Optimal Differentially Private Model Training with Public Data

Andrew Lowy [1] **Zeman Li** [2] **Tianjian Huang** [2] **Meisam Razaviyayn** [2]

Abstract

Differential privacy (DP) ensures that training a machine learning model does not leak private data. In practice, we may have access to auxiliary public data that is free of privacy concerns. In this work, we assume access to a given amount of public data and settle the following fundamental open questions: *1. What is the optimal (worst-case) error of a DP model trained over a private data set while having access to side public data? 2. How can we harness public data to improve DP model training in practice?* We consider these questions in both the local and central models of pure and approximate DP. To answer the first question, we prove tight (up to log factors) lower and upper bounds that characterize the optimal error rates of three fundamental problems: mean estimation, empirical risk minimization, and stochastic convex optimization. We show that the optimal error rates can be attained (up to log factors) by either discarding private data and training a public model, or treating public data like it is private and using an optimal DP algorithm. To address the second question, we develop novel algorithms that are "even more optimal" (i.e. better constants) than the asymptotically optimal approaches described above. For local DP mean estimation, our algorithm is optimal including constants. Empirically, our algorithms show benefits over the state-of-the-art.

1. Introduction

Training machine learning models on people's data can leak sensitive information, violating their privacy (Fredrikson et al., 2015; Shokri et al., 2017; Carlini et al., 2021). *Differential Privacy (DP)* prevents such leaks by providing a rigorous guarantee that no attacker can learn too much about any individual's data (Dwork et al., 2006). DP has been successfully deployed by various companies (Apple, 2016; Thakurta et al., 2017; Úlfar Erlingsson et al., 2014; Ding et al., 2017), and by government agencies (U.S. Census Bureau, 2020). However, a major hindrance to more widespread adoption of DP is that DP-trained models are less accurate than their non-private counterparts.

Leveraging *public data*—that is free of privacy concerns—appears to be a promising and practically important avenue for closing the accuracy gap between DP and non-private models (Papernot et al., 2017; Avent et al., 2017; Feldman et al., 2018; Amid et al., 2022). For example, large language models (LLMs) are often pre-trained on public data and fine-tuned on private data (Kerrigan et al., 2020b; Li et al., 2021b; Yu et al., 2021a). Public data may be provided by people who volunteer (e.g. product developers or early testers) (Church, 2005; Feldman et al., 2018) or sell their data. Data that is generated synthetically (Torkzadehmahani et al., 2019; Vietri et al., 2020; Boedihardjo et al., 2022; He et al., 2023) or released through a legal process (Klimt & Yang, 2004) may serve as additional sources of public data.

The power and limitations of public data vary depending on the particular learning problem and loss function/hypothesis class. To calibrate the effectiveness of a public-data-assisted DP algorithm, we can compare it against two *naïve baselines*: 1) "throw away" the private data and run an optimal non-private algorithm on the public data; 2) use an optimal DP algorithm on the full data set, treating the public data like it is private data. Some works have identified problems where significant improvements over the naïve approaches are possible. For example, (Alon et al., 2019) show that for agnostic PAC learning with a hypothesis class of finite VC-dimension, it is possible to achieve asymptotically smaller sample complexity than the naïve baselines. (Bassily et al., 2020) show a similar result for private query release with finite VC-dimension. On the other hand, for certain problems it is impossible to do better than the naïve approaches: e.g., releasing binary decision stumps (Bassily et al., 2020). Understanding what improvements, if any, over the naïve approaches are possible for other problems (e.g. optimization) and function classes is interesting.

This work considers DP *model training* with side access to

[1]University of Wisconsin-Madison, Wisconsin Institute of Discovery, Madison, WI, USA [2]Department of Industrial & Systems Engineering, University of Southern California, Los Angeles, CA, USA. Correspondence to: Andrew Lowy <alowy@wisc.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

public data: given a loss function, find model parameters to approximately minimize the expected training or test loss. We consider *empirical risk minimization* (ERM) and *stochastic convex optimization* (SCO), which correspond to minimizing training loss and expected test loss, respectively. For population mean estimation and SCO, we assume access to in-distribution public data. For ERM, the public data may be out-of-distribution. We answer a fundamental question:

Question 1. What is the optimal (minimax) error of DP model training with side access to public data? Is it possible to achieve smaller error than the naïve baselines?

Contribution 1: Limitations of Public Data for DP Model Training To answer **Question 1**, we characterize the optimal minimax error (up to constants or logarithms) of *semi-DP* (Beimel et al., 2013; Alon et al., 2019) algorithms—algorithms that are DP w.r.t. private data, but not necessarily DP w.r.t. public data (Definition 3). *We provide tight lower and upper bounds for three fundamental training problems*: mean estimation of bounded random variables[1], ERM and SCO with Lipschitz convex functions.[2] We consider both the *local* (Kasiviswanathan et al., 2011; Duchi et al., 2013) and *central* (Dwork et al., 2006) models of *pure* ($\delta = 0$) and *approximate* ($\delta > 0$) semi-DP. We prove nine sets of lower and upper bounds: see Figs. 1, 2 and 7. Our lower bounds imply that *it is impossible to obtain asymptotic improvements over the naïve approaches for semi-DP model training in the worst case.*

In light of these negative results, it is natural to wonder whether/how one can harness public data for more effective DP model training. This leads us to **Question 2**:

Question 2. Can we provide improved performance (e.g. theoretically smaller error and/or superior empirical results) over the naïve baselines?

Some prior works have tackled **Question 2** by imposing additional assumptions and/or shrinking the problem class. For example, (Amid et al., 2022) shows that under certain distributional assumptions, public data permits benefits over DP-SGD in linear regression. Also, (Zhou et al., 2020; Kairouz et al., 2021) show that by imposing certain "low-rank subspace" assumptions on the gradients in DP-SGD, public data can help attain dimension-independent rates for DP ERM. In contrast, we consider **Question 2** *without imposing any additional assumptions and without shrinking the loss function/data distribution class.*

Contribution 2: Power of Public Data for DP Model Training To address **Question 2**, we develop novel (central and local) semi-DP algorithms that add less noise than would be necessary to privatize the full data set (including public). By doing so, we can achieve *optimal (worst-case) error bounds with significantly improved constants* over the asymptotically optimal naïve algorithms. Our local semi-DP mean estimation algorithm is optimal including constants.

We complement our theoretical analyses with extensive numerical experiments. Our experiments show that *our algorithms outperform the naïve approaches, even when the optimal DP algorithm is pre-trained on the public data*. For example, our Algorithm 1 achieves a significant improvement in CIFAR-10 image classification tasks, reducing test error by $8-9\%$ for logistic regression and by at most 18.9% for Wide-ResNet, compared with the naïve approaches. We also identify a linear regression problem in which *DP-SGD diverges, but our algorithm converges* with small error using $n_{\text{pub}} = 0.1n$ public samples: see Figure 6. Moreover, *our algorithm consistently outperforms the state-of-the-art* public-data-assisted mirror-descent (PDA-MD) of (Amid et al., 2022).

1.1. Techniques and Challenges

Lower bounds: We develop and utilize a variety of techniques to prove our lower bounds.

To prove our central (ε, δ)-semi-DP SCO lower bound in Theorem 12, we build upon the techniques of Dwork et al. (2015); Bassily et al. (2020). A key challenge is finding a distribution whose mean has small norm and proving that this distribution is still hard enough for any semi-DP algorithm to estimate in ℓ_2-distance. To accomplish this, we make three main innovations: First, we modify the *tracing attack* of (Dwork et al., 2015) by incorporating more aggressive truncation; this is used to infer membership of many individuals in the data set, even under weak ℓ_2-accuracy guarantees. Second, we construct a novel distribution in which the prior (mean) is drawn from a shorter interval than the posterior (data). This innovation is crucial for obtaining our tight bounds, but also complicates the analysis. Thus, we provide a novel generalization of the *fingerprinting lemma* (Bun et al., 2014; 2017; Kamath et al., 2022b) that permits analysis of our construction. We provide more details on these proofs in Appendix E.

For our central pure ε-semi-DP population lower bounds in Theorems 32 and 42, we develop a *Semi-DP Fano's method* (Theorem 33). In combination with the reduction from estimation to testing, Theorem 33 generalizes DP Fano's method (Acharya et al., 2021) and strengthens classical Fano's method (Yu, 1997). We build on the tools of Barber & Duchi (2014) in our proof of Theorem 33.

Our ERM lower bound (Theorem 10) uses a novel semi-DP *packing argument*: we construct $2^{d/2}$ "hard" data sets with well-separated sample means, and use the group privacy

[1] Our ε-semi-DP analysis extends to unbounded/heavy-tailed distributions with bounded k-th order moment.

[2] Our results for ERM also cover non-convex loss functions.

Learning problem	Semi-DP error	When is semi-DP error less than DP error?
Mean Estimation (Pop. MSE)	$\min\left\{\frac{1}{n_{pub}}, \frac{d}{n^2\varepsilon^2} + \frac{1}{n}\right\}$ (Theorem 4)	$n_{\text{pub}} > \frac{n^2\varepsilon^2}{d}$ or $n_{\text{pub}} = \Theta(n)$
SCO (Excess pop. risk)	$\min\left\{\frac{1}{\sqrt{n_{pub}}}, \frac{\sqrt{d}}{n\varepsilon} + \frac{1}{\sqrt{n}}\right\}$ (Theorem 12)	$n_{\text{pub}} > \frac{n^2\varepsilon^2}{d}$ or $n_{\text{pub}} = \Theta(n)$

Learning problem	Semi-LDP error	When is semi-LDP error less than LDP?
Mean Estimation (Pop. MSE)	$\min\left\{\frac{1}{n_{pub}}, \frac{d}{n\varepsilon^2}\right\}$ (Theorem 14 & Remark 15)	$n_{\text{pub}} > \frac{\varepsilon^2 n}{d}$ or $n_{\text{pub}} = \Theta(n)$
SCO (Excess pop. risk)	$\min\left\{\frac{1}{\sqrt{n_{pub}}}, \sqrt{\frac{d}{n\varepsilon^2}}\right\}$ (Theorem 18 & Remark 15)	$n_{\text{pub}} > \frac{\varepsilon^2 n}{d}$ or $n_{\text{pub}} = \Theta(n)$

Figure 1. Minimax optimal error rates for central (ε, δ)-semi-DP (up to logs) and (local) (ε, δ)-semi-LDP. $n = n_{\text{priv}} + n_{\text{pub}}$, where n_{priv} (n_{pub}) denotes the number of private (public) samples. Dependence on δ, range and Lipschitz parameters, constraint set diameter omitted. See Appendix for strongly convex SCO results.

Learning problem	Semi-DP error	When is semi-DP error less than DP? error?
ERM (Excess emp. Risk)	$\min\left\{\frac{n_{priv}}{n}, \frac{d}{n\varepsilon}\right\}$ (Theorem 10)	$n_{\text{pub}} > n - d/\varepsilon$

Figure 2. Minimax optimal error rates for ε-semi-DP ERM. See Table 7 in Appendix for more ε-semi-(L)DP results (e.g. SCO).

property of DP to show that any semi-DP algorithm must make large error on at least one of these data sets. Such arguments have long been used to prove pure DP lower bounds (Hardt & Talwar, 2010; Bassily et al., 2014). However, to the best of our knowledge, our proof is the first to extend packing arguments to the semi-DP setting. The main challenge is in carefully constructing the private and public data sets so as to force semi-DP $\mathcal{A}$ to make large error, even when $\mathcal{A}$ has access to public data.

For our local semi-DP lower bounds (Theorems 14 and 18), we build on the sophisticated techniques of Duchi & Rogers (2019). The main idea of our proofs is to combine *Assouad's method* (Duchi, 2021) with bounds on the mutual information between the input and output of semi-DP algorithms.

Algorithms: We develop novel algorithms to obtain smaller error (including constants) than the asymptotically optimal naïve algorithms. Our algorithms are simple. For example, we propose a central (ε, δ)-semi-DP estimator that puts different weights on the private and public samples and adds (Gaussian) noise that is calibrated to the private weight. We choose the weights depending on the privacy level, dimension, and number of public samples to trade off smaller sensitivity (less noise) with larger variance on the public data. An optimal choice of weights minimizes the variance of our unbiased estimator, leading to smaller error than the optimal DP estimator. Our local semi-DP algorithm simply applies the optimal local randomizer of Bhowmick et al. (2018) *only* to the private samples, and averages the noisy private data with the raw public data. For SCO, we develop semi-DP stochastic gradient methods that use our mean estimation algorithms to estimate the gradient of the loss function in each iteration of training.

1.2. Preliminaries and Notation

Let $\|\cdot\|$ be the ℓ_2 norm and $\mathcal{X}$ denote a data universe. Function $g : \mathcal{W} \to \mathbb{R}$ is *μ-strongly convex* if $g(\alpha w + (1-\alpha)w') \leqslant \alpha g(w) + (1-\alpha)g(w') - \frac{\alpha(1-\alpha)\mu}{2}\|w - w'\|^2$ for all $\alpha \in [0,1]$ and all $w, w' \in \mathcal{W}$. If $\mu = 0$, we say g is *convex*. Function $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ is *uniformly L-Lipschitz* in w if $\sup_{x\in\mathcal{X}} |f(w,x) - f(w',x)| \leqslant L\|w - w'\|$. $\mathbb{B} = \{x \in \mathbb{R}^d | \|x\| \leqslant 1\}$ denotes the unit ℓ_2-ball in $\mathbb{R}^d$.

Definition 1 (Differential Privacy (Dwork et al., 2006))**.** *Let $\varepsilon \geqslant 0$, $\delta \in [0,1)$. Randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$ is (ε, δ)-differentially private (DP) if for all $X, X' \in \mathcal{X}^n$ differing in one sample and all measurable subsets $S \subseteq \mathcal{W}$, we have $\mathbb{P}(\mathcal{A}(X) \in S) \leqslant e^\varepsilon \mathbb{P}(\mathcal{A}(X') \in S) + \delta$.*

Definition 1 prevents attackers from learning much more about an individual's data than if the data had not been used for training. If $\delta = 0$, we write ε-DP and say "pure" DP.

Definition 2 (Zero-Concentrated Differential Privacy (zCDP) (Bun & Steinke, 2016))**.** *A randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$ satisfies ρ-zero-concentrated differential privacy (ρ-zCDP) if for all pairs of adjacent data sets $X, X' \in \mathcal{X}^n$ and all $\alpha \in (1, \infty)$, we have $D_\alpha(\mathcal{A}(X)||\mathcal{A}(X')) \leqslant \rho\alpha$, where $D_\alpha(\mathcal{A}(X)||\mathcal{A}(X'))$ is the α-Rényi divergence[3] between the distributions of $\mathcal{A}(X)$ and $\mathcal{A}(X')$.*

zCDP is weaker than ε-DP, but stronger than (ε, δ)-DP: see Proposition 20 in Appendix B.

We now define *semi-DP* (Beimel et al., 2013; Alon et al.,

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

2019), a relaxation of DP that permits $\mathcal{A}$ to violate the privacy of the public data:

Definition 3 (Semi-Differential Privacy (Beimel et al., 2013; Alon et al., 2019)). *Let $n = n_{pub} + n_{priv}$. Consider an algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$ that takes data $X = (X_{priv}, X_{pub}) \in \mathcal{X}^{n_{priv}} \times \mathcal{X}^{n_{pub}}$ as input. $\mathcal{A}$ is (centrally) (ε, δ)-semi-DP if $\mathcal{A}(\cdot, X_{pub})$ is (ε, δ)-DP for all $X_{pub} \in \mathcal{X}^{n_{pub}}$.*

We define *ρ-semi-zCDP* analogously. We define *semi-LDP* in Section 3, which is a similar relaxation of local DP (LDP) (Kasiviswanathan et al., 2011; Duchi et al., 2013). To distinguish Definition 3 from semi-LDP, we sometimes refer to algorithms that satisfy Definition 3 as *centrally* semi-DP.

1.3. Roadmap

We study the central model of semi-DP in Section 2. We give tight error bounds for mean estimation in Section 2.1 and Appendix E.1.2, and a novel algorithm with better constants than the asymptotically optimal algorithms in Section 2.2. In Sections 2.3 and 2.4, we characterize the optimal excess risk of semi-DP ERM and SCO respectively. We give an improved semi-DP algorithm for SCO in Section 2.5. In Section 3, we turn to the local model of semi-DP. We characterize the optimal error rates for semi-LDP mean estimation in Section 3.1 and SCO in Section 3.4. We give semi-LDP algorithms with improved error in Sections 3.2 and 3.5. We experimentally evaluate our algorithms in Section 4 and Appendix G. In Appendix A, we discuss related works in more detail. Due to the page limit, some results and all proofs are presented in the Appendix.

2. Optimal Centrally Private Model Training with Public Data

2.1. Optimal Semi-DP Mean Estimation

In this section, we determine the minimax optimal semi-DP error rates for estimating the mean of a bounded distribution.

Consider the following problem: given n_{priv} private samples $X_{\text{priv}} \subseteq \mathbb{B}$ and n_{pub} public samples $X_{\text{pub}} \subseteq \mathbb{B}$, drawn i.i.d. from an unknown distribution P on $\mathbb{B}$, estimate the population mean $\mathcal{A}(X) \approx \mathbb{E}_{x\sim P}[x]$ subject to the constraint that $\mathcal{A}$ satisfies semi-DP. Defining $n = n_{\text{priv}} + n_{\text{pub}}$, we will characterize the minimax squared error of population mean estimation under (ε, δ)-semi-DP:

$$\mathcal{M}_{\text{pop}}(\varepsilon, \delta, n_{\text{priv}}, n, d) := \inf_{\mathcal{A}\in\mathbb{A}_{\varepsilon,\delta}(\mathbb{B})} \sup_{P} \mathbb{E}_{\mathcal{A}, X\sim P^n}\left[\|\mathcal{A}(X) - \mathbb{E}_{x\sim P}[x]\|^2\right], \quad (1)$$

where $\mathbb{A}_{\varepsilon,\delta}(\mathbb{B})$ denotes the set of all (ε, δ)-semi-DP estimators $\mathcal{A} : \mathbb{B}^n \to \mathbb{B}$, and $|X_{\text{priv}}| = n_{\text{priv}}$.

Theorem 4. *Let $\varepsilon \lesssim 1/\log(nd)$, $\delta \ll 1/n_{priv}$. Then, there is a constant $C > 0$ such that*

$$\ell(d, n) \min\left\{\frac{1}{n_{pub}}, \frac{d}{n^2\varepsilon^2} + \frac{1}{n}\right\} \leqslant \mathcal{M}_{pop}(\epsilon, \delta, n_{priv}, n, d) \leqslant C \min\left\{\frac{1}{n_{pub}}, \frac{d\ln(1/\delta)}{n^2\varepsilon^2} + \frac{1}{n}\right\}, \quad (2)$$

where $1/\ell(d, n)$ is logarithmic in d and n.

Appendix E.1.2 has the proof of Theorem 4 and the pure ε-semi-DP result.

Remark 5. *Technically, our lower bound proof requires us to assume that $\mathcal{A} = (\mathcal{A}^1, \ldots, \mathcal{A}^d)$ is symmetric, meaning $\mathcal{A}^j = \mathcal{A}^l$ for all $j, l \in [d]$. This is a very reasonable assumption: to our knowledge, every algorithm that has been proposed in the literature (for ℓ_2 geometry) is symmetric. Further, the concurrent work of Ullah et al. (2024) gives an alternative proof that eliminates this assumption.*[4]

Naïve algorithms attain the optimal rates: the throw-away estimator $\mathcal{A}(X) = \frac{1}{n_{\text{pub}}}\sum_{x\in X_{\text{pub}}} x$ has MSE $O(1/n_{\text{pub}})$, and the DP (hence semi-DP) Gaussian mechanism has MSE $\widetilde{O}(d/(\varepsilon n)^2 + 1/n)$. In the next subsection, we show that these two algorithms have suboptimal constants: we provide improved (smaller error) estimators.

2.2. An "Even More Optimal" Semi-DP Algorithm for Mean Estimation

Before presenting our improved semi-DP algorithms, we precisely describe the worst-case error of the optimal naïve algorithms discussed in the preceding subsection. We will consider *ρ-semi-zCDP*, which facilitates a sharp characterization of the privacy of the Gaussian mechanism. Note that the lower bound in (2) also holds for semi-zCDP, since $\varepsilon^2/\ln(1/\delta)$-zCDP implies $(O(\varepsilon), \delta)$-DP, by Proposition 20.

Definition 6. *Let $\mathcal{P}(B, V)$ be the collection of all distributions P on $\mathbb{R}^d$ such that for any $x \sim P$, we have $Var(x) = V^2$ and $\|x\| \leqslant B$, P-almost surely.*

Lemma 7. *The error of the ρ-semi-zCDP throw-away algorithm $\mathcal{A}(X) = \frac{1}{n_{pub}}\sum_{x\in X_{pub}} x$ is*

$$\sup_{P\in\mathcal{P}(B,V)} \mathbb{E}_{X\sim P^n}\left[\|\mathcal{A}(X) - \mathbb{E}_{x\sim P}[x]\|^2\right] = \frac{V^2}{n_{pub}}.$$

Further, let $\bar{X}$ be the average of the public and private samples. The minimax error of the ρ-zCDP Gaussian mech-

[4] The work of Ullah et al. (2024) appeared on arXiv March 6, 2024. The first version of our paper appeared on arXiv on June 26, 2023, while v2 added the d-dimensional (ε, δ)-semi-DP lower bounds (Theorem 4 and Theorem 12) and appeared on February 14, 2024.

anism $\mathcal{G}(X) = \bar{X} + \mathcal{N}\left(0, \sigma^2 \mathbf{I}_d\right)$ *is*

$$\inf_{\rho\text{-zCDP } \mathcal{G}} \sup_{P \in \mathcal{P}(B,V)} \mathbb{E}_{\mathcal{G}, X\sim P^n}\left[\|\mathcal{G}(X) - \mathbb{E}_{x\sim P}[x]\|^2\right] = \frac{2dB^2}{\rho n^2} + \frac{V^2}{n}.$$

Intuitively, it seems like the naïve estimators in Lemma 7 do not harness the public and private data in the most effective way possible, despite being optimal up to constants: Throw-away fails to utilize the private data at all, while the Gaussian mechanism gives equal weight to X_{priv} and X_{pub} (regardless of ρ, d, n_{priv}), and provides unnecessary privacy for X_{pub}. We now present a ρ-semi-zCDP estimator that is "even more optimal" than the naïve estimators, meaning our estimator has smaller worst-case error (accounting for constants). We define the family of *Weighted-Gaussian* estimators:

$$\mathcal{A}_r(X) := \sum_{x\in X_{\text{priv}}} rx + \sum_{x\in X_{\text{pub}}} \left(\frac{1 - n_{\text{priv}} r}{n_{\text{pub}}}\right) x + \mathcal{N}\left(0, \frac{2B^2 r^2}{\rho}\mathbf{I}_d\right), \quad (3)$$

for $r \in [0, 1/n_{\text{priv}}]$. This estimator can recover both the throw-away and standard Gaussian mechanisms by choosing $r = 0$ or $r = 1/n$. Intuitively, as ρ/d shrinks, the accuracy cost of adding privacy noise grows, so we should choose smaller r to reduce the sensitivity of $\mathcal{A}_r$. On the other hand, smaller r increases the variance of $\mathcal{A}_r$ on X_{pub}. By choosing r optimally (depending on $\rho, d, n_{\text{priv}}, B, V$), $\mathcal{A}_r$ achieves smaller MSE than both throw-away and the Gaussian mechanism:[5]

Proposition 8. *$\mathcal{A}_r$ is ρ-semi-zCDP, and $\exists r > 0$ such that*

$$\sup_{P\in\mathcal{P}(B,V)} \mathbb{E}_{X\sim P^n}\left[\|\mathcal{A}_r(X) - \mathbb{E}_{x\sim P}[x]\|^2\right] < \min\left(\frac{V^2}{n_{pub}}, \frac{2dB^2}{\rho n^2} + \frac{V^2}{n}\right). \quad (4)$$

Further, if $\frac{V^2}{n_{pub}} \leqslant \frac{2dB^2}{\rho n^2}$, then the advantage of $\mathcal{A}_r$ is

$$\sup_{P\in\mathcal{P}(B,V)} \mathbb{E}_{X\sim P^n}\left[\|\mathcal{A}_r(X) - \mathbb{E}_{x\sim P}[x]\|^2\right] \leqslant \left(\frac{q}{q+s^2}\right)\min\left(\frac{V^2}{n_{pub}}, \frac{2dB^2}{\rho n^2} + \frac{V^2}{n}\right), \quad (5)$$

where $q = 2 + \frac{n_{priv}\rho V^2}{dB^2}$ and $s = \frac{V n_{priv}\sqrt{\rho}}{B\sqrt{d n_{pub}}}$.

When $\frac{V^2}{n_{\text{pub}}} \leqslant \frac{2dB^2}{\rho n^2}$, the throw-away estimator outperforms the DP Gaussian mechanism and our Weighted Gaussian estimator outperforms both of these estimators by a factor of at least $q/(q+s^2)$. Also, $q/(q+s^2) \in [1/2, 1]$ for allowable s, q. For example, if $n = 10,000$, $d = n/100$, $B = 25V$, $\rho = 0.1$, and $n_{\text{pub}} = 0.008n$, then the MSE of our Weighted Gaussian $\mathcal{A}_r$ is smaller than the MSE of throw-away and standard Gaussian by a multiplicative factor of ≈ 1.98.

Figures 12-14 in Appendix G.1.2 show that our estimator outperforms both naïve baselines for d-dimensional Bernoulli data with $\rho = 0.5$ (regardless of whether or not throw-away outperforms the Gaussian mechanism).

For pure ε-semi-DP, using Laplace noise instead of Gaussian noise in (3) yields an estimator with smaller error than the ε-DP Laplace mechanism and throw-away.

2.3. Optimal Semi-DP Empirical Risk Minimization

For a given (fixed) $X = (X_{\text{priv}}, X_{\text{pub}}) \in \mathcal{X}^n$ and parameter domain $\mathcal{W}$, consider the ERM problem:

$$\min_{w\in\mathcal{W}}\left(\widehat{F}_X(w) := \frac{1}{n}\sum_{j=1}^{n} f(w, x_j)\right),$$

where $f(\cdot, x)$ is a loss function and $n_{\text{priv}} = |X_{\text{priv}}|$ samples are private. We discuss practical applications of semi-DP ERM beyond ML in Appendix E.2. We measure the (in-sample) performance of a training algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$ on the data set X by its *excess empirical risk*

$$\mathbb{E}_{\mathcal{A}}\widehat{F}_X(\mathcal{A}(X)) - \widehat{F}^*_X = \mathbb{E}_{\mathcal{A}}\widehat{F}_X(\mathcal{A}(X)) - \min_{w'\in\mathcal{W}} \widehat{F}_X(w').$$

Definition 9. *Let $\mathcal{F}_{\mu,L,D}$ be the set of all functions $f : \mathcal{W}\times\mathcal{X} \to \mathbb{R}$ that are uniformly L-Lipschitz and μ-strongly convex ($\mu \geqslant 0$) in w for some convex compact $\mathcal{W} \subset \mathbb{R}^d$ with ℓ_2-diameter bounded by $D > 0$ and some set $\mathcal{X}$.*

Let $\mathbb{A}_\varepsilon$ contain all ε-semi-DP algorithms $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$ for some $\mathcal{X}, \mathcal{W}$. Define the minimax excess empirical risk of ε-semi-DP (strongly) convex ERM as

$$\mathcal{R}_{\text{ERM}}(\varepsilon, n_{\text{priv}}, n, d, L, D, \mu) := \inf_{\mathcal{A}\in\mathbb{A}_\varepsilon} \sup_{f\in\mathcal{F}_{\mu,L,D}} \sup_{\left\{X\in\mathcal{X}^{n_{\text{priv}}}\times\mathcal{X}^{n_{\text{pub}}}\right\}} \mathbb{E}_{\mathcal{A}}\widehat{F}_X(\mathcal{A}(X)) - \widehat{F}^*_X. \quad (6)$$

Theorem 10. *There are absolute constants $0 < c \leqslant C$ s.t.*

$$cLD\min\left\{\frac{n_{priv}}{n}, \frac{d}{n\varepsilon}\right\} \leqslant \mathcal{R}_{ERM}(\varepsilon, n_{priv}, n, d, L, D, \mu = 0) \leqslant CLD\min\left\{\frac{n_{priv}}{n}, \frac{d}{n\varepsilon}\right\}. \quad (7)$$

See Appendix E.2 for the $\mu > 0$ result and proofs.

Remark 11. *The same minimax risk bound (7) holds up to a logarithmic factor if we replace $\mathcal{F}_{\mu=0,L,D}$ by set of all Lipschitz non-convex loss functions in the definition (6). However, the optimal semi-DP algorithms are inefficient for non-convex loss functions. See Appendix E.2 for details.*

[5] We find the optimal choice of r^* explicitly in the proof of Proposition 8 in Appendix E.1.3.

2.4. Optimal Semi-DP Stochastic Convex Optimization

In stochastic convex optimization (SCO), we are given n i.i.d. samples from an unknown distribution $X \sim P^n$ (with n_{priv} of them being private), and aim to approximately minimize the expected *population loss* $F(w) := \mathbb{E}_{x\sim P}[f(w,x)]$. We measure the quality of a learner $\mathcal{A}$ by its *excess population risk*

$$\begin{aligned}&\mathbb{E}_{\mathcal{A},X\sim P^n} F(\mathcal{A}(X)) - F^* \\ &:= \mathbb{E}_{\mathcal{A},X\sim P^n}\mathbb{E}_{x\sim P}[f(\mathcal{A}(X),x)] - \min_{w'\in\mathcal{W}} \mathbb{E}_{x\sim P} f(w',x).\end{aligned}$$

Denote the minimax optimal semi-DP excess risk by

$$\begin{aligned}&\mathcal{R}_{\text{SCO}}(\varepsilon,\delta,n_{\text{priv}},n,d,L,D,\mu) \\ &:= \inf_{\mathcal{A}\in\mathbb{A}_{\varepsilon,\delta}} \sup_{f\in\mathcal{F}_{\mu,L,D}} \sup_{P} \mathbb{E}_{\mathcal{A},X\sim P^n} F(\mathcal{A}(X)) - F^*, \quad (8)\end{aligned}$$

where $\mathbb{A}_{\varepsilon,\delta}$ contains all (ε,δ)-semi-DP algorithms $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$ for some $\mathcal{X}, \mathcal{W}$, and $|X_{\text{priv}}| = n_{\text{priv}}$.

Theorem 12. *Let $\varepsilon \lesssim 1/\log(nd)$ and $\delta \ll 1/n$. Then, there is a constant $C > 0$ such that*

$$\begin{aligned}&\ell(d,n) LD \min\left\{\frac{1}{\sqrt{n_{pub}}}, \frac{\sqrt{d}}{n\varepsilon} + \frac{1}{\sqrt{n}}\right\} \\ &\quad\leqslant \mathcal{R}_{SCO}(\varepsilon,\delta,n_{priv},n,d,L,D,\mu=0) \\ &\quad\leqslant CLD\min\left\{\frac{1}{\sqrt{n_{pub}}}, \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon} + \frac{1}{\sqrt{n}}\right\},\end{aligned}$$

where $1/\ell(d,n)$ is logarithmic in d and n.

We provide the $\delta = 0$ and μ-strongly convex results ($\mu > 0$), and proofs in Appendix E.3. Remark 5 also applies to Theorem 12

Let us compare the semi-DP bound for SCO in Theorem 12 with the ERM bound in Theorem 10 when $d = 1 = L = D$. Depending on the values of ε and n_{priv}, the minimax excess population risk ("test loss") of SCO may either be larger *or smaller* than the excess empirical risk ("training loss") of ERM. For example, if $\varepsilon \approx 1$, then the semi-DP excess empirical risk $\Theta(1/n)$ is smaller than the excess population risk $\Theta(1/\sqrt{n})$. On the other hand, suppose $\varepsilon \approx 1/n$ and $n_{\text{priv}} \approx n^{2/3}$: then *the semi-DP excess empirical risk $\Theta(1/n^{1/3})$ is larger than the excess population risk* $\Theta(1/\sqrt{n})$. This is surprising: for both non-private learning and DP learning (with $n_{\text{pub}} = 0$), the optimal error of ERM is never larger than that of SCO. While it may seem counter-intuitive that minimizing the training loss can be harder than minimizing test loss, there is a natural explanation: For SCO, a small amount of public data gives us free information about the private data, since $X \sim P^n$ is i.i.d. by assumption. By contrast, for ERM, the public data does not give us any information about the private data, since X is not i.i.d.

2.5. Semi-DP SCO with an "Even More Optimal" Gradient Estimator

Our Algorithm 1 is a noisy stochastic gradient method that uses the "even more optimal" Weighted-Gaussian estimator (33) to estimate $\nabla F(w_t)$ in iteration t.[6] In Algorithm 1, $\text{clip}_C(x) := \text{argmin}_{y\in\mathbb{R}^d, \|y\|\leqslant C} \|x - y\|$ is the Euclidean projection onto the centered ℓ_2-ball of radius C.

We give privacy and excess risk guarantees for Algorithm 1 and describe an accelerated variant of Algorithm 1 in Appendix E.3.1. The excess risk of our algorithm is smaller than the state-of-the-art excess risk for a linear-time DP algorithm whose privacy analysis does not require convexity (Lowy & Razaviyayn, 2023b). We empirically evaluate our algorithm in Section 4.

Algorithm 1 Semi-DP-SGD via Weighted-Gaussian Gradient Estimation

Input: $T \in \mathbb{N}$, clip threshold $C > 0$, stepsizes $\{\eta_t\}_{t\in[T]}$, batch sizes $K_{\text{priv}} \in [n_{\text{priv}}]$, $K_{\text{pub}} \in [n_{\text{pub}}]$, weight parameter $\alpha \in [0,1]$, noise parameter $\sigma^2 > 0$.
Initialize $w_0 \in \mathcal{W}$.
for $t \in \{0,1,\cdots,T-1\}$ **do**
  Draw random batch of K_{priv} private samples B_t^{priv}.
  Draw random batch of K_{pub} public samples B_t^{pub}.
  Draw privacy noise $v_t \sim \mathcal{N}\left(0, \sigma^2 \mathbf{I}_d\right)$.
  $\widetilde{g}_t \leftarrow \alpha\left[\frac{1}{K_{\text{priv}}}\sum_{x\in B_t^{priv}} \text{clip}_C(\nabla f(w_t,x)) + v_t\right] + \frac{1-\alpha}{K_{\text{pub}}}\sum_{x\in B_t^{pub}} \nabla f(w_t,x)$.
  Update $w_{t+1} := \Pi_{\mathcal{W}}[w_t - \eta_t \widetilde{g}_t]$.
end for
Output: w_T or an average of the iterates $\{w_t\}_{t\in[T]}$.

3. Optimal Locally Private Model Training with Public Data

We now turn to a stronger privacy notion that we refer to as *semi-local DP* (semi-LDP). Semi-LDP guarantees *privacy for each private x_i, without requiring person i to trust others* (e.g. central server). Semi-LDP generalizes LDP (Kasiviswanathan et al., 2011; Duchi et al., 2013), which has been deployed in industry (Apple, 2016; Úlfar Erlingsson et al., 2014; Ding et al., 2017).

Following Duchi & Rogers (2019), we permit algorithms to be *fully interactive*: algorithms may adaptively query the same person i multiple times over the course of T communication rounds. For example, n cell phone users send messages to a server over T rounds, and message $Z_{i,t} \in \mathcal{Z}$ sent by user i in round t can depend on the previous messages $\{Z_{j,t}\}_{j\leqslant n, t'\leqslant t}$. Semi-LDP requires the messages

[6] In Algorithm 1, we re-parameterize by setting $r = \frac{\alpha}{K_{\text{priv}}}$ for $\alpha \in [0,1]$.

