebastien Bubeck, Qijia Jiang, Yin-Tat Lee, Yuanzhi Li, and Aaron Sidford. Complexity of highly parallel non-smooth convex optimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/3c0cd9bcd0686e8bc0a9047eae120cc5-Paper.pdf.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In Martin Hirt and Adam Smith, editors, *Theory of Cryptography*, pages 635–658, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg. ISBN 978-3-662-53641-4.

Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially Private Release and Learning of Threshold Functions . In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 634–649, Los Alamitos, CA, USA, October 2015. IEEE Computer Society. doi: 10.1109/FOCS.2015.45. URL https://doi.ieeecomputersociety.org/10.1109/FOCS.2015.45.

Mark Bun, Cynthia Dwork, Guy N. Rothblum, and Thomas Steinke. Composable and versatile privacy via truncated cdp. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, page 74–86, New York, NY, USA, 2018a. Association for Computing Machinery. ISBN 9781450355599. doi: 10.1145/3188745.3188946. URL https://doi.org/10.1145/3188745.3188946.

Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. *SIAM Journal on Computing*, 47(5):1888–1938, 2018b. doi: 10.1137/15M1033587. URL https://doi.org/10.1137/15M1033587.

Yair Carmon, Arun Jambulapati, Yujia Jin, Yin Tat Lee, Daogao Liu, Aaron Sidford, and Kevin Tian. Resqueing parallel and private stochastic convex optimization. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 2031–2058, 2023. doi: 10.1109/FOCS57990.2023.00124.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

Christopher A. Choquette-Choo, Arun Ganesh, and Abhradeep Guha Thakurta. Near-optimal rates for o(1)-smooth dp-sco with a single epoch and large batches. In Gautam Kamath and Po-Ling Loh, editors, *Proceedings of The 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pages 315–348. PMLR, 24–27 Feb 2025. URL https://proceedings.mlr.press/v272/choquette-choo25a.html.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

Rachel Cummings, Damien Desfontaines, David Evans, Roxana Geambasu, Yangsibo Huang, Matthew Jagielski, Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, Nicolas Papernot, Ryan Rogers, Milan Shen, Shuang Song, Weijie Su, Andreas Terzis, Abhradeep Thakurta, Sergei Vassilvitskii, Yu-Xiang Wang, Li Xiong, Sergey Yekhanin, Da Yu, Huanyu Zhang, and Wanrong Zhang. Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment. *Harvard Data Science Review*, 6(1), jan 16 2024. https://hdsr.mitpress.mit.edu/pub/sl9we8gh.

John C. Duchi and Martin J. Wainwright. Distance-based and continuum fano inequalities with applications to statistical estimation, 2013. URL https://arxiv.org/abs/1311.2669.

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.

Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 650–669, 2015. doi: 10.1109/FOCS.2015.46.

Fartash Faghri, Iman Tabrizian, Ilia Markov, Dan Alistarh, Daniel M. Roy, and Ali Ramezani-Kebrya. Adaptive gradient quantization for data-parallel sgd. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

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.

Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. Private convex optimization via exponential mechanism. In *Conference on Learning Theory*, pages 1948–1989. PMLR, 2022.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1225–1234. JMLR.org, 2016.

Xinmeng Huang, Yiming Chen, Wotao Yin, and Kun Yuan. Lower bounds and nearly optimal algorithms in distributed learning with communication compression. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 18955–18969. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/77f2d0c271e508278ea13e24cd8773d5-Paper-Conference.pdf.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.,

2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/ac1dd209cbcc5e5d1c6e28598e8cbbe8-Paper.pdf.

Janardhan Kulkarni, Yin Tat Lee, and Daogao Liu. Private non-smooth erm and sco in subquadratic steps. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4053–4064. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/211c1e0b83b9c69fa9c4bdede203c1e3-Paper.pdf.

Andrew Lowy and Meisam Razaviyayn. Private federated learning without a trusted server: Optimal algorithms for convex losses. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TVY6GoURrw.

Annie Marsden, Vatsal Sharan, Aaron Sidford, and Gregory Valiant. Efficient convex optimization requires superlinear memory. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2390–2430. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/marsden22a.html.

Prathamesh Mayekar and Himanshu Tyagi. Limits on gradient compression for stochastic optimization. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2658–2663, 2020a. doi: 10.1109/ISIT44484.2020.9174075.

Prathamesh Mayekar and Himanshu Tyagi. Ratq: A universal fixed-length quantizer for stochastic optimization. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1399–1409. PMLR, 26–28 Aug 2020b. URL https://proceedings.mlr.press/v108/mayekar20a.html.

A. S. Nemirovsky and D. B. Yudin. Problem complexity and method efficiency in optimization. *SIAM Review*, 27(2):264–265, 1985. doi: 10.1137/1027074. URL https://doi.org/10.1137/1027074.

Natalia Ponomareva, Sergei Vassilvitskii, Zheng Xu, Brendan McMahan, Alexey Kurakin, and Chiyaun Zhang. How to dp-fy ml: A practical tutorial to machine learning with differential privacy. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 5823–5824, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599561. URL https://doi.org/10.1145/3580305.3599561.

Ossi Räisä, Joonas Jälkö, and Antti Honkela. Subsampling is not magic: why large batch sizes work for differentially private stochastic optimisation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Sudeep Salgia, Nikola Pavlovic, Yuejie Chi, and Qing Zhao. Characterizing the accuracy-communication-privacy trade-off in distributed stochastic convex optimization. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings*

of Machine Learning Research, pages 4285–4293. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/salgia25a.html.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal convergence rates for convex distributed optimization in networks. *Journal of Machine Learning Research*, 20(159):1–31, 2019. URL http://jmlr.org/papers/v20/19-543.html.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. NIPS'18, page 4452–4463, Red Hook, NY, USA, 2018. Curran Associates Inc.

Enayat Ullah, Michael Menart, Raef Bassily, Cristóbal A Guzmán, and Raman Arora. Public-data assisted private stochastic optimization: Power and limitations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=j14wStqZni.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Zeqin Wang, Ming Wen, Yuedong Xu, Yipeng Zhou, Jessie Hui Wang, and Liang Zhang. Communication compression techniques in distributed deep learning: A survey. *Journal of Systems Architecture*, 142:102927, 2023. ISSN 1383-7621. doi: https://doi.org/10.1016/j.sysarc.2023.102927. URL https://www.sciencedirect.com/science/article/pii/S1383762123001066.

Blake Woodworth and Nathan Srebro. Tight complexity bounds for optimizing composite objectives. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3646–3654, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Blake E Woodworth, Brian Bullins, Ohad Shamir, and Nathan Srebro. The min-max complexity of distributed stochastic convex optimization with intermittent communication. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4386–4437. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/woodworth21a.html.

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Gradient perturbation is underrated for differentially private convex optimization. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021. ISBN 9780999241165.

Liang Zhang, Kiran Koshy Thekumparampil, Sewoong Oh, and Niao He. Bring your own algorithm for optimal differentially private stochastic minimax optimization. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022.

Appendix A. Supplementary Lemmas

The following lemmas and fact will be used several times throughout the appendix.

Fact 1 *Let $\mathcal{A}$ be an algorithm with expected running time T and which is α-accurate for $(\mathcal{F}_{1,\infty}, \mathcal{W}_1)$ and $\mathcal{W}_1$. Then one run of $\mathcal{A}$ can be used in a black box manner to obtain an αBL-accurate algorithm for $(\mathcal{F}_{L,\infty}, \mathcal{W}_B)$ with the same oracle complexity.*

The above is a standard fact and comes from running $\mathcal{A}$ on constraint set $\hat{\mathcal{W}} = \left\{\frac{w}{B} : w \in \mathcal{W}\right\}$ and loss $\hat{\mathcal{L}}(w) = \mathcal{L}(w/L)$, and then rescaling the output of $\mathcal{A}$ by B.

Lemma 13 *Let E, F be linear subspaces of $\mathbb{R}^d$ which are orthogonal to each other. Then $\Pi_E^\perp \Pi_F^\perp = \Pi_F^\perp \Pi_E^\perp = \Pi_{F,E}^\perp$.*

Proof Let $d_E = \mathsf{Dim}(E)$ and $d_F = \dim(F)$. Let $u_1, ..., u_{d_E}$ be an orthonormal basis for E, and $u_{d_E+1}, ..., u_{d_E+d_F}$ be an orthonormal basis for F. Let $u_{d_E+d_F+1}, ..., u_d$ be an orthonormal basis for the remaining space. For some vector $v \in \mathbb{R}^d$, let $\gamma_j = \langle v, u_j \rangle$. Now clearly

$$\Pi_E^\perp v = \Pi_E^\perp \sum_{j=1}^{d} \gamma_j u_j = \sum_{j=d_E+1}^{d} \gamma_j u_j,$$

and similarly for $\Pi_F^\perp$. It is now easy to see the projections commute. Because they commute, the product of projections is equal to the projection onto the intersection. ■

We will use the following result on privacy amplification via subsampling with replacement, which is a minor modification of (Bun et al., 2015, Lemma 4.14).

Lemma 14 *Let $\epsilon, \delta \in [0, 1]$ and let $\mathcal{M}$ be an (ϵ, δ)-DP algorithm for datasets of size $m > 0$. Then if $\epsilon \leq \min\left\{1, \frac{n}{2m}\right\}$, the algorithm $\tilde{\mathcal{M}}$, which on input dataset S of size n, first samples m points with replacement and then runs $\mathcal{M}$ on the result is (ϵ', δ')-DP with*

$$\epsilon' = 6\epsilon\frac{m}{n} \qquad \text{and} \qquad \delta' = 4e^{(6\epsilon m/n)}\frac{m}{n}\delta.$$

In contrast to the original statement, this lemma applies when $m > n$. Obviously, in this regime the result does not amplify privacy, and rather controls the impact of the likely event in which a datapoint gets copied into the sampled dataset many times. Nonetheless, this unified phrasing will be convenient. The proof is nearly identical. Bounding ϵ follows in exactly the same way, and to bound δ we leverage our additional assumption that $\epsilon \leq \min\left\{1, \frac{n}{2m}\right\}$. We have copied the proof from Bun et al. (2015), with the necessary modification, below.

Proof [Proof of Lemma 14] Let D, D' be adjacent databases of size n, and suppose without loss of generality that they differ on the last row. Let T be a random variable denoting the multiset of indices sampled by $\tilde{\mathcal{M}}$ and let $\ell(T)$ be the multiplicity of index n in T. Fix a subset S of the range of $\tilde{\mathcal{M}}$. For each $k = 0, 1, ..., m$ let

$$p_k = \mathbb{P}[\ell(T) = k] = \binom{m}{k} n^{-k}(1 - 1/n)^{m-k} = \binom{m}{k}(n-1)^{-k}(1-1/n)^m,$$
$$q_k = \mathbb{P}[\mathcal{M}(D|_T) \in S | \ell(T) = k],$$
$$q'_k = \mathbb{P}[\mathcal{M}(D'|_T) \in S | \ell(T) = k].$$

Here, $D|_T$ denotes the subsample of D consisting of the indices in T, and similarly for $D'|_T$. Note that $q_0 = q'_0$, since $D|_T = D'|_T$ if index n is not sampled. Our goal is to show that

$$\mathbb{P}[\tilde{\mathcal{M}}(D) \in S] = \sum_{k=0}^{m} p_k q_k \leq e^{\epsilon'} \sum_{k=0}^{m} p_k q'_k + \delta' = e^{\epsilon'} \mathbb{P}[\tilde{\mathcal{M}}(D') \in S] + \delta'.$$

To do this, observe that by privacy, $q_k \leq e^{\epsilon} q_{k-1} + \delta$ so

$$q_k \leq e^{k\epsilon} q_0 + \frac{e^{k\epsilon} - 1}{e^{\epsilon} - 1} \delta$$

Hence,

$$\begin{aligned}
\mathbb{P}[\tilde{\mathcal{M}}(D) \in S] &= \sum_{k=0}^{m} p_k q_k \\
&\leq \sum_{k=0}^{m} \binom{m}{k} (n-1)^{-k} (1-1/n)^m \left(e^{k\epsilon} q_0 + \frac{e^{k\epsilon} - 1}{e^{\epsilon} - 1} \delta \right) \\
&= q_0 (1-1/n)^m \sum_{k=0}^{m} \binom{m}{k} \left(\frac{e^{\epsilon}}{n-1} \right)^k + \frac{\delta}{e^{\epsilon} - 1} (1-1/n)^m \sum_{k=0}^{m} \binom{m}{k} \left(\frac{e^{\epsilon}}{n-1} \right)^k - \frac{\delta}{e^{\epsilon} - 1} \\
&= q_0 (1-1/n)^m \left(1 + \frac{e^{\epsilon}}{n-1} \right)^m + \frac{\delta}{e^{\epsilon} - 1} (1-1/n)^m \left(1 + \frac{e^{\epsilon}}{n-1} \right)^m - \frac{\delta}{e^{\epsilon} - 1} \\
&= q_0 \left(1 - \frac{1}{n} + \frac{e^{\epsilon}}{n} \right)^m + \frac{\left(1 - \frac{1}{n} + \frac{e^{\epsilon}}{n}\right)^m - 1}{e^{\epsilon} - 1} \delta. \quad (1)
\end{aligned}$$

Similarly, we also have that

$$\mathbb{P}[\tilde{\mathcal{M}}(D') \in S] \geq q_0 \left(1 - \frac{1}{n} + \frac{e^{-\epsilon}}{n} \right)^m - \frac{\left(1 - \frac{1}{n} + \frac{e^{-\epsilon}}{n}\right)^m - 1}{e^{-\epsilon} - 1} \delta,$$

Combining inequalities 1 and 2 we get that

$$\mathbb{P}[\tilde{\mathcal{M}}(D) \in S] \leq \left(\frac{1 - \frac{1}{n} + \frac{e^{\epsilon}}{n}}{1 - \frac{1}{n} + \frac{e^{-\epsilon}}{n}} \right)^m \cdot \left(\mathbb{P}[\tilde{\mathcal{M}}(D') \in S] + \frac{1 - \left(1 - \frac{1}{n} + \frac{e^{-\epsilon}}{n}\right)^m}{1 - e^{-\epsilon}} \delta \right) + \frac{\left(1 - \frac{1}{n} + \frac{e^{\epsilon}}{n}\right)^m - 1}{e^{\epsilon} - 1} \delta,$$

proving that $\tilde{\mathcal{M}}$ is (ϵ', δ')-DP for

$$\epsilon' \leq m \log \left(\frac{1 + \frac{e^{\epsilon} - 1}{n}}{1 + \frac{e^{-\epsilon} - 1}{n}} \right) \leq \frac{6\epsilon m}{n}.$$

Using $m \geq n/2$ and $\epsilon \leq \min\{1, \frac{n}{2m}\}$, we have,

$$\begin{aligned}
\delta' &\leq e^{6\epsilon\frac{m}{n}} \frac{1-\exp\left(\frac{2m}{n}(e^{-\epsilon}-1)\right)}{1-e^{-\epsilon}}\delta + \frac{\exp\left(\frac{m}{n}(e^{\epsilon}-1)\right)-1}{e^{\epsilon}-1}\delta \\
&\leq e^{6\epsilon\frac{m}{n}} \frac{1-\exp\left(-\frac{2m}{n}\epsilon\right)}{1-e^{-\epsilon}}\delta + \frac{\exp\left(\frac{2m}{n}\epsilon\right)-1}{e^{\epsilon}-1}\delta \\
&\leq e^{6\epsilon\frac{m}{n}} \left(\frac{2m}{n}\right)\delta + (\frac{2m}{n})\delta \\
&\leq e^{6\epsilon\frac{m}{n}} \left(\frac{4m}{n}\right)\delta.
\end{aligned}$$

■

Appendix B. Proof of Oracle Complexity Lower Bound (Theorem 5)

It suffices to consider the case when $L = 2$, $B = 1$. This is because, by a standard rescaling reduction, if $\mathcal{A}$ is α-accurate for $(\mathcal{F}^n_{L,\infty}, \mathcal{K}_B)$, it can be used to obtain an algorithm which is $\alpha/(BL)$-accurate for $(\mathcal{F}^n_{1,\infty}, \mathcal{K}_1)$. See Fact 1 in Appendix A.

