The Price of Locality in Label Privacy: Optimal Rates for Classification and Regression

Zongrui Zou
State Key Laboratory for Novel Software Technology
Nanjing University
zou.zongrui@smail.nju.edu.cn

Mina Dalirrooyfard
Morgan Stanley
mina.dalirrooyfard@morganstanley.com

Jingcheng Liu*
State Key Laboratory for Novel Software Technology
Nanjing University
liu@nju.edu.cn

Jalaj Upadhyay†
Management Science & Information Systems
Rutgers University
jalaj.upadhyay@rutgers.edu

Abstract

We study classification and regression tasks under label differential privacy (label-DP), a setting where only the labels are considered sensitive information. We quantify the "price of locality" by revealing fundamental statistical gaps between local and central privacy models. For classification, we establish a strict separation in accuracy. Specifically, we prove that the minimax excess risk in the local model is lower bounded by $\Omega(\frac{1}{\varepsilon}\sqrt{\mathrm{VC}(\mathcal{H})/n})$ for classification on size-n datasets, a fundamental limitation that holds even under relaxed approximate (ε, δ)-local DP. Overcoming this barrier, we propose the first *efficient* central-model algorithm that operates under the stricter pure ε-DP, yet achieves a significantly improved upper bound of $\tilde{O}(\sqrt{\mathrm{VC}(\mathcal{H})/(\varepsilon n)})$. For linear regression, we show that local randomizers or naive central-DP algorithms are statistically inefficient. Instead, we turn to the Matrix Mechanism and use an encoder-decoder framework built on ridge regression, which projects the data into a spherical latent space before injecting noise. This structural design effectively isolates and exponentially suppresses the severe noise inflation typically caused by ill-conditioned feature covariance matrices in full-DP (Cai et al., The Annals of Statistics, 49(5)), achieving an optimal privacy penalty of $\tilde{O}(d^2/(n^2\varepsilon^2))$ where d is the feature dimension.

*JL and ZZ have been supported in part by National Natural Science Foundation of China Grant No. 62472212.

†JU's research was funded by the NSF CNS 2433628, Google Seed Fund grant, Google Research Scholar Award, Dean Research Seed Fund, and Rutgers Decanal Grant no. 302918.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

The success of modern machine learning heavily relies on the collection and analysis of massive datasets, which simultaneously raises severe concerns regarding individual privacy. As a rigorous mathematical framework, *differential privacy* (DP) [16] has emerged as the *de facto* standard for privacy-preserving data analysis. Parameterized by privacy parameters ε and δ, a randomized algorithm $\mathcal{A}$ is (ε, δ)-*differentially private* if the probability of obtaining any subset of possible outputs S when run on two "neighboring" datasets D and D' is similar: $\mathbf{Pr}\left[\mathcal{A}(D) \in S\right] \leq e^{\varepsilon} \cdot \mathbf{Pr}\left[\mathcal{A}(D') \in S\right] + \delta$. When $\delta = 0$, we say that $\mathcal{A}$ preserves *pure differential privacy*; otherwise, $\mathcal{A}$ is *approximately differentially private*. In this definition, the specific notion of "neighboring" datasets plays a crucial role in determining what information is considered private.

In the context of supervised machine learning, a dataset D is typically composed of individual entries that pair an input feature with a corresponding label, denoted as $(\mathbf{x}_i, y_i)$. When applying traditional DP mechanisms to such tasks, two datasets are conventionally considered neighbors if they differ by a single full data entry $(\mathbf{x}_i, y_i)$. Consequently, these standard mechanisms attempt to protect the entire dataset by injecting noise into both the input features and the labels.

However, this full-data privatization often incurs a substantial cost to statistical utility, making it challenging to train highly accurate models in real industrial deployment, especially in high-dimensional settings. Fortunately, in many real-world applications, data sensitivity is fundamentally asymmetric. For instance, in medical diagnosis or online advertising, the feature vectors (e.g., patient demographics, user browsing history) are often publicly available or relatively insensitive, whereas the corresponding labels (e.g., clinical disease diagnosis, ad conversion events) constitute the highly sensitive information that requires rigorous protection [9, 30]. This practical asymmetry motivates the study of *label differential privacy*: two datasets are defined as "neighboring" if they differ exclusively in the label of a single data entry. Given a dataset $D = \{(\mathbf{x}_i, y_i)\}_{i\in[n]}$, label DP guarantees the privacy of the labels (i.e., $\{y_i\}_{i\in[n]}$) while allowing the features (i.e., $\{\mathbf{x}_i\}_{i\in[n]}$) to be publicly known. By focusing the privacy budget solely on the labels, label DP holds the promise of achieving significantly better statistical utility than standard DP models.

Within the label DP framework, there are two primary deployment paradigms: the local model and the central model [17]. In the local model, each user privatizes their own label before sending it to an untrusted server, typically via randomized response mechanisms. While offering strong privacy guarantees without relying on a trusted curator, local models often suffer severe statistical utility degradation. Conversely, the central model assumes a trusted curator that collects raw, unperturbed labels and injects noise only at the aggregation or output stage. This approach generally yields significantly higher statistical utility but hinges on the strong assumption that the central server is completely secure and trustworthy.

A line of work has made significant strides in understanding label DP. Chaudhuri et al. [12] first introduced the concept of label differential privacy within the central model, and recent works have investigated the theoretical limits and algorithms for classification under label DP in both the local and central models [7, 19, 20, 27, 39, 41, 42]. Concurrently, for regression tasks, optimal unbiased local randomizers [3] and sparse linear regression frameworks [44] have been explored within the local model. Despite these advances, a fundamental question remains open:

What is the "price of locality" in label DP?

More specifically, how much statistical utility is inherently sacrificed when "imposing" local privacy constraints on fundamental tasks such as classification and regression? *To what extent* can central algorithms bypass these local barriers?

In this paper, we answer the above question by quantifying the statistical gaps between local and central privacy models for two of the most fundamental learning tasks: binary classification and linear regression. Our results are summarized in Section 1.2.

1.1 Notations and Problem Definition

We begin by listing some necessary notations and provide formal definitions for the privacy frameworks used throughout this work. We also introduce the specific settings and learning objectives for the tasks of private binary classification and linear regression.

Notations. For any $\mathbf{v} \in \mathbb{R}^d$ and matrix Σ, we write $\|\mathbf{v}\|_\Sigma^2 = \mathbf{v}^\top \Sigma \mathbf{v}$. Given any real number $0 < x < 1$, we define the entropy function in terms of the binary entropy as $H(x) = -x \log_2 x - (1-x)\log_2(1-x)$. Similarly, we write $H_e(x) = -x \ln x - (1-x)\ln(1-x)$.

1.1.1 General Setup and Label Differential Privacy

Let $\mathcal{X}$ be the *abstract feature space* and $\mathcal{Y}$ be the general *label space*. A dataset $D \in (\mathcal{X} \times \mathcal{Y})^n$ consists of n entries, where we use $(\mathbf{x}_i, y_i)$ to denote the i-th entry of D. Let $\mathcal{P}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$, and we write $D \sim \mathcal{P}^n$ to denote that each element in D is i.i.d. sampled from $\mathcal{P}$. Let $\mathcal{H}$ denote a general *hypothesis class* or *parameter space* mapping from $\mathcal{X}$ to $\mathcal{Y}$. We first introduce the formal definition of label differential privacy in both the central and local models.

Definition 1. (Central Label-DP) Fix any $n \in \mathbb{N}_+$, $\varepsilon > 0$ and $0 \le \delta \le 1$. Let $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$ be a randomized learning algorithm. Then $\mathcal{A}$ preserves (ε, δ)-label (central) differential privacy if for any datasets $D, D' \in (\mathcal{X} \times \mathcal{Y})^n$ differing in the label of at most one data entry, and for any measurable set $S \subseteq \mathcal{H}$,

$$\mathbf{Pr}_{\mathcal{A}}[\mathcal{A}(D) \in S] \le e^\varepsilon \cdot \mathbf{Pr}_{\mathcal{A}}[\mathcal{A}(D') \in S] + \delta.$$

In the (non-interactive) local label differential privacy setting, the private learner does **not** have direct access to the true labels. Instead, alongside n public features $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$, the learner receives n private observations $Z_1, \ldots, Z_n \in \mathcal{Z}$. These observations are generated by independently applying a local randomizer Q to each data entry, such that $Z_i \sim Q(\mathbf{x}_i, y_i)$.

Definition 2. (Local Label-DP) Fix any $\varepsilon > 0$ and $0 \le \delta \le 1$. Let $Q : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a randomized mechanism. We say Q satisfies (ε, δ)-local label differential privacy if for any feature $\mathbf{x} \in \mathcal{X}$, any two labels $y, y' \in \mathcal{Y}$, and any measurable set $S \subseteq \mathcal{Z}$,

$$\mathbf{Pr}_Q[Q(\mathbf{x}, y) \in S] \le e^\varepsilon \cdot \mathbf{Pr}_Q[Q(\mathbf{x}, y') \in S] + \delta.$$

Given the privacy budget ε, δ, our goal is to bound the minimax excess risk for two learning tasks.

Task I: Private Binary Classification. For the binary classification task, we restrict the label space to $\mathcal{Y} = \{+1, -1\}$. Let $\mathcal{H}$ be a hypothesis class over the input space $\mathcal{X}$ with labels in $\{+1, -1\}$, where we measure its complexity by the VC-dimension, denoted as $\mathrm{VC}(\mathcal{H}) = d_{\mathrm{VC}}$ when $\mathcal{H}$ is clear from the context. Under this setting, the objective is to minimize the 0-1 loss. The corresponding minimax excess risk of the given hypothesis class under privacy budgets (ε, δ) is defined as:

$$R(\mathcal{H}, \varepsilon, \delta) = \inf_{\mathcal{A}} \sup_{\mathcal{P}} \left(\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{P}} \left[\hat{h}(\mathbf{x}) \neq y \,\middle|\, \hat{h} \leftarrow \mathcal{A}(D) \right] - \min_{h^* \in \mathcal{H}} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{P}} [h^*(\mathbf{x}) \neq y] \right),$$

where $D \sim \mathcal{P}^n$ and $\mathcal{A}$ is any (ε, δ)-label DP algorithm. We also use $R(\mathcal{H}, \varepsilon)$ to characterize the minimax risk for pure-DP algorithms.

Task II: Private Linear Regression. For the linear regression task, we consider a continuous label space $\mathcal{Y} \subseteq \mathbb{R}$ and a feature space $\mathcal{X} \subseteq \mathbb{R}^d$ for any integer $d \in \mathbb{N}_+$. Since the label space is unbounded, we follow the standard convention for differential privacy on real-valued data [18, 45]: two datasets $D = (\mathbf{X}, \mathbf{y})$ and $D' = (\mathbf{X}, \mathbf{y}')$ are considered neighbors if they share the same public features $\mathbf{X} \in \mathbb{R}^{n \times d}$ but their private label vectors satisfy that $\|\mathbf{y} - \mathbf{y}'\|_1 \le 1$. The hypothesis class is then given by the set of linear predictors $\mathcal{H} = \{\mathbf{x} \mapsto \theta^\top \mathbf{x} \mid \theta \in \mathbb{R}^d\}$.

For a distribution $\mathcal{P}$ on $\mathcal{X} \times \mathcal{Y}$, let $\mathcal{P}_\mathcal{X}$ be the marginal distribution of the feature vector $\mathbf{x} \in \mathcal{X}$. For the regression task we assume that $\mathcal{P}_\mathcal{X}$ is centered[3] (i.e., $\mathbb{E}_{\mathcal{P}_\mathcal{X}}[\mathbf{x}] = 0$). We then define the covariance matrix to be $\Sigma_x = \mathbb{E}_{\mathcal{P}_\mathcal{X}}[\mathbf{x}\mathbf{x}^\top]$. Recall that for any $\mathbf{v} \in \mathbb{R}^d$ we write $\|\mathbf{v}\|_{\Sigma_x}^2 = \mathbf{v}^\top \Sigma_x \mathbf{v}$. For the utility, given an algorithm $\mathcal{A} : \mathbb{R}^{n \times d} \times \mathbb{R}^n \to \mathbb{R}^d$ on selecting some $\theta \in \mathcal{H}$, we consider the *expected prediction risk* [10] to measure the error of $\mathcal{A}$ on input features $\mathbf{X} \in \mathbb{R}^{n \times d}$ and labels $\mathbf{y} \in \mathbb{R}^n$, defined as

$$R^{\mathrm{EPR}}(\mathbf{X}, \mathbf{y}, \mathcal{A}) = \mathop{\mathbb{E}}_{\theta \leftarrow \mathcal{A}(\mathbf{X}, \mathbf{y})} \left[\|\theta - \theta^*\|_{\Sigma_x}^2 \right] = \mathop{\mathbb{E}}_{\theta \leftarrow \mathcal{A}(\mathbf{X}, \mathbf{y})} \left[(\theta - \theta^*)^\top \Sigma_x (\theta - \theta^*) \right], \tag{1}$$

where θ^* is the underlying ground truth. For the sake of simplicity, when $\mathbf{X}, \mathbf{y}$ and $\mathcal{A}$ are clear from the context and that $\hat{\theta}$ is the random vector given by $\mathcal{A}(\mathbf{X}, \mathbf{y})$, we also write $R(\hat{\theta}) = R^{\mathrm{EPR}}(\mathbf{X}, \mathbf{y}, \mathcal{A})$.