$\{Z_{i,t}\}_{t\in[T]}$ to be semi-DP:

Definition 13 (Semi-Local Differential Privacy)**.** *The T-round interactive algorithm $\mathcal{A}$ is (ε,δ)-semi-LDP if the transcript $Z=\{Z_{i,t}\}_{i\leqslant n,t\leqslant T}$ is (ε,δ)-semi-DP: i.e. for all $X_{pub}\in\mathcal{X}^{n_{pub}}$, all adjacent $X_{priv}\sim X'_{priv}$ and all $S\subset\mathcal{Z}^{nT}$, $\mathbb{P}(Z\in S|X=(X_{priv},X_{pub}))\leqslant e^{\varepsilon}\mathbb{P}(Z\in S|X=(X'_{priv},X_{pub}))+\delta$.*

Definition 13 is stronger than Definition 3, since the latter only requires the final output of $\mathcal{A}$ to be semi-DP.

3.1. Optimal Semi-LDP Mean Estimation

We will characterize the minimax squared error of ε-semi-LDP d-dimensional mean estimation:

$$\mathcal{M}^{\text{loc}}_{\text{pop}}(\varepsilon,n_{\text{priv}},n,d) := \inf_{\mathbb{A}^{\text{loc}}_{\varepsilon}(\mathbb{B})}\sup_{P}\mathbb{E}_{\mathcal{A},X\sim P^n}\left[\|\mathcal{A}(X)-\mathbb{E}_{x\sim P}[x]\|^2\right], \quad (9)$$

where $\mathbb{A}^{\text{loc}}_{\varepsilon}(\mathbb{B})$ contains all (fully interactive) ε-semi-LDP estimators $\mathcal{A}:\mathbb{B}^n\to\mathbb{B}$, and $|X_{\text{priv}}|=n_{\text{priv}}$.

Theorem 14. *Let $\varepsilon\in(0,1]$. There are absolute constants $0<c\leqslant C$ s.t.*

$$c\min\left\{\frac{1}{n_{pub}},\frac{d}{n\varepsilon^2}\right\}\leqslant\mathcal{M}^{loc}_{pop}(\varepsilon,n_{priv},n,d)\leqslant C\min\left\{\frac{1}{n_{pub}},\frac{d}{n\varepsilon^2}\right\}. \quad (10)$$

Remark 15 (Approximate Semi-LDP)**.** *Theorem 14 still holds if we replace $\mathbb{A}^{loc}_{\varepsilon}$ in the definition of $\mathcal{M}^{loc}_{pop}(\epsilon,n_{priv},n,d)$ by the set of all (ε,δ)-semi-LDP estimators $\mathcal{A}$ for which either $\delta<1/2$ and $\mathcal{A}$ is "compositional" (Duchi & Rogers, 2019) (e.g. sequentially interactive (Duchi et al., 2013)) or $\delta<1/2^d$.*

The upper bound in Theorem 14 is the minimum of the error of the throw-away estimator and the optimal ε-LDP estimator of Duchi et al. (2013). The LDP estimator of Duchi et al. (2013) takes the form

$$\widetilde{\mathcal{M}}_{\text{Duchi}}(X)=\frac{1}{n}\sum_{i=1}^{n}\mathcal{M}_{\text{Duchi}}(x_i),$$

where $\mathcal{M}_{\text{Duchi}}(x_i)$ samples a vector uniformly from a carefully chosen subset of $\mathbb{B}$, depending on x_i.

3.2. An "Even More Optimal" Semi-LDP Estimator

By applying $\mathcal{M}_{\text{Duchi}}$ only to the private samples, we obtain an ε-semi-LDP algorithm with smaller error than the asymptotically optimal $\widetilde{\mathcal{M}}_{\text{Duchi}}$. Define the *Semi-LDP $\mathcal{A}_{Semi\text{-}Duchi}$*:

$$\mathcal{A}_{\text{Semi-Duchi}}(X)=\frac{1}{n}\left[\sum_{x\in X_{\text{priv}}}\mathcal{M}_{\text{Duchi}}(x)+\sum_{x'\in X_{\text{pub}}}x'\right]. \quad (11)$$

Let P be a distribution on $\mathbb{B}$ with $V^2:=\mathbb{E}\|x-\mathbb{E}_{x\sim P}[x]\|^2$.

Lemma 16. *Let $c>0$ such that $\mathbb{E}_{x\sim P}\|\mathcal{M}_{Duchi}(x)-\mathbb{E}_{x\sim P}[x]\|^2=\frac{cd}{n\varepsilon^2}$, so that $\mathbb{E}_{X\sim P^n}\|\widetilde{\mathcal{M}}_{Duchi}(X)-\mathbb{E}_{x\sim P}[x]\|^2=\frac{cd}{n\varepsilon^2}+\frac{V^2}{n}$. Then,*

$$\mathbb{E}_{X\sim P^n}\left[\|\mathcal{A}_{Semi\text{-}Duchi}(X)-\mathbb{E}_{x\sim P}[x]\|^2\right]=\frac{n_{priv}}{n}\cdot\frac{cd}{n\varepsilon^2}+\frac{n_{pub}}{n}\cdot\frac{V^2}{n}.$$

The constant c in Lemma 16 that bounds the error of $\mathcal{M}_{\text{Duchi}}$ may depend on the distribution P. Lemma 16 shows that given any P, the error of our semi-LDP estimator is smaller than the error of $\widetilde{\mathcal{M}}_{\text{Duchi}}$. Quantitatively, *the MSE of our estimator is smaller than the MSE of $\widetilde{\mathcal{M}}_{Duchi}$ by a factor of n_{priv}/n* if the privacy noise error term is dominant (e.g. if $d\gg\varepsilon^2$).

3.3. A Semi-LDP Estimator with Optimal Constants

In this subsection, we consider the task of estimating the average of data X on the unit sphere $\mathcal{S}^{d-1}\subset\mathbb{R}^d$. We give a semi-LDP estimator that is *truly optimal*—i.e. *our estimator has the smallest MSE, including constants*—among a large class of unbiased semi-LDP estimators of $\bar{X}$. Our semi-LDP estimator, $\mathcal{A}_{\text{semi-PrivU}}$ takes a similar shape to $\widetilde{\mathcal{A}}_{\text{semi-Duchi}}$, but uses *PrivUnit* (Bhowmick et al., 2018) instead of $\mathcal{M}_{\text{Duchi}}$ as the LDP randomizer in (11). We recall PrivUnit in Algorithm 3 in Appendix F.

Proposition 17. *Let $\mathcal{R}:\mathbb{S}^{d-1}\to\mathcal{Z}$ be an ε-LDP randomizer, $\mathcal{M}_{priv}$ and $\mathcal{M}_{pub}$ be aggregation protocols, and $\mathcal{A}(X)=\frac{1}{n}\left[\mathcal{M}_{priv}(\mathcal{R}(x_1),\ldots,\mathcal{R}(x_{n_{priv}}))+\mathcal{M}_{pub}(X_{pub})\right]$. Assume $\mathbb{E}\left[\mathcal{M}_{priv}(\mathcal{R}(x_1),\cdots,\mathcal{R}(x_{n_{priv}}))|X_{priv}\right]=\sum_{x\in X_{priv}}x$ and $\mathbb{E}\left[\mathcal{M}_{pub}(X_{pub})|X_{priv}\right]=\sum_{x\in X_{pub}}x\ \forall X=(X_{priv},X_{pub})\in\left(\mathbb{S}^{d-1}\right)^n$ Then,*

$$\sup_{X\in(\mathbb{S}^{d-1})^n}\mathbb{E}_{\mathcal{A}_{semi\text{-}PrivU}}\left[\|\mathcal{A}_{semi\text{-}PrivU}(X)-\bar{X}\|^2\right]\leqslant\sup_{X\in(\mathbb{S}^{d-1})^n}\mathbb{E}_{\mathcal{A}}\|\mathcal{A}(X)-\bar{X}\|^2.$$

Proposition 17 is proved by extending the analysis of Asi et al. (2022) to the semi-LDP setting.

3.4. Optimal Semi-LDP Stochastic Convex Optimization

We will characterize the minimax optimal excess population risk of semi-LDP SCO

$$\mathcal{R}^{\text{loc}}_{\text{SCO}}(\varepsilon,n_{\text{priv}},n,d,L,D,\mu) := \inf_{\mathcal{A}\in\mathbb{A}^{\text{loc}}_{\varepsilon}}\sup_{f\in\mathcal{F}_{\mu,L,D}}\sup_{P}\mathbb{E}_{\mathcal{A},X\sim P^n}F(\mathcal{A}(X))-F^*, \quad (12)$$

where $\mathbb{A}_\varepsilon^{\text{loc}}$ denotes the set of all algorithms $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$ that are ε-semi-LDP for some $\mathcal{X}$ and $\mathcal{W}$, and exactly n_{priv} samples in X are private.

Theorem 18. *Let $\varepsilon \in (0, 1]$ and $h(\varepsilon, n_{priv}, n, d, L, D) := LD \min\left\{\frac{1}{\sqrt{n_{pub}}}, \sqrt{\frac{d}{n\varepsilon^2}}\right\}$. There are absolute constants c and C with $0 < c \leqslant C$, such that*

$$c\, h(\varepsilon, n_{priv}, n, d, L, D) \leqslant \mathcal{R}^{loc}_{SCO}(\varepsilon, n_{priv}, n, d, L, D, \mu = 0) \leqslant C\, h(\varepsilon, n_{priv}, n, d, L, D).$$

See Appendix F.2 for the $\mu > 0$ result and proofs. The first term in the upper bound is achieved by throwing away X_{priv} and running SGD on X_{pub} (Nemirovski & Yudin, 1983). The second term in the upper bound is achieved by the one-pass *LDP-SGD* of Duchi et al. (2013). Remark 15 also applies to Theorem 18.

3.5. "Even More Optimal" Semi-LDP SCO Algorithm

We give a semi-LDP algorithm, called *Semi-LDP-SGD*, with smaller excess risk than the optimal LDP-SGD of Duchi et al. (2013). Essentially, Semi-LDP-SGD runs as follows: In each iteration $t \in [n]$, we draw a random sample $x_t \in X$ without replacement. If $x_t \in X_{\text{priv}}$, update $w_{t+1} = \Pi_{\mathcal{W}}\left[w_t - \eta \mathcal{M}_{\text{Duchi}}\left(\nabla f(w_t, x_t)\right)\right]$; if $x_t \in X_{\text{pub}}$, instead update $w_{t+1} = \Pi_{\mathcal{W}}\left[w_t - \eta \nabla f(w_t, x_t)\right]$. See Algorithm 4 in Appendix F.2.1 for pseudocode.

Proposition 19. *Let $f \in \mathcal{F}_{\mu=0,L,D}$, let P be any distribution and $\varepsilon \leqslant d$. Algorithm 4 is ε-semi-LDP. Further, there is an absolute constant c such that the output $\mathcal{A}(X) = \bar{w}_n$ of Algorithm 4 satisfies*

$$\mathbb{E}_{\mathcal{A}, X\sim P^n}\left[F(\bar{w}_n) - F^*\right] \leqslant c\frac{LD}{\sqrt{n}} \max\left\{\sqrt{\frac{d}{\varepsilon^2}}\sqrt{\frac{n_{priv}}{n}}, \sqrt{\frac{n_{pub}}{n}}\right\}. \quad (13)$$

Thus, Algorithm 4 has smaller excess risk than LDP-SGD, roughly by a factor of $\sqrt{n_{\text{priv}}/n}$.

4. Numerical Experiments

In this section, we empirically evaluate the performance of four different semi-DP algorithms: 1. *Throw-away* (i.e. minimize the public loss). 2. *DP-SGD* (Abadi et al., 2016; De et al., 2022). 3. *PDA-MD* (Amid et al., 2022), which is the state-of-the-art semi-DP algorithm for training convex models. 4. *Our Algorithm 1*. Unless otherwise noted, we evaluate all algorithms with *"warm start,"* which means finding a minimizer w_{pub} of the public loss and initializing training at w_{pub}. The hyperparameters of each algorithm were carefully tuned. See Appendix G for details on the experimental setups and additional results.

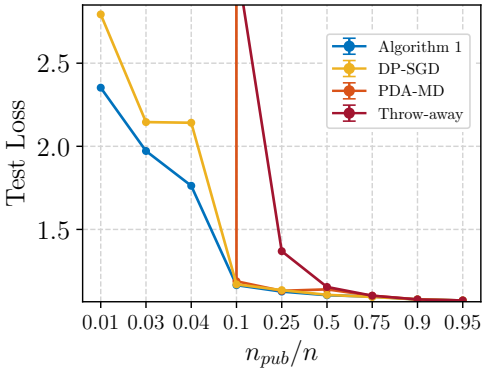

Figure 3. Test loss vs. n_{pub}/n. $\varepsilon = 2$.

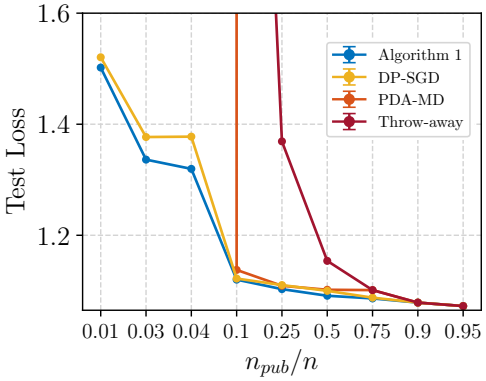

Figure 4. Test loss vs. n_{pub}/n. $\varepsilon = 4$.

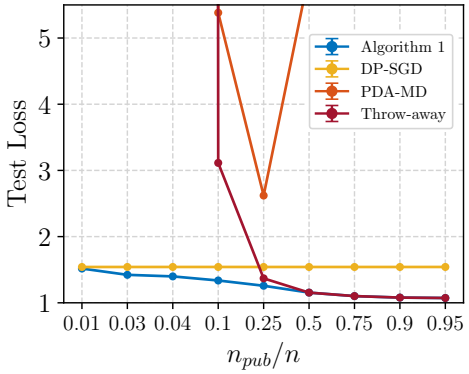

Figure 5. Test loss vs. n_{pub}/n. $\varepsilon = 4$, without warm-start.

Our Algorithm 1 achieves the smallest test loss among the semi-DP baselines across different levels of ε (privacy) and n_{pub}: Figures 3-4 show results for $(\varepsilon, \delta = 10^{-5})$-semi-DP linear regression with synthetic Gaussian data. In the Appendix, we evaluate the algorithms in several other tasks: e.g., logistic regression and Wide-ResNet: see Figures 15-18 and 19-20. Our results indicate that *Algorithm 1 consistently outperforms all baselines*.

Our Algorithm 1 can converge even when DP-SGD diverges: Figure 6 gives an example in which DP-SGD diverges but Algorithm 1 converges. We used the following

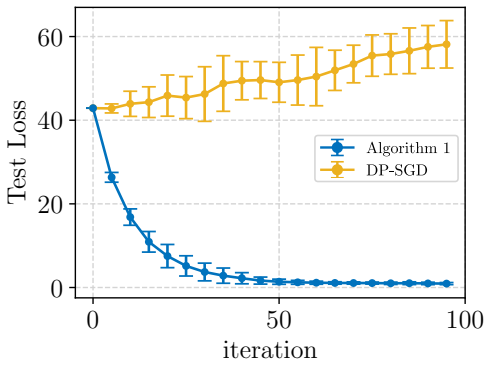

Figure 6. Test loss vs. iterations. $\frac{n_{\text{pub}}}{n} = 0.1$

parameters: $d = 50$, $n = 1000$, $\varepsilon = 0.01$, and $\frac{n_{\text{pub}}}{n} = 0.1$.

Algorithm 1 performs well even without "warm start", whereas PDA-MD performs poorly: e.g., Figures 5 and (in Appendix) 11 show that PDA-MD even performs worse than throw-away. By contrast, Algorithm 1 is resilient to the "cold start" condition and outperforms all baselines. In certain practical applications, such as advertising and healthcare, samples are often obtained in an online/streaming fashion, precluding the possibility of "warm start".

More public data always improves the performance of Algorithm 1, but does not always benefit PDA-MD. The main reason for this is that our algorithm more effectively handles the increasing privacy noise that is needed to maintain semi-DP with increasing n_{pub}/n. We give details and numerical evidence of this explanation in Appendix G.1.1.

Appendix G contains other findings, too. For example, Figure 8 shows that PDA-MD is sensitive to Hessian regularization parameter, which requires extra tuning on complicated tasks. By contrast, Algorithm 1 does not use any Hessian information.

5. Conclusion

We considered training DP models with side access to public data. Theoretically, we characterized the optimal error bounds (up to constants) for three fundamental problems: mean estimation, empirical risk minimization, and stochastic convex optimization. We show that it is impossible to improve over the naïve semi-DP algorithms asymptotically, in the worst case. Algorithmically, we developed new optimal methods for semi-DP learning that have smaller error than the asymptotically optimal algorithms. Empirically, we showed that our algorithms are effective in training semi-DP models. Our work raises interesting open questions. For instance, why do certain learning/optimization problems benefit more from public data than others? Is there some general underlying property that (don't) permit asymptotic benefits over the naïve baselines? Also, what can be said about semi-DP learning with *out-of-distribution* public data? Lastly, it would be useful to have an extensive empirical study that evaluates the efficacy of combining our semi-DP algorithms with various other techniques, such as dimensionality reduction (Yu et al., 2021b; Pinto et al., 2024).

Impact Statement

Our work provides algorithms for protecting the privacy of individuals who contribute training data. Privacy is commonly regarded in a positive light and is even enshrined as a fundamental right in various legal systems. However, there is a risk that corporations or governments could exploit our algorithms for nefarious purposes, such as unauthorized collection of personal data. Furthermore, the use of semi-privately trained models may result in decreased accuracy compared to non-private models, which can have adverse consequences. For instance, if a semi-DP model is utilized to forecast the effects of pollution, but yields less precise and overly optimistic outcomes, it could provide pretext for a government to unjustly dismantle environmental safeguards. Nonetheless, we firmly believe that the dissemination of privacy-preserving machine learning algorithms, coupled with enhanced understanding of these algorithms, ultimately offers a net benefit to society.

Another potential misuse of our work would be using the accuracy benefits of public data to argue for less stringent data privacy policies, laws, or regulations. However, we want to emphasize that even though public data can enhance the accuracy of models, we firmly believe that privacy laws and corporate policies should not be weakened. Differentially private synthetic data generation (Torkzadehmahani et al., 2019; Vietri et al., 2020; Boedihardjo et al., 2022; He et al., 2023) is one possible avenue for generating public data in an ethical, privacy-preserving manner.

Acknowledgements

The authors would like to thank Gautam Kamath, Adam Smith, Thomas Steinke, and Jonathan Ullman for very helpful pointers and explanations related to existing lower bound proof techniques. We thank Arnold Pereira for discussions and providing feedback at various stages of this project. We thank Michael Menart for helpful feedback on an earlier version of this manuscript. We thank the TPDP and ICML reviewers for their helpful feedback. This work was supported in part by a gift from Meta and the USC-Meta Center for Research and Education in AI & Learning. AL's work was supported in part by NSF award DMS-2023239 and AFOSR award FA9550-21-1-0084.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Oct 2016. doi: 10.1145/2976749.2978318. URL http://dx.doi.org/10.1145/2976749.2978318.

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

Alon, N., Bassily, R., and Moran, S. Limits of private learning with access to public data. *Advances in Neural Information Processing Systems*, 32, 2019.

Amid, E., Ganesh, A., Mathews, R., Ramaswamy, S., Song, S., Steinke, T., Suriyakumar, V. M., Thakkar, O., and Thakurta, A. Public data-assisted mirror descent for private model training. In *International Conference on Machine Learning*, pp. 517–535. PMLR, 2022.

Apple. Differential privacy overview, 2016.

Asi, H. and Duchi, J. C. Near instance-optimality in differential privacy. *arXiv preprint arXiv:2005.10630*, 2020a.

Asi, H. and Duchi, J. C. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14106–14117. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/a267f936e54d7c10a2bb70dbe6ad7a89-Paper.pdf.

Asi, H., Feldman, V., and Talwar, K. Optimal algorithms for mean estimation under local differential privacy. In *International Conference on Machine Learning*, pp. 1046–1056. PMLR, 2022.

Avent, B., Korolova, A., Zeber, D., Hovden, T., and Livshits, B. Blender: Enabling local search with a hybrid differential privacy model. In *USENIX Security Symposium*, pp. 747–764, 2017.

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

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

Bassily, R., Thakkar, O., and Guha Thakurta, A. Model-agnostic private learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Bassily, R., Feldman, V., Talwar, K., and Thakurta, A. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Bassily, R., Cheu, A., Moran, S., Nikolov, A., Ullman, J., and Wu, S. Private query release assisted by public data. In *International Conference on Machine Learning*, pp. 695–703. PMLR, 2020.

Beimel, A., Nissim, K., and Stemmer, U. Private learning and sanitization: Pure vs. approximate differential privacy. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques: 16th International Workshop, APPROX 2013, and 17th International Workshop, RANDOM 2013, Berkeley, CA, USA, August 21-23, 2013. Proceedings*, pp. 363–378. Springer, 2013.

Ben-David, S., Bie, A., Canonne, C. L., Kamath, G., and Singhal, V. Private distribution learning with public data: The view from sample compression. *arXiv preprint arXiv:2308.06239*, 2023.

Bhowmick, A., Duchi, J., Freudiger, J., Kapoor, G., and Rogers, R. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018.

Bie, A., Kamath, G., and Singhal, V. Private estimation with public data. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 18653–18666. Curran Associates, Inc., 2022.

Boedihardjo, M., Strohmer, T., and Vershynin, R. Covariance's loss is privacy's gain: Computationally efficient, private and accurate synthetic data. *Foundations of Computational Mathematics*, pp. 1–48, 2022.

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

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

Bun, M., Ullman, J., and Vadhan, S. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 1–10, 2014.

Bun, M., Steinke, T., and Ullman, J. Make up your mind: The price of online queries in differential privacy. In *Proceedings of the twenty-eighth annual ACM-SIAM symposium on discrete algorithms*, pp. 1306–1325. SIAM, 2017.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T. B., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6, 2021.

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

Church, G. M. The personal genome project, 2005.

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

Ding, B., Kulkarni, J., and Yekhanin, S. Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30, 2017.

Duchi, J. Lecture notes for statistics 311/electrical engineering 377. *URL: https://stanford.edu/class/stats311/Lectures/full_notes.pdf*, 2021.

Duchi, J. and Rogers, R. Lower bounds for locally private estimation via communication complexity. In *Conference on Learning Theory*, pp. 1161–1191. PMLR, 2019.

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

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

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

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

Dwork, C., Smith, A., Steinke, T., Ullman, J., and Vadhan, S. Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pp. 650–669. IEEE, 2015.

Fallah, A., Makhdoumi, A., Malekian, A., and Ozdaglar, A. Optimal and differentially private data acquisition: Central and local mechanisms. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, pp. 1141, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391504. doi: 10.1145/3490486.3538329. URL https://doi.org/10.1145/3490486.3538329.

Feldman, V., Mironov, I., Talwar, K., and Thakurta, A. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 521–532. IEEE, 2018.

Ferrando, C., Gillenwater, J., and Kulesza, A. Combining public and private data. *arXiv preprint arXiv:2111.00115*, 2021.

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

Ganesh, A., Thakurta, A., and Upadhyay, J. Langevin diffusion: An almost universal algorithm for private euclidean (convex) optimization. *arXiv preprint arXiv:2204.01585*, 2022.

Ganesh, A., Haghifam, M., Nasr, M., Oh, S., Steinke, T., Thakkar, O., Thakurta, A., and Wang, L. Why is public pretraining necessary for private model training?, 2023.

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

Golatkar, A., Achille, A., Wang, Y.-X., Roth, A., Kearns, M., and Soatto, S. Mixed differential privacy in computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8376–8386, 2022.

Gu, X., Kamath, G., and Wu, Z. S. Choosing public datasets for private machine learning via gradient subspace distance, 2023.

Hardt, M. and Talwar, K. On the geometry of differential privacy. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 705–714, 2010.

He, Y., Vershynin, R., and Zhu, Y. Algorithmically effective differentially private synthetic data. *arXiv preprint arXiv:2302.05552*, 2023.

Jorgensen, Z., Yu, T., and Cormode, G. Conservative or liberal? personalized differential privacy. In *2015 IEEE 31St international conference on data engineering*, pp. 1023–1034. IEEE, 2015.

Kairouz, P., Diaz, M. R., Rush, K., and Thakurta, A. (nearly) dimension independent private erm with adagrad rates via publicly estimated subspaces. In *Conference on Learning Theory*, pp. 2717–2746. PMLR, 2021.

Kamath, G., Liu, X., and Zhang, H. Improved rates for differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pp. 10633–10660. PMLR, 2022a.

Kamath, G., Mouzakis, A., and Singhal, V. New lower bounds for private estimation and a generalized fingerprinting lemma. *Advances in Neural Information Processing Systems*, 35:24405–24418, 2022b.

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

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Kerrigan, G., Slack, D., and Tuyls, J. Differentially private language models benefit from public pre-training. *arXiv preprint arXiv:2009.05886*, 2020a.

Kerrigan, G., Slack, D., and Tuyls, J. Differentially private language models benefit from public pre-training. *arXiv preprint arXiv:2009.05886*, 2020b.

Klimt, B. and Yang, Y. The enron corpus: A new dataset for email classification research. In *Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004. Proceedings 15*, pp. 217–226. Springer, 2004.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Li, X., Tramer, F., Liang, P., and Hashimoto, T. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*, 2021a.

Li, X., Tramer, F., Liang, P., and Hashimoto, T. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2021b.

Liu, J. and Talwar, K. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 298–309, 2019.

Liu, R., Bu, Z., Wang, Y.-x., Zha, S., and Karypis, G. Coupling public and private gradient provably helps optimization. *arXiv preprint arXiv:2310.01304*, 2023.

Liu, T., Vietri, G., Steinke, T., Ullman, J., and Wu, S. Leveraging public data for practical private query release. In *International Conference on Machine Learning*, pp. 6968–6977. PMLR, 2021.

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

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

Lowy, A. and Razaviyayn, M. Private federated learning without a trusted server: Optimal algorithms for convex losses. In *International Conference on Learning Representations*, 2023b. URL `https://openreview.net/pdf?id=TVY6GoURrw`.

Lowy, A., Ghafelebashi, A., and Razaviyayn, M. Private non-convex federated learning without a trusted server. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, 2023a.

Lowy, A., Gupta, D., and Razaviyayn, M. Stochastic differentially private and fair learning. In *International Conference on Learning Representations*, 2023b. URL `https://openreview.net/pdf?id=3nM5uhPlfv6`.

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

Mehta, H., Thakurta, A., Kurakin, A., and Cutkosky, A. Large scale transfer learning for differentially private image classification. *arXiv preprint arXiv:2205.02973*, 2022.

Mühl, C. and Boenisch, F. Personalized pate: Differential privacy for machine learning with individual privacy guarantees. *arXiv preprint arXiv:2202.10517*, 2022.

Nasr, M., Mahloujifar, S., Tang, X., Mittal, P., and Houmansadr, A. Effectively using public data in privacy preserving machine learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.),

Proceedings of the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, pp. 25718–25732. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/nasr23a.html.

Nemirovski, A. and Yudin, D. *Problem complexity and method efficiency in optimization.* Chichester, 1983.

Papernot, N. and Steinke, T. Hyperparameter tuning with renyi differential privacy. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=-70L8lpp9DF.

Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I., and Talwar, K. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations*, 2017.

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., and Erlingsson, U. Scalable private learning with pate. In *International Conference on Learning Representations*, 2018.

Pinto, F., Hu, Y., Yang, F., and Sanyal, A. Pillar: How to make semi-private learning more effective. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 110–139, 2024. doi: 10.1109/SaTML59370.2024.00014.

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

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1571–1578, 2012.

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

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Tang, X., Panda, A., Sehwag, V., and Mittal, P. Differentially private image classification by learning priors from random processes, 2023.

Thakurta, A. G., Vyrros, A. H., Vaishampayan, U. S., Kapoor, G., Freudiger, J., Sridhar, V. R., and Davidson, D. Learning new words, March 14 2017. US Patent 9,594,741.

Torkzadehmahani, R., Kairouz, P., and Paten, B. Dp-cgan: Differentially private synthetic data and label generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

Úlfar Erlingsson, Pihur, V., and Korolova, A. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 21st ACM Conference on Computer and Communications Security*, 2014.

Ullah, E., Menart, M., Bassily, R., Guzmán, C., and Arora, R. Public-data assisted private stochastic optimization: Power and limitations. *arXiv preprint arXiv:2403.03856*, 2024.

U.S. Census Bureau. Differential privacy and the 2020 census, 2020.

Vietri, G., Tian, G., Bun, M., Steinke, T., and Wu, S. New oracle-efficient algorithms for private synthetic data release. In *International Conference on Machine Learning*, pp. 9765–9774. PMLR, 2020.

Wang, J. and Zhou, Z.-H. Differentially private learning with small public data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6219–6226, 2020.

Yu, B. Assouad, fano, and le cam. *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pp. 423–435, 1997.

Yu, D., Naik, S., Backurs, A., Gopi, S., Inan, H. A., Kamath, G., Kulkarni, J., Lee, Y. T., Manoel, A., Wutschitz, L., et al. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*, 2021a.

Yu, D., Zhang, H., Chen, W., and Liu, T.-Y. Do not let privacy overbill utility: Gradient embedding perturbation for private learning. *arXiv preprint arXiv:2102.12677*, 2021b.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprintarXiv:1605.07146*, 2017.

Zhou, Y., Wu, Z. S., and Banerjee, A. Bypassing the ambient dimension: Private sgd with gradient subspace identification. *arXiv preprint arXiv:2007.03813*, 2020.

Appendix

A. Related Work

Many works have considered a variety of semi-DP learning problems, empirically and theoretically (Bassily et al., 2018; Feldman et al., 2018; Bassily et al., 2020; Kairouz et al., 2021; Wang & Zhou, 2020; Zhou et al., 2020; Li et al., 2021a; Liu et al., 2021; Papernot et al., 2017; 2018; Yu et al., 2021b; Alon et al., 2019; Bie et al., 2022; Ferrando et al., 2021; Amid et al., 2022). Here we discuss the works that are most closely related to our own.

Sample complexity bounds for semi-DP learning and estimation: The works of Alon et al. (2019); Bassily et al. (2020) give sample complexity bounds for semi-DP PAC learning and query release. Their upper bounds show that for hypothesis classes with finite VC-dimension, asymptotic improvements over the naïve approaches are possible. These results stand in contrast with our lower bounds for model training/optimization with Lipschitz loss functions. Both of these works also provide lower bounds. Below, we compare their results with our own lower bounds.

The negative result of Theorem 4.2 in (Alon et al., 2019) is similar in spirit to our lower bounds: no semi-DP algorithm can achieve error better than the minimum of the optimal DP error and a term that depends on n_{pub}. That being said, there are some significant and consequential differences between our lower bounds and (Alon et al., 2019, Theorem 4.2):

1. *Different learning problems*: (Alon et al., 2019) considers agnostic PAC learning/binary classification, whereas we consider model training (mean estimation and optimization). We are not aware of any way to obtain our lower bounds from the results of (Alon et al., 2019).
2. *Quantitative differences in the lower bounds*: The lower bound implied by (Alon et al., 2019, Theorem 4.2) is of the form $\Omega\left(\min\left\{1/n_{\text{pub}}, \text{optimal DP error}\right\}\right)$. Our lower bounds do not always take this form. For example, consider Theorem 10: the first term in our lower bound is n_{priv}/n, which can imply a much larger error than the bound in (Alon et al., 2019) (e.g., if $d = \varepsilon n$ and $n_{\text{priv}} = n/2$). This illustrates how different learning problems can benefit more from public data than others.
3. *Central vs. Local Semi-DP*: (Alon et al., 2019) only covers central semi-DP, whereas we cover both central and local semi-DP.
4. *Pure vs. Approximate Semi-DP*: (Alon et al., 2019)'s proof technique cannot handle approximate semi-DP because their Lemma 2.6 is limited to pure DP. By contrast, we give lower bounds for both pure and approximate semi-DP.
5. *New Techniques*: Our techniques differ substantially from those in (Alon et al., 2019). For example, we develop a novel semi-DP Fano's inequality and a novel semi-DP packing argument. For approximate semi-DP, we utilize fingerprinting proofs. Also, our semi-LDP lower bound techniques are completely different from (Alon et al., 2019)'s techniques. We hope that our novel techniques to find applications beyond those in our paper.

The result of Bassily et al. (2020, Theorem 13) showed (up to logarithmic factors) that no improvement over the naïve approaches is possible for *approximate* $(1, \delta)$-semi-DP releasing decision stumps. This implies a lower bound for $(1, \delta)$-semi-DP mean estimation in the ℓ_∞ norm, but does not imply the tight lower bound for the ℓ_2 setting that we provide in Theorem 4. Moreover, (Bassily et al., 2018)'s results and techniques do not lead to tight lower bounds for *pure* ε-semi-DP decision stumps or mean estimation. Thus, we develop novel techniques (e.g. semi-DP Fano and semi-DP packing arguments) for pure semi-DP estimation and model training.

For semi-DP d-dimensional Gaussian mean estimation, (Bie et al., 2022) gave sample complexity bounds that do not depend on the range parameters of the distribution if $n_{\text{pub}} \geqslant d + 1$; this is known to be impossible without public data. The concurrent and independent work of Ben-David et al. (2023) established a lower bound for this problem. (Ben-David et al., 2023) also explored the connection between semi-DP distribution learning of a class and the existence of a sample compression scheme for that class.

DP model training (ERM and SCO) with public data: The works of Kairouz et al. (2021); Zhou et al. (2020) considered DP ERM with public data and additional assumptions on the gradients lying in a certain low-dimensional subspace. Under these additional assumptions, (Kairouz et al., 2021; Zhou et al., 2020) show that nearly dimension-independent excess empirical risk bounds are possible (e.g. by using the public data to estimate the low-dimensional subspace and projecting

noisy gradients onto this subspace). Our lower bounds show that these additional assumptions are strictly necessary: in general, polynomial dependence on the dimension is necessary for semi-DP ERM and SCO. The work of Wang & Zhou (2020) used public data to adjust the parameters of DP-SGD. Empirically, pre-training on public data sets and privately fine-tuning the model (Li et al., 2021a; Kerrigan et al., 2020a; Mehta et al., 2022) has shown great promise for large-scale ML.

The work of Amid et al. (2022) developed a public data-assisted DP mirror descent (PDA-MD) algorithm that sometimes outperforms DP-SGD empirically in training ML models, and theoretically in terms of excess risk for linear regression under certain distributional assumptions. We use the PDA-MD of Amid et al. (2022) as a baseline in our experiments. (Amid et al., 2022) also gave an "efficient approximation" of their PDA-MD in (Amid et al., 2022, Equation 1), which they used for training non-convex models. While finalizing this manuscript, we became aware that this "efficient approximation" is nearly equivalent to our Algorithm 1. However, there are differences in the implementation of our algorithm. For example, we use a constant weight parameter α, whereas (Amid et al., 2022) uses decaying weights $\{\alpha_t\}_{t=1}^T$. Also, we clip both the private *and public* gradients in our implementation of Algorithm 1, which empirically improves performance: see Appendix G.1.1 for further discussion. Our Algorithm 1 was derived in a different fashion from the algorithm in (Amid et al., 2022, Equation 1): we derived our algorithm as an application of our "even more optimal" mean estimator, while theirs was derived as an approximation of their mirror descent method. Moreover, no theoretical analysis was provided for the "efficient approximation" in (Amid et al., 2022). In the linear regression setting, the PDA-MD algorithm can be viewed as Newton's algorithm where the Hessian is estimated via public data. Then the Hessian is inverted and multiplied by privatized gradients at each iteration. In other words, the update rule of the algorithm is given by $w^{t+1} = w^t - \alpha_t(X_{pub}^T X_{pub})^{-1}(g_t + n_t)$ where X_{pub} is the public data, g_t is the sampled gradient from private data, and n_t is the added noise. When the number of public data samples is small, the estimate of the Hessian becomes low rank (or inaccurate) and the (pseudo)inverse of it may introduce additional error (even after proper regularization). On the other hand, when the number of public data samples is large, but the number of private data is small, the PDA-MD algorithm can still suffer if it is not warm-started. This is because, although the Hessian $X^T X$ is estimated accurately in this case, but there is not enough private data to generate enough gradients to converge to optimal solution. This poor performance of PDA-MD when it is not warm started is also observed in our experiments.

