Lower Bounds For Public-Private Learning Under Distribution Shift

Amrith Setlur
CMU

Pratiksha Thaker
CMU

Jonathan Ullman
Northeastern University

Abstract

The most effective differentially private machine learning algorithms in practice rely on an additional source of purportedly public data. This paradigm is most interesting when the two sources combine to be more than the sum of their parts. However, there are settings such as mean estimation where we have strong lower bounds, showing that when the two data sources have the same distribution, there is no complementary value to combining the two data sources. In this work we extend the known lower bounds for public-private learning to a setting where the two data sources exhibit significant distribution shift. Our results apply to both Gaussian mean estimation where the two distributions have different means, and to Gaussian linear regression where the two distributions exhibit parameter shift. We find that when the shift is small (relative to the desired accuracy), either public or private data must be sufficiently abundant to estimate the private parameter. Conversely, when the shift is large, public data provides no benefit.

1 INTRODUCTION

Differential privacy (DP) [Dwork et al., 2006] is a formal privacy framework that makes it possible to perform statistical analysis or train large models while giving a strong guarantee of privacy to the individuals who contribute data. Informally, a differentially private algorithm does not reveal too much information about any one individual's data. While DP has been deployed successfully for statistics and machine learning, many of the most effective deployments work in a *public-private* setting in which there is access to an additional source of *public* data that can be used along with the private data. In this setting, the overall algorithm need not be fully differentially private—its output can reveal arbitrary information about the public data, but must limit the amount of information revealed about the private data. For example, the most effective differentially private predictive and generative models are often initialized with a large pretrained model that is not differentially private [Yu et al., 2021, Li et al., 2021], and the best models trained from scratch with only private data have much worse accuracy. This paradigm is most interesting when the public and private data combine to be more than the sum of their parts—by combining public and private data we can train an accurate model even though neither source on its own is large enough to train an accurate model.

However, not every application provides us with a useful source of public data. Moreover, when available, these purportedly public datasets may raise privacy concerns of their own [Tramèr et al., 2022]. As a result there is a growing body of theoretical work trying to understand whether the use of public data is *necessary* for private training, or can be replaced with better algorithms. These results give inconclusive answers, either showing that combining private and public data *does help* [Bie et al., 2022, Ganesh et al., 2023] or *doesn't help* [Bassily et al., 2020, Ullah et al., 2024, Lowy et al., 2024] depending on the specifics of the learning problem.

The prior work that most directly relates to ours studies *mean estimation* as a test case. Here we are given n *private samples* and m *public samples*, all drawn i.i.d. from a distribution P, and our goal is to design an algorithm that provides DP for the private samples but can do anything it wants with the public samples and estimates $\mu = \mathbb{E}[P]$. In this setting, it is known that combining public and private data doesn't help make more effective use of the private data [Bassily et al., 2020]. Throughout this work, when we say that combining private and public data *doesn't help*, we mean roughly

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

that whenever we are able to solve the learning problem using the combination of public and private data, then we could also have solved the learning problem (perhaps up to some small constant loss in accuracy) with just the public data or the private data alone.

Contributions. The above results raise two questions that we address in this work. First, prior work considers public and private data from the *same distribution*; we consider settings where the public and private data exhibit *distribution shift*, which is a more realistic setting for practical applications [Tramèr et al., 2022]. Second, we consider whether this phenomenon—that public data doesn't help make more effective use of private data—extends beyond mean estimation to more realistic supervised learning problems like *linear regression*.

Mean Estimation with Distribution Shift. In this setup, we assume the private samples are drawn from a distribution P and the public samples are drawn from some related distribution $Q \neq P$, and our goal is to estimate the mean of P. We know from prior work that when Q and P are the same, then there is no value to data from Q unless we already have enough samples from Q to solve the problem. We also know that holding the number of samples from Q fixed, shifting Q farther from P can only make the achievable accuracy worse. However, a priori it might be the case that shifting Q away from P, but giving many samples from Q, would allow us to find a sweet spot where the samples from Q are not enough on their own because of the shift, and samples from P are not enough because of privacy, but the two can be combined to achieve the desired accuracy.

Our first main result shows that this is not the case—that public data still doesn't help, even in settings where the public and private data come from different, but related, distributions. That is, we show that if we are able to estimate the mean up to some specified error, then *either* (1) the public data is abundant enough and its distribution Q is close enough to P that we can estimate the mean using only public data, *or* (2) the private data is abundant enough that we can privately estimate the mean using only the private data.

Specifically, we consider a setting where P and Q are both Gaussians with $P = N(\mu_P, \mathbf{I}_d)$ and $Q = N(\mu_Q, \mathbf{I}_d)$ we promise that $\|\mu_P - \mu_Q\|_2 \leq \tau$, and our goal is to find an estimate $\hat{\mu}_P$ such that $\|\hat{\mu}_P - \mu_P\|_2 \leq \alpha$. We assume access to n private samples from P and m public samples from Q. Note that if we have access to only private samples from P, then we could solve the problem if and only if $n = \Omega(d/\alpha^2 + d/\alpha\varepsilon)$. Also note that if $\tau \leq \alpha$ then, up to a factor of 2 in the error, we can simply use the public data as if it were from the same distribution as the private data. Thus, if we had access to only public samples from Q then we would need that both $\tau < \alpha$ and $m = \Omega(d/\alpha^2)$. Our result shows that, up to constant factors, we need either the private or public data to be sufficient on their own, and there is no complementarity between them.

Theorem 1.1 (Informal; mean estimation when $P \neq Q$)**.** *When the distribution shift is small, (i.e., $\tau \lesssim \alpha$) then to solve the mean estimation problem with error α in ℓ_2, we must have either enough public data alone to estimate the mean (i.e., $m = \Omega(d/\alpha^2)$), or enough total data to solve the mean estimation problem with differential privacy i.e., $n + m = \Omega(d/\alpha\varepsilon + d/\alpha^2)$. When the distribution shift is large, (i.e., $\tau \gtrsim \alpha$), then the entire sample complexity burden falls on private data and we must have enough private data to solve the mean estimation problem with differential privacy (i.e., $n = \Omega(d/\alpha\varepsilon + d/\alpha^2)$).*

Linear Regression without Distribution Shift. Here, we consider the richer setting of private linear regression with Gaussian covariates and Gaussian errors. That is, the data consist of pairs $(\vec{x}, y)$ where the features x are drawn from a standard multivariate Gaussian $N(0, \mathbf{I}_d)$ and $y|\vec{x} = \langle \beta, \vec{x} \rangle + \nu$ for some ground truth predictor β and error ν drawn from a standard Gaussian $N(0,1)$. Our goal is to estimate the optimal predictor β with small excess squared prediction error, which in this setting is equivalent to minimizing $\|\hat{\beta} - \beta\|_2^2$. For now we assume that the public and private data are both sampled from the same distribution.

Our second main result shows that the same qualitative phenomenon holds in this setting, that either the public data is sufficient to train a good predictor on its own, or the private data is sufficient to privately train a good predictor on its own.

Theorem 1.2 (Informal; linear regression when $\beta_P = \beta_Q$)**.** *To solve the linear regression problem with excess squared prediction error α^2, we either must have enough public samples alone to solve the linear regression problem (i.e., $m = \Omega(d/\alpha^2)$), or we must have enough total data to solve the problem with differential privacy (i.e., $n + m = \Omega(d/\alpha\varepsilon + d/\alpha^2)$).*

Linear Regression with Distribution Shift. Finally, we consider Gaussian linear regression with *parameter shift* between the public and private distributions. The only change to the setting above is that, for the private data, we have $y|\vec{x} = \langle \beta_P, \vec{x} \rangle + \nu$, and for the public data, we have $y|\vec{x} = \langle \beta_Q, \vec{x} \rangle + \nu$ for $\beta_P \neq \beta_Q$. Similar to shifted mean estimation, we parameterize the degree of distribution shift by a

parameter τ with the promise that $\|\beta_P - \beta_Q\|_2^2 \le \tau^2$. Note that, analogous to the case of mean estimation, if $\tau \le \alpha$, then up to a factor of 2 in the error, we can treat the public data as if it came from the same distribution as the private data. The goal is still to estimate β_P with small excess squared prediction error, or equivalently to minimize $\|\hat{\beta} - \beta_P\|_2^2$.

Analogous to the case of mean estimation with distribution shift and linear regression without distribution shift, our third main result shows that combining public and private data doesn't help in this setting.

Theorem 1.3 (Informal; linear regression when $\beta_P \neq \beta_Q$)**.** *When the distribution shift is small, (i.e., $\tau \lesssim \alpha$) then to solve linear regression with excess squared prediction error α^2, we must have either enough public data to solve the linear regression problem (i.e., $m = \Omega(d/\alpha^2)$), or enough total data to solve linear regression with differential privacy (i.e., $n + m = \Omega(d/\alpha\varepsilon + d/\alpha^2)$). When the distribution shift is large, (i.e., $\tau \gtrsim \alpha$), then the entire sample complexity burden falls on private data and we must have enough private data to solve the mean estimation problem with differential privacy (i.e., $n = \Omega(d/\alpha\varepsilon + d/\alpha^2)$).*

Proof Techniques. Our work builds on the fingerprinting method [Ullman, 2013, Bun et al., 2014, Dwork et al., 2015], which has been widely applied to proving lower bounds in differential privacy, including in the public-private setting [Bassily et al., 2020, Ullah et al., 2024, Lowy et al., 2024]. In particular, we adopt the recently introduced *Bayesian perspective on the fingerprinting method* [Narayanan, 2023]. As a starting point for our work, we can show that the Bayesian perspective gives an especially clean and simple way to establish lower bounds in the most basic setting of mean estimation with public-private data *without distribution shift,* recovering the result from [Bassily et al., 2020]. This new proof allows us to extend the results for mean estimation to settings with distribution shift (Theorem 1.1) by giving an easy way to choose a hard distribution for the public data and easily incorporating the influence of the public samples on the distribution of the private data using conjugate priors. This viewpoint also allows us to extend the lower bounds to settings like linear regression (Theorem 1.2), where lower bounds were previously only known for the private-only setting [Cai et al., 2021] and not the public-private setting. Finally, we can combine these two extensions to the setting of linear regression with parameter shift (Theorem 1.3). A key technique we employ in this setting is to reinterpret the parameter shift in linear regression as a form of non i.i.d. noise on the labels and then employ methods from generalized least-squares estimation. See Section 2 for a more thorough discussion of these techniques.

1.1 Related Work

The most closely related works to ours are [Bassily et al., 2020, Ullah et al., 2024, Lowy et al., 2024], which give lower bounds for mean estimation and stochastic convex optimization in the public-private setting using fingerprinting techniques, but only consider the setting where the public and private data come from the same distribution. We now review the related work on public-private estimation and machine learning as well as on fingerprinting lower bounds in differential privacy.

Public-private statistical estimation. A number of prior works have theoretically studied the benefits of public data in other settings, including mean estimation [Avent et al., 2020], query release [Liu et al., 2021, Bassily et al., 2020, Fuentes et al., 2024], convex optimization when gradients lie in a low-rank subspace [Kairouz et al., 2021, Amid et al., 2022], and non-convex optimization [Ganesh et al., 2023]. In particular [Ganesh et al., 2023] gives a non-convex optimization setting where public data is necessary to achieve high accuracy. In contrast to our work, many of these positive results formally assume that the public and private data come from the same distribution and do not directly address the question of how close the public and private data need to be in order for these positive results to hold.

The combination of public and private data has also been used effectively for practical machine learning in a number of domains [Abadi et al., 2016, Papernot et al., 2019, Tramer and Boneh, 2020, Luo et al., 2021, Yu et al., 2021, De et al., 2022, Ke et al., 2024, Ganesh et al., 2023, Li et al., 2021, Yu et al., 2021, Arora and Ré, 2022, Golatkar et al., 2022, Kurakin et al., 2022, He et al., 2022, Bu et al., 2023, Ginart et al., 2022], however these works are primarily empirical, whereas our work is focused on understanding whether combining public and private data from different distributions is necessary in the worst case.

Fingerprinting lower bounds in differential privacy. Fingerprinting methods have become the standard tool for establishing lower bounds on the cost of differentially private statistical estimation, particularly for high-dimensional problems. This method was first introduced in Ullman [2013], Bun et al. [2014], and was subsequently refined, used to prove lower bounds for a wide variety of problems [Dwork et al., 2014, Bassily et al., 2014, Dwork et al., 2015, Steinke

and Ullman, 2015a, 2017, Kamath et al., 2019, Cai et al., 2021, Narayanan et al., 2022, Kamath et al., 2022, Cai et al., 2023, Narayanan, 2023, Peter et al., 2024, Aliakbarpour et al., 2024, Portella and Harvey, 2024]. Relevant for our work is the so-called score attack for parameter estimation [Cai et al., 2023], which we adopt for our linear regression proofs. This general method has also been used for proving lower bounds outside of privacy, specifically for problems in adaptive data analysis [Hardt and Ullman, 2014, Steinke and Ullman, 2015b, Ullman et al., 2018, Lyu and Talwar, 2024] and the information-cost of learning [Attias et al., 2024]. Fingerprinting methods are also closely related to membership-inference attacks [Dwork et al., 2015] and privacy auditing [Jagielski et al., 2020].

2 TECHNICAL OVERVIEW

In this section, we describe the high-level structure of our proofs and highlight the new elements we introduce. We present a general technique for proving lower bounds for public-private computations under distribution shifts, starting from the private-only approach presented in Narayanan [2023] and showing how it can be naturally extended to proving lower bounds under distribution shift between public and private data.

Problem Setup and Prior Work: Private-Only Mean Estimation. In the private mean estimation problem we have a dataset X consisting of n samples $X_1,...,X_n$ drawn i.i.d. from a spherical multivariate normal distribution in $P = N(\mu, \mathbf{I}_d)$. Our goal is to design an (ε,δ)-differentially private algorithm $M(X)$ that returns an estimate $\hat{\mu}$ such that $\|\hat{\mu} - \mu\|_2 \leq \alpha$. Without privacy, the standard algorithm is to compute the empirical mean $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_i X_i$ and we will satisfy the accuracy goal as long as $n = \Omega(d/\alpha^2)$, which is also optimal. We want to know how the sample requirement changes when we require that the estimator is differentially private.

The main technique for proving such lower bounds is the so-called *fingerprinting method*. In this method we design a *test statistic* Z_i defined as:

$$Z_i = \langle X_i - \mu, \hat{\mu} - \mu \rangle \tag{1}$$

Intuitively, each Z_i measures the covariance between the estimate $\hat{\mu}$ and a single data point X_i. This test statistic is useful because: (1) an accurate $\hat{\mu}$ should depend on many data points X_i, so we would expect $\sum_i Z_i$ to be large on average, and (2) a private $\hat{\mu}$ should not depend significantly on the data points X_i, so we would expect the same sum to be small on average. Using these two observations, we can derive a contradiction showing accurate and private estimation is impossible for certain choices of parameters.

Formalizing and proving that the statistic $Z \stackrel{\text{def}}{=} \sum_i Z_i$ is *small* due to privacy is relatively straightforward using the definition of differential privacy. The tricky part is formalizing and proving that the statistic Z is large for any mechanism that is sufficiently accurate, and this is the crux of the argument, sometimes called the *fingerprinting lemma*. Although there are many proofs, we outline the main ideas in the proof of Narayanan [2023] as a starting point, which starts by writing the Markov chain:

$$\mu \longrightarrow X_1,...,X_n \longrightarrow \hat{\mu} \tag{2}$$

The key property of this Markov chain is that μ and $\hat{\mu}$ are independent conditioned on the data X. Now, a series of calculations gives us a lower bound on $\mathbb{E}_{\mu,X,\hat{\mu}}[Z]$ that will depend on specific properties of: (1) the *posterior mean* $\mathbb{E}[\mu \mid X]$, (2) the *posterior variance* $Var[\mu \mid X]$, and (3) the accuracy of the estimate $\|\hat{\mu} - \mu\|_2$. The third quantity is something we control by assumption and the first two quantities can be controlled by understanding the posterior distribution of μ conditioned on X. This framework makes it clear that we should choose a *conjugate prior* so that the posterior $\mu \mid X$ has a nice closed form. Now, we outline how we apply this purely-private framework above for establishing lower bounds in the presence of public data, especially when they are from a different distribution than private data.

Incorporating Public Data and Distribution Shift (Thm 1.1). We can discuss why the above Bayesian approach is ideal for incorporating public data into the lower bounds. Now we have the original distribution P with mean μ and samples $X_1,...,X_n$ that are private. We also have additional public data $X_{n+1},...,X_{n+m}$ samples i.i.d. from the same distribution, so the total dataset size is $N = n+m$. We want to design a mechanism $M(X_1,...,X_n,X_{n+1},...,X_{n+m})$ that is differentially private as a function of the first n samples.

Although our formal proof diverges slightly from this outline, for this overview, we can think about the same test statistic $Z = \sum_{i=1}^{n+m} Z_i$ computed over *both* the public and private samples. The main change in the proof is in the upper bound: we had used the fact that the mechanism is differentially private as a function of each of its samples, which is no longer true. However, we can separately bound $\sum_{i=1}^{n} Z_i$ using privacy and get a weaker bound on $\sum_{i=n+1}^{n+m} Z_i$ just using the variance of the distribution. Although the specific calculations are beyond the scope of this overview, we can use this outline to give a simpler proof of the lower bounds in [Bassily et al., 2020,

Ullah et al., 2024, Lowy et al., 2024] (captured as the special case of Theorem 1.1 where $\tau = 0$).

Now we turn to incorporating distribution shift. Here, we now assume that the private samples $X_1,...,X_n$ are drawn from $P = N(\mu_P, \mathbf{I}_d)$ and the public samples $X_{n+1},...,X_{n+m}$ are drawn from $Q = N(\mu_Q, \mathbf{I}_d)$ where $\|\mu_P - \mu_Q\|_2 \le \tau$. Note that increasing τ makes the public data less useful and thus should make the problem harder, and thus our goal is ultimately to prove stronger lower bounds on the number of private samples we need to solve the problem.

The main conceptual difference in the proof, and where we require a novel idea, is in the lower bound on Z. First, we think of $\mu_Q = \mu_P + v$ for some random variable v such that $\|v\|_2 \le \tau$ with high probability, and we will choose the distribution of v later. Now, if we elide the variable v, we have the Markov chain

$$\mu_P \longrightarrow X_1,...,X_n,X_{n+1},...,X_{n+m} \longrightarrow \hat{\mu}$$

where all these random variables are fully specified and where X contains all the public and private data. The key idea is to choose the random variable v determining the distribution shift in such a way that μ_P has a nice distribution. Specifically, we sample v according to a normal $N(0, \tau^2/d \cdot \mathbf{I}_d)$ so that $\|v\|_2 \le \tau$ with high probability and so that the distribution

$$(\mu_P, X_1,...,X_n,X_{n+1},...,X_{n+m})$$

becomes jointly normal in $d(n+m+1)$ variables with a particular block diagonal structure that makes it easy to work with. As a result the conditional distribution $\mu_P \,|\, X$ is now normal with a particular mean and covariance that we can write down and use to complete the analysis, although the exact calculations are again beyond the scope of this overview.

Extending to Linear Regression (Thms 1.2 and 1.3). An advantage of our proof structure is that it makes it easy to extend our lower bounds to richer problems like Gaussian linear regression. The main difference between our analysis of linear regression, compared to mean estimation, is in the choice of test statistic and in carefully defining which parts of the experiment we fix and which we don't.

First, in the private-only setting we have a distribution P over labeled samples $(\vec{x}, y)$ where we assume $\vec{x}$ has distribution $N(0, \mathbf{I}_d)$ and $y | \vec{x} = \langle \vec{x}, \beta \rangle + \nu$ where the error ν has distribution $N(0,1)$. Here β is the true parameter we want to estimate via a differentially private algorithm $\hat{\beta}$. To prove a tight lower bound for this setting, [Cai et al., 2021] introduced a suitable test statistic. We start by giving a simplified Bayesian version of their analysis. One key step in our analysis is to think of the feature vectors $\vec{x}_i$ as *fixed* and consider the Markov chain:

$$\beta \longrightarrow y_1,...,y_n \longrightarrow \hat{\beta} \tag{3}$$

Note that we will have to average over the vectors $\vec{x}_i$ at some point, but since they are normally distributed, the covariance of these vectors is tightly concentrated around $\mathbf{I}_d$, so we don't think of this as an important source of randomness. The advantage of considering this Markov chain is that, for every fixed choice of the feature vectors, $y \,|\, \beta$ is now normally distributed. Therefore, if we choose a normal prior for β, we will get that the distribution (β, y) is jointly Gaussian, and $\beta \,|\, y$ is Gaussian with a tractable closed form. These observations are enough to prove the lower bound for public-private linear regression with no distribution shift (Thm 1.2), albeit with significant technical challenges involved in completing the analysis.

Finally, we consider the setting with *parameter shift*. Here, as before there are n private samples and m public samples. All samples have features $\vec{x}$ chosen from $N(0, \mathbf{I}_d)$, but now we assume that for the private samples we have $y \,|\, \vec{x} = \langle \vec{x}, \beta_P \rangle + \nu$ and for the public samples we have $y \,|\, \vec{x} \langle \vec{x}, \beta_Q \rangle + \nu$, for $\|\beta_P - \beta_Q\|_2 \le \tau$. Here we can apply a similar idea as in the case of mean estimation. We consider the parameter shift $\beta_Q = \beta_P + v$ for a random variable v that is chosen from $N(0, \tau^2/d \cdot \mathbf{I}_d)$ so that $\|v\|_2 \le \tau$. Now we elide v and consider the joint distribution:

$$(\beta_P, y_1,...,y_n,y_{n+1},...,y_{n+m}) \tag{4}$$

This distribution is jointly normal in $n + m + d$ variables and therefore the distribution of β_P conditioned on all the private and public data is also normal with tractable mean and covariance, which we can use to complete the proof, albeit with significant technical complexities. We overcome these technical challenges by using techniques from *generalized least squares regression* where we think of the parameter shift as instead being a form of Gaussian label noise, but with a non i.i.d. structure, which helps us suitably modify the test statistic and complete the analysis.

3 PRELIMINARIES

We briefly review the definition of differential privacy and its application in a setting with public and private data. Let $X = (X_1,...,X_n) \in \mathcal{X}^n$ be a dataset consisting of n elements. We say two datasets $X, X' \in \mathcal{X}^n$ are *neighboring* if they differ on at most one element.

Definition 3.1 (Differential Privacy [Dwork et al., 2006]). An algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ is (ε, δ)-*differentially private* if for every pair of

neighboring datasets $X, X' \in \mathcal{X}^n$ and every $S \subseteq \mathcal{Y}$: $\Pr(M(X) \in S) \leq e^{\varepsilon}\Pr(M(X') \in S) + \delta$.

