Differentially Private Regression with Unbounded Covariates

Jason Milionis
Columbia University

Alkis Kalavasis
NTUA

Dimitris Fotakis
NTUA

Stratis Ioannidis
Northeastern University

Abstract

We provide computationally efficient, differentially private algorithms for the classical regression settings of Least Squares Fitting, Binary Regression and Linear Regression with unbounded covariates. Prior to our work, privacy constraints in such regression settings were studied under strong a priori bounds on covariates. We consider the case of Gaussian marginals and extend recent differentially private techniques on mean and covariance estimation (Kamath et al., 2019; Karwa and Vadhan, 2018) to the sub-gaussian regime. We provide a novel technical analysis yielding differentially private algorithms for the above classical regression settings. Through the case of Binary Regression, we capture the fundamental and widely-studied models of logistic regression and linearly-separable SVMs, learning an unbiased estimate of the true regression vector, up to a scaling factor.

1 INTRODUCTION

Ever since the introduction of Differential Privacy (DP) by Dwork et al. (2006), differentially private variants of statistical estimation procedures have been a research topic of intense interest. The work on learning linear models alone is vast (see Cai et al. (2020); Wang (2018) for two recent reviews). Empirical Risk Minimization is also the impetus for the development of a broad array of new methods for DP-mechanism design, including output perturbation (Iyengar et al., 2019; Zhang et al., 2017; Jain and Thakurta, 2014), objective perturbation (Chaudhuri et al., 2011; Kifer et al., 2012), and gradient perturbation (Bassily et al., 2014; Abadi et al., 2016), to name a few.

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

Nevertheless, despite the intense interest on this topic, *all of the existing work on regression provides differential-privacy guarantees assuming bounded covariates.* Intuitively, this can be explained by inspecting even the simple least squares estimator used in linear regression. It is easy to see that estimator's *sensitivity*, i.e., its variability under changes on a single sample, is determined by the design matrix (i.e., the matrix of samples). As sensitivity has a direct effect on differential privacy guarantees, bounding the design matrix's eigenvalues is the prevalent approach for bounding the sensitivity. For this reason, assuming bounded covariates is a ubiquitous assumption in DP literature on both linear regression and learning generalized linear models.

This assumption is quite restrictive, and is frequently identified as a deficiency of DP regression algorithms from a practical standpoint (Anonymous, 2019). It is also a significant drawback from a theoretical standpoint, as it precludes studying DP-estimators on data sampled from distributions of *unbounded support*. Even the Gaussian distribution, perhaps the most commonly used generative distribution in statistical machine learning literature (Deng et al., 2021; Daskalakis et al., 2020; Kini and Thrampoulidis, 2020; Diakonikolas et al., 2019b; Nakkiran, 2019; Kreidler et al., 2018), cannot be used in conjunction with the existing DP regression algorithms and maintain DP guarantees.

Our work aims to directly address this, by providing DP algorithms for regression assuming (unbounded) Gaussian covariates. In doing so, we leverage and extend the recent work of Kamath et al. (2019), who proposed differentially private mechanisms for estimating the mean and the covariance matrix of high-dimensional Gaussian random vectors.

1.1 Contributions

Our first major contribution is to answer the following question in the affirmative:

Question 1. *Is private regression analysis with unbounded covariates possible?*

We study this problem in the context of three scenarios (see Section 4): Least Squares Fitting, Binary Regression, and (standard) Linear Regression. In all three, we assume (unbounded) Gaussian covariates.

In the Least Squares Fitting setting, given a training set $\{(\boldsymbol{X}_i, y_i)\}$, our goal is to efficiently and privately compute an estimate that is close to the Least Squares Estimate (LSE), i.e., the coefficients of the best-fitting linear function. In this problem, we assume that *labels* y_i are bounded, but make no further assumptions on how they relate to the covariates $\boldsymbol{X}_i \in \mathbb{R}^d$. Our main result is the following:

Informal Theorem 1. *For accuracy $\alpha > 0$ and privacy guarantees $\epsilon, \delta > 0$, there exists an efficient (ϵ, δ)-DP algorithm that, with high probability, approximates arbitrarily α-closely the Least Squares Estimate using $n = \widetilde{O}\left(d/\alpha^2 + d^{3/2}\log(1/\delta)/(\alpha\epsilon)\right)$ samples.*

In our second setting, Binary Regression, we further assume that labels are binary (i.e., $y_i = \pm 1$) and that covariates are zero mean. Moreover, labels are generated by a generalized linear model of the form $\mathbf{Pr}[y_i = +1|\boldsymbol{X}_i] = f(\boldsymbol{\beta}^T\boldsymbol{X}_i)$, where $f : \mathbb{R} \to [0, 1]$ is the model function and $\boldsymbol{\beta} \in \mathbb{R}^d$ is the true regression coefficient. This setting captures some of the most fundamental machine learning tasks, such as logistic regression and learning linearly-separable Support Vector Machines (SVMs). Our second main result is that the same differentially private estimator we used in Least Squares Fitting scenario can be applied to Binary Regression to obtain the following guarantees:

Informal Theorem 2. *For accuracy $\alpha > 0$ and privacy guarantees $\epsilon, \delta > 0$, there exists an efficient (ϵ, δ)-DP algorithm that, with high probability, approximates arbitrarily α-closely the true Binary Regression coefficient up to a multiplicative factor using $n = \widetilde{O}\left(d/\alpha^2 + d^{3/2}\log(1/\delta)/(\alpha\epsilon)\right)$ samples.*

Finally, we turn our attention to the (standard) Linear Regression setting. Here, labels are given by $y_i = \boldsymbol{\beta}^T\boldsymbol{X}_i + \epsilon_i$, where ϵ_i are i.i.d. zero-mean Gaussian noise variables and $\boldsymbol{\beta} \in \mathbb{R}^d$ is again the true regression coefficient. Note that, in contrast to the two previous settings, labels y_i here are unbounded. Our result follows:

Informal Theorem 3. *For accuracy $\alpha > 0$ and privacy guarantees $\epsilon, \delta > 0$, there exists an efficient (ϵ, δ)-DP algorithm that, with high probability, approximates arbitrarily α-closely the true Linear Regression coefficient using $n = \widetilde{O}\left(d/\alpha^2 + d^{3/2}\log(1/\delta)/(\alpha\epsilon)\right)$ samples.*

To the best of our knowledge, these results constitute the first efficient and private algorithms for regression analysis with unbounded feature vectors. From a technical standpoint, our analysis for Informal Theorem 1 and 2 relies on the fact that the LSE requires the calculation of the inverse of a moment matrix, as well as the expectation of a central random quantity $y_i\boldsymbol{X}_i$. This latter random quantity had not appeared before in Gaussian mean and covariance estimation procedures, but is key to regression settings. Our main conceptual contribution is that this quantity has sub-gaussian tails, hence, by extending the work of Kamath et al. (2019) and Karwa and Vadhan (2018) to sub-gaussian vectors, we manage to estimate it in a private and sample-efficient way. Finally, utilizing the above results, we show that we are also able to resolve the fundamental case of Linear Regression with unbounded features, indicated in Informal Theorem 3.

2 RELATED WORK

Differentially Private Regression and GLMs with Bounded Covariates. Linear regression is of course a true workhorse of statistics, and there has been a significant body of work on the design of computationally and statistically efficient differentially private regression algorithms (see e.g., the recent surveys of Cai et al. (2020); Wang (2018) and the references therein). Approaches include objective perturbation (Iyengar et al., 2019; Kifer et al., 2012; Zhang et al., 2012; Chaudhuri et al., 2011), output perturbation (Asi and Duchi, 2020; Iyengar et al., 2019; Zhang et al., 2017; Jain and Thakurta, 2014), gradient perturbation (Abadi et al., 2016; Bassily et al., 2014), subsample-and-aggregate (Barrientos et al., 2019; Dwork and Smith, 2010), and sufficient statistics perturbation (Alabi et al., 2020; Wang, 2018; McSherry and Mironov, 2009). Additionally, several works study generalizations of such mechanisms to Generalized Linear Models (GLMs) (Kulkarni et al., 2021; Iyengar et al., 2019; Jain and Thakurta, 2014; Kifer et al., 2012). Approaches that are used in typical regression settings also include variants of differentially private Stochastic Gradient Descent (DP-SGD) or other form of stochastic convex optimization (Feldman et al., 2020; Bassily et al., 2019; Wang et al., 2017; Zhang et al., 2017; Abadi et al., 2016; Bassily et al., 2014), which commonly require the optimization domain to be of bounded diameter. All above works, thus, either operate under a random setting with bounded covariates, or use a fixed design matrix X with bounded minimum eigenvalue on X^TX. Such strong assumptions on the boundedness of feature vectors are precisely the kind of assumptions that our work aims to mend.

Mean and Covariance Estimation. The study of differentially private mechanisms for mean and covariance estimation under bounded covariates is classic

(see, e.g., Amin et al. (2019); Dwork et al. (2014); McSherry and Mironov (2009)). Sheffet (2017) studies covariance estimation under Gaussian samples, also applying it to the Least Squares Fitting problem we study here; nevertheless, their differential privacy guarantee assumes an upper bound on covariates. Sheffet (2019) obtains a collection of DP algorithms that approximate the second moment matrix of the given dataset using existing Linear Regression techniques. We remark that, in each provided algorithm, an upper bound on the ℓ_2 norm of each row of the data matrix $A = [X|\boldsymbol{y}]$ is required. This upper bound does not hold in our Linear Regression setting since the received data (both X and $\boldsymbol{y}$) could be unbounded. Karwa and Vadhan (2018) resolve, for the first time, the problem of differentially private univariate Gaussian mean estimation without strong a priori bounds and with almost optimal dependence on problem parameters. Also in the univariate setting, Bun et al. (2015) learn more general distributions w.r.t. Kolmogorov distance, which is weaker than the total variation considered by Karwa and Vadhan (2018); Diakonikolas et al. (2015) extend this work to total variation distance, again for univariate distributions.

Kamath et al. (2019) extend the work of Karwa and Vadhan (2018) to multivariate mean and covariance estimation for high-dimensional Gaussian random vectors – see Section 3 for a description of their guarantees. Related to our setting, Cai et al. (2020) provide lower bounds for the sample complexity of differentially-private learning the mean of Gaussian random vectors, though the estimation algorithms they propose operate over bounded covariates. Recently, Aden-Ali et al. (2021) and Brown et al. (2021) studied privately learning multivariate Gaussians from an informational theoretic standpoint; however, no computational methods presently match these sample complexity bounds. The latter underscores difficulties arising in the unbounded covariates setting.

LSE for GLMs. The differentially private algorithm we propose applies Least Squares Estimation (LSE) to learn the parameters of a binary Generalized Linear Model (GLM) (see Theorem 4) and, more generally, to perform Least Squares Fitting over bounded labels (c.f. Theorem 3). It is well known that, under Gaussian marginals, LSE is an unbiased estimator of the parameter vector of a GLM, up to a scaling factor (Kadioglu et al., 2021; Erdogdu, 2016; Sun et al., 2014; Brillinger, 2012a). This is a consequence of Stein's Lemma (Liu, 1994; Stein, 1981) – see also Appendix A. In the binary setting, LSE can also be seen as a special case of the Linear Discriminant Analysis (LDA) classification algorithm (Hastie et al., 2009). Our Theorem 4 can thus also be seen as a differentially private version of LDA.

Concurrent Work. There has been vibrant independent and concurrent work to ours on Differential Privacy with connections to high-dimensional statistics (Liu et al., 2021a,b; Hopkins et al., 2021; Kothari et al., 2021; Ashtiani and Liaw, 2021; Kamath et al., 2021a,b). Recent works study the problem of privately learning arbitrary Gaussians (Kamath et al., 2021b; Ashtiani and Liaw, 2021; Kothari et al., 2021); these papers provide (among other things) mean and covariance estimation for arbitrary Gaussians and their techniques can be potentially adopted to extend our results accordingly. Moreover, Liu et al. (2021b) examine various statistical tasks (including linear regression) and propose a novel (but computationally inefficient) algorithm that achieves optimal sample complexity under minimal assumptions for these problems using robust statistics tools (see also Liu et al. (2021a) for private mean estimation). The work of Kamath et al. (2021a) studies differentially private stochastic convex optimization with heavy-tailed data under classical structural assumptions (e.g., smoothness of the loss function and boundedness of the parameter space); their techniques could be applied to regression problems too. Finally, Hopkins et al. (2021) examine the problem of mean estimation under minimal assumptions and pure DP using the framework of Sum of Squares.

3 PRELIMINARIES

Notation. We use bold fonts for vectors (e.g., $\boldsymbol{\beta}, \boldsymbol{y}$) and denote the set $\{1, \ldots, n\}$ as $[n]$. When $\boldsymbol{X}_i \in \mathbb{R}^d$ for $i \in [n]$ are the (random) feature vectors and $y_i \in \mathbb{R}$ for $i \in [n]$ are the (random) labels of a regression setting, the matrix $X = [\boldsymbol{X}_1 \ \boldsymbol{X}_2 \ \ldots \ \boldsymbol{X}_n]^T \in \mathbb{R}^{n \times d}$ is called the (random) design matrix and the vector $\boldsymbol{y} = [y_1 \ y_2 \ \ldots \ y_n]^T \in \mathbb{R}^n$ is called the (random) response vector. An extended techical preliminary, with definitions required for our proofs, is in Appendix A.

Differential Privacy. We use standard (ϵ, δ)-DP:

Definition 1 (Differential Privacy (Dwork et al., 2006))**.** *A randomized algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies (ϵ, δ)-differential privacy (equivalently, is said to be (ϵ, δ)-DP) if for every pair of neighboring datasets $X, X' \in \mathcal{X}^n$ that differ on at most one element,*

$$\mathbf{Pr}[M(X) \in Y] \leq \exp(\epsilon)\, \mathbf{Pr}[M(X') \in Y] + \delta\,, \forall\, Y \subseteq \mathcal{Y}\,.$$

A crucial tool for differential privacy is the adaptive composition theorem, providing the privacy properties of a sequence of algorithms $M_1(X), \ldots, M_N(X)$, where the i-th algorithm may depend on the outcomes of the algorithms $M_1(X), \ldots, M_{i-1}(X)$, for $i \in [N]$.

Fact 1 (Composition of differentially private mechanisms (Dwork et al., 2006, 2010)). *If M is an adaptive composition of differentially private algorithms $M_1, \ldots, M_N$, where M_i is (ϵ, δ_i)-DP for any $i \in [N]$, then it holds that M is $(\epsilon N, \sum_{i=1}^N \delta_i)$-DP and, for every $\delta > 0$, M is $(\epsilon\sqrt{6N\log(1/\delta)}, \delta + \sum_{i=1}^N \delta_i)$-DP.*

DP Gaussian Parameter Estimation. At a technical level, our work extends the tools developed by Kamath et al. (2019) to privately estimate the mean $\boldsymbol{\mu}$ and covariance Σ of a d-dimensional Gaussian distribution. Their algorithm, which we call LearnGaussian-hd, has the following guarantee:

Theorem 2 (Multivariate Gaussian Estimation (Kamath et al., 2019)). *There exists a polynomial time $(\epsilon^2/2 + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-DP algorithm* LearnGaussian-hd *that takes at least*

$$n = \widetilde{O}\left(\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\epsilon} + \frac{d^{3/2}\log^{1/2}(\kappa) + d^{1/2}\log^{1/2}(R)}{\epsilon}\right)$$

i.i.d. samples $\boldsymbol{X}_i$, $i \in [n]$, from a d-dimensional Gaussian $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with unknown mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and unknown covariance $\Sigma \in \mathbb{R}^{d\times d}$ satisfying $\|\boldsymbol{\mu}\|_2 \le R$ and $\mathbb{I}_d \preceq \Sigma \preceq \kappa\mathbb{I}_d$, and outputs estimates $\widehat{\boldsymbol{\mu}}, \widehat{\Sigma}$ such that, with high probability, $\mathrm{TV}(\mathcal{N}(\boldsymbol{\mu}, \Sigma), \mathcal{N}(\widehat{\boldsymbol{\mu}}, \widehat{\Sigma})) \le \alpha$.

We remark that this TV distance bound is implied by the parameter estimation of the mean and covariance matrix in Mahalanobis distance. In short, LearnGaussian-hd produces differentially private estimates of the distribution's parameters using only $\widetilde{O}(d^2)$ samples. It operates under the following boundedness assumptions for the distributional parameters:

$$\|\boldsymbol{\mu}\|_2 \le R \quad \text{and} \quad \mathbb{I}_d \preceq \Sigma \preceq \kappa\mathbb{I}_d,$$

even though, crucially, *the samples $\boldsymbol{X}_i$ themselves are unbounded.* Moreover, both upper bounds (R and κ) are mild and well-motivated: even if LearnGaussian-hd is applied to a sequence of datasets where these grow sub-exponentially, the sample complexity remains polynomial. Additionally, notice that $\mathbb{I}_d \preceq \Sigma$ comes w.l.o.g.: as long as the smallest eigenvalue of Σ is non-zero, we can rescale the vectors $\boldsymbol{X}_i$ to ensure that this holds. If an eigenvalue of Σ is zero, then the distribution is degenerate: we can then apply LearnGaussian-hd in the subspace spanned by the features (in which Σ will have full rank).

Kamath et al. (2019) efficiently learn a symmetric matrix A, termed the *preconditioner* of the Gaussian distribution, that satisfies $\mathbb{I}_d \preceq A\Sigma A \preceq O(1)\mathbb{I}_d$. Multiplying the input samples with this preconditioner thus makes the Gaussian inputs nearly spherical, which reduces the geometry to the one-dimensional setting, previously studied by Karwa and Vadhan (2018).

4 PROBLEM FORMULATION

In this section, we formally define the regression settings we are interested in, namely, the Least Squares Fitting, the Binary Regression and the (standard) Linear Regression problems, as well as the associated technical assumptions we make.

Least Squares Fitting. In the Least Squares Fitting problem, we observe labeled examples $(\boldsymbol{X}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, and wish to produce an (ϵ, δ)-differentially private version of the Least Squares Estimator (LSE):

$$\boldsymbol{\beta}^\star = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} \sum_{i=1}^n \left(y_i - \boldsymbol{\beta}^T\boldsymbol{X}_i\right)^2 \tag{4.1}$$

$$= \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{X}_i\right) \tag{4.2}$$

$$= \left(\frac{1}{n}X^TX\right)^{-1}\frac{1}{n}X^T\boldsymbol{y}\,, \tag{4.3}$$

where $X = [\boldsymbol{X}_i]_{i=1}^n \in \mathbb{R}^{n\times d}$ is the matrix with feature vectors as rows and $\boldsymbol{y} = [y_i]_{i=1}^n \in \mathbb{R}^d$ is the vector of labels, respectively. In contrast to the Binary and Linear Regression problems below, we make no prior assumption on how labels y_i are linked to features $\boldsymbol{X}_i$; crucially, our differentially private algorithm *must not rely* on any presumed boundedness of features $\boldsymbol{X}_i$. We make the following technical assumption:

Assumption 1. *Labeled examples $(\boldsymbol{X}_i, y_i)$, $i = 1, \ldots, n$, are i.i.d. Moreover, $\boldsymbol{X}_i \in \mathbb{R}^d$ are sampled from a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ satisfying the following conditions:*

$$\|\boldsymbol{\mu}\|_2 \le R \quad \textit{and} \quad \mathbb{I}_d \preceq \Sigma \preceq \kappa\mathbb{I}_d\,, \tag{4.4}$$

while the labels satisfy $\frac{1}{\rho} \le |y_i| \le c$ for some universal parameters $\rho, c, \kappa, R > 0$.