The work of Nasr et al. (2023) used public data to train a generative model for data augmentation and to estimate the center of the clipping balls in DP-SGD. They did not provide any code for their experiments and thus we do not compare against them as a baseline in our experiments. However, combining our algorithm with the tricks used in (Nasr et al., 2023) could be a promising avenue for future empirical work.

Personalized DP: A related line of work is that of *personalized DP* (PDP) (Jorgensen et al., 2015; Golatkar et al., 2022; Mühl & Boenisch, 2022; Fallah et al., 2022), a generalization of DP in which each person may have different privacy parameters $(\varepsilon_i, \delta_i)$. By letting $\varepsilon_i = \infty$ for some person i, PDP also generalizes semi-DP. We leverage this connection to borrow techniques from the work of Fallah et al. (2022), which considers pure ($\delta_i = 0$) PDP estimation in one dimension ($d = 1$). We also note that the 1-dimensional pure (central) PDP mean estimation bound of Fallah et al. (2022) extends easily to a 1-dimensional ε-semi-DP bound. However, our d-dimensional semi-DP lower bounds require a different set of techniques. Additionally, the personalized LDP bound of Fallah et al. (2022) relies on the assumption $\varepsilon_i \leqslant 1$ and does not seem to extend to the ε-semi-LDP setting.

Concurrent and Subsequent Work: The work of Ullah et al. (2024) first appeared on arXiv a few weeks after v2 of our paper appeared.[7] (Ullah et al., 2024) proves results that are very similar to Theorems 4 and 12. However, their lower bound proofs do not require the (mild) symmetry assumption that our proof requires: see Remark 5. Moreover, in a restricted parameter regime, Ullah et al. (2024) also provide lower bounds that are tighter by a $\log(1/\delta)$ factor. Ullah et al. (2024) complement these lower bounds by giving novel algorithms for leveraging unlabeled public data in training private generalized linear models (GLM) with dimension-independent rates.

The work of Liu et al. (2023), which appeared on arXiv 3 months after v1 of our paper, couples private and public gradients in non-convex optimization via a similar weighting scheme to our own. They show benefits of their algorithm over standard DP approaches both theoretically and empirically.

[7] The first version of our paper appeared on arXiv more than eight months before (Ullah et al., 2024). However, v1 of our paper did not contain our tight high-dimensional approximate semi-DP population mean estimation and SCO lower bounds.

Tang et al. (2023), which appeared on arXiv 18 days before v1 of our paper, explores how to improve the privacy-utility tradeoff of DP-SGD by learning priors from images generated by random processes and transferring these priors to private data.

Other Related Works: The work of Gu et al. (2023) gave an algorithm for selecting an appropriate public dataset that can be used to enhance private optimization by projecting gradients onto a subspace prescribed by the this public datasaet.

Ganesh et al. (2023) provided an explanation for the empirically reported benefits of pre-training on public data, arguing that non-convex optimization algorithms must go through two phases: (i) selecting a good "basin" in the loss landscape; (ii) solving an easy optimization problem within that basin. They hypothesize that public pre-training can be helpful in selecting a good basin. They also demonstrated a separation between pretrained and non-pretrained models by constructing a non-convex optimization problem for which public pretraining is necessary to achieve non-trivial error.

B. Zero-Concentrated DP vs. Pure and Approximate DP

Proposition 20. *(Bun & Steinke, 2016) If $\mathcal{A}$ is ρ-zCDP, then $\mathcal{A}$ is $(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta)$-DP $\forall \delta > 0$. Moreover, if $\mathcal{A}$ is ε-DP, then $\mathcal{A}$ is $\varepsilon^2/2$-zCDP.*

C. Summary table of pure semi-DP and semi-LDP results

Learning problem	Semi-DP error	When is semi-DP error less than DP?
Mean Estimation (Pop. MSE)	$\min\left\{\frac{1}{n_{pub}}, \frac{d^2}{n^2\varepsilon^2} + \frac{1}{n}\right\}$ (Theorem 32)	$n_{\text{pub}} > \frac{n^2\varepsilon^2}{d^2}$ or $n_{\text{pub}} = \Theta(n)$
ERM (Excess emp. Risk)	$\min\left\{\frac{n_{priv}}{n}, \frac{d}{n\varepsilon}\right\}$ (Theorem 10)	$n_{\text{pub}} > n - d/\varepsilon$
SCO (Excess pop. risk)	$\min\left\{\frac{1}{\sqrt{n_{pub}}}, \frac{d}{n\varepsilon} + \frac{1}{\sqrt{n}}\right\}$ (Theorem 42)	$n_{\text{pub}} > \frac{n^2\varepsilon^2}{d^2}$ or $n_{\text{pub}} = \Theta(n)$

Learning problem	Semi-LDP error	When is semi-LDP error less than LDP?
Mean Estimation (Pop. MSE)	$\min\left\{\frac{1}{n_{pub}}, \frac{d}{n\varepsilon^2}\right\}$ (Theorem 14)	$n_{\text{pub}} > \frac{\varepsilon^2 n}{d}$ or $n_{\text{pub}} = \Theta(n)$
SCO (Excess pop. risk)	$\min\left\{\frac{1}{\sqrt{n_{pub}}}, \sqrt{\frac{d}{n\varepsilon^2}}\right\}$ (Theorem 18)	$n_{\text{pub}} > \frac{\varepsilon^2 n}{d}$ or $n_{\text{pub}} = \Theta(n)$

Figure 7. Minimax optimal error rates for central ε-semi-DP and (local) ε-semi-LDP SCO and mean estimation results. Dependence on range and Lipschitz parameters, and constraint set diameter omitted. Mean estimation and SCO lower bounds are only tight if $n_{\text{pub}} = O(n\varepsilon/d)$ or $d = O(1)$. Strongly convex ERM and SCO results are included later in this Appendix, but excluded from this table.

D. Notation

We recall notation from the main body and include some additional basic definitions below for convenience.

Let $\|\cdot\|$ be the ℓ_2 norm. $\mathcal{W}$ denotes a convex, compact subset of $\mathbb{R}^d$ with ℓ_2 diameter D. $\mathcal{X}$ denotes a data universe. Function $g : \mathcal{W} \to \mathbb{R}$ is *μ-strongly convex* if $g(\alpha w + (1-\alpha)w') \leqslant \alpha g(w) + (1-\alpha)g(w') - \frac{\alpha(1-\alpha)\mu}{2}\|w - w'\|^2$ for all $\alpha \in [0,1]$ and all $w, w' \in \mathcal{W}$. If $\mu = 0$, we say g is *convex*. For convex $f(\cdot, x)$, denote any *subgradient* of $f(w,x)$ w.r.t. w by $\nabla f(w,x) \in \partial_w f(w,x)$: i.e. $f(w',x) \geqslant f(w,x) + \langle \nabla f(w,x), w' - w\rangle$ for all $w' \in \mathcal{W}$. Function $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ is *uniformly L-Lipschitz* in w if $\sup_{x\in\mathcal{X}} |f(w,x) - f(w',x)| \leqslant L\|w - w'\|$. Let $\mathbb{B} = \{x \in \mathbb{R}^d | \|x\| \leqslant 1\}$ denote the unit ℓ_2-ball. For functions $a = a(\theta)$ and $b = b(\phi)$ of input parameter vectors θ and ϕ, we write $a \lesssim b$ or $a = O(b)$ if there is an absolute constant $C > 0$ such that $a \leqslant Cb$ for all values of input parameter vectors θ and ϕ.

E. Optimal Centrally Private Model Training with Public Data

E.1. Optimal Semi-DP Mean Estimation

We begin in Appendix E.1.1 with empirical mean estimation. This subsection was omitted from the main body due to space constraints. Theorem 21 will be useful for proving our ERM bounds (Theorem 10). Then, in Appendix E.1.2, we turn to population mean estimation (i.e. the proof of Theorem 4). Appendix E.1.3 contains proofs for Section 2.2.

E.1.1. Estimating the Empirical Mean

For a given data set $X \subset \mathbb{B} := \{x \in \mathbb{R}^d : \|x\| \leqslant 1\}$, consider the problem of estimating $\bar{X} = \frac{1}{n}\sum_{i=1}^n x_i$ subject to the constraint that the estimator satisfies semi-DP. We will characterize the minimax squared error of d-dimensional empirical mean estimation under ε-semi-DP:

$$\mathcal{M}_{\text{emp}}(\varepsilon, n_{\text{priv}}, n, d) := \inf_{\mathcal{A}\in\mathbb{A}_\varepsilon(\mathbb{B})} \sup_{X\in\mathbb{B}^n, |X_{\text{priv}}|=n_{\text{priv}}} \mathbb{E}_{\mathcal{A}}\left[\|\mathcal{A}(X) - \bar{X}\|^2\right], \tag{14}$$

where $\mathbb{A}_\varepsilon(\mathbb{B})$ denotes the set of all ε-semi-DP estimators $\mathcal{A} : \mathbb{B}^n \to \mathbb{B}$.

Theorem 21. *Let $\varepsilon > 0$, $n, d \in \mathbb{N}$, $n_{priv} \in [n]$. There exist absolute constants c and C with $0 < c \leqslant C$ such that*

$$c \min\left\{\frac{n_{priv}}{n}, \frac{d}{n\varepsilon}\right\}^2 \leqslant \mathcal{M}_{emp}(\epsilon, n_{priv}, n, d) \leqslant C \min\left\{\frac{n_{priv}}{n}, \frac{d}{n\varepsilon}\right\}^2.$$

Proof. **Lower bound:** We use a packing argument to prove our lower bound. Denote $\mathcal{X} = \left[\frac{-1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\right]^d$. Let $\mathcal{A} : \mathbb{B}^n \to \mathbb{B}$ be ε-semi-DP. Choose $K = 2^{d/2}$ private data points $\{x_i\}_{i=1}^K \subset \left\{\pm\frac{1}{\sqrt{d}}\right\}^d$ such that $\|x_i - x_j\| \geqslant 1/8$ for all $i \neq j$. The existence of such a set of points is well-known (see e.g. the Gilbert-Varshamov construction). Let $n^* = \frac{nd}{3\varepsilon n_{\text{priv}}}$.

Case 1: $n \leqslant n^*$ (i.e. $d \geqslant 3\varepsilon n_{\text{priv}}$). In this case, we'll show that $\mathbb{E}\|\mathcal{A}(X) - \bar{X}\|^2 \gtrsim \left(\frac{n_{\text{priv}}}{n}\right)^2$ for some $X \in \mathcal{X}^n$. For $i \in [K]$, let $X_i = (x_i, \cdots, x_i, \mathbf{0}_{n-n_{\text{priv}}})$ consist of n_{priv} copies of x_i followed by $n - n_{\text{priv}} = n_{\text{pub}}$ copies of $0 \in \mathbb{R}^d$. Suppose for the sake of contradiction that for every $i \in [K]$, with probability $\geqslant 1/3$ we have $\|\mathcal{A}(X) - \bar{X}\| < \frac{1}{32}\frac{n_{\text{priv}}}{n}$. That is, we are supposing $\mathbb{P}(\mathcal{A}(X_i) \in \mathcal{B}_i) \geqslant 1/3$ for all i, where $\mathcal{B}_i = \{x \in \mathbb{B} : \|x - \bar{X}_i\| \leqslant \frac{1}{32}\frac{n_{\text{priv}}}{n}\}$. Note that the sets $\{\mathcal{B}_i\}_{i=1}^K$ are disjoint by construction. Since $\mathcal{A}$ is ε-semi-DP, group privacy implies that $\mathbb{P}(\mathcal{A}(X_1) \in \mathcal{B}_i) \geqslant \frac{1}{3}e^{-\varepsilon n_{\text{priv}}}$ for all $i \in [K]$. Thus,

$$K\frac{1}{3}e^{-\varepsilon n_{\text{priv}}} \leqslant \sum_{i=1}^K \mathbb{P}(\mathcal{A}(X_1) \in \mathcal{B}_i) \leqslant 1,$$

where the last inequality follows from disjointness of the balls $\mathcal{B}_i$. Thus, we obtain $\ln(K/3) \leqslant \varepsilon n_{\text{priv}}$. Assume for now that $d \geqslant 8$. (A 1-dimensional lower bound that is tight up to constant factors can be shown easily by following the proof of the d-dimensional case but choosing $K = 16$ instead of $K = 2^{d/2}$.) Then $d/2 \leqslant d - 4 \leqslant \varepsilon n_{\text{priv}}$ implies $d \leqslant 2\varepsilon n_{\text{priv}}$, contradicting the assumption made in Case 1. Thus, we conclude that there exists a data set X_i such that with probability at least $2/3$, $\|\mathcal{A}(X_i) - \bar{X}_i\| > \frac{1}{32}\frac{n_{\text{priv}}}{n}$. Squaring both sides of this inequality and applying Markov's inequality yields the desired lower bound.

Case 2: $n > n^*$.

Additionally, suppose for now that $n^* \leqslant n_{\text{priv}}$. In this case, we'll show that $\mathbb{E}\|\mathcal{A}(X) - \bar{X}\|^2 \gtrsim \left(\frac{d}{n\varepsilon}\right)^2$ for some $X \in \mathcal{X}^n$. Let $\widetilde{X}_i = (x_i, \cdots, x_i, \mathbf{0}_{n-n^*})$ consist of n^* copies of x_i followed by $n - n^*$ copies of $0 \in \mathbb{R}^d$.[8] Denoting the mean of a dataset X by $q(X) := \frac{1}{n}\sum_{x\in X} x$ for convenience, we see that $q(\widetilde{X}_i) = \frac{n^*}{n}x_i = \frac{d}{3n_{\text{priv}}\varepsilon}x_i$. Define the algorithm $\hat{\mathcal{A}} : \mathbb{B}^{n^*} \to \mathbb{B}$ by $\hat{\mathcal{A}}(X) = \frac{n}{n^*}\mathcal{A}(X, \mathbf{0}_{n-n^*})$. Since $\mathcal{A}$ is ε-semi-DP, we see that $\hat{\mathcal{A}}$ is ε-semi-DP by post-processing. Also, the domain of $\hat{\mathcal{A}}$ is $\mathcal{X}^{n^*}$ and $n^* \leqslant n^*$, so the argument in Case 1 applies to $\hat{\mathcal{A}}$. Thus, by applying the result in Case 1, there exists $i \in [K]$ such that with probability at least $2/3$, $\|\hat{\mathcal{A}}(x_i, \cdots, x_i) - q(x_i, \cdots, x_i)\| \geqslant \frac{1}{32}\frac{n_{\text{priv}}}{n}$. (Here $(x_i, \cdots, x_i) \in \mathcal{X}^{n*}$.) But this implies $\|\mathcal{A}(\widetilde{X}_i) - q(\widetilde{X}_i)\| \geqslant \frac{1}{32}\frac{n_{\text{priv}}}{n}\frac{n^*}{n} = \frac{1}{96}\frac{d}{n\varepsilon}$ with probability at least $2/3$. Again, squaring both sides and applying Markov yields the desired lower bound.

Next, consider the complementary subcase where $n^* > n_{\text{priv}}$. Define the algorithm $\hat{\mathcal{A}} : \mathbb{B}^{n^*} \to \mathbb{B}$ by $\hat{\mathcal{A}}(X) = \mathcal{A}(X, \mathbf{0}_{n-n^*})$. Since $\mathcal{A}$ is ε-semi-DP, we see that $\hat{\mathcal{A}}$ is ε-semi-DP by post-processing. Also, the domain of $\hat{\mathcal{A}}$ is $\mathcal{X}^{n^*}$ and $n^* \leqslant n^*$, so the argument in Case 1 applies to $\hat{\mathcal{A}}$. Thus, by applying the result in Case 1, there exists $i \in [K]$ such that with probability at least $2/3$, $\|\hat{\mathcal{A}}(x_i, \cdots, x_i) - q(x_i, \cdots, x_i)\| \geqslant \frac{1}{32}\frac{n_{\text{priv}}}{n}$. (Here $(x_i, \cdots, x_i) \in \mathcal{X}^{n*}$.) But this implies that $\|\mathcal{A}(X_i) - \bar{X}_i\| \geqslant \frac{1}{32}\frac{n_{\text{priv}}}{n}$

[8] We assume without loss of generality that $n^* \in \mathbb{N}$. If n^* is not an integer, then choosing $\lceil n^* \rceil$ instead yields the same bound up to constant factors.

with probability at least $2/3$. Again, squaring both sides and applying Markov yields the desired lower bound. Thus, the lower bound holds in all cases.

Upper bound: For the first term in the minimum, consider the algorithm which throws away the private data and outputs $\mathcal{A}(X) = \frac{1}{n}\sum_{x\in X_{\text{pub}}} x$. Clearly, $\mathcal{A}$ is semi-DP. Moreover,

$$\|\mathcal{A}(X) - \bar{X}\|^2 = \frac{1}{n^2}\left\|\sum_{x\in X_{\text{priv}}} x\right\|^2 \leqslant \frac{n_{\text{priv}}}{n^2}\sum_{x\in X_{\text{priv}}}\|x\|^2 \leqslant \left(\frac{n_{\text{priv}}}{n}\right)^2.$$

For the second term, consider the Laplace mechanism $\mathcal{A}(X) = \bar{X} + (L_1, \cdots, L_d)$, where $L_i \sim \text{Lap}(2\sqrt{d}/n\varepsilon)$ are i.i.d. mean-zero Laplace random variables. We know $\mathcal{A}$ is ε-DP by (Dwork et al., 2014), since the ℓ_1-sensitivity is $\sup_{X\sim X'}\|\bar{X} - \bar{X}'\|_1 = \frac{1}{n}\sup_{x,x'}\|x - x'\|_1 \leqslant \frac{2\sqrt{d}}{n}$. Hence $\mathcal{A}$ is ε-semi-DP. Moreover, $\mathcal{A}$ has error

$$\mathbb{E}\|\mathcal{A}(X) - \bar{X}\|^2 = d\,\text{Var}(\text{Lap}(2\sqrt{d}/n\varepsilon)) = \frac{8d^2}{n^2\varepsilon^2}.$$

Combining the two upper bounds completes the proof. □

E.1.2. Estimating the Population Mean

Approximate (ε, δ)-Semi-DP Mean Estimation

Theorem 22 (Formal statement of Theorem 4)**.** *Let $\varepsilon \lesssim 1/\log(nd)$, $\delta \ll 1/n_{priv}$. Then, there is an absolute constant $C > 0$ such that*

$$\ell(d,n)\min\left\{\frac{1}{n_{pub}}, \frac{d}{n^2\varepsilon^2} + \frac{1}{n}\right\} \leqslant \mathcal{M}_{pop}(\epsilon, \delta, n_{priv}, n, d) \leqslant C\min\left\{\frac{1}{n_{pub}}, \frac{d\ln(1/\delta)}{n^2\varepsilon^2} + \frac{1}{n}\right\}, \tag{15}$$

where $1/\ell(d,n)$ is logarithmic in d and n. The lower bound holds for symmetric algorithms $\mathcal{A} = (\mathcal{A}^1, \ldots, \mathcal{A}^d)$ such that $\mathcal{A}^j = \mathcal{A}^l$ for all $j, l \in [d]$.

For the lower bound, we will first prove a stronger result in Theorem 23, in which we construct a hard distribution whose mean is small—scaling with the accuracy lower bound that we aim to prove. This "small mean" property will be needed for our semi-DP SCO lower bound (Theorem 12), even though it is not necessary for the proof of Theorem 22.

Theorem 23. *Let $\mathcal{X} = \left\{\pm\frac{1}{\sqrt{d}}\right\}^d$, $\varepsilon \lesssim 1/\log(nd)$, and $\delta \ll 1/n_{priv}$. Then, for any symmetric (ε, δ)-semi-DP $\mathcal{A}$, there exists a product distribution P on $\mathcal{X}$ with $\|\mathbb{E}_{x\sim P}[x]\| \leqslant \min\left(\frac{1}{\sqrt{n_{pub}}}, \frac{\sqrt{d}}{\varepsilon n_{priv}}\right)$ such that*

$$\mathbb{E}\left[\|\mathcal{A}(X) - \mathbb{E}_{x\sim P}[x]\|^2\right] = \widetilde{\Omega}\left(\min\left\{\frac{1}{n_{pub}}, \frac{d}{\varepsilon^2 n_{priv}^2} + \frac{1}{n}\right\}\right).$$

For any symmetric $\mathcal{A}$ and any distribution P with $X \sim P^n$, we have

$$\begin{aligned}\mathbb{E}\|\mathcal{A}(X) - \mathbb{E}_{x\sim P}[x]\|^2 &= \mathbb{E}\|\mathcal{A}(X) - \mathbb{E}\mathcal{A}(X)\|^2 + \|\mathbb{E}\mathcal{A}(X) - \mathbb{E}_{x\sim P}[x]\|^2 \\ &\geqslant \|\mathbb{E}\mathcal{A}(X) - \mathbb{E}_{x\sim P}[x]\|^2 \\ &= d|\mathbb{E}\mathcal{A}^1(X) - \mathbb{E}[x^1]|,\end{aligned} \tag{16}$$

where the last equality used assumption that $\mathcal{A}$ is symmetric. For $a \in \mathbb{R}$, scalar random variable $x^j \sim P_a$ with mean $\mathbb{E}[x^j] = a$ and $X^j \sim P_a^n$ denote

$$\text{Bias}_a(\mathcal{A}) := |\mathbb{E}\mathcal{A}^j(X^j) - a|.$$

Definition 24 (Low bias algorithms)**.** *We say symmetric $\mathcal{A}$ is low bias if for every $a \in \mathbb{R}$,*

$$\textit{Bias}_a(\mathcal{A})^2 \leqslant \frac{1}{d}\min\left(\frac{1}{n_{pub}}, \frac{1}{n} + \frac{d}{n^2\varepsilon^2}\right).$$

We can assume without loss of generality that $\mathcal{A}$ is low bias when we are proving the lower bound in Theorem 22: if $\mathcal{A}$ is not low bias, then inequality (16) implies that the worst-case MSE of $\mathcal{A}$ is lower bounded as in (15).

To prove Theorem 23 for low bias and symmetric $\mathcal{A}$, we will use Theorem 25.[9] This result shows that any sufficiently accurate $\mathcal{A}$ is vulnerable to an attack that traces many individuals in the data set with high probability.

Theorem 25. *Let $a \in (0, 1]$. Consider the product distribution on $\{\pm 1\}^d$ defined in the following way: for $j \in [d]$, independently draw $\theta^j \sim$ **Unif**$([-a, a])$ and $x_i^j \sim P_\theta$ such that $x_i^j \in \{\pm 1\}$ with mean $\mathbb{E}_{x_i^j \sim P_{\theta^j}}[x_i^j] = \theta^j$ for $i \in [n]$. Denote $X = (x_1, \ldots, x_n) \sim P_\theta^n$ where $\theta = (\theta^1, \ldots, \theta^d)$ and $P_\theta = \Pi_{j=1}^d P_{\theta^j}$. Let $\mathcal{A} : \{\pm 1\}^d \to [-a, a]^d$ satisfy $\mathbb{E}_{X \sim P_\theta^n}[\mathcal{A}(X)] = \theta$ and $\mathbb{E}\|\mathcal{A}(X) - \theta\|^2 \leqslant \alpha^2$ for $\sqrt{\frac{d}{n}} \leqslant \alpha \leqslant \frac{a\sqrt{d}}{100}$. Assume $d > 400\alpha n\sqrt{\ln(2/\delta)}$. Moreover, assume that $\mathcal{A}(X) = (\mathcal{A}^1(X), \ldots, \mathcal{A}^d(X))$ with $\mathcal{A}^j = \mathcal{A}^l$ for all $j, l \in [d]$. Then, the attack $\mathcal{I} : \{\pm 1\}^d \times [-a, a]^d \to \{IN, OUT\}$ described in Algorithm 2 satisfies the following properties: a) if $y \sim P_\theta$ independently of X, then $P(\mathcal{I}(y, \widetilde{\mathcal{A}}(X)) = IN) \leqslant \delta$; and b) $P(|\{i \in [n] : \mathcal{I}(x_i, \widetilde{\mathcal{A}}(X)) = IN\}| \geqslant \frac{d}{10^6\alpha^2}) \geqslant 1 - \delta$, where $\widetilde{\mathcal{A}}$ is $(\widetilde{O}(\varepsilon), \widetilde{O}(\zeta))$-semi-DP if and only if $\mathcal{A}$ is (ε, ζ)-semi-DP, and $\mathbb{E}_{X \sim P_\theta^n}[\widetilde{\mathcal{A}}(X)] = \theta$.*

The proof of Theorem 25 uses a convenient reduction in Lemma 27: for any low bias and symmetric $\mathcal{A}$ that has small expected mean squared error (in ℓ_2-norm), there exists another low bias mechanism $\widetilde{\mathcal{A}}$ that has even smaller ℓ_∞ accuracy with high probability. Moreover, $\mathcal{A}$ is (ε, δ)-semi-DP if and only if $\widetilde{\mathcal{A}}$ is $(\widetilde{O}(\varepsilon), \widetilde{O}(\delta))$-semi-DP. This lemma allows us to: construct a new product distribution whose mean is small, modify the attack of Dwork et al. (2015), and derive a generalized fingerprinting lemma (Lemma 26) in order to prove Theorem 25.

With Theorem 25 in hand, we leverage the proof technique of Bassily et al. (2020) to show that any sufficiently accurate $\widetilde{\mathcal{A}}$ must leak the data of more than n_{pub} individuals and thus cannot be $(\widetilde{O}(\varepsilon), \widetilde{O}(\delta))$-semi-DP. But by Lemma 27, this means that $\mathcal{A}$ cannot be (ε, δ)-semi-DP. Below, we discuss the attack that we use and then fill in the details of the proof.

The attack in Algorithm 2 is a modification of the robust tracing attack of Dwork et al. (2015). The attacker in Algorithm 2 receives as input the output $q \sim \mathcal{A}(X)$ of a mechanism, a target point y, and the mean of the data distribution P, $\theta = \mathbb{E}_{x \sim P}[x]$. (An estimate of the true mean or access to sufficiently many independent draws from P would also suffice in lieu of θ.) The target y is either a data point used by $\mathcal{A}(X)$ (i.e. $y \in X$) or an independent draw from the distribution P that X was drawn from. The attacker aims to infer whether or not y was in X. If the attacker outputs IN when y is in X and OUT when y is not in X (i.e. the attack succeeds) with high probability, then the algorithm $\mathcal{A}$ is not private. The truncation parameter is $\eta = 2\alpha/\sqrt{d}$, where α is the expected ℓ_2-error of the mechanism $\mathcal{A}$. The parameter δ can be chosen by the attacker.

Algorithm 2 Tracing Attack Against ℓ_2-Accurate Mechanisms

Input: Target $y \in \{\pm 1\}^d$, mean $\theta \in [-a, a]^d$, output of mechanism $q \sim \mathcal{A}(X)$.
Let $\eta := 2\alpha/\sqrt{d}$.
Let $\lfloor q - \theta \rceil_\eta \in [-\eta, \eta]^d$ denote the entrywise truncation of $q - \theta$.
Compute $\langle y - \theta, \lfloor q - \theta \rceil_\eta \rangle = \sum_{j=1}^d (y^j - \theta^j) \lfloor q - \theta \rceil_\eta^j$.
if $\langle y - \theta, \lfloor q - \theta \rceil_\eta \rangle > \tau := 2\eta\sqrt{d \ln(1/\delta)}$ **then**
    Output: IN.
else
    Output: OUT.
end if

Compared to (Dwork et al., 2015), we choose a smaller truncation parameter to handle the ℓ_2 case ($\eta = 2\alpha/\sqrt{d}$ instead of 2α). We also scale the "IN/OUT" decision threshold τ accordingly. Moreover, since we are not concerned with the size of the reference sample, we assume that the attacker knows the mean of the data distribution. Thus, we use θ in place of the average of reference samples to give a simpler attack than the one in (Dwork et al., 2015). This is not necessary for our attack to work: the attacker just needs a sufficiently close approximation to the true mean (which can be attained with access to enough reference samples) for our proof to go through.

Theorem 25 builds on (Dwork et al., 2015, Theorem 17), which gave a similar result for ℓ_∞-accurate mechanisms and θ drawn from a so-called *strong distribution* (see (Dwork et al., 2015, Definition 5)). By contrast, Theorem 25 only requires a

[9] For convenience, we work with data drawn from $\{\pm 1\}^d$ and then re-scale to obtain the final mean estimation lower bound.

weaker ℓ_2-accuracy guarantee and uses a distribution which is not "strong" in the sense required by (Dwork et al., 2015, Theorem 17). Our generalization of the *fingerprinting lemma* (Bun et al., 2017), given below, allows us to handle such a distribution, and is the first step towards proving Theorem 25:

Lemma 26 (Generalized Fingerprinting Lemma). *Let $a \in (0,1]$. Draw $\theta \sim$ **Unif**$([-a,a])$ and $x_i \sim P_\theta$ be such that $x_i \in \{\pm 1\}$ with mean $\mathbb{E}_{x_i \sim P_\theta}[x_i] = \theta$ for $i = 1, \dots, n$. Denote $X = (x_1, \dots, x_n) \sim P_\theta^n$. Then, for any $f : \{\pm 1\}^n \to \mathbb{R}$ that does not depend directly on θ, we have*

$$\mathbb{E}\left[(f(X)-\theta)\sum_{i=1}^n (x_i - \theta)\right] \geqslant \frac{a^2}{3} - \mathbb{E}\left[(f(X)-\theta)^2\right] + \frac{1-a^2}{2a}\left[\mathbb{E}_{X\sim P_a^n}[f(X)] - \mathbb{E}_{X\sim P_{-a}^n}[f(X)]\right],$$

where all expectations are over the random draw of θ and $X \sim P_\theta^n$ unless otherwise indicated.

Proof. Define $g : [-a,a] \to \mathbb{R}$ by $g(\theta) := \mathbb{E}_{X\sim P_\theta^n}[f(X)]$. Our first claim is

$$\mathbb{E}_{X\sim P_\theta^n}[(f(X)-\theta)\sum_{i=1}^n (x_i - \theta)] = g'(\theta)(1-\theta^2). \tag{17}$$

To prove (17), write

$$g(\theta) = \sum_{X\in\{\pm 1\}^n} f(X) \prod_{i=1}^n \frac{1+(\theta/x_i)}{2}$$

and

$$\begin{aligned}
g'(\theta) &= \sum_{X\in\{\pm 1\}^n} f(X) \frac{d}{d\theta}\left[\prod_{i=1}^n \left(\frac{1+(\theta/x_i)}{2}\right)\right] \\
&= \sum_{X\in\{\pm 1\}^n} f(X) \sum_{i=1}^n \frac{d}{d\theta}\left(\frac{1+(\theta/x_i)}{2}\right) \prod_{k\in[n]\setminus\{i\}} \left(\frac{1+(\theta/x_k)}{2}\right) \\
&= \sum_{X\in\{\pm 1\}^n} f(X) \sum_{i=1}^n \frac{1}{x_i+\theta} \prod_{i\in[n]} \left(\frac{1+(\theta/x_i)}{2}\right) \\
&= \mathbb{E}_{X\sim P_\theta^n}\left[f(X)\sum_{i=1}^n \frac{1}{x_i+\theta}\right] \\
&= \mathbb{E}_{X\sim P_\theta^n}\left[f(X)\sum_{i=1}^n \frac{x_i-\theta}{1-\theta^2}\right].
\end{aligned}$$

Since $\mathbb{E}_{X\sim P_\theta^n}\left[\theta\sum_{i=1}^n (x_i-\theta)\right] = 0$, (17) is proved.

Next, we claim: for any differentiable function $g : [-a,a] \to \mathbb{R}$, we have

$$\mathbb{E}_{\theta\sim\textbf{Unif}[-a,a]}[g'(\theta)(1-\theta^2)] = 2\mathbb{E}_{\theta\sim\textbf{Unif}[-a,a]}[g(\theta)\theta] + \frac{1-a^2}{2a}[g(a)-g(-a)]. \tag{18}$$

To prove (18), let $u(\theta) = 1-\theta^2$ and use the product rule and fundamental theorem of calculus to write

$$\begin{aligned}
\mathbb{E}_{\theta\sim\textbf{Unif}[-a,a]}[g'(\theta)(1-\theta^2)] &= \frac{1}{2a}\int_{-a}^a g'(\theta)u(\theta)d\theta \\
&= \frac{1}{2a}\int_{-a}^a \left[\frac{d}{d\theta}(g(\theta)u(\theta)) - g(\theta)u'(\theta)\right]d\theta \\
&= \frac{1}{2a}[g(a)u(a) - g(-a)u(-a)] - \frac{1}{2a}\int_{-a}^a g(\theta)\cdot(-2\theta)d\theta \\
&= \frac{1-a^2}{2a}[g(a)-g(-a)] + 2\mathbb{E}_{\theta\sim\textbf{Unif}[-a,a]}[g(\theta)\theta].
\end{aligned}$$

Now we will apply (17) and (18) with the differentiable function $g : [-a, a] \to \mathbb{R}$, $g(\theta) := \mathbb{E}_{X \sim P_\theta^n}[f(X)]$ to obtain

$$\mathbb{E}\left[(f(X) - \theta)\sum_{i=1}^n (x_i - \theta)\right] = \mathbb{E}_{\theta \sim \textbf{Unif}[-a,a]}[g'(\theta)(1-\theta^2)]$$
$$= 2\mathbb{E}_{\theta \sim \textbf{Unif}[-a,a]}[\theta g(\theta)] + \frac{1-a^2}{2a}[g(a) - g(-a)].$$

Moreover,

$$\mathbb{E}[(f(X) - \theta)^2] \geqslant -2\mathbb{E}[g(\theta)\theta] + \mathbb{E}[\theta^2]$$
$$= -2\mathbb{E}[g(\theta)\theta] + \frac{a^2}{3}.$$

Combining the above pieces completes the proof. □

We state a lemma that will allow us to conveniently assume without loss of generality that the given mechanism $\mathcal{A}$ is ℓ_∞-accurate:

Lemma 27. *Consider $\theta \sim$ **Unif**$([-a, a]^d)$ and $X \sim P_\theta^n$ be distributed as described above. Suppose $\mathbb{E}\|\mathcal{A}(X) - \theta\|^2 \leqslant \alpha^2$, where $\mathcal{A} : \{\pm 1\}^d \to [-a, a]^d$, $\mathcal{A}(X) = (\mathcal{A}^1(X) \ldots, \mathcal{A}^d(X))$. Assume that $\mathcal{A}(X)$ is a low bias estimator of θ and that $\mathcal{A}^j = \mathcal{A}^l$ for all j, l. Then, for any $0 < \beta \leqslant \min\left(\alpha^2/4d, 1/(n^2d^2)\right)$, there exists a low bias $\widetilde{\mathcal{A}}$ such that $P(|\widetilde{\mathcal{A}}^j(X) - \theta^j| > \frac{2\alpha}{\sqrt{d}}) \leqslant \beta$, $\widetilde{\mathcal{A}}^j = \widetilde{\mathcal{A}}^l$ for all $j, l \in [d]$, and $\mathbb{E}\|\widetilde{\mathcal{A}}(X) - \theta\|^2 \leqslant 5\alpha^2$. Also, $\mathcal{A}$ is (ε, δ)-semi-DP if and only if $\widetilde{\mathcal{A}}$ is $(\widetilde{O}(\varepsilon), \widetilde{O}(\delta))$-semi-DP.*

Proof. By the assumption that $\mathcal{A}^j = \mathcal{A}^l$ for all j, l and the fact that X is drawn from a product distribution, we have

$$\mathbb{E}\|\mathcal{A}(X) - \theta\|^2 = d\mathbb{E}|\mathcal{A}^1(X) - \theta^1|^2.$$

Thus, $\mathbb{E}\|M^j(X) - \theta^j|^2 \leqslant \frac{\alpha^2}{d}$ for all j. (We are also assuming without loss of generality here that $\mathcal{A}^j$ depends only on X^j: if this is not the case, then it's easy to see by the choice of distribution that there exists another algorithm with smaller error than $\mathcal{A}$ satisfying this assumption.) Also,

$$P\left(|\mathcal{A}^j(X) - \theta^j| > \frac{2\alpha}{\sqrt{d}}\right) \leqslant \frac{\mathbb{E}|\mathcal{A}^j(X) - \theta^j|^2}{(2\alpha/\sqrt{d})^2} \leqslant \frac{1}{4}, \tag{19}$$

by Chebyshev's inequality. To construct $\widetilde{\mathcal{A}}^j$ we will use the well-known "median trick": run each $\mathcal{A}^j$ for $m = O(\log(1/\beta))$ times, each time using a batch X_i of n/m independent samples in X. Then take the median of the $m = O(\log(1/\beta))$ outputs: $\widetilde{\mathcal{A}}^j(X) = \text{median}(\mathcal{A}_1^j(X_1), \ldots, \mathcal{A}_m^j(X_m))$. Then, $P(|\widetilde{\mathcal{A}}^j(X) - \theta^j| > \frac{2\alpha}{\sqrt{d}}) \leqslant \beta$ by a Chernoff/Hoeffding bound. Applying the law of total expectation with $\beta \leqslant \min\left(\alpha^2/4d, 1/(n^2d^2)\right)$ establishes the expected ℓ_2-accuracy claim for $\widetilde{\mathcal{A}}$. Moreover, $\widetilde{\mathcal{A}}$ is low bias since for all $j \in [d]$, we have $\widetilde{\mathcal{A}}^j(X) \in \mathcal{A}_i^j(X_i)$ for some $i \in [m]$ with probability 1 and $\mathcal{A}$ is low bias.