We consider learning algorithms that have access to both public and private data. Specifically, we have a dataset $X = (X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+m})$, where we think of the first n samples as *private* and the last m samples as *public*. In this setting we say that an algorithm $M : \mathcal{X}^{n+m} \to \mathcal{Y}$ is (ε, δ)-*differentially private with public and private data* if, for every possible fixed choice of the public samples $x_{n+1}, \ldots, x_{n+m}$, the algorithm $M(X_1, \ldots, X_n, x_{n+1}, \ldots, x_{n+m})$, is (ε, δ)-differentially private as a function of the first n elements. Note that M may depend in an arbitrary way as a function of the public samples $X_{n+1}, \ldots, X_{n+m}$. In our analysis we will rely heavily on the following technical lemma about random variables that are *indistinguishable* in the sense guaranteed by differential privacy.

Lemma 3.2 (See e.g.[Narayanan, 2023]). *Suppose $0 \leq \varepsilon \leq 1$ and $0 \leq \delta \leq \frac{1}{2}$, and that A, B are real-valued random variables such that for any set S, $e^{-\varepsilon} \cdot \Pr(A \in S) - \delta \leq \Pr(B \in S) \leq e^{\varepsilon} \cdot \Pr(A \in S) + \delta$. Then, $|\mathbb{E}[A - B]| \leq 2\varepsilon \cdot \mathbb{E}[|A|] + 2\sqrt{\delta} \cdot \mathbb{E}[A^2 + B^2]$.*

4 PUBLIC-PRIVATE MEAN ESTIMATION UNDER DISTRIBUTION SHIFT

In this section, we elaborate on our analysis of private mean estimation for a Gaussian distribution with unknown mean and identity covariance, given access to n private samples, as well as m public samples drawn from another Gaussian with a shifted mean.

First we give some notation to state our result and sketch the proof. We have a matrix comprised of m public examples $X_{\text{pub}} \in \mathbb{R}^{m \times d}$ and n private examples $X_{\text{priv}} \in \mathbb{R}^{n \times d}$. In particular, we assume that $X_{\text{priv},i} \sim \mathcal{N}(\mu_{\text{priv}}, \mathbf{I}_d)$, for every $i \in [n]$, and $X_{\text{pub},i} \sim \mathcal{N}(\mu_{\text{pub}}, \mathbf{I}_d)$, for every $i \in [m]$, such that $\mu_{\text{pub}} = \mu_{\text{priv}} + v$ for some vector v such that $\|v\|_2 \leq \tau$. Together, X_{pub} and X_{priv} give us $\mathcal{D}$ available to the learner M (where M must satisfy (ε, δ)-DP with respect to the private samples).

We split the proof of Theorem 1.1 in two parts. The most significant part is to analyze the case where the distribution shift dominates the overall error. That is, if we were to try using the empirical mean of the *public* data $\hat{\mu}_{\text{pub}} = \frac{1}{m}\sum_{i=1}^{m} X_{\text{pub},i}$ as an estimate of the mean of the private data, then the error would be $\|\mu_{\text{priv}} - \hat{\mu}_{\text{pub}}\|_2^2 = \tau^2 + d/m$, where the first term comes from the shift and the second term comes from the sampling error in the public data. When the sampling error dominates, the shift is largely irrelevant, so our analysis matches the case where $\tau = 0$ so there is no shift. Our analysis for this case is captured in Theorem 4.1.

Theorem 4.1 (Mean estimation lower bound for $\tau = 0$). *Fix any $\alpha > 0$ with $\alpha = O(\sqrt{d})$. Suppose M is an (ε, δ)-DP learner s.t. $\mathbb{E}_{\mathcal{D}}\|M(\mathcal{D}) - \mu_{\text{priv}}\|_2 \leq \alpha$, and $\mathcal{D}$ comprises X_{pub} and X_{priv} drawn from the same distribution with mean μ_{priv}. When $\delta \leq \varepsilon^2/d$, either $m = \Omega(d/\alpha^2)$ or $n + m = \Omega(d/\alpha\varepsilon + d/\alpha^2)$.*

Note that the above theorem remains true when $\tau > 0$, as increasing τ can only make it harder to obtain an accurate estimate. However, when $\tau \gg \sqrt{d/m}$ so that the distribution shift dominates the overall error, we want to prove a much stronger lower bound than what is in Theorem 4.1, showing that no amount of public data is useful for estimating the mean of the private data. This result is captured in Theorem 4.2.

Theorem 4.2 (Mean estimation lower bound for large $\tau > 0$). *Fix any $\alpha > 0, \tau > 0$, and $\alpha = O(\sqrt{d})$. Suppose M is an (ε, δ) private learner such that $\mathbb{E}_{\mathcal{D}}\|M(\mathcal{D}) - \mu_{\text{priv}}\|_2 \leq \alpha$, and additionally $\delta < \varepsilon^2/d$. When the shift is large, so that $\|\mu_{\text{pub}} - \mu_{\text{priv}}\|_2 = \tau = \omega(\sqrt{d/m})$, then either $\tau < \alpha$, which implies $m = \Omega(d/\alpha^2)$ or $n = \Omega(d/\alpha\varepsilon + d/\alpha^2)$.*

Proof Sketch. Our goal is to establish a minimax lower bound on the number of public and private samples needed for any (ε, δ) private learner M that is α-accurate in its estimation of the private mean, *i.e.*, $\|M(X) - \mu_P\|_2 \leq \alpha$, and for a fixed value of the shift parameter $\tau > 0$. Following the typical approach of minimax lower bounds, we lower bound the minimax risk with the Bayes risk under a carefully chosen prior over the unknown parameters: μ_P, v, i.e., the private mean we want to estimate and the shift vector. The private mean μ_P is sampled from a spherical Gaussian $\mu_P \sim \mathcal{N}(0, \sigma^2\mathbf{I}_d)$. Recall that $\mu_Q = \mu_P + v$. We define the prior over v as $v \sim \mathcal{N}(\mathbf{0}, \tau^2/d \cdot \mathbf{I}_d)$. Thus, in expectation over the sampling of v, for any value of μ_P, $\mathbb{E}\|\mu_P - \mu_Q\|_2^2 = \tau^2$. The fact that there is a small probability of $\|\mu_P - \mu_Q\|_2^2 \gg \tau^2$ will not impact our proof significantly.

Next, we define our test statistic Z_i for our fingerprinting method, matching the general form in Eq. 1. In particular, when $i \in [1, n]$, *i.e.*, Z_i tries to measure the correlation between the mechanism output $M(X)$ and a private sample X_i, the statistic is given by:

$$Z_i = \langle X_i - \mu_P, M(X) - \mu_P \rangle \quad i \in \{1 \ldots n\}.$$

More interestingly, the statistic that measures the correlation with a public datapoint X_j, where

$j \in \{n+1,...,n+m\}$ is now re-weighted by a value that depends on the degree of shift τ, *i.e.*,

$$Z_j = \frac{1}{m \cdot \tau^2/d + 1} \cdot \langle X_j - \mu_P, M(X) - \mu_P \rangle,$$

where $j \in \{n+1 \dots m+n\}$. Essentially, when we compute the sum $\sum_i Z_i$, the weight on Z_i corresponding to a public datapoint is $\Omega(1)$, when the shift is small: $\tau = O(\sqrt{d/m})$, and up to constants matches the setting with no distribution shift $\tau = 0$. On the other hand, as $\tau \to \infty$, and the error from the shift dominates the sampling error over public data $\tau = \omega(\sqrt{d/m})$, the weight on public Z_i is $o(1)$. Intuitively, we expect any accurate mechanism $M(X)$ to reduce its sensitivity over public data as the shift increases, and a similar effect is also captured by our test statistic. We upper bound the expected sum as

$$\mathbb{E}\left[\sum_i Z_i\right] = O\left(n\varepsilon\alpha + \frac{1}{1+m\cdot\tau^2/d} \cdot \left(\alpha\sqrt{md} + \alpha\tau\sqrt{m}\right)\right)$$

using differential privacy arguments we outline in Section 2. Then, we lower bound the expected value of the sum $\mathbb{E}[\sum_i Z_i]$ with $\Omega(d)$ when $\tau = O(\sqrt{d/m})$. For the lower bound, we use the Markov chain argument outlined in Section 2, and compute the posterior over $\mu_P \mid X$ given all the data, where the posterior mean has lower order dependence on empirical mean of the public data, as the shift grows. Together, with the upper bound on $\mathbb{E}[\sum_i Z_i]$, the $\Omega(d)$ lower bound gives us the final result we need on the sample complexity of n and m for small values of τ. For large values of τ, the weight over public data points in our test statistic vanishes. In addition, we also need that $m = \Omega(d/\alpha^2)$, for the public data to estimate its own mean μ_Q to an accuracy of α. Along with the condition on τ, we establish when $\alpha \gtrsim \tau$, only then we expect the public data to be helpful, when the shift is large ($\tau = \omega(\sqrt{d/m})$). □

5 PUBLIC-PRIVATE LINEAR REGRESSION

In this section, we analyze linear regression given access to n private and m public samples from the *same* distribution. Prior lower bounds in this setting were limited to the setting where the learner only has access to private samples [Cai et al., 2021]. Since the full proofs are technical, this section will only include theorem statements and a sketch of the key differences between the proofs in this section and those in Section 4, and the formal proofs will be deferred to the appendices.

5.1 Lower Bounds w/o Distribution Shift

First we consider the case where the public and private samples come from the same distribution. Once again, we have n private samples and m public samples where $N = n+m$. In particular, we assume $N > 100d$, i.e. we have enough samples to estimate the covariance matrix. Each we assume there is some optimal linear model $\beta \in \mathbb{R}^d$ and we assume that each sample (x_i, y_i) is drawn independently from the distribution defined by $y_i = x_i^\top \beta + \eta_i$, where $x_i \sim \mathcal{N}(0, I_d)$ and $\eta_i \sim \mathcal{N}(0, \sigma^2)$, and $\beta \in \mathbb{R}^d$ is the regression parameter of interest. Since we will need to perform linear algebra with the data, we will use the matrix $X \in \mathbb{R}^{n \times d} = [x_1, ..., x_N]^\top$ and the vector $y = (y_1, ..., y_N)$.

Theorem 5.1. *Let X, y be generated according to the model above. Let M be an (ε, δ)-differentially private algorithm such that $\mathbb{E}_{X,y}[\|M(X,y) - \beta\|_2] < \alpha$ for some $\alpha = O(1)$. When $\delta \le \varepsilon^2/d$, either $m = \Omega(d/\alpha^2)$, or $n + m = \Omega(d/\alpha\varepsilon + d/\alpha^2)$.*

Note that Theorem 5.1 corresponds directly to Theorem 4.1, except for the problem of linear regression instead of mean estimation. Specifically, it shows that when public and private samples are drawn from the same distribution, linear regression suffers the same fate as mean estimation: either the problem can be solved entirely using public samples, or we need enough samples to solve the problem privately. In the next section, we will extend this model to a shifted setting in which the private outcomes y_i are generated from a model with parameter β while the public outcomes are generated from a shifted model with parameter $\beta + v$ for some vector v such that $\|v\|_2 \le \tau$, which corresponds to Theorem 4.2.

Proof Sketch. The proof follows the same high-level outline that we employ for mean estimation lower bounds. Specifically, we will again use the Bayesian view of the fingerprinting lemma, where we consider β sampled from some prior $\mathcal{N}(0, 1/b \cdot I_d)$ for some parameter b, and analyze the posterior distribution of β conditioned on the public and private data. We also modify the test statistic using the *score attack* from [Cai et al., 2021], specifically defining

$$Z_i = \langle \hat{\beta} - \beta, (y_i - x_i^\top \beta) x_i \rangle.$$

The main ingredient in the proof is how we show that the test statistic is large when the estimated parameters $\hat{\beta}$ are close to the true parameter β, to obtain a lower bound on the error. We condition on a fixed, typical value of the features X and then compute the mean and covariance of the Gaussian posterior distribution $\beta \mid X$, which is significantly

more technical than in the case of mean estimation due to the presence of inverses of random matrices. □

5.2 Distribution Shift

We now consider a distribution shift setting for linear regression. In particular, we consider the effect of public samples drawn from a linear model with the same covariate and noise distribution as the private samples, but with a shifted *parameter* β_{pub}. This setting once again fits cleanly into the Bayesian framework we have outlined, and we reach a similar conclusion to that in mean estimation: when τ is small, the public data is as useful as it is when there is no shift, but if τ grows too large, the problem must be solved using private data alone. We again observe $N=n+m$ i.i.d. samples, but now the public data is drawn from a linear model with a potentially different set of parameters:

$$\begin{aligned} y_i &= x_i^\top \beta_{\text{priv}} + \eta_i,\ i \in [n], j \in [n+1:n+m] \\ y_j &= x_j^\top \beta_{\text{pub}} + \eta_j = x_j^\top (\beta_{\text{priv}} + v) + \eta_j. \end{aligned} \tag{5}$$

Again, we assume $x_i \sim \mathcal{N}(0, I_d)$ and $\eta_i \sim \mathcal{N}(0,\sigma^2)$ are independent. Here the *private* parameter is $\beta_{\text{priv}} \in \mathbb{R}^d$ and the *public* parameter is shifted by an unknown $v \in \mathbb{R}^d$ such that $\|v\|_2^2 = \tau^2$, so $\beta_{\text{pub}} = \beta_{\text{priv}} + v$. The learner receives the full data set $X = \{(x_1,y_1),\ldots,(x_N,y_N)\}$ but must be (ε,δ)-DP with respect to the private rows $\{1,\ldots,n\}$.

Throughout we assume $n,m > 100d$ so that the sample covariance matrices are well conditioned. Also, as in Section 5, the prior on the unknown regression vector is $\beta_{\text{priv}} \sim \mathcal{N}\big(0, 1/d \cdot I_d\big)$. The theorem below mirrors the structure of Theorem 4.2 for mean estimation and of Theorem 5.1 for in-distribution linear regression. As in mean estimation, for any $\tau > 0$, the lower bound in Theorem 5.1 applies. Again, we are interested in the case where τ is so large that no amount of public data will be useful and the entire burden falls on the private samples, i.e. $\tau = \omega(\sqrt{d/m})$.

Theorem 5.2 (Linear regression lower bound for large $\tau > 0$)**.** *Fix any $\alpha > 0, \varepsilon > 0, \tau > 0$ and assume $\alpha, \varepsilon = O(1)$. Suppose M is an (ε,δ)-DP mechanism such that $\mathbb{E}[\|M(D) - \beta_{\text{priv}}\|_2^2] \le \alpha^2$, with $\delta < \varepsilon^2/d$ When the shift is large, so that $\|\beta_{\text{pub}} - \beta_{\text{priv}}\|_2 = \omega(\sqrt{d/m})$, then either $\tau < \alpha$, which implies $m = \Omega(d/\alpha^2)$, or $n = \Omega(d/\alpha\varepsilon + d/\alpha^2)$.*

Proof sketch. While the techniques we use for shifted linear regression are similar to those for shifted mean estimation, the shift introduces a correlated (heteroskedastic) noise term into the linear regression model. Specifically, if we fix the covariates X, then we can write the labels as Gaussian with mean $X\beta_{\text{priv}}$, and label noise coming from a Gaussian with non-spherical covariance

$$\Sigma = \sigma^2 I_N + \tfrac{\tau^2}{d} P P^\top,$$

where $P = [0, X_{\text{pub}}]$ is defined from the public covariates. That is, the distribution of the label errors $y - X^\top\beta$ follows an arbitrary correlated Gaussian rather than i.i.d. Gaussian noise on each label. This setting is referred to as *generalized least squares (GLS)* in the literature, to contrast with the more standard ordinary least squares setup, and we borrow techniques from the analysis of this GLS problem to guide our analysis. Appendix D adapts the Bayesian fingerprinting framework of Section 2 to the shifted linear regression setting. To account for the non-spherical label noise, we modify the test statistic appropriately:

$$\sum_i Z_i = \langle M(D) - \beta_{\text{priv}}, X^\top \Sigma^{-1}(X\beta_{\text{priv}} - y)\rangle \tag{6}$$

Intuitively the purpose of multiplying by Σ^{-1} is that it restores the property that each of the terms in the inner product is distributed as a spherical Gaussian, which we relied on in the case of ordinary least squares. The proof then follows the same general structure but with additional technical details arising from the presence of this correction. The main feature of the analysis is that correction nicely interpolates between the case where τ is large where the public samples are too noisy to be useful and the case where τ is small and the public and private samples are very similar.

6 DISCUSSION

In this paper, we discuss lower bounds on the sample complexity requirements for private Gaussian mean estimation and private linear regression in the presence of public data under distribution shift between the public and private distributions. Across both mean estimation and linear regression, we establish that whenever the learning objective is achievable using the combination of public and private data, one of the two sources—public or private—must be sufficient on its own whenever the distribution shift is smaller than the accuracy budget. From a methodological standpoint, our work extends the fingerprinting lower bound technique to new regimes: (1) distribution shift in public-private mean estimation, and (2) private linear regression with and without parameter shift. By adopting a Bayesian lens on the fingerprinting method [Narayanan, 2023], we simplify existing proofs and open the door to analyzing more realistic learning settings within the learning framework of using public data to improve privacy-utility tradeoffs under differentially private learning.

7 ACKNOWLEDGMENTS

The authors would like to thank Virginia Smith, Moshe Shenfeld, and Zhiwei Steven Wu for many helpful discussions early on in this work. AS gratefully acknowledges the support of JP Morgan AI PhD Fellowship.

References

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

Alexander C Aitken. IV.—On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48, 1936.

Maryam Aliakbarpour, Konstantina Bairaktari, Adam Smith, Marika Swanberg, and Jonathan Ullman. Privacy in metalearning and multitask learning: Modeling and separations. *arXiv preprint arXiv:2412.12374*, 2024.

Ehsan Amid, Arun Ganesh, Rajiv Mathews, Swaroop Ramaswamy, Shuang Song, Thomas Steinke, Vinith M Suriyakumar, Om Thakkar, and Abhradeep Thakurta. Public data-assisted mirror descent for private model training. In *International Conference on Machine Learning*, pages 517–535. PMLR, 2022.

Simran Arora and Christopher Ré. Can foundation models help us achieve perfect secrecy? *arXiv preprint arXiv:2205.13722*, 2022.

Idan Attias, Gintare Karolina Dziugaite, Mahdi Haghifam, Roi Livni, and Daniel M Roy. Information complexity of stochastic convex optimization: Applications to generalization and memorization. *arXiv preprint arXiv:2402.09327*, 2024.

Brendan Avent, Yatharth Dubey, and Aleksandra Korolova. The power of the hybrid model for mean estimation. *Proceedings on Privacy Enhancing Technologies*, 4:48–68, 2020.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pages 464–473. IEEE, 2014.

Raef Bassily, Albert Cheu, Shay Moran, Aleksandar Nikolov, Jonathan Ullman, and Steven Wu. Private query release assisted by public data. In *International Conference on Machine Learning*, pages 695–703. PMLR, 2020.

Alex Bie, Gautam Kamath, and Vikrant Singhal. Private estimation with public data. *Advances in Neural Information Processing Systems*, 35: 18653–18666, 2022.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Differentially private optimization on large model at small cost. In *International Conference on Machine Learning*, pages 3192–3218. PMLR, 2023.

Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 1–10, 2014.

T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021.

T Tony Cai, Yichen Wang, and Linjun Zhang. Score attack: A lower bound technique for optimal differentially private learning. *arXiv preprint arXiv:2303.07152*, 2023.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

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

Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 11–20, 2014.

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

Miguel Fuentes, Brett C Mullins, Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Joint selection: Adaptively incorporating public information for private synthetic data. In *International Conference on Artificial Intelligence and Statistics*, pages 2404–2412. PMLR, 2024.

Arun Ganesh, Mahdi Haghifam, Milad Nasr, Sewoong Oh, Thomas Steinke, Om Thakkar, Abhradeep Guha Thakurta, and Lun Wang. Why is public pretraining necessary for private model training? In *International Conference on Machine Learning*, pages 10611–10627. PMLR, 2023.

Antonio Ginart, Laurens van der Maaten, James Zou, and Chuan Guo. Submix: Practical private prediction for large-scale language models. *arXiv preprint arXiv:2201.00971*, 2022.

Aditya Golatkar, Alessandro Achille, Yu-Xiang Wang, Aaron Roth, Michael Kearns, and Stefano Soatto. Mixed differential privacy in computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8376–8386, 2022.

Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *2014 IEEE 55th annual symposium on foundations of computer science*, pages 454–463. IEEE, 2014.

Jiyan He, Xuechen Li, Da Yu, Huishuai Zhang, Janardhan Kulkarni, Yin Tat Lee, Arturs Backurs, Nenghai Yu, and Jiang Bian. Exploring the limits of differentially private deep learning with group-wise clipping. In *The Eleventh International Conference on Learning Representations*, 2022.

Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 33: 22205–22216, 2020.

Peter Kairouz, Monica Ribero Diaz, Keith Rush, and Abhradeep Thakurta. (nearly) dimension independent private erm with adagrad via publicly estimated subspaces. In *Conference on Learning Theory*, pages 2717–2746. PMLR, 2021.

Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pages 1853–1902. PMLR, 2019.

Gautam Kamath, Argyris Mouzakis, and Vikrant Singhal. New lower bounds for private estimation and a generalized fingerprinting lemma. *Advances in neural information processing systems*, 35: 24405–24418, 2022.

Shuqi Ke, Charlie Hou, Giulia Fanti, and Sewoong Oh. On the convergence of differentially-private fine-tuning: To linearly probe or to fully fine-tune? *arXiv preprint arXiv:2402.18905*, 2024.

Alexey Kurakin, Shuang Song, Steve Chien, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. Toward training at imagenet scale with differential privacy. *arXiv preprint arXiv:2201.12328*, 2022.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2021.

Terrance Liu, Giuseppe Vietri, Thomas Steinke, Jonathan Ullman, and Steven Wu. Leveraging public data for practical private query release. In *International Conference on Machine Learning*, pages 6968–6977. PMLR, 2021.

Andrew Lowy, Zeman Li, Tianjian Huang, and Meisam Razaviyayn. Optimal differentially private model training with public data. In *Forty-first International Conference on Machine Learning*, 2024.

Zelun Luo, Daniel J Wu, Ehsan Adeli, and Li Fei-Fei. Scalable differential privacy with sparse network finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5059–5068, 2021.

Xin Lyu and Kunal Talwar. Fingerprinting codes meet geometry: Improved lower bounds for private query release and adaptive data analysis. *arXiv preprint arXiv:2412.14396*, 2024.

Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL http://probml.github.io/book1.

Shyam Narayanan. Better and simpler lower bounds for differentially private statistical estimation. *arXiv preprint arXiv:2310.06289*, 2023.

Shyam Narayanan, Vahab Mirrokni, and Hossein Esfandiari. Tight and robust private mean estimation with few users. In *International Conference on Machine Learning*, pages 16383–16412. PMLR, 2022.