[3] Note that this does not compromise the generality as the features are public in our label-DP setting and a non-zero mean can be absorbed into the bias term.

1.2 Our Contribution

In this section, we summarize our main theoretical results, which include several lower bounds and computationally efficient algorithms for private learning with label differential privacy. We consider both the local and central models, and our findings are also summarized in Table 1. Throughout this paper, $\widetilde{\Omega}(\cdot)$ and $\widetilde{O}(\cdot)$ hide factors polylogarithmic in n, and $\widetilde{\Omega}_\delta(\cdot)$ and $\widetilde{O}_\delta(\cdot)$ hide factors polylogarithmic in n and $\frac{1}{\delta}$.

Table 1: Summary of Main Results and Comparison with Prior Works. Note that $n \gg d$.

Task	Privacy Model	Privacy Penalty	Notes	Ref
Classification	Local ε-DP (Label)	$\Omega(\frac{1}{\varepsilon}\sqrt{\frac{d_{\text{VC}}}{n}})$	–	Zhao et al. [42]
	Local (ε,δ)-DP (Label)	$\Omega(\frac{1}{\varepsilon+n\delta}\sqrt{\frac{d_{\text{VC}}}{n}})$	$\delta = o(1/n)$	**Ours (Thm 3)**
	Central ε-DP (Label)	$\tilde{O}(\frac{1}{\sqrt{\varepsilon}}\sqrt{\frac{d_{\text{VC}}}{n}})$	Inefficient	Beimel et al. [7]
	Central ε-DP (Label)	$\tilde{O}(\frac{1}{\sqrt{\varepsilon}}\sqrt{\frac{d_{\text{VC}}}{n}})$	**Efficient**	**Ours (Thm 4)**
Regression	Central (ε,δ)-DP (Full)	$\tilde{O}_\delta(\kappa(\Sigma_x)\cdot\frac{d^2}{n^2\varepsilon^2})$	–	Cai et al. [10]
	Local (ε,δ)-DP (Label)	$\tilde{O}_\delta(\frac{d}{n\varepsilon^2})$	**Output Perturbation**	**Ours (Thm 50)**
	Central (ε,δ)-DP (Label)	$\widetilde{\Omega}_\delta(\frac{d^2}{n^2\varepsilon^2})$	**Isotropic Lower Bound**	**Ours (Thm 5)**
	Central (ε,δ)-DP (Label)	$\tilde{O}_\delta(\frac{d^2}{n^2\varepsilon^2})$	**Matrix Mechanism**	**Ours (Thm 6)**

1.2.1 On Optimal Rates for Private Classification

There are a number of gaps in prior work in binary classification among the four settings that come up from pure vs approx DP in local vs central settings. In the local setting, Zhao et al. [42] proved an $\Omega(\sqrt{\frac{d_{\text{VC}}}{\varepsilon^2 n}})$ lower bound for ε-locally label private learning when the hypothesis class $\mathcal{H}$ has VC dimension d_{VC} and the size of the dataset is n, showing that a simple randomized response algorithm is essentially optimal in this setting. However, it is not clear if this lower bound extends to the approximate DP setting, or algorithms with better bounds exist. In the central model, the best known *computationally inefficient* algorithm has an error bound of $\tilde{O}(\sqrt{d_{\text{VC}}/(\varepsilon n)})$ [7] and the best known computationally efficient algorithm achieves an error bound of $\tilde{O}(d_{\text{VC}}^{3/5}/(\varepsilon n)^{2/5})$ [6] (i.e., has a suboptimal dependence on the VC dimension, d_{VC} and sample size n). These results show a clear gap between central and local models in the *pure*-DP settings. However it is not clear if this gap shrinks when we take into account the computational complexity of the algorithms.

Together, these results show that the requirement of local privacy alone can incur a substantial penalty on statistical efficiency in classification tasks. However, it remains unclear whether 1) this gap is an artifact of requiring pure privacy, or whether it persists under the more permissive notion of approximate local DP, and 2) whether the *efficiency* of the algorithms in central DP has a role to play in this gap. Our first result is a minimax lower bound for approximate local label DP. We show in Appendix D.2 that relaxing the privacy requirement from pure DP to approximate DP ($\delta > 0$) *does not* eliminate the statistical inefficiency of the local model.

Theorem 3. (Lower bound for approximate local label DP). *Let $\mathcal{H}$ be a hypothesis class over an input space $\mathcal{X}$ with finite VC dimension $VC(\mathcal{H}) = d_{VC}$. Consider a binary classification task on $D = \{(\mathbf{x}_i, y_i)\}_{i\in[n]}$ where a learner receives n private observations $Z_1, \ldots, Z_n$ generated via an (ε,δ)-local differentially private mechanism Q that acts independently on each $(\mathbf{x}_i, y_i)$ ($i \in [n]$). Assume $\varepsilon < 1$, and that the privacy parameter δ is sufficiently small such that*

$$\tilde{\delta}\ln\left(\frac{M}{n\tilde{\delta}}\right) = O\left(\frac{d_{VC}}{n}\right) \quad \textit{where} \quad M = e^{\Theta(d_{VC})} \quad \textit{and} \quad \tilde{\delta} = \frac{\delta}{1-e^{-\varepsilon}}. \tag{2}$$

Then, the minimax excess risk is lower bounded by

$$R(\mathcal{H}, \varepsilon, \delta) = \Omega\left(\frac{1}{\varepsilon'}\sqrt{\frac{d_{VC}}{n}}\right), \quad \textit{where} \quad \varepsilon' = 3\varepsilon + \ln\left(\frac{1}{1-2n\tilde{\delta}}\right).$$

To get some intuition behind the lower bound, let us consider a simple example. Note that, for constant ε, eq. (2) is satisfied for the natural setting of $\delta \leq c/n$ for some constant c; this is because $\tilde{\delta}\ln\left(\frac{M}{n\tilde{\delta}}\right) = O\left(\frac{e^{\varepsilon} d_{\text{VC}}}{n(e^{\varepsilon}-1)}\right) = O\left(\frac{d_{\text{VC}}}{n}\right)$. This gives a lower bound of $\Omega(\frac{1}{\varepsilon+n\delta}\sqrt{\frac{d_{\text{VC}}}{n}})$.

To contrast this lower bound, we give an *efficient*[4] algorithm in the $(\varepsilon, 0)$-label DP in the central model with the same error bound of Beimel et al [7]. The proof of the following result is in Appendix E.

Theorem 4. *Let $\mathcal{H}$ be a hypothesis class over an input space $\mathcal{X}$ with finite VC dimension VC$(\mathcal{H}) = d_{VC}$. Consider an agnostic binary classification task with a data distribution $\mathcal{P}$ over $\mathcal{X} \times \{+1, -1\}$. For any $\varepsilon \in (10/n, 1]$, there exists a polynomial-time ε-central label differentially private algorithm (Algorithm 1) that, given n i.i.d. samples from $\mathcal{P}$, outputs a hypothesis $\hat{h} \in \mathcal{H}$ such that with a constant probability, the excess risk satisfies:*

$$R(\hat{h}) - \min_{h\in\mathcal{H}} R(h) = \tilde{O}\left(\sqrt{\frac{d_{VC}}{n\varepsilon}}\right), \quad \textit{where } R(h) = \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{P}}[h(\mathbf{x}) \neq y].$$

Theorem 3 and Theorem 4 thus imply a gap of $\sqrt{\frac{1}{\varepsilon}}$ factor between central and local label-DP.

1.2.2 On Efficient Central Mechanisms for Private Regression

We extend our study to linear regression to further show (i) the limitations of the local model; and (ii) the statistical advantages of label-DP over standard full-DP in the central model.

Prior work on linear regression in the central setting under the full-DP model has shown that optimal statistical rates are achievable only under strong assumptions, such as Gaussian or isotropic designs; otherwise, the privacy noise is inevitably amplified by the condition number of the feature covariance matrix [10]. In particular, Cai et al. [10] have shown that linear regression under Gaussian data with full-DP necessarily incurs an expected error of $\tilde{\Omega}_{\delta}(\kappa(\Sigma_x)\cdot\frac{d^2}{\varepsilon^2 n^2})$ on a d-dimensional dataset of size n, where $\kappa(\Sigma_x) = \frac{\lambda_{\max}}{\lambda_{\min}}$ denotes the condition number of the covariance matrix of the feature distribution ($\mathbb{E}[\mathbf{x}\mathbf{x}^\top] \in \mathbb{R}^{d\times d}$). Here $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues of Σ_x.

In this paper, for the distribution of the private labels, we consider the Gaussian linear model [10] where, for all $i \in [n]$ the private label $y_i \in \mathbb{R}$ is generated by $y_i = \mathbf{x}_i^\top\theta^* + \zeta_i$. Here, $\mathbf{x}_i \in \mathbb{R}^d$ are feature vectors, $\theta^* \in \mathbb{R}^d$ is the unknown parameter and $\zeta_i \sim \mathcal{N}(0, \sigma_\zeta^2)$.

As a warm-up to study regression under label DP, we first consider the local setting and restrict ourselves to a non-interactive setting[5]. We give a non-interactive private regression algorithm that ensures (ε, δ)-label local DP while incurring an error of $\tilde{O}_{\delta}(\frac{d}{n\varepsilon^2})$ (see Theorem 50 for formal result).

Moving to the central-label DP model, we show in Appendix F.2 that when only labels are private, we obtain qualitatively better results than previous results. In particular, for the regression task, we assume features are sampled from the more general sub-Gaussian distributions under a *linear sub-Gaussian design*[6] (which even allows for *non-isotropic* data with arbitrary covariance). We first show that the bound of Cai et al. [10] can be achieved in the central model under the label-DP by objective perturbation (i.e., mechanism that adds noise to the least squares estimate). That is, the accuracy loss due to objective perturbation is $\tilde{O}(\kappa(\Sigma_x)\cdot\frac{d^2}{\varepsilon^2 n^2})$ when $n = \Omega(d)$. However, this raises the question of whether we can achieve better accuracy w.r.t. any of these parameters.

[4]Specifically, our algorithm is oracle-efficient, assuming access to an efficient non-private oracle for ERM (empirical risk minimization), same as the setting considered by Esfandiari et al. [19].

[5]This is motivated by works discussing why interactive mechanisms are not that practical [24, 33].

[6]See Assumption 1 for details. Informally, this assumes that the observed features, which may be highly correlated, are generated by applying a linear transformation to a set of underlying, mutually independent sub-Gaussian components. Note that the standard multivariate Gaussian distribution automatically satisfies this assumption, as its uncorrelated components are naturally mutually independent.

We first show that for isotropic subgaussian data where $\Sigma_x = I_d$, the dependence on d, n, and ε is unavoidable even under label-DP for approximating the least squares estimate in linear regression (see Appendix F.3 for the proof).

Theorem 5. (Optimality of Output Perturbation for Isotropic sub-Gaussian). *Assume the public feature matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ has rows drawn i.i.d. from any zero-mean, isotropic sub-Gaussian distribution, where $n \gg d$. Let $\mathbf{y} \in \mathbb{R}^n$ be the private labels. For any fixed $\varepsilon \in (0,1)$ and sufficiently small $\delta > 0$, let $\mathcal{M}$ be any (ε,δ)-label DP mechanism outputting an estimate $\hat{\theta}$. Suppose $\max_{i\in[n]} \|\mathbf{x}_i\|_{\psi_2} \le K$. Then, with probability at least $1 - 2\exp(-cn/K^4)$ over the randomness of $\mathbf{X}$ for some constant $c > 0$, the expected privacy error (over the randomness of the mechanism $\mathcal{M}$) is lower bounded by:*

$$\max_{\mathbf{y}\in\mathbb{R}} \mathop{\mathbb{E}}_{\mathcal{M}} \left[\|\mathcal{M}(\mathbf{y}) - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}\|_2^2\right] = \widetilde{\Omega}\left(\frac{d^2}{n^2\varepsilon^2}\right),$$

Note that $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ is the ordinary least squares estimate and the expected risk reduces to the expected ℓ_2 error for linear regression, consistent with the choice of error metric in Theorem 5. Further, $\mathbf{X}^\top\mathbf{X}$ is full-rank with high probability [34], so $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ is well-defined.