We now prove the privacy claim. First note that $\mathcal{A}$ is (ε, δ)-semi-DP if and only if $\mathcal{A}^j$ is $(\widetilde{O}(\varepsilon/\sqrt{d}), O(\delta/d))$-semi-DP by the advanced composition theorem (Dwork et al., 2014), since $\mathcal{A}^j = \mathcal{A}^l$ for all $j, l \in [d]$. Also, $\mathcal{A}^j$ is (ε', δ')-semi-DP if and only if $\widetilde{\mathcal{A}}^j$ is $(\widetilde{O}(\varepsilon'), \widetilde{O}(\delta'))$-semi-DP by parallel composition of (semi) DP (McSherry, 2009) and the fact that $m = \widetilde{O}(1)$. Since $\widetilde{\mathcal{A}}^j = \widetilde{\mathcal{A}}^l$ for all j, l by construction, we can conclude that $\mathcal{A}$ is (ε, δ)-semi-DP if and only if $\widetilde{\mathcal{A}}$ is $(\widetilde{O}(\varepsilon), O(\delta))$-semi-DP. □

By Lemma 27 and the preceding discussion, *it suffices to assume that the given mechanism $\mathcal{A}$ is low bias, $2\alpha/\sqrt{d}$-ℓ_∞-accurate with probability at least $1 - \beta$ for any $\beta > 0$, $\mathcal{A}^j = \mathcal{A}^l$ for all j, l, and that $\mathcal{A}$ has expected ℓ_2-mean-squared-error bounded by α^2, for the remainder of the proof* of Theorem 25. We also assume in what follows that θ is drawn uniformly at random from $[-a, a]^d$ and that conditional on θ, the data X is drawn i.i.d. from the product distribution P_θ, as described in the statement of Theorem 25.

Now, we will use Lemma 26 to lower bound the sum of expected scores $\sum_{i=1}^n \mathbb{E}\langle x_i - \theta, \lfloor q - \theta \rceil_\eta \rangle$:

Lemma 28. *Let* $1 \geqslant a > 100\alpha/\sqrt{d}$. *Suppose* $\mathbb{E}\|\mathcal{A}(X) - \theta\|^2 \leqslant \alpha^2$, $\mathcal{A}^j = \mathcal{A}^l$ *is low bias for all* $j, l \in [d]$, *and* $P(|\mathcal{A}^j(X) - \theta^j| > \eta) \leqslant 1/48n$. *Then,*

$$\mathbb{E}\left[\sum_{i=1}^{n}(x_i^j - \theta^j)\lfloor\mathcal{A}(X) - \theta\rceil_\eta^j\right] \geqslant \frac{1}{24}$$

for all $j \in [d]$.

Proof. Note that $\mathbb{E}[(\mathcal{A}^j(X^j) - \theta^j)^2] \leqslant \frac{\alpha^2}{d}$ because $\mathcal{A}^j = \mathcal{A}^l$ for all j, l and $\mathbb{E}\|\mathcal{A}(X) - \theta\|^2 \leqslant \alpha^2$. By Lemma 26, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n}(x_i^j - \theta^j)(\mathcal{A}^j(X^j) - \theta^j)\right] &\geqslant \frac{a^2}{3} - \mathbb{E}[(\mathcal{A}^j(X^j) - \theta^j)^2] \\
&\quad + \frac{1-a^2}{2a}\left(\mathbb{E}_{X^j\sim P_a^n}[\mathcal{A}^j(X^j) - a] - \mathbb{E}_{X^j\sim P_{-a}^n}[\mathcal{A}^j(X^j) + a]\right) + 1 - a^2 \\
&\geqslant 1 - \frac{2}{3}a^2 - \frac{\alpha^2}{d} \\
&\geqslant 1/6,
\end{aligned}$$

since $\mathcal{A}^j$ is low bias and using the assumptions on the values of a and α.

Also, by the law of total expectation, we have

$$\begin{aligned}
\mathbb{E}\left[\left((\mathcal{A}(X)^j - \theta^j) - \lfloor\mathcal{A}(X) - \theta\rceil_\eta^j\right)\sum_{i=1}^{n}(x_i^j - \theta^j)\right] &\leqslant 2n\mathbb{E}\left[\left((\mathcal{A}(X)^j - \theta^j) - \lfloor\mathcal{A}(X) - \theta\rceil_\eta^j\right)\Big||\mathcal{A}(X)^j - \theta^j| > \eta\right]\frac{1}{48n} \\
&\leqslant \frac{1}{12}.
\end{aligned}$$

Combining the above inequalities completes the proof. □

As mentioned earlier, $P(|\mathcal{A}^j(X) - \theta^j| > \eta) \leqslant 1/48n$ (and the other assumptions in the lemma) can be ensured by Lemma 27.

Below we obtain a high probability lower bound on the sum of the scores:

Proposition 29. *Let* $1 \geqslant a > 100\alpha/\sqrt{d}$. *Suppose* $\mathbb{E}\|\mathcal{A}(X) - \theta\|^2 \leqslant \alpha^2$, $\mathcal{A}^j = \mathcal{A}^l$ *is low bias for all* $j, l \in [d]$, *and* $P(|\mathcal{A}^j(X) - \theta^j| > \eta) \leqslant 1/48n$. *If* $d \geqslant 200\alpha n\sqrt{\ln(1/\delta)}$, *then*

$$P\left[\sum_{i=1}^{n}\langle x_i - \theta, \lfloor\mathcal{A}(X) - \theta\rceil_\eta\rangle \geqslant \frac{d}{48}\right] \geqslant 1 - \delta.$$

Proof. We use the concentration inequality of Dwork et al. (2015, Theorem 36) and Lemma 28. Specifically, Lemma 28 implies that the assumptions of Dwork et al. (2015, Theorem 36) are satisfied with $X_{i,j} = x_i^j - \theta^j$, $c = 1/2$, $Y_j = \lfloor\mathcal{A}(X) - \theta\rceil_\eta^j$, $\gamma_j = 1/24$, $\alpha \to \eta$, and $\mathbf{a} = \mathbb{1}_n$ as the vector of all 1's. Thus, for any $\lambda > 0$, we have

$$\begin{aligned}
P\left[\sum_{i=1}^{n}\langle x_i - \theta, \lfloor\mathcal{A}(X) - \theta\rceil_\eta\rangle < \frac{d}{24} - \lambda\right] &\leqslant \exp\left(-\frac{\lambda^2}{8d\eta^2n^2}\right) \\
&= \exp\left(-\frac{\lambda^2}{8\alpha^2n^2}\right).
\end{aligned}$$

Choosing $\lambda = \alpha n\sqrt{8\ln(1/\delta)} \leqslant d/48$ completes the proof. □

Next, we provide a high probability upper bound on the sum of the squared scores:

Proposition 30. *(Dwork et al., 2015) Fix $\theta \in [-1,1]^d$ and let $X \sim P_\theta^n$. Assume $d \geqslant 64(n + \sqrt{\ln(1/\delta)})$. Then,*

$$P\left[\sum_{i=1}^n \langle x_i - \theta, \lfloor \mathcal{A}(X) - \theta \rceil_\eta \rangle^2 > 4\eta^2 d^2\right] \leqslant \delta.$$

Proof. This is immediate from Dwork et al. (2015, Lemma 22 and Proposition 23) and the proofs of these results. □

The next elementary lemma is taken verbatim from (Dwork et al., 2015, Lemma 24):

Lemma 31. *(Dwork et al., 2015) Let $\sigma \in \mathbb{R}^n$ satisfy $\sum_{i=1}^n \sigma_i \geqslant A$ and $\sum_{i=1}^n \sigma_i^2 \leqslant B^2$. Then,*

$$\left|\left\{i \in [n] : \sigma_i \geqslant \frac{A}{2n}\right\}\right| \geqslant \left(\frac{A}{2B}\right)^2.$$

We are now prepared to prove Theorem 25:

Proof of Theorem 25. For notational convenience, we will assume that $\tilde{\mathcal{A}} = \mathcal{A}$ is low bias, symmetric ($\mathcal{A}^j = \mathcal{A}^l$), and accurate in both expected ℓ_2 norm and high probability ℓ_∞ norm. This is without loss of generality, by Lemma and Lemma 27. We first prove a): Assume y is independent of X (drawn from the same distribution). By Hoeffding's inequality,

$$\begin{aligned} P\left(\mathcal{I}(y, \mathcal{A}(X)) = IN\right) &= P\left(\langle y - \theta, \lfloor \mathcal{A}(X) - \theta \rceil_\eta \rangle > \tau = 2\eta\sqrt{d\ln(1/\delta)}\right) \\ &\leqslant \exp\left(-\frac{2\tau^2}{d(4\eta)^2}\right) \leqslant \delta. \end{aligned}$$

Next, we prove b): Note that the assumptions on d and α imply that $d \geqslant 64(n + \sqrt{\ln(1/\delta)})$. Proposition 29 implies that

$$\sum_{i=1}^n \langle x_i - \theta, \lfloor \mathcal{A}(X) - \theta \rceil_\eta \rangle \geqslant \frac{d}{48} =: A$$

with probability at least $1 - \delta$. Proposition 30 implies that

$$\sum_{i=1}^n \langle x_i - \theta, \lfloor \mathcal{A}(X) - \theta \rceil_\eta \rangle^2 \leqslant 4\eta^2 =: B^2$$

with probability at least $1 - \delta$. By a union bound, both of the above events occur with probability at least $1 - 2\delta$. Then, Lemma 31 implies that

$$\left|\left\{i \in [n] : \langle x_i - \theta, \lfloor \mathcal{A}(X) - \theta \rceil_\eta \rangle \geqslant \frac{d}{96n}\right\}\right| \geqslant \left(\frac{A}{2B}\right)^2 \geqslant \frac{d}{10^6\alpha^2}.$$

Moreover, $\frac{d}{96n} \geqslant \tau = 4\alpha\sqrt{\ln(1/\delta)}$ by the assumption $d > 400\alpha n\sqrt{\ln(1/\delta)}$. This completes the proof. □

Now we are ready to prove Theorem 23:

Proof of Theorem 23. We will first prove a lower bound that is larger than the one stated in Theorem 23 by a factor of d for distributions on $\mathcal{X}' := \{\pm 1\}^d$ and $\|\mathbb{E}_{x\sim P}[x]\| \leqslant \sqrt{d}a$, where $a := \min\left(\frac{1}{\sqrt{n_{\text{pub}}}}, \frac{\sqrt{d}}{\varepsilon n_{\text{priv}}}\right)$. We will re-scale at the end to obtain Theorem 23.

By a standard reduction (see e.g. (Bun et al., 2014, Lemma 2.5)), it suffices to prove the lower bound for $\varepsilon = 1$. Let P_θ be the product distribution described in Theorem 25 with $\theta^j \sim \textbf{Unif}([-a, a])$ for $j \in [d]$ and $a = \min\left(\frac{1}{\sqrt{n_{\text{pub}}}}, \frac{\sqrt{d}}{\varepsilon n_{\text{priv}}}\right)$. Let $\mathcal{A} = (\mathcal{A}^1, \ldots, \mathcal{A}^d)$ be $(1, o(1/n_{\text{priv}}))$-semi-DP with $\mathbb{E}\|\mathcal{A}(X) - \theta\|^2 \leqslant \alpha^2 := \sup_P \mathbb{E}_{X\sim P^n}\|\mathcal{A}(X) - \mathbb{E}_{x\sim P}[x]\|^2$, where the supremum is taken over all distributions on $\mathcal{X}'$. By Lemma 27 and our earlier discussions, we assume without loss of

generality that $\mathcal{A}$ is low bias, $\mathcal{A}^j = \mathcal{A}^l$ for all $j, l \in [d]$, and that $\|\mathcal{A}(X) - \mathbb{E}_{x\sim P_\theta}[x]\|_\infty \leqslant 2\alpha/\sqrt{d}$ with arbitrarily high probability.

Note that $\alpha^2 \gtrsim \frac{d}{n}$ always holds by the non-private lower bound. Moreover, if $\alpha^2 > da^2/10000$, then we are done. Thus, we may assume $\alpha^2 \leqslant da^2/10000$. We will derive a contradiction under this assumption.

Let $X_{\text{priv}} = (x_1, \ldots, x_{n_{\text{priv}}})$ denote the private samples in X and $X_{\text{pub}} = (w_1, \ldots, w_{n_{\text{pub}}})$ denote the public samples in $X = (X_{\text{priv}}, X_{\text{pub}}) = (z_1, \ldots, z_n) \sim P_\theta^n$. Let $r = \frac{d}{400\alpha\sqrt{\ln(2/\gamma)}}$, $t = \frac{d}{10^6\alpha^2}$, and $\gamma \leqslant \frac{1}{(r+t)^2}$. We will show that if $\alpha \ll \frac{d}{n_{\text{priv}}}$ and $\alpha \ll \sqrt{\frac{d}{n_{\text{pub}}}}$, then $\mathcal{A}$ is not $(1, o(1/n_{\text{priv}}))$-semi-DP. Assume $n_{\text{priv}} < r$, $n_{\text{pub}} < t/2 - 1$, and $t < n < r + (t/2 - 1)$. Thus, the assumptions in Theorem 25 hold. We will show that the attack given in Algorithm 2 succeeds at identifying at least $n_{\text{pub}} + 1$ people in X with high probability: By part b) of Theorem 25 with $\delta = \gamma$, we have

$$\begin{aligned} P(|\{i \in [n] : \mathcal{I}(z_i, \mathcal{A}(X)) = IN\}| \geqslant n_{\text{pub}} + 1) &\geqslant P(|\{i \in [n] : \mathcal{I}(z_i, \mathcal{A}(X)) = IN\}| \geqslant t/2) \\ &\geqslant 1 - \frac{1}{(r+t)^2} \\ &\geqslant 1 - \frac{1}{n_{\text{priv}}^2}. \end{aligned}$$

That is, $\mathcal{I}$ identifies at least $n_{\text{pub}}+1$ individuals in the data set with high probability $\geqslant 1-1/n_{\text{priv}}^2$. Let $v_i = \mathbb{1}_{\{\mathcal{I}(z_i, \mathcal{A}(X)) = IN\}}$ be the indicator of the event $\mathcal{I}(z_i, \mathcal{A}(X)) = IN$. By Markov's inequality, we have

$$\mathbb{E}\left[\sum_{i=1}^n v_i\right] = \sum_{i=1}^{n_{\text{priv}}} P(\mathcal{I}(x_i, \mathcal{A}(X)) = IN) + \sum_{i=1}^{n_{\text{pub}}} P(\mathcal{I}(w_i, \mathcal{A}(X)) = IN) \geqslant (n_{\text{pub}} + 1)(1 - 1/n^2).$$

Now, $\sum_{i=1}^{n_{\text{pub}}} P(\mathcal{I}(w_i, \mathcal{A}(X)) = IN) \leqslant n_{\text{pub}}$, which implies that

$$\begin{aligned} \sum_{i=1}^n P(\mathcal{I}(x_i, \mathcal{A}(X)) = IN) &\geqslant (n_{\text{pub}} + 1)(1 - 1/n^2) - n_{\text{pub}} \\ &\geqslant 1 - 1/n_{\text{priv}} \\ &\geqslant 1/2. \end{aligned}$$

Thus, there exists a private sample $x_{i*} \in X_{\text{priv}}$ such that

$$P(\mathcal{I}(x_{i*}, \mathcal{A}(X)) = IN) \geqslant 1/2n_{\text{priv}}.$$

Now consider the adjacent data set X' obtained by replacing x_{i*} with an independent sample $y \sim P_\theta$. Then, part a) of Theorem 25 implies

$$P(\mathcal{I}(x_{i*}, \mathcal{A}(X')) = IN) \leqslant \frac{1}{(r+t)^2} \leqslant \frac{1}{n^2} \leqslant \frac{1}{n_{\text{priv}}^2},$$

where the probability is taken over the random independent draws of all the data points (including y) and the mechanism $\mathcal{A}$. Thus, $\mathcal{A}$ cannot satisfy $(\varepsilon, 1/4n_{\text{priv}})$-semi-DP unless

$$\frac{1}{2n_{\text{priv}}} \leqslant e^\varepsilon \frac{1}{n_{\text{priv}}^2} + \frac{1}{4n_{\text{priv}}} \implies \ln(n/4) \leqslant \varepsilon.$$

In particular, if $n_{\text{priv}} \geqslant 11$, then $\mathcal{A}$ cannot be $(1, o(1/n_{\text{priv}}))$-semi-DP, which contradicts our earlier hypothesis. This proves the unscaled lower bound on $\mathcal{X}'$: $\alpha^2 = \tilde{\Omega}\left(d \min\left(1/n_{\text{pub}}, d/(\varepsilon n_{\text{priv}})^2 + 1/n\right)\right)$.

Now we will scale the lower bound: Let $\mathcal{X} = \left\{\pm 1/\sqrt{d}\right\}^d$. Draw θ^j uniformly from $[-a/\sqrt{d}, a/\sqrt{d}]$ for all $j \in [d]$ and then let P_θ be the distribution on $\mathcal{X}$ that has mean $\theta \in [-a/\sqrt{d}, a/\sqrt{d}]^d$. Note that $\|\mathbb{E}_{x\sim P_\theta}[x]\| \leqslant \min\left(1/\sqrt{n_{\text{pub}}}, \sqrt{d}/\varepsilon n_{\text{priv}}\right)$ for any θ. Moreover, for any (ε, δ)-semi-DP $\mathcal{A} : \mathcal{X}^n \to [-a/\sqrt{d}, a/\sqrt{d}]^d$, there exists (ε, δ)-semi-DP $\mathcal{A}' : (\mathcal{X}')^n \to [-a, a]^d$

such that $\mathcal{A}(X) = \frac{1}{\sqrt{d}}\mathcal{A}'(X')$ for $X = \frac{1}{\sqrt{d}}X'$. Applying our lower bound for the MSE of $\mathcal{A}'$, we have

$$\begin{aligned}
\sup_{\theta\in[-a/\sqrt{d},a/\sqrt{d}]^d} \mathbb{E}_{X\sim P_\theta^n}\left[\|\mathcal{A}(X)-\theta\|^2\right] &= \sup_{\theta'\in[-a,a]^d} \mathbb{E}_{X\sim P_{\theta'}^n}\left\|\frac{1}{\sqrt{d}}(\mathcal{A}'(X)-\theta')\right\|^2 \\
&= \frac{1}{d}\sup_{\theta'\in[-a,a]^d} \mathbb{E}_{X\sim P_{\theta'}^n}\left\|\mathcal{A}'(X)-\theta'\right\|^2 \\
&\geqslant \frac{1}{d}\widetilde{\Omega}\left(d\min\left(1/n_{\text{pub}}, d/(\varepsilon n_{\text{priv}})^2\right)\right) \\
&= \widetilde{\Omega}\left(\min\left\{\frac{1}{n_{\text{pub}}}, \frac{d}{\varepsilon^2 n_{\text{priv}}^2}+\frac{1}{n}\right\}\right),
\end{aligned}$$

as desired. □

With Theorem 23 in hand, we can easily complete the proof of Theorem 22 by giving a matching upper bound (up to log factors):

Proof of Theorem 22. **Lower bound:** This was proved in Theorem 23.

Upper bound: First, the throw-away estimator $\mathcal{A}(X) = \frac{1}{n_{\text{pub}}}\sum_{x\in X_{\text{pub}}} x$ is clearly $(0,0)$-semi-DP and has MSE

$$\begin{aligned}
\mathbb{E}_{X\sim P^n}\left\|\mathcal{A}(X)-\mathbb{E}_{x\sim P}[x]\right\|^2 &= \mathbb{E}_{X\sim P^n}\left\|\frac{1}{n_{\text{pub}}}\sum_{x\in X_{\text{pub}}}(x-\mathbb{E}_{x\sim P}[x])\right\|^2 \\
&= \frac{1}{n_{\text{pub}}^2}\sum_{x\in X_{\text{pub}}}\mathbb{E}_{x\sim P}\left\|x-\mathbb{E}_{x\sim P}[x]\right\|^2 \\
&\leqslant \frac{1}{n_{\text{pub}}}.
\end{aligned}$$

To get the second term in the minimum in (15), consider the Gaussian mechanism $\mathcal{A}(X) = \bar{X} + \mathcal{N}(0,\sigma^2\mathbf{I}_d)$, where $\sigma^2 = \frac{8\ln(2/\delta)}{\varepsilon^2 n^2}$. We know $\mathcal{A}$ is (ε,δ)-DP (by e.g. (Dwork et al., 2014)) since the ℓ_2-sensitivity is $\sup_{X\sim X'}\|\bar{X}-\bar{X}'\|_2 \leqslant \frac{2}{n}$. Hence $\mathcal{A}$ is (ε,δ)-semi-DP. Moreover, the MSE of $\mathcal{A}$ is

$$\begin{aligned}
\mathbb{E}_{X\sim P^n}\left\|\mathcal{A}(X)-\mathbb{E}_{x\sim P}[x]\right\|^2 &= \mathbb{E}\|\mathcal{A}(X)-\bar{X}\|^2 + \mathbb{E}\left\|\bar{X}-\mathbb{E}_{x\sim P}[x]\right\|^2 \\
&\leqslant \frac{8d\ln(2/\delta)}{\varepsilon^2 n^2}+\frac{1}{n}.
\end{aligned}$$

This completes the proof of Theorem 22. □

Next, we turn to the pure ε-semi-DP case.

Pure ε-Semi-DP Mean Estimation

Theorem 32 (Pure ε-semi-DP mean estimation)**.** *Let $\varepsilon \leqslant d/8$ and either $n_{pub} \lesssim \frac{n\varepsilon}{d}$ or $d \lesssim 1$. Then, there exist absolute constants c and C, with $0 < c \leqslant C$, such that*

$$c\min\left\{\frac{1}{n_{pub}}, \frac{d^2}{n^2\varepsilon^2}+\frac{1}{n}\right\} \leqslant \mathcal{M}_{pop}(\epsilon,\delta=0,n_{priv},n,d) \leqslant C\min\left\{\frac{1}{n_{pub}}, \frac{d^2}{n^2\varepsilon^2}+\frac{1}{n}\right\}. \tag{20}$$

Moreover, the upper bound in (20) holds for any n_{pub}, d.

The restriction on n_{pub} is needed for our lower bound proof—via Theorem 33—to work. It seems challenging to remove this restriction, but we do believe the same lower bound holds in the complementary parameter regime. Proving this may require the invention of new techniques, making it an interesting direction for future work.

The proof of (20) will require the following intermediate result, Theorem 33, which can be viewed as a "semi-DP Fano's inequality."

Theorem 33. *Let $\{P_v\}_{v\in\mathcal{V}} \subset \mathcal{P}$ be a family of distributions on $\mathcal{X}$. Let P_0 be a distribution and $p \in [0,1]$ such that $P_{\theta_v} := (1-p)P_0 + pP_v \in \mathcal{P}$ for all $v \in \mathcal{V}$. Denote $\theta_v := \mathbb{E}_{x\sim P_{\theta_v}}[x]$ and $\rho^*(\mathcal{V}) := \min\{\|\theta_v - \theta_{v'}\| \mid v, v' \in \mathcal{V}, v \neq v'\}$. Let $\widehat{\theta} : \mathcal{X}^n \to \mathcal{X}$ be any ε-semi-DP estimator. Draw $V \sim$ Unif$(\mathcal{V})$; then conditional on $V = v$, draw an i.i.d. sample $X|V = v \sim P_{\theta_v}^n$ containing n_{priv} private samples and n_{pub} samples. Then,*

$$\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} P_{\theta_v}\left(\left\|\widehat{\theta}(X) - \theta_v\right\| \geqslant \rho^*(\mathcal{V})\right) \geqslant \frac{(|\mathcal{V}|-1)e^{-\varepsilon\lceil n_{priv}p\rceil}(1-p)^{n_{pub}}}{2\left(1+(|\mathcal{V}|-1)e^{-\varepsilon\lceil n_{priv}p\rceil}\right)} \tag{21}$$

Remark 34. *Note that the right-hand-side of (21) is similar to the second term on the right-hand-side of DP Fano's inequality (Acharya et al., 2021, Equation 5), after aligning notation. The main differences are that (21) has an extra factor of $(1-p)^{n_{pub}}$, and n_{priv} in place of n.*

Proof of Theorem 33. Our proof builds on the techniques of Barber & Duchi (2014). A key step in the proof is (22): if A is a measurable set and $v, v' \in \mathcal{V}$, then

$$P_{\theta_v}(\widehat{\theta}(X) \in A) \geqslant e^{\varepsilon\lceil n_{\text{priv}}p\rceil}\left[P_{\theta_{v'}}(\widehat{\theta} \in A) - 1 + \frac{(1-p)^{n_{\text{pub}}}}{2}\right]. \tag{22}$$

Let us now prove (22): We will use upper case letters to denote random variables and lower case letters to denote the values that the random variables take. Let $B = \{B_i\}_{i=1}^n$ be i.i.d. Bernoulli(p) random variables. Assume that the random variables $X = \{X_i\}_{i=1}^n$ are generated in the following way: first draw $W_1^0, \ldots, W_n^0 \sim P_0$ i.i.d. and draw $W_1^v, \ldots, W_n^v \sim P_v$ i.i.d. For each i, if $B_i = 0$, set $X_i = W_i^0$; if $B_i = 1$, set $X_i = W_i^v$. Thus, conditional on $V = v$, the random variables X_i are each distributed according to $P_{\theta_v} = (1-p)P_0 + pP_v$. For fixed $v' \in \mathcal{V}$, generate a different sample $X' = \{X_i'\}_{i=1}^n$ by drawing $W_i^{v'} \sim P_{v'}$ i.i.d. and setting $X_i' = W_i^0(1-B_i) + W_i^{v'}B_i$. Note that if $B_i = 0$, then $X_i = X_i'$. Thus, the hamming distance between X and X' is

$$d_{\text{ham}}(X, X') \leqslant B^T\mathbb{1} = \sum_{i=1}^n B_i.$$

Now let Q denote the conditional distribution of the ε-semi-DP estimator $\widehat{\theta}$ given input data (X or X'). For notational convenience, assume without loss of generality that $X = (X_1, \ldots, X_{n_{\text{priv}}}, X_{\text{pub}})$ and $X' = (X_1', \ldots, X_{n_{\text{priv}}}', X_{\text{pub}}')$. Then, for any fixed sequence $b = (b_1, \ldots, b_{n_{\text{priv}}}, \mathbf{0}_{n_{\text{pub}}}) \in \{0,1\}^{n_{\text{priv}}} \times \{0\}^{n_{\text{pub}}}$, we have

$$Q(\widehat{\theta} \in A | X_i = W_i^0(1-b_i) + W_i^v b_i \ \forall i \in [n]) \geqslant e^{-\varepsilon b^T\mathbb{1}} Q(\widehat{\theta} \in A | X_i' = W_i^0(1-b_i) + W_i^{v'} b_i \ \forall i \in [n]) \tag{23}$$

by group privacy, since $b_i = 0\ \forall i > n_{\text{priv}}$ implies $X_{\text{pub}} = X'_{\text{pub}}$. Thus,

$$
\begin{aligned}
&P_{\theta_v}(\widehat{\theta} \in A) \\
&= \sum_{b \in \{0,1\}^n} P(B = b) \int Q(\widehat{\theta} \in A | X_i = w_i^0(1-b_i) + w_i^v b_i \ \forall i \in [n]) dP_0^n(w_{1:n}^0) dP_v^n(w_{1:n}^v) \\
&\geqslant \sum_{\substack{b \in \{0,1\}^{n_{\text{priv}}} \times \{0\}^{n_{\text{pub}}}, \\ b^T \mathbb{1} \leqslant \lceil n_{\text{priv}} p \rceil}} P(B = b) \int Q(\widehat{\theta} \in A | X_i = w_i^0(1-b_i) + w_i^v b_i \ \forall i \in [n]) dP_0^n(w_{1:n}^0) dP_v^n(w_{1:n}^v) \\
&= \sum_{\substack{b \in \{0,1\}^{n_{\text{priv}}} \times \{0\}^{n_{\text{pub}}}, \\ b^T \mathbb{1} \leqslant \lceil n_{\text{priv}} p \rceil}} P(B = b) \int Q(\widehat{\theta} \in A | X_i = w_i^0(1-b_i) + w_i^v b_i \ \forall i \in [n]) dP_0^n(w_{1:n}^0) dP_v^n(w_{1:n}^v) dP_{v'}^n(w_{1:n}^{v'}) \\
&\geqslant \sum_{\substack{b \in \{0,1\}^{n_{\text{priv}}} \times \{0\}^{n_{\text{pub}}}, \\ b^T \mathbb{1} \leqslant \lceil n_{\text{priv}} p \rceil}} P(B = b) \int \left[e^{-\varepsilon b^T \mathbb{1}} Q(\widehat{\theta} \in A | X'_i = w_i^0(1-b_i) + w_i^{v'} b_i \ \forall i \in [n]) dP_0^n(w_{1:n}^0) \right. \\
&\qquad\qquad \left. \cdots dP_v^n(w_{1:n}^v) dP_{v'}^n(w_{1:n}^{v'}) \right] \\
&= \sum_{\substack{b \in \{0,1\}^{n_{\text{priv}}} \times \{0\}^{n_{\text{pub}}}, \\ b^T \mathbb{1} \leqslant \lceil n_{\text{priv}} p \rceil}} P(B = b) e^{-\varepsilon b^T \mathbb{1}} P_{\theta_{v'}} \left(\widehat{\theta} \in A | B = b \right) \\
&\geqslant \sum_{\substack{b \in \{0,1\}^{n_{\text{priv}}} \times \{0\}^{n_{\text{pub}}}, \\ b^T \mathbb{1} \leqslant \lceil n_{\text{priv}} p \rceil}} P(B = b) e^{-\varepsilon \lceil n_{\text{priv}} p \rceil} P_{\theta_{v'}} \left(\widehat{\theta} \in A | B = b \right) \\
&\geqslant e^{-\varepsilon \lceil n_{\text{priv}} p \rceil} P_{\theta_{v'}} \left(\widehat{\theta} \in A, B_{(i > n_{\text{priv}})} = \mathbf{0}_{n_{\text{pub}}}, B^T \mathbb{1} \leqslant \lceil n_{\text{priv}} p \rceil \right) \\
&\geqslant e^{-\varepsilon \lceil n_{\text{priv}} p \rceil} \left[P_{\theta_{v'}} \left(\widehat{\theta} \in A \right) - P \left(B_{(i > n_{\text{priv}})} \neq \mathbf{0}_{n_{\text{pub}}} \bigcup B^T \mathbb{1} > \lceil n_{\text{priv}} p \rceil \right) \right],
\end{aligned}
$$

where the second inequality used (23) and the last inequality used a union bound. Now, by independence of $\{B_i\}_{i=1}^n$, we have

$$
\begin{aligned}
&P \left(B_{(i > n_{\text{priv}})} \neq \mathbf{0}_{n_{\text{pub}}} \bigcup B^T \mathbb{1} > \lceil n_{\text{priv}} p \rceil \right) \\
=&1 - P \left(B_{(i > n_{\text{priv}})} = \mathbf{0}_{n_{\text{pub}}} \bigcap B^T \mathbb{1} \leqslant \lceil n_{\text{priv}} p \rceil \right) \\
=&1 - P \left(B_{(i > n_{\text{priv}})} = \mathbf{0}_{n_{\text{pub}}} \right) P \left(B^T \mathbb{1} \leqslant \lceil n_{\text{priv}} p \rceil \middle| B_{(i > n_{\text{priv}})} = \mathbf{0}_{n_{\text{pub}}} \right) \\
=&1 - P \left(B_{(i > n_{\text{priv}})} = \mathbf{0}_{n_{\text{pub}}} \right) P \left(B_{(i \leqslant n_{\text{priv}})}^T \mathbb{1} \leqslant \lceil n_{\text{priv}} p \rceil \right) \\
=&1 - (1-p)^{n_{\text{pub}}} P \left(B_{(i \leqslant n_{\text{priv}})}^T \mathbb{1} \leqslant \lceil n_{\text{priv}} p \rceil \right) \\
\leqslant&1 - (1-p)^{n_{\text{pub}}} \cdot \frac{1}{2},
\end{aligned}
$$

since the median of $B_{(i \leqslant n_{\text{priv}})}^T \mathbb{1} \sim \text{Binomial}(n_{\text{priv}}, p)$ is no larger than $\lceil n_{\text{priv}} p \rceil$. Putting together the pieces, we get

$$
P_{\theta_v}(\widehat{\theta} \in A) \geqslant e^{-\varepsilon \lceil n_{\text{priv}} p \rceil} \left[P_{\theta_{v'}} \left(\widehat{\theta} \in A \right) - 1 + (1-p)^{n_{\text{pub}}} \cdot \frac{1}{2} \right],
$$

establishing (22).