The assumptions in Eq. (4.4) are also made by Kamath et al. (2019) in the context of Gaussian estimation. As discussed in Section 3, both upper bounds are natural, while the lower bound on the covariance comes without any loss of generality. Crucially, in contrast to the majority of prior works on regression, samples $\boldsymbol{X}_i$ are indeed unbounded, as they are sampled from $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$. Finally, the boundedness of the outputs y_i, $i \in [n]$, is a requirement we share with other works (e.g., Alabi et al. (2020); Wang (2018); Kifer et al. (2012); Zhang et al. (2012)), and clearly applies to, e.g., binary classification; we also study unbounded labels in the Linear Regression setting.

Binary Regression. In the Binary Regression setting, we additionally assume that the labels y_i are binary (i.e., $y_i \in \{-1, +1\}$), and are produced by a

Generalized Linear Model (GLM) linking these binary labels to features. In contrast to the previous setting, this GLM is parameterized by a "true" $\boldsymbol{\beta} \in \mathbb{R}^d$ (see Assumption 2 below). Our goal is to give an estimate of this $\boldsymbol{\beta}$ again via *the same* (ϵ, δ)-differentially private version of the LSE given by Eq. (4.1). In particular, *in addition* to Assumption 1, we make the following assumption in the Binary Regression setting:

Assumption 2. *There exists a $\boldsymbol{\beta} \in \mathbb{R}^d$ such that, given $\boldsymbol{X}_i \in \mathbb{R}^d$ and for all $i \in [n]$,*

$$\mathbf{Pr}[y_i = +1 | \boldsymbol{X}_i] = f(\boldsymbol{\beta}^T \boldsymbol{X}_i), \tag{4.5}$$

where $f : \mathbb{R} \to [0,1]$ is a non-decreasing, continuously differentiable function satisfying $\lim_{x\to-\infty} f(x) = 0$ and $\lim_{x\to\infty} f(x) = 1$. Moreover, the features $\boldsymbol{X}_i$ are zero-mean, i.e., $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}_i] = \mathbf{0}$.

The probabilistic model defined by Eq. (4.5) holds for many important practical settings. For instance, it holds for logistic regression, where the link function is $f(x) = 1/(1+e^{-x})$. It also holds for Support Vector Machines (SVMs) with linearly separable data. We discuss this in more detail in Appendix B.

Finally, our assumption that $\boldsymbol{\mu} = \mathbf{0}$ is common (see, e.g., Kulkarni et al. (2021); Cai et al. (2020); Daskalakis et al. (2020); Bernstein and Sheldon (2019); Sheffet (2017); Erdogdu (2016)) and well-motivated in the context of our Binary Regression setting: even ignoring privacy considerations, the sample complexity guarantees of any estimator will degrade rapidly as $\boldsymbol{\mu}$ gets farther away from the origin. This is precisely because, under Gaussian covariates, the fraction of samples of one class will decrease exponentially as the distance of $\boldsymbol{\mu}$ from the separating hyperplane (that passes through the origin) increases.

Linear Regression. A natural question is whether we can extend our guarantees beyond bounded labels. To this end, we finally consider the standard Linear Regression setting (with Gaussian errors):

Assumption 3. *Labeled examples $(\boldsymbol{X}_i, y_i)$, $i = 1, \ldots, n$, are i.i.d., where $\boldsymbol{X}_i \in \mathbb{R}^d$ are sampled from the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ satisfying $\mathbb{I}_d \preceq \Sigma \preceq \kappa \mathbb{I}_d$ for some universal parameter $\kappa > 0$. Moreover, there exists a $\boldsymbol{\beta} \in \mathbb{R}^d$ and a $\sigma_\epsilon > 0$ such that, given $\boldsymbol{X}_i \in \mathbb{R}^d$,*

$$y_i = \boldsymbol{\beta}^T \boldsymbol{X}_i + \epsilon_i, \quad \text{for all } i = 1, \ldots, n\,, \tag{4.6}$$

where ϵ_i are i.i.d. samples from $\mathcal{N}(0, \sigma_\epsilon^2)$.

Note that, in this setting, labels y_i are themselves Gaussian and, therefore, unbounded. Our goal here is again to produce a differentially private estimate for the "ground truth" vector $\boldsymbol{\beta}$.

5 MAIN RESULTS

We formally state our results in this section. Our theorems provide $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-DP guarantees for the Least Squares Fitting, the Binary and Linear Regression settings. This guarantee is, in essence, equivalent to (ϵ, δ)-DP. For a more detailed discussion on this issue, we refer the reader to Appendix C.1. We focus here on the statement of our main results and conclusions drawn from them; an overview of the technical challenges we face when proving these results and the novel techniques we employ to address them can be found in Section 6.

5.1 Least Squares Fitting

Our differentially private LSE for the Least Squares Fitting setting is summarized in Algorithm 1. In short, we compute DP estimates of the quantities

$$(X^T X/n)^{-1} \quad \text{and} \quad X^T \boldsymbol{y}/n,$$

whose product, by Eq. (4.3), yields the LSE $\boldsymbol{\beta}^\star$.

The estimation of the first quantity proceeds as follows. Having access to the n i.i.d. samples $(\boldsymbol{X}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, where $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma), i \in [n]$, Algorithm 1 initially privately computes differentially private estimates $(\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}, \widehat{\Sigma}_{\boldsymbol{X}})$ of the mean and covariance matrix of the d-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, using the algorithm LearnGaussian-hd, discussed in Section 3. These estimates, that satisfy the guarantees indicated in Theorem 2, can be used to estimate $(X^T X/n)^{-1}$ via the relationship:

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T \approx \mathbb{E}[\boldsymbol{X}_i \boldsymbol{X}_i^T] \approx \widehat{\Sigma}_{\boldsymbol{X}} + \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}} \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}^T\,.$$

The second quantity, i.e., the term $X^T \boldsymbol{y}/n$, is somewhat harder to estimate in a differentially private fashion, as constituent terms $y_i \boldsymbol{X}_i$ are *not* Gaussian. The boundedness of variables y_i, however, ensures that these terms are sub-gaussian. As an important technical contribution, we design differentially private algorithms that operate in the sub-gaussian regime (see LearnSubGaussian-hd in Appendix C.3), extending the analysis of Kamath et al. (2019) and Karwa and Vadhan (2018), and obtain a private mean estimate $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X},y}$ for the sub-gaussian random vectors $y_i \boldsymbol{X}_i$.

Armed with these estimates, our differentially private LSE is finally given by:

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\Sigma}_{\boldsymbol{X}} + \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}} \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}^T\right)^{-1} \widehat{\boldsymbol{\mu}}_{\boldsymbol{X},y}\,, \tag{5.1}$$

whose privacy follows from appropriate composition rules. We refer to the resulting algorithm, summarized in Algorithm 1, as PrivLearnLSE. Our main

result with respect to the privacy and accuracy of this estimator is as follows:

Theorem 3 (Privacy and Accuracy of $\widehat{\boldsymbol{\beta}}$ in Private Least Squares Fitting)**.** *Under Assumption 1 with parameters (κ, c, ρ, R), for all privacy parameters $\epsilon, \delta > 0$, accuracy parameters $\alpha, \eta > 0$ and confidence $\gamma \in (0,1)$,* PrivLearnLSE *(defined in Algorithm 1) is $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-differentially private. Moreover, if the number of labeled examples is at least:*

$$n = \widetilde{O}\left(\frac{d^{3/2}\sqrt{\log(\kappa\rho c)}\cdot \mathrm{polylog}\left(\frac{1}{\gamma\delta}\right)}{\eta^2\epsilon}\right) + (1+R)\cdot\widetilde{O}\left(\frac{d^{3/2}\sqrt{\log\kappa}\cdot\mathrm{polylog}\left(\frac{1}{\gamma\delta}\right)}{\alpha^2\epsilon}\right),$$

then, PrivLearnLSE *runs in* $\mathrm{poly}(n)$ *time and, with probability at least $1 - O(\gamma)$, successfully returns an estimate $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$ that satisfies:*

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\right\|_2^2 \le O\left(\alpha^2\right)\cdot\left\|\Sigma^{1/2}\boldsymbol{\beta}^\star\right\|_2^2 + O\left(\eta^2\right)\cdot c^2\,,$$

with respect to the LSE $\boldsymbol{\beta}^*$.

We have provided a simplified bound in the number of samples; the precise sample complexity and the theorem's proof can be found in Appendix D. For a proof sketch, we refer to Section 6.1. Intuitively, the number of samples we require grows as $\widetilde{O}(d^{3/2})$, slightly more favorably than the covariance estimation case of Kamath et al. (2019). Moreover, the number of samples again grows polylogarithmically on κ (the bound on the covariance spectral norm) but linearly (rather than polylogarithmically) on R, the bound on the mean.

The LearnGaussian-HD routine (see also **Line 6** of Algorithm 1) is the algorithm of Kamath et al. (2019) (as discussed in Section 3 and Appendix C.2). The original algorithm by Kamath et al. (2019) requires knowledge of both upper bounds κ and R, but by switching the privacy guarantee from zero-concentrated DP to (ϵ, δ)-DP, we remove the requirement of prior knowledge of R, even though we still require κ as input. This adaptation can be found in Appendix C.1. In contrast, our routine LearnSubGaussian-HD (see **Line 8** of Algorithm 1), described in Appendix C.2, departs from the one of Kamath et al. (2019), the main difference being that it operates (and comes with guarantees for) sub-gaussian vectors. For further details about the modifications required to accomplish this, we refer the reader to Appendix C.3.

[1]The invertibility of the matrix in **Line 10** holds with high probability; we account for the bad non-invertibility event in the $O(\gamma)$ failure probability of Theorem 3.

Algorithm 1 Differentially Private LSE.

1: **Input:** $(X, \boldsymbol{y}) = (\boldsymbol{X}_i, y_i)_{i\in[n]}$ with $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu}, \Sigma$ are unknown and n satisfies Theorem 3.
2: **Parameters:** Privacy $\epsilon, \delta > 0$, accuracy $\alpha, \eta > 0$, confidence $\gamma \in (0,1)$, covariance spectral norm bound κ, upper bound of labels c.
3: **Output:** Estimate $\widehat{\boldsymbol{\beta}}$ that approaches the LSE $\boldsymbol{\beta}^\star$ in L_2 norm with high probability.
4: **procedure** PrivLearnLSE$((X, \boldsymbol{y}), \epsilon, \delta, \alpha, \eta, \gamma, \kappa)$
5: $\quad L \leftarrow \{\Theta(\epsilon), \Theta(\delta), \Theta(\alpha), \gamma, \kappa\}$
6: $\quad (\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}, \widehat{\Sigma}_{\boldsymbol{X}}) \leftarrow$ LearnGaussian-HD$(\{\boldsymbol{X}_i\}_i, L)$
7: $\quad L \leftarrow \{\Theta(\epsilon), \Theta(\delta), \Theta(\eta), \gamma, c^2\kappa\}$
8: $\quad \widehat{\boldsymbol{\mu}}_{\boldsymbol{X},y} \leftarrow$ LearnSubGaussian-HD$(\{y_i\boldsymbol{X}_i\}_i, L)$
9: $\quad \mathrm{M} \leftarrow \widehat{\Sigma}_{\boldsymbol{X}} + \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}^T$
10: $\quad$ **if** M is not invertible[1] **then** Output $\perp$
11: $\quad$ Output the private estimate $\widehat{\boldsymbol{\beta}} = \mathrm{M}^{-1}\widehat{\boldsymbol{\mu}}_{\boldsymbol{X},y}$

5.2 Binary Regression

We next turn our attention to the Binary Regression setting, in which *both* Assumption 1 and Assumption 2 apply. We study the properties of PrivLearnLSE (Algorithm 1) under these assumptions; the only (slight) modification of Algorithm 1, compared to the previous setting, is that we no longer need the estimate $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}$, as Assumption 2 states that $\boldsymbol{\mu} = \mathbf{0}$. Hence, we set $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}} = \mathbf{0}$ in Eq. (5.1), with the remaining terms computed as in the previous section. We show that the resulting algorithm has the following guarantees:

Theorem 4 (Privacy and Accuracy of $\widehat{\boldsymbol{\beta}}$ in Private Binary Regression)**.** *Under Assumption 1 with covariance parameter κ and Assumption 2 with true parameter $\boldsymbol{\beta} \in \mathbb{R}^d$, for every privacy parameters $\epsilon, \delta > 0$, accuracy parameters $\alpha, \eta > 0$ and confidence $\gamma \in (0,1)$,* PrivLearnLSE *(defined in Algorithm 1) with $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}} = \mathbf{0}$ is $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-differentially private. Moreover, if the number of labeled examples is at least:*

$$n = \widetilde{O}\left(\frac{d^{3/2}\cdot\mathrm{polylog}\left(\frac{1}{\gamma\delta}\right)}{\epsilon}\cdot\max\left\{\frac{\sqrt{\log\kappa}}{\eta^2}, \frac{1}{\alpha^2}\right\}\right),$$

then PrivLearnLSE *runs in* $\mathrm{poly}(n)$ *time and, with probability at least $1 - O(\gamma)$, successfully returns an output estimate $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$ that satisfies*

$$\|\widehat{\boldsymbol{\beta}} - k\boldsymbol{\beta}\|_2^2 \le O(\alpha^2)\left(1 + \left\|k\Sigma^{1/2}\boldsymbol{\beta}\right\|_2^2\right) + O(\eta^2)\,, \quad (5.2)$$

where $k = \frac{2n}{n-d-1}\,\mathbb{E}\left[f'\left(\boldsymbol{\beta}^T\boldsymbol{X}_i\right)\right]$.

As before, we have provided a simplified version of the exact number of samples. The exact expression and the proof of the theorem are in Appendix E. The proof sketch can be found at Section 6.2. As in Theorem 3,

the sample complexity grows as $d^{3/2}$, and is merely polylogarithmic on κ. Moreover, as in classic (non-DP) work on binary regression via LSE (Erdogdu, 2016; Sun et al., 2014; Brillinger, 2012a), our estimator learns the underlying "true" $\boldsymbol{\beta}$ up to a scaling factor k, that depends on the "sharpness" of the model function f (via its derivative f'). We note that, to discover the hyperplane separating positive from negative labels, it indeed suffices to learn only the direction of $\boldsymbol{\beta}$, not its magnitude, since a separating hyperplane is fully defined by this direction.

To further elaborate on the effect of k: by Assumption 2, $\boldsymbol{\beta}^T \boldsymbol{X}_i$ is a zero mean Gaussian, while f' tends to zero as its argument reaches either $+\infty$ or $-\infty$. Hence, the expectation that determines k very much depends by the behavior of f' around 0. That is, if f is relatively flat (i.e., binary labels are "noisy"), k will be small, and more samples will be needed to achieve a better numerical accuracy in Eq. (5.2); the converse is true when f is "sharp" (e.g., a sigmoid close to the sign function), and labels are less noisy. This dependence of the estimate accuracy on the noise inherent in the GLM (via the model function f) is natural.

5.3 Linear Regression

In this model, the labels y_i are assumed to be generated from an underlying "true" linear model $\boldsymbol{\beta}^T \boldsymbol{X}_i$ (with a Gaussian error), thus being unbounded, for some regression coefficient $\boldsymbol{\beta} \in \mathbb{R}^d$. Our goal is to estimate this "true" underlying $\boldsymbol{\beta}$ in a differentially private way. We provide the following algorithm for this task. For each drawn labeled example $(\boldsymbol{X}, y)$, the algorithm creates the vector $\boldsymbol{Z} = (\boldsymbol{X}, y)^T \in \mathbb{R}^{(d+1)}$. Observe that this random vector is also Gaussian with a covariance matrix $\Sigma' \in \mathbb{R}^{(d+1)\times(d+1)}$, given by:

$$\Sigma' = \begin{bmatrix} \Sigma & \Sigma\boldsymbol{\beta} \\ \boldsymbol{\beta}^T\Sigma & \sigma_\epsilon^2 + \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} \end{bmatrix}, \tag{5.3}$$

where $\boldsymbol{\mu}, \Sigma, \sigma_\epsilon^2$ are the parameters of Assumption 3. The algorithm, then, proceeds as follows. First, it computes a differentially private estimate $\widehat{\Sigma}$ of Σ using n samples of $\boldsymbol{X}_i$ via the routine LearnGaussian-hd, discussed in Section 3. Then, using n additional samples $\boldsymbol{Z}_i = (\boldsymbol{X}_i, y_i)^T$, it computes a differentially private estimate $\widehat{\Sigma'}$, again via LearnGaussian-hd. From Eq. (5.3), the first d elements of the last column of $\widehat{\Sigma'}$ can be used as a DP estimate $\widehat{\Sigma\boldsymbol{\beta}}$ of $\Sigma\boldsymbol{\beta}$.[2] Finally, the algorithm uses these two estimates to output:

$$\widehat{\boldsymbol{\beta}} = \widehat{\Sigma}^{-1}\widehat{\Sigma\boldsymbol{\beta}}.$$

[2]Note that the first d columns and rows of $\widehat{\Sigma'}$ can also be used as a DP estimate $\widehat{\Sigma'}$ of Σ; we nevertheless estimate this separately, to ensure the statistical independence of the two estimates.

A formal description of this algorithm can be found in Algorithm 2 in Appendix F. Our result with respect to its privacy and accuracy is as follows:

Theorem 5 (Privacy and Accuracy of $\widehat{\boldsymbol{\beta}}$ in Private Linear Regression)**.** *Under Assumption 3 with parameter κ and true vector $\boldsymbol{\beta} \in \mathbb{R}^d$, for all privacy parameters $\epsilon, \delta > 0$, accuracy parameters $\alpha, \eta > 0$ and confidence $\gamma \in (0,1)$, there exists an algorithm (see Algorithm 2) that is $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-differentially private, and if the number of samples is at least:*

$$n = \widetilde{O}\left(\frac{d^{3/2}\cdot \mathrm{polylog}\left(\frac{1}{\gamma\delta}\right)}{\epsilon}\max\left\{\frac{\sqrt{\log(\kappa(\Sigma'))}}{\eta^2}, \frac{1}{\alpha^2}\right\}\right),$$

then, it runs in $\mathrm{poly}(n)$ *time and, with probability at least $1 - O(\gamma)$, the output estimate $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$ and the "true" regression coefficient $\boldsymbol{\beta}$ satisfy:*

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2^2 \le O\left(\alpha^2\right)\cdot\left\|\Sigma^{1/2}\boldsymbol{\beta}\right\|_2^2 + O\left(\eta^2\right)\cdot\lambda_{max}^2(\Sigma'),$$

where $\kappa(\Sigma') = \frac{\lambda_{max}(\Sigma')}{\lambda_{min}(\Sigma')}$ is the condition number of the block matrix Σ' as in Eq. (5.3).