Hard problem instance. To prove our result, we construct a hard distribution over loss functions. Let $\mathcal{W} = \mathcal{B}(1)$. Letting $C_1 = 480$, set $K := \frac{1}{C_1^2\alpha^2}$. We will sample V as a uniformly random $d/2$-dimensional subspace and $X = \{X_1, ..., X_K\}$ as a uniformly random set of orthonormal vectors sampled orthogonal to V. Note this is possible since we have assumed $d \geq C_2/\alpha^2$. We then define $f_k(w) = |\langle w, X_k\rangle - C_1\alpha|$, $h(w) = 2\|\Pi_V w\|$ and, $\ell_i(w) = \ell(w) := \max\left\{\max_{k\in[K]}\{f_k(w)\}, h(w)\right\}, \forall i \in [n]$.

As the loss construction of interest is the same for each $i \in [n]$, from here on we will ignore the query indices, $\{I_{t,l}\}_{t\in[T],l\in[M_t]}$, in the queries made the oracle. At any $w \in \mathbb{R}^d$, we have the true oracle return $(f_k(w), \nabla f_k(w))$ for the smallest valid choice of k, and $(h(w), \nabla h(w))$ if no k is valid.

Proof notation. Before proceeding with the proof, we will need additional notation. First, we extend the random variables defined in Algorithms 1 and 2 by defining $Y_t = 0$ for $t > T$ and similarly for R_t, G_t, and M_t. In the following, let $Y = \{Y_1, Y_2, ...\}$ and similarly for M,X, G and R. Let $\{Q_{t,l}\}_{t,l\in\mathbb{Z}^+}$ be the random variables defined as,

$$Q_{t,l} = \begin{cases} 0 & \text{if } t > T \text{ or } l \geq M_t \\ k & \text{else if } \mathcal{O}(R_{t,l}) = (f_k(R_{t,l}), \nabla f_k(R_{t,l})) \\ K+1 & \text{else} \end{cases}.$$

We denote $Q_t = \{Q_{t,l}\}_{l\in[M_t]}$. For each $k \in [K]$, define, $\mathsf{Cnt}_k : \{0, \ldots, K+1\}^* \mapsto \mathbb{Z}$ as,

$$\mathsf{Cnt}_k(q) = |\{l \in \mathsf{length}(q) : q_l = k\}| \qquad \text{and} \qquad \overline{\mathsf{Cnt}}_k(q) = \min\left\{\mathsf{Cnt}_k(q), \sqrt{3C_1 d/\rho}\right\}.$$

In words, $\mathsf{Cnt}_k(Q_t)$ is the number of times $\mathcal{O}$ evaluates via f_k at iteration t. Finally, let $\bar{T}$ be a random variable corresponding to the number of unique vectors in $\{R_{t,l}\}_{t\in[T],l\in[M_t]}$.

As we will be bounded the information obtained about each X_k, it will be helpful to consider the mutual information with respect to a discretization of X_k. Let $\mathcal{C}$ be an α-packing and 2α-cover of $\mathcal{B}(C_1\alpha)$. Note such a set exists, as any maximal α-packing is also an 2α-cover (otherwise, one could find another point to add to the packing, a contradiction). We will then characterize the difficulty in estimating $\hat{X}_k = \arg\min_{c\in\mathcal{C}} \{\|C_1\alpha X_k - c\|\}$.

Overview. To establish Theorem 5, we will first in Section B.1 upper bound the information obtained during the optimization process. Then, in Section B.2, we will lower bound the information needed to solve the optimization problem. Doing so involves controlling what the optimizer learns about the subspace V, which is also analyzed in this section. Finally, in Section B.3 we will conclude the proof using these information bounds.

B.1. Bounding Information Obtained.

We will now bound the information obtained about a problem vector in terms of the number of times it is observed by the proxy oracle.

Lemma 15 *Under the assumptions of Theorem 5, for any $k \in [K]$ it holds that, $I(W; \hat{X}_k | X_{\neq k}, V) \leq \mathbb{E}\Big[\sum_{t=1}^{\infty} \overline{\mathsf{Cnt}}_k^2(Q_t)\Big]\rho + \mathbb{E}\left[\bar{T}\right]\log(K+1)$, where expectation is taken with respect to $\mathcal{A}$, $\tilde{\mathcal{O}}$, X, V.*

Proof In the following condition on $V = v$ and $X_{\neq k} = x_{\neq k}$ for some v, $x_{\neq k}$ in their support until otherwise stated. Let $\hat{R}_t$ denote the content sent to $\tilde{\mathcal{O}}$ at round t by $\mathcal{A}$. Since $I(W; \hat{X}_k) \leq I(Y, \hat{R}; \hat{X}_k) \leq I(Y, \hat{R}, Q; \hat{X}_k)$, we start by decomposing the information in Y, $\hat{R}$ and Q via the chain rule,

$$
\begin{aligned}
I(W; \hat{X}_k) &= \sum_{t=1}^{\infty} I(Y_t, \hat{R}_t, Q_t; \hat{X}_k | Y_{<t}, \hat{R}_{<t}, Q_{<t}) \\
&= \sum_{t=1}^{\infty} I(\hat{R}_t; \hat{X}_k | Y_{<t}, \hat{R}_{<t}, Q_{<t}) + I(Q_t; \hat{X}_k | Y_{<t}, \hat{R}_{\leq t}, Q_{<t}) + I(Y_t; \hat{X}_k | Y_{<t}, \hat{R}_{\leq t}, Q_{\leq t}) \\
&\overset{(i)}{=} \sum_{t=1}^{\infty} I(Q_t; \hat{X}_k | Y_{<t}, \hat{R}_{\leq t}, Q_{<t}) + I(Y_t; \hat{X}_k | Y_{<t}, \hat{R}_{\leq t}, Q_{\leq t}) \\
&\leq \sum_{t=1}^{\infty} \left(\sum_{l=1}^{M_t} H(Q_{t,l}; \hat{X}_k | Y_{<t}, \hat{R}_{\leq t}, Q_{<t}, Q_{t,<l}) \right) + I(Y_t; \hat{X}_k | Y_{<t}, \hat{R}_{\leq t}, Q_{\leq t}). \\
&\overset{(ii)}{\leq} \mathbb{E}\left[\sum_{t=1}^{\infty} \sum_{l=1}^{M_t} \log(K+1) \cdot \mathbb{1}_{[R_{t,l} \notin R_{\leq t, <l}]} \right] + \sum_{t=1}^{\infty} I(Y_t; \hat{X}_k | Y_{<t}, \hat{R}_{\leq t}, Q_{\leq t}) \\
&\overset{(iii)}{\leq} \mathbb{E}\left[\bar{T}\right]\log(K+1) + \sum_{t=1}^{\infty} I(Y_t; \hat{X}_k | Y_{<t}, \hat{R}_{\leq t}, Q_{\leq t}). \qquad (5)
\end{aligned}
$$

Step (i) uses the fact that, by data processing, $\hat{R}_t$ contains no information about $\hat{X}_k$ when conditioned on $Y_{<t}$. Step (ii) uses the fact that if a query $R_{t,l}$ is the same as a past query, its (conditional) entropy is zero, and otherwise the entropy is bounded by $\log(K+1)$. The final step (iii) uses the fact that the number of unique queries is $\bar{T}$.

What remains is to bound $\sum_{t=1}^{\infty} I(Y_t; \hat{X}_k | Y_{<t}, \hat{R}_{\leq t}, Q_{\leq t})$. To ease notation, define $P_t = (Y_{<t}, \hat{R}_{\leq t})$. Fix some $t \in [T]$ and note, $I(Y_t; \hat{X}_k | Q_t, P_t) = \mathbb{E}_{q_t \leftarrow Q_t}[I(Y_t; \hat{X}_k | Q_t = q_t, P_t)]$. Recall

G_t is the first order information returned by $\tilde{\mathcal{O}}$ during round t. Let us now condition on $Q_t = q_t$ and recall we have already conditioned on $\hat{X}_{\neq k} = \hat{x}_{\neq k}$. Consequently, the only randomness left in G_t is in X_k; let $G_t(x_k)$ denote the induced realization of G_t when $X_k = x_k$. Conditioning on $Q_t = q_t$ and $P_t = p_t$ throughout, and letting $\perp$ be the induced side information at round t, we have,

$$\begin{aligned} I(Y_t; \hat{X}_k | Q_t = q_t, P_t = p_t) &\overset{(i)}{\leq} I(Y_t; X_k | Q_t = q_t, P_t = p_t) \\ &\overset{(ii)}{=} \mathop{\mathbb{E}}_{x_k \leftarrow X_k} \left[\mathsf{KL}\Big(\tilde{\mathcal{O}}_\perp(G_t(x_k)) \;\Big|\Big|\; \tilde{\mathcal{O}}_\perp(G_t) \Big) \right] \\ &\overset{(iii)}{\leq} \mathop{\mathbb{E}}_{x_k, x_k' \leftarrow X_k} \left[\mathsf{KL}\Big(\tilde{\mathcal{O}}_\perp(G_t(x_k)) \;\Big|\Big|\; \tilde{\mathcal{O}}_\perp(G_t(x_k')) \Big) \right] \\ &\overset{(iv)}{\leq} \mathsf{Cnt}_k^2(q_t)\rho. \end{aligned} \tag{6}$$

The KL divergence is between the induced conditional distributions given $Q_t = q_t$ and $P_t = p_t$ (as well as $X_{\neq k} = x_{\neq k}$, and $V = v$). Above, (i) uses the fact that the mutual information for any quantization of two random variables is upper bounded by the mutual information between the original random variables. Step (ii) uses the fact that for any random variables A, B, $I(A; B) = \mathbb{E}_B[\mathsf{KL}(A|B \,||\, A)]$. Step (iii) uses the fact that the probability distribution of $\mathcal{M}_t(G_t)$ can be written as the expectation of the conditional distribution given X_k and the convexity of KL divergence. Step (iv) uses the definition of zCDP and it's group privacy properties; ρ-zCDP implies $s^2\rho$ group zCDP for groups of size s.

We have an additional upper bound on the mutual information via entropy,

$$I(Y_t; \hat{X}_k | Q_t = q_t, P_t = p_t) \leq H(\hat{X}_k | Q_t = q_t, P_t = p_t) \leq \log(|\mathcal{C}|) \leq 3C_1 d. \tag{7}$$

The last inequality comes from upper bounds on packing numbers. Recall we have defined $\overline{\mathsf{Cnt}}_k(q) = \min\{\mathsf{Cnt}_k(q), \sqrt{3C_1 d/\rho}\}$. After consolidating Eqns. (6) and (7) we further have,

$$\forall q_t, p_t: \;\; I(Y_t; \hat{X}_k | Q_t = q_t, P_t = p_t) \leq \overline{\mathsf{Cnt}}_k^2(q_t)\rho \implies I(Y_t; \hat{X}_k | Q_t, P_t) \leq \mathop{\mathbb{E}}_{Q} \left[\overline{\mathsf{Cnt}}_k^2(Q_t)\rho \right].$$

Plugging this into Eqn. (5), $I(Y, \hat{R}; \hat{X}_k) \leq \mathbb{E}\left[\sum_{t=1}^{\infty} \overline{\mathsf{Cnt}}_k^2(Q_t)\rho\right] + \mathbb{E}\left[\bar{T}\right] \log(K+1)$. Recall we have conditioned on $V = v$ and $X_{\neq k} = x_{\neq k}$ throughout. Since the above holds for arbitrary instantiations in their support, we have the same upper bound on $I(W; \hat{X}_k | X_{\neq k}, V)$. ■

B.2. Bounding Information Needed.

We now bound the information needed in the following lemma.

Lemma 16 *Let $d \geq \frac{C_2}{\alpha^2}$. Under the problem distribution given at the start of Appendix B, if $\mathcal{A}$ is of the form given by Algorithm 1 with $\mathbb{E}\left[\bar{T}\right] \leq \frac{d}{160 \log(1/\alpha)}$, then $\min_k\{I(W; \hat{X}_k | X_{\neq k}, V) \geq d/160$.*

We note that the lemma does not assume any structure of the proxy oracle, and thus holds even for the case when the optimizer interacts with the true oracle. Before giving the proof, we will require some additional notation. Let $Z = \{Z_{t,l}\}_{t \in [T], l \in [M_t]}$ be the random variables such that

$Z_{t,l} = \nabla h(R_{t,l})$. Further let $O = \{O_{t,l}\}_{t\in[T],l\in[M_t]}$ be such that $O_{t,l}$ is the unit vector orthogonal $Z_{t,l}$ in the plane spanned by $Z_{t,l}$ and $R_{t,l}$, taken with $\langle O_{t,l}, R_{t,l}\rangle \geq 0$ to break ambiguity. If $Z_{t,l}$ and $R_{t,l}$ are colinear, define $O_{t,l}$ to be the zero vector.

To prove the information lower bound, we will also need the following key sublemma.

Lemma 17 *Let $\mathcal{S}$ be the orthogonal complement of X_k inside $\mathsf{Span}(O)$. Then under the conditions of Lemma 16, $\mathbb{P}[\|\Pi^{\perp}_{X,\mathcal{S}}\Pi^{\perp}_{Z}W\| \geq 2\|\Pi_V\Pi^{\perp}_{X,\mathcal{S}}\Pi^{\perp}_{Z}W\| + 2\alpha] \leq \frac{1}{10}$.*

Proof [Proof of Lemma 17] Throughout we will use the random variables Z and O defined above and those defined by Algorithm 1. Let $\mathcal{V}$ be the set of all possible $d/2$ dimensional subspaces of $\mathbb{R}^d$. Conditioned on a set of gradient oracle queries and responses, let $\mathcal{V}_{good} = \mathcal{V}_{good}(R, Z, X) \subset \mathcal{V}$ denote the set of subspaces which result in the same set of oracle responses from the true oracle $\mathcal{O}$, and $\mathcal{V}_{bad} = \mathcal{V} \setminus \mathcal{V}_{good}$.