Nicolas Papernot, Steve Chien, Shuang Song, Abhradeep Thakurta, and Ulfar Erlingsson. Making the shoe fit: Architectures, initializations, and tuning for learning with privacy. *arXiv*, 2019.

Naty Peter, Eliad Tsfadia, and Jonathan Ullman. Smooth lower bounds for differentially private algorithms via padding-and-permuting fingerprinting codes. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4207–4239. PMLR, 2024.

Victor S Portella and Nick Harvey. Lower bounds for private estimation of gaussian covariance matrices under all reasonable parameter regimes. *arXiv preprint arXiv:2404.17714*, 2024.

Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *arXiv preprint arXiv:1501.06095*, 2015a.

Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Conference on learning theory*, pages 1588–1628. PMLR, 2015b.

Thomas Steinke and Jonathan Ullman. Tight lower bounds for differentially private selection. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 552–563. IEEE, 2017.

Florian Tramer and Dan Boneh. Differentially private learning needs better features (or much more data). In *International Conference on Learning Representations*, 2020.

Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Considerations for differentially private learning with large-scale public pretraining. *arXiv preprint arXiv:2212.06470*, 2022.

Enayat Ullah, Michael Menart, Raef Bassily, Cristóbal Guzmán, and Raman Arora. Public-data assisted private stochastic optimization: Power and limitations. *arXiv preprint arXiv:2403.03856*, 2024.

Jonathan Ullman. Answering $n^{2+o(1)}$ counting queries with differential privacy is hard. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 361–370, 2013.

Jonathan Ullman, Adam Smith, Kobbi Nissim, Uri Stemmer, and Thomas Steinke. The limits of post-selection generalization. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*, 2021.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A PRIVATE MEAN ESTIMATION ASSISTED WITH PUBLIC DATA

In this section, we analyze the hardness of private mean estimation for a Gaussian distribution with unknown mean and identity covariance. In addition to i.i.d. private samples, we are given i.i.d. public samples which can either be from the same distribution as the private samples, or from a different one. For this, we apply the Bayesian lower framework (from Section 1) to lower bound the expected mean of the fingerprinting statistics we will define shortly. The final lower bound on the error of any private mean estimation algorithm is then obtained by proving a tight upper bound on the expected mean of the fingerprinting statistics.

Setup. We have a matrix comprised of m public examples $X_{\text{pub}} \in \mathbb{R}^{m\times d}$ and n private examples $X_{\text{priv}} \in \mathbb{R}^{n\times d}$, where each row in them belong to the data universe $\mathcal{X} \subset \mathbb{R}^d$. In particular, we assume that $X_{\text{priv},i} \sim \mathcal{N}(\mu_{\text{priv}}, \mathbf{I}_d), \forall i \in [n]$, and $X_{\text{pub},i} \sim \mathcal{N}(\mu_{\text{pub}}, \mathbf{I}_d), \forall i \in [n,n+m]$, such that $\mu_{\text{pub}} = \mu_{\text{priv}} + \tau v$. When $\tau > 0$, then we are in the distribution shift setting, where the public and private data distributions observe a shifted mean, and when $\tau = 0$, then we are in the no-distribution shift setting. Together, X_{pub} and X_{priv} give us the dataset X available to the learner M. The learner M must run private computation over the dataset X and output $M(X)$ satisfying (ε,δ) differentially privacy with respect to the private dataset X_{priv}. We will use $N \stackrel{\text{def}}{=} n+m$ to denote the total number of examples.

Our goal is to establish a minimax lower bound on the number of public and private samples needed for any (ε,δ) private learner M that is α accurate in its estimation of the private mean, *i.e.*, $\|M(X) - \mu_{\text{priv}}\|_2 \leq \alpha$. Since minimax risk is lower bounded by the Bayes risk under any prior, we proceed by lower bounding the Bayes risk under the following prior. For a fixed and known τ, the private mean μ_{priv} is sampled from a spherical Gaussian $\mu_{\text{priv}} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$, and $\mu_{\text{pub}} = \mu_{\text{priv}} + \tau v$ where the displacement vector $v \sim \mathcal{N}(0, \mathbf{I}_d)$.

A.1 Lower Bound for Mean-Estimation when Public and Private Distributions are Identical

In Theorem A.1, we present an asymptotic lower bound on the number of public and private samples needed by any (ε,δ) private learner M, that is α accurate in estimating the private mean μ_{priv}, in ℓ_2 error.

Theorem A.1 (Mean estimation lower bound when public and private datapoints are i.i.d. $(\tau = 0)$)**.** *Fix any $\alpha > 0$. Suppose M is an (ε,δ) private learner that satisfies $\mathbb{E}_{X,\mu_{\text{priv}}} \|M(X) - \mu_{\text{priv}}\|_2 \leq \alpha$ for a prior distribution of $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$ over μ_{priv}. Here, the dataset X comprises of n i.i.d. private samples from $\mathcal{N}(\mu_{\text{priv}}, \mathbf{I}_d)$, and m i.i.d. public samples from $\mathcal{N}(\mu_{\text{priv}}, \mathbf{I}_d)$. When $\delta \leq \varepsilon^2/d$, either $m = \Omega(d/\alpha^2)$, or $n+m = \Omega(d/\alpha\varepsilon + d/\alpha^2)$.*

Proof. We begin by providing an overview of the proof technique via fingerprinting statistics we define shortly.

Overview. First, we define the fingerprinting statistics, which are random variables, that depend on the randomness in the learner M, the sampling of the dataset X and the unknown problem parameters μ_{priv}, v. Informally, fingerprinting statistics are random variables which cannot take very high values without violating the privacy or the accuracy of the learner. Next, we show that under appropriate priors, the sum of these fingerprinting statistics cannot be too low without violating statistical lower bounds. Together, the privacy and statistical error bounds give us the final result in Theorem A.1.

Fingerprinting statistics. We use $\{Z_i\}_{i=1}^N$, one for each element in the dataset (includes both private and public samples), to denote the fingerprinting statistic.

$$Z_i = \langle M(X) - \mu_{\text{priv}}, X_i - \mu_{\text{priv}} \rangle, \quad \forall i \in [N] \tag{7}$$

Similarly, we define the dataset X'_i that replaces the i^{th} element in X_{priv} with a random sample from $\mathcal{N}(\mathbf{0}, \mu_{\text{priv}})$ if $i \in [n]$ and from $\mathcal{N}(\mathbf{0}, \mu_{\text{pub}})$ otherwise:

$$Z'_i = \langle M(X'_i) - \mu_{\text{priv}}, X_i - \mu_{\text{priv}} \rangle, \quad \forall i \in [N] \tag{8}$$

Proposition A.2. *Assume the conditions in Theorem A.1. Then, $\mathbb{E}[Z'_i] = 0$ and $\mathbb{E}[(Z'_i)^2] = O(\alpha^2), i \in [n+m]$, where the expectation is taken over the sampling of μ_{priv}, X, and the randomness in the mechanism M.*

Proof. When conditioned on μ_{priv}, $X_i - \mu_{\text{priv}}$ is independent of $M(X'_i) - \mu_{\text{priv}}$, and $\mathbb{E}[X_i - \mu_{\text{priv}} \mid \mu_{\text{priv}}] = 0$. Thus, we have $\mathbb{E}[Z'_i] = \mathbb{E}[\mathbb{E}[Z'_i \mid \mu_{\text{priv}}]] = 0$. Next, we bound $\mathbb{E}[(Z'_i)^2 | \mu_{\text{priv}}]$ and similarly use the law of total

expectation to get the upper bound on $\mathbb{E}[(Z_i')^2]$.

$$\begin{aligned}
&\mathbb{E}[(Z_i')^2|\mu_{\text{priv}}] \\
=&\mathbb{E}\left[\sum_{j,k}(M(X_i')-\mu_{\text{priv}})_j(M(X_i')-\mu_{\text{priv}})_k(X_i-\mu_{\text{priv}})_j(X_i-\mu_{\text{priv}})_k\,|\,\mu_{\text{priv}}\right] \\
=&\sum_j \mathbb{E}\left[(M(X_i')-\mu_{\text{priv}})_j^2(X_i-\mu_{\text{priv}})_j^2\,|\,\mu_{\text{priv}}\right]=\mathbb{E}\left[\|M(X_i')-\mu_{\text{priv}}\|_2^2\,|\,\mu_{\text{priv}}\right]=O(\alpha^2).
\end{aligned}$$

The equality in the second step uses: (i) the conditional independence of $X_i-\mu_{\text{priv}}$ and $M(X_i')-\mu_{\text{priv}}$, given μ_{priv}, and (ii) the independence of $(X_i-\mu_{\text{priv}})_j$ and $(X_i-\mu_{\text{priv}})_j$ when $j\neq k$; and (iii) $\mathbb{E}[X_i-\mu_{\text{priv}}\,|\,\mu_{\text{priv}}]=0$. The third step uses the conditional indpendence and the fact that $\mathbb{E}[(X_i-\mu_{\text{priv}})_j^2]=1$. In the final step, we apply the assumption on the accuracy of the mechanism M. □

Proposition A.3. *Assume the conditions in Theorem A.1. Then, $\mathbb{E}[Z_i]=O(\varepsilon\alpha),\forall i\in[n]$, where the expectation is taken over the randomness of sampling μ_{priv}, X, and the randomness in the mechanism M.*

Proof. Since X and X_i' differ by a single element, and M is (ε,δ) differentially private, we can use the fact that the outputs $M(X)$ and $M(X_i')$ (and therefore, by the postprocessing property of DP, Z_i and Z_i') cannot be too different. In particular, we invoke Proposition 3.2.

$$|\mathbb{E}[Z_i]-\mathbb{E}[Z_i']|\leq 2\varepsilon\cdot\mathbb{E}[|Z_i'|]+2\sqrt{\delta}\cdot\mathbb{E}[Z_i^2+(Z_i')^2]\leq 2(\varepsilon+\sqrt{\delta})\cdot\sqrt{\mathbb{E}[(Z_i')^2]}+2\sqrt{\delta\cdot\mathbb{E}[Z_i^2]}.$$

Next, we apply Proposition A.2.

$$\mathbb{E}[Z_i]\leq O((\varepsilon+\sqrt{\delta})\alpha)+2\sqrt{\delta\cdot\mathbb{E}[Z_i^2]}.$$

Next, we bound $\mathbb{E}[Z_i^2]$ with Cauchy-Schwarz followed by the accuracy of M, and the concentration of X_i around the private mean:

$$\mathbb{E}[Z_i^2]\leq\sqrt{\mathbb{E}[\|M(X)-\mu_{\text{priv}}\|_2^4]}\sqrt{\mathbb{E}[\|X_i-\mu_{\text{priv}}\|_2^4]}=O(\alpha^2 d)$$

Plugging this result into the equation above, we conclude $\mathbb{E}[Z_i]=O(\varepsilon\alpha+\sqrt{\delta d}\alpha)$. Recall from the conditions in Theorem A.1, $\delta\leq\varepsilon^2/d$. Thus, $\mathbb{E}[Z_i]=O(\varepsilon\alpha)$. □

Proposition A.4. *Assume the conditions in Theorem A.1. Then, the following holds for the expected sum of the fingerprinting statistics defined for the public samples: $\mathbb{E}[\sum_{i=n}^{n+m}Z_i]=O(\alpha\sqrt{md})$.*

Proof. Recall the definition of the fingerprinting statistics: $\mathbb{E}[Z_j]=\mathbb{E}[\langle M(x)-\mu_{\text{priv}},x_j-\mu_{\text{priv}}\rangle]$. Then,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=n}^{n+m}Z_j\right]&=\mathbb{E}\left[\langle M(X)-\mu_{\text{priv}},\sum_{j=n}^{n+m}(x_j-\mu_{\text{priv}})\rangle\right] \\
&\leq\sqrt{\mathbb{E}[\|M(X)-\mu_{\text{priv}}\|_2^2]\cdot\sum_{j=n}^{n+m}\mathbb{E}[\|x_j-\mu_{\text{priv}}\|_2^2]} \\
&\leq\alpha\sqrt{\sum_{j=n}^{n+m}\mathbb{E}\left[\mathbb{E}[\|x_j-\mu_{\text{priv}}\|_2^2\,|\,\mu_{\text{priv}}]\right]}=O(\alpha\sqrt{md}).
\end{aligned}$$

The first inequality uses Cauchy-Schwarz, and the second uses the bound on the accuracy of the mechanism M. Finally, the last equality uses the tower law of expectations, followed by the variance of the public data around the public mean μ_{pub}, which matches μ_{priv} in the no distribution shift setting. □

Corollary A.5. *Under the conditions of Theorem A.1, for the expectation taken over the randomness of $\mu_{\text{priv}},X_{\text{priv}}$ and the mechanism M, $\sum_{i=1}^{n+m}\mathbb{E}[Z_i]=O(n\varepsilon\alpha+\alpha\sqrt{md})$.*

Next, we lower bound $\sum_{i=1}^{n+m} Z_i = (n+m)\langle M(X)-\mu_{\text{priv}}, \bar{\mu}-\mu_{\text{priv}}\rangle$, where $\bar{\mu} = {}^1\!/_{n+m}\sum_{i=1}^{n+m} Z_i$, *i.e.*, it is the empirical mean of the public and private samples combined. It is sufficient to lower bound $\langle M(X)-\mu_{\text{priv}}, \bar{\mu}-\mu_{\text{priv}}\rangle$.

Lemma A.6. *For $n+m>0$, we can lower bound $\langle M(X)-\mu_{\text{priv}}, \bar{\mu}-\mu_{\text{priv}}\rangle$, in the following way:*

$$\mathbb{E}\left[\langle M(X)-\mu_{\text{priv}}, \bar{\mu}-\mu_{\text{priv}}\rangle\right] \geq \mathbb{E}\left[\|\mu_{\text{priv}}-\bar{\mu}\|_2^2\right] - \sqrt{\mathbb{E}\left[\|M(X)-\bar{\mu}\|_2^2\right]\mathbb{E}_X\left[\|\mathbb{E}[\mu_{\text{priv}}\,|\,X]-\bar{\mu}\|_2^2\right]}. \tag{9}$$

Proof. First we add and subtract the empirical mean from the first term:

$$\langle M(X)-\mu_{\text{priv}}, \bar{\mu}-\mu_{\text{priv}}\rangle = \langle M(X)-\bar{\mu}, \bar{\mu}-\mu_{\text{priv}}\rangle + \|\bar{\mu}-\mu_{\text{priv}}\|_2^2. \tag{10}$$

Next, we bound the magnitude of $\mathbb{E}[\langle M(X)-\bar{\mu}, \bar{\mu}-\mu_{\text{priv}}\rangle]$, and for this we note that conditioned on the data X, the randomness in the mechanism M is independent of μ_{priv}. This conditional independence trick allows us to break the expectation down as follows:

$$\begin{aligned}
|\mathbb{E}[\langle M(X)-\bar{\mu}, \bar{\mu}-\mu_{\text{priv}}\rangle]| &= \mathbb{E}[\mathbb{E}[\langle M(X)-\bar{\mu}, \bar{\mu}-\mu_{\text{priv}}\rangle]\,|\,X] \\
&= \mathbb{E}[\langle \mathbb{E}[M(X)-\bar{\mu}\,|\,X)], \mathbb{E}[\bar{\mu}-\mu_{\text{priv}}\,|\,X]\rangle] \\
&\leq \sqrt{\mathbb{E}_X[\|\mathbb{E}[M(X)-\bar{\mu}\,|\,X)]\|_2^2]}\cdot\sqrt{\mathbb{E}_X[\|\mathbb{E}[\bar{\mu}-\mu_{\text{priv}}\,|\,X]\|_2^2]},
\end{aligned}$$

where the last inequality uses Cauchy-Schwarz. Applying Jensen inequality on the first term we get the desired lower bound on $\mathbb{E}[\langle M(X)-\bar{\mu}, \bar{\mu}-\mu_{\text{priv}}\rangle]$:

$$\mathbb{E}[\langle M(X)-\bar{\mu}, \bar{\mu}-\mu_{\text{priv}}\rangle] \geq -\sqrt{\mathbb{E}\left[\|M(X)-\bar{\mu}\|_2^2\right]\cdot\mathbb{E}_X\left[\|\mathbb{E}[\mu_{\text{priv}}\,|\,X]-\bar{\mu}\|_2^2\right]}.$$

Plugging the above inequality into *Eq.* 20 completes the proof. □

Lemma A.7. *Consider the setup in Theorem A.1. , Then,* $\mathbb{E}\left[\|\mathbb{E}[\mu_{\text{priv}}\,|\,X]-\bar{\mu}\|_2^2\right] = O\left(\frac{d}{(n+m)^3}\right)$.

Proof. Follows from $gN(\mathbf{0}, \sigma^2\mathbf{I}_d)$ being a conjugate prior over the private mean μ_{priv}, and the fact that $\bar{\mu} \sim \mathcal{N}(\mu_{\text{priv}}, {}^1\!/_{n+m}\cdot\mathbf{I}_d)$. The mean for the posterior distribution over $\mu_{\text{priv}}\,|\,X$ is given by $\bar{\mu}\cdot{}^{(n+m)\sigma^2}\!/_{(n+m)\sigma^2+1}$. Thus,

$$\mathbb{E}\left[\|\mathbb{E}[\mu_{\text{priv}}\,|\,X]-\bar{\mu}\|_2^2\right] = \mathbb{E}\left[\|\bar{\mu}\|_2^2\right]\cdot\frac{1}{((n+m)\sigma^2+1)^2} = O\left(\frac{d}{(n+m)^3}\right).$$

The last equality follows from the expected norm of $\bar{\mu}$ computed over X, μ_{priv}. Note that we can bound $\mathbb{E}\|\bar{\mu}\|_2 = \mathbb{E}_{\mu_{\text{priv}}}[\mathbb{E}_X[\|\bar{\mu}\|_2\,|\,\mu_{\text{priv}}]] = O(\sqrt{{}^d\!/_{(n+m)}})$ since $\mathbb{E}_X[\|\bar{\mu}\|_2\,|\,\mu_{\text{priv}}] = O(\sqrt{{}^d\!/_{(n+m)}})$ Wainwright [2019]. □

Lemma A.8. *Under the conditions of Theorem A.1, where $\mathbb{E}[\|M(X)-\mu_{\text{priv}}\|_2] \leq \alpha$, the number of total data points $n+m = \Omega({}^d\!/_{\alpha^2})$.*

Proof. Recall that for any random vector v with mean μ, $\mathbb{E}\left[\|v\|^2\right] = \mathbb{E}\left[\|v-\mu\|^2\right] + \|\mu\|^2 \geq \mathbb{E}\left[\|v-\mu\|^2\right]$. Conditioned on the dataset X, $M(X)-\mu_{\text{priv}}$ is a random sample from distribution with mean $M(X)-\mathbb{E}[\mu_{\text{priv}}\,|\,X]$. Applying the above inequality on this vector gives us:

$$\mathbb{E}\|\mu_{\text{priv}}-M(X)\|_2^2 \geq \mathbb{E}_X\left[\|\mathbb{E}[\mu_{\text{priv}}\,|\,X]-M(X)\|_2^2\right].$$

Also, note that:

$$\mathbb{E}\left[\|\mu_{\text{priv}}-\bar{\mu}\|_2^2\right] \leq 2\left(\mathbb{E}\left[\|\mathbb{E}[\mu_{\text{priv}}\,|\,X]-\mu_{\text{priv}}\|_2^2\right] + \mathbb{E}\left[\|\mathbb{E}[\mu_{\text{priv}}\,|\,X]-\bar{\mu}\|_2^2\right]\right),$$

which implies that $\mathbb{E}\|\mu_{\text{priv}}-M(X)\|_2^2 \geq \mathbb{E}\left[\|\mu_{\text{priv}}-\bar{\mu}\|_2^2\right] - 2\mathbb{E}\left[\|\mathbb{E}[\mu_{\text{priv}}\,|\,X]-\bar{\mu}\|_2^2\right]$. From Lemma A.7, we know that $\mathbb{E}\left[\|\mathbb{E}[\mu_{\text{priv}}\,|\,X]-\bar{\mu}\|_2^2\right] = O({}^d\!/_{(n+m)^3})$. Since $\mathbb{E}\left[\|\mu_{\text{priv}}-\bar{\mu}\|_2^2\right] = \Omega({}^d\!/_{(n+m)})$ Vershynin [2018], we conclude that: $\mathbb{E}\|\mu_{\text{priv}}-M(X)\|_2^2 = \Omega({}^d\!/_{(n+m)})$. Thus, for $\alpha^2 \geq \mathbb{E}\|\mu_{\text{priv}}-M(X)\|_2^2$, we need $n+m = \Omega({}^d\!/_{\alpha^2})$. □

Corollary A.9. *Under the conditions of Theorem A.1, when $\alpha = O(\sqrt{d})$, for the expectation taken over the randomness of $\mu_{\text{priv}}, X_{\text{priv}}$ and the mechanism M, $\sum_{i=1}^{n+m} \mathbb{E}[Z_i] = \Omega(d)$.*

Proof. We can write $\sum_{i=1}^{n+m} Z_i = (n+m) \cdot \mathbb{E}\big[\langle M(X) - \mu_{\text{priv}}, \bar{\mu} - \mu_{\text{priv}} \rangle\big]$, and from Lemma A.6 we know:

$$\mathbb{E}\big[\langle M(X) - \mu_{\text{priv}}, \bar{\mu} - \mu_{\text{priv}} \rangle\big] \geq \mathbb{E}\big[\|\mu_{\text{priv}} - \bar{\mu}\|_2^2\big] - \sqrt{\mathbb{E}\big[\|M(X) - \bar{\mu}\|_2^2\big] \cdot \mathbb{E}_X\big[\|\mathbb{E}[\mu_{\text{priv}} \mid X] - \bar{\mu}\|_2^2\big]}.$$

Next, we apply the result $\mathbb{E}\big[\|\mu_{\text{priv}} - \bar{\mu}\|_2^2\big] = \Theta(d/m+n)$, purely from known upper and lower bounds for the concentration of the empirical mean, when samples are drawn i.i.d. from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ Wainwright [2019]. Similarly, we can bound $\mathbb{E}\big[\|M(X) - \bar{\mu}\|_2^2\big] = \mathbb{E}\big[\|M(X) - \mu_{\text{priv}}\|_2^2\big] + \mathbb{E}\big[\|\mu_{\text{priv}} - \bar{\mu}\|_2^2\big] = O(\alpha^2 + d/n+m)$. Finally, we apply the Lemma A.7 on the concentration of the posterior mean, around μ_{priv} to get:

$$\mathbb{E}\big[\langle M(X) - \mu_{\text{priv}}, \bar{\mu} - \mu_{\text{priv}} \rangle\big] = \Theta\left(\frac{d}{n+m}\right) - O\left(\sqrt{\left(\alpha^2 + \frac{d}{n+m}\right) \cdot \frac{d}{(n+m)^3}}\right). \tag{11}$$

Now plugging in the condition that $\alpha = O(\sqrt{d})$, we get $\sum_{i=1}^{n+m} \mathbb{E}[Z_i] = \Omega(d)$ □

We are now ready to prove the main result in Theorem A.10. From Corollary A.5 and Corollary A.9, there exists constants $c_1, c_2 > 0, n_0 > 0$, such that for all $n \geq n_0, \delta \leq \varepsilon^2/d$, the following is true:

$$c_1 d \leq \mathbb{E}\left[\sum_{i=1}^{m+n} Z_i\right] \leq c_2\Big(n\varepsilon\alpha + \alpha\sqrt{md}\Big).$$

The above implies that either of the following conditions is true. First, $n\varepsilon\alpha = \Omega(d)$ or $n = \Omega(d/\varepsilon\alpha)$. Second, $\alpha\sqrt{md} = \Omega(d)$ or $m = \Omega(d/\alpha^2)$. Combining this result with Lemma A.8, we conclude that either $m = \Omega(d/\alpha^2)$ or $n + m = \Omega(d/\alpha\varepsilon + d/\alpha^2)$. We can interpret this result as the following: either we need enough public samples to estimate the mean μ_{priv} to accuracy α without the help of any private data, or we must have enough public and private data combined, to be able to treat the public data as private, and estimate the accuracy of μ_{priv} to accuracy α. □

A.2 Lower Bound for Mean Estimation when Public Distribution is Shifted from Private

First, we re-state a formal version of our Gaussian mean estimation result under distribution shift (Theorem 4.2). Then we provide a proof overview, that defines the test-statistic, followed by a presentation of the upper and lower bounds on the expected sum of the test-statistics.