The exact sample complexity bound and the theorem's proof can be found in Appendix F. A short proof sketch is provided in Section 6.3. As in our previous results, the sample complexity scales as $d^{3/2}$; also, it is polylogarithmic on the condition number of Σ'.[3]

6 TECHNICAL OVERVIEW

In this section, we provide a sketch of our technical contributions with respect to the proofs of Theorems 3, 4, and 5.

6.1 Theorem 3: Proof Sketch

We begin with Theorem 3, which deals with the Least Squares Fitting problem. Our goal is to privatize the Least Squares Estimator (see Eq. (4.1)) without significant accuracy loss. Hence, the differentially private algorithm (see Algorithm 1) computes a quantity $\widehat{\boldsymbol{\beta}}$ that is asymptotically the same as the Least Squares Estimate of Eq. (4.1):

$$\boldsymbol{\beta}^\star = \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{X}_i\right). \tag{6.1}$$

The structure of this estimate (product of two terms) suggests privatizing each term separately, thereby motivating Algorithm 1. To ensure that the desired

[3]Again, the non-invertibility of the matrix $\widehat{\Sigma}$ is a low probability event and is fully captured by the $O(\gamma)$ probability of failure, as also indicated in our proof.

privacy property holds, the key idea is to apply the composition of differentially private mechanisms (see Fact 1), hence affording privacy to the whole algorithm. It thus suffices to consider privatized estimates of the individual terms.

The key conceptual observation for our main result is that the second term in Eq. (6.1) consists, in fact, of sub-gaussian vectors. At a technical level, we have to expand the mean and covariance estimation procedures for Gaussian distributions to the sub-gaussian regime. More to that, in order to reduce as much as possible the dependence on the range of the mean value R of the feature vectors $\boldsymbol{X}_i$, we modify the multivariate mean estimation analysis of Kamath et al. (2019) to hold for unbounded mean feature vectors. As a technical tool, we use an alternative guarantee (see Lemma 16 in Appendix C.3) on mean estimation which allows us to disengage the concentration bounds from the bound on the mean, in the case of (ϵ, δ)-DP.

Even using those variants of the algorithms, we still have to satisfy a stronger privacy desideratum. In particular, Theorem 3 requires privacy guarantees for pairs $(\boldsymbol{X}_i, y_i)$. However, **Line 8** of Algorithm 1 affords privacy guarantees for the entire sub-gaussian terms $y_i \boldsymbol{X}_i$. So, it is not straightforward how to achieve the more general privacy guarantee of altering the individual $(\boldsymbol{X}_i, y_i)$ pairs. In Appendix D.1, we establish the desired privacy guarantee for $(\boldsymbol{X}_i, y_i)$.

For the desired accuracy guarantee on Algorithm 1, we have to control the quantity $\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\right\|_2^2$ (see Appendix D.2). At a first sight, the above expression cannot be handled by standard concentration of measure phenomena. However, we provide a non-trivial decomposition:

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star = \left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}} \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}^T\right)^{-1} \left(-Q_1 \boldsymbol{\beta}^\star + \boldsymbol{Q}_2\right) ,$$

using the below quantities that we introduce:

$$Q_1 = \widehat{\Sigma} + \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}} \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}^T - \frac{1}{n} X^T X, \quad \text{and} \quad \boldsymbol{Q}_2 = \widehat{\boldsymbol{\mu}}_{\boldsymbol{X},y} - \frac{1}{n} X^T \boldsymbol{y} ,$$

where $\widehat{\Sigma}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{X},y}$ are the private outputs of the algorithms described in Algorithm 1, and $X, \boldsymbol{y}$ are the design matrix and the labels vector, respectively. This decomposition, when altered in geometry for normalization purposes by a transformation $\boldsymbol{w} = \Sigma^{1/2} \boldsymbol{\beta}$ and $\widehat{\boldsymbol{w}} = \Sigma^{1/2} \widehat{\boldsymbol{\beta}}$, enables us to control each term individually and obtain the desired bounds. The intuition behind this decomposition lies in the fact that both Q_1 and $\boldsymbol{Q}_2$ vanish asymptotically (and so $\widehat{\boldsymbol{\beta}}$ tends to $\boldsymbol{\beta}^\star$), as the number of samples n increases.

The bounds on $Q_1, \boldsymbol{Q}_2$ are handled by further decomposing into the difference of private quantities and their actual values $(\Sigma, \boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\mu}_{\boldsymbol{X},y})$ and between empirical quantities and the actual values. To obtain tighter bounds on the individual terms of difference of private quantities and actual values, we use the private preconditioner matrix in our analysis, which allows us to avoid a strict dependence on the largest eigenvalue κ of the covariance matrix Σ in our bounds (see Theorem 3). For a detailed proof of Theorem 3, see Appendix D.

6.2 Theorem 4: Proof Sketch

As far as our second main result (Theorem 4) is concerned, the key conceptual contribution is to introduce a new estimator $\boldsymbol{\beta}_s^\star$ (solely for the purposes of the analysis) that is defined with the help of n additional samples $(\boldsymbol{X}_i, y_i)$ (for a total of $2n$ samples) as follows:

$$\boldsymbol{\beta}_s^\star = \left(\frac{1}{n} \sum_{i=n+1}^{2n} \boldsymbol{X}_i \boldsymbol{X}_i^T\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} y_i \boldsymbol{X}_i\right) .$$

This estimate resembles the Least Squares Estimate $\boldsymbol{\beta}^\star$ but *crucially introduces independence* between the two terms that constitute the Least Squares Estimate. This independence of the two terms is pivotal for proving that the estimate $\boldsymbol{\beta}_s^\star$ is an unbiased up to a multiplicative factor estimate of the true regression coefficient $\boldsymbol{\beta}$. In turn, this crucial observation is used to prove that our private estimate $\widehat{\boldsymbol{\beta}}$ (see Algorithm 1) is close to the true regression coefficient $\boldsymbol{\beta}$ up to a multiplicative factor, since the proof of Theorem 3 holds even for the Least-Squares-resembling estimate $\boldsymbol{\beta}_s^\star$ (because of the independent handling of the aforementioned quantities $Q_1, \boldsymbol{Q}_2$). At a technical level, the above discussion is a result of probabilistic tools, such as the high-dimensional geometry of Wishart matrices (Anderson, 2003). For a detailed proof of Theorem 4, see Appendix E.

6.3 Theorem 5: Proof Sketch

Finally, we briefly discuss the techniques behind Theorem 5. Recall that for the standard Linear Regression problem with true vector $\boldsymbol{\beta}$, our algorithm outputs the private estimate $\widehat{\boldsymbol{\beta}} = \widehat{\Sigma}^{-1} \widehat{\Sigma \boldsymbol{\beta}}$, as mentioned after Eq. (5.3). On one hand, the privacy guarantee follows from the composition theorems. On the other hand, for the accuracy guarantee, we have to control the quantity $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$. The main technical challenge for this step it to provide tight bounds for the eigenvalues of the block matrix Σ' of Eq. (5.3). In particular, we have to draw sufficiently many samples in order to control the quantities $\|\Sigma^{1/2} \widehat{\Sigma}^{-1} \Sigma^{1/2}\|_2^2$ and $\|\Sigma^{-1/2} (\widehat{\Sigma \boldsymbol{\beta}} - \Sigma \boldsymbol{\beta})\|_2^2$ dealing with our estimates $\widehat{\Sigma}$ and $\widehat{\Sigma \boldsymbol{\beta}}$. The first quantity is a constant, given

roughly $n = \Omega(d^{3/2}\sqrt{\log \kappa}/\epsilon)$ samples, using properties of the LearnGaussian-HD algorithm and concentration of random matrices. The second quantity is more challenging and is controlled by the maximum eigenvalue of Σ', with high probability, after roughly $n = \Omega(d^{3/2}\sqrt{\log(\kappa(\Sigma'))}/\epsilon)$ samples are drawn, where $\kappa(\Sigma')$ is the condition number of the block matrix of Eq. (5.3). To upper bound the condition number, we exploit bounds for eigenvalues of block matrices (Ma and Zarowski, 1995), and show that in our setting, these are tight for $\kappa(\Sigma')$ (see Appendix F.1).

7 CONCLUSION

We provide and analyze estimators for inference in three regression settings with unbounded covariates, formally proving that they are private and efficient. We believe that the line of work on unbounded covariates is of great interest with respect to both theory and practice. Potential future research based on this work includes, for instance, relaxing the i.i.d. assumptions on the provided data (to account for potential dependencies among feature vectors). In addition, lower bounds in differentially private regression regimes are either elusive or sub-optimal (see, e.g., Wang (2018)); examining possible lower bounds in unbounded regimes for regression-like environments is another promising future direction.

7.1 Limitations

For the above analysis, we have considered the case of Gaussian marginals and have extended recent differentially private techniques on mean and covariance estimation (Kamath et al., 2019; Karwa and Vadhan, 2018) to the sub-gaussian regime. For the detailed hypotheses upon which the aforementioned procedures were provided, the reader is encouraged to review Section 4, where all of the relevant assumptions are clearly indicated.

The focus of this work is in its nature theoretical. Supplementally to the theory, we believe that the community would benefit from additional experimental studies of the proposed methods. In fact, the design of practical algorithms is a strand of research of significant independent interest, since practical applications are able to immensely benefit from unbounded estimation procedures: see, e.g., the work of Biswas et al. (2020) that considers practical differentially private Gaussian mean and covariance estimation procedures. Thus, we believe that the practical extension of our results and relevant experiments are a natural and interesting premise for future work.

Acknowledgements

We thank the anonymous reviewers for useful remarks and comments on the presentation of our manuscript. The most significant part of this work was performed while Jason Milionis was an undergraduate student at the National Technical University of Athens. This work was partially supported by a research fellowship from the Costas M. Lemos Foundation. Dimitris Fotakis and Alkis Kalavasis were supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant," project BALSAM, HFRI-FM17-1424. Stratis Ioannidis was supported by the National Science Foundation (through grants 2112471, 2107062, and 1750539) and by the Niarchos Foundation, through the Greek Diaspora Fellowship Program.

References

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

I. Aden-Ali, H. Ashtiani, and G. Kamath. On the sample complexity of privately learning unbounded high-dimensional Gaussians. In *Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 185–216. PMLR, 2021.

D. Alabi, A. McMillan, J. Sarathy, A. Smith, and S. Vadhan. Differentially private simple linear regression, 2020. URL https://arxiv.org/abs/2007.05157.

K. Amin, T. Dick, A. Kulesza, A. Munoz, and S. Vassilvitskii. Differentially private covariance estimation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/4158f6d19559955bae372bb00f6204e4-Paper.pdf.

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, N.J, 3rd ed edition, 2003. ISBN 9780471360919.

Anonymous. Review #2 of "Differentially Private Covariance Estimation", 2019. URL https://papers.nips.cc/paper/2019/file/4158f6d19559955bae372bb00f6204e4-Reviews.html.

H. Ashtiani and C. Liaw. Private and polynomial time algorithms for learning Gaussians and beyond, 2021. URL https://arxiv.org/abs/2111.11320.

H. Asi and J. C. Duchi. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. In *Advances in Neural Information Processing Systems*, volume 33, pages 14106–14117. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/a267f936e54d7c10a2bb70dbe6ad7a89-Paper.pdf.

A. F. Barrientos, J. P. Reiter, A. Machanavajjhala, and Y. Chen. Differentially private significance tests for regression coefficients. *Journal of Computational and Graphical Statistics*, 28(2):440–453, 2019.

R. Bassily, A. Smith, and A. 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, Oct. 2014.

R. Bassily, V. Feldman, K. Talwar, and A. Guha Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/3bd8fdb090f1f5eb66a00c84dbc5ad51-Paper.pdf.

G. Bernstein and D. R. Sheldon. Differentially private bayesian linear regression. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/f90f2aca5c640289d0a29417bcb63a37-Paper.pdf.

S. Biswas, Y. Dong, G. Kamath, and J. Ullman. Coinpress: Practical private mean and covariance estimation. *Advances in Neural and Information Processing Systems*, 2020.

D. R. Brillinger. A generalized linear model with "Gaussian" regressor variables. In *Selected Works of David Brillinger*, pages 589–606. Springer, 2012a. URL https://doi.org/10.1007/978-1-4614-1344-8_34.

D. R. Brillinger. The identification of a particular nonlinear time series system. In *Selected Works of David Brillinger*, pages 607–613. Springer, 2012b.

G. Brown, M. Gaboardi, A. Smith, J. Ullman, and L. Zakynthinou. Covariance-aware private mean estimation without private covariance estimation. *Advances in Neural Information Processing Systems*, 34, 2021.

M. Bun and T. Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography*, pages 635–658, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg. ISBN 978-3-662-53641-4.

M. Bun, K. Nissim, U. Stemmer, and S. Vadhan. Differentially private release and learning of threshold functions. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 634–649. IEEE, 2015.

M. Bun, K. Nissim, and U. Stemmer. Simultaneous private learning of multiple concepts. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, page 369–380, New York, NY, USA, 2016. Association for Computing Machinery. URL https://doi.org/10.1145/2840728.2840747.

T. T. Cai, Y. Wang, and L. Zhang. The cost of privacy in generalized linear models: Algorithms and minimax lower bounds, 2020. URL https://arxiv.org/abs/2011.03900.

K. Chaudhuri, C. Monteleoni, and A. 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.

C. Daskalakis, D. Rohatgi, and E. Zampetakis. Truncated linear regression in high dimensions. In *Advances in Neural Information Processing Systems*, volume 33, pages 10338–10347. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/751f6b6b02bf39c41025f3bcfd9948ad-Paper.pdf.

A. Dembo. Bounds on the extreme eigenvalues of positive-definite toeplitz matrices. *IEEE Transactions on Information Theory*, 34(2):352–355, 1988. doi: 10.1109/18.2651.

Z. Deng, A. Kammoun, and C. Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, page iaab002, Apr. 2021. URL https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaab002/6209694.

I. Diakonikolas, M. Hardt, and L. Schmidt. Differentially private learning of structured discrete distributions. *Advances in Neural Information Processing Systems*, 28, 2015.

I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, Jan. 2019a. doi: 10.1137/17M1126680. URL https://epubs.siam.org/doi/10.1137/17M1126680.

I. Diakonikolas, W. Kong, and A. Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '19, page 2745–2754, USA, 2019b. Society for Industrial and Applied Mathematics.

C. Dwork and A. Smith. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2), 2010.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.

C. Dwork, G. N. Rothblum, and S. Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60. IEEE, 2010.

C. Dwork, K. Talwar, A. Thakurta, and L. 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.

M. A. Erdogdu. Newton-Stein method: An optimization method for GLMs via Stein's lemma. *Journal of Machine Learning Research*, 17(215): 1–52, 2016. URL http://jmlr.org/papers/v17/16-062.html.

V. Feldman, T. Koren, and K. Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020. URL https://doi.org/10.1145/3357713.3384335.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, NY, 2nd ed edition, 2009. ISBN 9780387848570 9780387848587.

S. B. Hopkins, G. Kamath, and M. Majid. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism, 2021. URL https://arxiv.org/abs/2111.12981.

R. Iyengar, J. P. Near, D. Song, O. Thakkar, A. Thakurta, and L. Wang. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 299–316, 2019. doi: 10.1109/SP.2019.00001.

P. Jain and A. Thakurta. (Near) dimension independent risk bounds for differentially private learning. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page I–476–I–484. JMLR.org, 2014.

B. Kadioglu, P. Tian, J. Dy, D. Erdogmus, and S. Ioannidis. On the sample complexity of rank regression from pairwise comparisons, 2021. URL https://arxiv.org/abs/2105.01463.

G. Kamath, J. Li, V. Singhal, and J. Ullman. Privately learning high-dimensional distributions. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1853–1902, Phoenix, USA, 25–28 Jun 2019. PMLR. URL: http://proceedings.mlr.press/v99/kamath19a.html.

G. Kamath, X. Liu, and H. Zhang. Improved rates for differentially private stochastic convex optimization with heavy-tailed data, 2021a. URL https://arxiv.org/abs/2106.01336.

G. Kamath, A. Mouzakis, V. Singhal, T. Steinke, and J. Ullman. A private and computationally-efficient estimator for unbounded Gaussians, 2021b. URL https://arxiv.org/abs/2111.04609.

V. Karwa and S. Vadhan. Finite Sample Differentially Private Confidence Intervals. In A. R. Karlin, editor, *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, volume 94 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 44:1–44:9, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-060-6. doi: 10.4230/LIPIcs.ITCS.2018.44. URL http://drops.dagstuhl.de/opus/volltexte/2018/8344.

D. Kifer, A. Smith, and A. Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *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. JMLR Workshop and Conference Proceedings. URL http://proceedings.mlr.press/v23/kifer12.html.

G. R. Kini and C. Thrampoulidis. Analytic study of double descent in binary classification: The impact of loss. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2527–2532, Los Angeles, CA, USA, June 2020. IEEE. ISBN 9781728164328. doi: 10.1109/ISIT44484.2020.9174344. URL https://ieeexplore.ieee.org/document/9174344/.

P. K. Kothari, P. Manurangsi, and A. Velingker. Private robust estimation by stabilizing convex relaxations, 2021. URL https://arxiv.org/abs/2112.03548.

S. M. Kreidler, B. M. Ringham, K. E. Muller, and D. H. Glueck. Calculating power for the general linear multivariate model with one or more Gaussian covariates. *Communications in Statistics - Theory and Methods*, Feb. 2018. ISSN 0361-0926. URL https://www.tandfonline.com/doi/full/10.1080/03610926.2018.1433849.

T. Kulkarni, J. Jälkö, A. Koskela, S. Kaski, and

A. Honkela. Differentially private bayesian inference for generalized linear models. In *International Conference on Machine Learning*, pages 5838–5849. PMLR, 2021.

J. S. Liu. Siegel's formula via Stein's identities. *Statistics & Probability Letters*, 21(3):247–251, 1994.

X. Liu, W. Kong, S. Kakade, and S. Oh. Robust and differentially private mean estimation. *Advances in Neural Information Processing Systems*, 34, 2021a.

X. Liu, W. Kong, and S. Oh. Differential privacy and robust statistics in high dimensions, 2021b. URL https://arxiv.org/abs/2111.06578.

E. Ma and C. Zarowski. On lower bounds for the smallest eigenvalue of a Hermitian positive-definite matrix. *IEEE Transactions on Information Theory*, 41 (2):539–540, 1995. doi: 10.1109/18.370166.

F. McSherry and I. Mironov. Differentially private recommender systems: Building privacy into the netflix prize contenders. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 627–636, 2009.

P. Nakkiran. More data can hurt for linear regression: Sample-wise double descent, 2019. URL https://arxiv.org/abs/1912.07242.

O. Sheffet. Differentially private ordinary least squares. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3105–3114. PMLR, 06–11 Aug 2017. URL http://proceedings.mlr.press/v70/sheffet17a.html.