Conditional distribution of V. We will first find the conditional density of v after conditioning on all randomness generated during the optimization process. We note the density is with respect to the rotation invariant measures on $d/2$-dimensional subspaces. Since O and G are completely determined given $X = x, R = r$ and $Z = z$, it suffices to find the conditional density given X, R and Z. We have,

$$\rho(v|x, r, z, w) = \rho(v|x, r, z) = \frac{\rho(r, z|v, x)\rho(v|x)}{\rho(r, z|x)}. \tag{8}$$

The first equality uses that $V \to (X, R, Z) \to W$ is a Markov chain. Further,

$$\begin{aligned}
\rho(r_{\leq T}, z_{\leq T}|v, x) &= \rho(z_T|r_{\leq T}, z_{<T}, v, x)\rho(r_T|r_{<T}z_{<T}, v, x)\rho(r_{<T}, z_{<T}|v, x)\\
&\vdots\\
&= \left(\prod_{t=1}^{T}\rho(z_t|r_{\leq t}, z_{<t}, v, x)\right)\left(\prod_{t=1}^{T}\rho(r_t|r_{<t}, z_{<t}, v, x)\right)\\
&= \left(\prod_{t=1}^{T}\rho(z_t|r_t, v)\right)\left(\prod_{t=1}^{T}\rho(r_t|r_{<t}, z_{<t}, x)\right).
\end{aligned}$$

Now plugging into Eqn. (8) and using $\rho(z_t|r_t, v) = \frac{\rho(z_t, r_t|v)}{\rho(r_t|v)}$ we have,

$$\rho(v|x, r, z, w) = \left(\rho(v|x)\prod_{t=1}^{T}\frac{\rho(z_t, r_t|v)}{\rho(r_t|v)}\right)\left(\frac{1}{\rho(r, z|x)}\prod_{t=1}^{T}\rho(r_t|r_{<t}, z_{<t}, x)\right).$$

Observe the second factor on the RHS is independent of v and is thus constant. For the first factor, since the true oracle $\mathcal{O}$ is deterministic, z_t is completely determined by r_t and v. Specifically, for any $v \in \mathcal{V}_{bad}$, $\left(\rho(v|x)\prod_{t=1}^{T}\frac{\rho(z_t, r_t|v)}{\rho(r_t|v)}\right) = 0$ because $\rho(z_t, r_t|v) = 0$ for some $t \in [T]$. Alternatively, if $v \in \mathcal{V}_{good}$, then $\frac{\rho(z_t, r_t|v)}{\rho(r_t|v)} = 1, \forall t \in [T]$, and $\left(\rho(v|x)\prod_{t=1}^{T}\frac{\rho(z_t, r_t|v)}{\rho(r_t|v)}\right)$ is constant for all $v \in \mathcal{V}_{good}$ since $\rho(v|x)$ is a uniform density on subspaces orthogonal to x. This establishes that $\rho(v|x, r, z)$ is the uniform distribution over $\mathcal{V}_{good}$.

Determining $\mathcal{V}_{good}$. Our aim is now to prove the following fact: $\mathcal{V}_{good}$ is exactly the set of $d/2$ dimensional linear subspaces which contain $\mathsf{Span}(z)$ and are orthogonal to $\mathsf{Span}(o)$ and $\mathsf{Span}(x)$. Let $v \in \mathsf{Supp}(V)$ and consider some individual query $\tilde{r} \in \mathbb{R}^d$ and denote $\tilde{z} = \Pi_v \tilde{r}$ and let $\tilde{o} \in \mathsf{Span}(\tilde{r}, \tilde{z})$ be the unit vector orthogonal to $\tilde{z}$ with $\langle \tilde{o}, \tilde{r} \rangle > 0$ or the zero vector if $\tilde{r} = \tilde{z}$. To prove the fact, it suffices to show that $\tilde{z} = \Pi_v \tilde{r}$ if and only if v is a subspace which is orthogonal to $\tilde{o}$ and contains $\tilde{z}$.

The claim is straightforward if $\tilde{r} = \tilde{z}$, and so we focus on the case where $\tilde{r} \neq \tilde{z}$. We first show that any other subspace, v', which is orthogonal to $\tilde{o}$ and contains $\tilde{z}$ still satisfies $\Pi_{v'} \tilde{r} = \tilde{z}$. Towards this end, we have the following:

$$\min_{u:\langle u, \tilde{o} \rangle = 0} \{\|u - \tilde{r}\|\} \overset{(i)}{=} \|\tilde{z} - \tilde{r}\| \overset{(ii)}{\geq} \min_{u \in \mathsf{Span}(v')} \{\|u - \tilde{r}\|\} \overset{(iii)}{\geq} \min_{u:\langle u, \tilde{o} \rangle = 0} \{\|u - \tilde{r}\|\}.$$

Equality (i) is because the projection of $\tilde{r}$ to the space orthogonal $\tilde{o}$ is obtained by removing the component along $\tilde{o}$, which results in some vector in the 1-dimensional space $\mathsf{Span}(\tilde{z})$. By the definition of projection, $\tilde{z}$ is closest point to $\tilde{r}$ in this one dimensional space. Inequality (ii) follows from the assumption that $\tilde{z} \in \mathsf{Span}(v')$, and (iii) follows from the fact that $\mathsf{Span}(v') \subseteq \{u \in \mathbb{R}^d : \langle u, \tilde{o} \rangle = 0\}$. Since the LHS and RHS above are equal, $\|\tilde{z} - \tilde{r}\| = \min_{u \in \mathsf{Span}(v')} \{\|u - \tilde{r}\|\}$ and thus $\tilde{z} = \Pi_{v'} \tilde{r}$ by the uniqueness of orthogonal projection onto a span.

Now we finish by proving the reverse implication, that if v' does not contain $\tilde{z}$ or is not orthogonal to $\tilde{o}$, then $\Pi_{v'} \tilde{r} \neq \tilde{z}$. If $\tilde{z} \notin \mathsf{Span}(v')$, clearly $\Pi_{v'} \tilde{r} \neq \tilde{z}$. If $\tilde{z} \in \mathsf{Span}(v')$, but $\mathsf{Span}(v')$ is not orthogonal to $\tilde{o}$, consider some $u \neq \tilde{z}$ as any vector in $\mathsf{Span}(v')$ such that $\langle u, \tilde{o} \rangle > 0$. We can assume positive inner product since $\mathsf{Span}(v')$ contains both u and $-u$. Now by definition v' contains the plane spanned by u and $\tilde{z}$. The properties of orthogonal projection ensure $\langle \tilde{r} - \Pi_{v'} \tilde{r}, u - \Pi_{v'} \tilde{r} \rangle \leq 0$ Assuming by contradiction that $\Pi_{v'} \tilde{r} = \tilde{z}$, and noting $\tilde{r} = \tilde{z} + a\tilde{o}$ for some $a \geq 0$ (recall we assume $\langle \tilde{o}, \tilde{r} \rangle > 0$), we have $\langle \tilde{z} + a\tilde{o} - \tilde{z}, u - \tilde{z} \rangle = \langle a\tilde{o}, u - \tilde{z} \rangle = a \langle \tilde{o}, u \rangle \leq 0$. But since $a \geq 0$, this contradicts the assumption that $\langle \tilde{o}, u \rangle > 0$, and thus $\Pi_{v'} \tilde{r} \neq \tilde{z}$.

Component of output in $\mathsf{Span}(V)$. We have now established $\mathcal{V}_{good}$ is the cartesian product of $\{\mathsf{Span}(Z)\}$ and some set of d' dimensional linear subspaces, for some $d' \geq d/2 - \bar{T}$, and that the posterior distribution of V given $R = r$, $Z = z$, $X = x$, and $W = w$ is uniform over $\mathcal{V}_{good}(r, z, x)$. Let $\bar{t} = \bar{t}(r)$ be the value of $\bar{T}$ induced by $R = r$. Define $U = \Pi^{\perp}_{X,\mathcal{S}} \Pi^{\perp}_{Z} W$ and let E denote the event $\frac{1}{2}\|U\| - \|\Pi_V U\| \geq \alpha$. We have,

$$\begin{aligned}
\mathbb{P}[E] &= \int_{\substack{r:\bar{t}(r) \leq d/4 \\ w,x,z}} \mathbb{P}[E|r,w,x,z] \, \rho(r,w,x,z) dr dw dx dz + \int_{\substack{r:\bar{t}(r) > d/4 \\ w,x,z}} \mathbb{P}[E|r,w,x,z] \, \rho(w,x,z|r)\rho(r) dr dw dx dz \\
&\leq \max_{\substack{r:\bar{t}(r) \leq d/4 \\ w,x,z}} \{\mathbb{P}[E|r,w,x,z]\} + \frac{1}{20}. \qquad (9)
\end{aligned}$$

The inequality uses $\mathbb{E}\left[\bar{T}\right] \leq \frac{d}{80}$ and Markov's inequality. Let $r \in \{r : \bar{t}(r) \leq d/4\}$ and w, x, z be any possible instantiations of W, X, Z given $R = r$. These quantities determine U, and so let u be its realization (recall O is determined given R and Z). To bound the worst case probability, observe that under conditioning, $d' \geq d/4$. Thus by the previous analysis, V is a uniformly random subspace of dimension at least $d/4$ supported on a linear subspace of dimension at most d, and we can apply the Johnson-Lindenstrauss lemma (see, e.g. (Vershynin, 2018, Lemma 5.3.2)). Concretely, for some

universal constant C, $\mathbb{P}_V\left[\frac{1}{2}\|u\| - \|\Pi_V u\| \geq \alpha \;\middle|\; R = r, W = w, X = x, Z = z\right] \leq \exp\left(-Cd\alpha^2\right)$. Thus when $d \geq \frac{\log(20)}{C\alpha^2}$ (which holds by assumption for C_2 large enough), we have via Eqn. (9),

$$\mathbb{P}\left[\|\Pi^{\perp}_{X,\mathcal{S}}\Pi^{\perp}_{Z}W\| \geq 2\|\Pi_V\Pi^{\perp}_{X,\mathcal{S}}\Pi^{\perp}_{Z}W\| + 2\alpha\right] \leq \exp(-Cd\alpha^2) + \frac{1}{20} \leq \frac{1}{10}.$$

■

We now prove the information lower bound, Lemma 16. We remark that Lemma 13 (Appendix A) will be used several times throughout the proof, and the results to reach Eqn. (10) are given in the subsequent subsection, Appendix B.4.

Proof [Proof of Lemma 16] Fix some $k \in [K]$. Our information lower bound starts by leveraging a variant of Fano's method, whose details we defer to Lemma 18 in Appendix B.4. Via this lemma, we have that any estimator $\hat{W}$ satisfies,

$$\mathbb{P}\left[\|\hat{W} - \hat{X}_k\| \geq 40\alpha\right] \geq \frac{1}{2} - 8 \cdot \frac{I(\hat{W}; \hat{X}_k | X_{\neq k}, V)) + 1}{d}. \tag{10}$$

The rest of the proof will be devoted to showing we can construct such an estimator from W and a bounded amount of additional information.

Constructing estimator, $\hat{W}$. Our first step is to transform the output of $\mathcal{A}$ into a vector which is close (in Euclidean distance) to $\hat{X}_k$. To do this we, roughly speaking, need to remove several components of W: the component contained in $\mathsf{Span}(X_{\neq k})$, the component in $\mathsf{Span}(O)$, and the component in $\mathsf{Span}(Z)$.

Let $\mathcal{C}_O$ be a minimal α-cover of $\mathsf{Span}(O)$ (chosen deterministically given O) and define $\tilde{X}_k = \min_{c \in \mathcal{C}_O}\{\|\Pi_O X_k - c\|\}$. Now let $\mathcal{S}$ be the orthogonal complement of X_k inside $\mathsf{Span}(O)$ and $\tilde{\mathcal{S}}$ be the orthogonal complement of $\tilde{X}_k$ inside $\mathsf{Span}(O)$. Finally, let $\hat{W} = \Pi^{\perp}_{X_{\neq k},\tilde{\mathcal{S}}}\Pi^{\perp}_{Z}W$. Intuitively, $\Pi^{\perp}_{X_{\neq k},\tilde{\mathcal{S}}}$ approximately removes the component of W in the subspace spanned by O that does not overlap with X_k. The projection $\Pi^{\perp}_{Z}$ can be interpreted as removing the component of W in the "known" span of V. It is worth noting that $\hat{W}$ is still an accurate solution in some sense, a fact we will now show in more detail.

Showing $\hat{W}$ is accurate. Analyzing now the distance term inside the probability on the LHS of Eqn. (10) above, we have,

$$\|\hat{W} - \hat{X}_k\| \leq \|\hat{W} - C_1\alpha X_k\| + 2\alpha \leq 2\max\left\{\|\Pi^{\perp}_{X_k}\hat{W}\|, |\langle \hat{W}, X_k\rangle - C_1\alpha|\right\} + 2\alpha. \tag{11}$$

The first inequality uses the fact that $\mathcal{C}$ is a 2α-cover. We will now show that both terms in the maximum are bounded with constant probability using the accuracy condition. Towards this end, observe that for every possible instantiation of X and V, there exists a minimizer, $w^* = C_1\alpha\sum_{k=1}^{K} X_k$, with 0 loss. Recall $K = \frac{1}{C_1^2\alpha^2}$, so indeed $w^* \in \mathcal{B}(1)$.

We now analyze the inner product term in the RHS of Eqn. (11). Since the minimizer has 0 loss, by the accuracy condition on $\mathcal{A}$, it must be that $\mathbb{E}\left[|\langle W, X_k\rangle - C_1\alpha|\right] \leq \mathbb{E}\left[F(W) - F(w^*)\right] \leq$

α, and so by Markov's inequality, $\mathbb{P}\left[|\langle W, X_k\rangle - C_1\alpha| \geq 10\alpha\right] \leq \frac{1}{10}$. Thus it suffices to show $\langle \hat{W}, X_k\rangle \approx \langle W, X_k\rangle$. We first have,

$$|\langle W, X_k\rangle - \langle \hat{W}, X_k\rangle| = |\langle W, X_k\rangle - \langle W, \Pi_{\mathcal{S}}^{\perp} X_k\rangle| = |\langle W, \Pi_{\mathcal{S}} X_k\rangle|.$$

The first equality follows from the fact that $\hat{W} = \Pi_Z^{\perp}\Pi_{X_{\neq k}}^{\perp}\Pi_{\mathcal{S}}^{\perp} W$ and Z and $X_{\neq k}$ are orthogonal to X_k. The second equality uses $X_k = \Pi_{\mathcal{S}} X_k + \Pi_{\mathcal{S}}^{\perp} X_k$. Continuing,

$$\|\Pi_{\mathcal{S}} X_k\| \overset{(i)}{=} \|\Pi_{\mathcal{S}}\Pi_O \tilde{X}_k + \Pi_{\mathcal{S}}\Pi_O(X_k - \tilde{X}_k)\| \overset{(ii)}{=} \|\Pi_{\mathcal{S}}\Pi_O(X_k - \tilde{X}_k)\| \leq \|\Pi_O(X_k - \tilde{X}_k)\| \overset{(iii)}{\leq} \alpha.$$

Here, (i) uses the fact that $\mathcal{S} \subseteq \mathsf{Span}(O)$, (ii) uses the fact that $\Pi_{\mathcal{S}}\tilde{X}_k = 0$, and (iii) uses the fact that $\tilde{X}_k \in \mathsf{Span}(O)$ and $\|\Pi_O X_k - \tilde{X}_k\| \leq \alpha$. We now obtain $|\langle \hat{W} - W, X_k\rangle| \leq 2\alpha$ from Cauchy Schwartz and the fact that $\|\hat{W} - W\| \leq 1$.