In view of these bounds, a natural question arises *whether we need to pay the price of privacy in terms of the condition number*? This is especially important because many real-world datasets are ill-conditioned, where these algorithms perform poorly (see also our empirical simulations in Section 2). To overcome this barrier when handling real-world data that is not normally isotropic, we propose in Section F.4 an algorithm based on the *matrix mechanism* [18, 26]. Our algorithm decouples the dependency between privacy and the condition number of Σ_x, achieving the optimal privacy penalty of $\tilde{O}(\frac{d^2}{n^2\varepsilon^2})$, while keeping a separate bias term that vanishes exponentially in n and only depends on the condition number but not the privacy. Since Σ_x can represent a lot of correlation, we work with the ridge-regression with a regularizer parameter $\lambda > 0$. We show the following in Appendix F.4:

Theorem 6 (Restatement of Theorem 53)**.** *For any privacy parameters $\varepsilon, \delta \in (0,1)$, suppose we are given n samples where the public features are drawn from a sub-Gaussian distribution (under a linear sub-Gaussian design, Assumption 1) with covariance matrix Σ_x, and the private labels are generated by an unknown underlying parameter. Let σ_ζ^2 be the variance of the observation noise. Define $M = \frac{\sigma_\zeta^2 d}{n}$ and $\Lambda = \frac{\lambda_{\max}^2}{\lambda_{\min}}$. There is an efficient (ε,δ)-central label DP algorithm (Algorithm F.4) with output $\hat{\theta}_{mm}$ and constant $c > 0$ such that, for sufficiently large n the expected prediction risk is*

$$R(\hat{\theta}_{mm}) = O\left(\frac{d\sigma_\zeta^2}{n}\right) + \tilde{O}_\delta\left(\frac{d^2}{n^2\varepsilon^2}\right) + e^{-cn/K^4}\left[3\left(M^2\Lambda\right)^{1/3} + \Lambda\right]$$

The first term $O\left(\frac{d\sigma_\zeta^2}{n}\right)$ is the *statistical accuracy loss*, the second term $\tilde{O}_\delta\left(\frac{d^2}{n^2\varepsilon^2}\right)$ is the price of privacy, while the last term is the *bias term*. Note that, the price of privacy in our algorithm does not suffer from dependency on the condition number and overall dependence on condition number is mild. In particular, if $\max\left\{M^{2/3}\lambda_{\max}^{1/3}\kappa(\Sigma_x)^{1/3}, \lambda_{\max}\kappa(\Sigma_x)\right\} \lesssim d\sigma_\zeta^2 e^{\Theta(n/K^4)}$, then the bias term is subsumed by the statistical error and the accuracy loss is $O\left(\frac{d\sigma_\zeta^2}{n}\right) + \tilde{O}_\delta\left(\frac{d^2}{n^2\varepsilon^2}\right)$. Under natural assumption that entries of Σ_x is polynomial in n, our bound is optimal when condition number is bounded by exponential in n. We leave it as an open problem whether we can remove this assumption.

Empirical Validation. We also complement the above theoretical findings with thorough empirical evaluations in Section 2. For classification, simulations visually confirm the sharp gap in sample complexity between local and central privacy. For regression, we highlight a critical phenomenon: while standard output perturbation fails on ill-conditioned datasets, our scheme based on the Matrix Mechanism maintains stable utility despite severe multicollinearity in real-world datasets. This validates that the independence from the condition number we theoretically established is not just an asymptotic artifact, but a vital necessity for deploying label DP in practical, non-idealized scenarios.

1.3 Technical overview

In this section, we highlight the key technical ingredients and analytical tools that underpin our upper and lower bounds across local and central label DP models.

Contraction Bounds and Hardness for Approximate Local DP (Theorem 3) To prove Theorem 3, we use a generalization of the *data processing inequalities* for local differential privacy introduced by Duchi et al. [15] from the pure DP setting to the approximate DP regime; we believe this may be of independent interest. We show that an (ε, δ)-approximate local DP mechanism, when viewed as an information-theoretic communication channel[7], induces an $O(\varepsilon^2 + n\delta)$ contraction in the mutual information between the underlying parameter distribution and the private outputs. For $\delta = o(n^{-1})$, by integrating this new contraction bound with Fano's inequality [40], we reduce the learning problem to a hypothesis testing problem and lower bound the minimax excess risk for binary classification in the *approximate* local label DP model by $\Omega(\frac{1}{\varepsilon}\sqrt{\frac{d_{\text{VC}}}{n}})$ for classification on hypothesis set $\mathcal{H}$ with $\text{VC}(\mathcal{H}) = d_{\text{VC}}$ and datasets of size n (see Section D.2).

Privacy Amplification for Efficient Central Classification (Theorem 4) To prove Theorem 4, we design a computationally efficient central-model algorithm (Algorithm 1) that achieves an optimal $\tilde{O}(1/\sqrt{\varepsilon})$ dependence. Our algorithm combines privacy amplification by subsampling [4] and PAC learning with noisy labels [19] to achieve this bound: instead of applying randomized response to the entire dataset with a small privacy budget ε, we independently subsample the dataset with probability ε and apply randomized response (with $\varepsilon_0 = 1$) only to the subsampled datapoints. The central curator manages to hide the uncertainty of user participation so that the privacy guarantee improves to $O(\varepsilon)$-central label DP, while the effective sample size is reduced to $n' = \Theta(n\varepsilon)$. Recall that applying randomized response on n'-size datasets with $O(1)$-privacy budgets gives an accuracy of $\tilde{O}(\sqrt{d_{\text{VC}}/(n')}) = \tilde{O}(\sqrt{d_{\text{VC}}/(n\varepsilon)})$ [19]. This implies a $1/\sqrt{\varepsilon}$ factor difference in the accuracy (i.e., excess risk) between central and local models.

Optimality and the Condition Number Barrier in Regression (Theorem 5) Recall that for the linear regression task, in Theorem 5, we analyze a trivial output perturbation algorithm and formally showed that its privacy penalty of $O(d^2/n^2\varepsilon^2)$ is asymptotically tight up to logarithmic factors for least squares estimation when the feature matrix is sampled from an *isotropic* sub-Gaussian distribution. The proof reduces the private estimation of the least squares weights to answering a set of linear queries defined by the workload matrix $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. By leveraging recent hardness results for linear queries [45], we establish a lower bound based on the Schatten-1 norm, $\|\mathbf{A}\|_1$. Using random matrix theory, we tightly bound the singular values of the sub-Gaussian matrix to explicitly compute $\|\mathbf{A}\|_1 = \Omega(d/\sqrt{n})$, verifying that in terms of the privacy error, simply using output perturbation is already theoretically optimal for isotropic designs.

Bypassing Ill-Conditioning via the Matrix Mechanism (Theorem 6) To establish Theorem 6, we factorize the regularized least squares operator $\mathbf{M} = (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top$ into two factors such that $\mathbf{M} = \mathbf{L}\mathbf{R}$. Specifically, we choose the encoder as $\mathbf{R} = (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}\mathbf{X}^\top$ and the corresponding decoder as $\mathbf{L} = (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}$. This factorization enables a two-step privatization process. First, the encoder $\mathbf{R}$ projects the raw sub-Gaussian features into a well-conditioned, spherical latent space. Because the feature matrix $\mathbf{X}$ is entirely public in the label-DP setting, we can compute the exact ℓ_2-sensitivity of the encoded labels $\mathbf{R}\mathbf{y}$ (given by the operator norm $\|\mathbf{R}\|_{1\to 2}$), and then explicitly calibrate the Gaussian noise without consuming any privacy budget. We then inject this calibrated noise directly into the latent space, and finally use the decoder $\mathbf{L}$ to project the noisy latent representation back to the original parameter space.

Note that in the naive output perturbation scheme, the sensitivity is dictated by the full inverse $(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}$, which severely penalizes ill-conditioned data. By contrast, our square-root factorization ties the sensitivity strictly to $(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}$. We give some intuition behind why matrix mechanism allows us to bypass the ill-conditioning. Intuitively, ill-conditioning means that the matrix is close to being singular. In that case, if we remove all the directions in which the singular values are small, then the restricted matrix reduces to the well-conditioned case covered earlier. At a high level, the

[7] See Duchi et al. [15] for more details on information theoretic communication channel.

matrix mechanism does it automatically. We formally prove that the spectral dampening by λI_d cleanly decouples the required noise scale from the condition number $\kappa(\Sigma_x)$, allowing our algorithm to maintain optimal utility even on highly skewed, real-world datasets. To the best of our knowledge, this work is the first to use matrix mechanism in label DP model to exploit the structure in the public feature matrix and it would be interesting to see if there are other applications of matrix mechanism.

2 Simulations and Experiments

While we focus on rigorous theoretical optimal rates for classification and regression under label-DP, we also conduct experiments on both synthetic and real-world datasets. Our empirical findings confirm not only the fundamental statistical gap between local and central model in label-DP for both classification (Theorem 3 & Theorem 4) and regression (Theorem 50 & Theorem 52), but also highlight the benefit of removing the dependency on $\kappa(\Sigma_x)$ in Matrix Mechanism based algorithm (Theorem 6), especially for ill-conditioned or real-world datasets. Our synthetic data are generated following the technical setups in our theorems, and our real-world datasets include California Housing dataset and the Communities and Crime dataset from UCI machine learning repository [2].

Due to space constraints, we defer our empirical evaluation for classification to Appendix G.

Regression on Standard Gaussian Data. To evaluate the convergence rates of the label differentially private regression algorithms, we first simulate data following a standard Gaussian linear model where $\Sigma_x = I_d$: (i) the feature vectors are drawn independently from $\mathcal{N}(0, I_d)$, and (ii) the labels are generated using a true parameter unit vector with additive Gaussian noise from $\mathcal{N}(0, 1)$.

For this experiment, we fix the feature dimension $d = 100$, the privacy budget $\varepsilon = 0.5$ while exponentially scaling the sample size n from 200 up to 10,000. We test the *ordinary least squares* (OLS) estimator for the non-private baseline, the basic local-DP approach (Input Perturbation, Algorithm F.1), and our two central-DP mechanisms (Output Perturbation, Algorithm F.2 and Matrix Mechanism, Algorithm F.4). To provide stronger state-of-the-art local DP baselines, we also include the optimal unbiased randomizer via linear programming [3], which feeds its unbiased discretized noisy labels into an OLS estimator, and the label-LDP-IHT algorithm by Wang and Xu [37]. The performance of each algorithm is measured using the Expected Prediction Risk introduced in Section 1.1. We visualize these risk values against the varying sample sizes using a log-log plot. The results are shown in Figure 1.

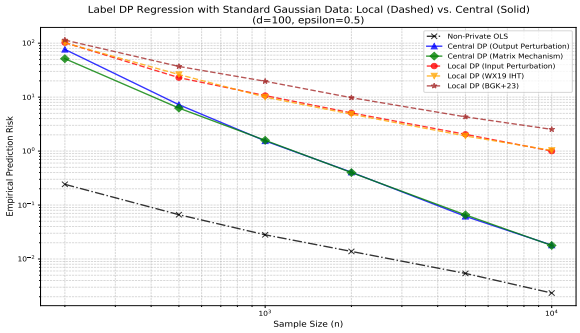

Figure 1: Empirical regression error with different data scale on standard Gaussian data.

The resulting trajectories provide an empirical validation of the theoretical bounds and the fundamental gap between local and central privacy models. As the data scale increases, the error of all local-DP variants including the advanced optimal unbiased LP [3] decreases slowly[8] and is bottlenecked by the $O(1/n)$ privacy cost shown in Theorem 50. On the other hand, both our central DP algorithms (Output Perturbation and Matrix Mechanism based) demonstrate a tighter $O(1/n^2)$ privacy cost as shown in Theorem 52 and Theorem 53, respectively.

