# Private Regression In Multiple Outcomes (PRIMO)

**Seth Neel***
Department of Technology and Operations Management
Harvard University
Boston, MA 02163
sneel@hbs.edu

## Abstract

We introduce a new differentially private regression setting we call *Private Regression in Multiple Outcomes* (PRIMO) inspired the common situation in the social and biomedical sciences where a data analyst wants to perform a set of $l$ regressions while preserving privacy, where in each of the regressions the covariates $X$ are shared, and each regression $i$ has a different vector of outcomes $y_i$. We show that while taking any one of the number of existing private linear regression techniques and naively applying it $l$ times leads to an increase in error of $\sqrt{l}$ for PRIMO over the standard linear regression setting, techniques based on sufficient statistics perturbation (SSP) Wang [2018] can be modified to yield greatly improved dependence on $l$ in a range of common parameter regimes. Our first key insight is that the data covariance matrix $X^T X$ is shared across the regressions, so when privately computing this covariance matrix is the main source of error we obtain PRIMO with no dependence on $l$ in the asymptotic error. In Section 4.2 via an equivalence from privately computing the $X^T Y$ term to the problem of private query release with low $l_2^2$ error, we adapt the geometric projection-based methods of Nikolov et al. [2013] for query release to the PRIMO setting. Under the assumption the labels $Y$ are public, the projection gives improved results over the Gaussian mechanism when $n < l\sqrt{d}$. In order to give high probability bounds that are required for analyzing the error of the regression, we give a high probability analysis of the error of the projection mechanism using a variant of the Hanson-Wright Inequality. To our knowledge this is the first application of query release algorithms to private regression. In Section 4.3 we analyze the complexity of our proposed algorithm, introduced several algorithmic tricks based on matrix decomposition that decrease the dependence of the running time on $l$. Finally in Section 4.4 we aim to speed up our method via sub-sampling, giving an analysis of sub-sampled SSP that makes heavy use of Matrix Chernoff bounds for sampling without replacement.

## 1 Introduction

Linear regression is one of the most fundamental statistical tools used across the applied sciences, for both inference and prediction. In genetics, polygenic risk scores Krapohl et al. [2018], Pattee and Pan [2020a] are computed by regressing phenotype (e.g. disease status) onto individual genomic data (SNPs) in order to identify genetic risk factors. In the social sciences, we might regress observed societal outcomes like income or marital status on a fixed set of covariates Agresti and Finlay [2009]. In many of these cases where the data records correspond to individuals, there are two aspects of the problem setting that often co-occur:

*personal website: https://www.hbs.edu/faculty/Pages/profile.aspx?facId=1326892

Preprint. Under review.

**Aspect 1.** The individuals may have a legal or moral right to privacy that has the potential to be compromised by their participation in a study.

**Aspect 2.** Often multiple regressions will be ran using the same set of individual characteristics across each regression with different outcomes, either within the same study or across many different studies.

Aspect 1 has been established as a legitimate concern through both theoretical and applied work. The seminal paper of Homer et al. [2008] showed that the presence of an individual in a genomic dataset could be identified given simple summary statistics about the dataset, leading to widespread concern over the sharing of the results of genomic analyses. In the machine learning setting, where what is being released is a model $w$ trained on the underlying data, there is a long line of research into so-called "Membership Inference Attacks" Hu et al. [2021], Shokri et al. [2016], which given a trained model are able to identify which points are in the training set. Since training set membership can itself be a sensitive quantity – it may reveal disease status in the genomic example, or income in the social science example – this is a potential privacy violation. Over the last decade, differential privacy Dwork and Roth [2014] has emerged as a rigorous solution to the privacy risk posed by Aspect 1. In the particular case of linear regression the problem of how to privately compute the optimal regressor has been studied in great detail, which we summarise in Subsection 3.

Aspect 2 has been studied extensively from the orthogonal perspective of multiple hypothesis testing, but until now has not been considered in the context of privacy. The problem of overfitting or "p-hacking" in the social and natural sciences has been referred to as the "statistical crisis in science" Gelman and Loken, and developing methods that quantify and mitigate the effects of overfitting has been the subject of much attention in the statistics and computer science communities Dwork et al. [2015a], Bassily et al. [2015], Korthauer et al. [2019]. Many methods in the genetic risk score literature attempt to explicitly combat overfitting Pattee and Pan [2020b], and a recent paper in the economics community points out that 2 of the most common datasets have been reused to predict many different outcome variables, leading to misleading estimates of statistical significance Heath [2022]. Given the ubiquity of Aspects 1 and 2, this raises an important question:

**When computing $i = 1 \ldots l$ distinct regressions with a common set of $X$'s and distinct $y_i$'s, what is the optimal accuracy-privacy tradeoff?**

Interestingly, at a technical level, the problem of multiple hypothesis testing is related to differential privacy. It has been shown that if each query (in this case regression which is a special class of query called an optimization query ) is computed subject to differential privacy, then we can obtain a provable tradeoff between the number of noisy query answers we provide about a dataset and the extent of overfitting that is possible Dwork et al. [2015a]. This gives a second motivation beyond privacy for our setting: even when the underlying data is not sensitive, our method can be viewed as a way to provably prevent overfitting when running multiple linear regressions.

### 1.1 Results

The primary contribution of this work is to introduce the novel PRIMO problem, and to provide a class of algorithms that trade off accuracy, privacy, and computation (Subsections 4.1, 4.2). In addition to introducing the PRIMO problem, to our knowledge we are the first to apply private query release methods to linear regression (Subsection 4.2), and to analyze sub-sampled private linear regression (Subsection 4.4). Our query release results make heavy use of an assumption that $Y$ is public and $X$ is private, and as such are only applicable to the "public label setting."

We will show that in a range of common parameter settings, our methods can obtain PRIMO at minimal cost. Bassily et al. [2014] shows that for private linear regression (PRIMO with $l = 1$), we have the following lower bound on the error:

$$f(w_{private}) - f(w_{opt}) \geq \min\{||\mathcal{Y}||^2, \frac{\sqrt{d}(||\mathcal{X}||^2||\mathcal{W}||^2 + ||\mathcal{X}||||\mathcal{W}||||\mathcal{Y}||}{n\epsilon}\} \quad (1)$$

Since this lower bound holds for the case when $l = 1$, it of course holds for the PRIMO setting where $l > 1$. Throughout the paper this lower bound will serve as a benchmark for the cost to accuracy of taking $l > 1$, and in settings in which the bounds for PRIMO match this lower bound we will say we have "PRIMO for Free."

Theorems 4.1, 4.3 imply that if $\alpha$ is the difference between mean squared error of the private estimator and the optimal estimator averaged over the $l$ regressions, then with high probability:

$$\alpha^2 = \tilde{O}\left(\frac{d||\mathcal{X}||^4||\hat{w}||^4}{n^2} + \min\left(\frac{||\hat{w}||^2\sqrt{d}||\mathcal{Y}||^2||\mathcal{X}||^2}{n}, \frac{||\hat{w}||^2 ld||\mathcal{Y}||^2||\mathcal{X}||^2}{n^2}\right)\right) \tag{2}$$

| Guarantees of ReuseCov | | | | |
|---|---|---|---|---|
| $\#\ l$ of regressions | $\alpha$ | cost$(l)$ | $\mathcal{M}$ | Privacy |
| $l > \frac{n}{\sqrt{d}}, n < \frac{\sqrt{d}||\mathcal{X}||^2||\mathcal{W}||^2}{||\mathcal{Y}||^2}$ | $\tilde{O}\left(\frac{\sqrt{d}||\mathcal{X}||^2||\hat{w}||^2}{n}\right)$ | 1 | Proj | Public $Y$ |
| $l > \frac{n}{\sqrt{d}}, n > \frac{\sqrt{d}||\mathcal{X}||^2||\mathcal{W}||^2}{||\mathcal{Y}||^2}$ | $\tilde{O}\left(\frac{||\hat{w}||d^{1/4}||\mathcal{Y}||||\mathcal{X}||}{\sqrt{n}}\right)$ | $\frac{n}{d^{1/4}}$ | Proj | Public $Y$ |
| $l < \min(\frac{n}{\sqrt{d}}, \frac{||\mathcal{X}||^2||\mathcal{W}||^2}{||\mathcal{Y}||^2})$ | $\tilde{O}\left(\frac{\sqrt{d}||\mathcal{X}||^2||\hat{w}||^2}{n}\right)$ | 1 | Gauss | Private $X, Y$ |
| $l \in (\frac{||\mathcal{X}||^2||\mathcal{W}||^2}{||\mathcal{Y}||^2}, \frac{n}{\sqrt{d}})$ | $\tilde{O}\left(\frac{||\hat{w}||\sqrt{ld}||\mathcal{Y}||||\mathcal{X}||}{n}\right)$ | $\sqrt{l}$ | Gauss | Private $X, Y$ |

Figure 1: Here cost$(l)$ is the ratio of the error for PRIMO given by Equation 2 to the lower bound in Equation 1, and $\mathcal{M}$ denotes the mechanism used to compute $X^TY$.

In Figure 1 we summarize the accuracy guarantees implied by Equation 2, where cost$(l)$ denotes the ratio of the error $\alpha$ to the lower bound in Equation 1. When cost$(l) = 1$ this corresponds to "PRIMO for Free", and when cost$(l) = \sqrt{l}$ our upper bound matches that of the naive private baseline (Equation 5).

The methods in this paper are a variant of "Sufficient Statistics Perturbation" Vu and Slavkovic [2009] that rely on perturbing $X^TX \approx \Sigma$ and $X^Ty \approx \Sigma_{Xy}$ separately, and then use these noisy estimates to compute the least squares estimator:

$$\tilde{w}_{i*} = \underbrace{(X^TX + E_1)^{-1}}_{\text{noisy covariance term } \tilde{\Sigma}} \times \underbrace{(X^TY_i + E_{2i})}_{\text{noisy association term } \tilde{\Sigma}_{Xy}} \tag{3}$$

In the error analysis for SSP adapted from Wang [2018] (Subsection 4.1) for each of the $i$ regressions we decompose the error of the private estimator $\tilde{w}_{i*}$ into the error attributable to the noise added to the covariance matrix $E_1$, and the noise added to the association term $E_{2i}$.

In Section 4.1 we adapt the previously proposed sufficient statistics perturbation (SSP) algorithm of Vu and Slavkovic [2009], Foulds et al. [2016] to the PRIMO setting. Via a novel accuracy analysis of SSP for the case when the privacy levels for $E_1, E_2$ differ (Theorem 4.1), we show that since the noisy covariance matrix is reused across the regressions (as it only depends on $X$), by allocating the majority of our privacy budget to computing this term, we are able to obtain PRIMO for Free when $l < \min(\frac{n}{\sqrt{d}}, \frac{||\mathcal{X}||^2||\mathcal{W}||^2}{||\mathcal{Y}||^2})$. When $l$ is sufficiently small, we obtain PRIMO for Free because the error term is dominated by the error in computing the noisy covariance matrix, which does not depend on $l$, rather than the error from computing the noisy association term.

Given that our PRIMO for Free result in Subsection 4.1 relies on taking $l$ small enough that the error is dominated by the covariance error term, it is natural to ask if, under parameter regimes where the error from the association term dominates, if we can obtain improved dependence of $\alpha$ on $l$ over the $\sqrt{l}$ given by the Gaussian Mechanism. This is the focus of Subsection 4.2, where we start by showing that computing $X^TY$ privately is an instance of releasing $dl$ "low sensitivity" queries Bassily et al. [2015]. Inspecting the dependence of $\alpha$ on the error $E_2$ in computing $X^TY$ (e.g. Equation 24) we see that in order to minimize the error $\alpha$ we want to bound $||E_2||^2$ with high probability; and so our focus should be on releasing queries with low mean squared error rather the conventional worst-case loss ($l_\infty$). Our Algorithm 2 is similar to the relaxed projection mechanism of Aydöre et al. [2021] except it projects onto a data domain that is in the $l_2$ ball rather than relaxing binary attributes to the hypercube. Via a refined analysis of this mechanism that obtains novel accuracy bounds by exploiting the geometric structure of our query class, we obtain the surprising result that it is possible to completely remove the dependence of the error in computing the noisy association term on $l$ and improve the dependence on $d$ by a factor of $d^{1/4}$ albeit at the cost of a factor of $\sqrt{n}$; if we assume that our labels $Y$ are public (Theorem 4.2).

Theorem 4.3 states the error for PRIMO of the variant of Algorithm 1 that uses this projection mechanism as a subroutine. Inspecting the accuracy bound, we see that when $l > \frac{n}{\sqrt{d}}$ but $\frac{n}{\sqrt{d}} < \frac{||\mathcal{X}||^2||\mathcal{W}||^2}{||\mathcal{Y}||^2}$ we also obtain PRIMO for Free. This is in contrast to when we obtained PRIMO for Free for small $l$ in Subsection 4.1. In this case it is because (i) $l$ is sufficiently large that the noisy association term is computed via the projection mechanism rather than the Gaussian Mechanism (Algorithm 2) and (ii) $n$ is sufficiently small such that the noisy covariance term dominates the error of the projection mechanism. It is an open question whether in the private $Y$ setting, there exists a solution to PRIMO that obtains non-trivial error, and improves upon the mechanism that computes $X^TY$ by adding Gaussian noise.