Now, with (22) in hand, we proceed as in the proof of Barber & Duchi (2014, Theorem 3): Let $\mathbb{B}_\alpha(\theta) := \{\theta' \in \mathcal{X} : \|\theta - \theta'\| \leqslant \alpha\}$. Note that the balls $\mathbb{B}_{\rho^*(\mathcal{V})}(\theta_v)$ are disjoint for $v \in \mathcal{V}$ by definition of $\rho^*(\mathcal{V})$. Denote the average probability of success for the estimator $\widehat{\theta}$ by

$$
P_{\text{succ}} := \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_{\theta_v} \left(\widehat{\theta} \in \mathbb{B}_{\rho^*(\mathcal{V})}(\theta_v) \right).
$$

Then by a union bound and disjointness of the balls $\{\mathbb{B}_{\rho^*(\mathcal{V})}(\theta_v)\}_{v\in\mathcal{V}}$, we have

$$P_{\text{succ}} \leqslant 1 - \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\sum_{v'\in\mathcal{V},v'\neq v} P_{\theta_v}\left(\widehat{\theta}\in\mathbb{B}_{\rho^*(\mathcal{V})}(\theta_{v'})\right).$$

An application of (22) yields

$$\begin{aligned}P_{\text{succ}} &\leqslant 1 - \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\sum_{v'\in\mathcal{V},v'\neq v}\left[e^{-\varepsilon\lceil n_{\text{priv}}p\rceil}P_{\theta_{v'}}\left(\widehat{\theta}\in\mathbb{B}_{\rho^*(\mathcal{V})}(\theta_{v'})\right) - \left(1-\frac{(1-p)^{n_{\text{pub}}}}{2}\right)\right]\\ &\leqslant 1 - e^{-\varepsilon\lceil n_{\text{priv}}p\rceil}(|\mathcal{V}|-1)P_{\text{succ}} + e^{-\varepsilon\lceil n_{\text{priv}}p\rceil}(|\mathcal{V}|-1)\left(1-\frac{(1-p)^{n_{\text{pub}}}}{2}\right).\end{aligned}$$

Re-arranging this inequality leads to

$$P_{\text{succ}} \leqslant \frac{1+(|\mathcal{V}|-1)\left(1-\frac{(1-p)^{n_{\text{pub}}}}{2}\right)e^{-\varepsilon\lceil n_{\text{priv}}p\rceil}}{1+(|\mathcal{V}|-1)e^{-\varepsilon\lceil n_{\text{priv}}p\rceil}}$$

and hence

$$1 - P_{\text{succ}} \geqslant \frac{(|\mathcal{V}|-1)e^{-\varepsilon\lceil n_{\text{priv}}p\rceil}\cdot\frac{(1-p)^{n_{\text{pub}}}}{2}}{1+(|\mathcal{V}|-1)e^{-\varepsilon\lceil n_{\text{priv}}p\rceil}}.$$

This last inequality is equivalent to the inequality stated in Theorem 33. □

While we state Theorem 33 for mean estimation, it holds more generally for estimating any population statistic $\theta : \mathcal{P}\to\Theta$. However, this additional generality will not be necessary for our purposes. With Theorem 33 in hand, we now turn to the proof of Theorem 22.

Proof of Theorem 32. **Lower Bounds:** We begin by proving (20). First suppose $n_{\text{pub}} \lesssim \frac{n\varepsilon}{d}$ and $d \geqslant 8$.

Choose a finite subset $\mathcal{V}\subset\mathbb{R}^d$ such that $|\mathcal{V}|\geqslant 2^{d/2}$, $\|v\|=1$, and $\|v-v'\|\geqslant\frac{1}{8}$ for all $v,v'\in\mathcal{V}, v\neq v'$. The existence of such a set of points is well-known (see e.g. the Gilbert-Varshamov construction). Define P_0 to be the point mass distribution on $\{X=0\}$ and P_v to be point mass on $\{X=v\}$ for $v\in\mathcal{V}$. For $v\in\mathcal{V}$, let $P_{\theta_v} := (1-p)P_0 + pP_v$ for some $p\in[0,1]$ to be specified later. Note that if $X\sim P_{\theta_v}$, then $\|X\|\leqslant 1$ with probability 1. Thus, P_{θ_v} is a valid distribution in the class $\mathcal{P}$ of bounded (by 1) distributions on $\mathbb{B}$ that we are considering. Also, note that $\theta_v := \mathbb{E}_{P_{\theta_v}}[X] = pv$. Further,

$$\rho^*(\mathcal{V}) := \min\left\{\|\theta_v-\theta_{v'}\| | v,v'\in\mathcal{V}, v\neq v'\right\} \geqslant \frac{p}{8}$$

by construction.

Now we use the classical reduction from estimation to testing (see (Barber & Duchi, 2014) for details) to lower bound the MSE of any ε-semi-DP estimator $\widehat{\theta}$ by

$$\begin{aligned}\sup_{P\in\mathcal{P}}\mathbb{E}_{X\sim P^n,\hat{\theta}}\left\|\widehat{\theta}(X)-\mathbb{E}_{x\sim P}[x]\right\|^2 &\geqslant \rho^*(\mathcal{V})^2\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}P_{\theta_v}\left(\|\widehat{\theta}(X)-\theta_v\|\geqslant\rho^*(\mathcal{V})\right)\\ &\geqslant \left(\frac{p}{8}\right)^2\frac{(|\mathcal{V}|-1)e^{-\varepsilon\lceil n_{\text{priv}}p\rceil}(1-p)^{n_{\text{pub}}}}{2\left(1+(|\mathcal{V}|-1)e^{-\varepsilon\lceil n_{\text{priv}}p\rceil}\right)}\\ &\geqslant \frac{p^2}{64}\frac{(2^{d/2}-1)e^{-\varepsilon\lceil n_{\text{priv}}p\rceil}(1-p)^{n_{\text{pub}}}}{2\left(1+(2^{d/2}-1)e^{-\varepsilon\lceil n_{\text{priv}}p\rceil}\right)}\end{aligned}$$

where we used Theorem 33 in the second inequality. Since we assumed $d\geqslant 8$, we have $2^{d/2}-1\geqslant e^{d/4}$ and hence

$$\begin{aligned}\sup_{P\in\mathcal{P}}\mathbb{E}_{X\sim P^n,\hat{\theta}}\left\|\widehat{\theta}(X)-\mathbb{E}_{x\sim P}[x]\right\|^2 &\geqslant \frac{p^2}{64}\cdot\frac{(1-p)^{n_{\text{pub}}}}{2}\cdot\frac{e^{d/4-\varepsilon\lceil n_{\text{priv}}p\rceil}}{1+e^{d/4-\varepsilon\lceil n_{\text{priv}}p\rceil}}\\ &\geqslant \frac{p^2}{64}\cdot\frac{(1-p)^{n_{\text{pub}}}}{4}.\end{aligned}$$

for any $p \leqslant \frac{d}{4n\varepsilon} - \frac{1}{n}$. Now choose

$$p = \min\left(\frac{d}{4n\varepsilon} - \frac{1}{n}, \frac{1}{2\sqrt{n_{\text{pub}}}}\right).$$

By assumption, there exists an absolute constant k such that $n_{\text{pub}} \leqslant k\frac{n\varepsilon}{d}$. Thus, if $n\varepsilon \geqslant 2$, then

$$\begin{aligned}(1-p)^{n_{\text{pub}}} &\geqslant \left(1 - \frac{d}{\varepsilon n}\right)^{kn\varepsilon/d} \\ &= \left[\left(1 - \frac{d}{\varepsilon n}\right)^{n\varepsilon/d}\right]^k \\ &\geqslant \left[\left(1 - \frac{1}{2}\right)^2\right]^k \\ &= \frac{1}{4^k}.\end{aligned}$$

On the other hand, if $n\varepsilon < 2$, then $n_{\text{pub}} \leqslant 2k \implies (1-p)^{n_{\text{pub}}} \geqslant \left(1 - \frac{1}{2}\right)^{2k} = \frac{1}{4^k}$. Also, note that $p \geqslant \min\left(\frac{d}{8\varepsilon n}, \frac{1}{2\sqrt{n_{\text{pub}}}}\right)$ by the assumption that $d \geqslant 8\varepsilon$. Therefore,

$$\sup_{P \in \mathcal{P}} \mathbb{E}_{X\sim P^n, \widehat{\theta}} \left\|\widehat{\theta}(X) - \mathbb{E}_{x\sim P}[x]\right\|^2 \geqslant \frac{1}{256 \cdot 4^k} \min\left(\frac{d}{8\varepsilon n}, \frac{1}{2\sqrt{n_{\text{pub}}}}\right)^2.$$

Combining the above inequality with the non-private lower bound of $\Omega(1/n)$ for mean estimation proves the lower bound in (20).

Now consider the alternative case in which $d \lesssim 1$ (i.e. $d \leqslant k$ for some absolute constant $k \in \mathbb{N}$), but $n_{\text{pub}} \in [n]$ is arbitrary. Then we will prove that the lower bound in (20) holds for $d = 1$, for any $\delta \in [0, \varepsilon]$ and $\varepsilon \leqslant 1$. By taking a k-fold product distribution, this will suffice to complete the proof of the lower bound in (20). To that end, we will use Le Cam's method and build on the techniques in (Barber & Duchi, 2014; Fallah et al., 2022). The key novel ingredient is the following extension of Fallah et al. (2022, Lemma 3) to the (ε, δ)-semi-DP setting:

Lemma 35. *Let $\widehat{\theta} : \mathcal{X}^n \to \mathcal{X}$ be (ε, δ)-semi-DP and let P_1, P_2 be distributions on $\mathcal{X}$ such that P_1 is absolutely continuous w.r.t. P_2. Denote the conditional distribution of $\widehat{\theta}$ given X by Q and let $Q_j(A) = \int Q(\widehat{\theta} \in A | x_{1:n}) dP_j^n(x_{1:n})$ for any measurable set A. Then*

$$\|Q_1 - Q_2\|_{TV} \leqslant \min\left\{\sqrt{\frac{n}{2} D_{KL}(P_1, P_2)},\ 2\|P_1 - P_2\|_{TV}\, n_{priv}(e^\varepsilon - 1 + \delta) + \sqrt{\frac{n_{pub}}{2} D_{KL}(P_1, P_2)}\right\}.$$

Let us defer the proof of Lemma 35 for now. We will now use Lemma 35 to prove the lower bound in (20) for $d = 1$ and $\varepsilon \leqslant 1$. Define distributions P_1, P_2 on $\{-1, 1\}$ as follows:

$$P_1(-1) = P_2(1) = \frac{1+\gamma}{2}, \quad P_1(1) = P_2(-1) = \frac{1-\gamma}{2}$$

for some $\gamma \in [0, 1/2]$ to be chosen later. Clearly $P_1, P_2 \in \mathcal{P}$ (i.e. they are bounded by 1 with probability 1). Also, $\mathbb{E}_{P_1}[x] = -\gamma$ and $\mathbb{E}_{P_2}[x] = \gamma$, so $\{P_1, P_2\}$ is a γ-*packing* of $\{-1, 1\}$. Thus, by Le Cam's method (see (Barber & Duchi, 2014) for details), for any (ε, δ)-semi-DP $\widehat{\theta}$, we have

$$\sup_{P \in \mathcal{P}} \mathbb{E}_{X\sim P^n, \widehat{\theta}} \left\|\widehat{\theta}(X) - \mathbb{E}_{x\sim P}[x]\right\|^2 \geqslant \frac{\gamma^2}{8}\left(1 - \|Q_1 - Q_2\|_{\text{TV}}\right).$$

Now, applying Lemma 35 and the assumption $\delta \leqslant \varepsilon \leqslant 1$ yields

$$\begin{aligned}
&\sup_{P\in\mathcal{P}} \mathbb{E}_{X\sim P^n,\widehat{\theta}} \left\|\widehat{\theta}(X) - \mathbb{E}_{x\sim P}[x]\right\|^2 \qquad (24)\\
\geqslant &\frac{\gamma^2}{8}\left[1 - \min\left\{\sqrt{\frac{n}{2}D_{\mathrm{KL}}(P_1,P_2)},\ 6\|P_1-P_2\|_{\mathrm{TV}}\, n_{\mathrm{priv}}\varepsilon + \sqrt{\frac{n_{\mathrm{pub}}}{2}D_{\mathrm{KL}}(P_1,P_2)}\right\}\right]\\
\geqslant &\frac{\gamma^2}{8}\left[1 - \min\left\{\sqrt{\frac{n}{2}3\gamma^2(P_1,P_2)},\ 6\gamma n_{\mathrm{priv}}\varepsilon + \sqrt{\frac{n_{\mathrm{pub}}}{2}3\gamma^2}\right\}\right]\\
= &\frac{\gamma^2}{8}\left[1 - \gamma\min\left\{\sqrt{\frac{3n}{2}},\ 6n_{\mathrm{priv}}\varepsilon + \sqrt{\frac{3n_{\mathrm{pub}}}{2}}\right\}\right]. \qquad (25)
\end{aligned}$$

In the second inequality we used the fact that $\|P_1 - P_2\|_{\mathrm{TV}} = \frac{1}{2}\left(\left|\frac{1+\gamma}{2} - \frac{1-\gamma}{2}\right|\cdot 2\right) = \gamma$ and $D_{\mathrm{KL}}(P_1,P_2) \leqslant 3\gamma^2$ for $\gamma \leqslant 1/2$.

Now we will choose γ to (approximately) maximize the right-hand side of (24). Suppose $\min\left\{\sqrt{\frac{3n}{2}}, 6n_{\mathrm{priv}}\varepsilon + \sqrt{\frac{3n_{\mathrm{pub}}}{2}}\right\} = \sqrt{\frac{3n}{2}}$. Then choosing $\gamma = \frac{1}{3}\sqrt{\frac{2}{3n}}$ yields

$$\sup_{P\in\mathcal{P}} \mathbb{E}_{X\sim P^n,\widehat{\theta}} \left\|\widehat{\theta}(X) - \mathbb{E}_{x\sim P}[x]\right\|^2 \geqslant \frac{k}{n}$$

for some absolute constant $k > 0$. Our assumption that $\min\left\{\sqrt{\frac{3n}{2}}, 6n_{\mathrm{priv}}\varepsilon + \sqrt{\frac{3n_{\mathrm{pub}}}{2}}\right\} = \sqrt{\frac{3n}{2}}$ implies that there exists $k' > 0$ such that $n \leqslant k' \max\left(n_{\mathrm{priv}}^2\varepsilon^2, n_{\mathrm{pub}}\right)$. Thus, $k/n \geqslant \frac{k}{k'}\min\left(\frac{1}{n_{\mathrm{priv}}^2\varepsilon^2}, \frac{1}{n_{\mathrm{pub}}}\right)$, which gives the desired lower bound in (20).

Suppose instead that $\min\left\{\sqrt{\frac{3n}{2}}, 6n_{\mathrm{priv}}\varepsilon + \sqrt{\frac{3n_{\mathrm{pub}}}{2}}\right\} = 6n_{\mathrm{priv}}\varepsilon + \sqrt{\frac{3n_{\mathrm{pub}}}{2}}$. Then choose $\gamma = \frac{2}{3}\left(6n_{\mathrm{priv}}\varepsilon + \sqrt{\frac{3n_{\mathrm{pub}}}{2}}\right)^{-1}$. Then, there are constants $k, c > 0$ such that

$$\begin{aligned}
\sup_{P\in\mathcal{P}} \mathbb{E}_{X\sim P^n,\widehat{\theta}} \left\|\widehat{\theta}(X) - \mathbb{E}_{x\sim P}[x]\right\|^2 &\geqslant \frac{\gamma^2}{8}\left[1 - \gamma\min\left\{\sqrt{\frac{3n}{2}}, 6n_{\mathrm{priv}}\varepsilon + \sqrt{\frac{3n_{\mathrm{pub}}}{2}}\right\}\right]\\
&\geqslant \frac{k}{n_{\mathrm{priv}}^2\varepsilon^2 + n_{\mathrm{pub}}}\\
&\geqslant c\min\left(\frac{1}{n_{\mathrm{priv}}^2\varepsilon^2}, \frac{1}{n_{\mathrm{pub}}}\right).
\end{aligned}$$

Combining the above inequality with the non-private lower bound $\Omega(1/n)$ completes the proof of the lower bound in (20), assuming the truth of Lemma 35.

It remains to prove Lemma 35. To that end, fix = $k \in \{0, n_{\mathrm{priv}}\}$ and denote by $\widetilde{Q}$ the marginal distribution of $\widehat{\theta}$ given $X_1,\ldots,X_k \sim P_1$ (i.i.d.) and $X_{k+1},\ldots,X_n \sim P_2$ (i.i.d.); i.e. for measurable A,

$$\widetilde{Q}(A) := \int Q(\widehat{\theta}\in A | X_{1:n} = x_{1:n}) dP_1^k(x_{1:k}) dP_2^{n-k}(x_{k+1:n}).$$

Note that if $k = 0$, then $\widetilde{Q} = Q_2$. We have

$$\|Q_1 - Q_2\|_{\mathrm{TV}} \leqslant \underbrace{\|Q_1 - \widetilde{Q}\|_{\mathrm{TV}}}_{\text{ⓐ}} + \underbrace{\|\widetilde{Q} - Q_2\|_{\mathrm{TV}}}_{\text{ⓑ}}.$$

Also,

$$\min_{k\in\{0,n_{\text{priv}}\}} \sqrt{\frac{n-k}{2} D_{\text{KL}}(P_1,P_2)} + 2\|P_1-P_2\|_{\text{TV}} \min_{k\in\{0,n_{\text{priv}}\}} k(e^\varepsilon - 1 + \delta) \tag{26}$$
$$\leqslant \min\left\{\sqrt{\frac{n}{2} D_{\text{KL}}(P_1,P_2)},\ 2\|P_1-P_2\|_{\text{TV}}\, n_{\text{priv}}(e^\varepsilon-1+\delta) + \sqrt{\frac{n_{\text{pub}}}{2} D_{\text{KL}}(P_1,P_2)}\right\},$$

so it suffices to upper bound the sum of the terms of ⓐ + ⓑ by the left-hand-side of (26). First, we deal with ⓐ: for any k, we have

$$\|Q_1 - \widetilde{Q}\|_{\text{TV}}^2 \leqslant \|P_1^n - P_1^k P_2^{n-k}\|_{\text{TV}}^2 \tag{27}$$
$$\leqslant \frac{1}{2} D_{\text{KL}}\left(P_1^n, P_1^k P_2^{n-k}\right) \tag{28}$$
$$\leqslant \frac{n-k}{2} D_{\text{KL}}(P_1,P_2), \tag{29}$$

by the data processing inequality for f-divergences, Pinsker's inequality, and the chain-rule for KL-divergences (see, e.g. (Duchi, 2021) for a reference on these facts). Thus, it remains to show

$$\|\widetilde{Q} - Q_2\|_{\text{TV}} \leqslant 2\|P_1-P_2\|_{\text{TV}} \min_{k\in\{0,n_{\text{priv}}\}} k(e^\varepsilon - 1 + \delta) \tag{30}$$

for $k \in \{0, n_{\text{priv}}\}$. If $k = 0$, (30) is trivial. Assume $k = n_{\text{priv}}$. Now, for any measurable A, we may write

$$\widetilde{Q}(A) - Q_2(A) = \int_{\mathbb{R}^{n_{\text{pub}}}} \Delta(x_{n_{\text{priv}}+1:n}) dP_2^{n_{\text{pub}}}(x_{n_{\text{priv}}+1:n}), \tag{31}$$

where

$$\Delta(x_{n_{\text{priv}}+1:n}) := \int_{\mathbb{R}^{n_{\text{priv}}}} Q(\widehat{\theta}\in A | X_{1:n} = x_{1:n}) \left(dP_1^{n_{\text{priv}}}(x_{1:n_{\text{priv}}}) - dP_2^{n_{\text{priv}}}(x_{1:n_{\text{priv}}})\right).$$

By (31), it suffices to show that $|\Delta(x_{n_{\text{priv}}+1:n})| \leqslant 2\|P_1-P_2\|_{\text{TV}}\, n_{\text{priv}}(e^\varepsilon-1+\delta)$ for all $x_{n_{\text{priv}}+1:n} \in \mathcal{X}^{n_{\text{pub}}}$. To do so, let $x_{1:n}^i := (x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)$ for some $i \leqslant n_{\text{priv}}$. Then

$$|Q(\widehat{\theta}\in A|X_{1:n}=x_{1:n}) - Q(\widehat{\theta}\in A|X_{1:n}=x_{1:n}^i)| \leqslant (e^\varepsilon-1)Q(\widehat{\theta}\in A|X_{1:n}=x_{1:n}^i) + \delta \tag{32}$$

since $\widehat{\theta}$ is (ε,δ)-semi-DP. Moreover, by the proof of Fallah et al. (2022, Lemma 3), we have

$$\Delta(x_{n_{\text{priv}}+1:n}) = \sum_{i=1}^{n_{\text{priv}}} \int_{\mathbb{R}^{n_{\text{priv}}}} \left[Q(\widehat{\theta}\in A|X_{1:n}=x_{1:n}) - Q(\widehat{\theta}\in A|X_{1:n}=x_{1:n}^i)\right]$$
$$\cdots dP_2^{i-1}(x_{1:i-1})\left(dP_1(x_i) - dP_2(x_i)\right) dP_1^{n_{\text{priv}}-i}(x_{i+1:n_{\text{priv}}}).$$

Applying the triangle inequality and (32), we get

$$\begin{aligned}
&|\Delta(x_{n_{\text{priv}}+1:n})| \\
\leqslant &\sum_{i=1}^{n_{\text{priv}}} \int_{\mathbb{R}^{n_{\text{priv}}}} \left[(e^\varepsilon-1)Q(\widehat{\theta}\in A|X_{1:n}=x_{1:n}^i) + \delta\right] dP_2^{i-1}(x_{1:i-1}) \left|dP_1(x_i) - dP_2(x_i)\right| dP_1^{n_{\text{priv}}-i}(x_{i+1:n_{\text{priv}}}) \\
\leqslant &2\|P_1-P_2\|_{\text{TV}}\, n_{\text{priv}}(e^\varepsilon-1+\delta),
\end{aligned}$$

as desired. This completes the proof of Lemma 35 and hence the proof of the lower bound in (20).

Upper bound: First, the throw-away estimator $\mathcal{A}(X) = \frac{1}{n_{\text{pub}}}\sum_{x\in X_{\text{pub}}} x$ is clearly $(0,0)$-semi-DP and has MSE

$$\begin{aligned}
\mathbb{E}_{X\sim P^n}\|\mathcal{A}(X) - \mathbb{E}_{x\sim P}[x]\|^2 &= \mathbb{E}_{X\sim P^n}\left\|\frac{1}{n_{\text{pub}}}\sum_{x\in X_{\text{pub}}}(x - \mathbb{E}_{x\sim P}[x])\right\|^2 \\
&= \frac{1}{n_{\text{pub}}^2}\sum_{x\in X_{\text{pub}}} \mathbb{E}_{x\sim P}\|x - \mathbb{E}_{x\sim P}[x]\|^2 \\
&\leqslant \frac{1}{n_{\text{pub}}}.
\end{aligned}$$

To get the second term in the minimum in (20), consider the Laplace mechanism $\mathcal{A}(X) = \bar{X} + (L_1, \dots, L_d)$, where $L_i \sim \text{Lap}(2\sqrt{d}/n\varepsilon)$ are i.i.d. mean-zero Laplace random variables. We know $\mathcal{A}$ is ε-DP by (Dwork et al., 2014), since the ℓ_1-sensitivity is $\sup_{X \sim X'} \|\bar{X} - \bar{X}'\|_1 = \frac{1}{n} \sup_{x,x'} \|x - x'\|_1 \leqslant \frac{2\sqrt{d}}{n}$. Hence $\mathcal{A}$ is ε-semi-DP. Moreover, $\mathcal{A}$ has MSE

$$\begin{aligned}
\mathbb{E}_{X \sim P^n} \|\mathcal{A}(X) - \mathbb{E}_{x \sim P}[x]\|^2 &\leqslant 2\mathbb{E}\|\mathcal{A}(X) - \bar{X}\|^2 + 2\mathbb{E}\left\|\bar{X} - \mathbb{E}_{x \sim P}[x]\right\|^2 \\
&\leqslant 2d \operatorname{Var}(\text{Lap}(2\sqrt{d}/n\varepsilon)) + \frac{2}{n} \\
&= \frac{16d^2}{n^2\varepsilon^2} + \frac{2}{n}.
\end{aligned}$$

This completes the proof of (20).

□

E.1.3. An "Even More Optimal" Semi-DP Algorithm for Mean Estimation

Lemma 36 (Re-statement of Lemma 7)**.** *Recall the definition of $\mathcal{P}(B, V)$ (Definition 6): $\mathcal{P}(B, V)$ denotes the collection of all distributions P on $\mathbb{R}^d$ such that for any $x \sim P$, we have $\|x\| \leqslant B$ P-almost surely and $\text{Var}(x) = V^2$. Then, the error of the ρ-semi-zCDP throw-away algorithm $\mathcal{A}(X) = \frac{1}{n_{pub}} \sum_{x \in X_{pub}} x$ is*

$$\sup_{P \in \mathcal{P}(B,V)} \mathbb{E}_{X \sim P^n} \left[\|\mathcal{A}(X) - \mathbb{E}_{x \sim P}[x]\|^2\right] = \frac{V^2}{n_{pub}}.$$

The minimax error of the ρ-semi-zCDP Gaussian mechanism $\mathcal{G}(X) = \bar{X} + \mathcal{N}\left(0, \sigma^2 \mathbf{I}_d\right)$ is

$$\inf_{\rho\text{-zCDP } \mathcal{G}} \sup_{P \in \mathcal{P}(B,V)} \mathbb{E}_{\mathcal{G}, X \sim P^n} \left[\|\mathcal{G}(X) - \mathbb{E}_{x \sim P}[x]\|^2\right] = \frac{2dB^2}{\rho n^2} + \frac{V^2}{n}. \tag{33}$$

Proof. For throw-away, the i.i.d. data assumption implies

$$\mathbb{E}\|\mathcal{A}(X) - \mathbb{E}x\|^2 = \frac{1}{n_{\text{pub}}^2} \sum_{x \in X_{\text{pub}}} \mathbb{E}\|x - \mathbb{E}x\|^2 = \frac{V^2}{n_{\text{pub}}}.$$

The Gaussian mechanism $\mathcal{G}^*(X) := \bar{X} + \mathcal{N}\left(0, \frac{2B^2}{\rho n^2} \mathbf{I}_d\right)$ is ρ-zCDP by (Bun & Steinke, 2016, Proposition 1.6) since the ℓ_2-sensitivity is bounded by $\Delta_2 = \sup_{X \sim X'} \|\bar{X} - \bar{X}'\|_2 \leqslant \frac{2B}{n}$. Moreover, this sensitivity bound is tight: consider any P such that $x = (B, 0_{d-1})$ and $x' = (-B, 0_{d-1})$ are in the support of P. Then fix any y in the support of P and consider the adjacent data sets $X = (x, y, \cdots, y)$ and $X' = (x', y, \cdots, y)$. We have $\|\bar{X} - \bar{X}'\|_2 = \frac{1}{n}\|x - x'\|_2 = \frac{2B}{n}$. Additionally, if the variance of the additive isotropic Gaussian noise σ^2 is smaller than $\frac{\Delta_2^2}{2\rho} = \frac{2B^2}{\rho n^2}$, then the Gaussian mechanism is not ρ-zCDP (Bun & Steinke, 2016). Thus, $\mathcal{G}^*(X)$ is the ρ-zCDP Gaussian mechanism with the smallest noise variance σ^2. Hence the infimum in (33) is attained by $\mathcal{G}^*$. Finally, for any $P \in \mathcal{P}(B, V)$, the MSE of $\mathcal{G}^*$ is

$$\begin{aligned}
\mathbb{E}_{\mathcal{G}, X \sim P^n} \left[\|\mathcal{G}^*(X) - \mathbb{E}_{x \sim P}[x]\|^2\right] &= \mathbb{E}\|\mathcal{G}^*(X) - \bar{X}\|^2 + \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n x_i - \mathbb{E}x_i\right\|^2 \\
&= \frac{2dB^2}{\rho n^2} + \frac{V^2}{n}.
\end{aligned}$$

□

Proposition 37 (Re-statement of Proposition 8)**.** *Recall the definition of $\mathcal{A}_r$ from (3). $\mathcal{A}_r$ is ρ-semi-zCDP. Also, there exists $r > 0$ such that*

$$\sup_{P \in \mathcal{P}(B,V)} \mathbb{E}_{X \sim P^n} \left[\|\mathcal{A}_r(X) - \mathbb{E}_{x \sim P}[x]\|^2\right] < \min\left(\frac{V^2}{n_{pub}}, \frac{2dB^2}{\rho n^2} + \frac{V^2}{n}\right). \tag{34}$$

Further, if $\frac{V^2}{n_{pub}} \leqslant \frac{2dB^2}{\rho n^2}$*, then the quantitative advantage of* $\mathcal{A}_r$ *is*

$$\sup_{P \in \mathcal{P}(B,V)} \mathbb{E}_{X\sim P^n}\left[\|\mathcal{A}_r(X) - \mathbb{E}_{x\sim P}[x]\|^2\right] \leqslant \left(\frac{q}{q+s^2}\right) \min\left(\frac{V^2}{n_{pub}}, \frac{2dB^2}{\rho n^2} + \frac{V^2}{n}\right), \tag{35}$$

where $q = 2 + \frac{n_{priv}\rho V^2}{dB^2}$ *and* $s = \frac{V n_{priv}\sqrt{\rho}}{B\sqrt{d n_{pub}}}$.

Proof. **Privacy:** Note that the ℓ_2-sensitivity of $M(X) = \sum_{x\in X_{\text{priv}}} rx + \sum_{x\in X_{\text{pub}}} \left(\frac{1-n_{\text{priv}}r}{n_{\text{pub}}}\right) x$ is

$$\begin{aligned}\Delta_2 &= \sup_{X_{\text{priv}}\sim X'_{\text{priv}}} \left\| \sum_{x\in X_{\text{priv}}} rx + \sum_{x\in X_{\text{pub}}} \left(\frac{1-n_{\text{priv}}r}{n_{\text{pub}}}\right) x - \sum_{x\in X'_{\text{priv}}} rx - \sum_{x\in X_{\text{pub}}} \left(\frac{1-n_{\text{priv}}r}{n_{\text{pub}}}\right) x \right\| \\ &= \sup_{x,x'} \|rx - rx'\| \leqslant 2rB.\end{aligned}$$

Recall that the Gaussian mechanism guarantees ρ-zCDP whenever $\sigma^2 \geqslant \frac{\Delta_2^2}{2\rho}$ (Bun & Steinke, 2016, Proposition 1.6). Thus, $\mathcal{A}_r$ is ρ-semi-zCDP for $\sigma_r^2 \geqslant \frac{(2rB)^2}{2\rho} = \frac{2B^2r^2}{\rho}$.

Error bounds: Let $P \in \mathcal{P}(B, V)$. We have

$$\begin{aligned}\mathbb{E}_{X\sim P^n}\left[\|\mathcal{A}_r(X) - \mathbb{E}_{x\sim P}[x]\|^2\right] &= \frac{2dB^2r^2}{\rho} + \mathbb{E}\left\| \sum_{x\in X_{\text{priv}}} r(x - \mathbb{E}x) + \sum_{x\in X_{\text{pub}}} \left(\frac{1-n_{\text{priv}}r}{n_{\text{pub}}}\right)(x - \mathbb{E}x)\right\|^2 \\ &= \frac{2dB^2r^2}{\rho} + \sum_{x\in X_{\text{priv}}} r^2V^2 + \sum_{x\in X_{\text{pub}}} \left(\frac{1-n_{\text{priv}}r}{n_{\text{pub}}}\right)^2 V^2 \\ &= \frac{2dB^2r^2}{\rho} + n_{\text{priv}}r^2V^2 + \frac{(1-n_{\text{priv}}r)^2}{n_{\text{pub}}}V^2,\end{aligned} \tag{36}$$

using independence of the Gaussian noise and the data, basic properties of variance, and the fact that the data is i.i.d.

To prove (34), let

$$J(r) := \frac{2dB^2r^2}{\rho} + n_{\text{priv}}r^2V^2 + \frac{(1-n_{\text{priv}}r)^2}{n_{\text{pub}}}V^2.$$

We compute first and second derivatives of J:

$$\frac{d}{dr}J(r) = 2r\left(\frac{2dB^2}{\rho} + n_{\text{priv}}V^2\right) - 2n_{\text{priv}}\frac{V^2}{n_{\text{pub}}}(1 - n_{\text{priv}}r)$$

and

$$\frac{d^2}{dr^2}J(r) = 2\left(\frac{2dB^2}{\rho} + n_{\text{priv}}V^2\right) + 2n_{\text{priv}}^2\frac{V^2}{n_{\text{pub}}}.$$

Since J is strongly convex, it has a unique minimizer r^* which satisfies $\frac{d}{dr}J(r^*) = 0$. We find

$$r^* = \frac{n_{\text{priv}}V^2}{n_{\text{pub}}}\left(\frac{2dB^2}{\rho} + n_{\text{priv}}V^2 + \frac{n_{\text{priv}}^2V^2}{n_{\text{pub}}}\right)^{-1}.$$

One can verify that $r^* \neq 0$ and $r^* \neq \frac{1}{n}$, since $1 \leqslant n_{\text{priv}} < n$ and $\rho < \infty$ by assumption. Thus, $J(r^*) < \min(J(0), J(1/n))$, which yields (34) by Lemma 7.

To prove (35), we will choose a different r: $r := \frac{K\sqrt{\rho}V}{B\sqrt{dn_{\text{pub}}}}$ for $K > 0$ to be determined. Then by (36), we have

$$\begin{aligned}\mathbb{E}_{X\sim P^n}\left[\|\mathcal{A}_r(X) - \mathbb{E}_{x\sim P}[x]\|^2\right] &= \frac{2dB^2r^2}{\rho} + n_{\text{priv}}r^2V^2 + \frac{(1-n_{\text{priv}}r)^2}{n_{\text{pub}}}V^2 \\ &= K^2\frac{V^2}{n_{\text{pub}}}\left(2 + \frac{n_{\text{priv}}\rho V^2}{dB^2}\right) + \frac{V^2}{n_{\text{pub}}}\left(1 - \frac{KV}{B}\frac{\sqrt{\rho}n_{\text{priv}}}{\sqrt{dn_{\text{pub}}}}\right)^2 \\ &= \frac{V^2}{n_{\text{pub}}}\left(qK^2 + \left(1 - \frac{KV}{B}\frac{\sqrt{\rho}n_{\text{priv}}}{\sqrt{dn_{\text{pub}}}}\right)^2\right),\end{aligned}$$

where $q = 2 + \frac{n_{\text{priv}}\rho V^2}{dB^2}$. Now, letting $s = \frac{Vn_{\text{priv}}\sqrt{\rho}}{B\sqrt{dn_{\text{pub}}}}$ and choosing $K = \frac{s}{q+s^2}$ gives

$$\begin{aligned}\mathbb{E}_{X\sim P^n}\left[\|\mathcal{A}_r(X) - \mathbb{E}_{x\sim P}[x]\|^2\right] &\leqslant \frac{V^2}{n_{\text{pub}}}\left(qK^2 + \left(1 - \frac{KV}{B}\frac{\sqrt{\rho}n_{\text{priv}}}{\sqrt{dn_{\text{pub}}}}\right)^2\right) \\ &\leqslant \frac{V^2}{n_{\text{pub}}}\left(\frac{q^2+qs^2}{q^2+2qs^2+s^4}\right).\end{aligned}$$

Finally, the assumption $\frac{V^2}{n_{\text{pub}}} \leqslant \frac{2dB^2}{\rho n^2}$ implies $\frac{V^2}{n_{\text{pub}}} = \min\left(\frac{V^2}{n_{\text{pub}}}, \frac{2dB^2}{\rho n^2} + \frac{V^2}{n}\right)$, completing the proof.

□

E.2. Optimal Semi-DP Empirical Risk Minimization

Practical Applications of Semi-DP ERM Beyond ML: Semi-DP ERM has numerous applications beyond training ML models. For example, consider semi-DP optimization of energy consumption in smart grids or semi-DP optimization of the total capacity of a multi-user wireless communication system. In these systems, the goal is to optimize the current performance of the system given existing users (e.g., optimize current beamforming strategies in wireless communications). Some users may opt-in to share their data (e.g. electricity consumption pattern) and some users may not. Thus, the problem is naturally a semi-DP ERM problem.