No Advantage for Matrix Mechanism on Standard Gaussian Data. We note that, the output perturbation (Algorithm F.2) and the matrix mechanism (Algorithm F.4) perform almost identically in

[8]We note that, the unbiased randomizer in Badanidiyuru et al. [3] is optimal on discrete domains. To apply to continuous Gaussian data and align with our theoretical analysis, we have to discretize it first, which may hinder its empirical performance.

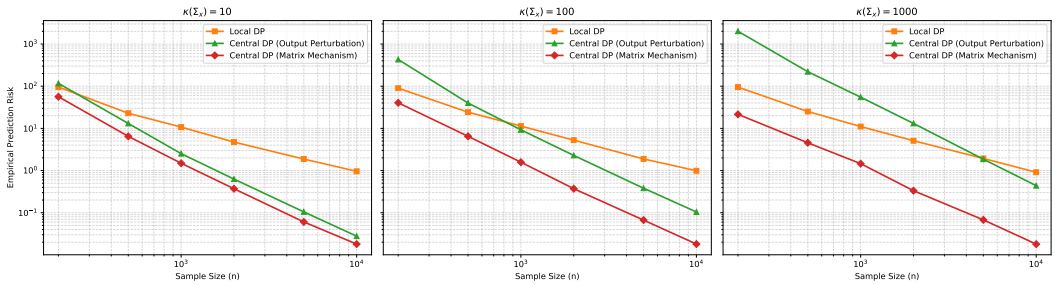

Figure 2: Empirical regression error with different data scale on ill-conditioned Gaussian data with varying condition numbers.

this experiment, which makes sense as our experimental data is already perfectly well-conditioned ($\kappa(\Sigma_x) = 1$). In other words, output perturbation suffers no baseline penalty, leaving the Matrix Mechanism's sophisticated structural advantages with nothing to correct. This aligns with Theorem 5, which states that the Output Perturbation (Algorithm F.2) is already optimal on isotropic data.

Regression on Ill-Conditioned Gaussian Data To show the strength of our Matrix Mechanism based algorithm, we evaluate how these algorithms handle realistic, skewed data distributions. These datasets may not always have well-conditioned data.

To isolate the impact of ill-conditioned data, we maintain the same baseline parameters ($d = 100$, $\varepsilon = 0.5$, and sample sizes n scaling up to 10,000) but introduce a structured covariance matrix Σ_x with varying condition numbers $\kappa(\Sigma_x) \in \{10, 100, 1000\}$. For the error metric we calculate the quadratic form $(\hat{\theta} - \theta^*)^\top \Sigma_x (\hat{\theta} - \theta^*)$ as defined in eq. (1), Section 1.1. The results shown in Figure 2 provide a strong empirical validation of the theoretical differences between the two central-DP mechanisms. As the condition number $\kappa(\Sigma_x)$ increases, the Output Perturbation algorithm rapidly deteriorates, forced to inject massive amounts of noise due to its direct dependence on the condition number. In contrast, the Matrix Mechanism remains highly robust even in extremely ill-conditioned scenarios, validating our Theorem 6.

Regression on Real-World Datasets Real-world data frequently exhibits strong multicollinearity and vastly different feature scales, leading to inherently ill-conditioned covariance matrices. Here, we evaluate the three different privacy mechanisms on two distinct real-world datasets: the *California Housing dataset* and the *Communities and Crime dataset* from UCI machine learning repository [2]. These datasets shows contrasting condition numbers in their empirical covariance matrices. Specifically, California Housing possesses a moderate condition number and Communities and Crime features severe multicollinearity among its many socio-economic indicators.

We apply a standard 80/20 train-test split to both datasets. The privacy budget is fixed at $\varepsilon = 0.5$ and $\delta = 10^{-5}$. We compute the empirical covariance matrix on the standardized training features as $\hat{\Sigma} = \frac{1}{n} X_{\text{train}}^\top X_{\text{train}}$, and calculate its condition number $\kappa(\hat{\Sigma})$ as the ratio of its maximum to minimum eigenvalues. Algorithm performance is then measured using the Mean Squared Error (MSE) on the held-out test set, providing a practical analog to the expected prediction risk.

The empirical results summarized in Table 2, confirm our theoretical findings. On the California Housing dataset, where the condition number is a moderate 4.06×10^1, the Matrix Mechanism (Algorithm F.4) already demonstrates a clear advantage over Output Perturbation (Algorithm F.2). Further, the Communities and Crime dataset exposes the fatal flaw of Output Perturbation: faced with an extreme condition number of $\kappa \approx 1.39 \times 10^{11}$, Output Perturbation is forced to inject a large amount of noise, rendering the estimator completely useless. In contrast, the Matrix Mechanism based algorithm successfully navigates this severe multicollinearity as predicted by Theorem 53, maintaining a highly stable test MSE and proving its robust utility in realistic, ill-conditioned scenarios.

3 Conclusion and Limitations

In this work, we rigorously quantified the "price of locality" in label-DP for fundamental classification and regression tasks. We propose algorithms that efficiently achieve optimal statistical rates, and show that unlike central full DP, central label DP inherently *decouples* the privacy penalty from the condition number of the feature covariance matrix. For limitation and future works, our algorithms

Table 2: Test MSE of Privacy Mechanisms on Real-World Datasets

Dataset	n	d	$\kappa(\hat{\Sigma})$	**Algorithm**	**Test MSE**
California Housing (*well-conditioned*)	2000	8	4.06×10^1	Non-Private Baseline	0.1530
				Local DP	2.1455
				Output Perturbation	1.8371
				Matrix Mechanism	**1.2858**
Communities & Crime (*ill-conditioned*)	937	49	1.39×10^{11}	Non-Private Baseline	0.9643
				Local DP	23.0947
				Output Perturbation	4.697×10^7
				Matrix Mechanism	**15.3181**

are designed strictly for either the local or central DP models. An interesting direction is adapting our techniques (such as the Matrix Mechanism) to intermediate trust models, like *shuffle DP* or *secure multi-party computation* (MPC) and study whether these intermediate models can offer the optimal utility of the central model alongside the decentralized security of the local model.

References

[1] Shahab Asoodeh, Maryam Aliakbarpour, and Flavio P. Calmon. Local differential privacy is equivalent to contraction of E_γ-divergence, 2021.

[2] Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

[3] Ashwinkumar Badanidiyuru Varadaraja, Badih Ghazi, Pritish Kamath, Ravi Kumar, Ethan Leeman, Pasin Manurangsi, Avinash V Varadarajan, and Chiyuan Zhang. Optimal unbiased randomizers for regression with label differential privacy. *Advances in Neural Information Processing Systems*, 36:60226–60246, 2023.

[4] Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 6280–6290, 2018.

[5] Raef Bassily, Mehryar Mohri, and Ananda Theertha Suresh. Differentially private learning with margin guarantees. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[6] Raef Bassily, Abhradeep Guha Thakurta, and Om Dipakbhai Thakkar. Model-agnostic private learning. *Advances in Neural Information Processing Systems*, 2018.

[7] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, pages 363–378. Springer, 2013.

[8] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. *Theory of Computing*, 12(1):1–61, 2016.

[9] Adrian Bussone, Bakita Kasadha, Simone Stumpf, Abigail C Durrant, Shema Tariq, Jo Gibbs, Karen C Lloyd, and Jon Bird. Trust, identity, privacy, and security considerations for designing a peer data sharing platform between people living with hiv. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–27, 2020.

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

[11] Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[12] Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 155–186. JMLR Workshop and Conference Proceedings, 2011.

[13] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

[14] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

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

[16] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *TCC*, pages 265–284, 2006.

[17] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[18] Alexander Edmonds, Aleksandar Nikolov, and Jonathan Ullman. The power of factorization mechanisms in local and central differential privacy. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 425–438, 2020.

[19] Hossein Esfandiari, Vahab Mirrokni, Umar Syed, and Sergei Vassilvitskii. Label differential privacy via clustering. In *International Conference on Artificial Intelligence and Statistics*, pages 7055–7075. PMLR, 2022.

[20] Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy. *Advances in neural information processing systems*, 34:27131–27145, 2021.

[21] Badih Ghazi, Pritish Kamath, Ravi Kumar, Ethan Leeman, Pasin Manurangsi, Avinash V. Varadarajan, and Chiyuan Zhang. Regression with label differential privacy. In *International Conference on Learning Representations*, 2023.

[22] Edgar N Gilbert. A comparison of signalling alphabets. *The Bell system technical journal*, 31(3):504–522, 1952.

[23] Robert M Gray. *Entropy and information theory*. Springer Science & Business Media, 2011.

[24] Matthew Joseph, Jieming Mao, Seth Neel, and Aaron Roth. The role of interactivity in local differential privacy. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 94–105. IEEE, 2019.

[25] Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pages 110–133, 2017.

[26] Chao Li, Gerome Miklau, Michael Hay, Andrew McGregor, and Vibhor Rastogi. The matrix mechanism: optimizing linear counting queries under differential privacy. *The VLDB journal*, 24(6):757–781, 2015.

[27] Mani Malek Esmaeili, Ilya Mironov, Karthik Prasad, Igor Shilov, and Florian Tramer. Antipodes of label differential privacy: Pate and alibi. *Advances in neural information processing systems*, 34:6934–6945, 2021.

[28] KV Mardia, JT Kent, and JM Bibby. Multivariate analysis academic press inc. *London Ltd*, 15:518, 1979.

[29] Jiri Matousek. *Geometric discrepancy: An illustrated guide*, volume 18. Springer Science & Business Media, 1999.

[30] Frank McSherry and Ilya 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. ACM, 2009.

[31] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[32] Jack Sherman and Winifred J Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, 1950.

[33] Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. Is interaction necessary for distributed private learning? In *IEEE Symposium on Security and Privacy*, 2017.

[34] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[35] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

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

[37] Di Wang and Jinhui Xu. On sparse linear regression in the local differential privacy model. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6628–6637. PMLR, 2019.

[38] Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *J. of the American Statistical Association*, 60(309):63–69, 1965.

[39] Shirong Xu, Chendi Wang, Will Wei Sun, and Guang Cheng. Binary classification under local label differential privacy using randomized response mechanisms. *Transactions on Machine Learning Research*, 2023.

[40] Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

[41] Puning Zhao, Rongfei Fan, Huiwen Wu, Qingming Li, Jiafei Wu, and Zhe Liu. Enhancing learning with label differential privacy by vector approximation. *arXiv preprint arXiv:2405.15150*, 2024.

[42] Puning Zhao, Chuan Ma, Li Shen, Shaowei Wang, and Rongfei Fan. On theoretical limits of learning with label differential privacy. *arXiv preprint arXiv:2502.14309*, 2025.

[43] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on mathematical software (TOMS)*, 23(4):550–560, 1997.

[44] Liyang Zhu, Meng Ding, Vaneet Aggarwal, Jinhui Xu, and Di Wang. Improved analysis of sparse linear regression in local differential privacy model. *arXiv preprint arXiv:2310.07367*, 2023.

[45] Zongrui Zou, Jingcheng Liu, and Jalaj Upadhyay. Optimality of matrix mechanism on ell_p^p-metric. In *13th International Conference on Learning Representations, ICLR 2025*, pages 34627–34653, 2025.

A Preliminaries

A.1 Differential Privacy

In this section, we introduce some basic technical definitions and lemmas for designing differentially private algorithms.

Lemma 7 (Post processing [16]). *Let $\mathcal{M} : \mathcal{X} \to \mathcal{O}$ be an (ε, δ)-differentially private algorithm. Let $f : \mathcal{O} \to \mathcal{O}'$ be an arbitrary randomized mapping, then $f \circ \mathcal{M} : \mathcal{X} \to \mathcal{O}'$ is also (ε, δ)-differentially private.*

Lemma 8 (Basic composition [16]). *Let $D \in \mathcal{X}$ be a dataset and $\mathcal{M}_1, \cdots, \mathcal{M}_k$ be k algorithms such that $\mathcal{M}_i : \mathcal{X} \to \mathcal{O}_i$ preserves $(\varepsilon_i, \delta_i)$-DP. Then the composed algorithm $\mathcal{M} : \mathcal{X} \to \prod_{i=1}^{k} \mathcal{O}_i$ defined as $\mathcal{M}(D) = (\mathcal{M}_1(D), \cdots, \mathcal{M}_k(D))$ preserves $(\sum_{i=1}^{k} \varepsilon_i, \sum_{i=1}^{k} \delta_i)$-DP.*

The following introduces the Laplace distribution and its standard tail bound:

Definition 9 (Laplace distribution). Given parameter b, the Laplace distribution (with scale b) is the distribution with probability density function

$$\mathrm{Lap}(x|b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right).$$

We use $\mathrm{Lap}(b)$ to denote the Laplace distribution with scale b.