In Subsections 4.3, 4.4 we consider the computational complexity of our methods. In Subsection 4.3 we show that given the QR decomposition of the noisy covariance matrix and the SVD of the label matrix $Y$, standard techniques give a simple way to compute the private estimators for all $l$. While SSP (and its adaptive variants) are known to achieve optimal error among private regression algorithms Wang [2018], iterative algorithms like NoisySGD are often preferred in practice because of their lower cost per iteration. When $n > d > l$ cost of SSP variants like our ReuseCov algorithm, are dominated by the cost of forming the covariance matrix, which is $O(nd^2)$ – prohibitively large for large $n, d$. To address this shortcoming, in Section 4.4 we develop a sub-sampling based version of ReuseCov which estimates the covariance matrix using a random sample of $s$ points – reducing the computational cost to $O(sd^2)$. Analyzing the accuracy of this procedure requires the use of Matrix-Chernoff bounds for sampling without replacement developed in Tropp [2010], which we must integrate with the error analysis of SSP in the stye of Wang [2018].

## 2 Preliminaries

We start by defining the standard linear regression problem. Given $X \in \mathcal{X}^n \subset \mathbb{R}^{n\times d}, y_i \in \mathcal{Y}^n$, and parameter space $\mathcal{W}$, for $w \in \mathcal{W}$ let

$$f(w) := \frac{1}{n}\sum_{k=1}^{n}(w \cdot x_k - y_{ik})^2$$

be the linear regression objective, and denote by $w_{i*} = \arg\min_w f(w) = (X^TX)^{-1}X^Ty_i$ the ordinary least squares estimator (OLS). Let $f_\lambda(w) = \frac{1}{n}\sum_{k=1}^{n}(w\cdot x_k - y_{ik})^2 + \lambda||w||_2^2$ the ridge regression objective, and let $w_{i*}^{\lambda} = \arg\min_w f_\lambda(w) = (\frac{1}{n}X^TX + \lambda I_d)^{-1}\frac{1}{n}X^Ty_i$.

The definition of data privacy we use throughout is the popular $(\varepsilon, \delta)$-differential privacy introduced in Dwork et al. [2006]. We refer the reader the to Dwork and Roth [2014] for an overview of the basic properties of $(\varepsilon, \delta) - DP$ including closure under post-processing and advanced composition.

**Definition 1.** *Let $\mathcal{M} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{O}$ a randomized algorithm taking as input a dataset of n records. We say that $(X, Y) \sim (X', Y') \in (\mathcal{X} \times \mathcal{Y})^n$ if $\exists x, yx', y'$ such that $(X, Y) \cup \{x', y'\}/\{x, y\} = (X', Y')$. Then we say that $\mathcal{M}$ is $(\varepsilon, \delta)$-DP if $\forall (X, Y) \sim (X', Y'), o \subset \mathcal{O}$:*

$$\Pr[\mathcal{M}(X, Y) \in o] \leq e^{\varepsilon}\Pr[\mathcal{M}(X', Y') \in o] + \delta \tag{4}$$

In the less restrictive case where Equation 4 holds only over adjacent $X \sim X'$ with the same fixed $Y$, we say that we have differential privacy in the *public label* setting, which we consider in Section 4.2.

We now formalize the *Private Regression In Multiple Outcomes (PRIMO)* problem.

**Definition 2.** *PRIMO. Let $x_i \in \mathcal{X} \subset \mathbb{R}^d, y_{ij} \in \mathcal{Y}$, for $i = 1 \ldots n, j = 1 \ldots l$. Let $X_{n\times d}$ the matrix with $i^{th}$ row $x_i$, and let $Y_{n\times l}$ the matrix with $j^{th}$ column $y_j = (y_{1j}, \ldots y_{nj})$. The optimal solution $W^*$ to the PRIMO problem is*

$$W^* = \inf_{W\in\mathcal{W}^l\subset\mathbb{R}^{d\times l}}\frac{1}{n}||XW - Y||_F^2$$

Given a randomized algorithm $\mathcal{M} : (\mathcal{X} \times \mathcal{Y}^l)^n \to \mathcal{W}^l$, we say that $\mathcal{M}$ is an $(\alpha, \beta, \varepsilon, \delta)$ solution to the PRIMO problem if (i) $\mathcal{M}$ is $(\varepsilon, \delta)$-DP with respect to $(X, Y)$ (or just $X$ in the public label setting), and (ii) with probability $1 - \beta$ over $\tilde{W} \sim \mathcal{M}$:

$$\frac{1}{n}||X\tilde{W} - Y||_1^2 - \frac{1}{n}||XW^* - Y||_1^2 < \alpha$$

While $y_j$ will denote the vector of $n$ outcomes for the $j^{th}$ outcome, throughout $y^i$ will denote the vector of $l$ outcomes corresponding to individual $i \in [n]$.

The most basic $(\varepsilon, \delta)$-mechanism is the Gaussian mechanism, which we make extensive use of throughout the paper.

**Lemma 1.** *GaussMech$(\varepsilon, \delta, \Delta)$ Dwork and Roth [2014] Let $f : \mathcal{X}^n \to \mathbb{R}^d$ an arbitrary d-dimensional function, and define it's sensitivity $\Delta_2(f) = \sup_{X \sim X'} ||f(X) - f(X')||_2$, where $X \sim X'$ are datasets that differ in exactly one element. Then the Gaussian mechanism releases $f(X) + \mathcal{N}(0, \sigma^2)$, and is $(\varepsilon, \delta)$-differentially private for $\sigma \geq c(\varepsilon, \delta)\Delta_2(f)$, where $c(\varepsilon, \delta) = \sqrt{2\log(1.25/\delta}/\varepsilon$.*

Throughout the remainder of the paper, and particularly in the proofs in the Appendix, we will make heavy use of common matrix and vector norms. For a vector $v \in \mathbb{R}^d$ and matrix $A \in \mathbb{R}^{d \times d}$, $||v||_A^2 := v^t A v$, $||A||_2^2 = \lambda_{\max}(A^T A)$, $||v||_2^2 = \sum_{k=1}^d v_k^2$, $||A||_F^2 = \sum_{i,j \in [d]} a_{ij}^2$.

## 3 Related Work

**Private Linear Regression.** Private linear regression is well-studied under a variety of different assumptions on the data generating distributions and parameter regimes. Typically analysis of private linear regression is done either under the fully agnostic setting where only parameter bounds $||\mathcal{X}||, ||\mathcal{Y}||, ||\mathcal{W}||$ are assumed, under the assumption of a fixed design matrix and $y$ generated by a linear Gaussian model (the so-called realizable case), or under the assumption of a random design matrix Milionis et al. [2022]. In this paper we focus on the first fully agnostic setting, because in our intended applications within the social and biomedical sciences in general we neither have realizability or Gaussian covariates. In the fully agnostic setting Wang [2018] provides a comprehensive survey of existing private regression approaches and bounds, including proposing a new adaptive technique.

Broadly speaking, techniques for private linear regression fall into $4$ classes, sufficient statistics perturbation (SSP) Vu and Slavkovic [2009], Foulds et al. [2016], Objective Perturbation (ObjPert) Kifer et al. [2012], Posterior sampling Dimitrakakis et al. [2013], and privatized (stochastic) gradient descent (NoisySGD) Chaudhuri et al. [2011]. The methods in this paper are a sub-class of SSP-based methods, which correspond to Algorithm 1 where $l = 1$.

In the fully agnostic case, there are two regimes for private linear regression, each with upper bounds on empirical risk achieved by the algorithms above, and corresponding lower bounds stated below. The regimes depend on how well-conditioned the covariance matrix $X^T X$ is. Letting $\alpha^*$ be the inverse condition number we have the following lower boundsBassily et al. [2014]:

- When $\alpha^* \geq \frac{d^1.5(||\mathcal{X}||||\mathcal{W}||_|\mathcal{Y}||}{n||\mathcal{X}||||\mathcal{W}||\varepsilon}$, then:

$$f(\hat{w}) - f(w^*) \geq \min\{||\mathcal{Y}||^2, \frac{d^2(||\mathcal{X}||||\mathcal{W}|| + ||\mathcal{Y}||)^2}{n^2\alpha^*\varepsilon^2}\}$$

- We always have (even when $X^T X$ is ill-conditioned):

$$f(\hat{w}) - f(w^*) \geq \min\{||\mathcal{Y}||^2, \frac{\sqrt{d}(||\mathcal{X}||^2||\mathcal{W}||^2 + ||\mathcal{X}||||\mathcal{W}||||\mathcal{Y}||}{n\varepsilon}\}$$

Two techniques, NoisySGD and ObjPert both achieve the minimax lower bounds in both settings, although in order to achieve the minimax rates their hyperparameters depend on the unknown $||\mathcal{W}||$ and $\alpha^*$, and so Wang [2018] proposes two adaptive methods that are based on Sufficient Statistics Perturbation, and are able to achieve optimal bounds in both settings. In Subsections 4.1,4.4 we state our theoretical results under the ill-conditioned setting as it is the most general, although analogous results hold in the well-conditioned setting as well.

Now given any $(\varepsilon, \delta)$-DP algorithm for computing $w_j$ privately, we can use it as a sub-routine to solve PRIMO by simply running it $l$ times to compute each row of $W$. Hence by running any of the optimal algorithms $l$ times with parameters $\varepsilon' \approx \varepsilon/\sqrt{l}, \delta' \approx \delta/l$, by advanced composition for differential privacy Dwork and Roth [2014] we can achieve, subject to $(\varepsilon, \delta)$-DP:

$$\alpha = \tilde{O}(\frac{\sqrt{ld}(||\mathcal{X}||^2||\mathcal{W}||^2 + ||\mathcal{X}||||\mathcal{W}||||\mathcal{Y}||}{n\varepsilon}) \tag{5}$$

So for a fixed privacy budget $\varepsilon$, this naive baseline is a factor of $\sqrt{l}$ worse than in the standard private regression setting where $l = 1$.

**Query Release.** In Subsection 4.2 we show that privately computing the association term $\frac{1}{n}X^TY$ is equivalent to the problem of differentially privately releasing a set of $l \cdot d$ low-sensitivity queries Bassily et al. [2015]. In the case where $X, Y$ are both private, this corresponds to releasing a subset of $l \cdot d$ 2-way marginal queries over $d+l$ dimensions, which is well-studied Dwork et al. [2015b], Thaler et al. [2012], Ullman and Vadhan [2011]. Theorem 5.7 in Dwork et al. [2015b] gives a polynomial time algorithm based on relaxed projections that achieves mean squared error $\tilde{O}(n\sqrt{d+l})$, which matches the best known information theoretic upper bound Dwork et al. [2015b], although there is a small gap to the existing lower bound $\min(n, (d+l)^2)$. This relaxed projection algorithm outperforms the Gaussian Mechanism when $n\sqrt{d+l} < ld \implies n < \frac{dl}{\sqrt{d+l}} \implies d\sqrt{d+l} > n$. Since the mean squared error of Algorithm 1 that used this projection as a subroutine is at least $\frac{d\sqrt{d+l}}{n}$, this means that in the regime where the projection outperforms the Gaussian Mechanism, we do not achieve mean squared error $< 1$ in our regression. However, in the less restrictive but still practically relevant setting where the labels are public, we are able to obtain greatly improved results by using a projection-based method instead of the Gaussian Mechanism, as summarised in Figure 1 and presented in Subsection 4.2.

**Linear Queries under $l_\infty$-loss.** Beyond 2-way marginals, the problem of privately releasing large numbers of linear queries (Definition 3) has been studied extensively. It is known that the worst case error is bounded by $\min(\frac{\sqrt{\log(|\mathcal{Q}|)(\log|\mathcal{X}|\log(1/\delta))^{1/4}}}{\sqrt{n\epsilon}}, \frac{\sqrt{|\mathcal{Q}|\log(1/\delta)}}{\epsilon n})$. The first term, which dominates in the so-called low-accuracy or "sparse" (Nikolov et al. [2013]) regime, is achieved by the *PrivateMultiplicativeWeights* algorithm of Hardt and Rothblum [2010], which is optimal over worst case workloads Bun et al. [2013]. However, this algorithm has running time exponential in the data dimension, which is unavoidable Vadhan [2017] over worst case $\mathcal{Q}$. The second term, which dominates for $n \gg \frac{|\mathcal{Q}|}{\log|\mathcal{Q}|\log|\mathcal{X}|^{1/4}}$, the "high accuracy" regime, is achieved by the simple and efficient Gaussian Mechanism Dwork and Roth [2014], which is also optimal over worst-case sets of queries $\mathcal{Q}$ Bun et al. [2013].