Theorem A.10 (Lower bound on public-assisted private Gaussian mean estimation when public & private data distributions differ)**.** *Fix any $d > 0$, $10\sqrt{d} > \alpha > 0$ and $\tau > 0$. Suppose M is an (ε, δ) private learner, with $\delta \leq \varepsilon^2/d$ that takes as input dataset X and satisfies an expected accuracy of $\mathbb{E}_X \|M(X) - \mu_{\text{priv}}\|_2 \leq \alpha$. Here, the dataset X available to M comprises of n private samples $X_1, \ldots, X_n \sim \mathcal{N}(\mu_{\text{priv}}, \mathbf{I}_d)$ and m public samples $X_{n+1}, \ldots, X_{n+m} \sim \mathcal{N}(\mu_{\text{pub}}, \mathbf{I}_d)$, where the public and private distributions differ, i.e., where $\mu_{\text{pub}} = \mu_{\text{priv}} + v$ and $\|v\|_2 = \tau$. Then, the following is true:*

- *when $\tau = O(\sqrt{d/m})$, then either $m = \Omega(d/\alpha^2)$ or $n + m = \Omega(d/\alpha\varepsilon + d/\alpha^2)$,*
- *and when $\tau = \omega(\sqrt{d/m})$, then either $\alpha \gtrsim \tau$ or $n = \Omega(d/\alpha\varepsilon + d/\alpha^2)$.*

Proof overview. As with any minimax lower bound, we step into the lower bound calculations by constructing a prior over μ_{priv}, v, and then computing the lower bound on sample complexity of n, m in expectation over this prior. The prior we choose over μ_{priv} is similar to the no distribution shift with $\mu_{\text{priv}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, and $v \sim \mathcal{N}(\mathbf{0}, \tau^2/d \cdot \mathbf{I}_d)$. Next, we construct our fingerprinting statistics that measure the correlation between mechanism output $M(X)$ and any private data point X_i. This test-statistic cannot be too high, else we violate the differential privacy guarantee, and cannot be too low, else accuracy cannot be guaranteed, since if it is too low, it means that the mechanism affords very low sensitivity on any given data point.

Remark on how distribution shift affects the design of test-statistic. As the distribution shift gets more severe, i.e., as τ increases, we expect public data to be less useful for any accurate mechanism M. Thus, we expect the correlation between $M(X)$, and any public data point X_j to get lower, i.e., more generally for any test-statistic that measures the presence of a private data point X_i in the dataset X vs. not, the sensitivity of the test-statistic should reduce as τ increases. Motivated by this, we define the following test-statistic Z_i for the Gaussian mean estimation with distribution shift setting:

$$\begin{aligned} Z_i &= \langle M(X)-\mu_{\text{priv}}, X_i-\mu_{\text{priv}}\rangle, \quad i\in\{1,\ldots,n\} \ \ (\text{private } X_i) \\ Z_i &= \frac{1}{m\tau^2/d+1}\cdot\langle M(X)-\mu_{\text{priv}}, X_i-\mu_{\text{priv}}\rangle, \quad i\in\{n+1,\ldots,n+m\} \ \ (\text{public } X_i) \end{aligned} \tag{12}$$

Similar to our proof of the lower bound in the identical distribution setting, we upper and lower bound the expected sum of the fingerprinting statistics above. We begin with the upper bound, on $\mathbb{E}[Z_i]$ in Eq. 12, and we bound the sum of this statistic separately for the public and private samples in Proposition A.11.

Proposition A.11. *Assume the conditions in Theorem A.10. The expected sum of the fingerprinting statistics corresponding to private samples satisfies:* $\mathbb{E}[\sum_{i=1}^{n} Z_i] = O(n\varepsilon\alpha)$*, and the one corresponding to public samples satisfies:* $\sum_{i=n+1}^{n+m}\mathbb{E}[Z_i] = O(\alpha\sqrt{md}+\alpha\tau\sqrt{md})$.

Proof. The proof for the upper bound over $\mathbb{E}[\sum_{i=1}^{n} Z_i]$ is similar to Proposition A.3, which uses the DP constraint to show $\mathbb{E}[\sum_{i=1}^{n} Z_i] = O(n\varepsilon\alpha)$. For the second term which sums the fingerprinting statistics corresponding to public samples, $\mathbb{E}[\sum_{i=n+1}^{n+m} Z_i]$ we need to additionally account for the randomness in the direction of shift v, when computing the expectation over the public mean $\bar{\mu}_{\text{pub}}$. Recall the definition of the fingerprinting statistics: $Z_j = \langle M(X)-\mu_{\text{priv}}, x_j-\mu_{\text{priv}}\rangle$. Then,

$$\begin{aligned} \mathbb{E}\left[\sum_{j=n+1}^{n+m} Z_j\right] &= \mathbb{E}\left[\langle M(X)-\mu_{\text{priv}}, \sum_{j=n+1}^{n+m}(x_j-\mu_{\text{priv}})\rangle\right] \\ &\le \sqrt{\mathbb{E}[\|M(X)-\mu_{\text{priv}}\|_2^2]}\sqrt{\sum_{j=n+1}^{n+m}\mathbb{E}[\|x_j-\mu_{\text{priv}}\|_2^2]} \end{aligned}$$

$$\begin{aligned} \mathbb{E}\left[\sum_{j=n+1}^{n+m} Z_j\right] &\le \alpha\sqrt{\sum_{j=n+1}^{n+m}\mathbb{E}\Big[\mathbb{E}[\|x_j-\mu_{\text{priv}}\|_2^2\,|\,\mu_{\text{priv}}]\Big]} \\ &= \alpha\sqrt{\sum_{j=n+1}^{n+m}\mathbb{E}[\mathbb{E}[\|x_j-\mu_{\text{priv}}-v\,|\,\mu_{\text{priv}}\|_2^2]]+\mathbb{E}[\|v\|_2^2]} = O(\alpha\sqrt{md}+\alpha\tau\sqrt{md}). \end{aligned}$$

The first inequality uses Cauchy-Schwarz, and the second uses the bound on the accuracy of the mechanism M. Finally, the last equality uses the tower law of expectations, followed by the independence between displacement v and private mean μ_{priv}. Then, we simply apply the definitions for the priors over public and private means to get the final result. □

Corollary A.12. *Under the conditions of Theorem A.10, over the randomness of* μ_{priv}, X, v *and mechanism* M, $\sum_{i=1}^{n}\mathbb{E}[Z_i] + \frac{1}{m/d\cdot\tau^2+1}\cdot\sum_{i=n+1}^{n+m}\mathbb{E}[Z_i] = O\Big(n\varepsilon\alpha + \frac{1}{m/d\cdot\tau^2+1}\cdot\Big(\alpha\sqrt{md}+\alpha\tau\sqrt{md}\Big)\Big)$.

Proof. Follows immediately from Proposition A.11 that upper bounds the second term in the summation and Proposition A.3 which upper bounds the first term. □

Next, we lower bound the expected sum over the fingerprinting statistics, for the expression:

$$\sum_{i=1}^{n}\mathbb{E}[Z_i] + \frac{1}{m/d\cdot\tau^2+1}\cdot\sum_{i=n+1}^{n+m}\mathbb{E}[Z_i].$$

Our approach for the lower bound is similar to the case where there is no shift, but we explicitly account for the distribution shift term v in our posterior calculations. Recall that our prior for v is $v \sim \mathcal{N}(0, \frac{\tau^2}{d} \cdot I_d)$. We will see that when $\tau = O(\sqrt{d/m})$, the lower bound on the above expression scales as the one where $\tau = 0$, i.e., when there is no shift.

Lemma A.13. *Consider the setup in Theorem A.10. Then,* $\mathbb{E}[\mu_{\text{priv}} \mid X] = w_{\text{priv}} \cdot \bar{\mu}_{\text{priv}} + w_{\text{pub}} \cdot \bar{\mu}_{\text{pub}}$, *where* $w_{\text{priv}} = \frac{\sigma^2(\tau^2/d + 1/m)}{\sigma^2(\tau^2/d + 1/m + 1/n) + \tau^2/dn + 1/mn}$ *and* $w_{\text{pub}} = \frac{\sigma^2/n}{\sigma^2(\tau^2/d + 1/m + 1/n) + \tau^2/nd + 1/mn}$.

Proof. Let $\bar{\mu}_{\text{pub}}$ be the empirical mean of the public data and $\bar{\mu}_{\text{priv}}$ be the empirical mean of the private data points. We can then derive the following joint distribution over μ_{priv}, $\bar{\mu}_{\text{pub}}, \bar{\mu}_{\text{priv}}$:

$$(\mu_{\text{priv}}, \bar{\mu}_{\text{pub}}, \bar{\mu}_{\text{priv}}) \sim \mathcal{N}(0, \Sigma), \Sigma \in \mathbb{R}^{3d \times 3d} \text{ and,} \tag{13}$$
$$\Sigma = \begin{pmatrix} \sigma^2 \mathbf{I}_d & \sigma^2 \mathbf{I}_d & \sigma^2 \mathbf{I}_d \\ \sigma^2 \mathbf{I}_d & (\sigma^2 + \tau^2/d + 1/m) \cdot \mathbf{I}_d & \sigma^2 \mathbf{I}_d \\ \sigma^2 \mathbf{I}_d & \sigma^2 \mathbf{I}_d & (\sigma^2 + 1/n) \cdot \mathbf{I}_d, \end{pmatrix}.$$

and $\mathbb{E}[\mu_{\text{priv}} \mid \bar{\mu}_{\text{pub}}, \bar{\mu}_{\text{priv}}]$ is given by $\Sigma_{12}\Sigma_{22}^{-1}\left([\bar{\mu}_{\text{pub}}, \bar{\mu}_{\text{priv}}]^\top - [\mathbb{E}[\mu_{\text{priv}}], \mathbb{E}[\mu_{\text{priv}}]]^\top\right)$, where Σ_{12}, and Σ_{22} are given by the top $d \times 2d$, and bottom $2d \times 2d$ submatrices of Σ. Based, on this, we get:

$$\Sigma_{12}\Sigma_{22}^{-1} = \left(\frac{\sigma^2/n}{(\sigma^2 + \tau^2/d + 1/m)(\sigma^2 + 1/n) - \sigma^4} I_d, \frac{\sigma^2(\tau^2/d + 1/m)}{(\sigma^2 + \tau^2/d + 1/m)(\sigma^2 + 1/n) - \sigma^4} I_d \right)$$

Then, the mean of the posterior distribution over μ_{priv} is given by:

$$\begin{aligned} \mathbb{E}[\mu_{\text{priv}} \mid \bar{\mu}_{\text{pub}}, \bar{\mu}_{\text{priv}}] &= \left(\frac{\sigma^2/n}{(\sigma^2 + \tau^2/d + 1/m)(\sigma^2 + 1/n) - \sigma^4} I_d \right) \cdot (\bar{\mu}_{\text{pub}} - \mathbb{E}[\mu_{\text{pub}}]) \\ &\quad + \left(\frac{\sigma^2(\tau^2/d + 1/m)}{(\sigma^2 + \tau^2/d + 1/m)(\sigma^2 + 1/n) - \sigma^4} I_d \right) \cdot (\bar{\mu}_{\text{priv}} - \mathbb{E}[\mu_{\text{priv}}]) \\ &= \frac{\sigma^2/n \cdot \bar{\mu}_{\text{pub}} + \sigma^2 \bar{\mu}_{\text{priv}} \cdot (\tau^2/d + 1/m)}{\sigma^2(\tau^2/d + 1/m + 1/n) + \tau^2/nd + 1/mn} \end{aligned}$$

From the expression above it it easy to derive w_{pub} and w_{priv} in the statement of Lemma A.13. □

Lemma A.14. *Under the conditions in Theorem A.10,*

$$\mathbb{E}_X \|\mathbb{E}[\mu_{\text{priv}} \mid X] - \bar{X}\|_2^2 = O\left(\frac{d}{(m/\kappa + n)^3} \right), \quad \text{where } \kappa \overset{def}{=} \frac{m\tau^2}{d} + 1,$$

and the inner expectation is taken over $\mu_{\text{priv}}, v \mid X$, *and* $\bar{X}$ *is defined as:*

$$\bar{X} \overset{def}{=} \bar{X}_{\text{pub}} \cdot \frac{m/(m\tau^2/d + 1)}{n + m/(m\tau^2/d + 1)} + \bar{X}_{\text{priv}} \cdot \frac{n}{n + m/(m\tau^2/d + 1)}, \tag{14}$$

with $\bar{X}_{\text{pub}}$ *and* $\bar{X}_{\text{priv}}$ *being the empirical mean over public and private data points.*

Proof. Directly applying Lemma A.13 we get:

$$\begin{aligned} &\mathbb{E}[\mu_{\text{priv}} \mid X] - \bar{X} \\ &= \frac{\sigma^2/n \cdot \bar{X}_{\text{pub}} + \sigma^2 \bar{X}_{\text{priv}} \cdot (\tau^2/d + 1/m)}{\sigma^2(\tau^2/d + 1/m + 1/n) + \tau^2/nd + 1/mn} - \frac{n}{m/(m\tau^2/d+1) + n} \cdot \bar{X}_{\text{priv}} - \frac{m/(m\tau^2/d+1)}{m/(m\tau^2/d+1) + n} \cdot \bar{X}_{\text{pub}} \qquad (15) \\ &= \bar{X}_{\text{pub}} \left(\frac{\sigma^2/n}{\sigma^2(\tau^2/d + 1/m + 1/n) + \tau^2/nd + 1/mn} - \frac{m/(m\tau^2/d+1)}{m/(m\tau^2/d+1) + n} \right) \\ &\quad + \bar{X}_{\text{priv}} \left(\frac{\sigma^2(\tau^2/d + 1/m)}{\sigma^2(\tau^2/d + 1/m + 1/n) + \tau^2/nd + 1/mn} - \frac{n}{m/(m\tau^2/d+1) + n} \right). \qquad (16) \end{aligned}$$

To bound the expected squared norm of the above term, we first need to bound terms α,β defined as follows:

$$\alpha \overset{\text{def}}{=} \frac{\sigma^2/n}{\sigma^2(\tau^2/d+1/m+1/n)+\tau^2/nd+1/mn} - \frac{m/(m\tau^2/d+1)}{m/(m\tau^2/d+1)+n}$$

$$\beta \overset{\text{def}}{=} \frac{\sigma^2(\tau^2/d+1/m)}{\sigma^2(\tau^2/d+1/m+1/n)+\tau^2/nd+1/mn} - \frac{n}{m/(m\tau^2/d+1)+n}$$

Let $\sigma^2=1$ and define $\kappa := 1+\frac{m\tau^2}{d}$. Then we can simplify in the following way:

$$\alpha = \frac{-m\kappa}{(n\kappa+\kappa+m)(n\kappa+m)}, \quad \beta = \frac{-n\kappa^2}{(n\kappa+\kappa+m)(n\kappa+m)} \tag{17}$$

Thus, $\alpha\bar{X}_{\text{pub}}+\beta\bar{X}_{\text{priv}}$ can be written as:

$$\alpha\bar{X}_{\text{pub}}+\beta\bar{X}_{\text{priv}} = -\frac{\left(\kappa^2\sum_{i=1}^{n}X_i+\kappa\sum_{j=n+m}^{n+1}X_j\right)}{(m+n\kappa+\kappa)(m+n\kappa)} \tag{18}$$

We can bound the $\|\cdot\|_2^2$ for the expression, by noting that $\|\sum_{i=1}^{n}X_i\|_2^2 = O(nd)$ and $\left\|\sum_{i=m+1}^{n+m}X_i\right\|_2^2 = O(md)$. Applying this in the above expression, and dividing both numerator and denominator by κ^2 we get the final result:

$$\begin{aligned}\left\|\mathbb{E}[\mu_{\text{priv}}\,|\,X]-\bar{X}\right\|_2^2 &= O\left(\frac{d(n+m/\kappa)}{(m/\kappa+n+1)^2(m/\kappa+n)^2}\right)\\ &= O\left(\frac{d}{(n+m/\kappa)^3}\right)\end{aligned} \tag{19}$$

This completes the proof of Lemma A.14. □

Lemma A.15. *Under the conditions in Theorem A.10, the expected sum of the fingerprinting statistics is lower bounded as follows:*

$$\mathbb{E}\left[\sum_i Z_i\right] = \Omega\left((n+m/\kappa)\cdot\left(\frac{d}{n+m/\kappa}-\sqrt{\left(\alpha^2+\frac{d}{n+m/\kappa}\right)\cdot\left(\frac{d}{(n+m/\kappa)^3}\right)}\right)\right),$$

where $\kappa=(m\tau^2/d+1)$ and the expectation is taken over the randomness of μ_{priv}, X, and the mechanism M.

Proof. Similar to our previous proofs on the lower bound for the expected sum of test statistics, we manipulate the following sum into two terms.

$$\begin{aligned}&\mathbb{E}_{\mu_{\text{priv}},X,v}\left[\sum_{i=1}^{n}Z_i+\frac{1}{m/d\cdot\tau^2+1}\cdot\sum_{i=n+1}^{n+m}Z_i\right]\\ &=(n+m/(m\tau^2/d+1))\cdot\left(\mathbb{E}\left[\|\bar{X}-\mu_{\text{priv}}\|_2^2\right]+\mathbb{E}\left\langle M(X)-\bar{X},\bar{X}-\mu_{\text{priv}}\right\rangle\right),\end{aligned} \tag{20}$$

Given, X, the random variable $M(X)$ is independent of μ_{priv}.

$$\begin{aligned}\mathbb{E}_{\mu_{\text{priv}},X,v}\left[\left\langle M(X)-\bar{X},\bar{X}-\mu_{\text{priv}}\right\rangle\right] &= \mathbb{E}_X\left[\left\langle\mathbb{E}_{\mu_{\text{priv}},v|X}[M(X)-\bar{X}],\mathbb{E}_{\mu_{\text{priv}},v|X}[\bar{X}-\mu_{\text{priv}}]\right\rangle\right]\\ &\le \sqrt{\mathbb{E}_X\|\mathbb{E}_{\mu_{\text{priv}},v|X}[M(X)-\bar{X}]\|_2^2}\cdot\sqrt{\mathbb{E}_X\|\mathbb{E}_{\mu_{\text{priv}},v|X}[\bar{X}-\mu_{\text{priv}}]\|_2^2},\end{aligned} \tag{21}$$

where the inequality uses Cauchy-Schwarz. We can further use the convexity of $\|\cdot\|_2^2$ in the first term above to arrive at the following upper bound.

$$\begin{aligned}&\mathbb{E}_{\mu_{\text{priv}},X,v}\left[\left\langle M(X)-\bar{X},\bar{X}-\mu_{\text{priv}}\right\rangle\right]\\&\leq\sqrt{\mathbb{E}_{\mu_{\text{priv}},v,X}\|[M(X)-\bar{X}]\|_2^2}\cdot\sqrt{\mathbb{E}_X\|\mathbb{E}_{\mu_{\text{priv}},v|X}[\mu_{\text{priv}}]-\bar{X}\|_2^2},\end{aligned}\tag{22}$$

Next, we proceed by bounding $\mathbb{E}_{\mu_{\text{priv}},v,X}\|[M(X)-\bar{X}]\|_2^2$.

$$\begin{aligned}\mathbb{E}_{\mu_{\text{priv}},v,X}\|[M(X)-\bar{X}]\|_2^2&=\mathbb{E}_{\mu_{\text{priv}},v,X}\|\mu_{\text{priv}}-\bar{X}]\|_2^2+\mathbb{E}_{\mu_{\text{priv}},v,X}\|[M(X)-\mu_{\text{priv}}]\|_2^2\\&\leq\mathbb{E}_{\mu_{\text{priv}},v,X}\|\mu_{\text{priv}}-\bar{X}\|_2^2+\alpha^2\end{aligned}\tag{23}$$

Recall the definition of $\bar{X}$:

$$\bar{X}\stackrel{\text{def}}{=}\bar{X}_{\text{pub}}\cdot\frac{m/(m\tau^2/d+1)}{n+m/(m\tau^2/d+1)}+\bar{X}_{\text{priv}}\cdot\frac{n}{n+m/(m\tau^2/d+1)},\tag{24}$$

with $\bar{X}_{\text{pub}}$ and $\bar{X}_{\text{priv}}$ being the empirical mean over public and private data points. Plugging this into , we get:

$$\begin{aligned}\mathbb{E}_{\mu_{\text{priv}},v,X}\|[M(X)-\bar{X}]\|_2^2&\leq\alpha^2+\frac{md/\kappa^2}{(n+m/\kappa)^2}+\frac{nd}{(n+m/\kappa)^2}\\&\leq\alpha^2+\frac{md/\kappa^2}{(n+m/\kappa)^2}+\frac{nd}{(n+m/\kappa)^2}\\&=O\left(\alpha^2+\frac{d}{(n+m/\kappa)}\right),\end{aligned}\tag{25}$$

where the second inequality follows from $\kappa\geq 1$. Finally, we do the following plugins: we plug in,

1. the result from Lemma A.14 for the bound on $\mathbb{E}_X\|\mathbb{E}[\mu_{\text{priv}}\,|\,X]-\bar{X}\|_2^2$ into Eq. 22,
2. the result on the bound on $\mathbb{E}\left[\|M(X)-\bar{X}\|_2^2\right]$ in Eq. 25 into Eq. 22,
3. the bound on $\mathbb{E}\left[\|\bar{X}-\mu_{\text{priv}}\|_2^2\right]$ derived as part of Eq. 25 into Eq. 20

Together, these complete the proof of the lower bound stated in Lemma A.15.