Lemma 10 (Tail bound on Laplace distribution). *Let x be a random variable with $Lap(b)$ distribution. Then, $\mathbf{Pr}[|x| \geq tb] \leq \exp(-t)$.*

We also describe mechanisms that are commonly used in differential privacy. We begin by defining the sensitivity of a function.

Definition 11 (ℓ_p-sensitivity). Let $f : \mathcal{X} \to \mathbb{R}^k$ be a query function on datasets. The ℓ_p-sensitivity of f (with respect to $\mathcal{X}$) is

$$\Delta_p(f) = \max_{\substack{D, D' \in \mathcal{X}, \\ D \sim D'}} \|f(D) - f(D')\|_p.$$

Lemma 12 (Laplace mechanism [16]). *Fix any $\varepsilon > 0$. Suppose $f : \mathcal{X} \to \mathbb{R}^k$ is a query function with ℓ_1-sensitivity $\Delta_1(f)$. Then the mechanism*

$$\mathcal{M}(D) = f(D) + (Z_1, \cdots, Z_k)^\top$$

is $(\varepsilon, 0)$-differentially private, where $Z_1, \cdots, Z_k$ are i.i.d random variables drawn from $Lap(\Delta_1(f)/\varepsilon)$.

Lemma 13 (Gaussian Mechanism [17]). *Fix any $\varepsilon \in (0, 1)$ and $\delta \in (0, 1)$. Suppose $f : \mathcal{X} \to \mathbb{R}^k$ is a query function with ℓ_2-sensitivity $\Delta_2(f)$. Then the Gaussian Mechanism $\mathcal{M}$, defined by*

$$\mathcal{M}(D) = f(D) + \mathcal{N}(0, \sigma^2 \mathbf{I}_k),$$

is (ε, δ)-differentially private if the noise scale σ satisfies

$$\sigma \geq \frac{\Delta_2(f)}{\varepsilon} \sqrt{2 \ln\left(\frac{1.25}{\delta}\right)}.$$

Lemma 14 (Randomized Response [38]). *Fix any $\varepsilon > 0$. Let $y \in \{+1, -1\}$ be a private binary value. The Randomized Response mechanism Q outputs a randomized value $\tilde{y} \in \{+1, -1\}$ such that:*

$$\mathbf{Pr}[Q(y) = \tilde{y}] = \begin{cases} \frac{e^\varepsilon}{e^\varepsilon + 1} & \text{if } \tilde{y} = y, \\ \frac{1}{e^\varepsilon + 1} & \text{if } \tilde{y} \neq y. \end{cases}$$

Then the mechanism Q satisfies ε-local differential privacy.

Our proof of Theorem 4 relies on the following result on privacy amplification by subsampling:

Theorem 15 (Balle et al. [4]). *Let $\mathcal{M} : \mathcal{X}^* \to \mathcal{R}$ be an ε-differentially private algorithm, and $\mathcal{S}_\gamma : \mathcal{X}^* \to \mathcal{X}^*$ be a Poisson subsampling algorithm that retains each element of the input independently with probability γ. Write $\mathcal{M}' = \mathcal{M} \circ \mathcal{S}_\gamma$. Then, $\mathcal{M}'$ is ε'-differentially private for $\varepsilon' = \ln(1+\gamma(e^\varepsilon - 1))$.*

We also need the following definition in our lower bound proof.

Definition 16 (Privacy Loss Distribution). Let $M : \mathcal{R} \to \mathcal{Z}$ be a randomized algorithm. For any two neighboring entries r, r' and any possible output $z \in \mathcal{Z}$, let *the privacy loss* be $L_{M,r,r'}(z) = \left|\ln \frac{\mathbf{Pr}[M(r)=z]}{\mathbf{Pr}[M(r')=z]}\right|$. Let $\text{PLD}_{M,r,r'}$ denote the distribution of privacy loss in terms of $M(r)$.

Proposition 17 (Dwork and Roth [17]). *Let $M : \mathcal{R} \to \mathcal{Z}$ be a randomized algorithm. If M is (ε, δ)-differentially private, then for all neighboring entries r, r',*

$$\mathop{\mathbb{E}}_{y \sim PLD_{M,r,r'}}[\max\left\{0, 1 - e^{\varepsilon - y}\right\}] \leq \delta.$$

Using Proposition 17, we prove the following bound:

Proposition 18. *Let $M : \mathcal{R} \to \mathcal{Z}$ be a randomized algorithm. If M is (ε, δ)-differentially private, then for all neighboring entries r, r',*

$$\mathop{\mathbf{Pr}}_{y \sim PLD_{M,r,r'}}[y > 2\varepsilon] \leq \frac{\delta}{1 - e^{-\varepsilon}}.$$

Proof. Note that for $y > \varepsilon$, $1 - e^{\varepsilon - y} > 0$ and is increasing in y. Therefore, for any $t > \varepsilon$, we have that

$$\mathbb{E}[\max\left\{0, 1 - e^{\varepsilon - y}\right\}] \geq \mathbb{E}[\max\left\{0, 1 - e^{\varepsilon - y}\right\} \cdot \mathbf{1}_{\{y > t\}}] \geq \mathbf{Pr}[y > t] \cdot (1 - e^{\varepsilon - t}).$$

According to Proposition 17, we have that $\mathbf{Pr}[y > t] \cdot (1 - e^{\varepsilon - t}) \leq \delta$. Setting $t = 2\varepsilon$ yields the desired result. □

We use the known results on linear queries to prove Theorem 5.

Definition 19 (Linear Query). Fix any $n \in \mathbb{N}_+$. Given a workload vector (linear query) $\mathbf{b} \in \mathbb{R}^n$, the answer to a linear query on test vector $\mathbf{y} \in \mathbb{R}^n$ is their dot product $\mathbf{b}^\top \mathbf{y}$.

Theorem 20 (Theorem 1.3 in Zou et al. [45]). *Fix any $0 < \varepsilon, \delta < 1$. Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a matrix representing m linear queries. Then for any (ε, δ)-differentially private algorithm $\mathcal{M}$,*

$$\max_{\mathbf{x} \in \mathbb{R}^n} \mathbb{E}\left[\|\mathcal{M}(\mathbf{x}) - \mathbf{A}\mathbf{x}\|_2^2\right] = \widetilde{\Omega}\left(\frac{\|\mathbf{A}\|_1^2}{\varepsilon^2 n}\right).$$

Here, $\|\mathbf{A}\|_1$ is the Schatten-1 norm of $\mathbf{A}$ and $\widetilde{\Omega}$ hides the logarithm factor.

A.2 Information-Theoretic Tools

We first recall standard information-theoretic quantities used throughout our analysis. Let P and Q be two probability distributions defined on a common measurable space $(\mathcal{Z}, \mathcal{F})$. Suppose they are absolutely continuous with respect to a base measure μ, with density functions p and q, respectively.

Definition 21 (Total Variation Distance). The total variation (TV) distance between P and Q is defined as

$$D_{\text{TV}}(P\|Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)| = \frac{1}{2}\int_{\mathcal{Z}} |p(z) - q(z)| d\mu(z).$$

Definition 22 (Kullback-Leibler Divergence). The Kullback-Leibler (KL) divergence between P and Q is defined as

$$D_{\text{KL}}(P\|Q) = \int_{\mathcal{Z}} p(z) \ln \frac{p(z)}{q(z)} d\mu(z),$$

where we use the convention that $0 \ln(0/a) = 0$. The divergence is defined to be $+\infty$ if P is not absolutely continuous with respect to Q.

Definition 23 (Mutual Information)**.** Let X and Y be two random variables taking values in $\mathcal{X}$ and $\mathcal{Y}$ respectively, with a joint distribution $P_{X,Y}$ and marginal distributions P_X and P_Y. The mutual information between X and Y is defined as the KL divergence between their joint distribution and the product of their marginals:

$$I(X;Y) = D_{\mathrm{KL}}(P_{X,Y} \| P_X \otimes P_Y) = \mathbb{E}_{P_{X,Y}}\left[\ln \frac{dP_{X,Y}}{d(P_X \otimes P_Y)}\right].$$

We also need a fundamental information-theoretic tool, Fano's Inequality, which lower bounds the probability of error in multiple hypothesis testing.

Lemma 24 (Fano's Inequality [23, 40])**.** *Let V be a random variable uniformly distributed over a finite set $\mathcal{V}$ with cardinality $|\mathcal{V}| = M \geq 2$. Let Z be any observation (or sequence of observations) related to V through a Markov chain $V \to Z$. For any estimator $\hat{V}$ which is a measurable function of Z, the probability of error satisfies:*

$$\mathbf{Pr}[\hat{V} \neq V] \geq 1 - \frac{I(V;Z) + \ln 2}{\ln M},$$

where $I(V;Z)$ denotes the mutual information between V and Z.

We also recall the Gilbert-Varshamov bound.

Proposition 25 (Gilbert-Varshamov bound [22, 29])**.** *Fix any $d_{VC} \in \mathbb{N}$ and $0 < \alpha < 1/2$. There exists a set of binary strings $\mathcal{S} \subseteq \{0,1\}^{d_{VC}}$ such that for any $s, s' \in \mathcal{S}$ with $s \neq s'$, the Hamming distance satisfies $d_H(s,s') \geq \alpha d_{VC}$, and the cardinality of $\mathcal{S}$ is at least $|\mathcal{S}| \geq \exp(d_{VC}(\ln 2 - H_e(\alpha)))$.*

A.3 Matrix Analysis

We use bold uppercase letters for matrices and bold lowercase letters for vectors. For a vector $\mathbf{v}$, we use $\|\mathbf{v}\|_p$ to denotes its ℓ_p norm.

For an invertible matrix A, we use the notation $\mathbf{A}^{-1}$ to denote its inverse and $\mathbf{A}^\top$ to denote its transpose. For a rank d matrix $\mathbf{A}$, we use the notation $\lambda_{\max}(\mathbf{A}) = \lambda_1(\mathbf{A}) \geq \cdots \lambda_d(\mathbf{A}) = \lambda_{\min}(A)$ to denote its eigenvalues. We drop $(\cdot)$ when it is clear from the context. We use the following two families of matrix norms for a given matrix A:

$$\|\mathbf{A}\|_{p\to q} = \max \frac{\|\mathbf{A}\mathbf{x}\|_q}{\|\mathbf{x}\|_p}$$

for $p, q \in [1, \infty)$. It is well know that when $p = 1$ and $q = 2$, then its value is the maximum Euclidean norm of the columns of $\mathbf{A}$. By continuity, we also define

$$\|\mathbf{A}\|_{2\to\infty} = \max \frac{\|\mathbf{A}\mathbf{x}\|_\infty}{\|\mathbf{x}\|_2}$$

The other family that we use is the family of Schatten-p norms, which is the ℓ_p norm of its singular value.

Proposition 26 (Sherman–Morrison-Woodbury [32])**.** *Suppose $\mathbf{A} \in \mathbb{R}^{d\times d}$ is an invertible square matrix and $\mathbf{v} \in \mathbb{R}^d$ is a column vector. Then $\mathbf{A} + \mathbf{v}\mathbf{v}^\top$ is invertible if $1 + \mathbf{v}^\top \mathbf{A}\mathbf{v} \neq 0$ and in this case:*

$$(\mathbf{A} + \mathbf{v}\mathbf{v}^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{v}\mathbf{v}^\top\mathbf{A}^{-1}}{1 + \mathbf{v}^\top\mathbf{A}^{-1}\mathbf{v}}.$$

The following proposition states that for two symmetric positive semi-definite (PSD) matrices A, B, the product ABA is also a symmetric PSD matrix.

Proposition 27. *Let $A, B \in \mathbb{R}^{d\times d}$ be symmetric positive semi-definite matrices. Then ABA is also symmetric and positive semi-definite.*

A.4 Probability Theory

Lemma 28 (Chernoff Bound for Lower Tail [13]). *Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables such that $\mathbf{Pr}[X_i = 1] = p_i$. Let $X = \sum_{i=1}^n X_i$ denote their sum and let $\mu = \mathbb{E}[X] = \sum_{i=1}^n p_i$ be its expected value. Then, for any $0 < \delta < 1$, we have:*

$$\mathbf{Pr}[X \leq (1-\delta)\mu] \leq \exp\left(-\frac{\delta^2\mu}{2}\right).$$

Next, we introduce formal definitions and useful properties regarding subgaussian random variables and vectors.