**Linear Queries under $l_2$-loss.** For the $l_2$ error, in the high accuracy $n \gg |\mathcal{Q}|$ regime the factorization mechanism achieves error that is exactly tight for any workload of linear queries $\mathcal{Q}$ up to a factor of $\log(1/\delta)$, although it is not efficient (Theorem Edmonds et al. [2019]). In the low-accuracy regime, the algorithm of Nikolov et al. [2013] that couples careful addition of correlated Gaussian noise (akin to the factorization mechanism) with an $l_1$-ball projection step achieves error within log factors of what is (a slight variant) of a quantity known as the *hereditary discrepancy* $\text{opt}_{\epsilon,\delta}(A, n)$ (Theorem Nikolov et al. [2013]). This quantity is a known lower bound on the error of any $(\epsilon, \delta)$ mechanism for answering linear queries Muthukrishnan and Nikolov [2012], and so the upper bound is tight up to log factors in $|\mathcal{Q}|, |\mathcal{X}|$. Theorem 21 in Nikolov et al. [2013] analyzes the simple projection mechanism that adds independent Gaussian noise and projects rather than first performing the decomposition step that utilizes correlated Gaussian noise, achieving error $O(nd\log(1/\delta)\sqrt{\log|X|}/\epsilon)$, which matches the best known (worst case over $\mathcal{Q}$) upper bound for the sparse $n < d$ case Gupta et al. [2011]. In our Theorem 4.2 we give such a universal upper bound, rather than one that depends on the hereditary discrepancy of the matrix $Y$. While the bound can of course be improved for a specific set of outcomes $Y$ by the addition of the decomposition step to the projection algorithm, we omit this step in favor of a simpler algorithm with more directly comparable bounds to existing private regression algorithms.

**Low-sensitivity queries.** When $Y$ is public, computing $\frac{1}{n}X^Ty$ is not an instance of releasing the answers to 2-way marginal or even linear queries, but is an instance of releasing the answers to $dl$

"low-sensitivity" queries Bassily et al. [2015], or queries that are not necessarily an average over points in the dataset like linear queries, but still have the property that changing any single point in the data set only changes the magnitude of the outcome by $O(1/n)$. For the general case of low-sensitivity queries, the Gaussian Mechanism achieves the same $l_\infty, l_2$ error, and in the $l_\infty$ sparse case the *MedianMechanism* of Blum et al. [2011] achieves error roughly $\tilde{O}(\frac{\log^2(|\mathcal{Q}|)\sqrt{\log|\mathcal{X}|}}{\sqrt{n}\epsilon})$ Dwork et al. [2010]. While less is known about the optimal $l_2$ error for arbitrary low-sensitivity queries, it clear that geometric techniques based on factorization and projection do not (at least obviously) apply, since there is no corresponding notion of the query matrix $A_\mathcal{Q}$.

While the information-theoretic $l_\infty$ or $l_2$ error achievable for linear queries is well-understood Edmonds et al. [2019], Nikolov et al. [2013], Bun et al. [2013], as synthetic data algorithms like *PrivateMultiplicativeWeights* and *MedianMechanism*, or the factorization or projection mechanisms are in general inefficient, there are many open problems pertaining to developing efficient algorithms for specific query classes, or heuristic approaches that work better in practice. Examples of these approaches along the lines of the factorization mechanism McKenna et al. [2018, 2019], efficient approximations of the projection mechanism Aydöre et al. [2021], Dwork et al. [2015b], and using heuristic techniques from distribution learning in the framework of iterative synthetic data algorithms Yoon et al. [2019], Torkzadehmahani et al. [2020], Beaulieu-Jones et al. [2019], Neel et al. [2019].

**Projection mechanisms and Algorithm 2.** Since we make heavy use of these projection mechanisms in Subsection 4.2 in the *public label* setting, we elaborate on the difference between the existing methods and our Algorithm 2 below.

The most technically related work to our Algorithm 2 is Aydöre et al. [2021], which is itself a variant of the projection mechanism of Nikolov et al. [2013]. There are several key differences in our analysis and application:

- The projection in Nikolov et al. [2013] is designed for linear queries over a discrete domain, and runs in time polynomial in the domain size. Algorithm 2 allows continuous $\mathcal{X}, \mathcal{Y}$ and runs in time polynomial in $n, d, l$. Like the relaxed projection mechanism Aydöre et al. [2021], when our data is discrete we relaxing our data domain to be continuous in order to compute the projection more efficiently. Unlike in their setting which attempts to handle general linear queries, due to the special structure of our queries our projection can be computed in polynomial time via linear regression over the $l_2$ ball, as opposed to solving a possibly non-convex optimization problem.
- Moreover, due to this special structure we can use the same geometric techniques as in Nikolov et al. [2013] to obtain theoretical accuracy guarantees (Theorem 4.2). Nikolov et al. [2013] bound the accuracy of their mechanism with respect to the *expected* mean squared error. Due to the necessity of analyzing the error of private regression with a high probability bound, rather than expected error, we prove a high probability error bound for the accuracy of the projection mechanism (Theorem 4.2). This requires application of the Hanson-Wright Inequality for anisotropic sub-Gaussian variables Vershynin [2019].

**Sub-sampled Linear Regression.** In Subsection 4.4 we analyze SSP where we first sub-sample a random set of $s$ points without replacement, and use this sub-sample to compute the noisy covariance matrix. Sub-sampled linear regression has been studied extensively absent privacy, where it is known that uniform sub-sampling is sub-optimal in that it produces biased estimates of the OLS estimator, and performs poorly in the presence of high-leverage points Derezinski et al. [2018]. To address these shortcomings, techniques based on leverage score sampling Drineas et al. [2006], volume-based sampling Derezinski and Warmuth [2017]Derezinski et al. [2018], and spectral sparsification Lee and Sun [2015] have been developed. Crucially, in these methods the probability of a point being sub-sampled is data-dependent, and so they are (not obviously) compatible with differential privacy.

## 4 Methods

In Subsection 4.1 we develop our SSP-variant ReuseCov as a solution to the PRIMO problem. We give an error analysis of SSP for private linear regression where the noise levels for the covariance and association terms differ, and use this to examine when the asymptotic error of PRIMO matches the error of private linear regression ("PRIMO for Free"). In Subsection 4.2 we focus on the noisy association term, and give a reduction to the problem of private release of low-sensitivity queries. We show that for sufficiently large $l$, computing the association term via projection Algorithm 2 gives a polynomial improvement in error over the Gaussian mechanism when $l$ is sufficiently large. In Subsection 4.3 we address the computational complexity of Algorithm 3. Finally, in Subsection 4.4 we address the high computational cost of computing the full covariance matrix $X^T X$ in $O(nd^2)$, and show how sub-sampling the covariance matrix allows us to trade-off computation with accuracy.

### 4.1 ReuseCov

**Algorithm 1** Input: $s, \lambda, \mathcal{X} \in \mathcal{X}^n \subset \mathbb{R}^{d\times n}$, $Y = [y_1, \dots y_l] \in \mathcal{Y}^{l\times n}$, privacy params: $\epsilon, \delta$

$\lambda -$ ReuseCov

1: Draw $E_1 \sim N_{d(d+1)/2}(0, \sigma_1^2)$, where $\sigma_1 = \frac{1}{s}2\sqrt{2\log(2.5/\delta)}||\mathcal{X}||^2/\epsilon$
2: Compute $\hat{I} = (\frac{1}{n}X^T X + E_1 + \lambda I)^{-1}$
3: Draw $\hat{v} = [\hat{v}_1, \dots \hat{v}_l] \sim \mathcal{M}(\epsilon/2, \delta/2, X, Y)$
4: **for** $i = 1 \dots l$
5: Set $\hat{w}_i = \hat{I}\hat{v}_i$
6: **end for**
7: Return $\hat{W} = [\hat{w}_1, \dots \hat{w}_l]$

In the analysis of the ridge regression variant of SSP, which corresponds to Algorithm 1 with $l = 1$, Equation 13 in Wang [2018] shows w.p. $1 - \rho$:

$$f(\tilde{w}_\lambda) - f(w_{i*}) = \tilde{O}\left(\underbrace{\frac{d}{(\lambda + \lambda_{\min})\epsilon^2}||\mathcal{X}||^2||\mathcal{Y}||^2}_{\text{association error term}} + \underbrace{\frac{d}{(\lambda + \lambda_{\min})\epsilon^2}||\mathcal{X}||^4||\mathcal{W}||^2}_{\text{covariance error term}} + \underbrace{\lambda||\mathcal{W}||^2}_{\text{error due to ridge penalty}}\right) \tag{6}$$

Inspecting these terms, we see that:

**Insight 4.0.1.** *When $||\mathcal{X}||||\mathcal{W}|| \gg ||\mathcal{Y}||$ the error is dominated by the cost of privately computing $X^T X$, rather than privately computing $X^T y$.*

We now turn back to our PRIMO setting, and imagine independently applying SSP to solve each of our $l$ regression problems. By the lower bounds in Bassily et al. [2014], given a fixed privacy budget this incurs at least a $\sqrt{l}$ multiplicative blow up in error. However, we notice that when our private linear regression subroutine is an SSP variant, this naive scheme, of running $l$ independent copies of SSP is grossly wasteful.

**Insight 4.0.2.** *Since the $X$ matrix is shared across the different regression, we can simply compute our noisy estimate of $X^T X$ once, and then share that noisy covariance matrix across all of our regressions.*

Combining insights 1 and 2, we present Algorithm 1, where $\mathcal{M}$ can be any $(\epsilon/2, \delta/2)$-DP algorithm for estimating $X^T Y$.

In line 2 we take $\epsilon_1 = \epsilon/2$, and $\epsilon_2 = \tilde{O}(\frac{\epsilon}{\sqrt{l}})$, allocating most of our privacy budget for the “harder" one-shot task of computing $X^T X$, and thus necessarily adding more noise to the association term. By advanced composition Dwork and Roth [2014] this ensures $(\epsilon, \delta)$ privacy overall, and propagating these noise terms into Equation 6 gives:

**Theorem 4.1.** *With $\mathcal{M}$ = GaussMech$(\epsilon/2, \delta/2, \Delta = \frac{1}{n}\sqrt{l}||\mathcal{X}||||\mathcal{Y}||)$, Algorithm 1 is an $(\alpha, \rho, \epsilon, \delta)$ solution to the PRIMO problem with*

$$\alpha = \tilde{O}\left(||\hat{w}||\sqrt{\frac{d||\mathcal{X}||^4||\hat{w}||^2}{n^2} + \frac{ld||\mathcal{Y}||^2||\mathcal{X}||^2}{n^2}}\right), \tag{7}$$

*where $||\hat{w}||^2 = \frac{1}{l}||W^*||_F^2$, and $\tilde{O}$ omits terms polynomial in $\frac{1}{\epsilon}, \log(1/\delta), \log(1/\rho)$.*

While we defer the formal proof to the Appendix, the proof follows the analysis in Wang [2018] that uses Lemma 8 (essentially the Woodbury matrix inversion formula) to expand the difference $||\tilde{w}_{i*} - w_{i*}||_{X^TX}$ into the form in Equation 23 in the Appendix:

$$O\left(\underbrace{||E_1 w_{i*}||^2_{(X^TX+\lambda I)^{-1}}}_{\text{covariance error}} + \underbrace{\lambda^2||w_{i*}||^2_{(X^TX+\lambda I)^{-1}}}_{\text{regularization error}} + \underbrace{||E_2||^2_{(X^TX+\lambda I)^{-1}}}_{\text{assocation error}}\right),$$

The bound follows from using a Johnson-Lindenstrauss type lemma (Lemma 11) to bound the $||E1||, ||E_2||$ terms, substituting in the noise levels in Algorithm 1, and optimizing over $\lambda$.

Inspecting Theorem 4.1, we see that when $l < \min(\frac{n}{\sqrt{d}}, \frac{||\mathcal{X}||^2||\mathcal{W}||^2}{||\mathcal{Y}||^2})$, Algorithm 1 with $\mathcal{M}$ the Gaussian mechanism achieves error $\tilde{O}\left(\frac{\sqrt{d}||\mathcal{X}||^2||\hat{w}||^2}{n}\right)$. Since this matches the lower bound in Equation 1 for a single private regression in this case we say that we have achieved "PRIMO for Free."

### 4.2 Private Query Release

The improvement of Algorithm 1 over the naive PRIMO baseline in the previous sections make heavy use of the fact that the asymptotic error is dominated by the covariance error term. In the case when $\mathcal{M}$ is the Gaussian mechanism, if we are in the regime where the association term dominates, ReuseCov still incurs a $\sqrt{l}$ multiplicative factor in the error term, which does not improve over the baseline. In this section we show that via a reduction from computing the association term to private query release via synthetic data, if we are only concerned with privacy in the $X$'s, the *public label* setting, we can obtain improved bounds for PRIMO over the naive baseline and ReuseCov for certain ranges of $(n, l, d)$.