This inner product difference with the previously derived fact that $\mathbb{P}\left[|\langle W, X_k\rangle - C_1\alpha| \geq 10\alpha\right] \leq \frac{1}{10}$ finally yields,

$$\mathbb{P}\left[|\langle \hat{W}, X_k\rangle - C_1\alpha| \geq 12\alpha\right] \leq \frac{1}{10}. \tag{12}$$

We now address the norm term in Eqn. (11). We have,

$$\begin{aligned}
\mathbb{E}\left[\|\Pi_V \Pi_{X,\mathcal{S}}^{\perp}\Pi_Z^{\perp} W\|\right] &= \mathbb{E}\left[\|\Pi_V \Pi_{X,\mathcal{S}}^{\perp}(\Pi_Z^{\perp} W - W + W)\|\right] \\
&\leq \mathbb{E}\left[\|\Pi_Z^{\perp} W - W\| + \|\Pi_V \Pi_{X,\mathcal{S}}^{\perp} W\|\right] \\
&\leq \mathbb{E}\left[\|\Pi_Z W\|\right] + \mathbb{E}\left[\|\Pi_V W\|\right] \\
&\leq \alpha.
\end{aligned}$$

Above, we have used the fact that V is orthogonal to $\mathsf{Span}(X) \cup \mathcal{S}$ and Lemma 13. The last inequality uses the accuracy condition, since $\mathcal{L}(w) - \mathcal{L}(w^*) \geq 2\|\Pi_V w\|$. Continuing,

$$\begin{aligned}
\mathbb{E}\left[\|\Pi_V \Pi_{X,\mathcal{S}}^{\perp}\Pi_Z^{\perp} W\|\right] \leq \alpha &\overset{(i)}{\implies} \mathbb{P}\left[\|\Pi_V \Pi_{X,\mathcal{S}}^{\perp}\Pi_Z^{\perp} W\| \geq 5\alpha\right] \leq 1/5 \\
&\overset{(ii)}{\implies} \mathbb{P}\left[\|\Pi_{X,\mathcal{S}}^{\perp}\Pi_Z^{\perp} W\| \geq 12\alpha\right] \leq 3/10 \\
&\overset{(iii)}{\iff} \mathbb{P}\left[\|\Pi_{X_k}^{\perp}\Pi_{X_{\neq k},\mathcal{S}}^{\perp}\Pi_Z^{\perp} W\| \geq 12\alpha\right] \leq 3/10
\end{aligned}$$

Implication (i) uses Markov's inequality, and (ii) results from Lemma 17. For implication (iii), we apply Lemma 13 since X_k is orthogonal to $X_{\neq k}$ and $\mathcal{S}$.

Now observe for any $u \in \mathcal{B}(1)$, $\|\Pi_{X_{\neq k},\mathcal{S}}^{\perp} u - \Pi_{X_{\neq k},\tilde{\mathcal{S}}}^{\perp} u\| \leq \alpha$ since $\mathsf{Span}(X_{\neq k}) \cup \mathcal{S}$ and $\mathsf{Span}(X_{\neq k}) \cup \tilde{\mathcal{S}}$ are close. That is, for any $v \in \mathcal{B}(1) \cap \mathcal{S}$, there exists $\xi \in \mathsf{Span}(O) \cap \mathcal{B}(\alpha)$ s.t. $v + \xi \perp \tilde{X}$, and thus $v + \xi \in \tilde{S}$. Thus we obtain,

$$\mathbb{P}\left[\|\Pi_{X_k}^{\perp}\Pi_{X_{\neq k},\tilde{\mathcal{S}}}^{\perp}\Pi_Z^{\perp}\hat{W}\| \geq 13\alpha\right] = \mathbb{P}\left[\|\Pi_{X_k}^{\perp}\hat{W}\| \geq 13\alpha\right] \leq 3/10. \tag{13}$$

Using this probability bound and Eqn. (12) above we obtain,

$$\mathbb{P}\left[\max\left\{\|\Pi_{\hat{X}_k}^{\perp}\hat{W}\|, |\langle\hat{W},\hat{X}_k\rangle - C_1\alpha|\right\} \geq 12\alpha\right] \leq \frac{1}{10} + \frac{3}{10} \leq \frac{2}{5}.$$

Combing this fact with Eqns. (10) and (11) we obtain,

$$\frac{2}{5} \geq \mathbb{P}\left[2\max\left\{\|\Pi_{\hat{X}}^{\perp}\hat{W}\|, |\langle\hat{W},X_k\rangle - C_1\alpha|\right\} + 2\alpha \geq 40\alpha\right] \geq \frac{1}{2} - 8\cdot\frac{I(\hat{W};\hat{X}_k|X_{\neq k},V)+1}{d}.$$

Consequently we have,

$$\frac{d}{80} \leq I(\hat{W};\hat{X}_k|X_{\neq k},V). \tag{14}$$

Showing $\hat{W}$ does not add much information. We now show that generating $\hat{W}$ does not add too much information beyond what is contained in W. Let $\hat{R}$ denote the information sent to $\tilde{\mathcal{O}}$ over the training run. Observe,

$$\begin{aligned}
I(\hat{W};\hat{X}_k|X_{\neq k},V) &\leq I(W,O,Z,\tilde{X};\hat{X}_k|X_{\neq k},V)\\
&\leq I(Y,\hat{R},O,Z,\tilde{X};\hat{X}_k|X_{\neq k},V)\\
&= I(Y,\hat{R};\hat{X}_k|X_{\neq k},V) + I(O,Z;\hat{X}_k|X_{\neq k},V,Y,\hat{R}) + I(\tilde{X}_k;\hat{X}_k|X_{\neq k},V,Y,\hat{R},O,Z)\\
&\overset{(i)}{\leq} I(Y,\hat{R};\hat{X}_k|X_{\neq k},V) + H(\tilde{X}_k|O)\\
&\overset{(ii)}{\leq} I(Y,\hat{R};\hat{X}_k|X_{\neq k},V) + \mathbb{E}\left[\bar{T}\right]\log(3/\alpha)\\
&\leq I(Y,\hat{R};\hat{X}_k|X_{\neq k},V) + \frac{d}{160}.
\end{aligned} \tag{15}$$

The first inequality uses the fact that $\hat{W}$ is determined by W, O, Z, and $\tilde{X}$. Step (i) uses that O and Z are deterministic conditioned on V and $\hat{R}$. Line (ii) uses $H(\tilde{X}_k|O) \leq \mathbb{E}\left[\log(|\mathcal{C}_O|)\right] \leq \mathbb{E}\left[d_O\right]\log(3/\alpha)$.

Finally, combining Eqns. (15) and (14), which hold for any choice of k, we obtain,

$$\frac{d}{160} = \min_k\left\{I(W;\hat{X}_k|X_{\neq k},V)\right\}.$$

■

B.3. Completing the proof of Theorem 5.

We conclude the proof of Theorem 5 by applying Lemmas 15 and 16, which yields,

$$\frac{d}{160} = \min_k\left\{I(W;X_k|X_{\neq k},V)\right\} \leq \mathop{\mathbb{E}}_{Q}\left[\sum_{t=1}^{\infty}\overline{\mathsf{Cnt}}_k^2(Q_t)\rho\right] + \mathbb{E}\left[\bar{T}\right]\log(K+1).$$

When $\mathbb{E}\left[\bar{T}\right] \leq \frac{d}{320\log(1/\alpha^2)}$ we obtain $\mathop{\mathbb{E}}_{Q}\left[\sum_{t=1}^{\infty}\overline{\mathsf{Cnt}}_k^2(Q_t)\rho\right] = \Omega(d)$. We now consider both cases of the theorem statement. For part 1, using $\overline{\mathsf{Cnt}}_k(\cdot) \leq \sqrt{3C_1 d/\rho}$ one can see that, $\mathbb{E}\left[\sum_{t=1}^{\infty}\overline{\mathsf{Cnt}}_k^2(Q_t)\right] = \Omega(d/\rho) \implies \mathbb{E}\left[\sum_{t=1}^{\infty}\overline{\mathsf{Cnt}}_k(Q_t)\right] = \Omega(\sqrt{d/\rho})$. Now we obtain,

$$\mathbb{E}\left[\|M\|_1\right] = \Big[\sum_{t=1}^{\infty}\sum_{k=1}^{K}\mathsf{Cnt}_k(Q_t)\Big] \geq \mathbb{E}\Big[\sum_{k=1}^{K}\sum_{t=1}^{\infty}\overline{\mathsf{Cnt}}_k(Q_t)\Big] = \Omega\Big(\frac{\sqrt{d}}{\alpha^2\sqrt{\rho}}\Big).$$

The first part of the theorem is obtained since $\max_{v,x}\{\mathop{\mathbb{E}}_{\mathcal{A},\tilde{\mathcal{O}}}[\|M\|_1|X=x,V=v]\} \geq \mathop{\mathbb{E}}_{\mathcal{A},\tilde{\mathcal{O}},X,V}[\|M\|_1]$.

If the max batch size is bounded, using the bound $\overline{\mathsf{Cnt}}_k(Q_t) \leq \bar{m}$ for any $k \in [K], t \in \mathbb{Z}^+$, and proceeding similarly to above we obtain $\max_{v,x}\{\mathop{\mathbb{E}}_{\mathcal{A},\tilde{\mathcal{O}}}[\|M\|_1|X=x,V=v]\} \geq \frac{d}{\alpha^2\bar{m}\rho}$ as desired.

B.4. Fano style bound used in proof of Lemma 16

In this section, we establish Eqn. (10), which is used in the proof of the information lower bound. Lemma 20 gives a variant of Fano's inequality which is useful due to the fact that multiple elements in the support of $\hat{X}_k$ are considered a success when estimating $\hat{X}_k$. Lemma 19 verifies that the entropy of $\hat{X}$ is high. Recall that $\hat{X}_k$ is essentially a uniformly random vector that has been snapped to a high resolution cover. Finally, Lemma 18 derives the actual information bound, and is largely algebraic in nature.

Lemma 18 *For any estimator $\hat{W}$, it holds that,*

$$\mathbb{P}\left[\|\hat{W}-\hat{X}_k\| \geq 40\alpha\right] \geq \frac{1}{2} - 8\cdot\frac{I(\hat{W};\hat{X}_k|X_{\neq k},V)) + 1}{d}.$$

Proof Recall $\hat{X}_k = \arg\min_{c\in\mathcal{C}}\{\|C_1\alpha X_k - c\|\}$ and note $\mathcal{C}$ has dimension $d_{\mathcal{C}} = d - d/2 - K + 1 \geq d/4$ (since $d \geq C_2/\alpha^2$). It can be shown that $\hat{X}_k$ has large entropy. We defer this fact to Lemma 19 given below, and here apply this lemma to obtain $H(\hat{X}|X_{\neq k} = x_{\neq k}, V = v) \geq (d_{\mathcal{C}}-1)\log(C_1\alpha/2\alpha) = (d_{\mathcal{C}}-1)\log(C_1/2)$. Define,

$$P_{err} = \mathbb{P}\left[\|\hat{W}-\hat{X}_k\| \geq 40\alpha \mid X_{\neq k} = x_{\neq k}, V = v\right] \quad \text{and} \quad N_{max} = \max_{c\in\mathcal{C}}\left\{|\{c'\in\mathcal{C} : \|c-c'\| \leq 40\alpha\}|\right\}.$$

Now by a variant of Fano's inequality (see Lemma 20 below),

$$P_{err} \geq \frac{H(\hat{X}_k|X_{\neq k}=x_{\neq k},V=v) - \log(N_{max})}{\log(|\mathcal{C}|/N_{max})} - \frac{I(\hat{W};\hat{X}_k|X_{\neq k}=x_{\neq k},V=v)+1}{\log(|\mathcal{C}|/N_{max}-1)}.$$

Observe N_{max} is at most the number of α-balls that can be packed into a 40α-ball in $d_{\mathcal{C}}$-dimensions. Bounds on the packing number then imply $N_{\max} \leq 120^{d_{\mathcal{C}}}$ and $|\mathcal{C}| \in [C_1^d, (3C_1)^{d_{\mathcal{C}}}]$. Now for $C_1 = 480$ and d larger than some constant (recall $d_{\mathcal{C}} \geq d/4$) we obtain,

$$\begin{aligned} P_{err} &\geq \frac{(d_{\mathcal{C}}-1)\log(C_1/2) - d_{\mathcal{C}}\log(120)}{d_{\mathcal{C}}\log(1440)} - \frac{I(\hat{W};\hat{X}_k|X_{\neq k}=x_{\neq k},V=v)+1}{d_{\mathcal{C}}\log(C_1/120-1)} \\ &\geq \frac{1}{2} - 8\cdot\frac{I(\hat{W};\hat{X}_k|X_{\neq k}=x_{\neq k},V=v)+1}{d} \end{aligned}$$

Taking the expectation over $X_{\neq k}$ and V then yields,

$$\mathbb{P}\left[\|\hat{W}-\hat{X}_k\| \geq 40\alpha\right] = \mathop{\mathbb{E}}_{X_{\neq k},V}\left[\mathbb{P}\left[\|\hat{W}-\hat{X}_k\| \geq 40\alpha | X_{\neq k}=x_{\neq k}, V=v\right]\right]$$
$$\geq \frac{1}{2} - 8\cdot\frac{I(\hat{W};\hat{X}_k|X_{\neq k},V))+1}{d}.$$

■

Lemma 19 *Let $\mathcal{C}$ be a maximal α-packing of the radius r ball in $d_{\mathcal{C}}$ dimensions. Let X be a uniformly random unit vector and $\hat{X} = \arg\min_{c\in\mathcal{C}}\{\|x-c\|\}$. Then $H(\hat{X}) \geq (d_{\mathcal{C}}-1)\log(r/2\alpha)$.*

Proof In the following $\mathcal{S}(r,c)$ be the surface of the $d_{\mathcal{C}}$-dimensional ball of radius r centered at c. Since $\mathcal{C}$ is a 2α cover and X is uniform over the surface of a $d_{\mathcal{C}}$-dimensional ball, for any $c\in\mathcal{C}$, $\mathbb{P}[\hat{X}=c]$ can be bounded using the surface area, i.e. $(d_{\mathcal{C}}-1)$-dimensional volume, of the spherical cap containing points within 2α distance of c on a r-radius ball. Now observe that the surface area of this cap is at most the surface area of a ball of radius 2α. To see this, let $\mathcal{X}$ be the convex hull of $\{x\in\mathcal{S}(r,0) : \|x-c\|\leq 2\alpha\}$. Taking the convex hull does not decrease the measure. Further, the mapping $f:\mathcal{S}(r,c)\mapsto\mathcal{X}$ given by $f(x)=\Pi_{\mathcal{X}}(x)$ is a contraction, and because $\mathcal{X}$ is convex the image of $\mathcal{S}(r,c)$ under f is $\mathcal{X}$. Thus $\mathsf{Vol}(\mathcal{S}(r,c)) \geq \mathsf{Vol}(\{f(x) : x\in\mathcal{S}(r,c)\}) = \mathsf{Vol}(\mathcal{X})$.