O. Sheffet. Old techniques in differentially private linear regression. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 789–827, Chicago, Illinois, 22–24 Mar 2019. PMLR. URL http://proceedings.mlr.press/v98/sheffet19a.html.

C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 1135–1151, 1981.

Y. Sun, S. Ioannidis, and A. Montanari. Learning mixtures of linear classifiers. In *International Conference on Machine Learning*, pages 721–729. PMLR, 2014.

T. Tao. *Topics in Random Matrix Theory*, volume 132. American Mathematical Society, 2012. ISBN 9780821874301.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018. ISBN 9781108415194.

D. Wang, M. Ye, and J. Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/f337d999d9ad116a7b4f3d409fcc6480-Paper.pdf.

Y. Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 93–103. AUAI Press, 2018. URL http://auai.org/uai2018/proceedings/papers/40.pdf.

J. Zhang, Z. Zhang, X. Xiao, Y. Yang, and M. Winslett. Functional mechanism: Regression analysis under differential privacy. *Proc. VLDB Endow.*, 5(11):1364–1375, July 2012. URL https://doi.org/10.14778/2350229.2350253.

J. Zhang, K. Zheng, W. Mou, and L. Wang. Efficient private ERM for smooth objectives. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3922–3928, 2017. URL https://doi.org/10.24963/ijcai.2017/548.

Supplementary Material: Differentially Private Regression with Unbounded Covariates

A Additional Preliminaries

In this section, we provide some tools required for our analysis.

Sub-gaussianity Tools. We begin with definitions of the sub-gaussian norm for univariate and multivariate random variables, and move on to some of their properties (Vershynin, 2018) that we later use.

Definition 6 (Sub-gaussian random variable)**.** *A random variable X is called a sub-gaussian random variable if there exists $K > 0$ such that, for all $\lambda : |\lambda| \leq 1/K$,*

$$\mathbb{E}\left[\exp\left(\lambda^2 X^2\right)\right] \leq \exp(\lambda^2 K^2)\,.$$

The smallest K for which the above property holds is called the sub-gaussian norm of X, and is denoted as $\|X\|_{\psi_2}$.

Definition 7 (Sub-gaussian random vector)**.** *A random vector $\boldsymbol{X} \in \mathbb{R}^d$ is called a sub-gaussian random vector if for all $\boldsymbol{u} \in \mathbb{R}^d$, the inner product $\langle \boldsymbol{X}, \boldsymbol{u} \rangle$ is a sub-gaussian random variable. The sub-gaussian norm of a sub-gaussian random vector is defined as follows:*

$$\|\boldsymbol{X}\|_{\psi_2} = \sup_{\boldsymbol{u} \in S^{d-1}} \|\langle \boldsymbol{X}, \boldsymbol{u} \rangle\|_{\psi_2}\,,$$

where $S^{d-1} = \{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_2 = 1\}$ is the d-dimensional unit sphere.

Lemma 8 (Properties of the sub-gaussian norm)**.** *Let $\boldsymbol{X}$ be a sub-gaussian random vector. Then, the following hold:*

- *For every constant $c > 0$, $c\boldsymbol{X}$ is a sub-gaussian random vector, with $\|c\boldsymbol{X}\|_{\psi_2} = c\,\|\boldsymbol{X}\|_{\psi_2}$.*
- *If $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu_X}$, then $\boldsymbol{X} - \boldsymbol{\mu_X}$ is a sub-gaussian random vector, with*

  $$\|\boldsymbol{X} - \boldsymbol{\mu_X}\|_{\psi_2} \leq C\,\|\boldsymbol{X}\|_{\psi_2}\,,$$

  for a universal constant $C > 0$.

Differential Privacy Tools. We continue with a slightly different definition of differential privacy that is roughly equivalent with the classical definition, according to Lemma 10 (Bun and Steinke, 2016).

Definition 9 (Zero-concentrated DP (zCDP))**.** *A randomized mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies ρ-zCDP if for every pair of neighboring datasets $X, X' \in \mathcal{X}^n$ that differ on at most one element, and for every $\alpha \geq 1$,*

$$D_\alpha\left(M(X)||M(X')\right) \leq \rho\alpha\,,$$

where $D_\alpha(P||Q) = \frac{1}{\alpha-1}\log\left(\mathbb{E}_{x\sim Q}\left[\left(\frac{P(x)}{Q(x)}\right)^\alpha\right]\right)$ is the α-Rényi divergence between the probability distributions P and Q.

Lemma 10 (An equivalence between zero-concentrated DP and "classical" DP)**.** *Let $M : \mathcal{X}^n \to \mathcal{Y}$ be a randomized mechanism. Then, the following results hold:*

- *If M is $\frac{\epsilon^2}{2}$-zCDP, then, for all $\delta > 0$, M is $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-DP.*
- *If M is $(\epsilon, 0)$-DP, then M is $\frac{\epsilon^2}{2}$-zCDP.*

Stein's Lemma and GLMs. Finally, we state a multivariate version of Stein's Lemma (Stein, 1981) due to Liu (1994).

Lemma 11 (Stein's Lemma (Liu, 1994)). *Let $\boldsymbol{Z} \in \mathbb{R}^p, \boldsymbol{W} \in \mathbb{R}^q$ be jointly Gaussian random vectors and let $f : \mathbb{R}^q \to \mathbb{R}$ be differentiable almost everywhere with $\mathbb{E}_{\boldsymbol{W}}\left[|\partial f(\boldsymbol{W})/\partial W_i|\right] < \infty$ for any $i \in [q]$. Then,* $\mathbf{Cov}[\boldsymbol{Z}, f(\boldsymbol{W})] = \mathbf{Cov}[\boldsymbol{Z}, \boldsymbol{W}]\,\mathbb{E}[\nabla f(\boldsymbol{W})]$.

The lemma has a direct application on Generalized Linear Models with Gaussian covariates (Kadioglu et al., 2021; Erdogdu, 2016; Brillinger, 2012a,b): for the GLM of Brillinger (2012a) with Gaussian covariates, whose model function satisfies the conditions of Lemma 11, one can show (Brillinger, 2012a,b) that the Ordinary Least Squares Estimator asymptotically converges to the true parameter vector $\boldsymbol{\beta}$ of the GLM with probability 1, up to a scaling factor k. This is analogous to the scaling factor k that appears in our analysis (see Theorem 4).

B Logistic Regression and SVMs as Models Satisfying Assumption 2

We will show here how the models of Logistic Regression and linearly-separable SVMs fit into our probabilistic model of Binary Regression (see Assumption 2). For the Logistic Regression model, applying the model function $f(x) = 1/(1+e^{-x})$ directly yields the conditional probabilistic model of $\mathbf{Pr}[Y = 1|\boldsymbol{X}] = \frac{1}{1+e^{-\boldsymbol{\beta}^T\boldsymbol{X}}}$ for the regression coefficients $\boldsymbol{\beta}$, which is precisely the desired Logistic Regression model.

In the second case, we consider linearly-separable SVMs, where the data $(\boldsymbol{X}_i, y_i)$ are completely separated by an underlying hyperplane that we are trying to uncover. That is, the model function would be $\mathrm{sgn}(\boldsymbol{\beta}^T\boldsymbol{X})$. However, such a function is neither smooth nor continuously differentiable near the origin, therefore we will apply the following trick: after receiving the perfectly linearly separable data, we will induce a minuscule amount of noise through a noisy model function f that is continuously differentiable and smooth everywhere (but crucially, near the origin). Intuitively, we smooth out the sign function. We can make infinitely good approximations of the sign function, and therefore passing the data through one of those before we apply our algorithm is sufficient to recover the true underlying $\boldsymbol{\beta}$ up to a scaling factor depending on the noise that we artificially introduced.

In order to overcome the model's non-smooth property, one way to smoothen this objective is to approximate the sign function of the model by a sigmoid

$$f(x) = \frac{2}{1+\exp(-\lambda x)} - 1\,,$$

which depends on a parameter λ. This sigmoid function would then be smooth and continuously differentiable near the origin, as desired. Our method could then be applied (in a similar way to the logistic regression). We note that the parameter would not affect the direction of to be estimated, but would affect the accuracy guarantee obtained.

C Overview of Kamath et al. (2019) and Extension to Sub-Gaussian Regime

We first review the modified covariance estimation result, and present a generalization of these results to the sub-gaussian case, whereupon we provide a subsequent discussion.

C.1 Equivalence of Privacy Guarantees of Kamath et al. (2019) to Classical DP

First of all, we note that the privacy guarantees hold for a variation of the differential privacy definition that is mentioned in Definition 9; specifically, $\frac{\epsilon^2}{2}$ "zero-concentrated DP." This variation is more lax than the classical (pure) ϵ-DP, but stronger than the (ϵ, δ)-DP that is commonly used in the privacy literature, as can be seen immediately from Lemma 10.

In our case, we prefer to keep the results of Theorem 3 and Theorem 4 in the classical (ϵ, δ)-DP definition, and therefore the respective algorithms are $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-DP. In essence, this guarantee is equivalent to (ϵ, δ)-DP, but this equivalence is worse as ϵ gets larger, i.e., in the case that little privacy is desired. Additionally, the fact that $\delta > 0$ allows us to bypass the requirement of an upper bound on $\|\boldsymbol{\mu}\|_2$ (role which was previously played by R), hence not requiring knowledge of R, as will be analyzed in Lemma 13.

C.2 Algorithm Overview

Here, we provide a high-level description of the algorithm LearnSubGaussian-hd (used in Algorithm 1) that learns the mean and covariance matrix of a high-dimensional Gaussian distribution, which differs slightly on its execution from the LearnGaussian-hd algorithm in a way that will be analyzed hereafter.

The building block of the covariance estimation algorithm is the NaivePCE algorithm (Algorithm 1 of the work), which intuitively would be the first try at inducing privacy in the covariance estimation procedure. More specifically, it truncates the input samples, adds a random Gaussian matrix to the empirical covariance that arises from these samples, and outputs the projection of the final matrix to the PSD cone. However, this naive "first try" algorithm exhibits a linear dependence of the accuracy to the largest eigenvalue κ of the covariance matrix Σ (intuitively, the largest variance across any direction), whereas we aim for a $\log \kappa$ dependence. Therefore, noticing that the accuracy dependence is optimal when the aforementioned largest eigenvalue is of constant order, we seek to transform the samples $\boldsymbol{X}_i$ to $A\boldsymbol{X}_i$ such that the largest eigenvalue of the covariance matrix of $A\boldsymbol{X}_i$ (which is $A\Sigma A$ for symmetric matrices A) satisfies the above condition.

The covariance estimation algorithm, thus, begins by efficiently finding such a matrix A (the "preconditioner") according to an algorithm (Algorithm 3 of the work) which does the following: it uses $O(\log \kappa)$ successive rounds of the NaivePCE algorithm such that every round "eliminates" the eigendirections of largest variance (through an eigenvector decomposition and keeping intact for the next rounds only the eigenvalues that are smaller than half the current upper bound) hence transforming each successive κ_j (for $1 \le j \le O(\log \kappa)$ the number of the current round) to $\kappa_{j+1} = 0.7\kappa_j$. After $O(\log \kappa)$ rounds, the final largest eigenvalue of $A\Sigma A$ will be of constant order, as desired. After this procedure which finds A, NaivePCE is run on the samples $A\boldsymbol{X}_i$ with a result of $\widetilde{\Sigma}$, and the covariance estimation algorithm finally outputs $\widehat{\Sigma} = A^{-1}\widetilde{\Sigma}A^{-1}$. For further consideration on the internal details of those algorithms, we refer the interested reader to Kamath et al. (2019). Note that for these steps, the knowledge of κ, the upper bound on the largest eigenvalue of the covariance matrix Σ of the initial samples $\boldsymbol{X}_i$ is necessary for the calibrated truncation and noise addition to occur correctly.

Had someone wanted to also estimate the mean, they would first get a matrix A as above through $2n$ samples $\frac{1}{\sqrt{2}}\left(\boldsymbol{X}_{2i} - \boldsymbol{X}_{2i-1}\right)$, $1 \le i \le n$ that are i.i.d. with the same covariance matrix Σ, and then draw n additional i.i.d. samples $\boldsymbol{X}_i$ (for a total of $3n$ samples) and apply the univariate mean estimation algorithm of Karwa and Vadhan (2018) to each coordinate of $A\boldsymbol{X}_i$ separately. Our algorithm's difference with Kamath et al. (2019) is that we call the algorithm of Karwa and Vadhan (2018) with $R = \infty$, which is allowable and efficient to do in our setting due to the privacy guarantee having $\delta > 0$ (see the guarantees on Appendix C.3). Once we are in a univariate sub-gaussian setting, and since we have a constant-order upper bound on the variance $\sigma^2 = O(1)$ of $(A\boldsymbol{X}_i)_j$, which denotes the j-th coordinate of the random vector $A\boldsymbol{X}_i$, the algorithm for univariate mean estimation works as follows: First, we find a differentially private estimation of an upper bound B on the data with high probability in the following way: we split the whole range that the mean might be located $(-\infty, \infty)$ to bins of width $\sigma = O(1)$, and taking advantage of the concentration of sub-gaussian random variables around their mean, we use a differentially private histogram algorithm (Bun et al., 2016) to locate the most frequent bin, which (along with its neighboring bins) should contain all data points with high probability. Second, we truncate the input data $\left(A\boldsymbol{X}_i\right)_j$ to a range calculated according to the above estimated bound, such that all input samples fall within that range with high probability, and then add Laplacian noise (calibrated according to the differentially-private calculated bound B) to the empirical mean of the input samples. We output as the result of the univariate mean estimation algorithm this noisy empirical mean of the (truncated) input samples.

Assuming the generic description of the algorithms above, we show how we extend the proofs to the sub-gaussian case below.

C.3 Differentially Private Sub-Gaussian Mean and Covariance Estimation

The modified algorithms that we presented in Appendix C.2 have the following guarantees, whereupon we will provide a proof sketch.

Lemma 12 (Private Covariance Estimation)**.** *For every $\epsilon, \delta, \gamma, \kappa, \alpha > 0$, there exists an $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-DP algorithm that, when given n i.i.d. samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from a sub-gaussian multivariate distribution with*

mean $\mathbb{E}[\boldsymbol{X}_i] = \mathbf{0}$ *and covariance matrix* $\mathbb{E}[\boldsymbol{X}_i \boldsymbol{X}_i^T] = \Sigma$ *with* $\mathbb{I}_d \preceq \Sigma \preceq \kappa \mathbb{I}_d$ *and*

$$n = O\left(\frac{d + \log(1/\gamma)}{\alpha^2} + \frac{d^{3/2}\mathrm{polylog}\left(\frac{d}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon} + \frac{d^{3/2}\sqrt{\log\kappa}\mathrm{polylog}\left(\frac{d\log\kappa}{\gamma\epsilon}\right)}{\epsilon}\right),$$

outputs $\widehat{\Sigma}$ *such that* $\left\|\Sigma^{-1/2}\left(\widehat{\Sigma} - \Sigma\right)\Sigma^{-1/2}\right\|_2 \leq O(\alpha)$ *with probability* $1 - O(\gamma)$.

Lemma 13 (Private Mean Estimation). *For every* $\epsilon, \delta, \gamma, \kappa, \alpha > 0$, *there exists an* $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$*-DP algorithm that, when given* n *i.i.d. samples* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ *from a sub-gaussian multivariate distribution with mean* $\mathbb{E}[\boldsymbol{X}_i] = \boldsymbol{\mu}$ *and covariance matrix* $\mathbb{E}[\boldsymbol{X}_i \boldsymbol{X}_i^T] = \Sigma$ *with* $\mathbb{I}_d \preceq \Sigma \preceq \kappa \mathbb{I}_d$ *and*

$$n = O\left(\frac{d\log(\frac{d}{\gamma})}{\alpha^2} + \frac{d\mathrm{polylog}(\frac{d\log(1/\delta)}{\alpha\gamma\epsilon})}{\alpha\epsilon} + \frac{\sqrt{d}\log(\frac{d}{\gamma\delta})}{\epsilon} + \frac{d^{3/2}\sqrt{\log\kappa}\mathrm{polylog}\left(\frac{d\log\kappa}{\gamma\epsilon}\right)}{\epsilon}\right),$$

outputs a (symmetric) matrix A *and a vector* $\widehat{\boldsymbol{\mu}}$ *such that* $\mathbb{I}_d \preceq A\Sigma A \preceq 1000\mathbb{I}_d$ *and* $\|A(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|_2 \leq \alpha$ *with probability* $1 - O(\gamma)$.

First of all, the respective algorithms adumbrated in Appendix C.2 hold for the case of sub-gaussian input random vectors too, because the concentration bounds that are utilized readily generalize to the sub-gaussian case. We provide here the variants of the concentration bounds that are needed for these algorithms, and we then show how the second modification with respect to the consideration of $R = \infty$ (see Appendix C.2 for this modification) alters the guarantees provided.

By Diakonikolas et al. (2019a), we have the following generalizations of concentration bounds in the sub-gaussian regime:

Lemma 14. *Let* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ *be* n *i.i.d. samples from a sub-gaussian multivariate distribution with mean* $\mathbb{E}[\boldsymbol{X}_i] = \mathbf{0}$ *and covariance matrix* $\mathbb{E}[\boldsymbol{X}_i \boldsymbol{X}_i^T] = \Sigma$. *Then, with probability* $1 - O(\gamma)$, *it holds that*

$$\left\|\Sigma^{-1/2}\boldsymbol{X}_i\right\|_2^2 \leq d\log(n/\gamma),\ \forall i \in [n].$$

Lemma 15 (Sub-gaussian covariance matrix estimation). *Let* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ *be* n *i.i.d. samples from a sub-gaussian multivariate distribution with mean* $\mathbb{E}[\boldsymbol{X}_i] = \mathbf{0}$ *and covariance matrix* $\mathbb{E}[\boldsymbol{X}_i \boldsymbol{X}_i^T] = \Sigma$. *Define* $\boldsymbol{Z}_i = \Sigma^{-1/2}\boldsymbol{X}_i$ *with covariance matrix* $\mathbb{E}[\boldsymbol{Z}_i \boldsymbol{Z}_i^T] = \mathbb{I}_d$. *Then, with probability* $1 - O(\gamma)$, *all the following hold:*

$$\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T - \mathbb{I}_d\right\|_2 \leq O\left(\sqrt{\frac{d + \log(1/\gamma)}{n}}\right)$$

$$\left(1 - O\left(\sqrt{\frac{d + \log(1/\gamma)}{n}}\right)\right)\cdot\mathbb{I}_d \preceq \frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T \preceq \left(1 + O\left(\sqrt{\frac{d + \log(1/\gamma)}{n}}\right)\right)\cdot\mathbb{I}_d$$

$$\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T - \mathbb{I}_d\right\|_F \leq O\left(\sqrt{\frac{d^2 + \log(1/\gamma)}{n}}\right)$$

where $\|A\|_F$ *is the Frobenius norm of a matrix* A, *defined as the square root of the sum of the squares of each of its entries.*

Note that the necessary bounds for the univariate case are obtained simply by setting $d = 1$ in the above lemmata. These are required for adapting the proofs of Karwa and Vadhan (2018) to the sub-gaussian regime.