Let us reconsider the problem of privately computing the association term $X^TY$ where $Y \in \mathcal{Y}^{n \times l}$, and $X \in \mathcal{X}^n \subset \mathbb{R}^{d \times n}$, and $Y$ is considered public, so our data release needs only be DP in $X$. Define the query $q_{kj} : \mathcal{X}^n \to [0, 1] = \frac{1}{n}x_k^T y_j$ be the function that takes input $X$, and outputs the $k^{th}$ element of $\frac{1}{n}X^T y_j$. Or if $e_k \in \mathbb{R}^d$ is the $k^{th}$ basis vector, $q_{kj}(X) = \frac{1}{n}e_k X^T y_j$. It will be convenient for us to write $q(\cdot)$ as a single inner product. Let $\text{vec}(X) = (x_{11}, \ldots, x_{1n}, x_{21}, \ldots x_{d1}, \ldots x_{dn}) \in \mathbb{R}^{nd}$, and given $y = \text{vec}(X)$, let $\text{mat}(y) = X$. Denote by $c_{kj} \in \mathbb{R}^{nd}$ the vector that has all zeros except in positions $(k-1)n + 1, \ldots kn$ it contains $\frac{1}{n}y_{1j}, \ldots \frac{1}{n}y_{nj}$. Then it is clear that $q_{kj}(X) = c_{kj}^T \text{vec}(X)$, so if we let $C \in \{0, 1\}^{dl \times dn}$ be the matrix with row $kj \in [dl]$ equal to $c_{kj}$:

$$\frac{1}{n}X^TY = C \cdot \text{vec}(X)$$

Given $x_{kj} \in X$, let $m = (k-1)n + j$ be the corresponding column of $C$, and let $c^m$ denote this column. Then $||c^m||_2 = \frac{1}{n}||(y_{1j}, \ldots y_{lj})||_2 = \frac{1}{n}||y^j||_2 \leq \frac{\sqrt{l}}{n}$ which will be useful in a moment.

The class $= (q_{kj})$ fall into what is known as a low-sensitivity queries Bassily et al. [2015], because changing one row of $i$ of $X$, changes at most one entry $\frac{1}{n}x_{ki}y_{ji}$ of $x_k^t y_j$, which changes the result by at most $\frac{||\mathcal{X}||||\mathcal{Y}||}{n}$. It is important that $Y$ is considered public here, as the queries are defined as a function of $Y$. If we wanted to guarantee privacy in the $Y$'s, we would have to define our query class as taking input $(X, Y) \in (\mathcal{X} \times \mathcal{Y}^l)^n$, making the dimension of our data $d + l$ rather than $d$. This blowup in dimension ends up negating any gains in accuracy from the synthetic data method, in the regime where the linear regressor obtains non-trivial accuracy, as discussed in Subsection 3. We now present our projection-based subroutine for privately computing $X^TY$.

**Algorithm 2** Input: $X \in \mathcal{X}^n \subset (\mathbb{R}^d)^n$, $Y = [y_1, \ldots y_l] \in \mathcal{Y}^{l \times n}$, privacy params: $\epsilon, \delta$,

Inner Product Projection Mechanism

.

1: Formulate $C = C(Y) \in [0,1]^{dl \times dn}$, $\text{vec}(X)$.
2: Let $r = \frac{c(\epsilon,\delta) \sup_{i \in [l]} ||y^i||_2 \cdot ||\mathcal{X}||_2}{n}$
3: Sample $w \sim N(0,1)^{dl}$
4: Let $\tilde{g} = C\text{vec}(X) + rw$
5: Let $\hat{g} = \text{argmin}_{g \in K} ||g - \tilde{g}||_2^2$, where $K = C(n||\mathcal{X}||_2^2 B_1)$.

Outputting $\tilde{g}$ in Line 4 corresponds to the Gaussian Mechanism, and has mean squared error $r^2 = O(\frac{l||\mathcal{Y}||||\mathcal{X}||^2}{n^2})$. Theorem 4.2 shows that the projection in Line 5 reduces error by a multiplicative factor of $\frac{n}{l\sqrt{d}}$. The crux of the proof is Lemma 2 which quantifies the reduction of error achieved by the projection step, which is the main workhorse behind the results in Nikolov et al. [2013], but has been folklore in the statistics community since at least Raskutti et al. [2009].

**Lemma 2** (Raskutti et al. [2009])**.** *Let $K \subset \mathbb{R}^d$ be a symmetric convex body, let $g \in K$, and $\tilde{g} = g + w$ for some $w \in \mathbb{R}^d$. Then if $\hat{g} = \text{argmin}_{g' \in K} ||\tilde{g} - g'||_2^2$, then*

$$||\hat{g} - g||_2^2 \leq \min\{4||w||_2^2, 4||w||_{K^\circ}\}$$

**Theorem 4.2.** *Let $\mathcal{M}$ denote Algorithm 2. $\mathcal{M}$ is $(\epsilon, \delta)$ differentially private, and if $\hat{g} \sim \mathcal{M}$, then with probability $1 - \rho$ :*

$$\frac{1}{dl}||g - \tilde{g}||_2^2 = O\left(c(\epsilon,\delta)\sqrt{\log(2/\rho)}\frac{||\mathcal{Y}||^2||\mathcal{X}||^2}{n\sqrt{d}}\right)$$

*Proof.* For $c(\epsilon, \delta) = \sqrt{2\log(1.25/\delta)}\epsilon$, $\mathcal{M}$ is $(\epsilon, \delta)$ differentially private by the Gaussian Mechanism and post-processing Dwork and Roth [2014]. So we focus on the high probability accuracy bound. By Lemma 4 from Nikolov et al. [2013] we have that:

$$||\hat{g} - g||_2^2 \leq 4||rw||_{K^\circ} = r \sup_{x \in K} x \cdot w \tag{8}$$

Since $K \subset (n||\mathcal{X}||^2)CB_1$, and using the fact that the $l_2$ norm is self-dual, we have:

$$||\hat{g} - g||_2^2 \leq 4r||w||_{K^\circ} \leq 4r||\mathcal{X}||\sqrt{n} \sup_{z \in B_1} (Cz) \cdot w = 4r||\mathcal{X}||\sqrt{n}||C^T w||_2 \tag{9}$$

So in order to bound $||\hat{g} - g||_2^2$ with high probability it suffices to bound $||C^T w||_2$ with high probability. This is the content of the Hanson-Wright Inequality for anisotropic random variables Vershynin [2019].

**Lemma 3** (Vershynin [2019])**.** *Let $C^T$ an $m \times n$ matrix, and $X \sim N(0,1)^n \in \mathbb{R}^n$. Then for a fixed constant $c > 0$:*

$$\mathbb{P}[\mid ||C^T X|| - ||C||_F \mid > t] \leq 2exp(\frac{-ct^2}{||C||_F})$$

Lemma 3 shows that $||C^T w||_2 = O(||C||_F \sqrt{\log(2/\rho)})$ with probability $1 - \rho$. Since $||c^k||_2 = \frac{||y^i||}{n}$ for every column $k = ij, i \in [n], j \in [d]$ of $C$, $||C||_F = \frac{\sqrt{d\sum_{i=1}^n ||y^i||^2}}{n}$.

Plugging in the value of $r$ gives, with probability $1 - \rho$:

$$\frac{1}{dl}\mathbb{E}[||\hat{g} - g||_2^2] = O\left(\frac{1}{dl} \cdot \sqrt{n}||\mathcal{X}||\frac{\sqrt{d\sum_{i=1}^n ||y^i||^2}}{n}\sqrt{\log(2/\rho)} \cdot \frac{c(\epsilon,\delta)||\mathcal{X}|| \sup_i ||y^i||_2}{n}\right) \tag{10}$$

$$= O\left(\frac{c(\epsilon,\delta)\sqrt{\log(2/\rho)}\sqrt{\frac{1}{n}\sum_{i=1}^n ||y^i||^2} \sup_i ||y^i||_2 ||\mathcal{X}||^2}{nl\sqrt{d}}\right), \tag{11}$$

Since both terms involving $Y$ in the numerator are $\leq \sqrt{l}||\mathcal{Y}||$ the bound follows. We note that since $Y$ is public, in practice we can compute these terms rather than using $||\mathcal{Y}||$, the worst case bound. ☐

We note that the mean squared error of the Gaussian mechanism without the projection is $O(r^2) = O(\frac{l||\mathcal{X}||^2||\mathcal{Y}||^2}{n^2})$, which for $l\sqrt{d} \gg n$ is strictly larger than the error of the projection mechanism. We also note that the bound in Theorem 4.3 is strictly better than the error given by applying the Median Mechanism algorithm of Blum et al. [2011] for low-sensitivity queries, which is tailored for $l_\infty$ error, and which also requires discrete $\mathcal{X}$. For example, when $\mathcal{Y} = \{0,1\}, \mathcal{X} = \{0,1\}^d$, then $||\mathcal{Y}|| = 1, ||\mathcal{X}|| = \sqrt{d}$, and so Theorem 4.2 gives a bound of $\tilde{O}(\frac{\sqrt{d}}{n})$, whereas the Median Mechanism gives $\tilde{O}(\frac{d^{2/3}\log(dl)^2}{n^{2/3}})$ for the mean squared error.

We now state the accuracy guarantees of Algorithm 1 with $\mathcal{M}$ given by Algorithm 2.

**Theorem 4.3.** *Let $\mathcal{A}$ denote the label-private variant of Algorithm 1, where $\mathcal{M}$ is Algorithm 2 with privacy parameters $(\epsilon/2, \delta/2)$. Then $\mathcal{A}$ is an $(\alpha, \rho, \epsilon, \delta)$ solution to the PRIMO problem with*

$$\alpha = \tilde{O}\left(||\hat{w}||\sqrt{\frac{d||\hat{w}||^2(||\mathcal{X}||^4)}{n^2} + \frac{\sqrt{d}||\mathcal{Y}||^2||\mathcal{X}||^2}{n}}\right),$$

*where $||\hat{w}||^2 = \frac{1}{l}||W^*||_F^2$, and $\tilde{O}$ omits terms polynomial in $\frac{1}{\epsilon}, \log(1/\delta), \log(1/\rho)$.*

*Proof.* We start with our usual expansion of $f(\hat{w}_i) - f(w_{i*})$, up until Equation 24, we have with probability $1-\rho$ for every $i \in [l]$:

$$n \cdot f(\hat{w}_i) - f(w_{i*}) = \tilde{O}\left(\frac{d}{\lambda_{\min}+\lambda}||w_{i*}||^2(||\mathcal{X}||^4/\epsilon^2)\log(2d^2/\rho) + \lambda||w_{i*}||^2 + \frac{1}{\lambda_{\min}+\lambda}||E_{2i}||_2^2\right) \tag{12}$$

Aggregating over $i$ and rearranging gives:

$$\frac{n}{l}\sum_{i=1}^{l} f(\hat{w}_i) - f(w_{i*}) = \tilde{O}\left(\frac{d}{\lambda_{\min}+\lambda}||\hat{w}||^2(||\mathcal{X}||^4/\epsilon^2)\log(2d^2/\rho) + \lambda||\hat{w}||^2 + \frac{1}{\lambda_{\min}+\lambda}\frac{1}{l}\sum_{i=1}^{l}||E_{2i}||_2^2\right), \tag{13}$$

where $||\hat{w}||^2 = \frac{1}{l}\sum_{i=1}^{l}||w_{i*}||^2 = \frac{1}{l}||W^*||_F^2$. Then by Theorem 4.2, we have with $\frac{1}{l}\sum_{i=1}^{l}||E_{2i}||_2^2 = O\left(c(\epsilon,\delta)\sqrt{\log(2/\rho)}n\sqrt{d}||\mathcal{Y}||^2||\mathcal{X}||^2\right)$ with probability $1-\rho$. So with probability $1-2\rho$, we have $\frac{n}{l}\sum_{i=1}^{l} f(\hat{w}_i) - f(w_{i*}) =$

$$\tilde{O}\left(\frac{d}{\lambda_{\min}+\lambda}||\hat{w}||^2(||\mathcal{X}||^4/\epsilon^2)\log(2d^2/\rho) + \lambda||\hat{w}||^2 + \frac{1}{\lambda_{\min}+\lambda}c(\epsilon,\delta)\sqrt{\log(2/\rho)}n\sqrt{d}||\mathcal{Y}||^2||\mathcal{X}||^2\right) \tag{14}$$

Finally optimizing over $\lambda$ gives the desired result:

$$\alpha = \tilde{O}\left(||\hat{w}||\sqrt{\frac{d||\hat{w}||^2(||\mathcal{X}||^4/\epsilon^2)\log(2d^2/\rho)}{n^2} + \frac{c(\epsilon,\delta)\sqrt{\log(2/\rho)}\sqrt{d}||\mathcal{Y}||^2||\mathcal{X}||^2}{n}}\right)$$

□

Inspecting Theorem 4.3 in the regime where the projection improves the error over the Gaussian Mechanim $l > \frac{n}{\sqrt{d}}$, when $n < \frac{\sqrt{d}||\mathcal{X}||^2||\mathcal{W}||^2}{||\mathcal{Y}||^2}$ is not too large, the dominant term in the error is $\tilde{O}\left(\frac{\sqrt{d}||\mathcal{X}||^2||\hat{w}||^2}{n}\right)$, and so we achieve PRIMO for Free. When $n$ is sufficiently large the dominant error term is $\tilde{O}\left(\frac{||\hat{w}||d^{1/4}||\mathcal{Y}||||\mathcal{X}||}{\sqrt{n}}\right)$ which is a factor of $\frac{n}{d^{1/4}}$ worse than the lower bound.