Now since X has uniform density of value $\frac{1}{\mathsf{Vol}(\mathcal{S}(r,0))}$, we have $\mathbb{P}[\hat{X}=c] \leq \frac{\mathsf{Vol}(\mathcal{S}(2\alpha,0))}{\mathsf{Vol}(\mathcal{S}(r,0))} = \frac{(2\alpha)^{d_{\mathcal{C}}-1}\mathsf{Vol}(\mathcal{S}(1,0))}{r^{d_{\mathcal{C}}-1}\mathsf{Vol}(\mathcal{S}(1,0))} = (2\alpha/r)^{d_{\mathcal{C}}-1}$. This establishes that $H(\hat{X}) = \mathbb{E}\left[\log(1/\mathbb{P}[\hat{X}=\hat{x}])\right] \geq (d_{\mathcal{C}}-1)\log(r/2\alpha)$. ■

Lemma 20 *(Restatement of (Duchi and Wainwright, 2013, Proposition 1)) Let X and Y be random variables supported on the discrete set $\mathcal{C}$. Let $\tau\geq 0$ and define $N_{max} = \max_{c\in\mathcal{C}}\{|\{c'\in\mathcal{C} : \|c-c'\|\leq\tau\}|\}$. Then*

$$\mathbb{P}\left[\|X-Y\|\geq\tau\right] \geq \frac{H(X)-\log(N_{max})}{\log\left(\frac{|\mathcal{C}|}{N_{max}}\right)} - \frac{I(Y;X)+1}{\log\left(\frac{|\mathcal{C}|}{N_{max}}-1\right)}.$$

Proof Define $N_{min} = \min_{c\in\mathcal{C}}\{|\{c'\in\mathcal{C} : \|c-c'\|\leq\tau\}|\}$. The claim is obtained from a simple manipulation of (Duchi and Wainwright, 2013, Proposition 1). Starting from that statement we have the following,

$$\mathbb{P}\left[\|X-Y\|\geq\tau\right] \geq \frac{H(X|Y)-\log(N_{max})-1}{\log\left(\frac{|\mathcal{C}|-N_{min}}{N_{max}}\right)}$$
$$\geq \frac{H(X)-\log(N_{max})}{\log\left(\frac{|\mathcal{C}|-N_{min}}{N_{max}}\right)} - \frac{I(Y;X)+1}{\log\left(\frac{|\mathcal{C}|-N_{min}}{N_{max}}\right)}$$
$$\geq \frac{H(X)-\log(N_{max})}{\log\left(\frac{|\mathcal{C}|}{N_{max}}\right)} - \frac{I(Y;X)+1}{\log\left(\frac{|\mathcal{C}|}{N_{max}}-1\right)}.$$

■

Appendix C. Additional Supplement to Section 3

C.1. Proof of Theorem 6 (upper bound via DP-SGD)

Algorithm 3 DP-SGD

Require: Oracle privacy $\rho \leq 1$, Batch size bound $\bar{m}$, Accuracy $\alpha \geq 0$, Oracle $\mathcal{O}_{\mathcal{L}}$, Constraint Set $\mathcal{W}$ of width B, Lipschitz constant L

Pick any $w_0 \in \mathcal{W}$

If $\alpha \geq \frac{BL}{3}$, stop and release w_0

Set $m = \min\left\{\sqrt{d/\rho}, \bar{m}\right\}$, and $\sigma = L \max\left\{\frac{1}{\sqrt{d}}, \frac{1}{\bar{m}\sqrt{\rho}}\right\}$

Set $T = \frac{B^2L^2}{\alpha^2} \cdot \max\left\{1, \frac{d}{\bar{m}^2\rho}\right\}$, $\eta = \frac{B}{L\sqrt{T}} \cdot \min\left\{1, \frac{\bar{m}\sqrt{\rho}}{\sqrt{d}}\right\}$,

for $t = 1...T$ **do**

  Sample minibatch $i_1, ..., i_m$ uniformly from $[n]$

  Obtain gradients G_t from $\{\mathcal{O}_{\mathcal{L}}(w_t, i_1), ..., \mathcal{O}_{\mathcal{L}}(w_t, i_m)\}$

  $w_{t+1} = \Pi_{\mathcal{W}}\left[w^t - \eta\left(\frac{1}{m}\sum_{g \in G_t} g + \xi^t\right)\right]$, where $\xi_t \sim \mathcal{N}(0, \mathbb{I}_d\sigma^2)$

end for

Output: $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w^t$

The fact that Algorithm 3 uses a ρ-zCDP oracle comes directly from the guarantees of the Gaussian mechanism Bun and Steinke (2016). Theorem 6 follows from the subsequent two lemmas.

Lemma 21 *Algorithm 3 is α'-accurate for $(\mathcal{F}^n_{L,\infty}, \mathcal{K}_B)$ with $\alpha' = O(\min\{BL, \alpha\})$. Further, the algorithm has oracle complexity $O\left(\frac{B^2L^2}{\alpha^2}\left(\frac{\sqrt{d}}{\sqrt{\rho}} + \frac{d}{\bar{m}\rho}\right)\right)$.*

Proof The result is essentially a corollary of known convergence results for SGD with noise. For example, by Lemma 3.3 of [BFTT19], Algorithm 3 obtains excess empirical risk,

$$\mathbb{E}[\mathcal{L}(\bar{w}) - \mathcal{L}(w^*)] = O\left(\frac{B^2}{\eta T} + \eta L^2 + \eta\sigma^2 d\right). \tag{16}$$

Plugging in the parameters settings verifies the utility guarantee. The BL term in the error comes from the trivial bound when w_0 is released. The oracle complexity is $Tm = O\left(\frac{B^2L^2}{\alpha^2}\left(\frac{\sqrt{d}}{\sqrt{\rho}} + \frac{d}{\bar{m}\rho}\right)\right)$. ■

Lemma 22 *Let $\delta \in [0, 1]$. Algorithm 3 run with $\alpha \geq 26\alpha^*_{1,\delta}$ and $\rho = \frac{1}{\log(1/\delta)}$ is (ϵ, δ)-DP with $\epsilon = 3\frac{\alpha^*_{1,\delta}}{\alpha}$.*

Proof We will use truncated differential privacy to perform the analysis, and will thus use several results from Bun et al. (2018a). By the guarantees of the Gaussian mechanism, the private oracle satisfies (ρ, ∞)-tCDP (or equivalently ρ-zCDP). Provided that $\log(n/m) = \log(n\sqrt{\rho/d}) \geq 3\rho(2 + \log_2(1/\rho))$, we can apply the privacy amplification via subsampling from (Bun et al., 2018a, Theorem 11). Note we can assume $\frac{\sqrt{d\log(1/\delta)}}{n} \leq 1/3$ (otherwise the algorithm releases

w_0). Thus for $\rho \leq \frac{1}{\log(1/\delta)}$, $\log(n\sqrt{\rho/d}) = \log(\frac{n}{\sqrt{d\log(1/\delta)}}) \geq 1$, and so our setting of ρ satisfies the subsampling lemma conditions. We thus obtain that each iteration satisfies (ρ', ω')-tCDP with $\rho' = 13(m/n)^2\rho = 13d/n^2$ and $\omega' = \frac{1}{4\rho}$. The composition properties of tCDP imply the overall algorithm satisfies (ρ'', ω'')-tCDP with $\rho'' = T\rho' = \frac{13B^2L^2d}{\alpha^2n^2}$ and $\omega'' = \omega' = \frac{1}{4\rho}$. We can now apply the following conversion to (ϵ, δ)-DP, given by (Bun et al., 2018a, Lemma 6),

$$\epsilon = \begin{cases} \rho'' + 2\sqrt{\rho''\log(1/\delta)} & \text{if } \log(1/\delta) \leq (\omega''-1)^2\rho'' \\ \rho''\omega'' + \frac{\log(1/\delta)}{\omega''-1} & \text{if } \log(1/\delta) \geq (\omega''-1)^2\rho'' \end{cases}$$

Observe that under our setting of ρ, the condition $\log(1/\delta) \leq (\omega''-1)^2\rho''$ is satisfied whenever $\alpha \geq 26\frac{BL\sqrt{d\log(1/\delta)}}{n} = 26\alpha^*_{1,\delta}$, in which case we obtain privacy $\epsilon = 3\frac{\alpha^*_{1,\delta}}{\alpha}$. ■

C.2. Extension to tCDP

In this section, we show how our lower and upper bound in the non-smooth setting hold under the notion of truncated CDP Bun et al. (2018a).

Definition 23 *(Truncated Concentrated Differential Privacy) Let $\rho > 0$ and $\omega \geq 1$. A randomized algorithm $\mathcal{M} : \mathcal{X}^n \mapsto \mathcal{Y}$ satisfies ω-truncated ρ-concentrated differential privacy, (ρ, ω)-tCDP, if for all datasets, S, S', differing in at most one element, it holds that for all $\alpha \in (1, \omega)$ that $D_\alpha(\mathcal{M}(X)||\mathcal{M}(S')) \leq \rho\alpha$, where D_α is the α-Rényi divergence.*

Lower bound. Consider the case where each $\mathcal{O}_\perp$ satisfies (ρ, ω)-tCDP for $\omega \geq \min\{\sqrt{d/\rho}, \bar{m}\}$. For comparison, note that ρ-zCDP is equivalent to (ρ, ∞)-tCDP. To extend the lower bound to this notion requires only a slight modification to Lemma 15, which upper bounds the information gained during the optimization procedure. Specifically, we recall Eqn. (6), which showed that under ρ-zCDP (using the same notation),

$$I(Y_t; \hat{X}_k | Q_t = q_t, P_t = p_t) \overset{(i)}{\leq} \mathop{\mathbb{E}}_{x_k, x'_k \leftarrow X_k} \left[\mathsf{KL}\Big(\tilde{\mathcal{O}}_\perp(G_{t,-k}(x_k))||\mathcal{M}(G_{t,-k}(x'_k))\Big)\right]$$
$$\overset{(ii)}{\leq} \mathsf{Cnt}_k^2(q_t)\rho.$$

Inequality (i) holds irregardless of any privacy notion, and so the consideration is inequality (ii). We observe that we only ever need to apply this bound for $\mathsf{Cnt}_k^2(q_t) \leq \sqrt{d/\rho}$. In the other regime, the proof upper bounds the information via entropy, $I(Y_t; \hat{X}_k | Q_t = q_t, P_t = p_t) = O(d)$. For $m \leq \omega$, (ρ, ω)-tCDP implies $(\rho m^2, \omega/m)$-group tCDP for groups of size m. Since tCDP also bounds KL divergence, this is sufficient to obtain inequality (ii), and the rest of the proof proceeds exactly the same as in the zCDP case. Similarly, in the case where $\bar{m} \leq \sqrt{d/\rho}$, we observe that we only have to use the group privay properties of tCDP for groups of size at most $\bar{m}$.

Upper bound. The fact that our upper bound, Algorithm 3, satisfied tCDP is already proved as an intermediate step in the proof of Lemma 22. Specifically, for any setting of ρ such that $\log(n\sqrt{\rho/d}) \geq 3\rho(2 + \log_2(1/\rho))$, the algorithm is (ρ', ω)-tCDP with $\rho' = O(\frac{1}{\alpha^2}\frac{d}{n^2})$ and $\omega = 1/\rho$. We remark that while many other upper bounds in the literature are stated for approximate DP, they often satisfy tCDP guarantees as they rely on zCDP oracles and subsampling.

Appendix D. Supplement to Section 4

D.1. Proof of Theorem 11 (oracle complexity lower bound for smooth losses)

Theorem 11 follows from two runtime lower bounds we prove in this section. The first is an $\Omega\left(\frac{BL\sqrt{d}}{\alpha}\right)$ lower bound for private optimizers, and a $\Omega\left(\min\left\{\frac{B^2L^2}{\alpha^2}, n\right\}\right)$ lower bound which holds even without privacy. Much of the private lower bound proof depends on a DP mean estimation lower bound for procedures which only access a limited number of samples from the dataset. These results are provided subsequently in Appendix D.1.1.

Theorem 24 *Let $\delta \leq \frac{1}{16nd}$, $\epsilon \leq \log(1/\delta)$, and d be larger than some constant. Let $\mathcal{A}$ be an α-accurate optimizer for $(\mathcal{F}^n_{L,\frac{\alpha}{B^2}}, \mathcal{K}_B)$ which is (ϵ,δ)-DP. Further assume that, in expectation, $\mathcal{A}$, makes at most s calls to the gradient oracle. Then, $s = \Omega\big(\frac{BL\sqrt{d}}{\alpha}\big)$.*

Proof Define the distribution $\mathcal{D}_\theta$, which for any vector $\theta \in [-1,1]^d$, is the product distribution where, for any $j \in [d]$, a sample has its j'th coordinate as 1 with probability $(1+\theta_j)/2$ and as -1 with probability $(1-\theta_j)/2$. Let $\Theta \sim \mathsf{Unif}([-1,1]^d)$. For $N = n\alpha/(BL)$, let $\tilde{x}_1, ..., \tilde{x}_N \sim \frac{L}{\sqrt{d}}\mathcal{D}^N_\Theta$. Now we take $x_1, ..., x_n$ to be a random permutation of $\tilde{x}_1, ..., \tilde{x}_N$ and $n-N$ zero vectors. Define the loss function,

$$\mathcal{L}(w) = \frac{1}{n}\sum_{i=1}^n \langle w, x_i\rangle + \lambda\|w\|^2. \tag{17}$$

We will use the constraint set $\mathcal{B}(72B)$, and define $w^* = \operatorname*{arg\,min}_{w\in\mathcal{B}(24B)} \mathcal{L}(w)$. Note that the *unconstrained* empirical minimizer is $\tilde{w}^* = \frac{\sum_{i=1}^n x_i}{2n\lambda}$. Since $\|\sum_{i=1}^n x_i\| \leq NL$, we set $\lambda \geq \frac{NL}{144nB}$ so that $\|\tilde{w}^*\| \leq 72B$ and thus $w^* = \tilde{w}^* = \frac{\sum_{i=1}^n x_i}{2n\lambda}$. Further, under the setting of N, we have $\lambda = \frac{\alpha}{144B^2} \leq \frac{L}{B}$, which ensures the loss is $2L$-Lipschitz over the set $\mathcal{B}(72B)$.

Now we will show that any w which achieves small excess risk is close to w^*. For any w we have,

$$\begin{aligned}
\mathcal{L}(w) - \mathcal{L}(w^*) &= \left\langle w - w^*, \frac{1}{n}\sum_{i=1}^n x_i\right\rangle + \lambda\left(\|w\|^2 - \|w^*\|^2\right)\\
&= 2\lambda\left\langle w - w^*, -w^*\right\rangle + \lambda\left(\|w\|^2 - \|w^*\|^2\right)\\
&= 2\lambda(\|w^*\|^2 - \langle w, w^*\rangle) + \lambda\left(\|w\|^2 - \|w^*\|^2\right)\\
&= 2\lambda\left(\|w^*\|^2 - \frac{1}{2}\|w^*\|^2 - \frac{1}{2}\|w\|^2 + \frac{1}{2}\|w - w^*\|^2\right) + \lambda\left(\|w\|^2 - \|w^*\|^2\right)\\
&= \lambda\|w - w^*\|^2.
\end{aligned}$$

where the fourth equality comes from $\langle a, b\rangle = \frac{1}{2}(\|a\|^2 + \|b\|^2 - \|a-b\|^2)$.

Now observe $\frac{1}{LN}\sum_{i=1}^N x_i = \frac{2n\lambda}{LN}w^*$ and consider the mean estimation candidate $\bar{\Theta} = \frac{2n\lambda}{LN}w$, where w is the output of the differentially private solver $\mathcal{A}$. Continuing from the above display we

then have,

$$\begin{aligned}\mathop{\mathbb{E}}_{\Theta,S,\mathcal{A}}\left[\mathcal{L}(w)-\mathcal{L}(w^*)\right] &= \mathbb{E}\left[\lambda\left\|w-w^*\right\|^2\right]\\ &= \frac{L^2N^2}{n^2\lambda}\mathbb{E}\left[\left\|\bar{\Theta}-\Theta_S\right\|^2\right]\\ &\geq \frac{L^2N^2}{4n^2\lambda}\mathbb{E}\left[\left\|\bar{\Theta}-\Theta_S\right\|\right]^2. \end{aligned} \tag{18}$$

Now because the nonzero vectors are randomly assigned indices in $[n]$, and there are at most N of them, the expected number of nonzero vectors in S accessed by $\mathcal{A}$ is $\frac{Ns}{n}=\frac{s\alpha}{BL}$. Assume by contradiction that $s<\frac{BL\sqrt{d}}{18\alpha\sqrt{\log(1/\delta)}}$, by Lemma 26 (given in the following section) this means $\mathbb{E}\left[\left\|\bar{\Theta}-\Theta_S\right\|\right]^2\geq\frac{1}{36}$.