Additionally, we remark that we do not need to modify the concentration bound arising from the Hanson-Wright inequality for bounding the norm of the noise matrix that is added according to NaivePCE, since, even in the case that the inputs are sub-gaussian, the added noise is purely Gaussian.

Proof Sketch. We now provide a proof sketch for the modification of the proof of Kamath et al. (2019). In this sketch, we will use the standard notation to move forward with our variation of the lemmata.

First, we provide an alternative lemma that removes the dependency on a prior bound for $\|\boldsymbol{\mu}\|_2$ when the DP guarantee that we desire to achieve is $\delta > 0$.

Lemma 16. *For every $\epsilon, \delta, \gamma, \kappa, \alpha > 0$, there exists an (ϵ, δ)-DP algorithm that, when given n i.i.d. samples $X_1, \ldots, X_n$ from a sub-gaussian univariate distribution with mean $\mathbb{E}[X_i] = \mu$ and variance $\mathbb{E}[(X_i - \mu)^2] = \sigma^2$ with $1 \leq \sigma^2 \leq \kappa$ and*

$$n = O\left(\frac{\log(1/\gamma)}{\alpha^2} + \frac{\text{polylog}\left(\frac{\log(1/\delta)}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon} + \frac{\log(1/\delta) + \log(1/\gamma)}{\epsilon}\right),$$

outputs $\widehat{\mu}$ such that $|\widehat{\mu} - \mu| \leq \alpha\kappa$ with probability $1 - \gamma$.

Generalizing this algorithm to the multivariate case, by following the NaivePME algorithm referenced in Appendix C.2, we obtain the following lemma, in the proof sketch of which we show only the modifications required.

Lemma 17. *For every $\epsilon, \delta, \gamma, \kappa, \alpha > 0$, there exists an $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-DP algorithm that, when given n i.i.d. samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from a sub-gaussian multivariate distribution with mean $\mathbb{E}[\boldsymbol{X}_i] = \boldsymbol{\mu}$ and covariance matrix $\mathbb{E}[\boldsymbol{X}_i \boldsymbol{X}_i^T] = \Sigma$ with $\mathbb{I}_d \preceq \Sigma \preceq \kappa\mathbb{I}_d$ and*

$$n = O\left(\frac{\kappa^2 d \log(d/\gamma)}{\alpha^2} + \frac{\kappa d \text{polylog}\left(\frac{\kappa d \log(1/\delta)}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon} + \frac{\sqrt{d}\left(\log(1/\delta) + \log(d/\gamma)\right)}{\epsilon}\right),$$

outputs $\widehat{\boldsymbol{\mu}}$ such that $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \leq \alpha$ with probability $1 - \gamma$.

Proof Sketch. Following the procedure of Lemma 16 for each dimension of the multivariate vectors $\boldsymbol{X}_i$ with the appropriate parameter settings as described in NaivePME, we obtain the following final result:

$$\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \leq \sqrt{d} \max_{1 \leq i \leq d} |\widehat{\mu}_i - \mu_i| \leq \sqrt{d}\left(\frac{\alpha}{\sqrt{d}}\right) = \alpha\,,$$

as desired. □

The final step is to use the private "preconditioner" A in order to reduce the condition number of Σ (which is at most κ) to at most a constant, and obtain the final bound of the lemma that we seek. Again, we show the modification of the bound, hinging on Lemma 17 which is used to output the final estimate $\widehat{\boldsymbol{\mu}} = A^{-1}\widetilde{\boldsymbol{\mu}}$ from the estimate $\widetilde{\boldsymbol{\mu}}$ of the mean of the variables $A\boldsymbol{X}_i$, as follows by the lemma with $\kappa = O(1)$:

$$\|A(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|_2 = \|\widetilde{\boldsymbol{\mu}} - A\boldsymbol{\mu}\|_2 \leq \alpha\,.$$

□

D Proof of Theorem 3

We divide the proof of Theorem 3 in a series of claims. For convenience, we restate the (stronger) version of the Theorem that we will prove here:

Theorem 18 (Privacy and Accuracy of $\widehat{\boldsymbol{\beta}}$ in Private Least Squares Fitting)**.** *Under Assumption 1 with parameters (κ, c, ρ, R), for all privacy parameters $\epsilon, \delta > 0$, accuracy parameters $\alpha, \eta > 0$ and confidence $\gamma \in (0, 1)$,* PrivLearnLSE *(defined in Algorithm 1) is $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-differentially private. Moreover, if the number of labeled examples is at least:*

$$n = O\left(\frac{d\log(\frac{d}{\gamma})}{\eta^2} + \frac{d\text{polylog}(\frac{d\log(1/\delta)}{\eta\gamma\epsilon})}{\eta\epsilon} + \frac{d^{3/2}\sqrt{\log(\kappa\rho c)}\text{polylog}\left(\frac{d\log(\kappa\rho c)}{\gamma\epsilon\delta}\right)}{\epsilon}\right)$$
$$+ O\left((1+R)\left(\frac{d\log(\frac{d}{\gamma})}{\alpha^2} + \frac{d^{3/2}\text{polylog}\left(\frac{d\log(1/\delta)}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon} + \frac{d^{3/2}\sqrt{\log\kappa}\text{polylog}\left(\frac{d\log\kappa}{\gamma\epsilon}\right)}{\epsilon}\right)\right), \tag{D.1}$$

then with probability at least $1 - O(\gamma)$ *an estimate* $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$ *is successfully output and along with the LSE* $\boldsymbol{\beta}^*$ *satisfies:*

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\right\|_2^2 \leq \|\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star\|_2^2 \leq O\left(\alpha^2\right) \cdot \|\boldsymbol{w}^\star\|_2^2 + O\left(\eta^2\right) \cdot c^2\,, \tag{D.2}$$

where $\widehat{\boldsymbol{w}} = \Sigma^{1/2}\widehat{\boldsymbol{\beta}}$ *and* $\boldsymbol{w}^\star = \Sigma^{1/2}\boldsymbol{\beta}^\star$. *Finally,* PrivLearnLSE *runs in* $\mathrm{poly}(n)$ *time.*

The outline of the proof is as follows. First, we prove the privacy guarantee (see Appendix D.1); the latter follows in a similar fashion as the privacy proofs of the algorithms in Appendix C.3, exploiting the extension to $\delta > 0$ that we described, and generalizing the obtained privacy for altering both y_i and $\boldsymbol{X}_i$ (see Section 6). In the case of accuracy (see Appendix D.2), we first establish the following inequality:

$$\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star\|_2^2 \leq 2\left\|\Sigma^{1/2}\left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T\right)^{-1}\Sigma^{1/2}\right\|_2^2 \left(\left\|\Sigma^{-1/2}Q_1\Sigma^{-1/2}\right\|_2^2 \cdot \|\boldsymbol{w}^\star\|_2^2 + \left\|\Sigma^{-1/2}\boldsymbol{Q}_2\right\|_2^2\right)\,.$$

Via a series of claims (see Claim 19, Claim 20, and Claim 21) we bound each of the constituent terms of the right-hand-side of this inequality, finally yielding Eq. (D.2) with as many samples as in Eq. (D.1).

D.1 Proof of Privacy Guarantee

We show first that Algorithm 1, that computes $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$ as in Theorem 3, is $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-DP. Our dataset consists of n pairs $(\boldsymbol{X}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, therefore we will look into what happens if one of those pairs is altered: specifically, consider that the pair $(\boldsymbol{X}_i, y_i)$ becomes $(\boldsymbol{X}_i', y_i')$ for some (specific) i. The algorithm for $\widehat{\boldsymbol{\beta}}$ uses three sub-algorithms, which we claim will be $\left(\frac{1}{3}\left(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}\right), \frac{\delta}{3}\right)$-DP each (as described in the algorithm above, it suffices to consider $O(\epsilon)$ and $O(\delta)$ as parameters of each one), thereby giving us the final result of the claim by the advanced composition properties of differentially private mechanisms (Fact 1).

For the covariance estimation algorithm (used both in the covariance estimation of the feature vectors $\boldsymbol{X}_i$ and in the mean estimation of the product $y_i\boldsymbol{X}_i$), it now suffices to show that the interface of the algorithms used in Appendix C.2 with the underlying data, i.e., the NaivePCE algorithm, is differentially private. Indeed, this result arises by computing the sensitivity of the (truncated) empirical covariance in the following way:

$$\left\|\frac{1}{n}\left(y_i^2\boldsymbol{X}_i\boldsymbol{X}_i^T - y_i'^2\boldsymbol{X}_i'\boldsymbol{X}_i'^T\right)\right\|_F \leq \frac{1}{n}\left(\|y_i\boldsymbol{X}_i\|_2^2 + \|y_i'\boldsymbol{X}_i'\|_2^2\right) \leq O\left(\frac{d\kappa c^2\log(n/\gamma)}{n}\right)\,,$$

since the truncation happens in accordance with Lemma 14, thereby allowing us to add the Gaussian noise of the magnitude prescribed in NaivePCE. Hence, the rest of the algorithms which hinge on NaivePCE are differentially private, due to the differential privacy composition theorems (see Fact 1).

In a similar fashion, the algorithm for mean estimation in Karwa and Vadhan (2018) depends upon the stability-based histogram learner of Bun et al. (2016) which is in turn based on the idea of introducing Laplacian noise to the empirical histogram of a dataset. Denoting X_{ij} as the j-th coordinate of the feature vector $\boldsymbol{X}_i$, we notice that the sensitivity of the empirical counting function is $2/n$ regardless of the input variations (at most 2 bins could have their counts altered in the worst case, if some y_iX_{ij} changed its location from the bin it was to another, whereas all the other products had the same locations in bins). This sensitivity is, thus, independent of whether both y_i and X_{ij} were changed in our model, thereby affording the desired level of privacy to the whole algorithm. □

D.2 Proof of Accuracy Guarantee

For simplicity in notation, in what follows we notate $\boldsymbol{\mu} = \boldsymbol{\mu}_{\boldsymbol{X}}$ and $\boldsymbol{\mu}' = \boldsymbol{\mu}_{\boldsymbol{X},y}$ (likewise, $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}$ and $\widehat{\boldsymbol{\mu}}' = \widehat{\boldsymbol{\mu}}_{\boldsymbol{X},y}$).

We begin by adding and subtracting the quantities of each factor of $\boldsymbol{\beta}^\star \in \mathbb{R}^d$, as follows:

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star = \left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T\right)^{-1}(-Q_1\boldsymbol{\beta}^\star + \boldsymbol{Q}_2) \Leftrightarrow \left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\right) = -Q_1\boldsymbol{\beta}^\star + \boldsymbol{Q}_2\,,$$

where

$$Q_1 = \widehat{\Sigma} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T - \frac{1}{n}X^TX, \quad \text{and} \quad \boldsymbol{Q}_2 = \widehat{\boldsymbol{\mu}}' - \frac{1}{n}X^T\boldsymbol{y}\,.$$

Then, substituting the quantities $\boldsymbol{w} = \Sigma^{1/2}\boldsymbol{\beta}$ and moving terms from the left side to the right of the equation, it holds that

$$\begin{aligned}
&\left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\right) = -Q_1\boldsymbol{\beta}^\star + \boldsymbol{Q}_2 \\
\Leftrightarrow\ &\left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T\right)\Sigma^{-1/2}\left(\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star\right) = -Q_1\Sigma^{-1/2}\boldsymbol{w}^\star + \boldsymbol{Q}_2 \\
\Leftrightarrow\ &\Sigma^{-1/2}\left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T\right)\Sigma^{-1/2}\left(\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star\right) = -\Sigma^{-1/2}Q_1\Sigma^{-1/2}\boldsymbol{w}^\star + \Sigma^{-1/2}\boldsymbol{Q}_2 \\
\Leftrightarrow\ &\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star = \Sigma^{1/2}\left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T\right)^{-1}\Sigma^{1/2}\left(-\Sigma^{-1/2}Q_1\Sigma^{-1/2}\boldsymbol{w}^\star + \Sigma^{-1/2}\boldsymbol{Q}_2\right).
\end{aligned}$$

Using Cauchy-Schwartz and the sub-multiplicative property of the spectral norm, we establish the following inequality:

$$\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star\|_2^2 \le 2\left\|\Sigma^{1/2}\left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T\right)^{-1}\Sigma^{1/2}\right\|_2^2\left(\left\|\Sigma^{-1/2}Q_1\Sigma^{-1/2}\right\|_2^2 \cdot \|\boldsymbol{w}^\star\|_2^2 + \left\|\Sigma^{-1/2}\boldsymbol{Q}_2\right\|_2^2\right).$$

We state three claims that bound the constituent terms of the right-hand-side of this inequality:

Claim 19. *When*

$$n = \Omega\left(d + \log(1/\gamma) + \frac{d^{3/2}\sqrt{\log\kappa}\,\mathrm{polylog}\left(\frac{d\log\kappa}{\gamma\epsilon}\right)}{\epsilon}\right),$$

the following inequality holds with probability $1 - O(\gamma)$*:*

$$\left\|\Sigma^{1/2}\left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T\right)^{-1}\Sigma^{1/2}\right\|_2^2 \le O(1).$$

Claim 20. *For every* $\alpha > 0$*, when*

$$n = \Omega\left((1+R)\left(\frac{d\log(\frac{d}{\gamma})}{\alpha^2} + \frac{d^{3/2}\mathrm{polylog}\left(\frac{d\log(1/\delta)}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon} + \frac{d^{3/2}\sqrt{\log\kappa}\,\mathrm{polylog}\left(\frac{d\log\kappa}{\gamma\epsilon}\right)}{\epsilon}\right)\right),$$

the following inequality holds with probability $1 - O(\gamma)$*:*

$$\left\|\Sigma^{-1/2}Q_1\Sigma^{-1/2}\right\|_2^2 \le O(\alpha^2).$$

Claim 21. *For every* $\eta > 0$*, when*

$$n = \Omega\left(\frac{d\log(\frac{d}{\gamma})}{\eta^2} + \frac{d\mathrm{polylog}(\frac{d\log(1/\delta)}{\eta\gamma\epsilon})}{\eta\epsilon} + \frac{\sqrt{d}\log(\frac{d}{\gamma\delta})}{\epsilon} + \frac{d^{3/2}\sqrt{\log(\kappa\rho c)}\,\mathrm{polylog}\left(\frac{d\log(\kappa\rho c)}{\gamma\epsilon}\right)}{\epsilon}\right),$$

the following inequality holds with probability $1 - O(\gamma)$*:*

$$\left\|\Sigma^{-1/2}\boldsymbol{Q}_2\right\|_2^2 \le O(\eta^2)\cdot c^2.$$

We prove each of these claims individually below (see Appendix D.2.1–Appendix D.2.3). When combined with a union bound of the respective probabilistic events, these claims give the desired Theorem 3. In particular, we directly obtain Theorem 3, since

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\right\|_2^2 = \left\|\Sigma^{-1/2}\Sigma^{1/2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\right)\right\|_2^2 \le \left\|\Sigma^{-1/2}\right\|_2^2 \cdot \|\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star\|_2^2 \le \|\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star\|_2^2,$$

by the sub-multiplicative property of the norm and because $\mathbb{I}_d \preceq \Sigma$. □

D.2.1 Proof of Claim 19

Following the covariance estimation procedure, we recall the "private preconditioner" matrix A that is used to reduce the effect of the condition number of Σ from at most $O(\kappa)$ to at most a constant order factor. More specifically, the following lemma holds:

Lemma 22 (Theorem 3.11 of Kamath et al. (2019)). *For every $\epsilon, \delta, \gamma, \alpha, \kappa > 0$, there exists an algorithm that, when given n i.i.d. samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from a sub-gaussian multivariate distribution with mean $\mathbb{E}[\boldsymbol{X}_i] = \boldsymbol{\mu}$ and covariance matrix $\mathbb{E}[\boldsymbol{X}_i \boldsymbol{X}_i^T] = \Sigma$ with $\mathbb{I}_d \preceq \Sigma \preceq \kappa \mathbb{I}_d$ and*

$$n = O\left(\frac{d^{3/2}\sqrt{\log \kappa}\text{polylog}\left(\frac{d \log \kappa}{\gamma \epsilon}\right)}{\epsilon}\right),$$

outputs a (symmetric) matrix A (the "private preconditioner") such that $\mathbb{I}_d \preceq A\Sigma A \preceq 1000\mathbb{I}_d$ with probability $1 - O(\gamma)$.

Afterwards, the naive private estimation by addition of Gaussian noise through a random Gaussian matrix perturbation to the sample covariance matrix estimate is run, taking as input the "normalized" samples $A\boldsymbol{X}_i$, and by denoting $\widetilde{\Sigma}$ this estimate (which is explained in the proof of Fact 3) and also $\widetilde{\boldsymbol{\nu}} = A\widehat{\boldsymbol{\mu}}$, we have that $\widehat{\Sigma} = A^{-1}\widetilde{\Sigma}A^{-1}$ and therefore with probability $1 - O(\gamma)$,

$$\left\|\Sigma^{1/2}\left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T\right)^{-1}\Sigma^{1/2}\right\|_2^2 \leq \left\|\Sigma^{1/2}A\right\|_2^2 \cdot \left\|\left(\widetilde{\Sigma} + \widetilde{\boldsymbol{\nu}}\widetilde{\boldsymbol{\nu}}^T\right)^{-1}\right\|_2^2 \cdot \left\|A\Sigma^{1/2}\right\|_2^2,$$

which gives the desired result, due to the following two facts.

Fact 2. *With probability $1 - O(\gamma)$, $\left\|\Sigma^{1/2}A\right\|_2^2 = \left\|A\Sigma^{1/2}\right\|_2^2 \leq O(1)$.*

Proof. Since A is a symmetric square matrix, $\left\|\Sigma^{1/2}A\right\|_2^2 = \left\|A\Sigma^{1/2}\right\|_2^2$. By using the definition of the spectral norm, the fact's statement is immediately obtained:

$$\left\|\Sigma^{1/2}A\right\|_2^2 = \sigma_{\max}^2\left(\Sigma^{1/2}A\right) = \lambda_{\max}\left(\left(\Sigma^{1/2}A\right)^T\Sigma^{1/2}A\right) = \lambda_{\max}\left(A\Sigma A\right) \leq 1000,$$

since by Lemma 22, $A\Sigma A \preceq 1000\mathbb{I}_d$ with probability $1 - O(\gamma)$. □

Fact 3. *With probability $1 - O(\gamma)$, when*

$$n = \Omega\left(d + \log(1/\gamma) + \frac{\sqrt{d}\text{polylog}\left(\frac{d}{\epsilon\gamma}\right)}{\epsilon}\right),$$

it holds that

$$\left\|\left(\widetilde{\Sigma} + \widetilde{\boldsymbol{\nu}}\widetilde{\boldsymbol{\nu}}^T\right)^{-1}\right\|_2^2 \leq 1 + O\left(\sqrt{\frac{d + \log(1/\gamma)}{n}} + \frac{\sqrt{d}\log(1/\gamma)\log(n/\gamma)}{n\epsilon}\right).$$

Proof. In order to prove this fact, we need to delve further into the procedure by which $\widetilde{\Sigma}$ is generated (see Appendix C.2), i.e., NaivePCE. In short, we will use that $\widetilde{\Sigma}$ is the projection into the PSD cone of the empirical covariance matrix of the inputs (which are the vectors $A\boldsymbol{X}_1, \ldots, A\boldsymbol{X}_n$ where A is the above "preconditioner" matrix) plus a symmetric random matrix N of small Gaussian perturbations (that serves to enforce the privacy guarantee). Hence, the following holds:

$$\widetilde{\Sigma} = \text{proj}_{\text{PSD}}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{Z}_i^T + N\right),$$

where $\boldsymbol{Z}_i = A\boldsymbol{X}_i$, and the matrix N is a symmetric random matrix with dimension $d \times d$ whose entries $N_{ij}, j \geq i$ are i.i.d. Gaussian random variables with zero mean and standard deviation $\sigma = \frac{d\log(n/\gamma)}{n\epsilon}$.