### 4.3 Computational Complexity

The computational complexity of Algorithm 1 can be broken down into 3 components:

- Step 1: Forming $X^T X$, $\big(nd^2$ or much faster$\big)$
- Step 2: the cost of computing $\hat{I}^{-1}\hat{v}_i = (X^T X + \lambda I + E_1)^{-1}\hat{v}_i \; \forall i \in [l]$, $\big(d^3 + ld^2\big)$
- Step 3: In the case where $\mathcal{M}$ is the projection algorithm, computing the projection $\text{argmin}_{v \in C(\mathcal{X}^n)} ||v - \hat{v}_i||_2^2$, $(nl\min(n,l) + nd + nld$ via diagonalization$)$

Forming the covariance matrix $X^T X$ is a matrix multiplication of two $d \times n$ matrices, which can be done via the naive matrix multiplication in time $O(nd^2)$, and via a long-line of "fast" matrix multiplication algorithms in time $O(d^{2+\alpha(n)})$; for example if $n < d^{.3}$ it can be done in time that is essentially $O(d^2)$ Gall [2012]. Step 2 can be completed by solving the equation $\hat{I}\hat{w}_i = \hat{v}_i, i = 1\ldots l$ via the conjugate gradient method, which takes time $O(\gamma(\hat{I})d^2\log(1/\epsilon))$ to compute an $\epsilon$-approximate solution Mahoney [2011] where $\gamma(\hat{I})$ is the condition number. We note that this has to be done separately for each $i = 1\ldots l$ giving total time $O(l \cdot \gamma(\hat{I})d^2\log(1/\epsilon))$. Alternatively, an exact solution $\hat{I}^{-1}\hat{v}_i$ can be computed directly using the QR decomposition of the matrix $\hat{I}$. The decomposition $\hat{I} = QR$ can be computed in time $O(d^3)$ Mahoney [2011] and does not depend on the $\hat{v}_i$, after which using $R\hat{w}_i = Q^T\tilde{v}_i$, $\hat{w}_i$ can be computed in time $O(d^2)$ via backward substitution. This gives a total time complexity of $O(d^3 + ld^2)$. So if $d$ and $\frac{1}{\gamma(\hat{I})}$ are sufficiently small relative to $l$, e.g. if $l = \tilde{\Omega}(d/(\gamma(\hat{I})))$, it will be faster to use the $QR$ decomposition based method. The projection in line 5 corresponds to minimizing a quadratic over a sphere. Setting $A = C^T C \in \mathbb{R}^{dn \times dn}, b = 2C^T\tilde{g} \in \mathbb{R}^{dl}$, then $\hat{v} = (Cn||\mathcal{X}||^2)x$, where $x \in \mathbb{R}^{nd}$ is the minimizer of:

$$\min_{x \in B_1} x^t A x - b^t x \tag{15}$$

$$\text{s.t. } ||x||_2 \leq \sqrt{n}||\mathcal{X}|| \tag{16}$$

Now, given the spectral decomposition of $A = U\Lambda U^T$, and the coordinates of $b$ in the eigenbasis $U^T b$, Lemma 2.2 in Hager [2001] gives a simple closed form for $x$ that computes each coordinate in constant time. Since there are $nd$ coordinates of $x$, this incurs an additional additive factor of $O(nd)$ in the complexity, which is dominated by the cost of diagonalizing $A$. So the complexity of this step is the complexity of diagonalizing $A = C^T C$, or equivalently finding the right singular vectors of $C$, plus the complexity of computing $U^T b$. This is seemingly bad news, as $C \in \mathbb{R}^{dn \times dl}$ is a very high-dimensional matrix, and the complexity for computing the SVD of $C$ without any assumptions about its structure is $O(d^3 ln\min(l,n))$ Golub and Van Loan [1996]. However, it is evident from the construction of $C$ in Subsection 3, that $C = I_d \otimes \frac{1}{n}Y^T$, where $\otimes$ is the Kronecker product. Then if $L\Lambda V^T$ is the SVD of $Y^T$, standard properties of the Kronecker product imply that the spectral decomposition of $C^T C$ is:

$$C = \frac{1}{n} \cdot (I_d \otimes L)(I_d \otimes \Lambda)(I_d \otimes V^T) \implies \tag{17}$$

$$C^T C = (I_d \otimes V)(I_d \otimes \Lambda^2)(I_d \otimes V^T) \tag{18}$$

Hence we can compute SPEC($C$) in the time it takes to compute SVD($Y$), or $O(nl\min(n,l))$. Similarly, to efficiently compute the $U^T b$ term required for Lemma 2.2 Hager [2001] we can again take advantage of properties of the Kronecker product, $U^T b =$

$$(I_d \otimes V)^T 2C^T\tilde{g} = 2(I_d \otimes V^T)(I_d \otimes \frac{1}{n}Y)\tilde{g} =$$

$$2(I_d \otimes \frac{1}{n}V^T y)\tilde{g} = \text{vec}(\frac{2}{n}V^T Y \text{mat}(\tilde{g})),$$

where mat($\tilde{g}$) is the $l \times d$ matrix with row $i$ given by elements $(l(i-1)+1, \ldots, l(i-1)+l$ of $\tilde{g}$, and the last equality follows properties of the Kronecker product. Now $V^T Y = \Lambda U^T$ which can be computed in $O(ln)$ since $\Lambda$ is diagonal. Multiplying by mat($\tilde{g}$) can be done in another $O(nld)$, for total complexity of $O(nl\max(\min(n,l),d))$.

Putting the complexity of these steps together we get:

**Theorem 4.4.** *The complexity of Algorithm 3 is* $O(\max(\min(nl^2, n^2 l), nld, nd^2, ld^2, d^3))$.

**Algorithm 3** Input: $\lambda, \mathcal{X} \in \mathcal{X}^n \subset \mathbb{R}^{d\times n}, Y = [y_1, \dots y_l] \in \mathcal{Y}^{l\times n}$, privacy params: $\epsilon, \delta$. We denote by $\mathcal{B}$ the Algorithm in Lemma 2.2 Hager [2001]

$\lambda -$ ReuseCov

Draw $E_1 \sim N_{d(d+1)/2}(0, \sigma_1^2)$, where $\sigma_1 = \frac{1}{n} 2\sqrt{2\log(2.5/\delta)}||\mathcal{X}||^2/\epsilon$
Compute $\hat{I} = (\frac{1}{n}X^T X + E_1 + \lambda I)$
Compute the QR decomposition $\hat{I} = QR$
Draw $\hat{v} = [\hat{v}_1, \dots \hat{v}_l] \sim$ GaussMech$(\epsilon/2, \delta/2, \Delta = \frac{1}{n}\sqrt{l}||\mathcal{X}||||\mathcal{Y}||)$
Compute SVD$(Y^T) = U\Lambda V^T$
Compute $\Pi_{C(Y)}\hat{v} = \mathcal{B}(V, \Lambda, \hat{v})$
**for** $i = 1 \dots l$
    Solve $R\hat{w}_i = Q^T \hat{v}_i$ by back substitution.
**end for**
Return $\hat{W} = [\hat{w}_1, \dots \hat{w}_l]$

### 4.4 Sub-sampling based ReuseCov

The discussion in the previous section shows that when $n > d > l$, the complexity of Algorithm 3 is $O(nd^2)$ or the cost of forming the covariance matrix. In this section we show how sub-sampling $s < n$ points can improve this to $O(sd^2)$ by giving an analysis of sub-sampled SSP. The key ingredient is marrying the convergence of the sub-sampled covariance matrix to $X^T X$ with the accuracy analysis of SSP we saw in Section 4.1.

**Algorithm 4** Input: $s, \lambda, \mathcal{X} \in \mathcal{X}^n \subset \mathbb{R}^{d\times n}, Y = [y_1, \dots y_l] \in \mathcal{Y}^{l\times n}$, privacy params: $\epsilon, \delta$

$\lambda -$ SubSampReuseCov

Sub-sample $s$ points without replacement from $\mathcal{X}$, we denote the sub-sampled design matrix by $X_S$.
Let $(\epsilon_1, \delta_1) = (\frac{n}{s}\epsilon/2, \delta/2)$
Draw $E_1 \sim N_{d(d+1)/2}(0, \sigma_1^2)$, where $\sigma_1 = \frac{1}{n} 2\sqrt{2\log(2.5/\delta)}||\mathcal{X}||^2/\epsilon_1$
Compute $\hat{I}_s = (\frac{1}{s}X_S^T X_S + E_1 + \lambda I)$
Draw $\hat{v} = [\hat{v}_1, \dots \hat{v}_l] \sim$ GaussMech$(\epsilon/2, \delta/2, \Delta = \frac{1}{n}\sqrt{l}||\mathcal{X}||||\mathcal{Y}||)$
**for** $i = 1 \dots l$
    Set $\hat{w}_i = \hat{I}_s^{-1}\hat{v}_i$
**end for**
Return $\hat{W} = [\hat{w}_1, \dots \hat{w}_l]$

Our algorithm is based on the observation that if we sub-sample $S \subset [n], |S| = s$ points without replacement then:

- The cost of computing the covariance matrix $\Sigma_S = \sum_{k\in S} x_k x_k^T$ is $O(sd^2)$
- By the "secrecy of the sub-sample" principle Dwork and Roth [2014], our privacy cost for estimating $\Sigma_S$ is scaled down by a factor of $s/n$
- With high probability for sufficiently large $s$, $\Sigma_S \to \Sigma$ by a matrix-Chernoff bound for sampling without replacement Tropp [2010]

**Theorem 4.5.** *With $\mathcal{M}$ = GaussMech$(\epsilon/2, \delta/2, \Delta = \frac{1}{n}\sqrt{l}||\mathcal{X}||||\mathcal{Y}||)$, Algorithm 4 is an $(\alpha, \rho, O(\epsilon), \delta)$ solution to the PRIMO problem with*

$$\alpha = O\left(||\hat{w}||\sqrt{||\mathcal{X}||||\mathcal{Y}||(\frac{\log(2d/\rho)}{s}) + \frac{d}{\lambda}(1 + \frac{\log(2d/\rho)}{s}) \cdot (\frac{||X||^4 n^2}{s^4\epsilon^2}||\mathcal{W}||^2 + \frac{l||\mathcal{X}||^2||\mathcal{Y}||^2}{n^2\epsilon^2})}\right)$$

**Proof Sketch.** Our analysis will hinge on the case where $l = 1$ e.g. that of standard private linear regression, which we will extend to the PRIMO case by our choice of $\epsilon$ as in the proof of Theorem 4.1. Now let:

- $w_{i*} = (X^TX)^{-1}X^TY$ the least squares estimator as before
- $w_{i*}^{\lambda} = (\frac{1}{n}X^TX + \lambda I)^{-1}\frac{1}{n}X^TY$ the ridge regression estimator
- $w_s = (\frac{1}{s}X_S^TX_S + \lambda I)^{-1}(\frac{1}{n}X^TY)$ the sub-sampled least squares estimator
- $\tilde{w}_s = (\frac{1}{s}X_S^TX_S + E_1 + \lambda I)^{-1}(\frac{1}{n}X^TY + E_2)$ our differentially private estimate of $w_s$

We note that $0 \leq f(w_{i*}^{\lambda}) - f(w_{i*}) \leq \lambda \left(||w_{i*}||^2 - ||w_{i*}^{\lambda}||^2\right) \leq \lambda||\mathcal{W}||^2$. Then by the Lemma 9 and Cauchy-Schwartz with respect to the norm $||\cdot||_{\frac{X^TX}{n}}$:

$$\begin{aligned}|f(\tilde{w}_s) - f(w_{i*})| = ||\tilde{w}_s - w_{i*}||^2_{\frac{X^TX}{n}} \leq \\ 3||w_{i*} - w_{i*}^{\lambda}||^2_{\frac{X^TX}{n}} + 3||w_{i*}^{\lambda} - w_s||^2_{\frac{X^TX}{n}} + 3||w_s - \tilde{w}_s||^2_{\frac{X^TX}{n}} \leq \\ 3\lambda||\mathcal{W}||^2 + \underbrace{3||w_s - w_{i*}^{\lambda}||^2_{\frac{X^TX}{n}}}_{\text{Matrix Chernoff}} + \underbrace{3||\tilde{w}_s - w_s||^2_{\frac{X^TX}{n}}}_{\text{SSP analysis+Matrix Chernoff}}\end{aligned} \tag{19}$$

So it suffices to bound each term with high probability. The second term, $||\tilde{w}_s - w_s||_{X^TX+\lambda I}$ can be bounded using the same arguments as in Theorem 4.1, with small differences due to scaling. Crucially though, as we need to bound this in the norm induced by $X^TX$ rather than $X_S^TX_S$, we will need to utilize the convergence of $X_S^TX_S \to X^TX$ via Matrix-Chernoff bounds.