Recalling $N=n\alpha/[BL]$ and $\lambda=\frac{NL}{144nB}$, under the assumption that $s<\frac{BL\sqrt{d}}{18\alpha\sqrt{\log(1/\delta)}}$ and $\mathbb{E}\left[\mathcal{L}(w)-\mathcal{L}(w^*)\right]\leq\alpha$, we have

$$\alpha\geq\frac{L^2N^2}{72n^2\lambda}=\frac{2BLN}{n}=2\alpha.$$

This is a contradiction and so it must be that $s\geq\frac{BL\sqrt{d}}{18\alpha\sqrt{\log(1/\delta)}}$. ■

The non-private component of the lower bound comes from the following result. We note a similar result was proved in Woodworth and Srebro (2016), and we provide the following only to extend it to algorithms with randomized running time.

Lemma 25 *Let $\mathcal{A}$ be an α-accurate optimizer for $(\mathcal{F}^n_{L,\frac{\alpha}{B^2}},\mathcal{K}_B)$. Further assume that, in expectation, $\mathcal{A}$, makes at most s calls to the gradient oracle. Then, $s=\Omega\big(\min\left\{\frac{B^2L^2}{\alpha^2},n\right\}\big)$.*

Proof We will first give a *distributional* mean estimation bound for estimators which use a variable number of samples from the distribution. Let T denote the number of samples used by the mean estimation procedure (which may be data dependent). Let $\mathcal{A}_t$ denote the set of all algorithms which w.p. 1 use at most t samples. We have for any estimator $\mathcal{M}$,

$$\begin{aligned}\mathbb{E}\left[\left\|\mathcal{M}(S)-\Theta\right\|\right] &= \sum_{\substack{\theta\in\mathsf{Supp}(\Theta)\\ s\in\mathsf{Supp}(S)}}\sum_{t=1}^{\infty}\mathop{\mathbb{E}}_{\mathcal{M}}\left[\left\|\mathcal{M}(S)-\Theta\right\|\,\middle|\,T=t,\Theta=\theta,S=s\right]\mathbb{P}[T=t|\Theta,S]\mathbb{P}[\Theta=\theta,S=s]\\ &\geq \sum_{\theta,s}\mathop{\mathbb{E}}_{\mathcal{M}}\left[\left\|\mathcal{M}(S)-\Theta\right\|\,\middle|\,T\leq 10s,\Theta=\theta,S=s\right]\mathbb{P}[T\leq 2s|\Theta=\theta,S=s]\mathbb{P}[\Theta=\theta,S=s]\\ &\geq \frac{1}{2}\sum_{\theta,s}\mathop{\mathbb{E}}_{\mathcal{M}}\left[\left\|\mathcal{M}(S)-\Theta\right\|\,\middle|\,T\leq 2s,\Theta=\theta,S=s\right]\mathbb{P}[\Theta=\theta,S=s]\\ &= \frac{1}{2}\mathop{\mathbb{E}}_{\mathcal{M},\Theta,S}\left[\left\|\mathcal{M}(S)-\Theta\right\|\,\middle|\,T\leq 2s\right]\\ &\geq \frac{1}{2}\min_{\mathcal{M}\in\mathcal{A}_{2s}}\left\{\mathop{\mathbb{E}}_{\mathcal{M},\Theta,S}\left[\left\|\mathcal{M}(S)-\Theta\right\|\right]\right\}.\end{aligned}$$

The second inequality uses the bound on expected running time and Markov's inequality. Now, the fact that $\min_{\mathcal{M}\in\mathcal{A}_{2s}}\{\mathop{\mathbb{E}}_{\mathcal{M},\Theta,S}[\|\mathcal{M}(S)-\Theta\|]\}=\Omega(\frac{L}{\sqrt{s}})$ follows from classic mean estimation lower bounds. For example, using $\Theta\sim\mathsf{Unif}([-2/\sqrt{s},2/\sqrt{s}]^d)$ and $S\sim\frac{L}{\sqrt{d}}\mathcal{D}_\Theta^n$ suffices by (Ullah et al., 2024, Theorem 13). Thus we have

$$\mathbb{E}\left[\|\mathcal{M}(S)-\Theta\|\right]=\Omega\left(\frac{L}{\sqrt{s}}\right). \tag{19}$$

We can now leverage this lower bound and the loss construction from Theorem 24. Letting $\bar{\Theta}=\frac{2n\lambda}{LN}w$ and $\Theta_S=\frac{1}{n}\sum_{x\in S}x$ and using the loss construction from Theorem 24/Eqn. (17) with $N=n$ and $\lambda=\alpha/B^2$, we obtain from Eqn. (18) that,

$$\alpha\geq\frac{B^2}{4\alpha}\mathbb{E}\left[\|\bar{\Theta}-\Theta_S\|\right]^2\implies\alpha\geq B\mathbb{E}\left[\|\bar{\Theta}-\Theta_S\|\right].$$

Clearly for any Θ, $\mathbb{E}\left[\|\Theta_S-\Theta\|\right]\leq\frac{L}{\sqrt{n}}$. Thus by Eqn. (19), for some constant C,

$$\alpha\geq B\mathbb{E}\left[\|\bar{\Theta}-\Theta_S\|\right]\geq B\left(\frac{CL}{\sqrt{s}}-\frac{L}{\sqrt{n}}\right)\implies s=\Omega\left(\min\left\{\frac{B^2L^2}{\alpha^2},n\right\}\right).$$

■

D.1.1. DP MEAN ESTIMATION WITH VARIABLE ACCESS

In this section, we provide lower bounds for DP mean estimation when the algorithm only accesses a random subset of the dataset. In the following, we will denote the distribution $\mathcal{D}_\theta$, which for any vector $\theta\in[-1,1]^d$, is the product distribution where, for any $j\in[d]$, a sample has its j'th coordinate as 1 with probability $(1+\theta_j)/2$ and as -1 with probability $(1-\theta_j)/2$.

Lemma 26 *Let $\delta\leq\frac{1}{12nd}$ and $\epsilon\leq\log(1/\delta)$. Let $\mathcal{A}$ be an (ϵ,δ)-DP algorithm such that for any dataset $S\in\mathcal{B}(1)^n$, in expectation $\mathcal{A}$ accesses at most s elements of S. Draw $\Theta\sim\mathsf{Unif}([-1,1]^d)$ and $S=\{X_1,\ldots,X_n\}\sim\frac{1}{\sqrt{d}}\mathcal{D}_\Theta^n$. It holds that,*

$$\mathop{\mathbb{E}}_{\Theta,\mathcal{A},S}\left[\left\|\mathcal{A}(S)-\frac{1}{n}\sum_{i=1}^n X_i\right\|\right]\geq\frac{1}{6}\quad\textit{or}\quad s\geq\frac{\sqrt{d}}{18\sqrt{\log(1/\delta)}}.$$

To prove the lemma, we will use a standard result in the privacy lower bound literature, often called the fingerprinting lemma. This result stems from Dwork et al. (2015) and can be obtained more directly from (Ullah et al., 2024, Lemma 4). Our accuracy assumption differs slightly from theirs. This modification can be obtained by simply avoiding an application of Jenson's inequality at the end of their proof, which we have copied below for completeness.

Lemma 27 *Let θ be sampled uniformly from $\mathsf{Unif}([-1,1]^d)$. Let $\mathcal{A}$ satisfy $\|\mathop{\mathbb{E}}_{S\sim\mathcal{D}_\theta^n}[\mathcal{A}(S)]-\theta\|\leq\sqrt{d}/6$ for any $\theta\in[-1,1]^d$. Then one has,*

$$\mathop{\mathbb{E}}_{\mathcal{A},S,\Theta}\left[\sum_{i=1}^n\langle\mathcal{A}(S),X_i-\Theta\rangle\right]\geq\frac{d}{3}.$$

Proof In the following we treat $\mathcal{A}$ as a deterministic function and bound $\mathop{\mathbb{E}}_{S,\Theta}\left[\sum_{i=1}^{n}\left\langle \mathcal{A}(S), X_i - \Theta\right\rangle\right]$. This is sufficient to bound $\mathop{\mathbb{E}}_{\mathcal{A},S,\Theta}\left[\sum_{i=1}^{n}\left\langle \mathcal{A}(S), X_i - \Theta\right\rangle\right]$ for randomized $\mathcal{A}$, since the analysis holds for any function (i.e. the distribution does not depend on $\mathcal{A}$). Further, we start with the one dimensional case such that $\Theta \in \mathbb{R}$. Define $g(\Theta) = \mathop{\mathbb{E}}_{S\sim\mathcal{D}_\Theta^n}\left[\mathcal{A}(S)\right]$. We start by applying results developed in Dwork et al. (2015),

$$\begin{aligned}
\mathop{\mathbb{E}}_{S,\Theta}\left[\mathcal{A}(S)\sum_{i=1}^{n}(X_i-\Theta)\right] &\overset{(i)}{=} \mathop{\mathbb{E}}_{\Theta}\left[g'(\Theta)(1-\Theta^2)\right] \\
&\overset{(ii)}{\geq} 1 - \mathbb{E}\left[\Theta^2\right] + 2\mathop{\mathbb{E}}_{\Theta}\left[(g(\Theta)-\Theta)\Theta\right] - \frac{|g(-1)+1|+|g(1)-1|}{2} \\
&\geq 2/3 + 2\mathop{\mathbb{E}}_{\Theta}\left[(g(\Theta)-\Theta)\Theta\right] - \frac{|g(-1)+1|+|g(1)-1|}{2}.
\end{aligned}$$

Above, (i) comes from (Dwork et al., 2015, Lemma 5) and (ii) comes from (Dwork et al., 2015, Lemma 14). We now have

$$\begin{aligned}
\mathop{\mathbb{E}}_{S,\Theta}\left[\mathcal{A}(S)\sum_{i=1}^{n}(X_i-\Theta)\right] &\geq 2/3 + \frac{|g(-1)+1|+|g(1)-1|}{2} + 2\mathop{\mathbb{E}}_{\Theta}\left[(g(\Theta)-\Theta)\Theta\right] \\
&\geq 2/3 + \frac{|g(-1)-1|+|g(1)-1|}{2} - 2\mathop{\mathbb{E}}_{\Theta}\left[|g(\Theta)-\Theta|\cdot|\Theta|\right] \\
&\geq 2/3 - \frac{|\mathbb{E}_{S\sim\mathcal{D}_{-1}}[\mathcal{A}(S)]+1| + |\mathbb{E}_{S\sim\mathcal{D}_1}[\mathcal{A}(S)]-1|}{2} \\
&\quad - 2\mathop{\mathbb{E}}_{\Theta}\left[\left|\mathop{\mathbb{E}}_{S\sim\mathcal{D}_\Theta}\left[\mathcal{A}(S)\right]-\Theta\right|\right].
\end{aligned}$$

Above we use the fact that $|\Theta| \leq 1$ and the definition of g.

We can now extend the above analysis to higher dimensions. For $\Theta \in \mathbb{R}^d$, the above holds for each Θ_j, $j \in [d]$. For convenience define $\bar{1} = (1,\ldots,1) \in \mathbb{R}^d$. Summing over d dimensions we have

$$\begin{aligned}
&\mathbb{E}_{S,\Theta}\left[\left\langle \mathcal{A}(S), \sum_{i=1}^{n}(X_i-\Theta)\right\rangle\right] \\
&\geq \frac{2d}{3} - \frac{1}{2}\left\|\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{-\bar{1}}}\left[\mathcal{A}(S)\right]+\bar{1}\right\|_1 - \frac{1}{2}\left\|\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\bar{1}}}\left[\mathcal{A}(S)\right]-\bar{1}\right\|_1 - 2\mathop{\mathbb{E}}_{\Theta}\left[\left\|\mathop{\mathbb{E}}_{S\sim\mathcal{D}_\Theta}\left[\mathcal{A}(S)\right]-\Theta\right\|_1\right] \\
&\geq \frac{2d}{3} - \frac{1}{2}\|\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{-\bar{1}}}\left[\mathcal{A}(S)\right]+\bar{1}\|_1 - \frac{1}{2}\|\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\bar{1}}}\left[\mathcal{A}(S)\right]-\bar{1}\|_1 - 2\mathop{\mathbb{E}}_{\Theta}\left[\left\|\mathop{\mathbb{E}}_{S\sim\mathcal{D}_\Theta}\left[\mathcal{A}(S)\right]-\Theta\right\|_1\right] \\
&\geq \frac{2d}{3} - \frac{\sqrt{d}}{2}\|\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{-\bar{1}}}\left[\mathcal{A}(S)\right]+\bar{1}\|_2 - \frac{\sqrt{d}}{2}\|\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\bar{1}}}\left[\mathcal{A}(S)\right]-\bar{1}\|_2 - 2\sqrt{d}\mathop{\mathbb{E}}_{\Theta}\left[\left\|\mathop{\mathbb{E}}_{S\sim\mathcal{D}_\Theta}\left[\mathcal{A}(S)\right]-\Theta\right\|_2\right] \\
&\geq \frac{d}{6}.
\end{aligned}$$

This proves the claim. ■

We can now prove the mean estimation lower bound. This proof follows a similar structure to existing proofs for DP mean estimation, although additional work must be done to account for the fact that $\mathcal{A}$ only accesses a subset of points in the dataset.

Proof [Proof of Lemma 26] For our proof we will use a dataset of vectors in $\{\pm 1\}^d$, and as such the ℓ_2 bound on the data is $\sqrt{d}$. The final result will follow from rescaling by $\frac{1}{\sqrt{d}}$. Condition on $\Theta = \theta$ and define the following random variables for each $i \in [n]$,

$$Z_i = \langle \mathcal{A}(S), X_i - \theta \rangle \qquad \text{and} \qquad Z_i' = \langle \mathcal{A}(S_{\sim i}), X_i - \theta \rangle ,$$

where $S_{\sim i}$ is the dataset formed by replacing i-th data point of S with $X_i' \sim \mathcal{D}_\theta$.