By Weyl's inequality for matrices, it is true for two real, symmetric matrices A, B and their sum $A+B$ that $\lambda_{\min}(A+B) \geq \lambda_{\min}(A)+\lambda_{\min}(B)$. Because $\left\| \left(\widetilde{\Sigma} + \widetilde{\boldsymbol{\nu}}\widetilde{\boldsymbol{\nu}}^T\right)^{-1} \right\|_2^2 = \frac{1}{\lambda^2_{\min}\left(\widetilde{\Sigma}+\widetilde{\boldsymbol{\nu}}\widetilde{\boldsymbol{\nu}}^T\right)}$, we will prove a lower bound about $\lambda_{\min}\left(\widetilde{\Sigma} + \widetilde{\boldsymbol{\nu}}\widetilde{\boldsymbol{\nu}}^T\right) \geq \lambda_{\min}\left(\widetilde{\Sigma}\right)$ (since $\widetilde{\boldsymbol{\nu}}\widetilde{\boldsymbol{\nu}}^T$ is a PSD matrix) in the following way: we will show a bound about the minimum eigenvalue of the inner sum $\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T + N$, and argue that it is positive with high probability $1-O(\gamma)$. This means that the projection of this matrix into the PSD cone is the same as the matrix itself with high probability, therefore the bound on the minimum eigenvalue will hold verbatim.

We begin by tailoring a lemma from random matrix theory referenced in Tao (2012) to the random matrix N:

Lemma 23 (Concentration of symmetric random matrices with Gaussian entries)**.** *Suppose that the entries N_{ij} for $j \geq i$ of a symmetric matrix N with dimensions $d \times d$ are i.i.d. Gaussian random variables with zero mean and variance σ^2. Then, there exist universal constants $C, c > 0$ such that the largest singular value of N satisfies for all $A \geq C$:*

$$\mathbf{Pr}\left[s_{max}(N) > A\sigma\sqrt{d}\right] \leq C\exp(-cAd)\,.$$

The above lemma means that with probability at least $1-\gamma$, we have that the largest singular value of N is at most

$$s_{\max}(N) \leq O\left(\frac{\sigma\log(1/\gamma)}{\sqrt{d}}\right).$$

Due to the matrix N being square symmetric with dimensions $d \times d$, it holds that its singular values are the absolute values of its eigenvalues, therefore $|\lambda_{\min}(N)|$ is one of the singular values of N (note that it could even be the largest), hence $|\lambda_{\min}(N)| \leq s_{\max}(N)$ and thus

$$\lambda_{\min}(N) \geq -s_{\max}(N) \geq -O\left(\frac{\sigma\log(1/\gamma)}{\sqrt{d}}\right) = -O\left(\frac{\sqrt{d}\log(1/\gamma)\log(n/\gamma)}{n\epsilon}\right), \tag{D.3}$$

since we remind the reader that $\sigma = \frac{d\log(n/\gamma)}{n\epsilon}$.

Lower bounds on the minimum eigenvalue of sample covariance matrices of the form are well-known in the literature, and we use here a version that appears in Diakonikolas et al. (2019a). Note that the vectors $\boldsymbol{Z}_i = A\boldsymbol{X}_i$, whose covariance matrix we are interested in, have a "normalized" distribution with covariance matrix $A\Sigma A^T = A\Sigma A$ (by the symmetry of A) that has at most a constant eigenvalue, since $A\Sigma A \preceq 1000\mathbb{I}_d$ by construction of the "preconditioner" A with probability $1-O(\gamma)$. Therefore, by classical covariance matrix estimation inequalities for eigenvalues of sample covariance matrices from Diakonikolas et al. (2019a), it holds that with probability $1-O(\gamma)$,

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T\right) \geq 1 - O\left(\sqrt{\frac{d+\log(1/\gamma)}{n}}\right). \tag{D.4}$$

To conclude, we combine the lower bounds in eigenvalues of Eq. (D.3) and Eq. (D.4). By Weyl's inequality, with probability $1-O(\gamma)$, we have that:

$$\begin{aligned}\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T + N\right) &\geq \lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T\right) + \lambda_{\min}\left(N\right) \\ &\geq 1 - O\left(\sqrt{\frac{d+\log(1/\gamma)}{n}} + \frac{\sqrt{d}\log(1/\gamma)\log(n/\gamma)}{n\epsilon}\right).\end{aligned}$$

Choosing n such that the above lower bound is positive[4], i.e., if

$$n = \Omega\left(d + \log(1/\gamma) + \frac{\sqrt{d}\text{polylog}\left(\frac{d}{\epsilon\gamma}\right)}{\epsilon}\right),$$

then the projection of the matrix sum into the PSD cone is equal to the matrix sum itself, therefore directly arriving at the final result by noting the additional fact that:

$$\left(\frac{1}{1-O(x)}\right)^2 \leq 1 + O(x),$$

obtainable by a Taylor expansion since we chose n to be at least such that the denominator of the fraction is positive. □

By combining Fact 2 and Fact 3, we arrive at the final result of Claim 19.

D.2.2 Proof of Claim 20

Initially, breaking Q_1 as

$$\begin{aligned}
Q_1 &= \widehat{\Sigma} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T - \frac{1}{n}X^TX \\
&= \left(\widehat{\Sigma} - \Sigma\right) - \left(\frac{1}{n}\sum_{i=1}^n (\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^T - \Sigma\right) \\
&\quad + \left(\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T - \boldsymbol{\mu}\boldsymbol{\mu}^T - \boldsymbol{\mu}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i - \boldsymbol{\mu}\right)^T - \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i - \boldsymbol{\mu}\right)\boldsymbol{\mu}^T\right) \\
&= \left(\widehat{\Sigma} - \Sigma\right) - \left(\frac{1}{n}\sum_{i=1}^n (\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^T - \Sigma\right) - \boldsymbol{\mu}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i - \boldsymbol{\mu}\right)^T - \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i - \boldsymbol{\mu}\right)\boldsymbol{\mu}^T \\
&\quad + \left((\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T + (\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})\boldsymbol{\mu}^T + \boldsymbol{\mu}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T\right),
\end{aligned}$$

it follows that

$$\begin{aligned}
\left\|\Sigma^{-1/2}Q_1\Sigma^{-1/2}\right\|_2^2 \leq\ & 2\left\|\Sigma^{-1/2}\left(\widehat{\Sigma} - \Sigma\right)\Sigma^{-1/2}\right\|_2^2 \\
& + 2\left\|\Sigma^{-1/2}\left(\frac{1}{n}\sum_{i=1}^n (\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^T - \Sigma\right)\Sigma^{-1/2}\right\|_2^2 \\
& + 2\left\|\boldsymbol{\mu}\right\|_2^2 \cdot \left(2\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{V}_i\right\|_2^2 + O\left(\left\|\Sigma^{-1/2}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})\right\|_2^2\right)\right),
\end{aligned}$$

where we have used the variance-normalized vectors $\boldsymbol{V}_i = \Sigma^{-1/2}(\boldsymbol{X}_i - \boldsymbol{\mu})$ which have covariance matrix $\mathbb{I}_d$.

We state and prove the two below facts which, along with Lemma 12, which we remind to the reader that it is applied to the $2n$ sample differences $\frac{1}{\sqrt{2}}(\boldsymbol{X}_{2i} - \boldsymbol{X}_{2i-1})$, so that they have zero mean, and with a similar procedure to Fact 2 and Lemma 13, leads to the desired result immediately.

Fact 4. *For every $\alpha > 0$, with probability $1 - O(\gamma)$, when*

$$n = \Omega\left(\frac{d + \log(1/\gamma)}{\alpha^2}\right),$$

[4] We remark that this eigenvalue lower bound also means that the eigenvalue of $\widetilde{\Sigma} + \widetilde{\nu}\widetilde{\nu}^T = A\left(\widehat{\Sigma} + \widehat{\mu}\widehat{\mu}^T\right)A$ is bounded away from 0 by means of the chosen sample size n. This directly implies, since A is always invertible by Kamath et al. (2019), that $\mathrm{M} = \widehat{\Sigma} + \widehat{\mu}\widehat{\mu}^T$ in Algorithm 1 is also invertible with the same high probability, $1 - O(\gamma)$.

it holds that

$$\left\| \Sigma^{-1/2} \left(\frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{X}_i - \boldsymbol{\mu} \right) \left(\boldsymbol{X}_i - \boldsymbol{\mu} \right)^T - \Sigma \right) \Sigma^{-1/2} \right\|_2^2 \leq O(\alpha^2) \, .$$

Proof. First of all, we restate the left hand side of the inequality in terms of the variance-normalized vectors $\boldsymbol{V}_i$, as follows:

$$\left\| \Sigma^{-1/2} \left(\frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{X}_i - \boldsymbol{\mu} \right) \left(\boldsymbol{X}_i - \boldsymbol{\mu} \right)^T - \Sigma \right) \Sigma^{-1/2} \right\|_2^2 = \left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{V}_i \boldsymbol{V}_i^T - \mathbb{I}_d \right\|_2^2 ,$$

which we can bound with high probability by the classical empirical covariance estimation concentration bounds in Diakonikolas et al. (2019a). More specifically, with probability $1 - O(\gamma)$, we have that:

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{V}_i \boldsymbol{V}_i^T - \mathbb{I}_d \right\|_2 = \lambda_{\max} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{V}_i \boldsymbol{V}_i^T - \mathbb{I}_d \right) \leq O \left(\sqrt{\frac{d + \log(1/\gamma)}{n}} \right) ,$$

which directly implies the desired fact. □

Fact 5. *For every $\alpha > 0$, with probability $1 - O(\gamma)$, when*

$$n = \Omega \left(\frac{d + \log(1/\gamma)}{\alpha^2} \right) ,$$

it holds that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{V}_i \right\|_2^2 \leq O(\alpha^2) \, .$$

Proof. The vector sum inside the desired term has a multivariate Gaussian distribution, and we can bound its ℓ_2-norm with high probability by the classical sub-gaussian concentration bounds in Diakonikolas et al. (2019a). More specifically, with probability $1 - O(\gamma)$, we have that:

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{V}_i \right\|_2 \leq O \left(\sqrt{\frac{d + \log(1/\gamma)}{n}} \right) ,$$

which directly implies the desired fact. □

Directly combining Lemma 12, Lemma 13 with Fact 4 and Fact 5, one obtains the stated Claim 20.

D.2.3 Proof of Claim 21

Initially, breaking $\boldsymbol{Q}_2$ as

$$\boldsymbol{Q}_2 = \widehat{\boldsymbol{\mu}}' - \frac{1}{n} X^T \boldsymbol{y} = \left(\widehat{\boldsymbol{\mu}}' - \boldsymbol{\mu}' \right) - \left(\frac{1}{n} X^T \boldsymbol{y} - \boldsymbol{\mu}' \right) ,$$

it follows that

$$\left\| \Sigma^{-1/2} \boldsymbol{Q}_2 \right\|_2^2 \leq \left\| \Sigma^{-1/2} A'^{-1} \right\|_2^2 \cdot \left\| A' \left(\widehat{\boldsymbol{\mu}}' - \boldsymbol{\mu}' \right) \right\|_2^2 + \left\| \frac{1}{n} \sum_{i=1}^{n} y_i \boldsymbol{V}_i - \Sigma^{-1/2} \boldsymbol{\mu}' \right\|_2^2 ,$$

where we have again used the notation of the variance-normalized vectors $\boldsymbol{V}_i = \Sigma^{-1/2} \boldsymbol{X}_i$, and the "private preconditioner" matrix A' (in this case, obtained for the mean estimation of the random vectors $y_i \boldsymbol{X}_i$), due to Lemma 22, as detailed below (see the first lines of the proof of Fact 6).

Before stating and proving the facts which lead to the desired result, we need to prove the sub-gaussianity of the vectors $y_i \boldsymbol{X}_i$ that are crucial for the conditions of Lemma 22 and the rest of our proof.

Proposition 24 (Sub-gaussianity of $y_i \boldsymbol{X}_i$)**.** *Let $\boldsymbol{X}_i$ be a random vector sampled according to a multivariate Gaussian distribution with mean value $\boldsymbol{\mu}$ and covariance matrix Σ such that $\mathbb{I}_d \preceq \Sigma$, and y_i be a random variable such that $\frac{1}{\rho} \leq |y_i| \leq c$. Then, $y_i \boldsymbol{X}_i$ are sub-gaussian random vectors with covariance matrix Σ' such that*

$$\frac{1}{\rho^2}\mathbb{I}_d \preceq \Sigma' \preceq c^2\Sigma\,,$$

and sub-gaussian norm

$$\|y_i \boldsymbol{X}_i\|_{\psi_2} \leq c\|\boldsymbol{X}_i\|_{\psi_2}\,.$$

Proof. First of all, we have that $\Sigma' = \mathbb{E}\left[y_i^2 \boldsymbol{X}_i \boldsymbol{X}_i^T\right] \preceq c^2\Sigma$, since the eigenvalues of $\mathbb{E}\left[(c^2 - y_i^2)\boldsymbol{X}_i \boldsymbol{X}_i^T\right]$ are non-negative, because the quantity inside the expectation is always a positive semi-definite matrix, and expectation is a linear operator, thus the eigenvalues of the matrix in expectation are also non-negative, by the Courant-Fischer min-max theorem.

Similarly, it can be seen that $\frac{1}{\rho^2}\mathbb{I}_d \preceq \frac{1}{\rho^2}\Sigma \preceq \Sigma'$. We proceed to prove the second part of the proposition, which is the sub-gaussian norm inequality.

We prove the sub-gaussianity of the desired vectors, by Definition 7: consider a unit vector $\boldsymbol{u} \in S^{d-1}$, then it holds that

$$\mathbb{E}\left[\exp(\lambda^2 y_i^2 \langle \boldsymbol{X}_i, \boldsymbol{u}\rangle^2)\right] \leq \mathbb{E}\left[\exp((\lambda c)^2 \langle \boldsymbol{X}_i, \boldsymbol{u}\rangle^2)\right] \leq \exp(\lambda^2 (cK)^2)\,,$$

for all $\lambda : |\lambda| \leq 1/(Kc)$, where K is the sub-gaussian norm of $\langle \boldsymbol{X}_i, \boldsymbol{u}\rangle$. The sub-gaussian norm follows:

$$\|y_i \boldsymbol{X}_i\|_{\psi_2} \leq c\|\boldsymbol{X}_i\|_{\psi_2}\,.$$

□

We are now ready to present the facts that lead to the claim.

Fact 6. *With probability $1 - O(\gamma)$, when*

$$n = \Omega\left(\frac{d^{3/2}\sqrt{\log(\kappa\rho c)}\mathrm{polylog}\left(\frac{d\log(\kappa\rho c)}{\gamma\epsilon}\right)}{\epsilon}\right)\,,$$

it holds that

$$\left\|\Sigma^{-1/2} A'^{-1}\right\|_2^2 \leq O(c^2)\,.$$

Proof. According to Proposition 24, the conditions of Lemma 22 apply to the variables $\rho y_i \boldsymbol{X}_i$ by a change of variables in the sample complexity of $\kappa' = \rho^2 c^2 \kappa$, where we remind to the reader that κ is the largest eigenvalue of the covariance matrix of the feature vectors: $\Sigma \preceq \kappa \mathbb{I}_d$.

Therefore, with

$$n = \Omega\left(\frac{d^{3/2}\sqrt{\log(\kappa\rho c)}\mathrm{polylog}\left(\frac{d\log(\kappa\rho c)}{\gamma\epsilon}\right)}{\epsilon}\right)\,,$$

we obtain a matrix A' (which is ρ times the A given by the algorithm of Lemma 22 as stated in the previous paragraph) such that with probability $1 - O(\gamma)$,

$$\mathbb{I}_d \preceq A'\Sigma' A' \preceq 1000\mathbb{I}_d\,. \tag{D.5}$$

Note that the knowledge of ρ is *not* required for Algorithm 1, since the change of variables that we did only affects the analysis that we performed here (the privacy of the algorithm is solely based on the upper bound c of the labels, and *not* on ρ).

Finally, the desired result holds with probability $1 - O(\gamma)$:

$$\left\|\Sigma^{-1/2} A'^{-1}\right\|_2^2 \leq \left\|\Sigma^{-1/2}\Sigma'^{1/2}\right\|_2^2 \cdot \left\|\Sigma'^{-1/2} A'^{-1}\right\|_2^2 \leq c^2 \cdot 1 = c^2\,,$$

since for the two quantities of interest we have separately the following:

By Proposition 24 and properties of the positive semi-definite order, we have that

$$\begin{aligned}
&\Sigma' \preceq c^2\Sigma \\
\Rightarrow\ &c^2\Sigma'^{-1} \succeq \Sigma^{-1} \\
\Rightarrow\ &\Sigma^{-1} - c^2\Sigma'^{-1} \preceq O \\
\Rightarrow\ &\Sigma'^{1/2}\Sigma^{-1}\Sigma'^{1/2} \preceq c^2\mathbb{I}_d \\
\Rightarrow\ &\left\|\Sigma^{-1/2}\Sigma'^{1/2}\right\|_2^2 \leq c^2\,.
\end{aligned}$$

At the same time, by Eq. (D.5), we obtain the final term:

$$\left\|\Sigma'^{-1/2}A'^{-1}\right\|_2^2 = \frac{1}{\sigma_{\min}^2\left(\Sigma'^{1/2}A'\right)} = \frac{1}{\lambda_{\min}\left(\left(\Sigma'^{1/2}A'\right)^T\Sigma'^{1/2}A'\right)} = \frac{1}{\lambda_{\min}\left(A'\Sigma'A'\right)} \leq 1\,.$$

□

Fact 7. *For every $\eta > 0$, with probability $1 - O(\gamma)$, when*

$$n = \Omega\left(\frac{d\log(\frac{d}{\gamma})}{\eta^2} + \frac{d\text{polylog}(\frac{d\log(1/\delta)}{\eta\gamma\epsilon})}{\eta\epsilon} + \frac{\sqrt{d}\log(\frac{d}{\gamma\delta})}{\epsilon} + \frac{d^{3/2}\sqrt{\log(\kappa\rho c)}\text{polylog}\left(\frac{d\log(\kappa\rho c)}{\gamma\epsilon}\right)}{\epsilon}\right)\,,$$

it holds that

$$\|A'\left(\widehat{\boldsymbol{\mu}}' - \boldsymbol{\mu}'\right)\|_2^2 \leq O(\eta^2)\,.$$

Proof. This fact is a direct implication of Proposition 24 which guarantees the conditions for Lemma 13 to hold. □

Fact 8. *For every $\eta > 0$, with probability $1 - O(\gamma)$, when*

$$n = \Omega\left(\frac{d + \log(1/\gamma)}{\eta^2}\right)\,,$$

it holds that

$$\left\|\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{V}_i - \Sigma^{-1/2}\boldsymbol{\mu}'\right\|_2^2 \leq O(\eta^2)\cdot c^2\,.$$

Proof. We will prove that, under the stated conditions,

$$\left\|\frac{1}{n}\sum_{i=1}^n \frac{y_i\boldsymbol{V}_i}{c} - \Sigma^{-1/2}\frac{\boldsymbol{\mu}'}{c}\right\|_2 \leq O(\eta)\,,$$

and the result will follow.