**Lemma 4.** *Under the assumption $||\mathcal{X}|| = O(n\lambda)$, then w.p. $1 - \rho/2$:*

$$||w_s - \tilde{w}_s||_{\frac{X^TX}{n}} = O(\frac{||X||^4}{n^2\epsilon^2}||\mathcal{W}||^2 \cdot \frac{d}{\lambda}\frac{\log(2d/\rho)}{s} + \frac{l||\mathcal{X}||^2||\mathcal{Y}||^2}{n^2\epsilon^2} \cdot \frac{d}{\lambda}\frac{\log(2d/\rho)}{s}) \tag{20}$$

Bounding the first term can be reduced to bounding $||I - (\frac{1}{n}X^TX)^{-1/2}(\frac{1}{s}X_S^TX_S)(\frac{1}{n}X^TX)^{-1/2}||_2$ which follows more directly via the Matrix-Chernoff bound for sub-sampling without replacement:

**Lemma 5.** *Under the assumption $||\mathcal{X}|| = O(n\lambda)$, then w.p. $1 - \rho/2$:*

$$||w_{i*}^{\lambda} - w_s||_{\frac{X^TX}{n}} = O\left(\frac{||\mathcal{X}||||\mathcal{Y}||\log(2d/\rho}{\lambda s}\right)$$

Substituting into Equation 20 and minimizing over $\lambda$ gives the desired result.

# References


Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction and estimation in unbounded domain, 2018. URL `https://arxiv.org/abs/1803.02596`.

Aleksandar Nikolov, Kunal Talwar, and Li Zhang. The geometry of differential privacy: the sparse and approximate cases. In *STOC '13*, 2013.

E. Krapohl, H. Patel, S. Newhouse, C J Curtis, S. von Stumm, P S Dale, D. Zabaneh, G. Breen, P F O'Reilly, and R. Plomin. Multi-polygenic score approach to trait prediction. *Molecular Psychiatry*, 23(5):1368–1374, 2018. doi: 10.1038/mp.2017.163. URL `https://doi.org/10.1038/mp.2017.163`.

Jack Pattee and Wei Pan. Penalized regression and model selection methods for polygenic scores on summary statistics. *PLOS Computational Biology*, 16(10):1–27, 10 2020a. doi: 10.1371/journal.pcbi.1008271. URL `https://doi.org/10.1371/journal.pcbi.1008271`.

Alan Agresti and Barbara Finlay. Statistical methods for the social sciences. *Pearson Prentice Hall*, 2009.

Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLOS Genetics*, 4(8):1–9, 08 2008. doi: 10.1371/journal.pgen.1000167. URL `https://doi.org/10.1371/journal.pgen.1000167`.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 2021.

Reza Shokri, Marco Stronati, and Vitaly Shmatikov. Membership inference attacks against machine learning models. *CoRR*, abs/1610.05820, 2016. URL http://arxiv.org/abs/1610.05820.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014. URL http://dblp.uni-trier.de/db/journals/fttcs/fttcs9.html#DworkR14.

Andrew Gelman and Eric Loken. The statistical crisis in science. 102:460+. ISSN 0003-0996. doi: 10.1511/2014.111.460. URL https://doi.org/10.1511/2014.111.460.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *CoRR*, abs/1506.02629, 2015a. URL http://arxiv.org/abs/1506.02629.

Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan R. Ullman. Algorithmic stability for adaptive data analysis. *CoRR*, abs/1511.02513, 2015. URL http://arxiv.org/abs/1511.02513.

Keegan Korthauer, Patrick K. Kimes, Claire Duvallet, Alejandro Reyes, Ayshwarya Subramanian, Mingxiang Teng, Chinmay Shukla, Eric J. Alm, and Stephanie C. Hicks. A practical guide to methods controlling false discoveries in computational biology. *Genome Biology*, 20(1):118, 2019. doi: 10.1186/s13059-019-1716-1. URL https://doi.org/10.1186/s13059-019-1716-1.

Jack Pattee and Wei Pan. Penalized regression and model selection methods for polygenic scores on summary statistics. *PLOS Computational Biology*, 16(10):1–27, 10 2020b. doi: 10.1371/journal.pcbi.1008271. URL https://doi.org/10.1371/journal.pcbi.1008271.

Matthew C. Ringgenberg Mehrdad Samadi Ingrid M. Werner Heath, Davidson. Reusing natural experiments. *International Finance Discussion Papers 1339. Washington: Board of Governors of the Federal Reserve System*, 2022. URL https://doi.org/10.17016/IFDP.2022.1339.

Raef Bassily, Adam D. Smith, and Abhradeep Thakurta. Private empirical risk minimization, revisited. *CoRR*, abs/1405.7085, 2014. URL http://arxiv.org/abs/1405.7085.

Duy Vu and Aleksandra Slavkovic. Differential privacy for clinical trial data: Preliminary evaluations. In *2009 IEEE International Conference on Data Mining Workshops*, pages 138–143, 2009. doi: 10.1109/ICDMW.2009.52.

James R. Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. On the theory and practice of privacy-preserving bayesian data analysis. *CoRR*, abs/1603.07294, 2016. URL http://arxiv.org/abs/1603.07294.

Sergül Aydöre, William Brown, Michael Kearns, Krishnaram Kenthapadi, Luca Melis, Aaron Roth, and Ankit A. Siva. Differentially private query release through adaptive projection. *CoRR*, abs/2103.06641, 2021. URL https://arxiv.org/abs/2103.06641.

Joel A. Tropp. Improved analysis of the subsampled randomized hadamard transform. *CoRR*, abs/1011.1595, 2010. URL http://arxiv.org/abs/1011.1595.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In Serge Vaudenay, editor, *Advances in Cryptology - EUROCRYPT 2006, 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28 - June 1, 2006, Proceedings*, volume 4004 of *Lecture Notes in Computer Science*, pages 486–503. Springer, 2006. doi: 10.1007/11761679_29. URL https://doi.org/10.1007/11761679_29.

Jason Milionis, Alkis Kalavasis, Dimitris Fotakis, and Stratis Ioannidis. Differentially private regression with unbounded covariates. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3242–3273. PMLR, 28–30 Mar 2022. URL `https://proceedings.mlr.press/v151/milionis22a.html`.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 25.1–25.40, Edinburgh, Scotland, 25–27 Jun 2012. PMLR. URL `https://proceedings.mlr.press/v23/kifer12.html`.

Christos Dimitrakakis, Blaine Nelson, Zhang, Zuhe, Aikaterini Mitrokotsa, and Benjamin Rubinstein. Bayesian differential privacy through posterior sampling, 2013. URL `https://arxiv.org/abs/1306.1066`.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(29):1069–1109, 2011. URL `http://jmlr.org/papers/v12/chaudhuri11a.html`.

Cynthia Dwork, Aleksandar Nikolov, and Kunal Talwar. Efficient algorithms for privately releasing marginals via convex relaxations. *Discret. Comput. Geom.*, 53(3):650–673, 2015b. doi: 10.1007/s00454-015-9678-x. URL `https://doi.org/10.1007/s00454-015-9678-x`.

Justin Thaler, Jonathan R. Ullman, and Salil P. Vadhan. Faster algorithms for privately releasing marginals. In Artur Czumaj, Kurt Mehlhorn, Andrew M. Pitts, and Roger Wattenhofer, editors, *Automata, Languages, and Programming - 39th International Colloquium, ICALP 2012, Warwick, UK, July 9-13, 2012, Proceedings, Part I*, volume 7391 of *Lecture Notes in Computer Science*, pages 810–821. Springer, 2012. doi: 10.1007/978-3-642-31594-7_68. URL `https://doi.org/10.1007/978-3-642-31594-7_68`.

Jonathan R. Ullman and Salil P. Vadhan. Pcps and the hardness of generating private synthetic data. In Yuval Ishai, editor, *Theory of Cryptography - 8th Theory of Cryptography Conference, TCC 2011, Providence, RI, USA, March 28-30, 2011. Proceedings*, volume 6597 of *Lecture Notes in Computer Science*, pages 400–416. Springer, 2011. doi: 10.1007/978-3-642-19571-6_24. URL `https://doi.org/10.1007/978-3-642-19571-6_24`.

Moritz Hardt and Guy N. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 61–70, 2010. doi: 10.1109/FOCS.2010.85.

Mark Bun, Jonathan R. Ullman, and Salil P. Vadhan. Fingerprinting codes and the price of approximate differential privacy. *CoRR*, abs/1311.3158, 2013. URL `http://arxiv.org/abs/1311.3158`.

Salil P. Vadhan. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, 2017.

Alexander Edmonds, Aleksandar Nikolov, and Jonathan R. Ullman. The power of factorization mechanisms in local and central differential privacy. *CoRR*, abs/1911.08339, 2019. URL `http://arxiv.org/abs/1911.08339`.

S. Muthukrishnan and Aleksandar Nikolov. Optimal private halfspace counting via discrepancy. In Howard J. Karloff and Toniann Pitassi, editors, *Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, New York, NY, USA, May 19 - 22, 2012*, pages 1285–1292. ACM, 2012. doi: 10.1145/2213977.2214090. URL `https://doi.org/10.1145/2213977.2214090`.

Anupam Gupta, Aaron Roth, and Jonathan Ullman. Iterative constructions and private data release, 2011.

Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. *CoRR*, abs/1109.2229, 2011. URL `http://arxiv.org/abs/1109.2229`.

Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60, 2010. doi: 10.1109/FOCS.2010.12.

Ryan McKenna, Gerome Miklau, Michael Hay, and Ashwin Machanavajjhala. Optimizing error of high-dimensional statistical queries under differential privacy. *CoRR*, abs/1808.03537, 2018. URL http://arxiv.org/abs/1808.03537.

Ryan McKenna, Daniel Sheldon, and Gerome Miklau. Graphical-model based estimation and inference for differential privacy. *CoRR*, abs/1901.09136, 2019. URL http://arxiv.org/abs/1901.09136.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1zk9iRqF7.

Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. DP-CGAN: differentially private synthetic data and label generation. *CoRR*, abs/2001.09700, 2020. URL https://arxiv.org/abs/2001.09700.

Brett Beaulieu-Jones, Zhiwei Wu, Chris Williams, Ran Lee, Sanjeev Bhavnani, James Byrd, and Casey Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12, 07 2019. doi: 10.1161/CIRCOUTCOMES.118.005122.

Seth V. Neel, Aaron L. Roth, and Zhiwei Steven Wu. How to use heuristics for differential privacy. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 72–93, 2019. doi: 10.1109/FOCS.2019.00014.

Roman Vershynin. High-dimensional probability. 2019. URL https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.pdf.

Michal Derezinski, Manfred K. Warmuth, and Daniel Hsu. Tail bounds for volume sampled linear regression. *CoRR*, abs/1802.06749, 2018. URL http://arxiv.org/abs/1802.06749.

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Sampling algorithms for l2 regression and applications. In *SODA '06*, 2006.

Michal Derezinski and Manfred K. Warmuth. Unbiased estimates for linear regression via volume sampling. *CoRR*, abs/1705.06908, 2017. URL http://arxiv.org/abs/1705.06908.

Yin Tat Lee and He Sun. Constructing linear-sized spectral sparsification in almost-linear time. *CoRR*, abs/1508.03261, 2015. URL http://arxiv.org/abs/1508.03261.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over $\ell_q$-balls, 2009. URL https://arxiv.org/abs/0910.2042.

François Le Gall. Faster algorithms for rectangular matrix multiplication. *CoRR*, abs/1204.1111, 2012. URL http://arxiv.org/abs/1204.1111.

Michael W. Mahoney. Randomized algorithms for matrices and data. *CoRR*, abs/1104.5557, 2011. URL http://arxiv.org/abs/1104.5557.

William W. Hager. Minimizing a quadratic over a sphere. *SIAM J. Optim.*, 12:188–208, 2001.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, third edition, 1996.

Michael J. Kearns. Efficient noise-tolerant learning from statistical queries. In S. Rao Kosaraju, David S. Johnson, and Alok Aggarwal, editors, *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing, May 16-18, 1993, San Diego, CA, USA*, pages 392–401. ACM, 1993. doi: 10.1145/167088.167200. URL https://doi.org/10.1145/167088.167200.

Shiva Prasad Kasiviswanathan, Mark Rudelson, Adam D. Smith, and Jonathan R. Ullman. The price of privately releasing contingency tables and the spectra of random matrices with correlated rows. In Leonard J. Schulman, editor, *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, pages 775–784. ACM, 2010. doi: 10.1145/1806689.1806795. URL https://doi.org/10.1145/1806689.1806795.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. *CoRR*, abs/1808.00087, 2018. URL http://arxiv.org/abs/1808.00087.