Let I denote the random variable corresponding to the subset of indices of data points accessed by $\mathcal{A}$. We have for some $\tau \geq 0$,

$$\begin{aligned}
\mathbb{P}\left[Z_i \geq \tau | i \in I\right] \mathbb{P}\left[i \in I\right] &= \mathbb{P}\left[Z_i \geq \tau\right] - \mathbb{P}\left[Z_i \geq \tau | i \notin I\right] \mathbb{P}\left[i \notin I\right] \\
&\leq e^\epsilon \mathbb{P}\left[Z_i' \geq \tau\right] + \delta \\
&\leq \exp\left(\epsilon - \frac{\tau^2}{8d}\right) + \delta.
\end{aligned}$$

The last inequality uses the Chernoff-Hoeffding bound. Since $\epsilon \leq \log(1/\delta)$, for $\tau = \sqrt{3d\log(1/\delta)}$ we obtain, $\mathbb{P}\left[Z_i \geq \tau | i \in I\right] \leq \frac{2\delta}{\mathbb{P}[i \in I]}$. Using this we can derive,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{A},S}[Z_i] &= \mathbb{E}\left[Z_j | i \in I\right] \mathbb{P}\left[i \in I\right] + \mathbb{E}\left[Z_j | i \notin I\right] \mathbb{P}\left[i \notin I\right] \\
&= \mathbb{E}\left[Z_j | i \in I\right] \mathbb{P}\left[i \in I\right] \\
&\leq \mathbb{P}\left[i \in I\right] \left(\sqrt{3d\log(1/\delta)} + 2d\mathbb{P}\left[Z_j > \tau | i \in I\right]\right) \\
&\leq \mathbb{P}\left[i \in I\right] \sqrt{3d\log(1/\delta)} + 4d\delta.
\end{aligned}$$

Above we use the fact that the expectation of Z_i is 0 when $\mathcal{A}$ does not access the i'th element. Now for the sum we have,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{A},S}\left[\sum_{i=1}^n Z_i\right] &\leq 4de^\epsilon\delta + \sqrt{3d\log(1/\delta)} \sum_{i=1}^n \mathbb{P}\left[i \in I\right] \\
&= 4nd\delta + \sqrt{3d\log(1/\delta)}\mathbb{E}\left[\sum_{i=1}^n \mathbb{1}_{[i \in I]}\right] \\
&= 4nd\delta + \sqrt{3d\log(1/\delta)}\mathbb{E}\left[|I|\right] \\
&= 4nd\delta + s\sqrt{3d\log(1/\delta)} \\
&\leq 3s\sqrt{d\log(1/\delta)}.
\end{aligned}$$

We then obtain the same upper bound for $\mathop{\mathbb{E}}_{\Theta,\mathcal{A},S}[\sum_{i=1}^n Z_i]$. We now use the fingerprinting lemma to lower bound the correlation. Specifically, in the case where $\mathcal{A}$ is at least $\sqrt{d}/6$ accurate (which we

note corresponds to $1/6$ accurate after rescaling), we have for any θ, $\|\mathbb{E}\left[\mathcal{A}(S)\right]-\theta\| = \|\mathbb{E}\left[\mathcal{A}(S) - \frac{1}{n}\sum_{i=1}^n X_i\right]\| \leq \mathbb{E}\left[\|\mathcal{A}(S) - \frac{1}{n}\sum_{i=1}^n X_i\|\right] \leq \frac{\sqrt{d}}{6}$. Thus by Lemma 27,

$$\mathbb{E}\Big[\sum_{i=1}^n Z_i\Big] = \mathbb{E}\Big[\Big\langle \mathcal{A}(S), \sum_{i=1}^n X_i - \Theta\Big\rangle\Big] \geq \frac{d}{6}.$$

Now using the upper and lower bounds on $\mathbb{E}\Big[\sum_{i=1}^n Z_i\Big]$ we obtain,

$$3s\sqrt{d\log(1/\delta)} \geq d/6 \implies s \geq \frac{\sqrt{d}}{18\sqrt{\log(1/\delta)}}.$$

■

D.2. Upper bounds for smooth losses

The lower bound in Theorem 11 is mostly matched by a modification of the Phased SGD algorithm of Feldman et al. (2020).

Theorem 28 *Let $\alpha > 0$, $\delta \in [0,1]$, and $\beta \leq L\sqrt{d\log(1/\delta)}/B$. There exists an algorithm which is $O(\alpha)$-accurate for $(\mathcal{F}^n_{L,\beta}, \mathcal{K}_B)$ and uses at most $O\big(\max\big\{\frac{BL\sqrt{d\log(1/\delta)}}{\alpha}, \frac{B^2L^2}{\alpha^2}\big\}\big)$ oracle evaluations. Further, for $\epsilon \in [0,1]$, if $\alpha \geq 6\alpha^*_{\epsilon,\delta}$ it satisfies (ϵ,δ)-DP.*

We provide a proof subsequently. This nearly matches the lower bound when $\alpha \geq 1/\sqrt{n}$. In the low error regime with β-smooth losses, an alternative algorithm based on solving a series of regularized ERM problems with accelerated ERM solvers can achieve similar results in roughly $O((n + \beta n/\sqrt{d})\log^2(n/\alpha))$ running time. A similar approach for DP-SCO was given in Zhang et al. (2022). We provide details for this result in D.2.1. In aggregate, these upper bounds imply the lower bound is essentially tight. Given that our upper bound does not use a private oracle, one question that arises is whether the lower bound, which holds for general DP algorithms, can still be matched by *private oracle* algorithms. At least in the case of 1-smooth losses and $\alpha = \alpha^*_{\epsilon,\delta}$, the results of Choquette-Choo et al. (2025) show the answer is yes. For $\omega(1)$-smooth losses it is unclear.

Algorithm 4 Phased SGD

Require: Accuracy $\alpha \geq 0$, Oracle $\mathcal{O}$ for losses $\ell_1, ..., \ell_n$, Constraint Set $\mathcal{W}$ of width B, Lipschitz constant L, Privacy parameter $\delta \in [0,1]$

Pick any $w_0 \in \mathcal{W}$

Set $R = \frac{1}{2}\log_2(1/\alpha)$

Set $T = \max\left\{\frac{BL\sqrt{d\log(n/\delta)}}{\alpha}, \frac{B^2L^2}{\alpha^2}\right\}$ and $\eta = \frac{B}{L}\min\left\{\frac{1}{\sqrt{d\log(n/\delta)}}, \frac{\alpha}{BL}\right\}$

for $r = 1...R$ **do**

  Set $T_r = 2^{-r}T$ and $\eta_r = 4^{-r}\eta$

  Run SGD over $\mathcal{W}$ from w_{r-1} for T_r steps with learning rate η_i. Let $\bar{w}_r$ be the average iterate

  $w_r = \bar{w}_r + \xi_r$, with $\xi_r \sim \mathcal{N}(0, \mathbb{I}_d\sigma_r^2)$ and $\sigma_r = \frac{4B}{4^r\sqrt{d}}$

end for

Output: w_R

Proof [Proof of Theorem 28] We first note that the SGD algorithm used as a subroutine in Algorithm 4 starts at some point $w_0 \in \mathcal{W}$, and at each step samples $i \sim \mathsf{Unif}([n])$ and performs the update $w_t = \Pi_{\mathcal{W}}(w_{t-1} - \eta \nabla \ell_i(w))$.

For notation, define $\bar{w}_0 = w^*$ and $\xi_0 = w_0 - w^*$. Using the convergence results of SGD. The error can be decomposed via,

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}(w_R) - \mathcal{L}(w^*)\right] &= \mathbb{E}\left[\mathcal{L}(w_R) - \mathcal{L}(\bar{w}_R)\right] + \sum_{i=1}^{R} \mathbb{E}\left[\mathcal{L}(\bar{w}_r) - \mathcal{L}(\bar{w}_{r-1})\right] \\
&\leq L\mathbb{E}\left[\|\xi_R\|\right] + \sum_{r=1}^{R} \frac{\mathbb{E}\left[\|\xi_{r-1}\|^2\right]}{2\eta_i T_i} + \frac{\eta_i L^2}{2} \\
&\leq \frac{BL}{2^{-2R}} + \sum_{r=1}^{R} 2^{-r}\left(\frac{8B^2}{\eta T} + \frac{\eta L^2}{2}\right) \\
&\leq \alpha + \frac{8B^2}{\eta T} + \frac{\eta L^2}{2} \\
&= O(\alpha).
\end{aligned}$$

The first inequality comes from standard convergence guarantees of projected SGD, see e.g. (Shalev-Shwartz and Ben-David, 2014, Theorem 14.8).

For the privacy analysis, we will leverage privacy amplification via subsampling (without replacement) results. Specifically we will use an extended version of (Bun et al., 2015, Lemma 4.14), restated as Lemma 14 in Appendix A.

Consider the mechanism which, upon receiving $m > 0$ losses, $\ell_1, ..., \ell_m$, performs one-pass SGD over the losses, then adds isotropic Gaussian noise with variance σ_r^2. Assuming each ℓ_i, $i \in [n]$, is at least $1/(2\eta)$-smooth, $\bar{w}_r$ has sensitivity (w.r.t. changing one of $\{\ell_1, \ldots, \ell_m\}$) at most $2L\eta_r$, see e.g. (Hardt et al., 2016, Lemma 3.6). As such, the Gaussian mechanism is $(\epsilon_r, \delta/n)$-DP with respect to a change in one sampled loss function, with $\epsilon_r = \frac{4L\eta_r\sqrt{\log(n/\delta)}}{\sigma_r}$.

Now observe that Algorithm 4, at each phase, applies the previously described mechanism to a batch of T_i losses, sampled with replacement from $\{\ell_1, \ldots, \ell_n\}$. In the regime $\alpha \geq \frac{BL}{\sqrt{d\log(n/\delta)}}$, we have $T = \frac{BL\sqrt{d\log(n/\delta)}}{\alpha}$ and $\eta = \frac{B}{L\sqrt{d\log(n/\delta)}}$, and thus $\forall r \in [R], T_r \leq n/2$ and $\epsilon_r \leq \frac{1}{2^r}$. Alternatively, in the other regime we have $T = \frac{B^2L^2}{\alpha^2}$ and $\eta = \frac{\alpha}{L^2}$, and thus $\epsilon_r \leq \frac{1}{2^r}\frac{\alpha\sqrt{d}}{BL}$. Observe $\frac{1}{2^r}\frac{\alpha\sqrt{d}}{BL} \leq \frac{n}{2T_i}$ for any $\alpha \geq \alpha^*_{1,\delta}$. In either case, $\epsilon_r \leq \frac{n}{2T_i}$, and thus we can apply the amplification via subsampling result from Lemma 14. Specifically, this implies that each round of the algorithm is (ϵ'_r, δ'_r) with,

$$\begin{aligned}
\epsilon'_r &= \frac{6T_i}{n}\epsilon_r \leq \frac{6}{R}\max\left\{\frac{BL\sqrt{d\log(n/\delta)}}{n\alpha}, \frac{B^2L^2}{\alpha^2 n}\cdot\frac{\alpha\sqrt{d\log(n/\delta)}}{BL}\right\} = 6\frac{\alpha^*_\delta}{2^r\alpha}, \\
\delta'_r &= e^{6\epsilon' T_i/n}\frac{4T_i}{n}\frac{\delta}{n} \leq e^6 2^{-r}\delta.
\end{aligned}$$

By composition, the overall privacy of the algorithm satisfies (ϵ, δ)-DP with $\epsilon \leq 6\sum_{r=1}^{R} 2^{-r}\frac{\alpha^*_{1,\delta}}{\alpha} \leq 6\frac{\alpha^*_{1,\delta}}{\alpha}$. ■

Remark 29 *The Phased SGD algorithm also shows why one must assume the queries sent to the private proxy oracle are non-adaptive for our lower bound in the non-smooth case to be hold. An inspection of the privacy analysis in the proof of Theorem 28 shows that, if one only cares that the algorithm is private with respect to its dataset of gradients, smoothness is not needed. We emphasize that being private with respect to the dataset of gradients does not imply a general DP optimizer however, and indeed Phased SGD is not known to be DP is non-smooth case.*

D.2.1. $\tilde{O}(n)$ RUNNING TIME ALGORITHM WHEN $\alpha = O(\frac{BL}{\sqrt{n}})$.

To achieve near linear running time, one case use the Phased-ERM algorithm of Feldman et al. (2020) (Algorithm 3 therein) in conjunction with accelerated ERM solvers. This was essentially shown by Zhang et al. (2022), although because they studied DP-SCO, they only explicitly stated results for error $\alpha \geq \frac{BL}{\sqrt{n}}$. Nonetheless, their technique translates just as well for smaller error when considering DP-ERM. We describe this in the following.

Given some (non-private) ERM solver, $\mathcal{A}$, which solves a strongly convex ERM problem to high accuracy, the Phased-ERM algorithm, Algorithm 5, is differentially private and yields an accurate solution. Precisely, we have the following result.

Lemma 30 *Let $\delta \in [0, B\alpha^2]$ and $\epsilon \in [0, 1]$. Algorithm 5 is $O(\alpha)$-accurate for $(\mathcal{F}^n_{L,\infty}, \mathcal{K}_B)$ and for $\alpha \geq \log(n)\alpha^*_{\epsilon,\delta}$ it satisfies $(\epsilon, 2\log(n)\delta)$-DP.*

The proof follows similarly to the one in Feldman et al. (2020), and is given below. When the loss is additionally β-smooth, there are solvers for the regularized subproblem (such as SVRG, Johnson and Zhang (2013)) which achieve the accuracy condition in $O((n+B^2\beta/\alpha)\log(n/\alpha))$ oracle calls. For $\alpha \geq \alpha^*_{1,\delta}$, we get near linear running time if $\beta \leq L\sqrt{d}/B\}$.

Algorithm 5 Phased ERM

Require: Accuracy $\alpha \geq 0$, Oracle $\mathcal{O}$ for losses $\ell_1, ..., \ell_n$, Constraint Set $\mathcal{W}$ of width B, Lipschitz constant L, Parameter $\delta \in [0, 1]$

1: Pick any $w_0 \in \mathcal{W}$
2: Set $R = \log_2(LB/\alpha)$
3: Set $\lambda_r = 2^r \frac{\alpha}{B^2}$ for all $r \in [R]$
4: **for** $r = 1...R$ **do**
5: Define $\mathcal{L}_r(w) = \frac{1}{n}\sum_{i=1}^n \ell_i(w) + \lambda_r\|w - w_{r-1}\|^2$ and $w^*_r = \arg\min_{w\in\mathcal{W}}\{\mathcal{L}_r(w)\}$
6: Compute $\bar{w}_r$ such that w.p. at least $1-\delta$, $\mathcal{L}_r(w) - \mathcal{L}_r(w^*_r) \leq \min\left\{\frac{L^2}{\lambda_r n^2}, 2^{-r}\alpha\right\}$
7: $w_r = \bar{w}_r + \xi_r$ where $\xi_r \sim \mathcal{N}(0, \mathbb{I}_d\sigma_r^2)$ and $\sigma_r = \frac{4B}{2^r\sqrt{d}}$
8: **end for**
9: **Output:** w_R

Proof [Proof of Lemma 30] For notation, let $\bar{w}_0 = w^*$. We have,

$$\mathbb{E}\left[\mathcal{L}(w_R) - \mathcal{L}(w^*)\right] = \mathbb{E}\left[\mathcal{L}(w_R) - \mathcal{L}(\bar{w}_R)\right] + \sum_{i=1}^R \mathbb{E}\left[\mathcal{L}(\bar{w}_r) - \mathcal{L}(\bar{w}_{r-1})\right]$$

$$\leq 4\alpha + \sum_{i=1}^R \mathbb{E}\left[\mathcal{L}_r(\bar{w}_r) - \mathcal{L}_r(\bar{w}_{r-1})\right].$$

The inequality uses $\sigma_R = \frac{4\alpha}{L\sqrt{d}}$ to bound $L\mathbb{E}\left[\|\xi_R\|\right]$. We have for any $r \in [R]$,

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\bar{w}_r) - \mathcal{L}(\bar{w}_{r-1})\right] &= \mathbb{E}\left[\mathcal{L}_r(\bar{w}_r) - \mathcal{L}_r(\bar{w}_{r-1}) + \mathcal{L}(\bar{w}_r) - \mathcal{L}_r(\bar{w}_r) + \mathcal{L}_r(\bar{w}_{r-1}) - \mathcal{L}(\bar{w}_{r-1})\right] \\ &\le (2^{-r}\alpha + \delta L) + \lambda_r \mathbb{E}\left[\|\bar{w}_{r-1} - w_{r-1}\|^2\right] \\ &\le 2^{-r}\alpha + \delta L + 2^r \frac{\alpha}{B^2}\left(\frac{16B^2}{2^{2r}}\right).\end{aligned}$$

Thus we have $\mathbb{E}\left[\mathcal{L}(\bar{w}_r) - \mathcal{L}(\bar{w}_{r-1})\right] \le 2^{-r}18\alpha$ provided $\delta \le B\alpha^2$, and the accuracy guarantee follows by combining the both displays.