Definition 29 (Subgaussian Random Variable). Let Y be a random variable on $\mathbb{R}$. Define its subgaussian norm to be $\|Y\|_{\psi_2} = \inf\{K > 0 : \mathbb{E}[\exp(Y^2/K^2)] \leq 2\}$. Then Y is subgaussian if $\|Y\|_{\psi_2} < \infty$.

Definition 30 (Subgaussian Random Vector). Fix any $d \in \mathbb{N}_+$. A random vector V in $\mathbb{R}^d$ is a subgaussian vector if its one-dimensional marginals $\langle V, \theta\rangle$ are subgaussian variables for all $\theta \in \mathbb{R}^d$. The subgaussian norm of V is defined by

$$\|V\|_{\psi_2} = \sup_{\|\theta\|_2=1} \|V^\top \theta\|_{\psi_2}.$$

Definition 31 (Isotropic Random Vector). A random vector V in $\mathbb{R}^d$ is isotropic if $\mathbb{E}[VV^\top] = I_d$, where I_d denotes the $d \times d$ identity matrix.

Remark 32. Let V be the vector of d i.i.d. random variables from Gaussian distribution $\mathcal{N}(0,1)$, then it is well known that V is an isotropic random vector with $\|V\|_{\psi_2} = O(1)$.

We recall the following fact regarding the high-probability bound on the singular values of isotropic subgaussian matrices.

Proposition 33 (Chapter 4.6 in [35]). *Let $Z \in \mathbb{R}^{n\times d}$ be a random matrix whose rows $\{Z_i\}_{i\in[n]}$ are i.i.d sampled from a mean-zero, subgaussian, isotropic distribution on $\mathbb{R}^d$. Suppose that $\forall i, \|Z_i\|_{\psi_2} \leq K$. Let s_{min} be its smallest singular value, then for any $t > 0$:*

$$\mathbf{Pr}\left[\sqrt{n} - CK^2(\sqrt{d}+t) \leq s_{min} \leq s_{\max} \leq \sqrt{n} + CK^2(\sqrt{d}+t)\right] > 1 - 2\exp(-t^2)$$

for some universal constant C.

Proposition 34 (Hanson-Wright inequality [31]). *Let $\mathbf{z} = (z_1, \ldots, z_d)^\top$ be a random vector with independent, zero-mean, sub-gaussian coordinates. Let $K = \mathbf{x}_i\|z_i\|_{\psi_2}$ and $\mathbf{A} \in \mathbb{R}^{d\times d}$ be any deterministic matrix. Then, there exists an absolute constant $c > 0$ such that for any $t > 0$:*

$$\mathbf{Pr}\left[\left|\mathbf{z}^\top\mathbf{A}\mathbf{z} - \mathbb{E}[\mathbf{z}^\top\mathbf{A}\mathbf{z}]\right| > t\right] \leq 2\exp\left(-c\min\left(\frac{t^2}{K^4\|\mathbf{A}\|_F^2}, \frac{t}{K^2\|\mathbf{A}\|_2}\right)\right).$$

We also introduce the following fact about the inverse Wishart distribution, which is useful when the underlying data distribution is Gaussian.

Proposition 35 (Mardia, Kent, and Bibby [28]). *Let $x_1, x_2, \cdots, x_n$ be i.i.d. samples drawn from $\mathcal{N}(0,\Sigma)$ where Σ is a positive definite covariance matrix. Let X be the $n \times d$ data matrix where the i-th row is $\mathbf{x}_i^\top$. Then, we have that $(\mathbf{X}^\top\mathbf{X})^{-1}$ follows an inverse Wishart distribution with parameters Σ^{-1} and $n-d-1$, and thus for $n > d+1$,*

$$\mathop{\mathbb{E}}_{X\sim\mathcal{N}(0,\Sigma)}[(\mathbf{X}^\top\mathbf{X})^{-1}] = \frac{\Sigma^{-1}}{n-d-1}.$$

B Auxiliary Results

We need the following generalization of isotropic property of Gaussian distribution to sub-Gaussian distribution:

Proposition 36. *Let $\mathbf{z}$ be a random vector on $\mathbb{R}^d$ with covariance matrix $\mathbb{E}[\mathbf{z}\mathbf{z}^\top] = \Sigma$. Then $\mathbf{v} = \Sigma^{-1/2}\mathbf{z}$ is isotropic and that for any $\theta \in \mathbb{R}^d$,*

$$\|\mathbf{z}^\top\theta\|_{\psi_2} \le \|\mathbf{v}\|_{\psi_2} \cdot \|\Sigma^{1/2}\theta\|_2.$$

Proof. Since Σ is symmetric PSD, we have

$$\mathbb{E}[\mathbf{v}\mathbf{v}^\top] = \mathbb{E}[\Sigma^{-1/2}\mathbf{z}\mathbf{z}^\top\Sigma^{-1/2}] = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = I_d$$

and thus $\mathbf{v}$ is isotropic. Further, we notice that for any vector $\mathbf{u} \in \mathbb{R}^d$:

$$\|\mathbf{v}^\top\mathbf{u}\|_{\psi_2} \le \|\mathbf{u}\|_2 \cdot \sup_{\theta\in\mathbb{R}^d;\, \theta^\top\theta=1} \|\mathbf{v}^\top\theta\|_{\psi_2} = \|\mathbf{v}\|_{\psi_2}\|u\|_2$$

where the equality comes from Definition 30. Therefore,

$$\|\mathbf{z}^\top\theta\|_{\psi_2} = \|\mathbf{v}^\top(\Sigma^{1/2}\theta)\|_{\psi_2} \le \|\mathbf{v}\|_{\psi_2} \cdot \|\Sigma^{1/2}\theta\|_2.$$

completing the proof of Proposition 36. □

The proof of Theorem 52 relies on that fact that $\frac{1}{n}(\mathbf{X}^\top\mathbf{X}) \succeq \frac{1}{2}\Sigma_x$ holds with high probability, which we show below:

Claim 37. *For $n, d \in \mathbb{N}_+$, let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be the subgaussian matrix whose rows $\{\mathbf{x}_i\}_{i\in[n]}$ are i.i.d. random subgaussian vectors with zero mean and covariance Σ_x, satisfying that*

$$\|\langle\mathbf{x}_i, \mathbf{v}\rangle\|_{\psi_2} \le K\|\Sigma_x^{1/2}\mathbf{v}\|_2, \quad \textit{for all } \mathbf{v} \in \mathbb{R}^d$$

(i.e., following Properties (1) and (2)). Then if $n \ge c'dK^4$ for some universal constant c', we have the following:

$$\mathbf{Pr}\left[\frac{n}{2}\Sigma_x \preceq \mathbf{X}^\top\mathbf{X}\right] \ge 1 - \exp\left(-\frac{c_2 n}{K^4}\right),$$

where the probability is over the randomness of $\mathbf{X}$.

Proof. Fix an arbitrary $i \in [n]$. We write $\mathbf{x}_i = \Sigma_x^{1/2}\mathbf{z}_i$. We note that $\mathbf{z}_i$ is isotropic as

$$\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^\top] = \mathbb{E}[\Sigma_x^{-1/2}\mathbf{x}_i\mathbf{x}_i^\top\Sigma_x^{-1/2}] = \Sigma_x^{-1/2}\Sigma_x\Sigma_x^{-1/2} = I_d.$$

Further, $\mathbf{z}_i$ is subgaussian with $\|\mathbf{z}_i\|_{\psi_2} \le K$ since

$$\|\mathbf{z}_i\|_{\psi_2} = \sup_{\mathbf{u}\in\mathbb{R}^d;\, \mathbf{u}^\top\mathbf{u}=1} \|\mathbf{z}_i^\top\mathbf{u}\|_{\psi_2} = \sup_{\mathbf{u}\in\mathbb{R}^d;\, \mathbf{u}^\top\mathbf{u}=1} \|\mathbf{x}_i^\top(\Sigma_x^{-1/2}\mathbf{u})\|_{\psi_2} \le K\|\Sigma_x^{-1/2}\Sigma_x^{1/2}\mathbf{u}\|_2 = K.$$

Write $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n)^\top$, then $\mathbf{X} = \mathbf{Z}\Sigma_x^{1/2}$ and thus $\frac{1}{n}\mathbf{X}^\top\mathbf{X} = (\Sigma_x^{1/2})^\top\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{Z}\right)\Sigma_x^{1/2}$. Multiplying $\Sigma_x^{-1/2}$ on both sides gives

$$\frac{1}{n}(\Sigma_x^{-1/2})^\top\mathbf{X}^\top\mathbf{X}\Sigma_x^{-1/2} = \frac{1}{n}\mathbf{Z}^\top\mathbf{Z}.$$

Therefore, as long as we can show that $\frac{1}{n}\lambda_{\min}(\mathbf{Z}^\top\mathbf{Z}) \ge 1/2$, it is enough to see that

$$\frac{1}{n}(\Sigma_x^{-1/2})^\top\mathbf{X}^\top\mathbf{X}\Sigma_x^{-1/2} \succeq \frac{1}{2}I_d$$

and thus $\frac{1}{n}(\mathbf{X}^\top\mathbf{X}) \succeq \frac{1}{2}\Sigma_x$. From Proposition 33, by letting $t = O(\sqrt{n}/K^2)$, we see that there exist constants $c_1, c_2 > 0$ such that if $n > c_1 dK^4$, then

$$\mathbf{Pr}\left[\lambda_{\min}(\mathbf{Z}^\top\mathbf{Z}) \ge \frac{n}{2}\right] = \mathbf{Pr}\left[s_{\min}(\mathbf{Z}) \ge \sqrt{\frac{n}{2}}\right] \ge 1 - \exp(-c_2 n/K^4)$$

where $s_{\min}(\mathbf{Z})$ is the least singular value of the $n \times d$ isotropic sub-gaussian matrix $\mathbf{Z}$. This means that with probability at least $1 - e^{-c_2 n/K^4}$, we have that

$$\frac{1}{n}\lambda_{\min}(\mathbf{Z}^\top\mathbf{Z}) \ge 1/2$$

and thus $\frac{1}{n}(\mathbf{X}^\top\mathbf{X}) \succeq \frac{1}{2}\Sigma_x$. □

We need the following bound on the concentration property of subGaussian random vector in the proof of Theorem 52.

Lemma 38 (Concentration of the Euclidean norm of a sub-Gaussian vector)**.** *Let* $\mathbf{u} = (u_1, \ldots, u_d)^\top \in \mathbb{R}^d$ *be a random vector whose components* u_i *are independent, satisfying* $\mathbb{E}[u_i] = 0$ *and Var*$(u_i) = 1$*. Suppose each* u_i *is a sub-gaussian random variable with a bounded sub-gaussian norm, meaning* $x_i \|u_i\|_{\psi_2} \leq K$*. Then, there exists an absolute constant* $C > 0$ *such that for any* $t > 0$*:*

$$\mathbf{Pr}\left[\|\mathbf{u}\|_2^2 \geq d + CK^2\sqrt{dt} + CK^2 t\right] \leq \exp(-t).$$

Proof. We express the squared norm as a quadratic form, $\|\mathbf{u}\|_2^2 = \mathbf{u}^\top I_d \mathbf{u}$, with expectation $\mathbb{E}[\mathbf{u}^\top I_d \mathbf{u}] = d$. Applying the Hanson-Wright inequality (Proposition 34) to I_d (where $\|I_d\|_F^2 = d$ and $\|I_d\|_2 = 1$), we bound the deviation $s > 0$ as:

$$\mathbf{Pr}\left[\|\mathbf{u}\|_2^2 - d \geq s\right] \leq \exp\left(-c \min\left(\frac{s^2}{K^4 d}, \frac{s}{K^2}\right)\right).$$

Setting the argument of the exponent to t, we get $c \min\left(\frac{s^2}{K^4 d}, \frac{s}{K^2}\right) = t$. Solving for s requires $s \geq K^2\sqrt{\frac{dt}{c}}$ and $s \geq \frac{K^2 t}{c}$, meaning $s = \max\left(K^2\sqrt{\frac{dt}{c}}, \frac{K^2 t}{c}\right)$. Using the inequality $\max(a, b) \leq a + b$, we can relax this to $s \leq CK^2\sqrt{dt} + CK^2 t$, where $C = \max(1/\sqrt{c}, 1/c)$. Substituting this s back into the probability bound completes the proof. □

Proposition 39. *Let* $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{d \times d}$ *be three matrices such that* $\mathbf{A} - \mathbf{B}$ *and* $\mathbf{C}$ *are positive semi-definite. If* $\mathbf{A} \succeq \mathbf{B}$*, then tr*$(\mathbf{CA}) \geq$ *tr*$(\mathbf{CB})$.