# 5 Appendix

## 5.1 Lemmas and Definitions

**Definition 3.** *Statistical Query Kearns [1993] Let $D \in \mathcal{X}^n$ a dataset. A linear query is a function $q : \mathcal{X} \to [0,1]$, where $q(D) := \frac{1}{n}\sum_{x_i \in D}^{n} q(x_i)$.*

**Definition 4.** *Kasiviswanathan et al. [2010] Let $\mathcal{X} = \{0,1\}^m$, and $\mathcal{Q}_k = \{q_{ij}\}_{1 \leq i_1 < i_2 < i_k \leq m}$, where $q_{ij}(x) := \prod_{u=1}^{k} x_{i_u}$. Then the class $\mathcal{Q}_k$ are called $k-$way marginals.*

**Lemma 6.** *Tropp [2010] Let $Z_i \in \mathbb{S}_d^+$, and sample $Z_1, \ldots Z_s$ without replacement from $\{Z_1, \ldots Z_n\}$. Suppose $\mathbb{E}[Z_i] = I_d$, and $\max_{i \in [n]} \lambda_{\max}(Z_i) \leq B$. Then for $\delta = \sqrt{\frac{2B\log(2d/\rho)}{s}}$, with probability $1-\rho$:*

$$\lambda_{\max}(\frac{1}{s}\sum_{i=1}^{s} Z_i) < 1 + \delta, \quad \lambda_{\min}(\frac{1}{s}\sum_{i=1}^{s} Z_i) > 1 - \delta$$

**Lemma 7** (folklore e.g. Wang et al. [2018])**.** *Given a dataset $\mathcal{X}^n$ of $n$ points and an $(\epsilon,\delta)-$DP mechanism $M$. Let the procedure **subsample** take a random subset of $s$ points from $\mathcal{X}^n$ without replacement. Then if $\gamma = s/n$, the procedure $M \circ$ **subsample** is $(O(\gamma\epsilon), \gamma\delta)-$DP for sufficiently small $\epsilon$.*

## 5.2 Proofs from Subsection 4.1

The following lemma is used repeatedly in analyzing the accuracy of all $SSP$ variants.

**Lemma 8.** *Let $A, B$ invertible matrices in $\mathbb{R}^{n \times n}$, and $v, c$ vectors $\in \mathbb{R}^n$.*

*Then*

$$A^{-1}v - (A+B)^{-1}(v+c) = (A+B)^{-1}BA^{-1}v - (A+B)^{-1}c$$

*Proof.* Expanding we get that:

$$A^{-1}v - (A+B)^{-1}(v+c) = (A^{-1} - (A+B)^{-1})v - (A+B)^{-1}c,$$

so it suffices to show that:

$$(A^{-1} - (A+B)^{-1})v = (A+B)^{-1}BA^{-1}v$$

Now the Woodbury formula tells us that $(A+B)^{-1} = A^{-1} - (A + AB^{-1}A)^{-1}$, hence

$$(A^{-1} - (A+B)^{-1})v = (A + AB^{-1}A)^{-1}v = (A(I + B^{-1}A))^{-1}v$$

Then since:

$$(A(I + B^{-1}A))^{-1} = (I + B^{-1}A)^{-1}A^{-1} = (B^{-1}(B+A))^{-1A^{-1}} = (B+A)^{-1}BA^{-1},$$

we are done. □

**Proof of $(\epsilon, \delta)$-DP in Theorem 4.1**:

*Proof.* The privacy proof follows from a straightforward application of the Gaussian mechanism. We note that releasing each $\hat{v}_i$ privately, is equivalent to computing $X^TY + E_2$, where $E_2 \sim N_{d\times l}(0, \sigma_2^2)$. Now it is easy to compute $l$-sensitivity $\Delta(f)$ of $f(X) = X^TY$. Fix an individual $i$, and an adjacent dataset $X' = X/\{x_i, y_i\} \cup \{x_i', y_i'\}$. Then $f(X) - f(X') = \Delta_V = [y_{i1}x_i - y_{i1}'x_{i1}', \ldots y_{il}x_i - y_{il}'x_{il}']$. Then:

$$||f(X) - f(X')|| = \sqrt{||\Delta_V||_F^2} = \sqrt{\sum_{j=1}^{l} ||y_{ij}x_i - y_{ij}'x_{ij}'||^2} \leq \sqrt{l \cdot 4||\mathcal{X}||^2||\mathcal{Y}||^2} = 2\sqrt{l}||\mathcal{X}||||\mathcal{Y}||$$

Hence setting $\sigma_2 = 2\sqrt{2\log(2.5/\delta)}2\sqrt{l}||\mathcal{X}||||\mathcal{Y}||/\epsilon$ by the Gaussian mechanism Dwork and Roth [2014] publishing $\hat{V}$ satisfies $(\epsilon/2, \delta/2) - DP$. Similarly if $g = X^TX$, $\Delta(g) \leq ||X||^2$, and so setting $\sigma_1 = 2\sqrt{2\log(2.5/\delta)}||X||^2/\epsilon$, means publishing $\hat{I}$ is $(\epsilon/2, \delta/2)$-DP. By basic composition for DP, the entire mechanism is $(\epsilon, \delta)$-DP. □

**Proof of Accuracy in Theorem 4.1:**

*Proof.* We follow the general proof technique developed in Wang [2018] analysing the accuracy guarantees of the ridge regression variant of SSP in the case $\lambda_{\min}(X^TX) = 0$, adding some mathematical detail to their exposition, and doing the appropriate book-keeping to handle our setting where the privacy level (as a function of the noise level) guaranteed by $E_1$ and $E_2$ differ. The reader less interested in these details can skip to Equation 25 below for the punchline.

Fix a specific index $i \in [l]$, and let $y = y_i$. We will analyze the prediction error of $\hat{w}_i$ e.g. $F(w_i) - F(w_i^*)$. Then the following result is stated in Wang [2018] for which provide a short proof:

**Lemma 9.**

$$F(\hat{w}_i) - F(w_{i*}) = ||y - X\hat{w}_i||^2 - ||y - Xw_{i*}|| = ||\hat{w}_i - w_{i*}||_{X^TX}^2$$

*Proof.* We note that all derivatives of orders higher than 2 of $f(w) = ||y - Xw||^2$ are zero, and that $\nabla f_{w_{i*}} = 0$ by the optimality of $w_{i*}$. We also note that the Hessian $\nabla^2 f_w = X^TX$ at all points $w$. Then by the Taylor expansion of $f(w)$ around $w_{i*}$:

$$f(\hat{w}_i) = f(w_{i*}) + (\hat{w}_i - w_{i*}) \cdot \nabla f_{w_{i*}} + (\hat{w}_i - w_{i*})'X^TX(\hat{w}_i - w_{i*})$$

Which using $\nabla f_{w_{i*}} = 0$ and rearranging terms gives the result. □

Now Corollary 7 in the Appendix of Wang [2018] states (without proof) the below identity, which we provide a proof of for completeness via Lemma 8:

$$\hat{w}_i - w_{i*} = (-X^TX + \lambda I + E_1)^{-1}E_1w_{i*} - \lambda(X^TX + \lambda I + E_1)^{-1}w_{i*} + (X^TX + \lambda I + E_1)^{-1}E_2 \tag{21}$$

Hence, still following Wang [2018], for any psd matrix $A$, $||\hat{w}_i - w_{i*}||_A^2 \leq$

$$3||(X^TX + \lambda I + E_1)^{-1}E_1w_{i*}||_A^2 + 3\lambda^2||(X^TX + \lambda I + E_1)^{-1}w_{i*}||_A^2 + 3||(X^TX + \lambda I + E_1)^{-1}E_2||_A^2 \tag{22}$$

**Lemma 10.** *Wang [2018] With probability* $1 - \rho$, $||E_1|| \leq (\lambda_{\min}(X^TX) + \lambda)/2$, *and hence*

$$X^TX + \lambda I + E_1 \succ .5(X^TX + \lambda I)$$

We also remark that $||By||_A^2 = (By)^TABy = ||y||_{B^TAB}^2$ for any vector $y$, and matrices $A, B$.

Hence, Inequality 22, with $A = X^TX$ can be further simplified to:

$$\begin{aligned}
&3||(X^TX + \lambda I + E_1)^{-1}E_1w_{i*}||_A^2 + 3\lambda^2||(X^TX + \lambda I + E_1)^{-1}w_{i*}||_A^2 + 3||(X^TX + \lambda I + E_1)^{-1}E_2||_A^2 \leq \\
&O\left(||E_1w_{i*}||_{(X^TX + \lambda I + E_1)^{-1}}^2 + \lambda^2||w_{i*}||_{(X^TX + \lambda I + E_1)^{-1}}^2 + ||E_2||_{(X^TX + \lambda I + E_1)^{-1}}^2\right) \leq \\
&\text{(Lemma 10)} \;\; O\left(||E_1w_{i*}||_{(X^TX + \lambda I)^{-1}}^2 + \lambda^2||w_{i*}||_{(X^TX + \lambda I)^{-1}}^2 + ||E_2||_{(X^TX + \lambda I)^{-1}}^2\right)
\end{aligned} \tag{23}$$

By basic properties of the trace we have: $\mathrm{tr}((\lambda I + X^TX)^{-1}) \leq d\lambda_{max}(\lambda I + X^TX^{-1}) = \frac{d}{\lambda_{\min}(\lambda I + X^TX)} = \frac{d}{(\lambda_{\min} + \lambda)}$, and $||w_{i*}||_{(X^TX + \lambda I)^{-1}}^2 \leq \frac{||w_{i*}||^2}{\lambda}$. Continuing from Wang [2018] by their Lemma 6, we can bound each $||E_1w_{i*}||_{(X^TX + \lambda I)^{-1}}^2$ and $||E_2||_{(X^TX + \lambda I)^{-1}}^2$.

**Lemma 11** (Wang [2018]). *Let $\theta \in \mathbb{R}^d$ and let $E$ a symmetric Gaussian matrix where the upper triangular region is sampled from $N(0, \sigma^2)$ and let $A$ be any psd matrix. Then with probability $1 - \rho$:*

$$||E\theta||_A^2 \leq \sigma^2 tr(A)||\theta||^2 \log(2d^2/\rho)$$

Then recalling that:

- $\sigma_1^2 = \tilde{O}(||\mathcal{X}||^4/\epsilon^2)$
- $\sigma_2^2 = \tilde{O}(l||\mathcal{X}||^2||\mathcal{Y}||^2/\epsilon^2)$

Plugging into Lemma 11, and bringing it all together we get:

$$\begin{aligned} |F(\hat{w}_i) - F(w_{i*})| \leq ||w_{i*} - \hat{w}_i||_{X^TX} = \\ O\left(||E_1 w_{i*}||^2_{(X^TX+\lambda I)^{-1}} + \lambda^2||w_{i*}||^2_{(X^TX+\lambda I)^{-1}} + ||E_2||^2_{(X^TX+\lambda I)^{-1}}\right) = \\ \tilde{O}\left(\frac{d}{\lambda_{\min}+\lambda}||w_{i*}||^2(||\mathcal{X}||^4/\epsilon^2)\log(2d^2/\rho) + \lambda||w_{i*}||^2 + \frac{d}{\lambda_{\min}+\lambda}l||\mathcal{X}||^2||\mathcal{Y}||^2/\epsilon^2\log(2d^2/\rho)\right) \end{aligned} \tag{24}$$

Upper bounding Equation 24 by taking $\lambda_{\min} = 0$, we minimize over $\lambda$ setting

$$\lambda = \tilde{O}\left(\frac{1}{\epsilon}\sqrt{d\log(2d^2/\rho)}||X||\sqrt{||\mathcal{X}||^2 + \frac{l||\mathcal{Y}||^2}{||w_{i*}||^2}}\right) \implies$$

$$|F(\hat{w}_i) - F(w_{i*})| = \tilde{O}\left(\frac{1}{\epsilon}\sqrt{d\log(2d^2/\rho)}||X||\sqrt{||\mathcal{X}||^2||w_{i*}||^4 + l||\mathcal{Y}||^2||w_{i*}||^2}\right) \tag{25}$$

Now if we are in the $\eta-$small regime, we have $||\mathcal{Y}|| \leq \eta||\mathcal{X}||||w_{i*}||$, and so

$$l||\mathcal{Y}||^2 w_{i*}^2 \leq l\eta^2||\mathcal{X}||^2||w_{i*}||^4,$$

which reduces Equation 25 to:

$$|F(\hat{w}_i) - F(w_{i*})| = \tilde{O}\left(\frac{1}{\epsilon}\sqrt{d\log(2d^2/\rho)}||X||^2||w_{i*}||^2(\eta\sqrt{l})\right),$$

as desired. □

### 5.3 Proofs from Subsection 4.4

**Proof of Theorem 4.5**:

*Proof.* Our analysis will hinge on the case where $l = 1$ e.g. that of standard private linear regression, which we will extend to the PRIMO case by our choice of $\epsilon$ as in the proof of Theorem 4.1. The fact that the Algorithm is $(O(\epsilon), \delta)$ private follows immediately from the Gaussian mechanism, and the secrecy of the sub-sample lemma (Lemma 7), which is why we can set $\epsilon_1 = \frac{n}{s}\epsilon/2$ in Line 2. We proceed with the accuracy analysis.