Theorem 38 (Complete statement of Theorem 10)**.** *There exist absolute constants c_0 and C_0, with $0 < c_0 \leqslant C_0$, such that*

$$c_0 LD\min\left\{\frac{n_{priv}}{n}, \frac{d}{n\varepsilon}\right\} \leqslant \mathcal{R}_{ERM}(\varepsilon, n_{priv}, n, d, L, D, \mu = 0) \leqslant C_0 LD\min\left\{\frac{n_{priv}}{n}, \frac{d}{n\varepsilon}\right\}.$$

Further, if $\mu > 0$, then there exist absolute constants $0 < c_1 \leqslant C_1$ such that

$$c_1 LD\min\left\{\frac{n_{priv}}{n}, \frac{d}{n\varepsilon}\right\}^2 \leqslant \mathcal{R}_{ERM}(\varepsilon, n_{priv}, n, d, L, D, \mu) \leqslant C_1\frac{L^2}{\mu}\min\left\{\frac{n_{priv}}{n}, \frac{d\sqrt{\ln(n)}}{n\varepsilon}\right\}^2.$$

Proof. **Lower Bounds:** Given a lower bound for empirical mean estimation, Bassily et al. (Bassily et al., 2014) show how to prove excess risk lower bounds for convex and strongly convex ERM by reducing these problems to mean estimation. Thus, our lower bounds follow immediately by combining the lower bound in Theorem 21 with the reduction in (Bassily et al., 2014). Roughly, the reduction works as follows:

In the strongly convex case, we simply take $f(w, x) = \frac{1}{2}\|w - x\|^2$ on $\mathcal{W} = \mathcal{X} = \mathbb{B}$; $f(\cdot, x)$ is 1-uniformly-Lipschitz and 1-strongly convex. Moreover, for any ε-semi-DP $\mathcal{A}$ with output $w_{\text{priv}} = \mathcal{A}(X)$, we have

$$\mathbb{E}\widehat{F}_X(w_{\text{priv}}) - \widehat{F}_X^* = \frac{1}{2}\mathbb{E}\|\mathcal{A}(X) - \bar{X}\|^2.$$

Applying Theorem 21 and then scaling $f \to \frac{L}{D}f$ and $\mathcal{W} \to D\mathcal{W}$ and $\mathcal{X} \to D\mathcal{X}$ completes the proof.

For the convex case, we take $f(w,x) = -\langle w, x\rangle$ on $\mathcal{W} = \mathcal{X} = \mathbb{B}$. Then $w^* := \operatorname{argmin}_{w\in\mathcal{W}} \widehat{F}_X(w) = \frac{\bar{X}}{\|\bar{X}\|}$ and

$$\widehat{F}_X(w_{\text{priv}}) - \widehat{F}_X^* \geqslant \frac{1}{2}\|\bar{X}\|\|w_{\text{priv}} - w^*\|^2.$$

Also, the proof of Theorem 21 shows that there exists a dataset $X \in \mathcal{X}^n$ such that $\|\bar{X}\| = M/n := \min\left(\frac{n_{\text{priv}}}{n}, \frac{d}{3n\varepsilon}\right)$ and $\mathbb{E}\|\mathcal{A}'(X) - \bar{X}\|^2 \gtrsim \min\left(\frac{n_{\text{priv}}}{n}, \frac{d}{n\varepsilon}\right)^2$ for any ε-semi-DP $\mathcal{A}'$. Note that $\mathcal{A}' := \frac{M}{n} w_{\text{priv}}$ is ε-semi-DP by post-processing. Thus,

$$\begin{aligned}
\mathbb{E}\widehat{F}_X(w_{\text{priv}}) - \widehat{F}_X^* &\geqslant \frac{M}{2n}\mathbb{E}\left[\|w_{\text{priv}} - w^*\|^2\right] \\
&= \frac{M}{2n}\left(\frac{n}{M}\right)^2 \mathbb{E}\left[\|\mathcal{A}'(X) - \bar{X}\|^2\right] \\
&\gtrsim \frac{n}{M}\min\left(\frac{n_{\text{priv}}}{n}, \frac{d}{n\varepsilon}\right)^2 \\
&\gtrsim \min\left(\frac{n_{\text{priv}}}{n}, \frac{d}{n\varepsilon}\right).
\end{aligned}$$

A standard scaling argument (see (Bassily et al., 2014) for details) completes the proof.

Upper Bounds: The second terms in each minimum follows by running the ε-DP algorithms in (Bassily et al., 2014): these achieve the desired excess empirical risk bounds and are automatically ε-semi-DP.

We now prove the first term in each respective minimum. Denote $\widehat{F}_{pub}(w) = \frac{1}{n}\sum_{x\in X_{pub}} f(w,x)$ and $\widehat{F}_{priv}(w) = \frac{1}{n}\sum_{x\in X_{priv}} f(w,x)$, so that $\widehat{F}_X = \widehat{F}_{pub} + \widehat{F}_{priv}$. The algorithm we will use simply returns any minimizer of the public empirical loss: $\mathcal{A}(X) = w^*_{pub} \in \operatorname{argmin}_{w\in\mathcal{W}} \widehat{F}_{pub}(w)$. (It will be easy to see from the proof that any approximate minimizer would also suffice.) $\mathcal{A}$ is clearly ε-semi-DP. Next, we bound the excess risk of $\mathcal{A}$. Let $w^* \in \operatorname{argmin}_{w\in\mathcal{W}} \widehat{F}_X(w)$.

Convex Upper Bound: We have

$$\begin{aligned}
\widehat{F}_X(w^*_{pub}) - \widehat{F}_X(w^*) &= \widehat{F}_X(w^*_{pub}) - \widehat{F}_{pub}(w^*_{pub}) + \widehat{F}_{pub}(w^*_{pub}) - \widehat{F}_{pub}(w^*) \\
&\quad + \widehat{F}_{pub}(w^*) - \widehat{F}_X(w^*) \\
&\leqslant \frac{1}{n}\sum_{x\in X_{priv}} f(w^*_{pub}, x) + 0 - \frac{1}{n}\sum_{x\in X_{priv}} f(w^*, x) \\
&\leqslant \frac{1}{n}\sum_{x\in X_{priv}} L\|w^*_{pub} - w^*\| \\
&= L\|w^*_{pub} - w^*\|\frac{n_{\text{priv}}}{n} \\
&\leqslant LD\frac{n_{\text{priv}}}{n}.
\end{aligned}$$

Strongly Convex Upper Bound: By the above, $\widehat{F}_X(w^*_{pub}) - \widehat{F}_X(w^*) \leqslant L\|w^*_{pub} - w^*\|\frac{n_{\text{priv}}}{n}$. Now we will use strong convexity to bound $\|w^*_{pub} - w^*\|$. To do so, we use the following lemma, versions of which have appeared, e.g. in (Lowy & Razaviyayn, 2021; Chaudhuri et al., 2011):

Lemma 39. *(Lowy & Razaviyayn, 2021) Let $H(w), h(w)$ be convex functions on some convex closed set $\mathcal{W} \subseteq \mathbb{R}^d$ and suppose that $H(w)$ is μ_H-strongly convex. Assume further that h is L_h-Lipschitz. Define $w_1 = \arg\min_{w\in\mathcal{W}} H(w)$ and $w_2 = \arg\min_{w\in\mathcal{W}}[H(w) + h(w)]$. Then $\|w_1 - w_2\| \leqslant \frac{L_h}{\mu_H}$.*

We apply the lemma with $h(w) = \widehat{F}_{priv}(w)$ and $H(w) = \widehat{F}_{pub}(w)$. Then the conditions of the lemma are satisfied with $L_h \leqslant \frac{n_{\text{priv}}}{n} L$ and $\mu_H = \frac{n_{\text{pub}}}{n}\mu$. Thus,

$$\|w^* - w^*_{pub}\| \leqslant \frac{L_h}{\mu_H} \leqslant \frac{L}{\mu}\frac{n_{\text{priv}}}{n_{\text{pub}}}$$

This leads to

$$\widehat{F}_X(w^*_{pub}) - \widehat{F}_X(w^*) \leqslant \frac{L^2}{\mu}\frac{n_{\text{priv}}^2}{nn_{\text{pub}}}.$$

Combining the two strongly convex upper bounds with the upper bound $LD\frac{n_{\text{priv}}}{n} \leqslant \frac{L^2}{\mu}\frac{n_{\text{priv}}}{n}$ (which holds for any convex function), we have an algorithm $\mathcal{A}$ with the following excess risk:

$$\mathbb{E}\widehat{F}_X(\mathcal{A}(X)) - \widehat{F}_X^* \lesssim \frac{L^2}{\mu}\min\left(\frac{n_{\text{priv}}}{n}, \frac{n_{\text{priv}}^2}{nn_{\text{pub}}}, \frac{d^2\ln(n)}{n^2\varepsilon^2}\right). \tag{37}$$

We will show that (37) is equal to the strongly convex upper bound stated in Theorem 10 up to constant factors. First, suppose $n_{\text{pub}} \gtrsim n$: i.e. there is a constant $k > 0$ such that $n_{\text{pub}} \geqslant kn$ for all $n \geqslant 1$. Then, clearly (37) and the strongly convex upper bound stated in Theorem 10 are both equal to $\Theta\left(\frac{L^2}{\mu}\min\left(\frac{n_{\text{priv}}^2}{n^2}, \left(\frac{d}{n\varepsilon}\right)^2\ln(n)\right)\right)$.

Next, suppose $n_{\text{pub}} \ll n$: i.e. for any $k > 0$, there exists $n \geqslant 1$ such that $n_{\text{pub}} < kn$. Then we claim that $\min\left\{\frac{n_{\text{priv}}}{n}, \frac{d\sqrt{\ln(n)}}{n\varepsilon}\right\}^2 \gtrsim \min\left\{1, \frac{d\sqrt{\ln(n)}}{n\varepsilon}\right\}^2$. If we prove this claim, then we are done. There are two subcases to consider: A) $n_{\text{priv}}/n < \frac{d\sqrt{\ln(n)}}{\varepsilon n}$; and B) $n_{\text{priv}}/n \geqslant \frac{d\sqrt{\ln(n)}}{\varepsilon n}$. In subcase B), the claim is immediate. Consider subcase A): if $n_{\text{priv}} \gtrsim n$, then we're done. If not, then we have $n_{\text{priv}} \ll n$ and $n_{\text{pub}} \ll n$, so $n = n_{\text{priv}} + n_{\text{pub}} \ll n$, a contradiction. This completes the proof. □

Remark 40 (Details of Remark 11). *The same minimax risk bound (7) holds up to a logarithmic factor if we replace $\mathcal{F}_{\mu=0,L,D}$ by the larger class of all Lipschitz non-convex (or convex) loss functions in the definition (6): First, the lower bound in Theorem 10 clearly still holds for non-convex loss functions. For the upper bound, the ε-DP (hence semi-DP) exponential mechanism achieves error $O(LD\frac{d}{n\varepsilon}\ln(n))$ (Bassily et al., 2014; Ganesh et al., 2022). Further, the proof of Theorem 10 reveals that convexity is not necessary for the throw-away algorithm to achieve error $O(LDn_{priv}/n)$. However, the optimal algorithms are inefficient for non-convex loss functions: to the best of our knowledge, all existing polynomial time implementations of the exponential mechanism require convexity for their runtime guarantees to hold. Further, computing $\approx \text{argmin}_{w\in\mathcal{W}}\widehat{F}_{pub}(w)$ in the implementation of throw-away may not be tractable in polynomial time for non-convex $\widehat{F}_{pub}$.*

E.3. Optimal Semi-DP Stochastic Convex Optimization

Approximate (ε, δ)-Semi-DP SCO

Theorem 41 (Complete Version of Theorem 12). *Let $\varepsilon \lesssim 1/\log(nd)$ and $\delta \ll 1/n$. Then, there is a constant $C > 0$ such that*

$$\ell(d,n)LD\min\left\{\frac{1}{\sqrt{n_{pub}}}, \frac{\sqrt{d}}{n\varepsilon} + \frac{1}{\sqrt{n}}\right\} \leqslant \mathcal{R}_{SCO}(\varepsilon, \delta, n_{priv}, n, d, L, D, \mu = 0) \leqslant CLD\min\left\{\frac{1}{\sqrt{n_{pub}}}, \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon} + \frac{1}{\sqrt{n}}\right\},$$

and

$$\ell(d,n)LD\min\left\{\frac{1}{\sqrt{n_{pub}}}, \frac{\sqrt{d}}{n\varepsilon} + \frac{1}{\sqrt{n}}\right\}^2 \leqslant \mathcal{R}_{SCO}(\varepsilon, \delta, n_{priv}, n, d, L, D, \mu) \leqslant C\frac{L^2}{\mu}\min\left\{\frac{1}{\sqrt{n_{pub}}}, \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon} + \frac{1}{\sqrt{n}}\right\}^2,$$

where $1/\ell(d,n)$ is logarithmic in d and n. Our lower bounds hold for symmetric $\mathcal{A} = (\mathcal{A}^1, \ldots, \mathcal{A}^d)$.

Proof. **Lower bounds:** Let $\mathcal{A}$ be (ε, δ)-semi-DP and symmetric, and denote $w_{\text{priv}} = \mathcal{A}(X)$.

Strongly convex lower bounds: We begin with the strongly convex lower bounds, which can be proved by reducing strongly convex SCO to mean estimation and applying Theorem 4. In a bit more detail, let $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ be given by

$$f(w, x) = \frac{L}{2D}\|w - x\|^2,$$

where $\mathcal{W} = \mathcal{X} = D\mathbb{B}$. Note that $f(\cdot, x)$ is L-uniformly Lipschitz and $\frac{L}{D}$-strongly convex in w for all x. Further, $w^* := \text{argmin}_{w\in\mathcal{W}}\{F(w) = \mathbb{E}_{x\sim P}[f(w,x)]\} = \mathbb{E}_{x\sim P}[x]$. By a direct calculation (see e.g. (Kamath et al., 2022a, Lemma 6.2)), we have

$$\mathbb{E}F(w_{\text{priv}}) - F(w^*) = \frac{L}{2D}\mathbb{E}\|w_{\text{priv}} - w^*\|^2. \tag{38}$$

We can lower bound $\mathbb{E}\|w_{\text{priv}} - w^*\|^2 = \mathbb{E}\|\mathcal{A}(X) - \mathbb{E}_{x\sim P}[x]\|^2$ via Theorem 4 (and its proof, to account for the re-scaling). Specifically,

$$\mathbb{E}\|w_{\text{priv}} - w^*\|^2 \geqslant \ell(d,n)D^2 \min\left\{\frac{1}{n_{\text{pub}}}, \frac{d}{n^2\varepsilon^2} + \frac{1}{n}\right\},$$

for a logarithmic function $\ell(d,n)$ of d and n, by Theorem 4. Applying (38) yields the desired excess risk lower bound for $\delta > 0$.

Convex lower bounds: We will begin by proving the lower bounds for the case in which $L = D = 1$, and then scale our construction to get the lower bounds for arbitrary L, D.

Let $\mathcal{X} = \left\{\pm\frac{1}{\sqrt{d}}\right\}^d \subset \mathbb{R}^d$ and $\mathcal{W} = \mathbb{B}$. Define

$$f(w,x) = -\langle w, x\rangle,$$

which is convex and 1-uniformly-Lipschitz in w on $\mathcal{X}$. Let P_θ be the hard distribution used to prove Theorem 23, which satisfies $\mathbb{E}_{x\sim P_\theta}[x] = \theta \in [-a/\sqrt{d}, a/\sqrt{d}]^d$ and $\|\theta\| \leqslant a = \min\left(\frac{1}{\sqrt{n_{\text{pub}}}}, \frac{\sqrt{d}}{n_{\text{priv}}\varepsilon} + \frac{1}{n}\right)$. Further, $w_\theta^* = \operatorname{argmin}_{w\in\mathcal{W}}[F_\theta(w) := \mathbb{E}_{x\sim P_\theta} f(w,x) = -\langle w, \theta\rangle] = \frac{\theta}{\|\theta\|}$. A direct calculation (see e.g. (Kamath et al., 2022a, Equation 14)) shows

$$\sup_\theta \mathbb{E}F_\theta(w_{\text{priv}}) - F_\theta^* \geqslant \frac{1}{2}\mathbb{E}\left[\|\theta\|\|w_{\text{priv}} - w_\theta^*\|^2\right] \tag{39}$$

$$= \frac{1}{2}\sup_\theta \mathbb{E}\left[\frac{1}{\|\theta\|}\|\tilde{w}_{\text{priv}} - \theta\|^2\right], \tag{40}$$

where $\tilde{w}_{\text{priv}} = \widetilde{\mathcal{A}}_\theta(X)$ is the output of the algorithm $\widetilde{\mathcal{A}}_\theta : \mathcal{X}^n \to \|\theta\|\mathcal{W}$ defined by $\widetilde{\mathcal{A}}_\theta(X) := \|\theta\|\mathcal{A}(X)$. Note that $\widetilde{\mathcal{A}}_\theta$ is (ε,δ)-semi-DP by post-processing, for any θ. Now, we invoke Theorem 23 to obtain

$$\sup_\theta \mathbb{E}\left[\|\tilde{w}_{\text{priv}} - \theta\|^2\right] \geqslant \tilde{\Omega}\left(\min\left\{\frac{1}{n_{\text{pub}}}, \frac{d}{n_{\text{priv}}^2\varepsilon^2} + \frac{1}{n}\right\}\right)$$

for $\|\theta\| \leqslant \min\left(\frac{1}{\sqrt{n_{\text{pub}}}}, \frac{\sqrt{d}}{n_{\text{priv}}\varepsilon} + \frac{1}{n}\right)$. This implies the desired lower bound when $L = D = 1$.

For general L and D, we scale the problem instance as follows: let $\widetilde{W} = D\mathcal{W}$, $\widetilde{X} = L\mathcal{X}$, and $\widetilde{x} \sim \widetilde{P_\theta} \iff \widetilde{x} = Lx$ for $x \sim P_\theta$. Define $\widetilde{f} : \widetilde{\mathcal{W}} \times \widetilde{\mathcal{X}} \to \mathbb{R}$ by $\widetilde{f}(\widetilde{w},\widetilde{x}) := f(\widetilde{w},\widetilde{x}) = -\langle\widetilde{w},\widetilde{x}\rangle$. Then $\widetilde{f}(\cdot,\widetilde{x})$ is L-Lipschitz and convex. Moreover, if $F(w) = \mathbb{E}_{x\sim P_\theta}[f(w,x)]$, $\widetilde{F}(\widetilde{w}) = \mathbb{E}_{\widetilde{x}\sim\widetilde{P}}[f(\widetilde{w},\widetilde{x})]$, $\widetilde{w} = Dw$, and $\widetilde{\theta} = \mathbb{E}_{\widetilde{x}\sim\widetilde{P_\theta}}[\widetilde{x}] = L\theta$, then $Dw^* \in \operatorname{argmin}_{\widetilde{w}\in\widetilde{\mathcal{W}}}\widetilde{F}(\widetilde{w})$ and

$$\begin{aligned}\widetilde{F}(\widetilde{w}) - \widetilde{F}^* &= -\langle\widetilde{w},\widetilde{\theta}\rangle + \langle\widetilde{w}^*,\widetilde{\theta}\rangle \\ &= D\langle\widetilde{\theta}, w^* - w\rangle \\ &= LD\langle\theta, w^* - w\rangle \\ &= LD\left[F(w) - F^*\right].\end{aligned}$$

This shows that excess risk scales by LD, completing the lower bound proofs.

Upper bounds: Convex upper bounds: Consider the 0-semi-DP throw-away algorithm that discards X_{priv} and runs n_{priv} steps of one-pass SGD (stochastic approximation) using X_{pub}. This algorithm has excess risk $O\left(LD/\sqrt{n_{\text{pub}}}\right)$ (Nemirovski & Yudin, 1983). To obtain the second term in the convex (ε,δ)-semi-DP upper bound, one can use, e.g. (ε,δ)-DP-SGD (Bassily et al., 2019).

Strongly convex upper bounds: Consider the 0-semi-DP throw-away algorithm that discards X_{priv} and runs n_{priv} steps of one-pass SGD (stochastic approximation) using X_{pub}. This algorithm has excess risk $O\left(\frac{L^2}{\mu n_{\text{pub}}}\right)$ (Nemirovski & Yudin, 1983). The second term in the strongly convex (ε,δ)-semi-DP upper bound can be attained, e.g. by (ε,δ)-DP-SGD (Lowy & Razaviyayn, 2023b). □

Next, we provide minimax optimal excess risk bounds for pure ε-semi-DP SCO.

Pure ε-Semi-DP SCO

Theorem 42 (Pure ε-Semi-DP SCO). *Suppose $\varepsilon \leqslant d/8$, and either $n_{pub} \lesssim \frac{n\varepsilon}{d}$ or $d \lesssim 1$. If $\mu = 0$ (convex case), then there exist absolute constants $0 < c_0 \leqslant C_0$ such that*

$$c_0 LD \min\left\{\frac{1}{\sqrt{n_{pub}}}, \frac{d}{n\varepsilon} + \frac{1}{\sqrt{n}}\right\} \leqslant \mathcal{R}_{SCO}(\varepsilon, \delta = 0, n_{priv}, n, d, L, D, \mu = 0) \leqslant C_0 LD \min\left\{\frac{1}{\sqrt{n_{pub}}}, \frac{d}{n\varepsilon} + \frac{1}{\sqrt{n}}\right\}.$$

If $\mu > 0$, there are constants $0 < c_1 \leqslant C_1$ such that

$$c_1 LD \min\left\{\frac{1}{n_{pub}}, \frac{d^2}{n^2\varepsilon^2} + \frac{1}{n}\right\} \leqslant \mathcal{R}_{SCO}(\varepsilon, \delta = 0, n_{priv}, n, d, L, D, \mu) \leqslant C_1 \frac{L^2}{\mu} \min\left\{\frac{1}{n_{pub}}, \frac{d^2 \ln(n)}{n^2\varepsilon^2} + \frac{1}{n}\right\}.$$

The above upper bounds hold for any n_{pub}, d.

Proof of Theorem 42. **Lower bounds:** Let $\mathcal{A}$ be ε-semi-DP and denote $w_{\text{priv}} = \mathcal{A}(X)$.

Strongly convex lower bound: We begin with the strongly convex lower bound, which can be proved by reducing strongly convex SCO to mean estimation and applying Theorem 32. In a bit more detail, let $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ be given by

$$f(w, x) = \frac{L}{2D}\|w - x\|^2,$$

where $\mathcal{W} = \mathcal{X} = D\mathbb{B}$. Note that f is L-uniformly Lipschitz and $\frac{L}{D}$-strongly convex in w for all x. Further, $w^* := \operatorname{argmin}_{w\in\mathcal{W}} \{F(w) = \mathbb{E}_{x\sim P}[f(w, x)]\} = \mathbb{E}_{x\sim P}[x]$. By a direct calculation (see e.g. (Kamath et al., 2022a, Lemma 6.2)), we have

$$\mathbb{E}F(w_{\text{priv}}) - F(w^*) = \frac{L}{2D}\mathbb{E}\|w_{\text{priv}} - w^*\|^2. \tag{41}$$

We can lower bound $\mathbb{E}\|w_{\text{priv}} - w^*\|^2 = \mathbb{E}\|\mathcal{A}(X) - \mathbb{E}_{x\sim P}[x]\|^2$ via Theorem 32 (and its proof, to account for the re-scaling). Specifically, if $\delta = 0$, $\varepsilon \leqslant \max(1, d/8)$, and either $d = O(1)$ or $n_{\text{pub}} = O(n\varepsilon/d)$, then Theorem 32 and its proof imply

$$\mathbb{E}\|w_{\text{priv}} - w^*\|^2 \geqslant cD^2 \min\left(\frac{1}{n_{\text{pub}}}, \frac{d^2}{n^2\varepsilon^2} + \frac{1}{n}\right).$$

Combining this with (41) leads to the desired excess risk lower bound for $\delta = 0$.

Convex lower bound: We will begin by proving the lower bound for the case in which $L = D = 1$, and then scale our construction to get the lower bounds for arbitrary L, D.

Assume $\delta = 0$, $\varepsilon \leqslant d/8$, and either $n_{\text{pub}} \lesssim n\varepsilon/d$ or $d \lesssim 1$. Let $\mathcal{X} = \{0\} \bigcup \left\{\pm\frac{1}{\sqrt{d}}\right\}^d \subset \mathbb{R}^d$ and $\mathcal{W} = \mathbb{B}$. Define

$$f(w, x) = -\langle w, x\rangle,$$

which is convex and 1-uniformly-Lipschitz in w on $\mathcal{X}$. Choose $\mathcal{V}$ to be a finite subset of $\mathbb{R}^d$ such that $|\mathcal{V}| \geqslant 2^{d/2}$, $\|v\| = 1$ for all v, and $\|v - v'\| \geqslant 1/8$ for all $v \neq v'$ (see e.g. the Gilbert-Varshamov construction). Following the proof of Theorem 4, we define $P_{\theta_v} = (1 - p)P_0 + pP_v$ for all $v \in \mathcal{V}$, where $p \in [0, 1]$ will be chosen later, P_0 is point mass on $\{X = 0\}$ and P_v is point mass on $\{X = v\}$. Denote the mean $\theta_v := \mathbb{E}_{x\sim P_{\theta_v}}[x] = pv$. Note that $\|\theta_v\| = p$ for all v. Let $F_v(w) = \mathbb{E}_{x\sim P_{\theta_v}}[f(w, x)]$ and $w_v^* \in \operatorname{argmin}_{w\in\mathcal{W}} F_v(w) = \frac{\theta_v}{\|\theta_v\|} = v$. A direct calculation (see e.g. (Kamath et al., 2022a, Equation 14)) shows

$$\mathbb{E}F_v(w) - F_v^* \geqslant \frac{1}{2}\mathbb{E}\left[\|\theta_v\|\|w - w_v^*\|^2\right] \tag{42}$$

for any $w \in \mathcal{W}, v \in \mathcal{V}$. Also,

$$\rho^*(\mathcal{V}) = \min\left\{\|w_v^* - w_{v'}^*\| : v, v' \in \mathcal{V}, v \neq v'\right\} = \min\left\{\|v - v'\| : v, v' \in \mathcal{V}, v \neq v'\right\} \geqslant \frac{1}{8}.$$

Thus, by combining (42) with the reduction from estimation to testing and Theorem 33 (see the proof of Theorem 4 for details), we have

$$\begin{aligned}
\sup_{v\in\mathcal{V}} \mathbb{E}\left[F_v(w_{\text{priv}}) - F_v^*\right] &\geqslant \frac{1}{2}\sup_{v\in\mathcal{V}} \mathbb{E}\left[\|\theta_v\|\|w_{\text{priv}} - w_v^*\|^2\right] \\
&= \frac{p}{2}\sup_{v\in\mathcal{V}} \mathbb{E}\left[\|w_{\text{priv}} - w_v^*\|^2\right] \\
&\geqslant \frac{p}{2}\rho^*(\mathcal{V})^2 \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} P_{\theta_v}\left(\|\widehat{\theta}(X) - \theta_v\| \geqslant \rho^*(\mathcal{V})\right) \\
&\geqslant \frac{p}{128}\frac{(|\mathcal{V}|-1)e^{-\varepsilon\lceil n_{\text{priv}}p\rceil}(1-p)^{n_{\text{pub}}}}{2\left(1+(|\mathcal{V}|-1)e^{-\varepsilon\lceil n_{\text{priv}}p\rceil}\right)} \\
&\geqslant \frac{p}{128}\frac{(2^{d/2}-1)e^{-\varepsilon\lceil n_{\text{priv}}p\rceil}(1-p)^{n_{\text{pub}}}}{2\left(1+(2^{d/2}-1)e^{-\varepsilon\lceil n_{\text{priv}}p\rceil}\right)} \\
&\geqslant \frac{p}{512}(1-p)^{n_{\text{pub}}}\min\left\{1, \frac{2^{d/2}-1}{e^{\varepsilon(n_{\text{priv}}p+1)}}\right\}.
\end{aligned}$$

Now, assume $d \geqslant 4$ so that $2^{d/2} - 1 \geqslant e^{d/4}$. Then, as detailed in the proof of Theorem 4, choosing

$$p = \min\left(\frac{d}{4n\varepsilon} - \frac{1}{n}, \frac{1}{2\sqrt{n_{\text{pub}}}}\right)$$

and assuming $n_{\text{pub}} \leqslant kn\varepsilon/d$ for some absolute constant k implies

$$\sup_{v\in\mathcal{V}} \mathbb{E}\left[F_v(w_{\text{priv}}) - F_v^*\right] \geqslant c\min\left(\frac{1}{\sqrt{n_{\text{pub}}}}, \frac{d}{n\varepsilon}\right)$$

for some absolute constant $c > 0$. Combining this with the non-private SCO lower bound (Nemirovski & Yudin, 1983) yields

$$\sup_{P} \mathbb{E}\left[F(w_{\text{priv}}) - F^*\right] \geqslant c'\min\left(\frac{1}{\sqrt{n_{\text{pub}}}}, \frac{d}{n\varepsilon} + \frac{1}{\sqrt{n}}\right),$$

where $F(w) := \mathbb{E}_{x\sim P}[f(w,x)]$.

Suppose instead that $0 \leqslant \delta \leqslant \varepsilon$ and $d \lesssim 1$ (i.e. $d \leqslant k$ for some absolute constant $k \geqslant 1$), but $n_{\text{pub}} \in [n]$ is arbitrary. We will prove the lower bound for $d = 1$; by taking the k-fold product distribution, this is sufficient to complete the proof of the unscaled ε-semi-DP lower bound. Define distributions P_1, P_2 on $\{-1, 1\}$ as follows:

$$P_1(-1) = P_2(1) = \frac{1+\gamma}{2}, \quad P_1(1) = P_2(-1) = \frac{1-\gamma}{2}$$

for some $\gamma \in (0, 1/2]$ to be chosen later. Note $\theta_1 := \mathbb{E}_{P_1}[x] = -\gamma$ and $\theta_2 := \mathbb{E}_{P_2}[x] = \gamma$, so $|\theta_j| = \gamma$ for $j = 1,\ 2$. Let $F_j(w) = \mathbb{E}_{x\sim P_j} f(w,x)$ and $w_j^* = \frac{\theta_j}{|\theta_j|} = \frac{\theta_j}{\gamma} \in \operatorname{argmin}_{w\in\mathcal{W}} F_j(w)$. Then by (42), we have

$$\begin{aligned}
\max_{j\in\{1,2\}} \mathbb{E}\left[F_j(w_{\text{priv}}) - F_j^*\right] &\geqslant \frac{1}{2}\max_{j\in\{1,2\}} \mathbb{E}\left[|\theta_j||w_{\text{priv}} - w_j^*|^2\right] \\
&= \frac{\gamma}{2}\max_{j\in\{1,2\}} \mathbb{E}\left[|w_{\text{priv}} - w_j^*|^2\right] \\
&= \frac{1}{2\gamma}\max_{j\in\{1,2\}} \mathbb{E}\left[|w'_{\text{priv}} - \theta_j|^2\right],
\end{aligned}$$

where $w'_{\text{priv}} := \gamma w_{\text{priv}}$ is semi-DP iff w_{priv} is semi-DP (by post-processing). Thus, by applying Le Cam's method and Lemma 35 (see the proof of Theorem 4 for details), we get

$$\max_{j\in\{1,2\}} \mathbb{E}\left[F_j(w_{\text{priv}}) - F_j^*\right] \geqslant \frac{1}{2\gamma}\left[\frac{\gamma^2}{8}\left(1 - \gamma\min\left(\sqrt{\frac{3n}{2}}, 6n_{\text{priv}}\varepsilon + \sqrt{\frac{3n_{\text{pub}}}{2}}\right)\right)\right].$$

Now we will choose γ to (approximately) maximize the right-hand side of the above inequality. If $\min\left\{\sqrt{\frac{3n}{2}}, 6n_{\text{priv}}\varepsilon + \sqrt{\frac{3n_{\text{pub}}}{2}}\right\} = \sqrt{\frac{3n}{2}}$, then choosing $\gamma = \frac{1}{3}\sqrt{\frac{2}{3n}}$ yields

$$\max_{j\in\{1,2\}} \mathbb{E}\left[F_j(w_{\text{priv}}) - F_j^*\right] \geqslant \frac{k}{\sqrt{n}}$$

for some absolute constant $k > 0$. If instead $\min\left\{\sqrt{\frac{3n}{2}}, 6n_{\text{priv}}\varepsilon + \sqrt{\frac{3n_{\text{pub}}}{2}}\right\} = 6n_{\text{priv}}\varepsilon + \sqrt{\frac{3n_{\text{pub}}}{2}}$, then we choose $\gamma = \frac{2}{3}\left(6n_{\text{priv}}\varepsilon + \sqrt{\frac{3n_{\text{pub}}}{2}}\right)^{-1}$. This choice implies

$$\max_{j\in\{1,2\}} \mathbb{E}\left[F_j(w_{\text{priv}}) - F_j^*\right] \geqslant k' \min\left(\frac{1}{n_{\text{priv}}\varepsilon}, \frac{1}{\sqrt{n_{\text{pub}}}}\right)$$

for some absolute constant $k' > 0$. Combining the pieces above with the non-private SCO lower bound (Nemirovski & Yudin, 1983) yields

$$\sup_P \mathbb{E}\left[F(w_{\text{priv}}) - F^*\right] \geqslant c \min\left(\frac{1}{\sqrt{n_{\text{pub}}}}, \frac{1}{n\varepsilon} + \frac{1}{\sqrt{n}}\right),$$

where $F(w) := \mathbb{E}_{x\sim P}[f(w,x)]$.

A standard scaling argument completes the lower bound proofs (see e.g. the proof of Theorem 41 for details).

Upper bounds: Convex upper bounds: Consider the 0-semi-DP throw-away algorithm that discards X_{priv} and runs n_{priv} steps of one-pass SGD (stochastic approximation) using X_{pub}. This algorithm has excess risk $O\left(LD/\sqrt{n_{\text{pub}}}\right)$ (Nemirovski & Yudin, 1983). To obtain the second term in the convex ε-semi-DP upper bound, use the ε-DP (hence semi-DP) regularized exponential mechanism of Ganesh et al. (2022).

Strongly convex upper bounds: Consider the 0-semi-DP throw-away algorithm that discards X_{priv} and runs n_{priv} steps of one-pass SGD (stochastic approximation) using X_{pub}. This algorithm has excess risk $O\left(\frac{L^2}{\mu n_{\text{pub}}}\right)$ (Nemirovski & Yudin, 1983). To obtain the second term in the strongly convex ε-semi-DP upper bound, one can use, e.g. the ε-DP (hence semi-DP) iterated exponential mechanism of Ganesh et al. (2022). □

E.3.1. Semi-DP SCO with an "Even More Optimal" Gradient Estimator

Proposition 43. *We provide privacy guarantees for Algorithm 1:*

1. *Suppose we sample with replacement in line 4 of Algorithm 1. Then, there exist constants c_1, c_2 such that for any $\varepsilon < c_1\left(\frac{K_{priv}}{n_{priv}}\right)^2 T$, Algorithm 1 is (ε,δ)-semi-DP for any $\delta > 0$ if we choose $\sigma^2 \geqslant c_2 \frac{C^2 \ln(1/\delta) T}{\varepsilon^2 n_{priv}^2}$.*
2. *Suppose we sample without replacement in line 4 and choose $T \leqslant \frac{n}{K_{priv}}$. Then Algorithm 1 is ρ-semi-zCDP if $\sigma^2 \geqslant \frac{2C^2}{\rho K_{priv}^2}$.*

Proof. Note that the ℓ_2-sensitivity of the private stochastic gradient query is

$$\Delta = \sup_{X_{\text{priv}}\sim X'_{\text{priv}}} \left\| \frac{\alpha}{K_{\text{priv}}} \sum_{x\in B_t^{priv}} \text{clip}_C(\nabla f(w_t, x)) - \frac{\alpha}{K_{\text{priv}}} \sum_{x'\in B_t^{priv}} \text{clip}_C(\nabla f(w_t, x')) \right\|_2 \leqslant \frac{2\alpha C}{K_{\text{priv}}}.$$

1. Consider sampling with replacement. Then we are randomly subsampling from the private data uniformly with sampling ratio K_{priv}/a. Thus, the theorem follows from (Abadi et al., 2016, Theorem 1).