Finally, we are ready to put together the upper and lower bounds on $\mathbb{E}[\sum_i Z_i]$ to arrive at the final result in Theorem A.10. From Corollary A.12 and Lemma A.15 we conclude that whenever $\tau=O(\sqrt{d/m})$, there exists positive constants c_1,c_2 and values n_0,m_0, such that for $n\geq n_0,m\geq m_0$:

$$c_2\cdot d\leq\sum_{i=1}^{n}\mathbb{E}[Z_i]+\frac{1}{\kappa}\cdot\sum_{i=n+1}^{n+m}\mathbb{E}[Z_i]\leq c_1\left(n\varepsilon\alpha+\alpha\sqrt{md}\right)$$

This means that when either $n=\Omega(d/\varepsilon\alpha)$ or $m=\Omega(d^2/\alpha)$. Note that the lower bound in Lemma A.8 is still applicable for n. Together, this tells us that when $\tau=O(\sqrt{d/m})$, then either $n=\Omega(d/\varepsilon\alpha+d/\alpha^2)$ or $m=\Omega(d/\alpha^2)$. This completes the first result in Theorem A.10. For the second part, let us consider the case where $\tau=\omega(\sqrt{d/m})$. Then, either there exists constants c_1,c_2, such that for all $n\geq n_0$:

$$c_1 d\leq c_2 n\varepsilon\alpha,$$

in which case $n=\Omega(d/\varepsilon\alpha)$ or there exists constants c_1,c_2 and values n_0,m_0, such that for all $n\geq n_0,m\geq m_0$ the following holds:

$$d\leq\frac{\alpha\tau\sqrt{md}+\alpha\sqrt{md}}{1+m\tau^2/d},$$

which implies:

$$c_1 d\leq\alpha\tau\sqrt{md}/m\tau^2/d.$$

The above holds when $\alpha\gtrsim\tau$. This completes all conditions in the proof of Theorem A.10. □

B USEFUL LEMMAS FOR LINEAR REGRESSION

We reuse some eigenvalue tail bounds throughout the linear regression lower bounds.

Lemma B.1 (Wainwright [2019], Theorem 6.1)**.** *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be drawn according to the Σ-Gaussian ensemble. Then, for $N \geq d$, the minimum singular value $\sigma_{\min}(\mathbf{X})$ satisfies the lower deviation inequality*

$$\mathbb{P}\left[\frac{\sigma_{\min}(\mathbf{X})}{\sqrt{N}} \leq \lambda_{\min}(\sqrt{\Sigma})(1-\delta) - \sqrt{\frac{\mathrm{tr}(\Sigma)}{N}}\right] \leq e^{-N\delta^2/2}.$$

Specializing to the case of $\Sigma = I_d$, we have

Corollary B.2.

$$\mathbb{P}\left[\frac{\sigma_{\min}(\mathbf{X})}{\sqrt{N}} \leq 1 - \sqrt{\frac{d}{N}} - \delta\right] \leq e^{-N\delta^2/2}.$$

and thus,

$$\mathbb{P}\left[\sigma_{\min}(\mathbf{X})^2 = \lambda_{\min}(X^\top X) \leq N\left(1 - \sqrt{\frac{d}{N}} - \delta\right)^2\right] \leq e^{-N\delta^2/2}.$$

We also have the following upper tail variant:

Lemma B.3 (Wainwright [2019], Theorem 6.1)**.** *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be drawn according to the Σ-Gaussian ensemble. Then, for $N \geq d$, the maximum singular value $\sigma_{\max}(\mathbf{X})$ satisfies the upper deviation inequality*

$$\mathbb{P}\left[\frac{\sigma_{\max}(\mathbf{X})}{\sqrt{N}} \geq \lambda_{\max}(\sqrt{\Sigma})(1+\delta) + \sqrt{\frac{\mathrm{tr}(\Sigma)}{N}}\right] \leq e^{-N\delta^2/2}.$$

Specializing to the case of $\Sigma = I_d$, we have

Corollary B.4.

$$\mathbb{P}\left[\frac{\sigma_{\max}(\mathbf{X})}{\sqrt{N}} \geq 1 + \sqrt{\frac{d}{N}} + \delta\right] \leq e^{-N\delta^2/2}.$$

similarly,

$$\mathbb{P}\left[\sigma_{\max}(\mathbf{X})^2 = \lambda_{\max}(X^\top X) \geq N\left(1 + \sqrt{\frac{d}{N}} + \delta\right)^2\right] \leq e^{-N\delta^2/2}.$$

The following simple lemma will be useful in converting exponential tail bounds to upper bounds involving N.

Fact B.5. *For $N \geq 400$ the following is true:*

$$e^{-N/8} \leq 1/N^8. \tag{26}$$

The following calculations are also useful in several parts of the proof, so we black-box them here.

Lemma B.6. *Let $z > 100d$ (thus $0 < d < z/100$), $\psi = 1/10$ and consider $t = z(1 - \sqrt{d/z} - \psi)^2$. Then $t \geq 63/100 \cdot z$.*

Proof. We have

$$z(1-\sqrt{d/z}-\psi)^2 = z(9/10-\sqrt{d/z})^2 \tag{27}$$
$$= z(81/100-(18/10)\sqrt{d/z}+d/z) \tag{28}$$
$$= z(81/100)-(18/10)\sqrt{dz}+d \tag{29}$$
$$\geq z(81/100)-(18/10)\sqrt{z^2/100}+d \tag{30}$$
$$\geq z(81-18)/100 = 63/100\cdot z. \tag{31}$$

□

Lemma B.7. *Let $z > 100d$ (thus $0 < d < z/100$), $\psi = 1/10$ and consider $t = z(1+\sqrt{d/z}+\psi)^2$. Then $t \leq 64/100\cdot z$.*

Proof. We have

$$z(1+\sqrt{d/z}+\psi)^2 = z(9/10+\sqrt{d/z})^2 \tag{32}$$
$$= z(81/100+(18/10)\sqrt{d/z}+d/z) \tag{33}$$
$$= z(81/100)+(18/10)\sqrt{dz}+d \tag{34}$$
$$\leq z(81/100)+(18/10)\sqrt{z^2/100}+z/100 \tag{35}$$
$$\leq z(81-18+1)/100 = 64/100\cdot z. \tag{36}$$

□

C PROOFS FOR LINEAR REGRESSION

C.1 Upper bound

We first upper bound $\mathbb{E}[\sum_i Z_i]$. We can split this into a bound over the sum of public samples and private samples. We will bound the private samples using the differential privacy constraint, and the public samples using concentration.

For the private samples, we can use Proposition 3.2 as in the mean estimation case.

Proposition C.1. *Assume the conditions in Theorem 5.1. Then, $\mathbb{E}[Z_i'] = 0$ and $\mathbb{E}[(Z_i')^2] = O(\alpha^2)$, where the expectation is taken over the randomness of sampling β, $\mathcal{D}$, and the randomness in the mechanism M.*

Proof. When conditioned on β, $(y_i - x_i^\top \beta)x_i$ is independent of $M(\mathcal{D}_i') - \beta$, and $\mathbb{E}[(y_i - x_i^\top \beta)x_i \,|\, \beta] = 0$ (because $\mathbb{E}[\eta_i x_i] = 0$). Thus, we have $\mathbb{E}[Z_i'] = \mathbb{E}[\mathbb{E}[Z_i' \,|\, \beta]] = 0$.

Next, we bound $\mathbb{E}[(Z_i')^2 | \beta]$ and similarly use the law of total expectation to get the upper bound on $\mathbb{E}[(Z_i')^2]$. Note that we have $y_i = x_i^\top \beta + \eta_i$, thus

$$\mathbb{E}[(Z_i')^2|\beta] = \mathbb{E}\left[\sum_{j,k}(M(\mathcal{D}_i')-\beta)_j(M(\mathcal{D}_i')-\beta)_k(\eta_i x_i)_j(\eta_i x_i)_k|\beta\right]$$
$$= \sum_j \mathbb{E}\left[(M(\mathcal{D}_i')-\beta)_j^2(\eta_i x_i)_j^2|\beta\right]$$
$$= \sum_j \mathbb{E}\left[(M(\mathcal{D}_i')-\beta)_j^2|\beta\right]\mathbb{E}\left[(\eta_i x_i)_j^2|\beta\right]$$
$$= \mathbb{E}\left[\left\|M(\mathcal{D}_i')-\beta)_j^2\right\|_2^2|\beta\right] \leq \alpha^2.$$

The second equality uses the fact the coordinates of η and x_i are sampled independently, and the third equality uses that $M(\mathcal{D}_i') - \beta$ is conditionally independent of $\eta_i x_i$ given β (because x_i and y_i are resampled in $\mathcal{D}_i'$). The final step uses the accuracy of the mechanism M and the variance of η_i and x_i. □

Proposition C.2. *Assume the conditions in Theorem 5.1. Then, for $i \in [n]$, $\mathbb{E}[Z_i] = O(\varepsilon\alpha)$, where the expectation is taken over the randomness of sampling β, $\mathcal{D}$, and the randomness in the mechanism M.*

Proof. Since $\mathcal{D}$ and $\mathcal{D}'_i$ differ by a single element, and M is (ε,δ) differentially private, we can use the fact that the outputs $M(\mathcal{D})$ and $M(\mathcal{D}'_i)$ (and therefore, by the postprocessing property of DP, Z_i and Z'_i) cannot be too different. In particular, we invoke Proposition 3.2.

$$|\mathbb{E}[Z_i]-\mathbb{E}[Z'_i]| \le 2\varepsilon\cdot\mathbb{E}[|Z'_i|]+2\sqrt{\delta}\cdot\mathbb{E}[Z_i^2+(Z'_i)^2] \le 2(\varepsilon+\sqrt{\delta})\cdot\sqrt{\mathbb{E}[(Z'_i)^2]}+2\sqrt{\delta\cdot\mathbb{E}[Z_i^2]}.$$

Next, we apply Proposition C.1.

$$\mathbb{E}[Z_i] \le O((\varepsilon+\sqrt{\delta})\alpha)+2\sqrt{\delta\cdot\mathbb{E}[Z_i^2]}.$$

Next, we bound $\mathbb{E}[Z_i^2]$ with Cauchy-Schwarz followed by the accuracy of M, and the concentration of $\eta_i x_i$:

$$\begin{aligned}\mathbb{E}[Z_i^2] &\le \sqrt{\mathbb{E}[\|M(\mathcal{D})-\mu_{\text{priv}}\|_2^4]}\sqrt{\mathbb{E}[\|\eta_i x_i\|_2^4]} \\ &= O(\alpha^2 d)\end{aligned}$$

Plugging this result into the equation above, we conclude $\mathbb{E}[Z_i] = O(\varepsilon\alpha+\sqrt{\delta d}\alpha)$. Recall from the conditions in Theorem 5.1, $\delta \le \varepsilon^2/d$. Thus, $\mathbb{E}[Z_i] = O(\varepsilon\alpha)$. □

Finally, we bound the value of Z_i for the public samples.

Proposition C.3. *Assume the conditions in Theorem A.1. Then, the following holds for the expected sum of the fingerprinting statistics defined for the public samples: $\mathbb{E}[\sum_{i=n}^{n+m} Z_i] = O(\alpha\sqrt{md})$.*

Proof. Recall the definition of the fingerprinting statistics: $\mathbb{E}[Z_j] = \mathbb{E}\left[\langle M(x)-\beta, (y_i - x_i^\top\beta)x_i\rangle\right]$. Then,

$$\mathbb{E}\left[\sum_{i=n}^{n+m} Z_i\right] = \mathbb{E}\left[\langle M(\mathcal{D})-\beta_{\text{priv}}, \sum_{i=n}^{n+m}(y_i - x_i^\top\beta)x_i\rangle\right] \tag{37}$$

$$\le \sqrt{\mathbb{E}[\|M(\mathcal{D})-\beta\|_2^2]}\sqrt{\sum_{i=n}^{n+m}\mathbb{E}[\|\eta_i x_i\|_2^2]} \tag{38}$$

$$= O(\alpha\sqrt{md}). \tag{39}$$

The first inequality uses Cauchy-Schwarz, and the second uses the bound on the accuracy of the mechanism M and the variance of $\eta_i x_i$. □

Corollary C.4. *Under the conditions of Theorem 5.1, for the expectation taken over the randomness of $\beta, X_{\text{priv}}, X_{\text{pub}}, \eta$ and the mechanism M, $\sum_{i=1}^{n+m}\mathbb{E}[Z_i] = O(n\varepsilon\alpha+\alpha\sqrt{md})$.*

C.2 Fingerprinting lemma (lower bound)

Theorem C.5. *Assume the conditions of Theorem 5.1. Then*

$$\mathbb{E}\left[\sum_i Z_i\right] = \Omega(d) \tag{40}$$

where the sum is taken over both the public and private samples.

To prove this theorem, we follow the same structure as in the mean estimation proof (conditioning on X, y instead of just X).

Let $\hat{\beta} \stackrel{\text{def}}{=} (X^T X + \lambda I_d)^{-1} X^T y$ for some $\lambda > 0$. Then, we again have the breakdown:

$$\mathbb{E}\left[\sum_i Z_i\right] = \mathbb{E}\left[\langle M(X,y) - \beta, X^T y - X^T X \beta \rangle\right] \tag{41}$$

$$= \mathbb{E}\left[\langle M(X,y) - \hat{\beta} + \hat{\beta} - \beta, X^T X(\hat{\beta} - \beta) \rangle\right] \tag{42}$$

$$= \mathbb{E}\left[\|X\beta - X\hat{\beta}\|_2^2\right] - \mathbb{E}\left[\langle M(X,y) - \hat{\beta}, X^T X(\beta - \hat{\beta}) \rangle\right] \tag{43}$$

$$= \mathbb{E}\left[\left\|X(\hat{\beta} - \beta)\right\|_2^2\right] - \mathbb{E}\left[\langle M(X,y) - \hat{\beta}, X^T (X\beta) - X^\top y \rangle\right] \tag{44}$$

The first term corresponds to the expected error of the nonprivate OLS estimate. We include a short proof for completeness.

Lemma C.6. *Under the assumptions of Theorem 5.1,*

$$\mathbb{E}_{\beta, X, \eta}\left[\left\|X(\hat{\beta} - \beta)\right\|_2^2\right] = \sigma^2 d$$

Proof. Rewrite $\hat{\beta} = (X^\top X)^{-1} X^\top y = (X^\top X)^{-1} X^\top (X\beta + \eta) = \beta + (X^\top X)^{-1} X^\top \eta$.

Then $\hat{\beta} - \beta = (X^\top X)^{-1} X^\top \eta$. Hence, $X(\hat{\beta} - \beta) = X(X^\top X)^{-1} X^\top \eta$.

Finally

$$\mathbb{E}\left[\left\|X(X^\top X)^{-1} X^\top \eta\right\|_2^2\right] = \mathbb{E}\left[\mathrm{Tr}(\eta^T X (X^\top X)^{-1} X^\top \eta)\right] \tag{45}$$

$$= \sigma^2 d \tag{46}$$

□

We now proceed similarly to the mean estimation proof in order to upper bound the second term. Conditioning on X, η we have:

$$\mathbb{E}_{M,\beta}\left[\langle M(\mathcal{D}) - \hat{\beta}, X^\top X\beta - X^\top y \rangle \,|\, X, y\right] \tag{47}$$

$$= \langle \mathbb{E}_M\left[M(\mathcal{D}) - \hat{\beta} \,|\, X, y\right], \mathbb{E}_\beta\left[X^\top X\beta - X^\top y \,|\, X, y\right] \rangle \tag{48}$$

$$= \left\|\mathbb{E}_M\left[M(\mathcal{D}) - \hat{\beta} \,|\, X, y\right]\right\|_2 \cdot \left\|\mathbb{E}_\beta\left[X^\top X\beta - X^\top y \,|\, X, y\right]\right\|_2 \tag{49}$$

$$\leq \sqrt{\mathbb{E}_M\left[\left\|M(\mathcal{D}) - \hat{\beta}\right\|_2^2 \,|\, X, y\right]} \cdot \left\|\mathbb{E}_\beta\left[X^\top X\beta - X^\top y \,|\, X, y\right]\right\|_2 \tag{50}$$

Removing the conditioning on X, y and applying Cauchy-Schwarz, we have

$$\underbrace{\sqrt{\mathbb{E}_{X,y,M}\left\|M(X,y) - \hat{\beta}\right\|_2^2}}_{①} \cdot \underbrace{\mathbb{E}_{X,y}\|X^\top X \mathbb{E}_\beta[\beta | X, y] - X^\top y\|_2^2}_{②} \tag{51}$$

To bound term (1), we use the triangle inequality as before:

$$\mathbb{E}_{X,\eta}\mathbb{E}_{\hat{\beta}}\left\|M(X,y)-\hat{\beta}\right\|_2^2=\mathbb{E}_{X,\eta}\mathbb{E}_{\hat{\beta}}\left\|M(X,y)-\beta_{\text{priv}}+\beta_{\text{priv}}-\hat{\beta}\right\|_2^2 \tag{52}$$

$$\leq\alpha^2+\mathbb{E}_{X,\eta}\mathbb{E}_{\hat{\beta}}\left\|\hat{\beta}-\hat{\beta}\right\|_2^2$$

$$\leq\alpha^2+\sigma^2\cdot d/n. \tag{53}$$

We will next bound the term (2).

We need the posterior mean $\mathbb{E}[\beta|X,y]$. We choose the prior $\beta\sim\mathcal{N}(0,b^{-1}I_d)$ where b is the precision. This gives the posterior $\beta|X,y\sim\mathcal{N}(\mu,\Lambda^{-1})$ where

$$\Lambda=aX^\top X+bI,\quad \mu=a\Lambda^{-1}X^\top y. \tag{54}$$

The above is a standard derivation for computing the posterior of the linear regression parameter under a Bayesian prior (see Chapter 11.7 in Murphy [2022] for reference). The parameter a here is the noise variance σ^2.

Then we have,

$$\mathbb{E}_\beta[\beta|X,y]=a\Lambda^{-1}X^\top y \tag{55}$$

Now to remove the conditioning, we can expand

$$\mathbb{E}_{X,y}\left\|X^\top X\mathbb{E}_\beta[\beta|X,y]-X^\top y\right\|_2^2=\mathbb{E}\left[\left\|X^\top X\mathbb{E}[\beta|X,y]-X^\top y\right\|_2^2\right] \tag{56}$$

$$=\mathbb{E}\left[\left\|X^\top Xa\Lambda^{-1}X^\top y-X^\top y\right\|_2^2\right] \tag{57}$$

$$=\mathbb{E}\left[\left\|(X^\top Xa\Lambda^{-1}-I_d)X^\top y\right\|_2^2\right] \tag{58}$$

Then

$$=\mathbb{E}\left[\left\|(X^\top Xa\Lambda^{-1}-I_d)X^\top y\right\|_2^2\right] \tag{59}$$

$$\leq\sqrt{\mathbb{E}\left[\|(X^\top Xa\Lambda^{-1}-I_d)\|_2^4\right]\mathbb{E}\left[\|X^\top y\|_2^4\right]} \tag{60}$$

$$=\sqrt{\mathbb{E}\left[\|(X^\top Xa\Lambda^{-1}-\Lambda\Lambda^{-1})\|_2^4\right]\mathbb{E}\left[\|X^\top y\|_2^4\right]} \tag{61}$$

$$\leq\sqrt{\mathbb{E}\left[\|(X^\top Xa-\Lambda)\|_2^4\|\Lambda^{-1}\|_2^4\right]\mathbb{E}\left[\|X^\top y\|_2^4\right]} \tag{62}$$

$$=\sqrt{\mathbb{E}\left[\|(X^\top Xa-(X^\top Xa+bI))\|_2^4\|\Lambda^{-1}\|_2^4\right]\mathbb{E}\left[\|X^\top y\|_2^4\right]} \tag{63}$$

$$=\sqrt{b^4\mathbb{E}\left[\|\Lambda^{-1}\|_2^4\right]\mathbb{E}\left[\|X^\top y\|_2^4\right]} \tag{64}$$

We now derive upper bounds on $\mathbb{E}\left[\left\|\Lambda^{-1}\right\|_2^4\right]$ and $\mathbb{E}\left[\left\|X^\top y\right\|_2^4\right]$.

Lemma C.7. *Under the conditions of Theorem 5.1,* $\mathbb{E}\left[\left\|\Lambda^{-1}\right\|_2^4\right]\leq O\left(\frac{1}{(aN+b)^4}\right)$, *when* $N=m+n=\Omega(d)$ *and* $b>1/N$.

Proof. First, note that we can rewrite $\mathbb{E}\left[\left\|\Lambda^{-1}\right\|_2^4\right]$ as follows:

$$\mathbb{E}\left[\left\|\Lambda^{-1}\right\|_2^4\right]=\mathbb{E}\left[\left\|(aX^\top X+bI)^{-1}\right\|_2^4\right] \tag{65}$$

$$=\mathbb{E}\left[\frac{1}{(a\sigma_{\min}(X^\top X)+b)^4}\right] \tag{66}$$

We will split the expectation into two conditional expectations:

$$\mathbb{E}\left[\left\|\Lambda^{-1}\right\|_2^4\right] = \mathbb{E}\left[\left\|\Lambda^{-1}\right\|_2^4 \mid \sigma_{\min}(X^\top X) \in (0,t]\right] \Pr[\sigma_{\min}(X^\top X) \in (0,t)] + \tag{67}$$

$$\mathbb{E}\left[\left\|\Lambda^{-1}\right\|_2^4 \mid \sigma_{\min}(X^\top X) \in [t,\infty)]\right] \Pr[\sigma_{\min}(X^\top X) \in [t,\infty)]. \tag{68}$$

for some $0 < t < 1$.

For any $\sigma \geq 0$ we can bound $\mathbb{E}\left[\left\|\Lambda^{-1}\right\|_2^4\right] < 1/b^4$, and when $\sigma \geq t$ we additionally have $\mathbb{E}\left[\left\|\Lambda^{-1}\right\|_2^4\right] \leq 1/(at+b)^4$. Then the above simplifies to

$$\mathbb{E}\left[\left\|\Lambda^{-1}\right\|_2^4 \mid \sigma_{\min}(X^\top X) \in (0,t]\right] \Pr[\sigma_{\min}(X^\top X) < t] + \tag{69}$$

$$\mathbb{E}\left[\left\|\Lambda^{-1}\right\|_2^4 \mid \sigma_{\min}(X^\top X) \in [t,\infty)]\right] \Pr[\sigma_{\min}(X^\top X) \in [t,\infty)] \tag{70}$$

$$= \frac{1}{b^4} \Pr[\sigma_{\min}(X^\top X) < t] + \frac{1}{(at+b)^4} \Pr[\sigma_{\min}(X^\top X) \geq t] \tag{71}$$

We approximate $\Pr[\sigma_{\min}(X^\top X) \geq t] \leq 1$, giving

$$= \frac{1}{b^4} \Pr[\sigma_{\min}(X^\top X) < t] + \frac{1}{(at+b)^4} \tag{72}$$

It then remains to bound $\Pr[\sigma_{\min}(X^\top X) < t]$.