Proof. Given $\mathbf{A} \succeq \mathbf{B}$, we have $\mathbf{A} - \mathbf{B} \succeq 0$. Let $\mathbf{D} = \mathbf{A} - \mathbf{B}$, which means $\mathbf{D}$ is a symmetric positive semi-definite matrix ($\mathbf{D} \succeq 0$). The inequality $\text{tr}(\mathbf{CA}) \geq \text{tr}(\mathbf{CB})$ is equivalent to $\text{tr}(\mathbf{C}(\mathbf{A} - \mathbf{B})) = \text{tr}(\mathbf{CD}) \geq 0$. Since $\mathbf{C} \succeq 0$, there exists a unique symmetric positive semi-definite square root matrix $\mathbf{C}^{1/2}$ such that $\mathbf{C} = \mathbf{C}^{1/2}\mathbf{C}^{1/2}$. Substituting this into the trace and applying the cyclic property of the trace operator, we obtain $\text{tr}(\mathbf{CD}) = \text{tr}(\mathbf{C}^{1/2}\mathbf{C}^{1/2}\mathbf{D}) = \text{tr}(\mathbf{C}^{1/2}\mathbf{D}\mathbf{C}^{1/2})$.

Let $\mathbf{M} = \mathbf{C}^{1/2}\mathbf{D}C^{1/2}$. For any arbitrary column vector $\mathbf{x}$, its quadratic form evaluates to $\mathbf{x}^\top\mathbf{M}\mathbf{x} = \mathbf{x}^\top\mathbf{C}^{1/2}\mathbf{D}\mathbf{C}^{1/2}\mathbf{x} = (\mathbf{C}^{1/2}\mathbf{x})^\top\mathbf{D}(\mathbf{C}^{1/2}\mathbf{x})$. By defining $\mathbf{y} = \mathbf{C}^{1/2}\mathbf{x}$, this simplifies to $\mathbf{y}^\top\mathbf{D}\mathbf{y}$. Because $\mathbf{D} \succeq 0$, the definition of positive semi-definiteness dictates that $\mathbf{y}^\top\mathbf{D}\mathbf{y} \geq 0$ for all $\mathbf{y}$. Consequently, $\mathbf{x}^\top\mathbf{M}\mathbf{x} \geq 0$ for all $\mathbf{x}$, proving that the matrix M is also positive semi-definite ($\mathbf{M} \succeq 0$). The trace of a positive semi-definite matrix is the sum of its non-negative eigenvalues, ensuring that $\text{tr}(\mathbf{M}) \geq 0$. Therefore, $\text{tr}(\mathbf{CD}) = \text{tr}(\mathbf{M}) \geq 0$, which yields $\text{tr}(\mathbf{C}(\mathbf{A} - \mathbf{B})) \geq 0$ and finally concludes the proof that $\text{tr}(\mathbf{CA}) \geq \text{tr}(\mathbf{CB})$. □

The following lemma provides an upper bound on the expectation of the inverse of the regularized empirical covariance matrix.

Lemma 40. *For* $n, d \in \mathbb{N}_+$*, let* $\mathbf{X} \in \mathbb{R}^{n \times d}$ *be a random matrix whose rows* $\{\mathbf{x}_i\}_{i \in [n]}$ *are i.i.d. subgaussian random vectors with zero mean and covariance* Σ_x*, satisfying that* $\|\langle \mathbf{x}_i, \mathbf{v}\rangle\|_{\psi_2} \leq K\|\Sigma_x^{1/2}\mathbf{v}\|_2$ *for all* $\mathbf{v} \in \mathbb{R}^d$ *(i.e., following Properties (1) and (2)). Then we have that*

$$\mathop{\mathbb{E}}_{\mathbf{X}}[(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}] \preceq \frac{2}{n}\Sigma_x^{-1} + \frac{e^{-cn/K^4}}{\lambda} I_d.$$

Proof. Note that Claim 37 implies that with probability at least $1 - e^{-cn/K^4}$ for some constant $c > 0$, we have

$$(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1} \preceq (\mathbf{X}^\top\mathbf{X})^{-1} \preceq \frac{2}{n}\Sigma_x^{-1}.$$

Since $(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1} \preceq \frac{1}{\lambda} I_d$ holds unconditionally, we can bound the expectation by conditioning on the high-probability event and its complement. This yields

$$\mathop{\mathbb{E}}_{\mathbf{X}}[(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}] \preceq \frac{2}{n}\Sigma_x^{-1} + \frac{e^{-cn/K^4}}{\lambda} I_d,$$

completing the proof of Lemma 40. □

The following fact is well known, we include a self-contained proof.

Proposition 41. *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be any real matrix. For any parameter $\lambda > 0$ and any $i \in [n]$, the i-th ridge **leverage score**, defined as*

$$h_i = x_i^\top (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} x_i,$$

satisfies $h_i \leq 1$.

Proof. We begin by isolating the contribution of the i-th sample. Let $\mathbf{X}_{-i} \in \mathbb{R}^{(n-1)\times d}$ denote the design matrix with the i-th row removed, and define the corresponding leave-one-out covariance matrix as $M_{-i} = \mathbf{X}_i{-}i^\top \mathbf{X}_i{-}i + \lambda I_d$. Because the parameter satisfies $\lambda > 0$ and $\mathbf{X}_i{-}i^\top \mathbf{X}_i{-}i$ is inherently positive semi-definite, their sum M_{-i} is strictly positive definite and therefore invertible. Further we note that

$$\mathbf{X}^\top \mathbf{X} + \lambda I_d = \left(\sum_{i\in[n]} \mathbf{x}_i \mathbf{x}^\top \right) + \lambda I_d = M_{-i} + \mathbf{x}_i \mathbf{x}_i^\top.$$

By applying the Sherman-Morrison Theorem (Proposition 26), we see that

$$(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} = (\mathbf{M}_{-i} + \mathbf{x}_i \mathbf{x}_i^\top)^{-1} = \mathbf{M}_{-i}^{-1} - \frac{\mathbf{M}_{-i}^{-1} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{M}_{-i}^{-1}}{1 + \mathbf{x}_i^\top \mathbf{M}_{-i}^{-1} \mathbf{x}_i}$$

as $1 + \mathbf{x}_i^\top \mathbf{M}_{-i}^{-1} \mathbf{x}_i > 0$. Substituting this exact expansion back into the original definition of the leverage score gives

$$h_i = \mathbf{x}_i^\top \left(\mathbf{M}_{-i}^{-1} - \frac{\mathbf{M}_{-i}^{-1} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{M}_{-i}^{-1}}{1 + \mathbf{x}_i^\top \mathbf{M}_{-i}^{-1} \mathbf{x}_i} \right) \mathbf{x}_i = \mathbf{x}_i^\top \mathbf{M}_{-i}^{-1} \mathbf{x}_i - \frac{(\mathbf{x}_i^\top \mathbf{M}_{-i}^{-1} \mathbf{x}_i)^2}{1 + \mathbf{x}_i^\top \mathbf{M}_{-i}^{-1} \mathbf{x}_i}.$$

Write $u = \mathbf{x}_i^\top \mathbf{M}_{-i}^{-1} \mathbf{x}_i$. Since $\mathbf{M}_{-i}^{-1}$ is positive definite, the scalar u is non-negative for any vector $\mathbf{x}_i \in \mathbb{R}^d$. Rewriting the leverage score strictly in terms of u, the equation collapses to

$$h_i = u - \frac{u^2}{1+u} = \frac{u(1+u) - u^2}{1+u} = \frac{u}{1+u} \leq 1$$

for every $u \geq 0$, which completes the proof of Proposition 41. □

Proposition 42 (Variance Contraction of Ridge Estimators)**.** *Let $X \in \mathbb{R}^{n\times d}$ be a matrix with full column rank, and let $\Sigma_x \in \mathbb{R}^{d\times d}$ be a symmetric positive definite covariance matrix. For any parameter $\lambda > 0$, the following trace inequality holds:*

$$\mathbf{tr}\left(\mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \Sigma_x (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} X^\top \right) \leq \mathbf{tr}\left(\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \Sigma_x (\mathbf{X}^\top \mathbf{X})^{-1} X^\top \right).$$

Proof. Applying the cyclic permutation property of the trace operator, we shift the rightmost matrix $X^\top$ to the front on both sides of the inequality. The right-hand side reduces to

$$\text{tr}(\Sigma_x (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}) = \text{tr}(\Sigma_x (\mathbf{X}^\top \mathbf{X})^{-1}),$$

while the left-hand side becomes

$$\text{tr}(\Sigma_x (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} (\mathbf{X}^\top \mathbf{X}) (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1}).$$

Let $A = \mathbf{X}^\top \mathbf{X}$. Since X possesses full column rank, A is symmetric positive definite. To establish the trace inequality under multiplication by any positive definite Σ_x, it suffices to prove the underlying matrix inequality $(A + \lambda I_d)^{-1} A (A + \lambda I_d)^{-1} \preceq A^{-1}$. Because A and the resolvent $(A + \lambda I_d)^{-1}$ share the exact same eigenbasis, we can verify this domination directly via their spectra. For any eigenvalue $\mu > 0$ of A, the corresponding eigenvalue of the left-hand matrix is structurally given by $\frac{\mu}{(\mu+\lambda)^2}$, whereas the corresponding eigenvalue for A^{-1} is $\frac{1}{\mu}$. Since both the eigenvalue μ and the regularization parameter λ are strictly positive, it immediately follows that $\frac{\mu}{(\mu+\lambda)^2} < \frac{\mu}{\mu^2} = \frac{1}{\mu}$. As this holds uniformly across all eigenvalues, the strict matrix inequality $(A+\lambda I_d)^{-1} A (A+\lambda I_d)^{-1} \prec A^{-1}$ is established. By Proposition 39, multiplying by $\Sigma_x \succ 0$ and taking the trace preserves this ordering, which completes the proof upon restoring $A = \mathbf{X}^\top \mathbf{X}$. This completes the proof of Proposition 42. □

Proposition 43 (Quadratic Bound in Matrix Partial Order)**.** *Let $\mathbf{M}, \Sigma \in \mathbb{R}^{d\times d}$ be symmetric positive definite matrices, and let $c > 0$ be a scalar. If $\mathbf{M} \preceq \mathbf{c}\Sigma^{-1}$, then $\mathbf{M}\Sigma\mathbf{M} \preceq c^2\Sigma^{-1}$.*

Proof. Assume $\mathbf{M} \preceq c\Sigma^{-1}$. Because Σ is symmetric positive definite, its principal square root $\Sigma^{1/2}$ and inverse square root $\Sigma^{-1/2}$ are well-defined, symmetric, and positive definite, which implies that $\Sigma^{1/2}\mathbf{M}\Sigma^{1/2} \preceq cI$. Let $\mathbf{K} = \Sigma^{1/2}\mathbf{M}\Sigma^{1/2}$. Since $\mathbf{M}$ is positive definite and congruence transformations preserve positive definiteness, $\mathbf{K}$ is a symmetric positive definite matrix satisfying $0 \prec \mathbf{K} \preceq cI$. This implies that all eigenvalues of $\mathbf{K}$ are strictly positive and bounded above by the scalar c. Consequently, the eigenvalues of the squared matrix K^2 are bounded by c^2, which establishes the matrix inequality $\mathbf{K}^2 \preceq c^2 I$. Substituting the definition of K back into this squared inequality yields

$$(\Sigma^{1/2}\mathbf{M}\Sigma^{1/2})(\Sigma^{1/2}\mathbf{M}\Sigma^{1/2}) \preceq c^2 I,$$

which implies that $\Sigma^{1/2}\mathbf{M}\Sigma\mathbf{M}\Sigma^{1/2} \preceq c^2 I$. Finally, we multiply both sides by $\Sigma^{-1/2}$ on the left and right. This operation yields

$$\Sigma^{-1/2}(\Sigma^{1/2}\mathbf{M}\Sigma\mathbf{M}\Sigma^{1/2})\Sigma^{-1/2} \preceq \Sigma^{-1/2}(c^2 I)\Sigma^{-1/2},$$

which implies $\mathbf{M}\Sigma\mathbf{M} \preceq c^2\Sigma^{-1}$, completing the proof of Proposition 43. □

C Related Works

Approximate local-DP lower bound. Our lower bound follows the standard broad contraction-plus-testing blueprint for local privacy, but the technical setting here differs from that of the closest prior work. In particular, Asoodeh et al. [1] study approximate local differential privacy in the full privacy setting where the *entire* datum is private, and develop a general divergence-based contraction framework for deriving Le Cam/Fano lower bounds. By contrast, our model is *label-private*: the feature vectors are public and only the labels are randomized. As a result, lower bounds for full local DP do not directly imply lower bounds for local label DP, and our proof must control mixtures of public and private information. Moreover, our argument is not a minor extension of Asoodeh et al.: their proof is built around a general E_γ-divergence characterization of approximate local DP, whereas our lower bound is derived through a contraction argument tailored to the label-private classification setting and then combined with a VC-dimension packing and Fano reduction.