2. Consider sampling without replacement. Then by the ρ-zCDP guarantee of the Gaussian mechanism (Bun & Steinke, 2016, Proposition 1.6) and the sensitivity bound above, $\widetilde{g}_t$ is ρ-semi-zCDP for every t. Moreover, since we are sampling without replacement, the privacy of every $x \in X_{\text{priv}}$ is only affected by $\widetilde{g}_t$ for a single $t \in [T]$. Thus, semi-zCDP of Algorithm 1 follows by parallel composition (McSherry, 2009). □

Excess risk of By Proposition 8, there exists a choice of α such that the variance of our unbiased estimator in line 7 is always less than the variance of both the throw-away gradient estimator $\frac{1}{K_{\text{pub}}}\sum_x \nabla f(w_t, x)$ and the DP-SGD estimator $\frac{1}{K_{\text{priv}}+K_{\text{pub}}}\sum_{x\in B_t}\nabla f(w_t,x) + u_t$, where u_t is appropriately scaled (to ensure DP) Gaussian noise. Consequently, if we choose T and $K = K_{\text{priv}} + K_{\text{pub}}$ such that $n = TK$ and sample without replacement (i.e. one pass), then *Algorithm 1 always has smaller excess risk than both the throw-away SCO algorithm and one-pass DP-SGD*. Moreover, if the loss function has Lipschitz continuous gradient, then one can combine the stochastic gradient estimator of line 7 with *acceleration* (Ghadimi & Lan, 2012) to obtain a linear-time semi-DP algorithm that always outperforms the accelerated DP algorithm of Lowy & Razaviyayn (2023b). This is because the variance of our gradient estimator (hence our excess risk) is strictly smaller than that of Lowy & Razaviyayn (2023b), by Theorem 8. For example, for β-smooth, μ-strongly convex loss functions, one-pass Algorithm 1 achieves excess risk that is optimal up to a factor of $O(\sqrt{\beta/\mu})$ and improves over (Lowy & Razaviyayn, 2023b). Moreover, (Lowy & Razaviyayn, 2023b) has the smallest excess risk among *linear-time* (one-pass) DP algorithms whose *privacy analysis does not require convexity*. Thus, our algorithm can be used for deep learning. In our numerical experiments, we implement the with-replacement sampling version of Algorithm 1.

We also note that near-optimal excess risk bounds for *non-convex* loss functions that satisfy the *(Proximal) PL inequality* (Polyak, 1963; Karimi et al., 2016) can be derived by combining a proximal variation of Algorithm 1 with the techniques of Lowy et al. (2023a). Further, if $f(\cdot, x)$ is not uniformly Lipschitz, but has stochastic gradients with bounded k-th order moment for some $k \geqslant 2$, then excess risk bounds can still be derived for Algorithm 1 via techniques in (Lowy & Razaviyayn, 2023a). Our algorithm can also be extended to a variation of noisy stochastic gradient descent ascent, which could be used, e.g. for *fair* semi-DP model training (Lowy et al., 2023b). We leave it as future work to explore these and other potential applications of our gradient estimator in efficiently training private ML models with public data.

F. Optimal Locally Private Model Training with Public Data

Notation and Setup: Following Duchi & Rogers (2019), we permit algorithms to be *fully interactive*. That is, algorithms may adaptively query the same individual i multiple times over the course of T "communication rounds." We denote i's message in round t by $Z_{i,t} \in \mathcal{Z}$. Person i's message $Z_{i,t} \in \mathcal{Z}$ in round t may depend on all previous communications $B^{(t)} := (Z_{\leqslant n,t}, B^{(t-1)})$ and on i's own data: $Z_{i,t} \sim Q_{i,t}(\cdot|x_i, Z_{<i,t}, B^{(t-1)})$. If i's data is private, then $Z_{i,t}$ is a randomized view of x_i distributed (conditionally) according to $Q_{i,t}$. If i's data is public, then $Z_{i,t}$ may be deterministic. Full interactivity is the most general notion of interactivity. If $T = 1$, then we say the algorithm is *sequentially interactive*. If, in addition, each person's message $Z_{i,1}$ depends only on x_i and not on $x_{j\neq i}$, then we say the algorithm is *non-interactive*. Semi-LDP (Definition 13) essentially requires that the messages $\{Z_{i,t}\}_{t\in[T]}$ be DP for all private $x_i \in X_{\text{priv}}$.

F.1. Optimal Semi-LDP Mean Estimation

Theorem 44 (Re-statement of Theorem 14)**.** *Let $\varepsilon \in (0, 1]$. There are absolute constants $0 < c \leqslant C$ s.t.*

$$c \min\left\{\frac{1}{n_{pub}}, \frac{d}{n\varepsilon^2}\right\} \leqslant \mathcal{M}^{loc}_{pop}(\varepsilon, n_{priv}, n, d) \leqslant C \min\left\{\frac{1}{n_{pub}}, \frac{d}{n\varepsilon^2}\right\}.$$

Proof. **Lower bound:** We will actually prove a more general lower bound than the one in Theorem 14; namely, we will show a lower bound on the minimax ℓ_1-error for estimation of distributions on $\mathcal{X}_r = \{\pm r\}^d$ for $r > 0$. To that end, let $\gamma \in (0, 1)$ and

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

and

$$P_{-1} := \begin{cases} r & \text{with probability } \frac{1-\gamma}{2} \\ -r & \text{with probability } \frac{1+\gamma}{2} \end{cases}.$$

We define our hard distribution on $\mathcal{X}_r$ by first drawing $V \sim \text{Unif}(\{\pm 1\}^d)$ and then—conditional on $V = v$—drawing $X_{i,j} \sim P_v = \Pi_{j=1}^d P_{v_j}$ for $i \in [n], j \in [d]$, where P_v denotes the product distribution. We have Markov chains $V_j \to X_{i,j} \to Z$ for all $j \in [d], i \in [n]$, where Z is the semi-LDP transcript. Note that $\left|\ln\left(\frac{dP_1}{dP_{-1}}\right)\right| \leqslant \ln\left(\frac{1+\gamma}{1-\gamma}\right) \triangleq b$ and $e^b \leqslant 3$ for any $\gamma \in (0, 1/2]$. Now we will use the following lemma from Duchi & Rogers (2019):

Lemma 45. *(Duchi & Rogers, 2019, Lemma 24) Let $V \to X \to Z$ be a Markov chain, where $X \sim P_v$ conditional on $V = v$. If $\left|\ln \frac{dP_v}{dP_{v'}}\right| \leqslant \alpha$ for all v, v', then*

$$I(V; Z) \leqslant 2(e^\alpha - 1)^2 I(X; Z).$$

Thus, for $V_j \sim \text{Unif}(\{\pm 1\})$, Lemma 45 implies $I(V_j; Z) \leqslant \frac{8\gamma^2}{(1-\gamma)^2} I(X_{i,j}; Z)$. Hence the strong data processing constant (Duchi & Rogers, 2019, Definition 9) is $\beta := \beta(P_1, P_{-1}) \leqslant \frac{8\gamma^2}{(1-\gamma)^2}$.

Now, $\theta_{v_j} := \mathbb{E}_{x \sim P_{v_j}}[x] = \gamma r v_j$ for any $v_j \in \{\pm 1\}$. Moreover, letting $\theta_v = (\theta_{v_1}, \cdots, \theta_{v_d})$ for $v \in \{\pm 1\}^d$ and $\theta \in \mathbb{R}^d$, we have

$$\|\theta - \theta_v\|_1 = \sum_{j=1}^d |\theta_j - r\gamma v_j| = r\gamma \sum_{j=1}^d \left|\frac{\theta_j}{r\gamma} - v_j\right| \geqslant r\gamma \sum_{j=1}^d \mathbb{1}_{\{\text{sign}(\theta_j) \neq v_j\}}.$$

Thus, $\{\pm 1\}^d$ induces an $r\gamma$-Hamming separation, so Assouad's lemma (Duchi et al., 2018, Lemma 1) yields

$$\inf_{\mathcal{A} \in \mathbb{A}_\varepsilon^{\text{loc}}} \sup_{P \in \mathcal{P}_r} \mathbb{E}\|\mathcal{A}(X) - \theta(P)\|_1 \geqslant r\gamma \sum_{j=1}^d \inf_{\hat{V}} \mathbb{P}(\hat{V}_j(Z) \neq V_j),$$

where Z is the communication transcript of $\mathcal{A}$, the infimum on the RHS is over all estimators of V, $\theta(P) = \mathbb{E}_{x \sim P}[x]$, and $\mathcal{P}_r$ is the set of distributions on $\mathcal{X}_r$

Assume WLOG that the private samples are the first n_{priv} samples of X: $X_{\text{priv}} = (x_1, \cdots, x_{n_{\text{priv}}})$. To lower bound $\sum_{j=1}^d \inf_{\hat{V}} \mathbb{P}(\hat{V}_j(Z) \neq V_j)$, we use a slight extension of Duchi & Rogers (2019, Theorem 10):

$$\sum_{j=1}^d \inf_{\hat{V}} \mathbb{P}(\hat{V}_j(Z) \neq V_j) \geqslant \frac{d}{2}\left[1 - \sqrt{\frac{7(e^b+1)}{d}\beta\left(I(X_{\text{priv}}; Z|V) + I(X_{\text{pub}}; Z|V)\right)}\right].$$

This follows since $V \to X_{\text{priv}} \to Z$ and $V \to X_{\text{pub}} \to Z$ are both Markov chains and the other assumptions in (Duchi & Rogers, 2019, Theorem 10) all hold. Combining this bound with Assouad's lemma (Duchi et al., 2018, Lemma 1) and substituting the definitions of b and β given above gives us

$$\inf_{\mathcal{A} \in \mathbb{A}_\varepsilon^{\text{loc}}} \sup_{P \in \mathcal{P}_r} \mathbb{E}\|\mathcal{A}(X) - \theta(P)\|_1 \geqslant \frac{r\gamma d}{2}\left[1 - \sqrt{\frac{896}{d}\gamma^2\left(I(X_{\text{priv}}; Z|V) + I(X_{\text{pub}}; Z|V)\right)}\right]$$

for any $\gamma \in (0, 1/2]$. It remains to upper bound the conditional mutual information $I(X_{\text{priv}}; Z|V)$ and $I(X_{\text{pub}}; Z|V)$.

Now for any ε-semi-LDP algorithm with communication transcript Z, we have $I(X_{\text{priv}}; Z|V) \leqslant n_{\text{priv}} \min(\varepsilon, 4\varepsilon^2)$, by an easy extension of Duchi & Rogers (2019, Lemma 12). Also, $I(X_{\text{pub}}; Z|V) \leqslant H(X_{\text{pub}}|V) \leqslant \log(|\mathcal{X}^{n_{\text{pub}}}|) = dn_{\text{pub}}$, where $H(\cdot|\cdot)$ denotes conditional entropy. Thus,

$$\inf_{\mathcal{A} \in \mathbb{A}_\varepsilon^{\text{loc}}} \sup_{P \in \mathcal{P}_r} \mathbb{E}\|\mathcal{A}(X) - \theta(P)\|_1 \geqslant \frac{r\gamma d}{2}\left[1 - \sqrt{\frac{4000}{d}\gamma^2\left(n_{\text{priv}} \min(\varepsilon, \varepsilon^2) + dn_{\text{pub}}\right)}\right].$$

Choosing $\gamma^2 = c \min\left(\frac{1}{n_{\text{pub}}}, \frac{d}{n_{\text{priv}} \min(\varepsilon, \varepsilon^2)}\right)$ for some small constant $c > 0$ yields

$$\inf_{\mathcal{A} \in \mathbb{A}_\varepsilon^{\text{loc}}} \sup_{P \in \mathcal{P}_r} \mathbb{E}\|\mathcal{A}(X) - \theta(P)\|_1 \gtrsim rd \min\left(\frac{1}{\sqrt{n_{\text{pub}}}}, \sqrt{\frac{d}{n_{\text{priv}} \min(\varepsilon, \varepsilon^2)}}\right),$$

whence

$$\inf_{\mathcal{A} \in \mathbb{A}_\varepsilon^{\text{loc}}} \sup_{P \in \mathcal{P}_r} \mathbb{E}\|\mathcal{A}(X) - \theta(P)\|_2 \gtrsim r\sqrt{d} \min\left(\frac{1}{\sqrt{n_{\text{pub}}}}, \sqrt{\frac{d}{n_{\text{priv}} \min(\varepsilon, \varepsilon^2)}}\right).$$

By applying the non-private mean estimation lower bound, we get

$$\inf_{\mathcal{A}\in\mathbb{A}_\varepsilon^{\text{loc}}}\sup_{P\in\mathcal{P}_r}\mathbb{E}\|\mathcal{A}(X)-\theta(P)\|_2 \gtrsim r\sqrt{d}\min\left(\frac{1}{\sqrt{n_{\text{pub}}}},\sqrt{\frac{d}{n_{\text{priv}}\min(\varepsilon,\varepsilon^2)}}+\frac{1}{\sqrt{n}}\right).$$

Choosing $r=1/\sqrt{d}$ ensures that $\mathcal{P}_r\subset\mathcal{P}(\mathbb{B})$ and yields

$$\inf_{\mathcal{A}\in\mathbb{A}_\varepsilon^{\text{loc}}}\sup_{P\in\mathcal{P}(\mathbb{B})}\mathbb{E}\|\mathcal{A}(X)-\theta(P)\|_2 \gtrsim \min\left(\frac{1}{\sqrt{n_{\text{pub}}}},\sqrt{\frac{d}{n\min(\varepsilon,\varepsilon^2)}}+\frac{1}{\sqrt{n}}\right).$$

Applying Jensen's inequality completes the proof of the lower bound in Theorem 14. Since we assumed $\varepsilon\leqslant 1\leqslant d$, the minimum in the denominator simplifies to ε^2 and the $1/\sqrt{n}$ term is non-dominant.

Upper bound: The first term in the minimum can be realized by the algorithm that throws away the private data and returns $\mathcal{A}(X)=\frac{1}{n_{\text{pub}}}\sum_{x\in X_{\text{pub}}}x$, which is 0-semi-LDP. Also,

$$\mathbb{E}\|\mathcal{A}(X)-\mathbb{E}x\|^2=\frac{1}{n_{\text{pub}}^2}\sum_{x\in X_{\text{pub}}}\mathbb{E}\|x-\mathbb{E}x\|^2\leqslant\frac{1}{n_{\text{pub}}}.$$

The second term in the upper bound can be realized by the ε-LDP (hence ε-semi-LDP) estimator of Duchi et al. (2013), which has worst-case MSE upper bounded by $O\left(\frac{d}{n\varepsilon^2}\right)$. □

Algorithm 3 PrivUnit(p,γ) (Bhowmick et al., 2018)

Input: $v\in\mathbb{S}^{d-1}$, $\gamma\in[0,1]$, $p\in[0,1]$. $B(\cdot;\cdot,\cdot)$ below is the incomplete Beta function $B(x;a,b)=\int_0^x t^{a-1}(1-t)^{b-1}\mathrm{d}t$ and $B(a,b)=B(1;a,b)$.
Draw $z\sim$ Bernoulli(p)
if $z=1$ **then**
  Draw $V\sim$ Unif$\{u\in\mathbb{S}^{d-1}:\langle u,v\rangle\geqslant\gamma\}$
else
  Draw $V\sim$ Unif$\{u\in\mathbb{S}^{d-1}:\langle u,v\rangle<\gamma\}$
end if
Set $\alpha=\frac{d-1}{2}$ and $\tau=\frac{1+\gamma}{2}$
Calculate normalization constant

$$m=\frac{(1-\gamma^2)^\alpha}{2^{d-2}(d-1)}\left(\frac{p}{B(\alpha,\alpha)-B(\tau;\alpha,\alpha)}+\frac{1-p}{B(\tau;\alpha,\alpha)}\right)$$

Return $\frac{1}{m}\cdot V$

F.1.1. An "Even More Optimal" Semi-LDP Estimator

Lemma 46 (Re-statement of Lemma 16)**.** *Let P be a distribution on $\mathbb{B}$ with $V^2=\mathbb{E}\|x-\mathbb{E}_{x\sim P}[x]\|$. Let $c>0$ such that $\mathbb{E}_{x\sim P}\|\mathcal{M}_{Duchi}(x)-\mathbb{E}_{x\sim P}[x]\|^2=\frac{cd}{n\varepsilon^2}$, so that $\mathbb{E}_{X\sim P^n}\|\widetilde{\mathcal{M}}_{Duchi}(X)-\mathbb{E}_{x\sim P}[x]\|^2=\frac{cd}{n\varepsilon^2}+\frac{V^2}{n}$. Then,*

$$\mathbb{E}_{X\sim P^n}\left[\|\mathcal{A}_{Semi\text{-}Duchi}(X)-\mathbb{E}_{x\sim P}[x]\|^2\right]=\frac{n_{priv}}{n}\cdot\frac{cd}{n\varepsilon^2}+\frac{n_{pub}}{n}\cdot\frac{V^2}{n}.$$

Proof. We have

$$\begin{aligned}\mathbb{E}\|\widetilde{\mathcal{A}}_{\text{semi-Duchi}}(X)-\mathbb{E}_{x\sim P}[x]\|^2&=\frac{1}{n^2}\left[\sum_{x\in X_{\text{priv}}}\mathbb{E}\|\mathcal{M}_{\text{Duchi}}(x)-\mathbb{E}_{x\sim P}[x]\|^2+\sum_{x\in X_{\text{pub}}}\mathbb{E}\|x-\mathbb{E}_{x\sim P}[x]\|^2\right]\\&=\frac{ncd}{\varepsilon^2n^2}+\frac{n_{\text{pub}}V^2}{n^2},\end{aligned}$$

by independence of the data and the assumptions in the statement of the lemma. □

F.1.2. A Semi-LDP Estimator with Optimal Constants

Proposition 47 (Re-statement of Proposition 17)**.** *Let $\mathcal{A}(X) = \frac{1}{n}\left[\mathcal{M}_{priv}(\mathcal{R}(x_1), \cdots, \mathcal{R}(x_{n_{priv}})) + \mathcal{M}_{pub}(X_{pub})\right]$ be a ε-semi-LDP algorithm, where $\mathcal{R} : \mathbb{S}^{d-1} \to \mathcal{Z}$ is an ε-LDP randomizer and $\mathcal{M}_{priv} : \mathcal{Z}^{n_{priv}} \to \mathbb{R}^d$ and $\mathcal{M}_{pub} : \mathcal{Z}^{n_{pub}} \to \mathbb{R}^d$ are aggregation protocols such that $\mathbb{E}_{\mathcal{M}_{priv},\mathcal{R}}\left[\mathcal{M}_{priv}(\mathcal{R}(x_1), \cdots, \mathcal{R}(x_{n_{priv}}))\right] = \sum_{x\in X_{priv}} x$ and $\mathbb{E}_{\mathcal{M}_{pub}}\left[\mathcal{M}_{pub}(X_{pub})\right] = \sum_{x\in X_{pub}} x$ for all $X = (X_{priv}, X_{pub}) \in \left(\mathbb{S}^{d-1}\right)^n$. Then,*

$$\sup_{X\in(\mathbb{S}^{d-1})^n} \mathbb{E}_{\mathcal{A}_{\textit{semi-PrivU}}}\|\mathcal{A}_{\textit{semi-PrivU}}(X) - \bar{X}\|^2 \leqslant \sup_{X\in(\mathbb{S}^{d-1})^n} \mathbb{E}_{\mathcal{A}}\|\mathcal{A}(X) - \bar{X}\|^2.$$

Proof. First, Asi et al. (Asi et al., 2022, Proposition 3.4) showed that PrivUnit (with a proper choice of (p, γ)) has the smallest worst-case variance among all unbiased ε-LDP randomizers:

$$\sup_{x\in\mathbb{S}^{d-1}} \mathbb{E}\|\mathcal{R}(x) - x\|^2 \geqslant \sup_{x\in\mathbb{S}^{d-1}} \mathbb{E}\|\text{PrivUnit}(x) - x\|^2 \tag{43}$$

for all ε-LDP randomizers $\mathcal{R}$ such that $\mathbb{E}[\mathcal{R}(x)] = x$ for all $x \in \mathbb{S}^{d-1}$.

Now, let $\mathcal{R}$ be a ε-LDP randomizer and $\mathcal{M}_{\text{priv}}$ and $\mathcal{M}_{\text{pub}}$ be aggregation protocols such that the assumptions in Proposition 17 are satisfied. We claim that there exists an unbiased ε-LDP randomizer $\mathcal{R}' : \mathbb{S}^{d-1} \to \mathcal{Z}$ such that

$$\sup_{X\in(\mathbb{S}^{d-1})^n} \mathbb{E}_{\mathcal{A}}\left\|n\mathcal{A}(X) - \sum_{x\in X} x\right\|^2 \geqslant \sup_{X\in(\mathbb{S}^{d-1})^n} \mathbb{E}_{\mathcal{R}'}\left\|\sum_{x\in X_{\text{priv}}} \mathcal{R}'(x) - \sum_{x\in X_{\text{priv}}} x\right\|^2. \tag{44}$$

To prove (44), we follow the idea in the proof of Asi et al. (2022, Proposition 3.3). Let P denote the uniform distribution on $\mathbb{S}^{d-1}$. We have

$$\begin{aligned}
&\sup_{X\in(\mathbb{S}^{d-1})^n} \mathbb{E}_{\mathcal{A}}\left\|n\mathcal{A}(X) - \sum_{x\in X} x\right\|^2 \geqslant \mathbb{E}_{X\sim P^n,\mathcal{A}}\left\|n\mathcal{A}(X) - \sum_{x\in X} x\right\|^2 \\
\geqslant &\mathbb{E}\left\|\mathcal{M}_{\text{priv}}(\mathcal{R}(x_1), \cdots, \mathcal{R}(x_{n_{\text{priv}}})) - \sum_{x\in X_{\text{priv}}} x\right\|^2 + \mathbb{E}\left\|\mathcal{M}_{\text{pub}}(X_{\text{pub}}) - \sum_{x\in X_{\text{pub}}} x\right\|^2,
\end{aligned}$$

since the cross-term (inner product) vanishes by independence of X_{pub} and X_{priv}, and unbiasedness of $\mathcal{M}_{\text{pub}}$. Now, (Asi et al., 2022, Lemma A.1) shows that there exist ε-LDP randomizers $\{\hat{\mathcal{R}}_x\}_{x\in X_{\text{priv}}}$ such that $\mathbb{E}[\hat{\mathcal{R}}_x(v)] = v$ for all $v \in \mathbb{S}^{d-1}$ and

$$\mathbb{E}_{X_{\text{priv}}\sim P^{n_{\text{priv}}}}\left\|\mathcal{M}_{\text{priv}}(\mathcal{R}(x_1), \cdots, \mathcal{R}(x_{n_{\text{priv}}})) - \sum_{x\in X_{\text{priv}}} x\right\|^2 \geqslant \sum_{x\in X_{\text{priv}}} \mathbb{E}_{v\sim P}\|\hat{\mathcal{R}}_x(v) - v\|^2.$$

Hence

$$\sup_{X\in(\mathbb{S}^{d-1})^n} \mathbb{E}_{\mathcal{A}}\left\|n\mathcal{A}(X) - \sum_{x\in X} x\right\|^2 \geqslant \sum_{x\in X_{\text{priv}}} \mathbb{E}_{v\sim P}\|\hat{\mathcal{R}}_x(v) - v\|^2.$$

Define $\mathcal{R}'_x(v) := U^T\hat{\mathcal{R}}_x(Uv)$ for $v \in \mathbb{S}^{d-1}$, where U is a uniformly random rotation matrix such that $U^TU = \mathbf{I}_d$. Note that $\mathcal{R}'_x$ is an ε-LDP randomizer such that $\mathbb{E}[\mathcal{R}'(v)] = v$ for all $v \in \mathbb{S}^{d-1}, x \in X_{\text{priv}}$. Moreover, for any fixed $v \in \mathbb{S}^{d-1}, x \in X_{\text{priv}}$, we have

$$\begin{aligned}
\mathbb{E}\|\mathcal{R}'_x(v) - v\|^2 &= \mathbb{E}_U\|\hat{\mathcal{R}}_x(Uv) - Uv\|^2 \\
&= \mathbb{E}_{v'\sim P}\|\hat{\mathcal{R}}_x(v') - v'\|^2
\end{aligned}$$

Let $x^* := \operatorname{argmin}_{x\in X_{\text{priv}}} \sup_{v\in\mathbb{S}^{d-1}} \mathbb{E}\|\mathcal{R}'_x(v) - v\|^2$ and $\mathcal{R}'(v) := \mathcal{R}'_{x^*}(v)$. Then putting the pieces together, we have

$$\begin{aligned}\sup_{X\in(\mathbb{S}^{d-1})^n} \mathbb{E}_{\mathcal{A}} \left\| n\mathcal{A}(X) - \sum_{x\in X} x \right\|^2 &\geqslant \sum_{x\in X_{\text{priv}}} \sup_{v\in\mathbb{S}^{d-1}} \mathbb{E}\|\mathcal{R}'_x(v) - v\|^2 \\ &\geqslant n_{\text{priv}} \sup_{v\in\mathbb{S}^{d-1}} \mathbb{E}\|\mathcal{R}'(v) - v\|^2 \\ &= \sup_{X_{\text{priv}}\in(\mathbb{S}^{d-1})^{n_{\text{priv}}}} \mathbb{E}_{\mathcal{R}'} \left\| \sum_{x\in X_{\text{priv}}} \mathcal{R}'(x) - \sum_{x\in X_{\text{priv}}} x \right\|^2,\end{aligned}$$

by conditional independence of $\{\mathcal{R}'(x)\}_{x\in X_{\text{priv}}}$ given X. This establishes (44). Thus,

$$\begin{aligned}n^2 \sup_{X\in(\mathbb{S}^{d-1})^n} \mathbb{E}_{\mathcal{A}} \left\|\mathcal{A}(X) - \bar{X}\right\|^2 &= \sup_{X\in(\mathbb{S}^{d-1})^n} \mathbb{E}_{\mathcal{A}} \left\| n\mathcal{A}(X) - \sum_{x\in X} x \right\|^2 \\ &\geqslant \sup_{X\in(\mathbb{S}^{d-1})^n} \mathbb{E}_{\mathcal{R}'} \left\| \sum_{x\in X_{\text{priv}}} \mathcal{R}'(x) - \sum_{x\in X_{\text{priv}}} x \right\|^2 \\ &\geqslant \sup_{X\in(\mathbb{S}^{d-1})^n} \mathbb{E} \left\| \sum_{x\in X_{\text{priv}}} \text{PrivUnit}(x) - \sum_{x\in X_{\text{priv}}} x \right\|^2 \\ &= n^2 \sup_{X\in(\mathbb{S}^{d-1})^n} \mathbb{E}_{\mathcal{A}_{\text{semi-PrivU}}} \|\mathcal{A}_{\text{semi-PrivU}}(X) - \bar{X}\|^2,\end{aligned}$$

where we used (43) in the last inequality. Dividing both sides of the above inequality by n^2 completes the proof. □

F.2. Optimal Semi-LDP Stochastic Convex Optimization

If $\mu = 0$ (convex case), we denote $\mathcal{R}^{\text{loc}}_{\text{SCO}}(\varepsilon, n_{\text{priv}}, n, d, L, D) := \mathcal{R}^{\text{loc}}_{\text{SCO}}(\varepsilon, n_{\text{priv}}, n, d, L, D, \mu = 0)$.

Theorem 48 (Complete statement of Theorem 18)**.** *Let $\varepsilon \in (0, 1]$. There exist absolute constants c and C, with $0 < c \leqslant C$, such that*

$$cLD \min\left\{\frac{1}{\sqrt{n_{pub}}}, \sqrt{\frac{d}{n\varepsilon^2}}\right\} \leqslant \mathcal{R}^{loc}_{SCO}(\varepsilon, n_{priv}, n, d, L, D) \leqslant CLD \min\left\{\frac{1}{\sqrt{n_{pub}}}, \sqrt{\frac{d}{n\varepsilon^2}}\right\},$$

and

$$cLD \min\left\{\frac{1}{n_{pub}}, \frac{d}{n\varepsilon^2}\right\} \leqslant \mathcal{R}^{loc}_{SCO}(\varepsilon, n_{priv}, n, d, L, D, \mu) \leqslant C\frac{L^2}{\mu} \min\left\{\frac{1}{n_{pub}}, \frac{d}{n\varepsilon^2}\right\}.$$

Proof. **Lower bounds:** Let $\mathcal{A}$ be ε-semi-LDP and denote $w_{\text{priv}} = \mathcal{A}(X)$.

Strongly convex lower bound: We begin with the strongly convex lower bounds, which can be proved straightforwardly by reducing strongly convex SCO to mean estimation and applying Theorem 14. In a bit more detail, let $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ be given by

$$f(w, x) = \frac{L}{2D}\|w - x\|^2,$$

where $\mathcal{W} = \mathcal{X} = D\mathbb{B}$. Note that f is L-uniformly Lipschitz and $\frac{L}{D}$-strongly convex in w for all x. Further, $w^* := \operatorname{argmin}_{w\in\mathcal{W}} \{F(w) = \mathbb{E}_{x\sim P}[f(w, x)]\} = \mathbb{E}_{x\sim P}[x]$. By a direct calculation (see e.g. (Kamath et al., 2022a, Lemma 6.2)), we have

$$\mathbb{E}F(w_{\text{priv}}) - F(w^*) = \frac{L}{2D}\mathbb{E}\|w_{\text{priv}} - w^*\|^2. \tag{45}$$

We can lower bound $\mathbb{E}\|w_{\text{priv}} - w^*\|^2 = \mathbb{E}\|\mathcal{A}(X) - \mathbb{E}_{x\sim P}[x]\|^2$ via Theorem 14 (and its proof, to account for the re-scaling). Specifically, there is a distribution P on $\mathcal{X}$ such that

$$\mathbb{E}_{X\sim P^n, w_{\text{priv}}}\|w_{\text{priv}} - w^*\|^2 \geqslant cD^2 \min\left(\frac{1}{n_{\text{pub}}}, \frac{d}{n\varepsilon^2}\right).$$

Combining this with (45) leads to the desired excess risk lower bound.

Convex lower bound: We will begin by proving the lower bounds for the case in which $L = D = 1$, and then scale our construction to get the lower bounds for arbitrary L, D.

Let $\mathcal{W} = \mathbb{B}$, $\mathcal{X} = \left\{\pm\frac{1}{\sqrt{d}}\right\}^d$, and

$$f(w, x) = -\langle w, x\rangle,$$

which is convex and 1-uniformly-Lipschitz in w on $\mathcal{X}$. For any ε-semi-LDP $\mathcal{A}'(X) = w'_{\text{priv}}$ and any $\gamma \in (0, 1)$, the proof of Theorem 14 constructs a distribution P_γ on $\mathcal{X}$ with mean $\mathbb{E}_{x\sim P_\gamma}[x] = \theta \in \{\pm\frac{\gamma}{\sqrt{d}}\}^d$ such that

$$\mathbb{E}_{w'_{\text{priv}}, X\sim P_\gamma^n}\left[\left\|w'_{\text{priv}} - \theta\right\|^2\right] \geqslant \frac{\gamma^2}{4}\left[1 - \sqrt{4000\gamma^2\left(\frac{n_{\text{priv}}\varepsilon^2}{d} + n_{\text{pub}}\right)}\right]. \tag{46}$$

Now, let $F_\gamma(w) = \mathbb{E}_{x\sim P_\gamma}[f(w, x)]$ and $w^*_\gamma = \operatorname{argmin}_{w\in\mathcal{W}} F_\gamma(w) = \frac{\theta}{\|\theta\|}$. A direct calculation (see e.g. (Kamath et al., 2022a, Equation 14)) shows

$$\begin{aligned}
\mathbb{E}F_\gamma(w) - F^*_\gamma &\geqslant \frac{1}{2}\mathbb{E}\left[\|\theta\|\|w - w^*_\gamma\|^2\right] \\
&\geqslant \frac{1}{2\|\theta\|}\mathbb{E}\left[\|w' - \theta\|^2\right] \\
&= \frac{1}{2\gamma}\mathbb{E}\left[\|w' - \theta\|^2\right]
\end{aligned} \tag{47}$$

for any $w \in \mathcal{W}, w' := w\|\theta\|$. Note that $\mathcal{A}(X) = w_{\text{priv}}$ is ε-semi-DP if and only if $\mathcal{A}'(X) = w'_{\text{priv}} := \|\theta\| w_{\text{priv}} = \gamma w_{\text{priv}}$ is ε-semi-DP, by post-processing. Thus, (46) and (47) together imply that any ε-semi-DP w'_{priv} has worst-case excess risk that is lower bounded by

$$\mathbb{E}[F_\gamma(w'_{\text{priv}}) - F^*_\gamma] \geqslant \frac{\gamma}{8}\left[1 - \sqrt{4000\gamma^2\left(\frac{n_{\text{priv}}\varepsilon^2}{d} + n_{\text{pub}}\right)}\right].$$

Choosing $\gamma^2 = c\min\left(\frac{d}{n_{\text{priv}}\varepsilon^2}, \frac{1}{n_{\text{pub}}}\right)$ for some small $c \in (0, 1/16000)$ implies

$$\mathbb{E}[F_\gamma(w'_{\text{priv}}) - F^*_\gamma] \geqslant c'\min\left(\sqrt{\frac{d}{n_{\text{priv}}\varepsilon^2}}, \frac{1}{\sqrt{n_{\text{pub}}}}\right)$$

for some $c' > 0$. This proves the desired lower bound for the case when $L = D = 1$. In the general case, we scale our hard instance, as in the proof of Theorem 4: Let $\widetilde{W} = D\mathcal{W}$, $\widetilde{X} = L\mathcal{X}$, and $\widetilde{x} \sim \widetilde{P} \iff \widetilde{x} = Lx$ for $x \sim P$. Define $\widetilde{f} : \widetilde{\mathcal{W}} \times \widetilde{\mathcal{X}} \to \mathbb{R}$ by $\widetilde{f}(\widetilde{w}, \widetilde{x}) := f(\widetilde{w}, \widetilde{x}) = -\langle\widetilde{w}, \widetilde{x}\rangle$. Then $\widetilde{f}(\cdot, \widetilde{x})$ is L-Lipschitz and convex. Moreover, if $F(w) = \mathbb{E}_{x\sim P}[f(w, x)]$, $\theta := \mathbb{E}_{x\sim P}[x]$, $w^* := \operatorname{argmin}_{w\in\mathcal{W}} F(w) = \frac{\theta}{\|\theta\|}$, $\widetilde{F}(\widetilde{w}) = \mathbb{E}_{\widetilde{x}\sim\widetilde{P}}[f(\widetilde{w}, \widetilde{x})]$, $\widetilde{w} = Dw$, and $\widetilde{\theta} := \mathbb{E}_{\widetilde{x}\sim\widetilde{P}}[\widetilde{x}] = L\theta$, then $\widetilde{w}^* = Dw^* \in \operatorname{argmin}_{\widetilde{w}\in\widetilde{\mathcal{W}}} \widetilde{F}(\widetilde{w})$ and

$$\begin{aligned}
\widetilde{F}(\widetilde{w}) - \widetilde{F}^* &= -\langle\widetilde{w}, \widetilde{\theta}\rangle + \langle\widetilde{w}^*, \widetilde{\theta}\rangle \\
&= D\langle\widetilde{\theta}, w^* - w\rangle \\
&= LD\langle\theta, w^* - w\rangle \\
&= LD\left[F(w) - F^*\right].
\end{aligned}$$

This shows that the excess risk of the scaled instance scales by LD, completing the lower bound proofs.

Upper bounds: The first term in each of the upper bounds ($LD/\sqrt{n_{\text{pub}}}$ for convex, and $L^2/(\mu n_{\text{pub}})$ for strongly convex) is attained by the throw-away algorithm that runs n_{pub} steps of (one-pass) SGD on X_{pub} (Nemirovski & Yudin, 1983).

The second term in the convex upper bound follows from the ε-LDP (hence semi-LDP) upper bound of Duchi et al. (2013, Proposition 3).