As $X^\top X$ is symmetric and PSD, we have $\sigma_{\min}(X^\top X) = \lambda_{\min}(X^\top X) = \sigma_{\min}(X)^2$.

Now we have

$$\Pr[\sigma_{\min}(X^\top X) \in (0,t]] \leq \Pr(\sigma_{\min}(X^\top X) < t) \tag{73}$$

$$= \Pr(\sigma_{\min}(X)^2 < t) \tag{74}$$

We can now use Lemma B.1. Setting $t = N(1 - \sqrt{d/N} - \psi)^2$ for some $0 < \psi < 1 - \sqrt{d/N}$, $\Pr(\sigma_{\min}(X)^2 < t) \leq e^{-N\psi^2/2}$.

Using Fact B.5, we can upper bound this by, e.g., $O(1/N^8)$.

Thus, we have $\frac{1}{b^4} \Pr[\sigma_{\min}(X^\top X) < t] = O(\frac{1}{b^4 N^8})$, so in total we have

$$\mathbb{E}\left[\left\|\Lambda^{-1}\right\|_2^4\right] \leq \frac{1}{b^4} \Pr[\sigma_{\min}(X^\top X) < t] + \frac{1}{(at+b)^4} \tag{75}$$

$$\leq O\left(\frac{1}{b^4 N^8}\right) + O\left(\frac{1}{(at+b)^4}\right) \tag{76}$$

Assuming $N > 100d$, we have the constraint $0 < \psi < 9/10$. We set $\psi = 1/10$. Then this gives

$$t = N(1 - \sqrt{d/N} - 1/10)^2 \tag{77}$$

$$= N(9/10)^2 - 2(9/10)\sqrt{dN} + d \tag{78}$$

Since we assume $N > 100d$ (i.e. $d < N/100$),

$$\geq N(81/100) - (18/10)\sqrt{N^2/100} + d \tag{79}$$

$$= N(81-18)/100 + d = \Omega(N). \tag{80}$$

Thus, for $b > 1/N$, we have

$$O\left(\frac{1}{b^4 N^8}\right) + O\left(\frac{1}{(at+b)^4}\right) \leq O\left(\frac{1}{b^4 N^8}\right) + O\left(\frac{1}{(aN+b)^4}\right) \leq O\left(\frac{1}{(aN+b)^4}\right). \tag{81}$$

□

Lemma C.8. *Under the conditions of Theorem 5.1,* $\mathbb{E}\left[\|X^T y\|_2^4\right] = \mathcal{O}\left(N^2d^2 + N^4d^2/b^2\right)$, *where* $N = m+n$.

Proof. Let $N = m+n$ be the total number of datapoints (including both public and private). First, we simplify $X^T y = X^T(X\beta + \vec{\eta})$. Then, we use the identity $(a+b)^2 \le 2(a^2+b^2)$ twice and get:

$$\mathbb{E}\left[\|X^T y\|_2^4\right] \le \mathbb{E}\left[\|X^T X\beta\|_2^4\right] + \mathbb{E}\left[\|X^T\vec{\eta}\|_2^4\right]. \tag{82}$$

Now, we bound each of the above two terms separately. First, we start with the second term $\mathbb{E}\left[\|X^T\eta\|_2^4\right]$.

$$\begin{aligned}\mathbb{E}\left[\|X^T\vec{\eta}\|_2^4\right] &= \mathbb{E}\left[\left\|\sum\nolimits_{i\in[N]} x_i\eta_i\right\|_2^4\right] \\ &= \mathbb{E}\left[\sum\nolimits_{i,j,k,l\in[N]} x_i^T x_j x_k^T x_l \cdot \eta_i\eta_j\eta_k\eta_l\right]\end{aligned}$$

Since $\eta_i \perp \eta_j$ whenever $i \neq j$ and $\mathbb{E}[\eta] = 0$, it is easy to see that the only non-zero terms in the expectation above are when: (i) $i=j=k=l$; (ii) $i=j, k=l, i\neq k$; (iii) $(i,j)=(k,l), i\neq j$. Thus, the summation breaks down into:

$$\begin{aligned}\mathbb{E}\left[\|X^T\vec{\eta}\|_2^4\right] &= N\mathbb{E}\left[\|x\|_2^4\right]\mathbb{E}\left[\eta^4\right] + N(N-1)\left(\mathbb{E}\left[\|x\|_2^2\right]\mathbb{E}\left[\eta^2\right]\right)^2 \\ &\quad + N(N-1)\mathbb{E}\left[\|x\|_2^2\right]\left(\mathbb{E}\left[\eta^2\right]\right)^2 \\ &= 3N\mathbb{E}\left[\|x\|_2^4\right] + N(N-1)\left(\mathbb{E}\left[\|x\|_2^2\right]\right)^2 + \mathbb{E}\left[\|x\|_2^2\right] \\ &= O(N^2d^2).\end{aligned} \tag{83}$$

In the above equations, the first equality uses $x_i \perp x_j$ when $i \neq j$, and $x_i \perp \eta_i, \forall i$. The second uses the fact that $\mathbb{E}\left[\eta^4\right] = 3$ and $\mathbb{E}\left[\eta^2\right] = 1$. Finally, the last equality uses $\mathbb{E}\left[\|x\|_2^2\right] = d$ and $\mathbb{E}\left[\|x\|_2^4\right] = O(d^2)$.

Next, we look at the first term in Eq. 82. Using the independence of X and β, we bound this as follows:

$$\mathbb{E}\left[\left\|X^\top X\beta\right\|_2^4\right] \le \mathbb{E}\left[\left\|X^\top X\right\|_2^4 \|\beta\|_2^4\right] \tag{84}$$

$$= \mathbb{E}\left[\left\|X^\top X\right\|_2^4\right]\mathbb{E}\left[\|\beta\|_2^4\right] \tag{85}$$

Now we have $\mathbb{E}\left[\|\beta\|_2^4\right] = O(\frac{d^2}{b^2})$. Further, we can bound $\mathbb{E}\left[\left\|X^\top X\right\|_2^4\right] = O(N^4)$.

This gives an overall bound of $\mathbb{E}\left[\left\|X^\top X\beta\right\|_2^4\right] = O(N^4d^2/b^2)$. Combining with the previous term gives the result.

□

Lemma C.9. *When* $b = 1/d > 1/N$, $m+n = \Omega(d)$, *and* $\alpha = O(1)$, *the expression:*

$$\sqrt{\mathbb{E}_{X,\eta}\mathbb{E}_{\hat{\beta}}\left\|M(X,y) - \hat{\beta}\right\|_2^2 \cdot \mathbb{E}_{X,\eta}\|X^\top X\mathbb{E}[\beta|X,\eta] - X^\top y\|_2^2} = O(1).$$

Proof. From Lemma C.7 and Lemma C.8, we derive the following:

$$\sqrt{\mathbb{E}_{X,\eta}\mathbb{E}_{\hat{\beta}}\left\|M(X,y) - \hat{\beta}\right\|_2^2 \cdot \sqrt{b^4\mathbb{E}\left[\|\Lambda^{-1}\|_2^4\right]\mathbb{E}\left[\|X^\top y\|_2^4\right]}} \tag{86}$$

$$\le \sqrt{\mathbb{E}_{X,\eta}\mathbb{E}_{\hat{\beta}}\left\|M(X,y) - \hat{\beta}\right\|_2^2 \cdot b^2(1/(aN+b)^2) \cdot \sqrt{N^4d^2/b^2 + N^2d^2}}$$

$$\le \sqrt{(\sigma^2 d/N + \alpha^2)b^2(1/(aN+b)^2) \cdot (\sqrt{N^4d^2/b^2} + \sqrt{N^2d^2})} \tag{87}$$

$$\le \sqrt{(\sigma^2 d/N + \alpha^2)b^2(1/(aN)^2) \cdot (N^2d/b + Nd)} \tag{88}$$

$$\le \sqrt{(\sigma^2 d/N + \alpha^2)b \cdot \left(\frac{d}{a^2} + \frac{db}{a^2N}\right)} \tag{89}$$

Note that generally we will set $a=\sigma^2$. Let $b=1/d$. Also, note $\alpha=O(1)$. Then:

$$\leq \sqrt{(\sigma^2 d/N+\alpha^2)(1/d)\cdot(\sigma^2 d+\sigma^4/N)} \tag{90}$$
$$\leq \sqrt{(\sigma^2 d/N+\alpha^2)\cdot(\sigma^4+\sigma^4/(Nd))} \tag{91}$$
$$=O(1), \tag{92}$$

since $\alpha^2=O(1)$. □

Now, we are ready to prove the final result in Theorem 1.3. For this we invoke the upper bound on $\mathbb{E}[\sum_i Z_i]$ in Corollary C.4 and the lower bound of $\mathbb{E}[\sum_i Z_i]=\Omega(d)$ induced by Eq. 9, Lemma A.6 and Lemma 86. This tells us that there exist positive constants c_1,c_2 and values m_0,n_0 such that for $m\geq m_0,n\geq n_0$, the following holds:

$$c_1\cdot d\leq \mathbb{E}\left[\sum_{i\in[N]} Z_i\right]\leq c_2\cdot\left(n\varepsilon\alpha+\alpha\sqrt{md}\right) \tag{93}$$

Thus, either $m=\Omega(\frac{d}{\alpha^2})$ or $n\geq\Omega(\frac{d}{\varepsilon\alpha})$. Note that, similar to Lemma A.8 for Gaussian mean estimation, a similar statistical lower bound also applies for linear regression which implies $n=\Omega(\frac{d}{\alpha^2})$. Together this tells us, that when public and private data come from the same distribution, then for privately solving linear regression with an ε,δ-DP mechanism, where $\delta\leq\varepsilon^2/d$, to a parameter estimation accuracy of α in ℓ_2, we either need $m=\Omega(d/\alpha^2)$ public samples or the number of public and private samples together must satisfy: $n+m=\Omega(\frac{d}{\alpha\varepsilon}+\frac{d}{\alpha^2})$. This completes the proof of our result in Theorem 5.1.

D LINEAR REGRESSION WITH SHIFTED PARAMETER

We now consider a variation of linear regression where the public parameter β_{pub} is shifted by a random vector v sampled from $\mathcal{N}(\mathbf{0},\tau^2/d\cdot I_d)$. The goal is to use both the public and private samples to estimate the private parameter β_{priv}.

That is, we have n samples (x_i,y_i) such that

$$y_i=x_i^\top\beta_{\text{priv}}+\eta_i \tag{94}$$

and m samples such that

$$y_i=x_i^\top\beta_{\text{pub}}+\eta_i \tag{95}$$
$$=x_i^\top(\beta_{\text{priv}}+v)+\eta_i \tag{96}$$

Let $X^\top=[X_{\text{pub}}^\top\ X_{\text{priv}}^\top]$. Note that $y=X\beta+Pv+\eta$ where

$$P=\begin{bmatrix}X_{\text{pub}}\\ 0_{n\times d}\end{bmatrix} \tag{97}$$

We can define $\Psi=Pv+\eta$ to be noise correlated with X_{pub}.

Set $\hat{\beta}$ to be the generalized least squares (GLS) estimator [Aitken, 1936] $(X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}y$, where $\Sigma=\text{Cov}(\Psi|X)$. In particular because P,v,η are independent and mean zero,

$$\Sigma=\mathbb{E}\left[(Pv+\eta)(Pv+\eta)^\top|X\right]=\mathbb{E}\left[Pvv^\top P^\top\right]+\mathbb{E}\left[\eta\eta^T\right] \tag{98}$$
$$=(\tau^2/d)PP^\top+\sigma^2 I_N \tag{99}$$

Then we have $(X^\top\Sigma^{-1}X)\hat{\beta}=X^\top\Sigma^{-1}y$.

With this in mind, we will use a modified test statistic:

$$\mathbb{E}\left[\sum_i Z_i\right]=\mathbb{E}\left[\langle M(X,y)-\beta,X^\top\Sigma^{-1}(X\beta-y)\rangle\right] \tag{100}$$

D.1 Upper bound

As in the non-shifted proof, for the upper bound, we split the sum over fingerprinting statistics into public and private samples. For the private samples we will use the differential privacy constraint to bound the sum over samples (as in the non-shifted case), while for the public samples we can use concentration while taking into account the parameter shift.

Note that Σ is block-diagonal, so that

$$\Sigma^{-1} = \begin{bmatrix} (\frac{\tau^2}{d} X_{\text{pub}} X_{\text{pub}}^\top + \sigma^2 I_m)^{-1} & 0 \\ 0 & 1/\sigma^2 I_n \end{bmatrix} \tag{101}$$

Hence, $X^\top \Sigma^{-1} = [X_{\text{pub}}^T (\frac{\tau^2}{d} X_{\text{pub}} X_{\text{pub}}^\top + \sigma^2 I_m)^{-1} \ \ \frac{1}{\sigma^2} X_{\text{priv}}^\top]$. This implies that we can analyze the sum over private samples using the same analysis as in the non-shifted setting, using C.2, up to a factor of $1/\sigma^2$.

It then remains to bound the sum over public samples using concentration.

Lemma D.1. $\mathbb{E}\left[\sum_{i=n}^{n+m} Z_i\right] \leq O\left(n\varepsilon\alpha + \alpha\sqrt{\frac{1}{\sigma^2} md - \frac{1}{\sigma^4}\left(\frac{m^2 d}{\frac{d}{\tau^2} + \frac{1}{\sigma^2} m}\right)}\right)$.

Proof. Let $\Sigma_{\text{pub}}^{-1} = (\Sigma^{-1})_{11} = (\frac{\tau^2}{d} X_{\text{pub}} X_{\text{pub}}^\top + \sigma^2 I_m)^{-1}$. Then we have

$$\mathbb{E}\left[\sum_{i=n}^{n+m} Z_i\right] = \mathbb{E}\left[\langle M(X,y) - \beta, X_{\text{pub}}^\top \Sigma_{\text{pub}}^{-1} (X_{\text{pub}}\beta - y_{\text{pub}})\rangle\right] \tag{102}$$

$$= \mathbb{E}\left[\langle M(X,y) - \beta, X_{\text{pub}}^\top \Sigma_{\text{pub}}^{-1} (X_{\text{pub}}\beta - (X_{\text{pub}}\beta + X_{\text{pub}} v + \eta_{\text{pub}}))\rangle\right] \tag{103}$$

$$\leq \sqrt{\mathbb{E}\left[\|M(X,y) - \beta\|_2^2\right]} \sqrt{\mathbb{E}\left[\left\| X_{\text{pub}}^\top \Sigma_{\text{pub}}^{-1} (X_{\text{pub}} v + \eta_{\text{pub}})\right\|_2^2\right]} \tag{104}$$

We first bound just the second term.

$$\mathbb{E}\left[\left\| X_{\text{pub}}^\top \Sigma_{\text{pub}}^{-1} (X_{\text{pub}} v + \eta_{\text{pub}})\right\|_2^2\right] \tag{105}$$

$$= \mathbb{E}\left[\text{Tr}\left((X_{\text{pub}} v + \eta_{\text{pub}})^\top \Sigma_{\text{pub}}^{-1} X_{\text{pub}} X_{\text{pub}}^\top \Sigma_{\text{pub}}^{-1} (X_{\text{pub}} v + \eta_{\text{pub}})\right)\right] \tag{106}$$

$$= \mathbb{E}\left[\text{Tr}\left(\Sigma_{\text{pub}}^{-1} X_{\text{pub}} X_{\text{pub}}^\top \Sigma_{\text{pub}}^{-1} (X_{\text{pub}} v + \eta_{\text{pub}})(X_{\text{pub}} v + \eta_{\text{pub}})^\top\right)\right] \tag{107}$$

$$= \text{Tr}(\mathbb{E}_X\left[\mathbb{E}\left[\Sigma_{\text{pub}}^{-1} X_{\text{pub}} X_{\text{pub}}^\top \Sigma_{\text{pub}}^{-1} | X\right] \mathbb{E}\left[(X_{\text{pub}} v + \eta_{\text{pub}})(X_{\text{pub}} v + \eta_{\text{pub}})^\top | X\right]\right]) \tag{108}$$

$$= \text{Tr}(\mathbb{E}_X\left[\Sigma_{\text{pub}}^{-1} X_{\text{pub}} X_{\text{pub}}^\top \Sigma_{\text{pub}}^{-1} \Sigma_{\text{pub}}\right]) \tag{109}$$

$$= \mathbb{E}_X\left[\text{Tr}(X_{\text{pub}}^\top \Sigma_{\text{pub}}^{-1} X_{\text{pub}})\right] \tag{110}$$

Applying the Woodbury identity to Σ_{pub}^{-1}, we get

$$= \mathbb{E}_X\left[\text{Tr}(X_{\text{pub}}^\top (\frac{1}{\sigma^2} I_m - \frac{1}{\sigma^4} X_{\text{pub}}^\top (\frac{d}{\tau^2} I_d + \frac{1}{\sigma^2} X_{\text{pub}}^\top X_{\text{pub}})^{-1} X_{\text{pub}}^\top) X)\right] \tag{111}$$

$$= \mathbb{E}_X\left[\text{Tr}(X_{\text{pub}}^\top \frac{1}{\sigma^2} X)\right] - \mathbb{E}_X\left[\text{Tr}(X_{\text{pub}}^\top (\frac{1}{\sigma^4} X_{\text{pub}} (\frac{d}{\tau^2} I_d + \frac{1}{\sigma^2} X_{\text{pub}}^\top X_{\text{pub}})^{-1} X_{\text{pub}}^\top) X)\right] \tag{112}$$

The first term is $\frac{1}{\sigma^2}md$ by standard Frobenius norm bounds. Then we need to lower bound the second term:

$$\mathbb{E}_X\left[\mathrm{Tr}(\frac{1}{\sigma^4}(\frac{d}{\tau^2}I_d+\frac{1}{\sigma^2}X_{\text{pub}}^\top X_{\text{pub}})^{-1}X_{\text{pub}}^\top X_{\text{pub}}X_{\text{pub}}^\top X)\right] \tag{113}$$

$$\geq\mathbb{E}_X\left[\frac{1}{\sigma^4}\left\|(\frac{d}{\tau^2}I_d+\frac{1}{\sigma^2}X_{\text{pub}}^\top X_{\text{pub}})^{-1}X_{\text{pub}}^\top X_{\text{pub}}X_{\text{pub}}^\top X\right\|_2\right] \tag{114}$$

$$\geq\frac{1}{\sigma^4}\mathbb{E}_X\left[\lambda_{\min}(\frac{d}{\tau^2}I_d+\frac{1}{\sigma^2}X_{\text{pub}}^\top X_{\text{pub}})^{-1}\mathrm{Tr}(X_{\text{pub}}^\top X_{\text{pub}}X_{\text{pub}}^\top X_{\text{pub}})\right] \tag{115}$$

$$=\frac{1}{\sigma^4}\mathbb{E}_X\left[\frac{1}{\left\|(\frac{d}{\tau^2}I_d+\frac{1}{\sigma^2}X_{\text{pub}}^\top X_{\text{pub}})\right\|_2}\left\|X_{\text{pub}}^\top X_{\text{pub}}\right\|_F^2\right] \tag{116}$$

$$\geq\frac{1}{\sigma^4}\mathbb{E}_X\left[\frac{1}{\left\|(\frac{d}{\tau^2}I_d+\frac{1}{\sigma^2}X_{\text{pub}}^\top X_{\text{pub}})\right\|_2}\left(\sum_{j=1}^{d}\lambda_j(X_{\text{pub}}^\top X_{\text{pub}})^2\right)\right] \tag{117}$$

$$\geq\frac{1}{\sigma^4}\mathbb{E}_X\left[\frac{1}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}\lambda_{\max}(X_{\text{pub}}^\top X_{\text{pub}})}\left(d\lambda_{\min}(X_{\text{pub}}^\top X_{\text{pub}})^2\right)\right] \tag{118}$$

We can break the above conditional expectation into two components, one where $\lambda_{\min}(X_{\text{pub}}^\top X_{\text{pub}})\in(0,t_1]$,$\lambda_{\max}(X_{\text{pub}}^\top X_{\text{pub}})\in[t_2,\infty)$,$t_1\leq t_2$, and the other given by its complement, for some appropriately chosen $t_1,t_2>0$.

$$\frac{1}{\sigma^4}\mathbb{E}_X\left[\frac{1}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}\lambda_{\max}(X_{\text{pub}}^\top X_{\text{pub}})}\left(d\lambda_{\min}(X_{\text{pub}}^\top X_{\text{pub}})^2\right)\right]$$
$$\geq\frac{1}{\sigma^4}\cdot\frac{dt_1^2}{d/\tau^2+t_2/\sigma^2}\Pr\left(\lambda_{\min}(X_{\text{pub}}^\top X_{\text{pub}})\geq t_1\right)\cdot\Pr\left(\lambda_{\max}(X_{\text{pub}}^\top X_{\text{pub}})\leq t_2\right)$$

The two tail bounds on the eigen values of $X_{\text{pub}}^\top X_{\text{pub}}$ that appear in the above equation can be resolved using Lemma B.1, and Lemma B.3. When $N\geq 100d$, then with constant probability $\lambda_{\min}(X_{\text{pub}}^\top X_{\text{pub}})=m-\Omega(1)$ and $\lambda_{\max}(X_{\text{pub}}^\top X_{\text{pub}})=m+O(1)$. Plugging this result into the above lower bound, we conclude:

$$\frac{1}{\sigma^4}\mathbb{E}_X\left[\frac{1}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}\lambda_{\max}(X_{\text{pub}}^\top X_{\text{pub}})}\left(d\lambda_{\min}(X_{\text{pub}}^\top X_{\text{pub}})^2\right)\right]\geq\frac{1}{\sigma^4}\Omega\left(\frac{dm^2}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}m}\right) \tag{119}$$

Thus, in total, we have:

$$\mathbb{E}\left[\left\|X_{\text{pub}}^\top\Sigma_{\text{pub}}^{-1}(X_{\text{pub}}v+\eta_{\text{pub}})\right\|_2^2\right]\leq\frac{1}{\sigma^2}md-\frac{1}{\sigma^4}\left(\frac{m^2d}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}m}\right). \tag{120}$$

Finally, from the accuracy assumption of the mechanism M we have $\sqrt{\mathbb{E}\left[\|M(X,y)-\beta\|_2^2\right]}\leq\alpha$. Putting it all together, we arrive at the inequality:

$$\mathbb{E}\left[\sum_{i=n}^{n+m}Z_i\right]=O\left(\alpha\sqrt{\frac{1}{\sigma^2}md-\frac{1}{\sigma^4}\left(\frac{m^2d}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}m}\right)}\right). \tag{121}$$

The statement of Lemma D.1 follows from applying the non-shifted upper bound over private samples in Proposition C.2 along with the above result.