First of all, we prove that $\frac{y_i\boldsymbol{V}_i}{c}$ is sub-gaussian and calculate a bound on its sub-gaussian norm. Similarly to the sub-gaussianity of Proposition 24, and by Lemma 8, we have that

$$\left\|\frac{y_i\boldsymbol{V}_i}{c}\right\|_{\psi_2} = \frac{\|y_i\boldsymbol{V}_i\|_{\psi_2}}{c} \leq C_1\left\|\boldsymbol{V}_i\right\|_{\psi_2} \leq C_2\,,$$

for some universal constants $C_1, C_2 > 0$, since $\boldsymbol{V}_i = \Sigma^{-1/2}\boldsymbol{X}_i$ are variance-normalized random vectors.

Then, noting that

$$\mathbb{E}\left[\frac{y_i\boldsymbol{V}_i}{c}\right] = \Sigma^{-1/2}\frac{\boldsymbol{\mu}'}{c}\,,$$

and by Lemma 8, it holds that the (centered) quantity $\frac{y_i \boldsymbol{V}_i}{c} - \Sigma^{-1/2}\frac{\boldsymbol{\mu}'}{c}$ is also sub-gaussian with sub-gaussian norm at most a constant times the sub-gaussian norm of the non-centered random vector $\frac{y_i \boldsymbol{V}_i}{c}$. Notice that the covariance matrix of the centered quantity above is $\preceq \mathbb{I}_d$. We will leverage this relationship, alongside the sub-gaussianity of the quantity, to prove the final concentration inequality, from which the fact follows:

Lemma 25. *There exist universal constants $A, B > 0$ such that, for all $t > 0$,*

$$\mathbf{Pr}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{y_i \boldsymbol{V}_i}{c} - \mathbb{E}\left[\frac{y_i \boldsymbol{V}_i}{c}\right]\right\|_2 > t\right] \leq 4\exp\left(Ad - Bnt^2\right)$$

Proof. We denote the covariance matrix of $\frac{y_i \boldsymbol{V}_i}{c}$ as Σ'', for which it holds that $\Sigma'' \preceq \mathbb{I}_d$, and we also name the variance-normalized random vectors $(\Sigma'')^{-1/2}\frac{y_i \boldsymbol{V}_i}{c}$ as $\boldsymbol{W}_i$ (therefore, $\mathbb{E}[\boldsymbol{W}_i \boldsymbol{W}_i^T] = \mathbb{I}_d$).

By a classical result of sub-gaussian concentration inequalities (see, e.g., Lemma 2.21 of Diakonikolas et al. (2019a)), we have that there exist universal constants $A, B > 0$ such that, for all $t > 0$,

$$\mathbf{Pr}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{W}_i - \mathbb{E}\left[\boldsymbol{W}_i\right]\right\|_2 > t\right] \leq 4\exp\left(Ad - Bnt^2\right) .$$

Additionally, by definition of the spectral norm, and since $\Sigma'' \preceq \mathbb{I}_d$, we have that:

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{W}_i - \mathbb{E}\left[\boldsymbol{W}_i\right]\right\|_2 &\geq \left\|(\Sigma'')^{1/2}\right\|_2 \left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{W}_i - \mathbb{E}\left[\boldsymbol{W}_i\right]\right\|_2 \\ &\geq \left\|(\Sigma'')^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{W}_i - \mathbb{E}\left[\boldsymbol{W}_i\right]\right)\right\|_2 ,\end{aligned}$$

namely, that:

$$\begin{aligned}\mathbf{Pr}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{y_i \boldsymbol{V}_i}{c} - \mathbb{E}\left[\frac{y_i \boldsymbol{V}_i}{c}\right]\right\|_2 > t\right] &= \mathbf{Pr}\left[\left\|(\Sigma'')^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{W}_i - \mathbb{E}\left[\boldsymbol{W}_i\right]\right)\right\|_2 > t\right] \\ &\leq \mathbf{Pr}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{W}_i - \mathbb{E}\left[\boldsymbol{W}_i\right]\right\|_2 > t\right] \\ &\leq 4\exp\left(Ad - Bnt^2\right) ,\end{aligned}$$

as desired. □

Utilizing Lemma 25, Fact 8 follows. □

Directly combining Fact 6, Fact 7, and Fact 8, we obtain the stated Claim 21.

E Proof of Theorem 4

Again, for convenience, we restate here the (stonger) version of Theorem 4 that we will prove here:

Theorem 26 (Privacy and Accuracy of $\widehat{\boldsymbol{\beta}}$ in Private Binary Regression)**.** *Under Assumption 1 with covariance parameter κ and Assumption 2 with true parameter $\boldsymbol{\beta} \in \mathbb{R}^d$, for every privacy parameters $\epsilon, \delta > 0$, accuracy parameters $\alpha, \eta > 0$ and confidence $\gamma \in (0,1)$,* PrivLearnLSE *(defined in Algorithm 1) with $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}} = \mathbf{0}$ is $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-differentially private. Moreover, if the number of labeled examples is at least:*

$$\begin{aligned}n = O\left(\frac{d\log(\frac{d}{\gamma})}{\eta^2} + \frac{d\text{polylog}(\frac{d\log(1/\delta)}{\eta\gamma\epsilon})}{\eta\epsilon} + \frac{d^{3/2}\sqrt{\log\kappa}\text{polylog}\left(\frac{d\log\kappa}{\gamma\epsilon\delta}\right)}{\epsilon}\right) \\ + O\left(\frac{d\log(\frac{d}{\gamma})}{\alpha^2} + \frac{d^{3/2}\text{polylog}\left(\frac{d\log(1/\delta)}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon}\right) ,\end{aligned}$$

then with probability at least $1 - O(\gamma)$ *an estimate* $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$ *is successfully output and satisfies*

$$\|\widehat{\boldsymbol{\beta}} - k\boldsymbol{\beta}\|_2^2 \leq \|\widehat{\boldsymbol{w}} - k\boldsymbol{w}\|_2^2 \leq O\left(\alpha^2\right) \cdot \left(1 + \|k\boldsymbol{w}\|_2^2\right) + O\left(\eta^2\right) ,$$

where $\widehat{\boldsymbol{w}} = \Sigma^{1/2}\widehat{\boldsymbol{\beta}}$, $\boldsymbol{w} = \Sigma^{1/2}\boldsymbol{\beta}$ *and* $k = \frac{2n}{n-d-1}\,\mathbb{E}\left[f'\left(\boldsymbol{\beta}^T\boldsymbol{X}_i\right)\right]$. *Finally,* PrivLearnLSE *runs in* $\mathrm{poly}(n)$ *time.*

Proof. The privacy of the algorithm in this Theorem arises directly from the privacy of Theorem 3, since the algorithm is the same.

For the accuracy guarantee, as discussed in the Technical overview (Section 6), we first break the norm of the vector difference $\left\|\widehat{\boldsymbol{\beta}} - k\boldsymbol{\beta}\right\|_2^2$ into the distance from the estimate $\boldsymbol{\beta}_s^\star$, for which we remind to the reader that we define as

$$\boldsymbol{\beta}_s^\star = \left(\frac{1}{n}\sum_{i=n+1}^{2n} \boldsymbol{X}_i\boldsymbol{X}_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\boldsymbol{X}_i\right) , \tag{E.1}$$

that resembles the Least Squares Estimate but *crucially introduces independence* between the two terms that constitute the Least Squares Estimate (to which we can apply our result from Theorem 3), and the distance of the Least-Squares-resembling estimate from a constant multiplicative factor of the true regression coefficient $\boldsymbol{\beta}$ (respectively, of the estimate $\boldsymbol{w}_s^\star$ from $\boldsymbol{w} = \Sigma^{1/2}\boldsymbol{\beta}$), as follows:

$$\|\widehat{\boldsymbol{w}} - k\boldsymbol{w}\|_2^2 \leq \|\widehat{\boldsymbol{w}} - \boldsymbol{w}_s^\star\|_2^2 + \|\boldsymbol{w}_s^\star - k\boldsymbol{w}\|_2^2 .$$

The first term gets bounded by Theorem 3, since as we also noted in the Technical overview (Section 6), the independence between Q_1 and $\boldsymbol{Q}_2$ in our proof of Theorem 3 allows us to prove the same claim for $\boldsymbol{\beta}_s^\star$ as we did for $\boldsymbol{\beta}^\star$ above (see Appendix D). In the remainder of the proof, we focus on bounding the second term.

We first supply the following central Lemma, which uncovers the (unbiased up to a multiplicative factor) relation between the Least-Squares-resembling estimate $\boldsymbol{\beta}_s^\star$ and the true regression coefficient $\boldsymbol{\beta}$, following from Stein's Lemma:

Lemma 27. *There exists a multiplicative factor* $k \in \mathbb{R}_+$ *that depends on the model function* f*, where* f *as defined in Assumption 2, such that the estimate* $\boldsymbol{\beta}_s^\star$*, as in Eq. (E.1), is an unbiased up to a multiplicative factor estimate of the true parameter* $\boldsymbol{\beta}$ *of Assumption 2, i.e.,*

$$\mathbb{E}\left[\boldsymbol{\beta}_s^\star\right] = k\boldsymbol{\beta} .$$

Proof. First, we note the following equality following from the definitions of Assumption 2:

$$\mathbb{E}\left[y_i|\boldsymbol{X}_i\right] = 2f\left(\boldsymbol{\beta}^T\boldsymbol{X}_i\right) - 1 . \tag{E.2}$$

Also, by a classical result on Wishart matrices (for instance, see Anderson (2003)), it is true that the inverse sample covariance matrix is proportional to the true covariance matrix for multivariate Gaussian random vectors:

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=n+1}^{2n} \boldsymbol{X}_i\boldsymbol{X}_i^T\right)^{-1}\right] = \frac{n}{n-d-1}\Sigma^{-1} . \tag{E.3}$$

By the independence of the first n samples $(1 \ldots n)$ from the next $(n+1 \ldots 2n)$, the law of iterated expectations, using Eq. (E.2), Eq. (E.3) and the zero-mean property of the feature vectors $\boldsymbol{X}_i$, we have that:

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\beta}_s^\star\right] &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=n+1}^{2n} \boldsymbol{X}_i\boldsymbol{X}_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\boldsymbol{X}_i\right)\right] \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=n+1}^{2n} \boldsymbol{X}_i\boldsymbol{X}_i^T\right)^{-1}\right]\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i\,\mathbb{E}[y_i|\boldsymbol{X}_i]\right] \\
&= \frac{n}{n-d-1}\Sigma^{-1}\,\mathbb{E}\left[\boldsymbol{X}_i\left(2f\left(\boldsymbol{\beta}^T\boldsymbol{X}_i\right) - 1\right)\right] \\
&= \frac{n}{n-d-1}\Sigma^{-1}\,\mathbf{Cov}\left[\boldsymbol{X}_i,\ 2f\left(\boldsymbol{\beta}^T\boldsymbol{X}_i\right) - 1\right] ,
\end{aligned}$$

Now, an application of Stein's Lemma (see Lemma 11), since $\boldsymbol{X}_i$ and $\boldsymbol{\beta}^T\boldsymbol{X}_i$ are jointly Gaussian, suggests that

$$\mathbf{Cov}\left[\boldsymbol{X}_i,\ 2f\left(\boldsymbol{\beta}^T\boldsymbol{X}_i\right)-1\right]=2\,\mathbf{Cov}\left[\boldsymbol{X}_i,\boldsymbol{\beta}^T\boldsymbol{X}_i\right]\mathbb{E}\left[f'\left(\boldsymbol{\beta}^T\boldsymbol{X}_i\right)\right]=2\,\mathbb{E}\left[f'\left(\boldsymbol{\beta}^T\boldsymbol{X}_i\right)\right]\ \Sigma\,\boldsymbol{\beta}\,,$$

and combining with the above equality yields

$$\mathbb{E}\left[\boldsymbol{\beta}_s^\star\right]=k\boldsymbol{\beta}\,,$$

where

$$k=\frac{2n}{n-d-1}\,\mathbb{E}\left[f'\left(\boldsymbol{\beta}^T\boldsymbol{X}_i\right)\right]\,.$$

□

Continuing to the proof of the result, we use the form as written with the expectation, breaking the term into three sub-terms by adding and subtracting the same quantities (defining the variance-normalized vectors $\boldsymbol{V}_i=\Sigma^{-1/2}\boldsymbol{X}_i$), to deduce that

$$\begin{aligned}
\|\boldsymbol{w}_s^\star-k\boldsymbol{w}\|_2^2&=\|\boldsymbol{w}_s^\star-\mathbb{E}\left[\boldsymbol{w}_s^\star\right]\|_2^2\\
&=\left\|\Sigma^{1/2}\left(\frac{1}{n}\sum_{i=n+1}^{2n}\boldsymbol{X}_i\boldsymbol{X}_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}y_i\boldsymbol{X}_i\right)-\frac{n}{n-d-1}\Sigma^{-1/2}\,\mathbb{E}\left[y_j\boldsymbol{X}_j\right]\right\|_2^2\\
&\le 2\left\|\left(\frac{1}{n}\sum_{i=n+1}^{2n}\boldsymbol{V}_i\boldsymbol{V}_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}y_i\boldsymbol{V}_i\right)-\left(\frac{1}{n}\sum_{i=1}^{n}y_i\boldsymbol{V}_i\right)\right\|_2^2\\
&+2\left\|\left(\frac{1}{n}\sum_{i=1}^{n}y_i\boldsymbol{V}_i\right)-\mathbb{E}\left[y_j\boldsymbol{V}_j\right]\right\|_2^2+2\left\|\frac{d+1}{n-d-1}\,\mathbb{E}\left[y_j\boldsymbol{V}_j\right]\right\|_2^2\\
&=2\left\|\left(\frac{1}{n}\sum_{i=n+1}^{2n}\boldsymbol{V}_i\boldsymbol{V}_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=n+1}^{2n}\boldsymbol{V}_i\boldsymbol{V}_i^T-\mathbb{I}_d\right)\left(\frac{1}{n}\sum_{i=1}^{n}y_i\boldsymbol{V}_i\right)\right\|_2^2\\
&+2\left\|\frac{1}{n}\sum_{i=1}^{n}\left(y_i\boldsymbol{V}_i-\mathbb{E}\left[y_j\boldsymbol{V}_j\right]\right)\right\|_2^2+2\left\|\frac{d+1}{n-d-1}\,\mathbb{E}\left[y_j\boldsymbol{V}_j\right]\right\|_2^2\,.
\end{aligned}$$

When we have at least as many samples as required for Theorem 3, it follows from the sub-proofs presented in the proof of this Theorem above (specifically, Fact 3, Fact 4, Fact 5, and Fact 8 of Appendix D) that the whole quantity is bounded as follows:

$$\|\boldsymbol{w}_s^\star-\mathbb{E}\left[\boldsymbol{w}_s^\star\right]\|_2^2\le O\left(\alpha^2\right)+O\left(\eta^2\right)\,.$$

□

F Algorithm and Guarantees on Standard Linear Regression

We begin this section with a description of Algorithm 2. First, we can deduce from Eq. (4.6) that the (marginal) distribution of labels y_i is a Gaussian distribution $\mathcal{N}(\boldsymbol{\beta}^T\boldsymbol{\mu},\boldsymbol{\beta}^T\Sigma\boldsymbol{\beta}+\sigma_\epsilon^2)$, therefore, defining the vectors $\boldsymbol{Z}_i\in\mathbb{R}^{d+1}$ as

$$\boldsymbol{Z}_i=\begin{bmatrix}\boldsymbol{X}_i\\ y_i\end{bmatrix}\,,\tag{F.1}$$

we can see that they similarly follow a Gaussian distribution, with a covariance matrix Σ' that may be written in a block matrix form:

$$\Sigma'=\mathbb{E}\left[\boldsymbol{Z}_i\boldsymbol{Z}_i^T\right]=\begin{bmatrix}\Sigma & \Sigma\boldsymbol{\beta}\\ \boldsymbol{\beta}^T\Sigma & \sigma_\epsilon^2+\boldsymbol{\beta}^T\Sigma\boldsymbol{\beta}\end{bmatrix}\,.\tag{F.2}$$

Eq. (F.2) indicates that a natural way to estimate $\boldsymbol{\beta}$ would be to estimate the covariance matrices Σ,Σ' and then, extracting the last column of Σ' (ExtractLastColumn in Algorithm 2), which (without the last element) is

Algorithm 2 Private Estimation of Linear Regression Coefficient.

Input: $(X, \boldsymbol{y}) = (\boldsymbol{X}_i, y_i)_{i\in[n]}$ with $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu}, \Sigma$ are unknown and n satisfies Theorem 5.
Parameters: Privacy $\epsilon, \delta > 0$, accuracy $\alpha, \eta > 0$, confidence $\gamma \in (0,1)$, covariance spectral norm bound κ.
Output: Estimate $\widehat{\boldsymbol{\beta}}$ that approaches the true vector $\boldsymbol{\beta}$ in L_2 norm with high probability.
procedure PrivLearnLinear($(X, \boldsymbol{y}), \epsilon, \delta, \alpha, \eta, \gamma, \kappa$)
  Set $\boldsymbol{Z}_i \leftarrow [\boldsymbol{X}_i, y_i]^T$ for $i \in [n]$
  $L \leftarrow \{\Theta(\epsilon), \Theta(\delta), \Theta(\alpha), \gamma, \kappa\}$
  $\widehat{\Sigma'} \leftarrow$ LearnGaussian-hd($\{\boldsymbol{Z}_i\}_{i\in[n]}, L$)
  $[\widehat{\Sigma\boldsymbol{\beta}},\ \widehat{\sigma}] \leftarrow$ ExtractLastColumn($\widehat{\Sigma'}$) ▷ *See Eq. (5.3),* $\widehat{\Sigma\boldsymbol{\beta}} \in \mathbb{R}^d, \widehat{\sigma} = \sigma_\epsilon^2 + \widehat{\boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}} \in \mathbb{R}$.
  Draw $\boldsymbol{X}_i, i \in \{n+1, \dots, 2n\}$ from $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$
  $\widehat{\Sigma} \leftarrow$ LearnGaussian-hd($\{\boldsymbol{X}_i\}_{i\in[n+1..2n]}, L$)
  if $\widehat{\Sigma}$ is not invertible[5] **then** Output $\perp$
  Output the private estimate $\widehat{\boldsymbol{\beta}} = \widehat{\Sigma}^{-1}\widehat{\Sigma\boldsymbol{\beta}}$

$\Sigma\boldsymbol{\beta}$, left multiply by the inverse of the estimate of Σ that we have. Indeed, we show that this approach works and, when the estimation of the above covariance matrices is made according to the differentially private algorithms of Gaussian covariance estimation, the end result is a differentially private estimator of $\boldsymbol{\beta}$ for the "standard linear regression" model of Assumption 3.