For the second term in the strongly convex upper bound, we run ε-LDP-SGD as in (Duchi et al., 2013). We also return a non-uniform weighted average $\hat{w}_n$ of the iterates $w_1, \ldots, w_n$ as in (Rakhlin et al., 2012) to obtain

$$\mathbb{E}F(\hat{w}_n) - F^* \leqslant C\frac{G^2}{\mu n},$$

where $G^2 = \sup_{t\in[n]} \mathbb{E}\|\mathcal{M}_{\text{Duchi}}(\nabla f(w_t, x_t))\|^2 \lesssim L^2\left(1 + \frac{d}{\varepsilon^2}\right)$ (Duchi et al., 2013). Thus,

$$\mathbb{E}F(\hat{w}_n) - F^* \leqslant C'\frac{L^2}{\mu}\left(\frac{d}{n\varepsilon^2}\right),$$

since $d \geqslant 1 \geqslant \varepsilon^2$. This completes the proof. □

F.2.1. An "Even More Optimal" Semi-LDP Algorithm for SCO

Algorithm 4 Semi-LDP-SGD

Input: clip threshold $C > 0$, stepsize η, privacy parameter $\varepsilon \in (0, 1]$.
Initialize $w_0 \in \mathcal{W}$.
for $t \in \{0, 1, \cdots, n-1\}$ **do**
  Draw random sample x_t from X without replacement.
  if $x_t \in X_{\text{priv}}$ **then**
    $\widetilde{g}_t \leftarrow \mathcal{M}_{\text{Duchi}}(\text{clip}_C(\nabla f(w_t, x_t)))$.
  else
    $\widetilde{g}_t \leftarrow \nabla f(w_t, x_t)$
  end if
  Update $w_{t+1} := \Pi_{\mathcal{W}}[w_t - \eta\widetilde{g}_t]$.
end for
Output: $\bar{w}_n = \frac{1}{n}\sum_{i=1}^n w_i$.

Proposition 49 (Re-statement of Proposition 19)**.** *Let $f \in \mathcal{F}_{\mu=0,L,D}$ and P be any distribution and $\varepsilon \leqslant d$. Algorithm 4 is ε-semi-LDP. Further, there is an absolute constant c such that the output $\mathcal{A}(X) = \bar{w}_n$ of Algorithm 4 satisfies*

$$\mathbb{E}_{\mathcal{A}, X\sim P^n} F(\bar{w}_n) - F^* \leqslant c\frac{LD}{\sqrt{n}}\max\left\{\sqrt{\frac{d}{\varepsilon^2}}\sqrt{\frac{n_{priv}}{n}}, \sqrt{\frac{n_{pub}}{n}}\right\}.$$

Proof. **Privacy:** Since $\mathcal{M}_{\text{Duchi}}$ is an ε-LDP randomizer and we are applying $\mathcal{M}_{\text{Duchi}}$ to the gradients of all the private samples $x \in X_{\text{priv}}$, Algorithm 4 is ε-semi-LDP.

Excess risk: Choose $C > L$: i.e. we don't clip, since stochastic subgradients are already uniformly bounded by the L-Lipschitz assumption. By the classical analysis of the stochastic subgradient method (see e.g. (Bubeck et al., 2015)), we can obtain

$$\begin{aligned}\mathbb{E}F(\bar{w}_n) - F^* &\leqslant \frac{1}{n}\sum_{t=1}^n \mathbb{E}[\widetilde{g}_t^T(w_t - w^*)] \\ &\leqslant \frac{D^2}{\eta n} + \frac{\eta}{n}\left(n_{\text{priv}}G_a^2 + n_{\text{pub}}G_b^2\right),\end{aligned}$$

where $G_a^2 := \sup_t \mathbb{E}\|\mathcal{M}_{\text{Duchi}}(\nabla f(w_t, x_t))\|^2$ and $G_b^2 := \sup_t \mathbb{E}\|\nabla f(w_t, x_t)\|^2$. By the uniform Lipschitz assumption, we have $G_b^2 \leqslant L^2$. By (Duchi et al., 2013), we have $G_a^2 \leqslant c^2L^2\frac{d}{\varepsilon^2}$ for some absolute constant $c > 0$. Thus, choosing $\eta = \frac{D}{L}\min\left(\sqrt{\frac{\varepsilon^2}{n_{\text{priv}}c^2d}}, \frac{1}{\sqrt{n_{\text{pub}}}}\right)$ yields

$$\mathbb{E}F(\bar{w}_n) - F^* \leqslant 3\frac{LD}{\sqrt{n}}\max\left(\sqrt{\frac{c^2d}{\varepsilon^2}}\sqrt{\frac{n_{\text{priv}}}{n}}, \sqrt{\frac{n_{\text{pub}}}{n}}\right).$$

This completes the proof. □

G. Numerical Experiments

Code for all of the experiments is available here: https://github.com/optimization-for-data-driven-science/DP-with-public-data.

G.1. Central Semi-DP Experiments

G.1.1. Semi-DP Linear Regression with Gaussian Data

Data Generation: We implement Algorithm 1 on synthetic data designed for a linear regression problem of dimension $2,000$, using the squared error loss: $f(w,x) = (\langle w, x^{(1)}\rangle - x^{(2)})^2$, where $x^{(1)} \in \mathbb{R}^d$ denotes the feature vector and $x^{(2)} \in \mathbb{R}$ is the target. Here, $d = 2,000$. Our synthetic dataset consists of $n = 30,000$ training samples, 7500 validation samples, and $37,500$ test samples. The feature vectors $x_i := x_i^{(1)}$ and the optimal parameter vector w^* are drawn i.i.d. from a multivariate Gaussian distribution $\mathcal{N}(0, \mathbf{I}_{2000})$. We generate predicted values $x_i^{(2)}$ from a Gaussian distribution $\mathcal{N}(\langle w^*, x_i\rangle, 1)$. Thus, the optimal linear regression model has which ensures an optimal mean squared error of 1.

Experimental Setup: Our experiments investigate two phenomena: 1) the effect of the ratio $\frac{n_{\text{pub}}}{n}$ on test loss when $\varepsilon \in \{2, 4\}$ is fixed, for values of $\frac{n_{\text{pub}}}{n}$ ranging from 0.01 to 0.95, see Figures 3-4; and 2) the effect of privacy (quantified by ε) on test loss, for fixed $\frac{n_{\text{pub}}}{n} \in \{0.1, 0.25\}$, and varying $\varepsilon \in \{0.1, 0.5, 1.0, 2.0, 4.0, 8.0\}$, see Figures 9-10 . We maintain the privacy parameter δ at a constant value of 10^{-5} throughout our experiments. We set the private batch size $K_{\text{priv}} = 500$, public batch size $K_{\text{pub}} = 200$, and iterations $T = 5000$. All algorithms undergo extensive hyperparameter tuning using the validation dataset, and the performance of each tuned algorithm is subsequently assessed using the test dataset. (See "**Hyperparameter Tuning**" paragraph below for details on the tuning process.)

Details on Implementations of Algorithms: We compare four different semi-DP algorithms: 1. *Throw-away*. 2. *DP-SGD* (Abadi et al., 2016; De et al., 2022). 3. *PDA-MD* (Amid et al., 2022, Algorithm 1). 4. *Our Algorithm 1*—specifically, the *sample-with-replacement* version of our algorithm. If $n_{\text{pub}} \geqslant d$, the *Throw-away* algorithm simply returns a minimizer w_{pub} of the public loss: $w_{\text{pub}} = (X_{\text{pub}}^T X_{\text{pub}})^{-1} X_{\text{pub}}^T y_{\text{pub}}$. Otherwise, we used pretrained warm-start models with all public samples. (See "**warm-start**" paragraph below for details.) *DP-SGD* adds noise to all (public or private) gradients. We use the state-of-the-art (for image classification) implementation of DP-SGD of (De et al., 2022). We adopt the re-parameterization of DP-SGD in (De et al., 2022, Equation 3) to ease the hyperparameter tuning. For *PDA-MD*, we implement (Amid et al., 2022, Algorithm). We use their exact Mirror descent form by multiplying the inverse of the hessian $X_{\text{pub}}^T X_{\text{pub}}$ by the private gradient (Amid et al., 2022). In their original implementation, they added a small constant, Hessian Regularization, times the identity matrix to the Hessian before calculating the inverse for numerically stable. We choose the hessian regularization constant as 0.01, the same value in their original implementation.

Effect of increasing ratio n_{pub}/n_{priv}: More public data always improves the performance of Algorithm 1, but does not always benefit PDA-MD. The primary reason for this is that our algorithm more effectively handles the increasing privacy noise that is needed to maintain semi-DP with increasing ratio n_{pub}/n. Our algorithm achieves this by using the weight parameter to reduce the variance of the increasingly noisy private gradients and leverage public gradients. By contrast, as n_{pub}/n ratio rises, the efficacy of PDA-MD may diminish due to its over-reliance on increasingly noisy private gradients. We verify our reasoning numerically in Appendix G.3: Tables 11 and 12 record the standard deviation σ of the privacy noise for different ratios.

Effect of hessian regularization parameter on PDA-MD: Upon reproducing the results of PDA-MD, we discovered a high sensitivity to the choices of Hessian Regularization, especially when the ratio of the largest and the smallest eigenvalue of the data matrix is large. We test PDA-MD on the dataset proposed in the original study as well as on our own dataset. The results of these tests are displayed in Fig. 8. We see PDA-MD is sensitive to hessian regularization parameter, which requires extra tuning on complicated tasks. We implement PDA-MD with the optimal regularization value of 0.01.

Details on warm-start and cold-start: In our evaluation, we adopt a "warm start" strategy for all algorithms: we first find a minimizer w_{pub} of the public loss, and then initialize the training process with w_{pub}. Note that there are two cases: $n_{\text{pub}} \geqslant n$ and $n_{\text{pub}} < n$. In the case of $n_{\text{pub}} \geqslant n$, the minimizer w_{pub} of the public loss can be obtained via $w_{\text{pub}} = \left(X_{\text{pub}}^T X_{\text{pub}}\right)^{-1} X_{\text{pub}}^T y_{\text{pub}}$. In the case of $n_{\text{pub}} < n$, we run SGD on linear regression with all n_{pub}. Specifically, we minimize w_{pub} of the public loss by running the SGD with public batch size $K_{\text{pub}} = 200$, stepsize $\eta_t = 0.5$, and iterations $T = 50000$ to allow the models fully converge. For cold start scenarios, we initialize all model parameters w to 0. The cold

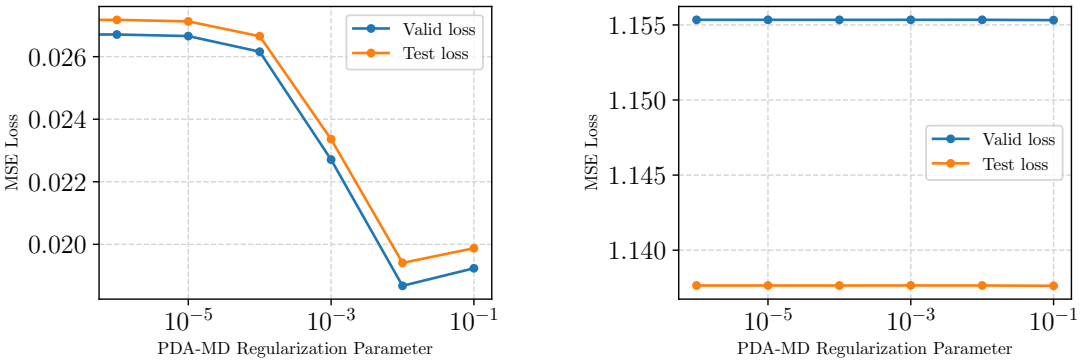

Figure 8. Loss v.s. PDA-MD Regularization Parameter. Left: Results on proposed dataset in (Amid et al., 2022). Right: Results on our dataset. PDA-MD is sensitive to hessian regularization parameter, which requires extra tuning on complicated tasks.

start experiment results can be seen in Figure 5 and Figure 11.

Clipping public gradients improves performance: We have empirically found that a slight modification of Algorithm 1 offers performance benefits. Namely, it is beneficial to project the public gradients onto the ℓ_2-sphere of radius C; that is, we re-scale the public gradients to have ℓ_2-norm equal to C. Note that semi-DP still holds regardless of whether or not the public gradients are re-scaled. Re-scaling the public gradients helps balance the effects of the public and private gradients on the optimization trajectory. In the original method stated in Algorithm 1, if unclipped public gradients and the private gradient are of very different magnitudes, then one gradient direction might dominate the optimization procedure, leading to a sub-optimal model. Thus, our public gradient re-scaling technique promotes a more balanced update, which gracefully combines the public and private data in each iteration. To the best of our knowledge, this technique is novel.

Privacy accounting: We compute the privacy loss of each algorithm by using the moments acccountant of Abadi et al. (Abadi et al., 2016). For a fixed clip threshold C, privacy level (ε, δ) is determined by three parameters: the variance of privacy noise σ^2, the private sampling ratio $q := K_{\text{priv}}/n_{\text{priv}}$, and the total number of iterations T. In our setting, the privacy parameters (ε, δ) are given, and we use the moments accountant to compute an approximation of σ^2 for any choice of hyperparameters T and q. We utilize the implementation of the privacy accountant provided by the Pytorch privacy framework, Opacus.

Hyperparameter Tuning: The results reported are for each algorithm with the hyperparameters (step size and α in Semi-DP) that attain the best performance for a given experiment. For simplicity and computation efficiency, we keep clip threshold $C = 1$ for all of our experiments. Preliminary experiments found that a clip threshold of $C = 1$ worked well for all algorithms. To tune all algorithms, we use grid search. See Tables 1 and 2 in Appendix G.2 for detailed descriptions of the hyperparameter search grids.

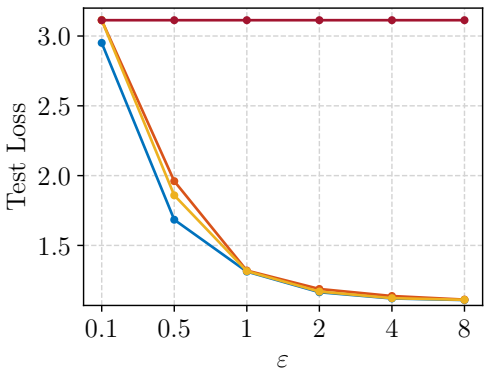

Figure 9. Test loss vs. ε. $\frac{n_{\text{pub}}}{n} = 0.1$.

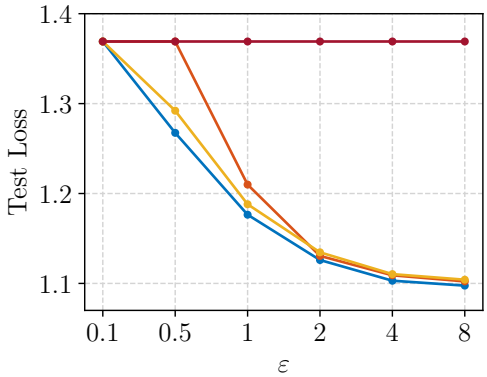

Figure 10. Test loss vs. ε. $\frac{n_{\text{pub}}}{n} = 0.25$.

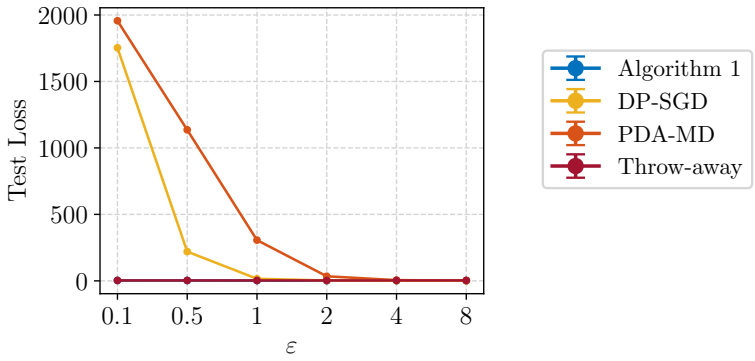

Figure 11. Test loss vs. ε. $\frac{n_{\text{pub}}}{n} = 0.1$, without warm-start.

G.1.2. Semi-DP Mean Estimation

We present an experiment with mean estimation: We fix ρ-semi-zCDP with privacy parameter $\rho = 0.5$. We draw n i.i.d. samples from a $(d = 1000)$-dimensional **Bernoulli**$(p = 1/2)$ product distribution. We investigated the effect of the ratio $\frac{n_{\text{pub}}}{n}$ on mean ℓ_2 error, for values of $\frac{n_{\text{pub}}}{n}$ ranging from 0.05 to 0.95. We presented the experiment in three different high-dimensional and low-dimensional settings: $n < d$, $n = d$, and $n > d$. Fig. 12 shows that when $n < d$, throw-away outperforms Gaussian mechanism. Fig. 13 and Fig. 14 show that when $d \leqslant n$, throw-away outperforms Gaussian mechanism except when $\frac{n_{\text{pub}}}{n} = 0.05$. Moreover, **our Weighted Gaussian estimator outperforms both throw-away and the Gaussian mechanism in every case**.

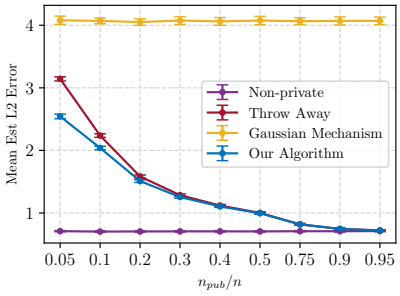

Figure 12. $d = 1000$, $n = 500$

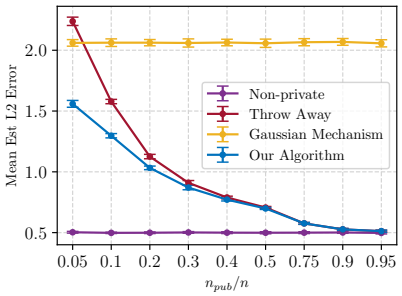

Figure 13. $d = 1000$, $n = 1000$

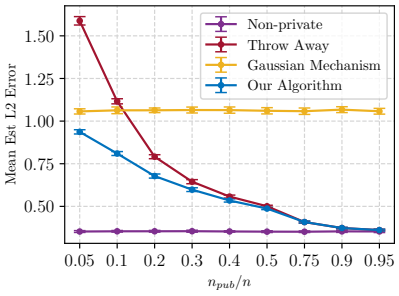

Figure 14. $d = 1000$, $n = 2000$

G.1.3. Semi-DP Logistic Regression with CIFAR-10

We evaluate the performance of Algorithm 1 in training a logistic regression model to classify digits in the CIFAR-10 dataset (Krizhevsky et al., 2009). We compare Algorithm 1 against DP-SGD and throw-away. Note that PDA-MD does not have an efficient implementation for logistic regression. This is because there does not exist a closed form update rule for the mirror descent step. Thus, we do not compare against PDA-MD.

We flatten the images and feed them to the logistic (softmax) model. Cross-entropy loss is used here; therefore, the model is convex. Implementations of the algorithms are similar to the linear regression case. However, in this case, throw-away consists of running non-private SGD on X_{pub} to find an approximate minimizer w_{pub}. For all three algorithms, we fixed batch-size 256 and privacy parameter $\delta = 10^{-6}$. The remaining hyperparameters are tuned by grid search, using the same grid for each algorithm. Also, in contrast to the linear regression experiments, *we do not use warm-start* for any of the algorithms and *we use SGD* for all algorithms in these experiments. See Table 3 in Appendix G.2 for detailed descriptions of the hyperparameter search grids.

Results are reported in Figs. 15 to 18. *Our Algorithm 1 always outperforms the baselines* in terms of minimizing test accuracy/error. The advantage of Algorithm 1 over these baselines is even more pronounced than it was for linear regression. This might be partially due to the fact that we did not use warm-start.

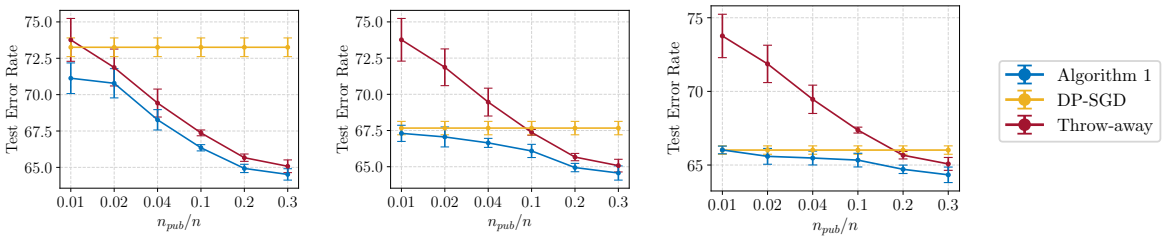

Figure 15. Test error rate vs. $\frac{n_{\text{pub}}}{n}$. Left: $\varepsilon = 0.1$. Middle: $\varepsilon = 0.5$. Right: $\varepsilon = 1.0$

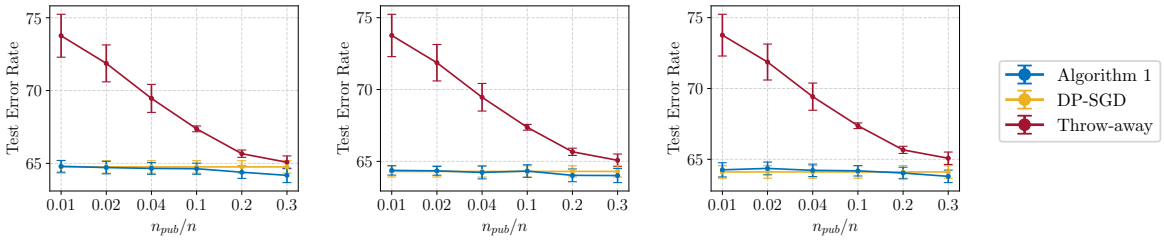

Figure 16. Test error rate vs. $\frac{n_{\text{pub}}}{n}$. Left: $\varepsilon = 2.0$. Middle: $\varepsilon = 4.0$. Right: $\varepsilon = 8.0$

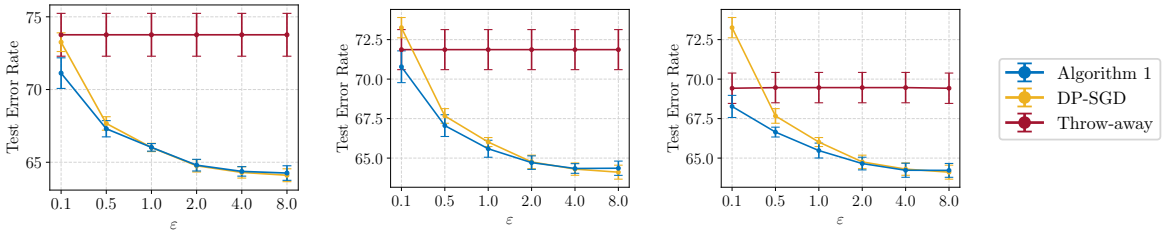

Figure 17. Test error rate vs. ε. Left: $\frac{n_{\text{pub}}}{n} = 0.01$. Middle: $\frac{n_{\text{pub}}}{n} = 0.02$. Right: $\frac{n_{\text{pub}}}{n} = 0.04$

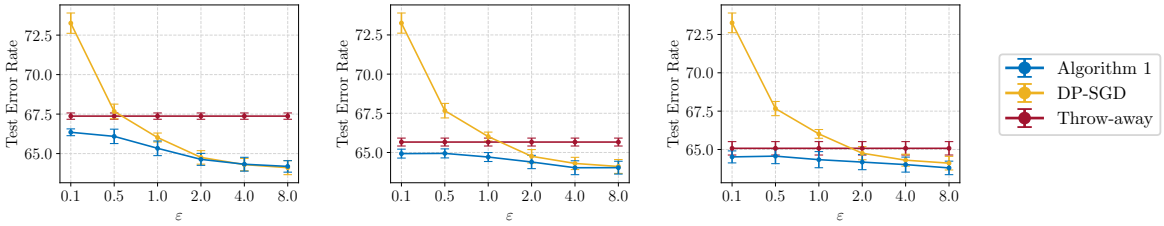

Figure 18. Test error rate vs. ε. Left: $\frac{n_{\text{pub}}}{n} = 0.1$. Middle: $\frac{n_{\text{pub}}}{n} = 0.2$. Right: $\frac{n_{\text{pub}}}{n} = 0.3$

G.1.4. Semi-DP Wide-ResNet-16-4 with CIFAR-10

To evaluate the performance of Algorithm 1 in training non-convex (neural) models, we use the Wide-ResNet with 16 layers and a width factor of 4 (WRN-16-4) (Zagoruyko & Komodakis, 2017). When trained non-privately on CIFAR-10 dataset, this model achieves an error rate of 5.02 (Zagoruyko & Komodakis, 2017).

Experimental Setup: We followed the same experimental setup as (De et al., 2022). For all algorithms, we used a constant learning rate and did not use momentum, weight decay, or dropout. We fixed the batch size of 256 and privacy parameter $\delta = 10^{-5}$. For simplicity and to isolate the effects of the weighted gradient estimator, we used the WRN-16-4

without batch/group normalization or data augmentation.[10] We split the CIFAR-10 dataset (Krizhevsky et al., 2009) of 50,000 training examples into 40,000 training samples and 10,000 validation samples. We used their 10,000 test images as our test set. Same as Appendix G.1.3, we did not use warm-start here.

Hyperparameter Tuning: We selected the hyperparameters settings with the highest validation accuracy for all algorithms and reported their test accuracy on the official test set. To tune the hyperparameters, we used the bayesian hyperparameter optimization technique. That is, we build a probability model of the objective function and use it to select the most promising hyperparameters to evaluate in the true objective function. See Table 4 in Appendix G.2 for detailed descriptions of the hyperparameter search range.

Results: We investigate two cases: 1) the effect of the ratio of public samples $\frac{n_{\text{pub}}}{n}$ on accuracy when $\varepsilon = 8$ is fixed. We test values of $\frac{n_{\text{pub}}}{n}$ ranging from 0.01 to 0.3; and 2) the effect of privacy (quantified by ε) on accuracy, for fixed $\frac{n_{\text{pub}}}{n} = 0.04$. We vary $\varepsilon \in \{0.1, 0.5, 1.0, 2.0, 4.0, 8.0\}$. Results are reported in Figs. 19 and 20. *Our Algorithm 1 always outperforms the baselines* (DP-SGD and Throw-away) in terms of minimizing test error, across all experimental setups.

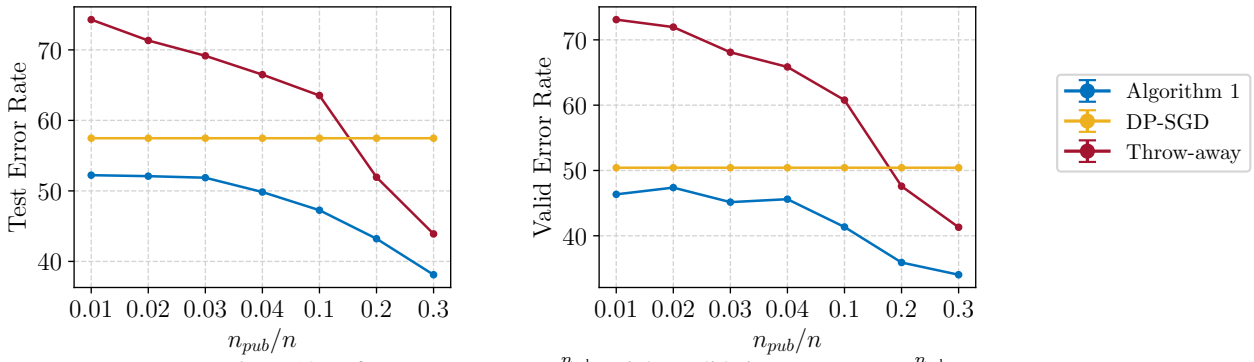

Figure 19. Left: Test error rate vs. $\frac{n_{\text{pub}}}{n}$. Right: Validation error rate vs. $\frac{n_{\text{pub}}}{n}$. $\varepsilon = 8$

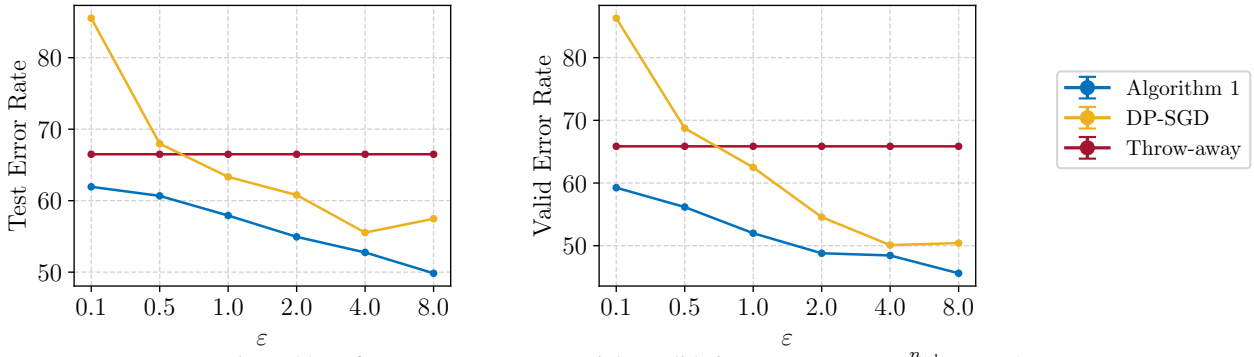

Figure 20. Left: Test error rate vs. ε. Right: Validation error rate vs. ε. $\frac{n_{\text{pub}}}{n} = 0.04$

G.2. Hyperparameters Search Grids

hyperparameter	learning-rate
value	$0, 0.01, 0.03, 0.05, 0.07, 0.09, 0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5, 1.7, 1.9$

Table 1. Grid used for hyperparameter search for PDA-MD and DP-SGD in Appendix G.1.1

[10] We believe that accuracy could be improved by combining these tricks with our semi-DP gradient estimator (De et al., 2022; Nasr et al., 2023).

hyperparameter	learning-rate	α
value	$0, 0.01, 0.03, 0.05, 0.07, 0.09, 0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5, 1.7, 1.9$	$0, 0.1, 0.2 \ldots 1$

Table 2. Grid used for hyperparameter search for Algorithm 1 in Appendix G.1.1

hyperparameter	value
iterations	$2000, 4000, 6000, \ldots, 20000$
learning-rate	$0.005, 0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 5$
α	$0, 0.1, 0.2, \ldots, 1$

Table 3. Grid used for hyperparameter search for DP-SGD and Semi-DP in Appendix G.1.3

hyperparameter	iterations	learning-rate	α
value	$[3000, 7000]$	$[0.1, 3]$	$[0, 1]$

Table 4. Range used for Bayesian hyperparameter search for DP-SGD and Semi-DP in Appendix G.1.4

G.3. Exact Numerical Results

n_{pub}/n	Algorithm 1	PDA-MD	DP-SGD	Throw-away
0.01	2.3526	1689.5905	2.7935	1689.5905
0.03	1.9719	1084.2257	2.1457	1084.2256
0.04	1.7626	776.1302	2.1417	776.1301
0.1	1.1648	1.1880	1.1695	3.1133
0.25	1.1260	1.1306	1.1346	1.3691
0.5	1.1043	1.1394	1.1065	1.1539
0.75	1.0946	1.1013	1.0937	1.1013
0.9	1.0787	1.0787	1.0787	1.0788
0.95	1.0725	1.0725	1.0725	1.0725

Table 5. Exact training results of curves reported in Figure 3.

n_{pub}/n	Algorithm 1	PDA-MD	DP-SGD	Throw-away
0.01	1.5020	1689.5905	1.5206	1689.5905
0.03	1.3363	1084.2257	1.3769	1084.2256
0.04	1.3196	776.1302	1.3776	776.1302
0.1	1.1201	1.1376	1.1221	3.1133
0.25	1.1030	1.1090	1.1103	1.3691
0.5	1.0911	1.1019	1.0999	1.1539
0.75	1.0863	1.1013	1.0877	1.1013
0.9	1.0787	1.0787	1.0787	1.0788
0.95	1.0725	1.0725	1.0725	1.0725

Table 6. Exact training results of curves reported in Figure 4.

ε	Algorithm 1	PDA-MD	DP-SGD	Throw-away
0.10	2.9509	3.1133	3.1133	3.1133
0.50	1.6841	1.9602	1.8588	3.1133
1.00	1.3125	1.3201	1.3158	3.1133
2.00	1.1648	1.1880	1.1695	3.1133
4.00	1.1201	1.1376	1.1221	3.1133
8.00	1.1088	1.1120	1.1104	3.1133

Table 7. Exact training results of curves reported in Figure 9.

ε	Algorithm 1	PDA-MD	DP-SGD	Throw-away
0.10	1.3691	1.3691	1.3691	1.3691
0.50	1.2647	1.3691	1.2679	1.3691
1.00	1.1764	1.2099	1.1880	1.3691
2.00	1.1260	1.1306	1.1346	1.3691
4.00	1.1030	1.1090	1.1103	1.3691
8.00	1.0976	1.1023	1.1041	1.3691

Table 8. Exact training results of curves reported in Figure 10.

n_{pub}/n	Algorithm 1	PDA-MD	DP-SGD	Throw-away
0.01	1.5175	2010.2422	1.5411	1689.5905
0.03	1.4225	2010.2422	1.5411	1084.2256
0.04	1.3989	2010.2422	1.5413	776.1302
0.1	1.3371	5.3822	1.5413	3.1133
0.25	1.2574	2.6205	1.5411	1.3691
0.5	1.1539	5.9698	1.5411	1.1539
0.75	1.1013	78.9970	1.5411	1.1013
0.9	1.0787	1053.6185	1.5411	1.0788
0.95	1.0725	1662.3572	1.5413	1.0725

Table 9. Exact training results of curves reported in Figure 5.

ε	Algorithm 1	PDA-MD	DP-SGD	Throw-away
0.10	3.1133	1956.6069	1830.1820	3.1133
0.50	3.1133	1092.1338	213.6728	3.1133
1.00	2.1843	306.5518	15.8089	3.1133
2.00	1.6313	34.4665	2.7226	3.1133
4.00	1.3359	5.3648	1.4746	3.1133
8.00	1.1986	2.4319	1.2472	3.1133

Table 10. Exact training results of curves reported in Figure 11.

n_{pub}/n	0.01	0.03	0.04	0.1	0.25	0.5	0.75	0.9	0.95
σ	2.490	2.529	2.568	2.744	3.252	4.805	9.531	23.672	47.344

Table 11. Standard Deviation σ of the private noise v_t used in experiment shown in Figure 3.

n_{pub}/n	0.01	0.03	0.04	0.1	0.25	0.5	0.75	0.9	0.95
σ	1.470	1.489	1.509	1.597	1.860	2.671	5.176	12.812	25.586

Table 12. Standard Deviation σ of the private noise v_t used in experiment shown in Figure 4 and 5.

H. Limitations

Limitations of Theoretical Results: Our theoretical results rely on certain assumptions (e.g convex, Lipschitz loss, i.i.d. data for SCO), that may be violated in certain applications. We leave it as future work to investigate the questions considered in this work under different assumptions (e.g. non-convexity, semi-DP SCO with *out-of-distribution* public data). Also, we reiterate that our theoretical results describe the optimal *worst-case* error. It might be the case that the worst-case distributions we construct in our lower bound proofs are unlikely to appear in practice. Thus, another interesting direction for future work would be to analyze "instance-optimal" (Asi & Duchi, 2020a;b) semi-DP error rates.

Limitations of Experiments: It is important to note that pre-processing and hyperparameter tuning were not done in a DP manner, since we did not want to detract focus from evaluation of the (fully tuned) semi-DP algorithms.[11] As a consequence, the overall privacy loss for the entire experimental process is higher than the ε indicated in the plots: the ε indicated in the plots solely reflects the privacy loss from running the algorithms with fixed hyperparameters and (pre-processed) data.

[11] See, e.g. (Liu & Talwar, 2019; Papernot & Steinke, 2022) and the references therein for discussion of DP hyperparameter tuning.