For the privacy analysis, consider some $r \in [R]$. We have by standard results on the stability of regularized ERM that each w_r^* is $\frac{L}{\lambda n}$-stable (i.e. changing one loss in $\mathcal{L}$ perturbs w_r^* by at most $\frac{L}{\lambda n}$) Bousquet and Elisseeff (2002). The λ_r strong-convexity of $\mathcal{L}_r$ and accuracy condition also implies that, conditioning on the randomness in previous rounds, with probability at least $1-\delta$, $\|\bar{w} - w_r^*\| \le \sqrt{\frac{\mathcal{L}_r(w) - \mathcal{L}_r(w_r^*)}{\lambda_r}} \le \frac{L}{\lambda_r n}$. So each $\bar{w}_r$ is Δ-stable, with $\Delta = \frac{2L}{\lambda_r n} = \frac{2B^2}{2^r \alpha n} \le \frac{2B\epsilon}{2^r\sqrt{\log(n) d \log(1/\delta)}}$, where the inequality follows from $\alpha \ge \log(n)\alpha^*_{\epsilon,\delta}$. Thus the Gaussian mechanism ensures each round r is $(\epsilon/\log(n), 2\delta)$-DP and by composition the overall algorithm is $(\epsilon, 2\log(n)\delta)$-DP. ■

Appendix E. Non-smooth optimization with information limited oracles

Our proof techniques from Section 3 can easily be adapted to provide lower bounds for information limited oracles. In this section, we consider an individual loss $\mathcal{L}$, i.e. $n = 1$, such that the objective is to approximate $\arg\min_{w \in \mathcal{W}} \{\mathcal{L}(w)\}$. We are interested in algorithms of the form of Algorithm 1 (ignoring the query indices) interacting with a proxy oracle of bounded information capacity.

Theorem 31 *(Restatement of Theorem 12) Let $\tilde{\mathcal{O}}$ be a Γ-information limited proxy oracle and $\mathcal{A}$ an algorithm of the form given by Algorithm 1 with $\mathbb{E}\left[\bar{T}\right] \le \frac{d}{640\log(BL/\alpha)}$. If $d \ge \frac{C_2 B^2 L^2}{\alpha^2}$ then $\overline{\mathsf{Time}}(\mathcal{A}, \tilde{\mathcal{O}}, \alpha, B, L, \infty) = \Omega\left(\frac{B^2L^2d}{\alpha^2\Gamma}\right)$.*

We give the proof after the following discussion. We can drop the unique query limit and more simply lower bound the runtime as $\Omega\left(\min\{\frac{d}{\alpha^2\Gamma}, \frac{d}{\log(1/\alpha)}\}\right)$. Note also that it is possible to have $\Gamma = \omega(d)$, which can be reasonable when the batch sizes are $\omega(1)$. The proof of this result follows from a simplified version of the proof of Theorem 5.

As a corollary of our result, consider the case where we fix the batch size in Algorithm 1 to be 1 and the proxy oracle is instantiated to be $\tilde{\mathcal{O}}(w) = \nabla\mathcal{L}(w) + \mathcal{N}(0, \mathbb{I}_d\frac{\sigma^2}{d})$. A standard fact on Gaussian channels implies that for 1-Lipschitz losses and $\sigma \ge 1$ this oracle has information capacity $\Gamma \le d/\sigma^2$. Theorem 31 thus recovers the oracle complexity lower bound for stochastic oracles, $\Omega\left(\frac{\sigma^2}{\alpha^2}\right)$ (Nemirovsky and Yudin, 1985), at least for certain parameter regimes. This $\frac{\sigma^2}{\alpha^2}$ lower bound is achieved by SGD, which incidentally also means our lower bound is tight when $\Gamma \ge 1/\alpha^2$. That said, it is perhaps more interesting to consider whether the bound can be matched via a quantization scheme; (Mayekar and Tyagi, 2020a,b) shows this is the case up to log factors.

Despite recovering stochastic and quantized stochastic oracle complexity lower bounds, we emphasize that our lower bound also meaningfully diverges from such results. Such lower bounds have been obtained via a reduction to mean estimation, where each oracle response is a noisy version

of this mean, e.g. Nemirovsky and Yudin (1985); Agarwal et al. (2012)), and in the stochastic quantized oracle setting, strong data processing techniques are used to get better bounds (Mayekar and Tyagi, 2020a; Acharya et al., 2021). Clearly, we cannot hope to obtain Theorem 31 from such constructions, as transmitting the mean vector to α accuracy requires only $O(d\log(1/\alpha))$ bits of information. Put another way, the difficulty in previous lower bound constructions largely stems from the difficulty of mean estimation. Our bound relaxes the stochastic oracle assumption by leveraging structure unique to solving optimization problems. For similar reasons, it is not a-priori obvious that allowing the proxy oracle use a batch size larger than 1, and thus transmit messages about multiple gradients, would not help improve oracle complexity. Our lower bound shows this is indeed the case.

Proof of Theorem 31. The proof leverages the same loss construction and distribution as Theorem 5. In particular, we will apply Lemma 16 verbatim. Upper bounding the information is in fact simpler than in Lemma 15.

Lemma 32 *Under the conditions of Theorem 31, for any $k \in [K]$ it hold that*

$$I(W; X_k | X_{\neq k}, V) \leq \mathbb{E}\Big[\sum_{t=1}^{\infty} \mathbb{1}_{[\mathsf{Cnt}_j(Q_t)\geq 1]}\Big]\Gamma + \mathbb{E}\left[\bar{T}\right]\log(K+1). \tag{20}$$

Proof As in the proof of Lemma 15, we condition on $V = v$ and $X_{\neq k} = x_{\neq k}$ throughout and recall the definitions of Q, Cnt, and $P_t = (Y_{<t}, \hat{R}_{\leq t})$. Using the same derivation as Eqn. (5) we obtain,

$$I(Y, R; X_k) \leq \mathbb{E}\left[\bar{T}\right]\log(K+1) + \sum_{t=1}^{T} I(Y_t; X_k | Q_{\leq t}, P_t). \tag{21}$$

Further by the assumption on the oracle, when $Q_t = q_t$,

$$I(Y_t; X_k | Q_t = q_t, P_t) \leq \Gamma \cdot \mathbb{1}_{[\mathsf{Cnt}_j(Q_t)\geq 1]} \implies I(Y_t; X_k | Q_t, P_t) \leq \Gamma \cdot \mathbb{E}\left[\mathbb{1}_{[\mathsf{Cnt}_j(Q_t)\geq 1]}\right].$$

Recalling we have conditioned on $V = v$, we take expectation and plug into Eqn. (21) to obtain the claim. ■

Proof [Theorem 31] Under the assumption that $\mathbb{E}\left[\bar{T}\right] \leq \frac{d}{320\log(1/\alpha^2)}$, applying Lemmas 16 and 32 obtains,

$$\begin{aligned}
\Omega(dK) \leq \sum_{j=1}^{K} I(Y, R; X_j | X_{\neq j} V) &\leq \mathbb{E}\left[\sum_{t=1}^{\infty}\sum_{j=1}^{K} \mathbb{1}_{[\mathsf{Cnt}_j(Q_t)\geq 1]}\right]\Gamma + K\mathbb{E}\left[\bar{T}\right]\log(K+1) \\
&\leq \mathbb{E}\left[\sum_{t=1}^{\infty}\sum_{j=1}^{K} M_t\right]\Gamma + K\mathbb{E}\left[\bar{T}\right]\log(K+1) \\
&= \mathbb{E}\left[\|M\|_1\right]\Gamma + K\mathbb{E}\left[\bar{T}\right]\log(K+1).
\end{aligned}$$

We can then finish similarly to Theorem 5. ■

Appendix F. Between DP-SCO to DP-ERM

In this appendix, we show via a reduction that (DP)-ERM and (DP)-SCO are equally hard in terms of runtime, up to log factors. This implies similar reductions for non-private settings as well.

F.1. Reducing DP-SCO to DP-ERM

The following shows that DP-SCO is no harder than DP-ERM (up to log factors) via a reduction. Specifically, given an algorithm, $\mathcal{A}$, which solves DP-ERM for $B = L = 1$, we show how to construct a DP-SCO algorithm which has similar running time, using only black box access to $\mathcal{A}$.

Theorem 33 *Let $n, B, L \geq 0$, $\beta \in \mathbb{R}^+ \cup \{\infty\}$, $\beta' = \beta + \frac{L}{B}$, and $n' = \frac{n}{\log(n)}$. Let $\mathcal{A}$ be an (ϵ, δ)-DP algorithm which is α-accurate for $(\mathcal{F}^{n'}_{5L,\beta'}, \mathcal{K}_B)$ for DP-ERM and has expected running time T. Then there exists an algorithm which, using black box access to $\mathcal{A}$, is $(\tilde{O}(\epsilon), \tilde{O}(\delta))$-DP and $\tilde{O}(\alpha + \frac{BL}{\sqrt{n}} + \frac{BL}{n\epsilon})$-accurate for $(\mathcal{F}^n_{L,\beta}, \mathcal{K}_B)$ for DP-SCO, and has expected running time $\tilde{O}(T + \frac{B^2L^2}{\alpha^2})$.*

Note that any general DP algorithm for minimizing the population risk of non-smooth losses must incur error at least $\frac{BL}{\sqrt{n}} + \frac{BL}{n\epsilon}$ and runtime at least $\frac{B^2L^2}{\alpha^2}$. Thus these additive factors are no worse than what one would obtain with a "direct" algorithm for DP-SCO. In the smooth case, the running time lower bound for finite sums is $\min\{\frac{B^2L^2}{\alpha^2}, n\}$, and since $\alpha \geq \frac{BL}{\sqrt{n}}$, we again see that their is no asymptotic loss in runtime.

We now show how to obtain the theorem using the following result from Bassily et al. (2023). We borrow parameter definitions from the above theorem statement. As an aside, we note that statements similar to Theorem 33 in different geometries are likely obtainable using a generalization of this statement provided in Bassily et al. (2024).

Theorem 34 *(Bassily et al., 2023, Theorem 1) Let $\mathcal{A}$ be an algorithm which, given $D \in [B\sqrt{\frac{\log(n)}{n}}, B]$ and randomly generated point $w' \in \mathcal{W}$, satisfies $\mathop{\mathbb{E}}_{w',\mathcal{A}}[\mathcal{L}(\mathcal{A}(\mathcal{O}_\mathcal{L})) - \mathcal{L}(w^*)] \leq \hat{\alpha}D$ whenever $\mathop{\mathbb{E}}_{w'}[\|w' - w^*\|] \leq D$ and any $\mathcal{L} \in \mathcal{F}^{n'}_{5L,\beta'}$.Then there exists an algorithm, which interacts with $\mathcal{L}$ through at most $\log(n)$ calls to $\mathcal{A}$ and is α-accurate for $(\mathcal{F}^n_{L,\beta}, \mathcal{K}_B)$ for SCO with $\alpha = O\big(\log(n)B\hat{\alpha} + \frac{\log^{3/2}(n)BL}{\sqrt{n}}\big)$.*

The original statement in Bassily et al. (2023) assumes the accuracy condition holds for all $D > 0$, but an inspection of their proof shows that the relative accuracy condition is only used in their Eqn. 12 and for $D \in [B\sqrt{\frac{\log(n)}{n}}, B]$. Further, Bassily et al. (2023) studied the more general case of saddle point problems, but ERM can be recovered by assuming range of the dual parameter is a singleton. Finally, we note that their algorithm only requires running the subroutine $\mathcal{A}$ on regularized version of the loss, which, under their level of regularization, increases the smoothness parameter of the loss by at most $\frac{L}{B}$.

It has essentially already been shown in (Arora et al., 2022, Section 5) how to obtain a DP algorithm satisfying the accuracy condition of Theorem 34 using black box access to a DP constrained optimizer, although their setting differs slightly. We provide a self contained version of their argument below. We will also make use of Fact 1 in Appendix A several times.

Proof [Proof of Theorem 33] In the following, let T denote the expected running time of an α-accurate DP-ERM algorithm, $\mathcal{A}$, in the case where $B = L = 1$.

We fist boost the expected empirical risk guarantee of $\mathcal{A}$ into a high probability guarantee. Let $u_0, \ldots, u_K$ be the result of $K = \log(n)$ independent runs of $\mathcal{A}$ on S. By Markov's inequality, at least one of these runs achieves excess risk 2α with probability at least $1 - \frac{1}{2^K} = 1 - \frac{1}{n}$. For each run $j \in [K]$, we generate a loss estimate, E_j, by sampling (without replacement) a minibatch of $1/\alpha^2$ losses from $\mathcal{L}$ and computing the average loss on u_j. Since the range of the losses is 1-bounded, we have by Chernoff-Hoeffding that, $\mathbb{P}[E_j - \mathcal{L}(u_j) \geq \sqrt{\frac{8\log(n)}{n}}] \leq \frac{1}{n^2}$. We then apply the exponential mechanism with privacy parameter ϵ over the scores $E_1, \ldots, E_K$ to select the solution candidate from $u_1, \ldots, u_K$. The guarantees of the exponential mechanism ensures that with probability at least $1 - 1/n$ the selected solution has loss within $\frac{4\log(n)}{n\epsilon}$ of the minimal loss candidate. Thus we obtain an accurate solution with probability at least $1 - O(1/n)$ via a procedure that is via an algorithm that is $((\log(n) + 1)\epsilon, \log(n)\delta)$-DP. The expected running time of this procedure is $\log(n)T + \frac{\log(n)}{\alpha^2}$.

Applying Fact 1, we can assume access to an algorithm $\tilde{\mathcal{A}}$, which with probability at least $1 - O(1/n)$ achieves accuracy $BL\tilde{\alpha}$ where $\tilde{\alpha} = \big(\alpha + \sqrt{\frac{8\log(n)}{n}} + \frac{4\log(n)}{n\epsilon}\big)$ on problems which are L-Lipschitz and have constraint set of radius at most B.

We now describe how to use $\tilde{\mathcal{A}}$ to obtain relative accuracy. Letting $R = \frac{1}{2}\log(n)$, we run $\mathcal{A}$ on $\mathcal{W}_0, ..., \mathcal{W}_R$, to obtain candidate solutions $w_0, ..., w_R$, where $\mathcal{W}_r = \mathcal{W} \cap \{w : \|w - w'\| \leq 2^{-r}B\}$. Observe $w^* \in \mathcal{W}_r$ for any $r \leq \frac{\log(1/\|w' - w^*\|)}{\log(2B)}$. We then pick the best candidate using the same loss estimate/exponential mechanism procedure used in the boosting argument. It is then easy to see that the solution selected by the exponential mechanism achieves excess empirical risk that is $O\Big(\|w' - w^*\|L\tilde{\alpha} + BL\big(\sqrt{\frac{\log(n)}{n}} + \frac{\log(n)}{n\epsilon}\big)\Big)$. Converting the high probability guarantee to expectation we obtain excess empirical risk $O\Big(\|w' - w^*\|L\tilde{\alpha} + BL\big(\sqrt{\frac{\log(n)}{n}} + \frac{\log(n)}{n\epsilon} + \frac{1}{n}\big)\Big)$. By taking expectation w.r.t. w' we see the condition of the theorem is satisfied with $\hat{\alpha} = O(L\tilde{\alpha})$. Theorem 33 then follows by applying Theorem 34 to the previously described algorithm. ■

F.2. Reducing DP-ERM ot DP-SCO

The reverse direction was given by (Bassily et al., 2019, Appendix C), and we here note that this direction can be performed without loss of log factors via a slightly different analysis. Specifically, for $\epsilon \leq 1/6$ and $\delta \in [0, 1]$, given an (ϵ, δ)-DP-SCO solver $\mathcal{A}_{SCO}$, we first sample n points from the empirical distribution over $\{\ell_1, \ldots, \ell_n\}$, then runs the DP-SCO solver on the resampled dataset. By results for privacy amplification via subsampling with replacement (see Lemma 14 in Appendix A), the result is $(6\epsilon, e\delta)$-DP. The bound on excess empirical risk follows directly from the accuracy guarantees of $\mathcal{A}_{SCO}$, since it is run on i.i.d. samples from the empirical distribution.