Define:

- $w_{i*} = (X^TX)^{-1}X^TY$: the least squares estimator as before
- $w_{i*}^\lambda = (\frac{1}{n}X^TX + \lambda I)^{-1}\frac{1}{n}X^TY$: the ridge regression estimator
- $w_s = (\frac{1}{s}X_S^TX_S + \lambda I)^{-1}(\frac{1}{n}X^TY)$: the sub-sampled least squares estimator
- $\tilde{w}_s = (\frac{1}{s}X_S^TX_S + E_1 + \lambda I)^{-1}(\frac{1}{n}X^TY + E_2)$: differentially private estimate of $w_s$

We note that $0 \leq f(w_{i*}^\lambda) - f(w_{i*}) \leq \lambda \left(||w_{i*}||^2 - ||w_{i*}^\lambda||^2\right) \leq \lambda||\mathcal{W}||^2$. Then by the Lemma 9 and Cauchy-Schwartz with respect to the norm $||\cdot||_{\frac{X^TX}{n}}$:

$$
\begin{aligned}
&|f(\tilde{w}_s) - f(w_{i*})| = ||\tilde{w}_s - w_{i*}||^2_{\frac{X^TX}{n}} \leq \\
&\qquad 3||w_{i*} - w_{i*}^\lambda||^2_{\frac{X^TX}{n}} + 3||w_{i*}^\lambda - w_s||^2_{\frac{X^TX}{n}} + 3||w_s - \tilde{w}_s||^2_{\frac{X^TX}{n}} \leq \\
&\qquad\qquad 3\lambda||\mathcal{W}||^2 + 3||w_s - w_{i*}^\lambda||^2_{\frac{X^TX}{n}} + 3||\tilde{w}_s - w_s||^2_{\frac{X^TX}{n}}
\end{aligned} \tag{26}
$$

Lemmas 12, 13 bound these terms with high probability.

**Lemma 12.** *Under the assumption $||\mathcal{X}|| = O(n\lambda)$, then w.p. $1 - \rho/2$:*

$$
||w_{i*}^\lambda - w_s||_{\frac{X^TX}{n}} = O\left(\frac{||\mathcal{X}||||\mathcal{Y}||\log(2d/\rho)}{\lambda s}\right)
$$

**Lemma 13.** *Under the assumption $||\mathcal{X}|| = O(n\lambda)$, then w.p. $1 - \rho/2$:*

$$
||w_s - \tilde{w}_s||_{\frac{X^TX}{n}} = O(\frac{||X||^4}{n^2\epsilon^2}||\mathcal{W}||^2 \cdot \frac{d}{\lambda}\frac{\log(2d/\rho)}{s} + \frac{l||\mathcal{X}||^2||\mathcal{Y}||^2}{n^2\epsilon^2} \cdot \frac{d}{\lambda}\frac{\log(2d/\rho)}{s}) \tag{27}
$$

Then by Equation 26 and Lemmas 12, 13, we get that w.p. $1 - \rho$:

$$
\begin{aligned}
&f(\tilde{w}_s) - f(w_{i*}) = \\
&O\left(\lambda||w_{i*}||^2 + \frac{||\mathcal{X}||||\mathcal{Y}||}{\lambda}(\frac{\log(2d/\rho)}{s}) + \frac{||X||^4n^2}{s^4\epsilon^2}||\mathcal{W}||^2 \cdot \frac{d}{\lambda}(1 + \frac{\log(2d/\rho)}{s}) \;+\; \frac{l||\mathcal{X}||^2||\mathcal{Y}||^2}{n^2\epsilon^2} \cdot \frac{d}{\lambda}(1 + \frac{\log(2d/\rho)}{s})\right),
\end{aligned} \tag{28}
$$

Summing over $i$ and minimizing over $\lambda$ we set

$$
\lambda = \frac{\sqrt{||\mathcal{X}||||\mathcal{Y}||(\frac{\log(2d/\rho)}{s}) + \frac{||X||^4n^2}{s^4\epsilon^2}||\mathcal{W}||^2 \cdot \frac{d}{\lambda}(1 + \frac{\log(2d/\rho)}{s}) \;+\; \frac{l||\mathcal{X}||^2||\mathcal{Y}||^2}{n^2\epsilon^2} \cdot \frac{d}{\lambda}(1 + \frac{\log(2d/\rho)}{s})}}{\sqrt{1}l||W||_F},
$$

which completes the result. □

**Proof of Lemma 12**:

*Proof.* Now:

$$
||w_{i*}^\lambda - w_s||^2_{\frac{1}{n}X^TX} = n||w_{i*}^\lambda - w_s||^2_{\frac{1}{n}X^TX} \leq n||w_{i*}^\lambda - w_s||^2_{\frac{1}{n}X^TX+\lambda I}
$$

We will focus on $||w_{i*}^\lambda - w_s||^2_{\frac{1}{n}X^TX+\lambda I}$. Let $\Sigma = \frac{1}{n}X^TX + \lambda I$, $\Sigma_s = \frac{1}{s}\sum_{j\in S} x_sx_s^T + \lambda I$, and $v = \frac{1}{n}X^TY$. Expanding $||w_{i*}^\lambda - w_s||_\Sigma =$

$$
(\Sigma_s^{-1}v - \Sigma^{-1}v)^T\Sigma(\Sigma_s^{-1}v - \Sigma^{-1}v) = v^T \overbrace{(\Sigma_s^{-1} - \Sigma^{-1})\Sigma(\Sigma_s^{-1} - \Sigma^{-1})}^{A} v = v^TAv \tag{29}
$$

Now since $A$ is Hermitian, we know $||v||_A \leq ||v||||A||_2$. Since $||v|| \leq ||\mathcal{X}||||\mathcal{Y}||$, it suffices to bound $||A||_2$ with high probability. Noting that $A = (\Sigma_s^{-1} - \Sigma^{-1})\Sigma(\Sigma_s^{-1} - \Sigma^{-1}) = \Sigma_s^{-1}\Sigma^{1/2}(I - \Sigma^{-1/2}\Sigma_s\Sigma^{-1/2})(I - \Sigma^{-1/2}\Sigma_s\Sigma^{-1/2})\Sigma^{1/2}\Sigma_s^{-1}$, we have by the sub-multiplicativity of the operator norm:

$$
\begin{aligned}
&||A||_2 \leq \left(||\Sigma_S^{-1}\Sigma^{1/2}||_2^2\right) \cdot \left(||I - \Sigma^{-1/2}\Sigma_s\Sigma^{-1/2}||_2^2\right) \leq \\
&\qquad\qquad \frac{\lambda_{\max}(\Sigma)}{\lambda^2} \cdot \left(||I - \Sigma^{-1/2}\Sigma_s\Sigma^{-1/2}||_2^2\right)
\end{aligned} \tag{30}
$$

Now consider $\Sigma^{-1/2}\Sigma_s\Sigma^{-1/2} = \Sigma^{-1/2}\frac{1}{s}\sum_{j\in S}(x_sx_s^T + \lambda I)\Sigma^{-1/2} = \frac{1}{s}\sum_{j\in s} Z_i$. Then note that $\mathbb{E}[Z_i] = \Sigma^{-1/2}(\frac{1}{n}\mathbb{E}[x_ix_i^T] + \lambda I)\Sigma^{-1/2} = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = I$, and that $\lambda_{max}(Z_i) \leq ||\Sigma^{-1}||_2||\frac{1}{n}x_ix_i^T + \lambda I||_2 \leq 1 + \frac{||\mathcal{X}||}{n\lambda}$.

Now we can bound $||I - \frac{1}{s}\sum_{j\in s} Z_i||_2$ by Theorem 2.2 in Tropp [2010]:

**Lemma 14.** *Tropp [2010] Let $Z_1, \dots Z_s$ sampled without replacement from $\{Z_1, \dots Z_n\}$. Then if $Z_i \in \mathbb{S}_d^+, \mathbb{E}[Z_i] = I_d$, and $\max_{i\in[n]} \lambda_{\max}(Z_i) \le B$ w.p. $1-\rho$, for $\delta = \sqrt{\frac{2B\log(2d/\rho)}{s}}$:*

$$\lambda_{\max}(\frac{1}{s}\sum_{i=1}^{s} Z_i) < 1+\delta, \;\; \lambda_{\min}(\frac{1}{s}\sum_{i=1}^{s} Z_i) > 1-\delta$$

So by Lemma 14, we know that with probability $1-\rho$: $|\lambda_{\min}(I - \Sigma^{-1/2}\Sigma_s\Sigma^{-1/2})| = 1 - \lambda_{\max}(\Sigma^{-1/2}\Sigma_s\Sigma^{-1/2}) \le \delta$, and similarly $|\lambda_{\max}(I - \Sigma^{-1/2}\Sigma_s\Sigma^{-1/2})| \le \delta$, thus $||I - \Sigma^{-1/2}\Sigma_s\Sigma^{-1/2})||_2 \le \delta$. Substituting this all into Equation 30 and noting $\lambda_{\max}(\Sigma) \le \frac{||\mathcal{X}||}{n} + \lambda$ we get with probability $1-\rho$:

$$||w_{i*}^{\lambda} - w_s||_{\frac{1}{n}X^TX} \le ||\tilde{w}_s - w_s||_{\Sigma} \le ||\mathcal{X}||||\mathcal{Y}||n||A||_2 \le ||\mathcal{X}||||Y||\frac{\lambda_{\max}(\Sigma)}{\lambda^2}\delta^2 = \tag{31}$$

$$2||\mathcal{X}||||\mathcal{Y}||\frac{(\frac{||\mathcal{X}||}{n}+\lambda)^2}{\lambda^3}\frac{\log(2d/\rho)}{s}, \tag{32}$$

which under the assumption $||\mathcal{X}|| = O(n\lambda)$ gives $O(\frac{||\mathcal{X}||||\mathcal{Y}||\log(2d/\rho)}{\lambda s})$. as desired. □

**Proof of Lemma 13:**

*Proof.* By Lemma 8, with $A = \frac{1}{s}X_S^TX + \lambda I$, $B = E_1, c = E_2, v = \frac{1}{n}^TY$, we get $w_s - \tilde{w}_s = (\frac{1}{s}X_S^TX + \lambda I + E_1)E_1w_s - (\frac{1}{s}X_S^TX + \lambda I + E_1)^{-1}E_2$, and so

$$||w_s - \tilde{w}_s||^2_{\frac{X^TX}{n}} \le 2||(\frac{1}{s}X_S^TX + \lambda I + E_1)^{-1}E_1w_s||^2_{\frac{X^TX}{n}} + 2||(\frac{1}{s}X_S^TX + \lambda I + E_1)^{-1}E_2||^2_{\frac{X^TX}{n}}$$

Under the assumption $||E_1||_2 \le \lambda/2$, this becomes

$$\begin{aligned}||w_s - \tilde{w}_s||^2_{\frac{X^TX}{n}} = O\left(||E_1w_s||_{(\frac{1}{s}X_S^TX_S+\lambda I)^{-1}(X^TX/n+\lambda I)(\frac{1}{s}X_S^TX_S+\lambda I)^{-1}}\right) + \\ O\left(||E_2||_{(\frac{1}{s}X_S^TX_S+\lambda I)^{-1}(X^TX/n+\lambda I)(\frac{1}{s}X_S^TX_S+\lambda I)^{-1}}\right)\end{aligned} \tag{33}$$

Now to apply Lemma 11, we need to bound

$$\begin{aligned}\mathrm{Tr}((\frac{1}{s}X_S^TX_S + \lambda I)^{-1}(X^TX/n + \lambda I)(\frac{1}{s}X_S^TX_S + \lambda I)^{-1}) \le d\lambda_{\max}(\Sigma_S)^{-1}(\Sigma)(\Sigma_S)^{-1}) = \\ d\lambda_{\max}(\Sigma^{-1/2}(\Sigma^{1/2}\Sigma_S^{-1}\Sigma^{1/2})^2\Sigma^{-1/2}) \le \\ d\lambda_{\max}(\Sigma^{-1})\frac{1}{\lambda_{\min}(\Sigma^{-1/2}\Sigma^{1/2}\Sigma_S^{-1}\Sigma^{1/2})}^2 \le \frac{d}{\lambda}(\frac{1}{1-\delta})^2,\end{aligned} \tag{34}$$

where the last inequality follows from Lemma 14. Applying Lemma 11 we get that with probability $1-2\rho$:

$$||w_{i*}^{\lambda} - w_s||_{X^TX/n} = O\left(\sigma_1^2 \cdot \frac{d}{\lambda}(\frac{1}{1-\delta})^2 \cdot ||w_s||^2\log(2d^2/\rho) + \sigma_2^2 \cdot \frac{d}{\lambda}(\frac{1}{1-\delta})^2\log(2d^2/\rho)\right), \tag{35}$$

From Lemma 14, $\delta = \sqrt{\frac{2(1+\frac{||\mathcal{X}||}{n\lambda})\log(2d/\rho)}{s}}$, which under the assumption $||\mathcal{X}|| = O(n\lambda)$ gives $\frac{1}{(1-\delta)}^2 = O(1 + \frac{\log(2d/\rho)}{s})^2$. Substituting in the value of $\sigma_1, \sigma_2$ gives:

$$\frac{||X||^4n^2}{s^4\epsilon^2}||\mathcal{W}||^2 \cdot \frac{d}{\lambda}(1+\frac{\log(2d/\rho)}{s}) \;+\; \frac{l||\mathcal{X}||^2||\mathcal{Y}||^2}{n^2\epsilon^2} \cdot \frac{d}{\lambda}(1+\frac{\log(2d/\rho)}{s}),$$

as desired. □