Description of Algorithm 2. Having access to the n i.i.d. samples $(\boldsymbol{X}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, where $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma), i \in [n]$, the algorithm initially computes a differentially private estimate $\widehat{\Sigma}$ of the covariance matrix of the d-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, using the algorithm LearnGaussian-hd, discussed in Section 3. Then, the algorithm forms the random vectors $\boldsymbol{Z}_i \in \mathbb{R}^{d+1}$ as in Eq. (F.1) and computes a differentially private estimate $\widehat{\Sigma'}$ of the covariance matrix of the $(d+1)$-dimensional Gaussian distribution with covariance matrix of the block form of Eq. (F.2), again using the algorithm LearnGaussian-hd. From the matrix $\widehat{\Sigma'}$, the algorithm obtains only the first d elements of the last column of that matrix, naming them as $\widehat{\Sigma\boldsymbol{\beta}} \in \mathbb{R}^d$ (hinting at the form of Eq. (F.2)). Armed with these estimates, the differentially private estimate of $\boldsymbol{\beta}$ is finally given by:

$$\widehat{\boldsymbol{\beta}} = \widehat{\Sigma}^{-1}\widehat{\Sigma\boldsymbol{\beta}}\,, \tag{F.3}$$

whose privacy follows from appropriate composition rules. We now proceed to prove the accuracy guarantee of this estimate.

Theorem 28 (Accuracy of $\widehat{\boldsymbol{\beta}}$ in Private Standard Linear Regression)**.** *Under Assumption 3 with parameters (κ, Σ') where Σ' is defined as in Eq. (F.2), for all privacy parameters $\epsilon, \delta > 0$, accuracy parameters $\alpha, \eta > 0$ and confidence $\gamma \in (0,1)$, there exists an algorithm (Algorithm 2) that is $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-differentially private, and if the number of labeled examples is at least:*

$$n = O\left(\frac{d + \log(1/\gamma)}{\eta^2} + \frac{d^{3/2}\mathrm{polylog}\left(\frac{d}{\eta\gamma\epsilon}\right)}{\eta\epsilon} + \frac{d^{3/2}\sqrt{\log\left(\kappa(\Sigma')\right)}\mathrm{polylog}\left(\frac{d\log\left(\kappa(\Sigma')\right)}{\gamma\epsilon}\right)}{\epsilon}\right)$$
$$+ O\left(\frac{d + \log(1/\gamma)}{\alpha^2} + \frac{d^{3/2}\mathrm{polylog}\left(\frac{d}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon}\right),$$

then with probability at least $1 - O(\gamma)$ an estimate $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$ is successfully output and along with the "true" regression coefficient $\boldsymbol{\beta}$ satisfies:

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2^2 \le \|\widehat{\boldsymbol{w}} - \boldsymbol{w}\|_2^2 \le O\left(\alpha^2\right) \cdot \|\boldsymbol{w}\|_2^2 + O\left(\eta^2\right) \cdot \lambda_{max}^2(\Sigma')\,, \tag{F.4}$$

[5] The invertibility of the matrix in **Line 10** holds with high probability and the non-invertibility bad event is captured by the $O(\gamma)$ failure probability of Theorem 28.

where $\kappa(\Sigma') = \frac{\lambda_{max}(\Sigma')}{\lambda_{min}(\Sigma')}$ *is the condition number of the block matrix* Σ' *as in Eq. (F.2),* $\widehat{\boldsymbol{w}} = \Sigma^{1/2}\widehat{\boldsymbol{\beta}}$ *and* $\boldsymbol{w} = \Sigma^{1/2}\boldsymbol{\beta}$*. Finally, the algorithm runs in* $\text{poly}(n)$ *time.*

Proof. The privacy of the algorithm in this Theorem arises directly from the privacy of the differentially private covariance estimation algorithm and the composition theorems.

For the accuracy guarantee, we begin by adding and subtracting the quantities of each factor of $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$, as follows:

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\widehat{\Sigma}^{-1} - \Sigma^{-1}\right)\Sigma\boldsymbol{\beta} + \widehat{\Sigma}^{-1}\left(\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}\right) .$$

Then, substituting the quantities $\boldsymbol{w} = \Sigma^{1/2}\boldsymbol{\beta}$ and left multiplying both sides of the equation by $\Sigma^{1/2}$, we obtain that

$$\begin{aligned} \widehat{\boldsymbol{w}} - \boldsymbol{w} &= \left(\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2} - \mathbb{I}_d\right)\boldsymbol{w} + \Sigma^{1/2}\widehat{\Sigma}^{-1}\left(\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}\right) \\ \Leftrightarrow \widehat{\boldsymbol{w}} - \boldsymbol{w} &= \left(\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2}\right)\left[-\left(\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - \mathbb{I}_d\right)\boldsymbol{w} + \Sigma^{-1/2}\left(\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}\right)\right] . \end{aligned}$$

Using Cauchy-Schwartz and the sub-multiplicative property of the spectral norm, we establish the following inequality:

$$\|\widehat{\boldsymbol{w}} - \boldsymbol{w}\|_2^2 \leq 2\left\|\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2}\right\|_2^2\left(\left\|\Sigma^{-1/2}\left(\widehat{\Sigma} - \Sigma\right)\Sigma^{-1/2}\right\|_2^2 \cdot \|\boldsymbol{w}\|_2^2 + \left\|\Sigma^{-1/2}\left(\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}\right)\right\|_2^2\right) .$$

In order to bound the constituent terms of the right-hand-side of this inequality, we need a very similar claim to Claim 19 which we state below without proof (since it is almost the same as Appendix D.2.1), Lemma 12 (which we remind to the reader that it is applied to the $2n$ sample differences $\frac{1}{\sqrt{2}}(\boldsymbol{X}_{2i} - \boldsymbol{X}_{2i-1})$, so that they have zero mean), and Claim 30, which is the main subject that we elaborate on in Appendix F.1.

Claim 29 (Similar to Claim 19)**.** *When*

$$n = \Omega\left(d + \log(1/\gamma) + \frac{d^{3/2}\sqrt{\log\kappa}\text{polylog}\left(\frac{d\log\kappa}{\gamma\epsilon}\right)}{\epsilon}\right) ,$$

the following inequality holds with probability $1 - O(\gamma)$*:*

$$\left\|\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2}\right\|_2^2 \leq O\left(1\right) .$$

Claim 30. *When*

$$n = \Omega\left(\frac{d + \log(1/\gamma)}{\eta^2} + \frac{d^{3/2}\text{polylog}\left(\frac{d}{\eta\gamma\epsilon}\right)}{\eta\epsilon} + \frac{d^{3/2}\sqrt{\log\left(\kappa(\Sigma')\right)}\text{polylog}\left(\frac{d\log\left(\kappa(\Sigma')\right)}{\gamma\epsilon}\right)}{\epsilon}\right) ,$$

the following inequality holds with probability $1 - O(\gamma)$*:*

$$\left\|\Sigma^{-1/2}\left(\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}\right)\right\|_2^2 \leq O\left(\eta^2\right) \cdot \lambda_{max}^2(\Sigma') .$$

Combining Claim 29, Lemma 12, and Claim 30 with a union bound of the respective events, we directly obtain Theorem 28, since

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2^2 = \left\|\Sigma^{-1/2}\Sigma^{1/2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right\|_2^2 \leq \left\|\Sigma^{-1/2}\right\|_2^2 \cdot \|\widehat{\boldsymbol{w}} - \boldsymbol{w}\|_2^2 \leq \|\widehat{\boldsymbol{w}} - \boldsymbol{w}\|_2^2 ,$$

by the sub-multiplicative property of the norm and because $\mathbb{I}_d \preceq \Sigma$. □

F.1 Proof of Claim 30

First of all, we note that, according to the first lines of the proof of Fact 6, in order to apply the (accuracy) results of Lemma 22 and Lemma 12 to the covariance estimation of the random vectors $\boldsymbol{Z}_i \in \mathbb{R}^{d+1}$ as in Eq. (F.1), a change of variables is needed, that affects solely the analysis of the algorithm (and more specifically, appears in a change of the sample complexity). Therefore, the specific result which applies in our case here is stated in the following fact.

Fact 9 (Covariance $\widehat{\Sigma}'$ estimation accuracy)**.** *For every $\eta > 0$, the output $\widehat{\Sigma}'$ of algorithm* LearnGaussian-hd *when given at least n samples $\boldsymbol{Z}_i$ with*

$$n = O\left(\frac{d+\log(1/\gamma)}{\eta^2} + \frac{d^{3/2}\mathrm{polylog}\left(\frac{d}{\eta\gamma\epsilon}\right)}{\eta\epsilon} + \frac{d^{3/2}\sqrt{\log\left(\kappa(\Sigma')\right)}\mathrm{polylog}\left(\frac{d\log\left(\kappa(\Sigma')\right)}{\gamma\epsilon}\right)}{\epsilon}\right),$$

where $\kappa(\Sigma') = \frac{\lambda_{max}(\Sigma')}{\lambda_{min}(\Sigma')}$ is the condition number of the block matrix Σ' as in Eq. (F.2), satisfies the following accuracy guarantee with probability $1 - O(\gamma)$:

$$\left\|\Sigma'^{-1/2}\left(\widehat{\Sigma}' - \Sigma'\right)\Sigma'^{-1/2}\right\|_2 \leq O(\eta)\,. \tag{F.5}$$

We remind to the reader the form of the block matrix Σ' which is as follows:

$$\Sigma' = \begin{bmatrix} \Sigma & \Sigma\boldsymbol{\beta} \\ \boldsymbol{\beta}^T\Sigma & \sigma_\epsilon^2 + \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} \end{bmatrix}. \tag{F.6}$$

By definition of the spectral norm, from Eq. (F.5) we have that for every vector $\boldsymbol{u} \in \mathbb{R}^{d+1} : \|\boldsymbol{u}\|_2 \leq 1$, it holds that $\left\|\Sigma'^{-1/2}\left(\widehat{\Sigma}' - \Sigma'\right)\Sigma'^{-1/2}\boldsymbol{u}\right\|_2 \leq O(\eta)$. Taking advantage of the spectral decomposition of the (real symmetric, PSD) matrix $\Sigma' = U\Lambda U^T$ for some unitary orthogonal matrix U and diagonal matrix Λ, it is well-known that $\Sigma'^{-1/2} = \Lambda^{-1/2}U^T$, and because $\|U^T\boldsymbol{u}\|_2 = \|\boldsymbol{u}\|_2$ for every vector $\boldsymbol{u} \in \mathbb{R}^{d+1}$ (since U is an orthonormal matrix), and since $\sqrt{\lambda_{\max}(\Sigma')}\Lambda^{-1/2} \succeq \mathbb{I}_{d+1}$, we conclude that by choosing $\boldsymbol{u} \in \mathbb{R}^{d+1} : \sqrt{\lambda_{\max}(\Sigma')}\Lambda^{-1/2}\boldsymbol{u} = \boldsymbol{e}_{d+1}$ (where $\boldsymbol{e}_{d+1}$ is the unit vector that has only the $(d+1)$-th coordinate 1 and all other coordinates 0), it is true that $\left\|\Sigma'^{-1/2}\left(\widehat{\boldsymbol{v}}_{d+1} - \boldsymbol{v}_{d+1}\right)\right\|_2^2 \leq O(\eta^2)\cdot\lambda_{\max}(\Sigma')$, where $\widehat{\boldsymbol{v}}_{d+1}, \boldsymbol{v}_{d+1}$ are the last columns of the matrices $\widehat{\Sigma}'$ and Σ' respectively (see Eq. (F.6) for what the last column looks like). Therefore, it is immediate that

$$\begin{aligned}
\|\widehat{\boldsymbol{v}}_{d+1} - \boldsymbol{v}_{d+1}\|_2^2 &= \left\|\Sigma'^{1/2}\Sigma'^{-1/2}\left(\widehat{\boldsymbol{v}}_{d+1} - \boldsymbol{v}_{d+1}\right)\right\|_2^2 \\
&\leq \left\|\Sigma'^{1/2}\right\|_2^2 \cdot \left\|\Sigma'^{-1/2}\left(\widehat{\boldsymbol{v}}_{d+1} - \boldsymbol{v}_{d+1}\right)\right\|_2^2 \\
&\leq O(\eta^2)\cdot\lambda_{\max}^2(\Sigma')\,,
\end{aligned}$$

and since the first d coordinates of $\widehat{\boldsymbol{v}}_{d+1} - \boldsymbol{v}_{d+1} \in \mathbb{R}^{d+1}$ are the vector $\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta} \in \mathbb{R}^d$ (see the structure of Eq. (F.6)), we have that $\left\|\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}\right\|_2^2 \leq \|\widehat{\boldsymbol{v}}_{d+1} - \boldsymbol{v}_{d+1}\|_2^2 \leq O(\eta^2)\cdot\lambda_{\max}^2(\Sigma')$.

Claim 30 follows, since

$$\left\|\Sigma^{-1/2}\left(\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}\right)\right\|_2^2 \leq \left\|\Sigma^{-1/2}\right\|_2^2 \cdot \left\|\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}\right\|_2^2 \leq \left\|\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}\right\|_2^2 \leq O(\eta^2)\cdot\lambda_{\max}^2(\Sigma')\,,$$

by the sub-multiplicative property of the norm and because $\mathbb{I}_d \preceq \Sigma$.

The proof has now been completed. Of course, the condition number $\kappa(\Sigma')$ is an interesting quantity that merits consideration to examine what it depends upon. From Theorem 1 of Dembo (1988), one may deduce the following upper bound on the largest eigenvalue of Σ':

$$\lambda_{\max}(\Sigma') \leq 2\left(\boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} + \max\left(\kappa, \sigma_\epsilon^2\right)\right), \tag{F.7}$$

where we remind to the reader that $\kappa = \lambda_{\max}(\Sigma)$.

The lower bound on the smallest eigenvalue provided by the above work (Dembo, 1988) is non-optimal, since it may be negative at certain cases, while the matrix itself only ever exhibits non-negative eigenvalues (since it is PSD, by definition of being a covariance matrix). A better bound may be deduced by block matrix eigenvalue approaches (Ma and Zarowski, 1995), as follows:

$$\begin{aligned}\lambda_{\min}(\Sigma') &\geq \frac{\sigma_\epsilon^2 + \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} + \lambda_{\min}(\Sigma)}{2} - \sqrt{\left(\frac{\sigma_\epsilon^2 + \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} + \lambda_{\min}(\Sigma)}{2}\right)^2 - \sigma_\epsilon^2\lambda_{\min}(\Sigma)} \\ &= \frac{\sigma_\epsilon^2\lambda_{\min}(\Sigma)}{\frac{1}{2}\left(\sigma_\epsilon^2 + \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} + \lambda_{\min}(\Sigma) + \sqrt{\left(\sigma_\epsilon^2 + \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} + \lambda_{\min}(\Sigma)\right)^2 - 4\sigma_\epsilon^2\lambda_{\min}(\Sigma)}\right)} \\ &\geq \frac{\sigma_\epsilon^2\lambda_{\min}(\Sigma)}{\sigma_\epsilon^2 + \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} + \lambda_{\min}(\Sigma)}\,. \end{aligned} \tag{F.8}$$

The first form of the lower bound given above is tight, as we will now prove by examining the specific case of $\Sigma = \kappa\mathbb{I}_d$. In this case, one would have that:

$$\Sigma' = \begin{bmatrix} \kappa\mathbb{I}_d & \kappa\boldsymbol{\beta} \\ \kappa\boldsymbol{\beta}^T & \sigma_\epsilon^2 + \kappa\|\boldsymbol{\beta}\|_2^2 \end{bmatrix} = \kappa \begin{bmatrix} \mathbb{I}_d & \boldsymbol{\beta} \\ \boldsymbol{\beta}^T & \frac{\sigma_\epsilon^2}{\kappa} + \|\boldsymbol{\beta}\|_2^2 \end{bmatrix},$$

reducing our calculations to the simple case of covariance matrix $\mathbb{I}_d$, for which the second matrix written in the above equation has eigenvalues t according to the roots of the equation

$$\begin{aligned} &(1-t)^d\left(\frac{\sigma_\epsilon^2}{\kappa} + \|\boldsymbol{\beta}\|_2^2 - t\right) - \sum_{i=1}^d \beta_i^2(1-t)^{d-1} = 0 \\ \Leftrightarrow\ &(1-t)^{d-1}\left(t^2 - \left(1 + \|\boldsymbol{\beta}\|_2^2 + \frac{\sigma_\epsilon^2}{\kappa}\right)t + \frac{\sigma_\epsilon^2}{\kappa}\right) = 0\,, \end{aligned}$$

where β_i is the i-th coordinate of the vector $\boldsymbol{\beta}$. Therefore Σ' has the following smallest eigenvalue (the smallest of the two roots of the quadratic equation, which is guaranteed to be ≤ 1, i.e., smaller than the other eigenvalues):

$$\begin{aligned}\lambda_{\min}(\Sigma') &= \frac{\kappa}{2}\left(1 + \|\boldsymbol{\beta}\|_2^2 + \frac{\sigma_\epsilon^2}{\kappa} - \sqrt{\left(1 + \|\boldsymbol{\beta}\|_2^2 + \frac{\sigma_\epsilon^2}{\kappa}\right)^2 - 4\frac{\sigma_\epsilon^2}{\kappa}}\right) \\ &= \frac{\sigma_\epsilon^2 + \kappa\|\boldsymbol{\beta}\|_2^2 + \kappa}{2} - \sqrt{\left(\frac{\sigma_\epsilon^2 + \kappa\|\boldsymbol{\beta}\|_2^2 + \kappa}{2}\right)^2 - \kappa\sigma_\epsilon^2}\,, \end{aligned}$$

which neatly matches the first form of the lower bound given in Eq. (F.8), since $\Sigma = \kappa\mathbb{I}_d$.

Therefore, the condition number $\kappa(\Sigma')$ (which is at most the ratio of the right-hand-sides of Eq. (F.7) to Eq. (F.8)) and the largest eigenvalue $\lambda_{\max}(\Sigma')$ depend on both $\boldsymbol{\beta}$ and σ_ϵ^2 besides the usual dependence on the smallest and largest eigenvalues of the feature vector covariance matrix Σ, i.e., $\lambda_{\max}(\Sigma) \leq \kappa$ and $\lambda_{\min}(\Sigma) \geq 1$ respectively. We note, in particular, that this means that, when $\|\boldsymbol{\beta}\|_2$ is large, more samples will be necessary to achieve a fixed additive accuracy, as indicated by Eq. (F.4).