□

D.2 Lower bound

Transforming the test statistic appropriately gives:

$$\mathbb{E}\left[\sum_i Z_i\right] = \mathbb{E}\left[\left\|\Sigma^{-1/2}X(\hat{\beta}-\beta)\right\|_2^2\right] - \mathbb{E}\left[\langle M(X,y)-\hat{\beta}, X^\top\Sigma^{-1}(X\beta-y)\rangle\right] \tag{122}$$

In the above, we lower bound the first term with the non-private statistical error. For the second term, since Σ only depends on X_{pub} and σ^2, we can still break the dependence in the terms as usual by conditioning on X and y.

$$\mathbb{E}_{X,y}\left[\mathbb{E}_{M,\beta,v}\left[\langle M(X,y)-\hat{\beta}, X^\top\Sigma^{-1}X\beta - X^\top\Sigma^{-1}y\rangle \,\middle|\, X,y\right]\right] \tag{123}$$

$$= \mathbb{E}_{X,y}\left[\langle \mathbb{E}_M[M(X,y)-\hat{\beta}\,|\,X,y], \mathbb{E}_{\beta,v}\left[X^\top\Sigma^{-1}X\beta - X^\top\Sigma^{-1}y\,|\,X,y\right]\rangle\right] \tag{124}$$

$$\le \sqrt{\mathbb{E}_{X,y}\mathbb{E}_M\left\|M(X,y)-\hat{\beta}\right\|_2^2}\cdot\sqrt{\mathbb{E}_{X,y}\left\|\mathbb{E}_{\beta,v}\left[X^\top\Sigma^{-1}X\beta - X^\top\Sigma^{-1}y\,|\,X,y\right]\right\|_2^2} \tag{125}$$

The final inequality uses Cauchy-Schwarz. Removing the conditioning on X,y we get:

$$\underbrace{\sqrt{\mathbb{E}_{X,y,M}\left\|M(X,y)-\hat{\beta}\right\|_2^2}}_{①}\cdot\underbrace{\mathbb{E}_{X,y}\|X^\top\Sigma^{-1}X\cdot\mathbb{E}_{\beta,v}[\beta|X,y]-X^\top\Sigma^{-1}y\|_2^2}_{②} \tag{126}$$

Next, we present a lemma that lower bounds the first term in Eq. 122.

Lemma D.2. $\mathbb{E}\left[\left\|\Sigma^{-1/2}X(\hat{\beta}-\beta)\right\|_2^2\right] = \Theta(d)$.

Proof. First, we simplify $\hat{\beta}-\beta$:

$$\hat{\beta}-\beta = (X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}(X\beta+Pv+\eta)-\beta \tag{127}$$

$$= (X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}(Pv+\eta) \tag{128}$$

Then we have,

$$\mathbb{E}\left[\left\|\Sigma^{-1/2}X(\hat{\beta}-\beta)\right\|_2^2\right] = \mathrm{Tr}(\mathbb{E}\left[(Pv+\eta)^\top\Sigma^{-1}X(X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}(Pv+\eta)\right]) \tag{129}$$

$$= \mathrm{Tr}(\mathbb{E}\left[\Sigma^{-1}X(X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}(Pv+\eta)(Pv+\eta)^\top\right]) \tag{130}$$

This is equivalent to:

$$= \mathrm{Tr}(\mathbb{E}_X\left[\mathbb{E}\left[\Sigma^{-1}X(X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}|X\right]\mathbb{E}\left[(Pv+\eta)(Pv+\eta)^\top|X\right]\right]) \tag{131}$$

$$= \mathrm{Tr}(\mathbb{E}_X\left[\Sigma^{-1}X(X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}\Sigma\right]) \tag{132}$$

$$= \mathrm{Tr}(\mathbb{E}_X\left[(X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}X\right]) \tag{133}$$

$$= \mathrm{Tr}(I_d) = d. \tag{134}$$

□

Now, we move to upper bound the second term in Eq. 122. For this, we compute the posterior mean over β.

Lemma D.3 (posterior for generalized least squares)**.** *For a fixed covariates matrix X, we are given responses y drawn from the following Bayesian model for generalized linear regression:*

$$\beta \sim \mathcal{N}\left(0, \Lambda_0^{-1}\right), \qquad y\,|\,\beta \sim \mathcal{N}\left(X\beta, \Sigma\right),$$

where $X \in \mathbb{R}^{N\times d}, y \in \mathbb{R}^N$, $\Lambda_0 \succeq 0$ is the prior precision, and $\Sigma \succ 0$ is noise covariance. Then the mean of the posterior distribution over $\beta \mid X,y$ is given by:

$$\mathbb{E}[\beta \mid X,y] = \left(X^\top \Sigma^{-1} X + \Lambda_0\right)^{-1} X^\top \Sigma^{-1} y$$

Proof. First, lets apply the Bayes' rule in its canonical form:

$$\log p(\beta \mid y) = -\tfrac{1}{2}(y - X\beta)^\top \Sigma^{-1}(y - X\beta) - \tfrac{1}{2}\beta^\top \Lambda_0 \beta + \text{const.} \tag{135}$$

Then, we can expand the quadratic in $y - X\beta$ to get:

$$(y - X\beta)^\top \Sigma^{-1}(y - X\beta) = y^\top \Sigma^{-1} y - 2\beta^\top X^\top \Sigma^{-1} y + \beta^\top X^\top \Sigma^{-1} X \beta. \tag{136}$$

Hence, we can conclude

$$\log p(\beta \mid y) = -\tfrac{1}{2}\left[\beta^\top \left(X^\top \Sigma^{-1} X + \Lambda_0\right)\beta - 2\beta^\top X^\top \Sigma^{-1} y\right] + \text{const} \tag{137}$$

Next, we apply "completing the squares" trick, where we subtract and add $y^\top \Sigma^{-1} X (X^\top \Sigma^{-1} X + \Lambda_0)^{-1} X^\top \Sigma^{-1} y$. Note that this term does not contain β. So, we can treat it as a constant when computing $\log p(\beta \mid X,y)$.

$$\begin{aligned}\log p(\beta \mid y) = -\tfrac{1}{2}\Big[&\beta^\top \left(X^\top \Sigma^{-1} X + \Lambda_0\right)\beta \\ &- 2\beta^\top X^\top \Sigma^{-1} y + y^\top \Sigma^{-1} X (X^\top \Sigma^{-1} X + \Lambda_0)^{-1} X^\top \Sigma^{-1} y\Big] + \text{const.}\end{aligned} \tag{138}$$

For simplicity of calculations lets first define: Define the *posterior precision*

$$\Lambda_{\text{post}} := X^\top \Sigma^{-1} X + \Lambda_0, \quad \mu_{\text{post}} := \Lambda_{\text{post}}^{-1} X^\top \Sigma^{-1} y.$$

Substituting the above into Eq. 138 yields:

$$\log p(\beta \mid y) = -\tfrac{1}{2}(\beta - \mu_{\text{post}})^\top \Lambda_{\text{post}} (\beta - \mu_{\text{post}}) + \text{const}, \tag{139}$$

since $\Lambda_{\text{post}} \mu_{\text{post}} = X^\top \Sigma^{-1} y$.

The above posterior distribution over $\beta \mid X, y$ matches the functional form of a Gaussian distribution $\beta \mid y \sim \mathcal{N}\left(\mu_{\text{post}}, \Lambda_{\text{post}}^{-1}\right)$ with mean:

$$\mu_{\text{post}} = \left(X^\top \Sigma^{-1} X + \Lambda_0\right)^{-1} X^\top \Sigma^{-1} y,$$

□

With the above posterior mean over β, we are now ready to upper bound the second term in Eq. 126. Let $\tilde{\Lambda} = \left(X^\top \Sigma^{-1} X + \Lambda_0\right)$. Now to remove the conditioning, we can expand:

$$\mathbb{E}_{X,y}\left\|X^\top \Sigma^{-1} X \mathbb{E}_\beta[\beta|X,y] - X^\top \Sigma^{-1} y\right\|_2^2 = \mathbb{E}\left[\left\|X^\top \Sigma^{-1} X \mathbb{E}[\beta|X,y] - X^\top \Sigma^{-1} y\right\|_2^2\right] \tag{140}$$

$$= \mathbb{E}\left[\left\|X^\top \Sigma^{-1} X \tilde{\Lambda}^{-1} X^\top \Sigma^{-1} y - X^\top \Sigma^{-1} y\right\|_2^2\right]$$

$$= \mathbb{E}\left[\left\|(X^\top \Sigma^{-1} X \tilde{\Lambda}^{-1} - I_d) X^\top \Sigma^{-1} y\right\|_2^2\right] \tag{141}$$

Then,

$$= \mathbb{E}\left[\left\|(X^\top \Sigma^{-1} X \tilde{\Lambda}^{-1} - I_d) X^\top \Sigma^{-1} y\right\|_2^2\right] \tag{142}$$

$$\leq \sqrt{\mathbb{E}\left[\left\|(X^\top \Sigma^{-1} X \tilde{\Lambda}^{-1} - I_d)\right\|_2^4\right] \mathbb{E}\left[\|X^\top \Sigma^{-1} y\|_2^4\right]} \tag{143}$$

$$= \sqrt{\mathbb{E}\left[\left\|(X^\top \Sigma^{-1} X \tilde{\Lambda}^{-1} - \tilde{\Lambda}\tilde{\Lambda}^{-1})\right\|_2^4\right] \mathbb{E}\left[\|X^\top \Sigma^{-1} y\|_2^4\right]} \tag{144}$$

$$\leq \sqrt{\mathbb{E}\left[\left\|(X^\top \Sigma^{-1} X - \tilde{\Lambda})\right\|_2^4 \left\|\tilde{\Lambda}^{-1}\right\|_2^4\right] \mathbb{E}\left[\|X^\top \Sigma^{-1} y\|_2^4\right]} \tag{145}$$

$$= \sqrt{\mathbb{E}\left[\|(X^\top X \Sigma^{-1} - (X^\top \Sigma^{-1} X + \Lambda_0)))\|_2^4 \left\|\tilde{\Lambda}^{-1}\right\|_2^4\right] \mathbb{E}\left[\|X^\top \Sigma^{-1} y\|_2^4\right]} \tag{146}$$

$$= \sqrt{\|\Lambda_0\|_2^4 \mathbb{E}\left[\left\|\tilde{\Lambda}^{-1}\right\|_2^4\right] \mathbb{E}\left[\|X^\top \Sigma^{-1} y\|_2^4\right]} \tag{147}$$

Lemma D.4 (re-stating the posterior over β)**.** *Given the following matrices:*

$$S \overset{def}{=} X_{\text{pub}}^\top X_{\text{pub}} \quad W \overset{def}{=} \left(\frac{d}{\tau^2}\mathbf{I}_d + \frac{1}{\sigma^2} S\right)^{-1},$$

we can rewrite the posterior $\mathbb{E}[\beta \,|\, X,y]$ *as:*

$$\mathbb{E}[\beta \,|\, X,y] = (M + b\mathbf{I}_d)^{-1} m,$$

where we can inturn write the matrix M *and vector* m *as:*

$$M = \frac{1}{\sigma^2} X^\top X - \frac{1}{\sigma^4} SWS, \quad m = \frac{1}{\sigma^2} X^\top y - \frac{1}{\sigma^4} SW X_{\text{pub}}^\top y_{\text{pub}}.$$

Proof. Matching the form of $\mathbb{E}[\beta \,|\, X,y]$ to that in Lemma D.3, we need to show: $X^\top \Sigma^{-1} X = \frac{1}{\sigma^2} X^\top X - \frac{1}{\sigma^4} SWS$, and $X^\top \Sigma^{-1} y = \frac{1}{\sigma^2} X^\top y - \frac{1}{\sigma^4} SW X_{\text{pub}}^\top y_{\text{pub}}$. We first rewrite the noise covariance so that the Woodbury identity applies. Let

$$U := \begin{bmatrix} X_{\text{pub}} \\ 0 \end{bmatrix}, \qquad C := \frac{\tau^2}{d} I_d, \tag{149}$$

This means that our Σ matrix is given by:

$$\Sigma = \sigma^2 I_{m+n} + UCU^\top. \tag{150}$$

Applying Woodbury to the above we get,

$$\Sigma^{-1} = \frac{1}{\sigma^2} I_{m+n} - \frac{1}{\sigma^4} U\left(C^{-1} + \sigma^{-2} U^\top U\right)^{-1} U^\top. \tag{151}$$

Because $U^\top U = X_{\text{pub}}^\top X_{\text{pub}} =: S$, we define W as

$$W := \left(\frac{d}{\tau^2} I_d + \sigma^{-2} S\right)^{-1}, \tag{152}$$

this abbreviates the inverse above and yields

$$\Sigma^{-1} = \frac{1}{\sigma^2} I - \frac{1}{\sigma^4} UWU^\top.$$

With the above calculations, we start with the matrix term $X^\top \Sigma^{-1} X$,

$$X^\top \Sigma^{-1} X = \frac{1}{\sigma^2} X^\top X - \frac{1}{\sigma^4} X^\top U W U^\top X \tag{153}$$

$$= \frac{1}{\sigma^2} X^\top X - \frac{1}{\sigma^4} SWS. \tag{154}$$

This matches the definition of M in the statement of Lemma D.4.

$$M = \frac{1}{\sigma^2} X^\top X - \frac{1}{\sigma^4} SWS.$$

Next, we look at the vector term $X^\top \Sigma^{-1} y$.

$$X^\top \Sigma^{-1} y = \frac{1}{\sigma^2} X^\top y - \frac{1}{\sigma^4} X^\top U W U^\top y \tag{155}$$

$$= \frac{1}{\sigma^2} X^\top y - \frac{1}{\sigma^4} SW X_{\text{pub}}^\top y_{\text{pub}}. \tag{156}$$

This matches the m term in the statement of Lemma D.4.

$$m = \frac{1}{\sigma^2} X^\top y - \frac{1}{\sigma^4} SW X_{\text{pub}}^\top y_{\text{pub}}.$$

Plugging this into the posterior mean for β, we get:

$$\mathbb{E}[\beta \mid X, y] = (X^\top \Sigma^{-1} X + bI_d)^{-1} X^\top \Sigma^{-1} y = (M + bI_d)^{-1} m.$$

□

Lemma D.5. *Under the conditions of Theorem 5.2,* $\mathbb{E}\left[\left\|\tilde{\Lambda}^{-1}\right\|_2^4\right] \le O\left(\frac{1}{(\frac{1}{\sigma^2}N - \frac{1}{\sigma^4}m^2 \frac{1}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}m} + b)^4}\right)$, *when* $n, m > 100d$ *and* $b > 1/N$.

Proof. We have $\tilde{\Lambda} = \left(X^\top \Sigma^{-1} X + \Lambda_0\right)$ where we set $\Lambda_0 = bI_d$ for some b. Using the previous transformations,

$$\tilde{\Lambda} = \left(\frac{1}{\sigma^2} X^\top X - \frac{1}{\sigma^4} X_{\text{pub}}^\top X_{\text{pub}} \left(\frac{d}{\tau^2} I_d + \frac{1}{\sigma^2} X_{\text{pub}}^\top X_{\text{pub}}\right)^{-1} X_{\text{pub}}^\top X_{\text{pub}} + bI_d\right) \tag{157}$$

To bound the spectral norm, we will use a series of transformations to express the spectral norm in terms of eigenvalues of the matrices in the expression.

We have

$$\left\|\tilde{\Lambda}^{-1}\right\|_2 = \frac{1}{\lambda_{\min}\left(\frac{1}{\sigma^2} X^\top X - \frac{1}{\sigma^4} X_{\text{pub}}^\top X_{\text{pub}} \left(\frac{d}{\tau^2} I_d + \frac{1}{\sigma^2} X_{\text{pub}}^\top X_{\text{pub}}\right)^{-1} X_{\text{pub}}^\top X_{\text{pub}} + bI_d\right)} \tag{158}$$

(Note that we will eventually take the expectation over the fourth power.)

We want to upper bound this quantity which involves lower bounding the denominator.

Weyl's inequality gives that $\lambda_{\min}(A - B) \ge \lambda_{\min}(A) - \lambda_{\max}(B)$ for real symmetric A and B. Moreover, the original expression $\tilde{\Lambda}$ is PSD and therefore the eigenvalues are lower bounded by 0, giving a stronger guarantee than the Weyl lower bound. Hence,

$$\lambda_{\min}\left(\frac{1}{\sigma^2} X^\top X - \frac{1}{\sigma^4} X_{\text{pub}}^\top X_{\text{pub}} \left(\frac{d}{\tau^2} I_d + \frac{1}{\sigma^2} X_{\text{pub}}^\top X_{\text{pub}}\right)^{-1} X_{\text{pub}}^\top X_{\text{pub}} + bI_d\right) \tag{159}$$

$$= \lambda_{\min}\left(\frac{1}{\sigma^2} X^\top X - \frac{1}{\sigma^4} X_{\text{pub}}^\top X_{\text{pub}} \left(\frac{d}{\tau^2} I_d + \frac{1}{\sigma^2} X_{\text{pub}}^\top X_{\text{pub}}\right)^{-1} X_{\text{pub}}^\top X_{\text{pub}}\right) + b \tag{160}$$

$$\ge \max\left(0, \lambda_{\min}\left(\frac{1}{\sigma^2} X^\top X\right) - \lambda_{\max}\left(\frac{1}{\sigma^4} X_{\text{pub}}^\top X_{\text{pub}} \left(\frac{d}{\tau^2} I_d + \frac{1}{\sigma^2} X_{\text{pub}}^\top X_{\text{pub}}\right)^{-1} X_{\text{pub}}^\top X_{\text{pub}}\right)\right) + b. \tag{161}$$

Focusing now on the last term, which we want to upper bound in order to lower bound the previous expression:

$$\lambda_{\max}\left(\frac{1}{\sigma^4}X_{\text{pub}}^\top X_{\text{pub}}(\frac{d}{\tau^2}I_d+\frac{1}{\sigma^2}X_{\text{pub}}^\top X_{\text{pub}})^{-1}X_{\text{pub}}^\top X_{\text{pub}}\right) \tag{162}$$

$$=\left\|\left(\frac{1}{\sigma^4}X_{\text{pub}}^\top X_{\text{pub}}(\frac{d}{\tau^2}I_d+\frac{1}{\sigma^2}X_{\text{pub}}^\top X_{\text{pub}})^{-1}X_{\text{pub}}^\top X_{\text{pub}}\right)\right\|_2 \tag{163}$$

$$\leq\frac{1}{\sigma^4}\left\|X_{\text{pub}}^\top X_{\text{pub}}\right\|_2^2\left\|(\frac{d}{\tau^2}I_d+\frac{1}{\sigma^2}X_{\text{pub}}^\top X_{\text{pub}})^{-1}\right\|_2 \tag{164}$$

The last line follows from the submultiplicativity of operator norm.

Bounding even further,

$$\left\|(\frac{d}{\tau^2}I_d+\frac{1}{\sigma^2}X_{\text{pub}}^\top X_{\text{pub}})^{-1}\right\|_2=\frac{1}{\lambda_{\min}((\frac{d}{\tau^2}I_d+\frac{1}{\sigma^2}X_{\text{pub}}^\top X_{\text{pub}}))} \tag{165}$$

$$=\frac{1}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}\lambda_{\min}(X_{\text{pub}}^\top X_{\text{pub}})}. \tag{166}$$

Putting it all together, we arrive at:

$$\left\|\tilde{\Lambda}^{-1}\right\|_2\leq\frac{1}{\max\left(0,\frac{1}{\sigma^2}\lambda_{\min}(X^\top X)-\frac{1}{\sigma^4}\left\|X_{\text{pub}}^\top X_{\text{pub}}\right\|_2^2\frac{1}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}\lambda_{\min}(X_{\text{pub}}^\top X_{\text{pub}})}\right)+b} \tag{167}$$

Note $\left\|X_{\text{pub}}^\top X_{\text{pub}}\right\|_2=\lambda_{\max}(X_{\text{pub}}^\top X_{\text{pub}})$. We now have the expression in terms of min and max eigenvalues of $X^\top X$ and $X_{\text{pub}}^\top X_{\text{pub}}$.

Moreover since the expression is positive and monotonic, we also have

$$\left\|\tilde{\Lambda}^{-1}\right\|_2^4\leq\frac{1}{\left(\max\left(0,\frac{1}{\sigma^2}\lambda_{\min}(X^\top X)-\frac{1}{\sigma^4}\left\|X_{\text{pub}}^\top X_{\text{pub}}\right\|_2^2\frac{1}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}\lambda_{\min}(X_{\text{pub}}^\top X_{\text{pub}})}\right)+b\right)^4}. \tag{168}$$

This expression is upper bounded by $1/b^4$ when the max term goes to 0.

To bound this, we will do something similar to the non-shifted setting but instead of bounding a single constant c, we will instead want to bound the three eigenvalue terms with $t_1,t_2,t_3>0$. (Note that all the involved matrices are PSD.) Specifically: Let $\lambda_1:=\lambda_{\min}(X^\top X)$, $\lambda_2:=\lambda_{\max}(X_{\text{pub}}^\top X_{\text{pub}})$, $\lambda_3:=\lambda_{\min}(X_{\text{pub}}^\top X_{\text{pub}})$. Then

$$\mathbb{E}\left[\left\|\tilde{\Lambda}^{-1}\right\|_2^4\right] \tag{169}$$

$$=\mathbb{E}\left[\left\|\tilde{\Lambda}^{-1}\right\|_2^4|\lambda_1\in[t_1,\infty)\wedge\lambda_2\in[0,t_2]\wedge\lambda_3\in[t_3,\infty)\right]$$
$$\cdot\Pr(\lambda_1\in[t_1,\infty)\wedge\lambda_2\in[0,t_2]\wedge\lambda_3\in[t_3,\infty)) \tag{170}$$

$$+\mathbb{E}\left[\left\|\tilde{\Lambda}^{-1}\right\|_2^4|\lambda_1\in[0,t_1)\vee\lambda_2\in(t_2,\infty)\vee\lambda_3\in[0,t_3)\right]$$
$$\cdot\Pr(\lambda_1\in[0,t_1)\vee\lambda_2\in(t_2,\infty)\vee\lambda_3\in[0,t_3)) \tag{171}$$

Again, we will coarsely upper bound the first probability by 1 and use a union over tail bounds to bound the second probability.

$$\leq\frac{1}{\max(0,\frac{1}{\sigma^2}t_1-\frac{1}{\sigma^4}t_2^2\frac{1}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}t_3})+b)^4}+\frac{1}{b^4}\Pr(\lambda_1\in[0,t_1)\vee\lambda_2\in(t_2,\infty)\vee\lambda_3\in[0,t_3)) \tag{172}$$

We will eventually want $t_1 = \Omega(N), t_2 = O(m), t_3 = \Omega(m)$, and we would like the second union over tail bounds to be vanishingly small. Following Lemma B.6 and B.7, we set $t_1 = N(1-\sqrt{d/N}-\psi)^2$, $t_2 = m(1+\sqrt{d/m}+\psi)^2$, and $t_3 = m(1-\sqrt{d/m}-\psi)^2$. We assume $n, m > 100d$ and set $\psi = 1/10$ in all cases. The lemmas then give $t_1 > 63/100 \cdot N, t_2 < 64/100 \cdot m, t_3 > 63/100 \cdot m$.

We can lastly confirm that with these values of t_1, t_2, t_3 we can drop the max in the denominator of the first term, as follows:

$$\frac{1}{\sigma^2}t_1 - \frac{1}{\sigma^4}t_2^2 \frac{1}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}t_3} \geq \frac{1}{\sigma^2}\frac{63}{100}N - \frac{1}{\sigma^4}(\frac{64}{100}m)^2 \frac{1}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}(63/100)m} \overset{?}{>} 0 \tag{173}$$

$$\tag{174}$$

Since $d/\tau^2 > 0$,

$$\frac{1}{\sigma^2}\frac{63}{100}N - \frac{1}{\sigma^4}(\frac{64}{100}m)^2 \frac{1}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}(63/100)m} > \frac{1}{\sigma^2}\frac{63}{100}N - \frac{1}{\sigma^4}(\frac{64}{100}m)^2 \frac{1}{\frac{1}{\sigma^2}(63/100)m} \overset{?}{>} 0 \tag{175}$$

Omitting algebra, and noting that $N = n+m$, this leads to the condition

$$n > ((64/100)^2(100/63)-1)m = -0.349m \tag{176}$$

which is true. Given that this term is therefore strictly greater than 0 for the values of t_1, t_2, t_3 we have chosen, we can drop the $\max(0,\cdot)$.

We can then bound the relevant tail probabilities using Lemma B.1, Lemma B.3, and Fact B.5 and take a union bound to conclude that

$$\Pr(\lambda_1 \in [0,t_1) \vee \lambda_2 \in (t_2,\infty) \vee \lambda_3 \in [0,t_3)) = O\left(\frac{1}{N^8}\right).$$

Thus, we have

$$\mathbb{E}\left[\left\|\tilde{\Lambda}^{-1}\right\|_2^4\right] \leq O\left(\frac{1}{(\frac{1}{\sigma^2}N - \frac{1}{\sigma^4}m^2\frac{1}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}m}+b)^4}\right) + O\left(\frac{1}{b^4N^8}\right) \tag{177}$$

and the first term dominates, giving the result.

□

Lemma D.6 (bound on $\mathbb{E}\left[\left\|X^\top\Sigma^{-1}y\right\|_2^4\right]$). *We can bound* $\mathbb{E}\left[\left\|X^\top\Sigma^{-1}y\right\|_2^4\right]$ *with the following expression:*

$$\mathbb{E}\left[\left\|X^\top\Sigma^{-1}y\right\|_2^4\right] = O\left(\left(\frac{1}{\sigma^2}N - \frac{1}{\sigma^4}\frac{m^2}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}m}\right)^4 (d^2/b^2+d^2)\right). \tag{178}$$

Proof. Using the block-diagonal structure of Σ (and therefore Σ^{-1}), we can decompose $X^\top\Sigma^{-1}y$ into $X_{\text{pub}}^\top\Sigma_{\text{pub}}^{-1}y_{\text{pub}} + X_{\text{priv}}^\top\Sigma_{\text{priv}}^{-1}y_{\text{priv}}$. ($\Sigma_{\text{pub}}$ is the upper-left block of Σ corresponding to public data and Σ_{priv} the corresponding lower-right block for private data.) Here, $\Sigma_{\text{priv}}^{-1}$ is simply $\frac{1}{\sigma^2}I_n$, so we can reuse the upper bound from Lemma C.8 with an additional factor of $\frac{1}{\sigma^8}$.

The term incorporating public data is more complicated.

We first note the following useful transformation (applying the Woodbury identity only to Σ_{pub}^{-1}):

$$\Sigma_{\text{pub}}^{-1} = \frac{1}{\sigma^2}I_m - \frac{1}{\sigma^4}X_{\text{pub}}\left(\frac{d}{\tau^2}I_d + \frac{1}{\sigma^2}X_{\text{pub}}^\top X_{\text{pub}}\right)^{-1}X_{\text{pub}}^\top. \tag{179}$$

Note: For notational convenience, in the following, we will drop the "pub" term from the subscript since we only need to bound $X_{\text{pub}}^\top \Sigma_{\text{pub}}^{-1} y_{\text{pub}}$.

We will use the following bound in several places:

$$\mathbb{E}\Big[\big\|X^\top\Sigma^{-1}X\big\|_2^4\Big] \le \mathbb{E}\left[\left\|X^\top\left(\frac{1}{\sigma^2}I_m - \frac{1}{\sigma^4}X\left(\frac{d}{\tau^2}I_d+\frac{1}{\sigma^2}X^\top X\right)^{-1}X^\top\right)X\right\|_2^4\right] \tag{180}$$

$$= \mathbb{E}\left[\left\|\frac{1}{\sigma^2}X^\top X - \frac{1}{\sigma^4}X^\top X\left(\frac{d}{\tau^2}I_d+\frac{1}{\sigma^2}X^\top X\right)^{-1}X^\top X\right\|_2^4\right] \tag{181}$$

$$\le \mathbb{E}\left[\left(\frac{1}{\sigma^2}\big\|X^\top X\big\|_2 - \frac{1}{\sigma^4}\lambda_{\min}\left(X^\top X\left(\frac{d}{\tau^2}I_d+\frac{1}{\sigma^2}X^\top X\right)^{-1}X^\top X\right)\right)^4\right] \tag{182}$$

$$\le \mathbb{E}\left[\left(\frac{1}{\sigma^2}\big\|X^\top X\big\|_2 - \frac{1}{\sigma^4}\lambda_{\min}(X^\top X)^2\frac{1}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}\|X^\top X\|_2}\right)^4\right] \tag{183}$$

Note that the third inequality follows from the fact that $X^\top\Sigma^{-1}X \succeq 0$, so we can upper bound the full term by applying an upper bound on the term within $(\cdot)^4$ since the term inside it is always positive. Next, we can break the expectation into two conditional expectations.

$$\mathbb{E}\left[\left(\frac{1}{\sigma^2}\big\|X^\top X\big\|_2 - \frac{1}{\sigma^4}\lambda_{\min}(X^\top X)^2\frac{1}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}\|X^\top X\|_2}\right)^4\right]$$

$$\le \mathbb{E}\left[\left(\frac{1}{\sigma^2}\big\|X^\top X\big\|_2 - \frac{1}{\sigma^4}\lambda_{\min}(X^\top X)^2\frac{1}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}\|X^\top X\|_2}\right)^4 \mid \lambda_{\min}(X^\top X)\ge t_1, \lambda_{\max}(X^\top X)\le t_2\right]$$

$$\cdot \Pr\big(\lambda_{\min}(X^\top X)\ge t_1, \lambda_{\max}(X^\top X)\le t_2\big)$$

Using Lemma B.1, we can upper bound the above expression by applying a high probability tail bound on $\lambda_{\min}(X^\top X)$. When $N \ge 100d$, then with constant probability $\lambda_{\min}(X^\top X) = m - \Omega(1)$ and $\lambda_{\max}(X^\top X) = m - O(1)$. Plugging this result into the above expression, we conclude:

$$\mathbb{E}\left[\left(\frac{1}{\sigma^2}\big\|X^\top X\big\|_2 - \frac{1}{\sigma^4}\lambda_{\min}(X^\top X)^2\frac{1}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}\|X^\top X\|_2}\right)^4\right] \le \left(\frac{1}{\sigma^2}m - \frac{1}{\sigma^4}\frac{m^2}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}m}\right)^4. \tag{184}$$

When $\tau \to 0$ the term $\|X^\top\Sigma^{-1}X\beta\|^4$ goes to m^4 (giving the correct contribution of m^4d^2 when combined with the contribution of β) and when $\tau \to \infty$ the second part of the term cancels sending the public sample contribution to 0.

Once again, we decompose the term

$$\mathbb{E}\Big[\big\|X^\top\Sigma^{-1}y\big\|_2^4\Big] = \mathbb{E}\Big[\big\|X^\top\Sigma^{-1}X\beta + Xv + \eta)\big\|_2^4\Big] \tag{185}$$

$$\le \mathbb{E}\Big[\big\|X^\top\Sigma^{-1}X\beta\big\|_2^4\Big] + \mathbb{E}\Big[\big\|X^\top\Sigma^{-1}(Xv+\eta)\big\|_2^4\Big] \tag{186}$$

$$\le \mathbb{E}\Big[\big\|X^\top\Sigma^{-1}X\beta\big\|_2^4\Big] + \mathbb{E}\Big[\big\|X^\top\Sigma^{-1}Xv\big\|_2^4\Big] + \mathbb{E}\Big[\big\|X^\top\Sigma^{-1}\eta\big\|_2^4\Big] \tag{187}$$

$$\le \mathbb{E}\Big[\big\|X^\top\Sigma^{-1}X\big\|_2^4\|\beta\|_2^4\Big] + \mathbb{E}\Big[\big\|X^\top\Sigma^{-1}X\big\|_2^4\|v\|_2^4\Big] + \mathbb{E}\Big[\big\|X^\top\Sigma^{-1}\big\|_2^4\|\eta\|_2^4\Big] \tag{188}$$

$$= \mathbb{E}\Big[\big\|X^\top\Sigma^{-1}X\big\|_2^4\Big]\mathbb{E}\Big[\|\beta\|_2^4\Big] + \tag{189}$$

$$\mathbb{E}\Big[\big\|X^\top\Sigma^{-1}X\big\|_2^4\Big]\mathbb{E}\Big[\|v\|_2^4\Big] + \mathbb{E}\Big[\big\|X^\top\Sigma^{-1}\big\|_2^4\Big]\mathbb{E}\Big[\|\eta\|_2^4\Big] \tag{190}$$

Since $\beta \sim \mathcal{N}(0, \frac{1}{b} I_d)$ we have $\mathbb{E}\left[\|\beta\|_2^4\right] \le d^2/b^2$. Similarly, for $v \sim \mathcal{N}(0, I_d)$ we have $\mathbb{E}[\|v\|_2] \le d^2$. Finally, $\eta \sim \mathcal{N}(0, I_m)$ gives $\mathbb{E}[\|\eta\|_2] \le m^2$.

We also need the following bound:

$$\mathbb{E}\left[\left\|X^\top \Sigma^{-1}\right\|_2^4\right] \le \mathbb{E}\left[\|X\|_2^4 \left\|\Sigma^{-1}\right\|_2^4\right] \tag{191}$$

From Gordon's theorem [Vershynin, 2018] we have $\mathbb{E}\left[\|X\|_2^4\right] \le (\sqrt{m}+\sqrt{d})^4 \le 8(m^2+d^2)$. With a small variation on the earlier analysis of $X^\top \Sigma^{-1} X$, bounding $\|X\|_2^2 \le (\sqrt{m}+\sqrt{d})^2$ using Gordon's theorem, we also have $\mathbb{E}\left[\left\|\Sigma^{-1}\right\|_2^4\right] \le \left(\frac{1}{\sigma^2} - \frac{1}{\sigma^4} \frac{(\sqrt{m}+\sqrt{d})^2}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}m}\right)^4 \le \left(\frac{1}{\sigma^2} - \frac{1}{\sigma^4} \frac{O(m)}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}m}\right)^4$.

Overall, this gives:

$$\mathbb{E}\left[\left\|X^\top \Sigma^{-1} y\right\|_2^4\right] \le O\Big(\frac{1}{\sigma^8}(n^2 d^2 + n^4 d^2/b^2) + \tag{192}$$

$$\left(\frac{1}{\sigma^2} m - \frac{1}{\sigma^4} \frac{m^2}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}m}\right)^4 (d^2/b^2 + d^2) + \left(\frac{1}{\sigma^2} - \frac{1}{\sigma^4} \frac{m}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}m}\right)^4 (m^2+d^2)(m^2)\Big). \tag{193}$$

The first two terms follow from using Lemma C.8 with only private samples (as they are not shifted and the same bound applies).

We can simplify, combining the first two terms and dropping lower-order terms (note that we assume $m \gg d$) to arrive at a form that will later be more convenient:

$$\le O\left(\left(\frac{1}{\sigma^2}(n+m) - \frac{1}{\sigma^4} \frac{m^2}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}m}\right)^4 (d^2/b^2 + d^2)\right). \tag{194}$$

□

Lastly, in order to bound $\mathbb{E}\left[\left\|M(X,y) - \hat{\beta}\right\|_2^2\right]$, we will need a bound on $\mathbb{E}\left[\left\|\hat{\beta} - \beta\right\|_2^2\right]$.

Lemma D.7. $\mathbb{E}\left[\left\|\hat{\beta} - \beta\right\|_2^2\right] \le O\left(\frac{d}{n+m/\kappa}\right)$, *where* $\kappa = m\tau^2/d + 1$.

Proof. Recall that $\hat{\beta}$ is the GLS estimator, where $\hat{\beta} = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} y$. Throughout the proof, we assume $d, n \gg d$, and that the matrix $X^\top \Sigma^{-1} X$ is full rank with constant probability.

Let $\Lambda \stackrel{\text{def}}{=} X^\top \Sigma^{-1} X$. Fixing β, and taking the expectation over X, y, we can simplify the expression as follows:

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\beta} - \beta\right\|_2^2\right] &= \mathbb{E}\left[\|(X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} y - \beta\|_2^2\right] \\
&= \mathbb{E}\left[\|\Lambda^{-1} X^\top \Sigma^{-1} y - \Lambda^{-1} \Lambda \beta\|_2^2\right] \\
&\le \mathbb{E}\left[\|\Lambda^{-1}(X^\top \Sigma^{-1}(X\beta + \eta) - \Lambda\beta)\|_2^2\right] \\
&= \mathbb{E}\left[\|\Lambda^{-1}(X^\top \Sigma^{-1}(X\beta + \eta) - X^\top \Sigma^{-1} X \beta)\|_2^2\right] \\
&= \mathbb{E}\left[\|\Lambda^{-1}(X^\top \Sigma^{-1} \eta)\|_2^2\right]
\end{aligned} \tag{195}$$

$$\mathbb{E}\left[\|\Lambda^{-1} X^\top \Sigma^{-1} \eta\|_2^2\right] = \mathbb{E}_X\left[\|\Lambda^{-1}\|_2^2 \mathbb{E}_\eta\left[\|X^\top \Sigma^{-1} \eta\|_2^2 \,|\, X\right]\right] \tag{196}$$

$$= \mathbb{E}_X\left[\|\Lambda^{-1}\|_2^2 \operatorname{tr}\left(X^\top \Sigma^{-1} \Sigma \Sigma^{-1} X\right)\right]$$

$$= \mathbb{E}_X\left[\|\Lambda^{-1}\|_2^2 \operatorname{tr}(\Lambda)\right], \tag{197}$$

where the second equality follows from the fact that $\mathbb{E}[\eta\eta^\top \mid X] = \Sigma$. Since $\|\Lambda^{-1}\|_2^2 = 1/\lambda_{\min}(\Lambda)^2$ and $\mathrm{tr}(\Lambda) \le d\lambda_{\max}(\Lambda)$,

$$\|\Lambda^{-1}\|_2^2 \mathrm{tr}(\Lambda) \le \frac{d\lambda_{\max}(\Lambda)}{\lambda_{\min}(\Lambda)^2}. \tag{198}$$

For Gaussian designs whitened by $\Sigma^{-1/2}$ one has, with constant probability,

$$\lambda_{\min}(\Lambda) = \Omega\big(n+m/\kappa\big), \qquad \lambda_{\max}(\Lambda) = O\big(n+m/\kappa\big). \tag{199}$$

Substituting Eq. 199 into Eq. 198 and taking expectations,

$$\mathbb{E}\big[\|\hat{\beta}-\beta\|_2^2\big] \le C\frac{d(n+m/\kappa)}{\big(n+m/\kappa\big)^2} = O\Big(\tfrac{d}{n+m/\kappa}\Big),$$

for an absolute constant $C>0$. This concludes the proof. □

Lemma D.8. *When $b=1/d$, $m+n=\Omega(d)$, and $\alpha=O(1)$, the expression:*

$$\sqrt{\mathbb{E}_{X,\eta}\mathbb{E}_{\hat{\beta}}\Big\|M(X,y)-\hat{\beta}\Big\|_2^2 \cdot \mathbb{E}_{X,\eta}\|X^\top\Sigma^{-1}X\mathbb{E}[\beta|X,\eta]-X^\top\Sigma^{-1}y\|_2^2} = O(1).$$

Proof. Let $\gamma(\tau) = d/\tau^2 + m/\sigma^2$. Note that when $\tau\to\infty$, $\gamma(\tau)$ approaches m/σ^2 (such that $m/\sigma^2 - \frac{1}{\gamma(\tau)} \to 0$ negating the effect of public samples), and when $\tau\to 0$, $\gamma(\tau)\to\infty$ (such that $m/\sigma^2 - \frac{1}{\gamma(\tau)} \to m/\sigma^2$ so that public and private samples contribute equally).

From Lemma D.5 and Lemma D.6, we derive the following:

$$\sqrt{\mathbb{E}_{X,\eta}\mathbb{E}_{\hat{\beta}}\Big\|M(X,y)-\hat{\beta}\Big\|_2^2 \cdot \sqrt{b^4\mathbb{E}\left[\left\|\tilde{\Lambda}^{-1}\right\|_2^4\right]\mathbb{E}\left[\|X^\top\Sigma^{-1}y\|_2^4\right]}} \tag{200}$$

We first simplify just the inner square root. Using the previous lemmas and applying the identity $\sqrt{x+y} \le C\cdot(\sqrt{x}+\sqrt{y})$:

$$\le b^2 O\left(\frac{1}{(\frac{1}{\sigma^2}N - \frac{1}{\sigma^4}\frac{m^2}{\gamma(\tau)}+b)^2}\right)\sqrt{O\left(\left(\frac{1}{\sigma^2}N - \frac{1}{\sigma^4}\frac{m^2}{\gamma(\tau)}\right)^4 (d^2/b^2+d^2)\right)} \tag{201}$$

$$\le b^2 O\left(\frac{1}{(\frac{1}{\sigma^2}N - \frac{1}{\sigma^4}\frac{m^2}{\gamma(\tau)}+b)^2}\right) O\left(\left(\frac{1}{\sigma^2}N - \frac{1}{\sigma^4}\frac{m^2}{\gamma(\tau)}\right)^2 (d/b+d)\right) \tag{202}$$

$$\le b^2 O\left(\frac{1}{(\frac{1}{\sigma^2}N - \frac{1}{\sigma^4}\frac{m^2}{\gamma(\tau)})^2}\right) O\left(\left(\frac{1}{\sigma^2}N - \frac{1}{\sigma^4}\frac{m^2}{\gamma(\tau)}\right)^2 (d/b+d)\right) \tag{203}$$

Substituting $b=1/d$ and making cancellations:

$$\le (d^2+d)/d^2 \le 1+1/d \tag{204}$$

Finally, we substitute into the original expression, combining with Lemma D.7 (recall that $\kappa = m\tau^2/d+1$ and $\alpha = O(1)$):

$$\sqrt{\mathbb{E}_{X,\eta}\mathbb{E}_{\hat{\beta}}\Big\|M(X,y)-\hat{\beta}\Big\|_2^2 \cdot \sqrt{b^4\mathbb{E}\left[\left\|\tilde{\Lambda}^{-1}\right\|_2^4\right]\mathbb{E}\left[\|X^\top\Sigma^{-1}y\|_2^4\right]}} \tag{205}$$

$$\le \sqrt{\left(O\left(\frac{d}{n+m/\kappa}\right)+\alpha^2\right)\cdot O(1+1/d)} = O(1) \tag{206}$$

since $\alpha^2 = O(1)$. □

Now, we are ready to prove the final result in Theorem 1.3. For this we invoke the upper bound on $\mathbb{E}[\sum_i Z_i]$ in Lemma D.1 and the lower bound of $\mathbb{E}[\sum_i Z_i] = \Omega(d)$ induced by Lemma D.2 and Lemma 200. This tells us that there exist positive constants t_1, t_2 and values m_0, n_0 such that for $m \geq m_0, n \geq n_0$, the following holds:

$$t_1 \cdot d \leq \mathbb{E}\left[\sum_{i\in[N]} Z_i\right] \leq t_2 \cdot \left(n\varepsilon\alpha + \alpha\sqrt{\frac{1}{\sigma^2}md - \frac{1}{\sigma^4}\left(\frac{m^2 d}{\frac{d}{\tau^2}+\frac{1}{\sigma^2}m}\right)}\right) \tag{207}$$

The upper bound interpolates between the private-sample-only regime when $\tau \to \infty$ and the all-samples regime when $\tau \to 0$ (no distribution shift). In particular, we can analyze the term inside the square root. It can be written as:

$$md - \frac{m^2\tau^2}{\kappa},$$

when $\sigma = 1$ and $\kappa = m\tau^2/d + 1$. Now, consider two regimes for the shift parameter τ:

- **Large shift regime:** Suppose $\kappa = \Omega\left(\frac{m\tau^2}{d}\right)$, which corresponds to the condition $\tau = \omega\left(\sqrt{d/m}\right)$. Then:

  $$\left|md - \frac{m^2\tau^2}{\kappa}\right| = o(1).$$

  Hence, the only dominant term on the right-hand side becomes $n\varepsilon\alpha$, yielding the bound:

  $$n \geq \frac{d}{\varepsilon\alpha}.$$

  Note, that $n = \Omega(d/\alpha^2)$ is implied by the statistical non-private linear regression lower bound. Thus, in all $n = \Omega(d/\varepsilon\alpha + d/\alpha^2)$.

- **Small shift regime:** Suppose $\tau = \mathcal{O}\left(\sqrt{d/m}\right)$. Then:

  $$md - \frac{m^2\tau^2}{\kappa} = \mathcal{O}(md),$$

  and therefore the right-hand side becomes:

  $$\mathcal{O}\left(n\varepsilon\alpha + \alpha\sqrt{md}\right),$$

  which matches the bound in the no-shift case. Thus, the sample complexity requirements in the small shift setting of $\tau = O(\sqrt{d/m})$ matches the no-shift setting.

- Finally, we note that in the large shift setting, when $\alpha \gtrsim \tau$, then the OLS estimator that only uses public samples already satisfies an accuracy of α in the ℓ_2 norm.

This completes the proof of our result in Theorem 5.2.