The remaining prior work is also technically different. Xu et al. [39] study binary classification under *pure* local label DP achieved by randomized response and analyze a specific randomized-response-based classification pipeline, while Zhao et al. [42] give a broader minimax treatment of label DP but, in the local setting, focusing on pure ε-label local DP rather than the approximate (ε, δ) regime considered here. Our contribution is therefore best viewed as a lower bound specialized to *approximate local label-DP classification*, with techniques adapted to this model rather than obtained by routine modification of prior full-local-DP or pure-label-DP analyses.

Efficient label-DP algorithms in the central model. Beimel, Nissim, and Stemmer [8] imply the same central label-private classification rate when their result is translated into convergence-rate form. The relevant difference for our purposes is computational: while their result shows that this guarantee is achievable, the corresponding learner is not a generic efficient algorithm for arbitrary classes, and later work explicitly characterizes this route as inefficient [19]. By contrast, our result gives an efficient central label-DP learner with the same qualitative guarantee in a general concept-class setting. Other related works differ in a different way: Esfandiari et al. [19] give an efficient centralized mechanism under additional clustering assumptions, while Bassily, Mohri, and Suresh [5] provide a general label-DP learner with margin-based guarantees rather than the same excess-risk guarantee.

Label-DP algorithms for regression. The closest prior work is Wang and Xu [37], who study sparse linear regression under local differential privacy and also analyze a restricted response-private setting, but their guarantees are for the high-dimensional sparse regime and are obtained via a label-private version of iterative hard thresholding rather than the dense low-dimensional OLS setting considered here. Ghazi et al. [21] study regression with label differential privacy as a mechanism-design problem and characterize optimal label-randomization mechanisms for regression losses, but they do not provide the same kind of statistical prediction-risk bound for Gaussian linear regression. Classical local-DP regression results such as Duchi et al. [14] concern the stronger model in which the

entire datum is private, and therefore do not directly imply this theorem, while central-DP regression results such as Cai, Wang, and Zhang [10] are more relevant to our central-model theorems than to the local label-private one.

D Hardness of Classification in Local Model with Approximate DP

In this section, we first generalize the contraction inequality in Duchi et al. (2013) [14] for local differential privacy from pure-DP to approximate-DP, conditional on the public features[9]. Then, with this tool, we prove the lower bound on the minimax risk of binary classification under approximate label differential privacy.

D.1 Contraction by Approximate Differential Privacy

Before stating the main contraction result, we first introduce the background of parameter testing under local differential privacy. Let $\mathcal{V}$ be a finite set of parameters. Suppose we draw a parameter V uniformly at random from $\mathcal{V}$. Conditioned on $V = \nu$, we draw n independent and identically distributed data entries $R_1, R_2, \ldots, R_n$ from a distribution P_ν defined over a general data domain $\mathcal{R}$. To preserve privacy, each data entry R_i is passed through an independent local randomized mechanism $Q : \mathcal{R} \to \mathcal{Z}$, producing private observations $Z_1, Z_2, \ldots, Z_n$. This generative process naturally forms a Markov chain $V \to \{R_i\}_{i\in[n]} \to \{Z_i\}_{i\in[n]}$. We recall the following contraction inequality for pure local-DP channel:

Theorem 44 (Corollary 1 in [14])**.** *Let V be a uniformly random parameter in $\mathcal{V}$. For any ε-locally differentially private channel Q, we have*

$$I(Z_1, Z_2, \cdots, Z_n; V) \leq \frac{4n(e^\varepsilon - 1)^2}{M^2} \sum_{\nu,\nu'\in\mathcal{V}} D_{TV}^2(P_\nu || P_{\nu'}).$$

To precisely analyze the approximate local differential privacy, we define the neighboring relationship on the data domain. For any data entry $r \in \mathcal{R}$, let $\Gamma(r) \subseteq \mathcal{R}$ denote the set of all possible neighboring entries to r. We define the maximum neighborhood size across the domain as $\Gamma(\mathcal{R}) = \max_{r\in\mathcal{R}} |\Gamma(r)|$. The following theorem extends the mutual information contraction result in Duchi et al. (2013) [14] to approximate local label-DP, conditional on the public features.

Theorem 45 (Approximate DP Mutual Information Contraction)**.** *Fix $0 < \varepsilon \leq 1$ and $0 \leq \delta \leq 1$. Consider the Markov chain $V \to R = \{R_i\}_{i\in[n]} \to Z = \{Z_i\}_{i\in[n]}$ as described above, with $R_i = (X_i, Y_i) \in \mathcal{X} \times \{+1, -1\}$ and $Z_i \leftarrow Q(X_i, Y_i)$, where Q satisfies (ε, δ)-local label-DP. Assume that $I(V; X_{1:n}) = 0$, so all P_ν have the same feature marginal P_X, and write $P_\nu(\cdot \mid X)$ for the conditional label distribution. Let $T = (X_{1:n}, Z_{1:n})$ be the complete learner transcript and $\tilde{\delta} = \frac{\delta}{1-e^{-\varepsilon}}$. Suppose δ is sufficiently small such that $n\tilde{\delta} < 1/2$.Then there exist universal constants $C, C' > 0$ such that*

$$\begin{aligned} I(V;T) &= I(V; Z_{1:n} \mid X_{1:n}) \\ &\leq \frac{Cn(\varepsilon')^2}{|\mathcal{V}|^2} \sum_{\nu,\nu'\in\mathcal{V}} \mathop{\mathbb{E}}_{X\sim P_X} \left[D_{TV}^2(P_\nu(\cdot \mid X) \| P_{\nu'}(\cdot \mid X)) \right] \\ &\quad + C' n\tilde{\delta} \ln\left(\frac{|\mathcal{V}|}{n\tilde{\delta}}\right), \end{aligned}$$

where $\varepsilon' = 3\varepsilon + \ln\left(\frac{1}{1-2n\tilde{\delta}}\right)$, and the last term is interpreted as zero when $\delta = 0$.

Proof. By the chain rule and the independence assumption,

$$I(V;T) = I(V; X_{1:n}) + I(V; Z_{1:n} \mid X_{1:n}) = I(V; Z_{1:n} \mid X_{1:n}).$$

Fix $X_{1:n} = x_{1:n}$ and write $Q_i(y) = Q(x_i, y)$. Each Q_i is an ordinary binary (ε, δ)-LDP channel. Let $q_{i,y}$ be its output density with respect to a common dominating measure μ_i. According to

[9] Taking the public features to be constant also gives a contraction bound for ordinary approximate local DP on a binary input domain.

Proposition 18, the bad set $B_{i,y} = \{z : q_{i,y}(z) > e^{2\varepsilon} q_{i,-y}(z)\}$ has $Q_i(y)$-probability at most $\tilde{\delta}$. Define the truncated and normalized channel by

$$q^0_{i,y}(z) = \min\{q_{i,y}(z), e^{2\varepsilon} q_{i,-y}(z)\}, \qquad a_{i,y} = 1 - \int q^0_{i,y}\, d\mu_i \le \tilde{\delta},$$

$$\widehat{q}_{i,y}(z) = \frac{q^0_{i,y}(z)}{1 - a_{i,y}}.$$

The truncated densities have common support and likelihood ratios between $e^{-2\varepsilon}$ and $e^{2\varepsilon}$. Hence each normalized channel $\widehat{Q}_i$ is ε_0-pure LDP, where

$$\varepsilon_0 = 2\varepsilon + \ln\left(\frac{1}{1-\tilde{\delta}}\right) \le \varepsilon', \qquad \varepsilon_0 \le 2 + \ln 2.$$

Moreover, $D_{\mathrm{TV}}(Q_i(y) \| \widehat{Q}_i(y)) \le a_{i,y} \le \tilde{\delta}$. Let $\widehat{Z}_{1:n}$ be the conditionally independent outputs of these channels given the original labels. Conditional on $X_{1:n} = x_{1:n}$, V remains uniform and the labels are independent given V. Applying Theorem 44 coordinatewise and using conditional independence gives

$$\begin{aligned} I(V; \widehat{Z}_{1:n} \mid X_{1:n} = x_{1:n}) &\le \sum_{i=1}^{n} I(V; \widehat{Z}_i \mid X_{1:n} = x_{1:n}) \\ &\le \frac{C_2(\varepsilon')^2}{|\mathcal{V}|^2} \sum_{i=1}^{n} \sum_{\nu,\nu' \in \mathcal{V}} D^2_{\mathrm{TV}}(P_\nu(\cdot \mid x_i) \| P_{\nu'}(\cdot \mid x_i)), \end{aligned}$$

for a universal constant C_2, since $(e^{\varepsilon_0} - 1)^2 = O(\varepsilon_0^2) \le O((\varepsilon')^2)$ on the bounded range above.

Couple Z_i and $\widehat{Z}_i$ independently across coordinates, conditional on the inputs, so that their mismatch probability is at most $\tilde{\delta}$. Let E indicate any mismatch. By the union bound, $\mathbf{Pr}[E = 1 \mid X_{1:n} = x_{1:n}] \le n\tilde{\delta} = \delta'$. Using the chain rule of mutual information, and the fact that $Z_{1:n} = \widehat{Z}_{1:n}$ when $E = 0$, we obtain

$$\begin{aligned} I(V; Z_{1:n} \mid X_{1:n} = x_{1:n}) &\le I(V; \widehat{Z}_{1:n}, E, Z_{1:n} \mid X_{1:n} = x_{1:n}) \\ &\le I(V; \widehat{Z}_{1:n} \mid X_{1:n} = x_{1:n}) + h(\delta') + \delta' \ln |\mathcal{V}|, \end{aligned}$$

where h is the binary entropy function. Averaging over $X_{1:n} \sim P_X^n$ and using $h(\delta') \le C_1 \delta' \ln(1/\delta')$ for $\delta' < 1/2$ yields the stated bound. □

Remark on the Neighborhood Size. In our label-DP setting, the abstract data entry is defined as $r = (\mathbf{x}, y) \in \mathcal{X} \times \{+1, -1\}$. Since the features are public and fixed, the privacy mechanism only acts on the label. Therefore, for any given entry r, the only valid neighboring entry r' is the one where the label y is flipped. This implies that the neighborhood size is exactly $|\Gamma(r)| = 1$ for all r, yielding $\Gamma(\mathcal{R}) = 1$. The contraction bound from Theorem 45 applies conditional on the public features, with $I(V; X_{1:n}) = 0$ ensuring that they reveal no information about V.

D.2 Minimax Lower Bound for Classification (Proof of Theorem 3)

With the generalized contraction bound in hand, we formalize the minimax lower bound for the binary classification task under approximate label-DP in the local model. For the ease of reader, we recall the lower bound.

Theorem 3. (Lower bound for approximate local label DP). *Let $\mathcal{H}$ be a hypothesis class over an input space $\mathcal{X}$ with finite VC dimension $VC(\mathcal{H}) = d_{VC}$. Consider a binary classification task on $D = \{(\mathbf{x}_i, y_i)\}_{i \in [n]}$ where a learner receives n private observations $Z_1, \ldots, Z_n$ generated via an (ε, δ)-local differentially private mechanism Q that acts independently on each $(\mathbf{x}_i, y_i)$ $(i \in [n])$. Assume $\varepsilon < 1$, and that the privacy parameter δ is sufficiently small such that*

$$\tilde{\delta} \ln\left(\frac{M}{n\tilde{\delta}}\right) = O\left(\frac{d_{VC}}{n}\right) \quad \textit{where} \quad M = e^{\Theta(d_{VC})} \quad \textit{and} \quad \tilde{\delta} = \frac{\delta}{1 - e^{-\varepsilon}}. \tag{2}$$

Then, the minimax excess risk is lower bounded by

$$R(\mathcal{H}, \varepsilon, \delta) = \Omega\left(\frac{1}{\varepsilon'}\sqrt{\frac{d_{VC}}{n}}\right), \quad \textit{where} \quad \varepsilon' = 3\varepsilon + \ln\left(\frac{1}{1 - 2n\tilde{\delta}}\right).$$

Proof. The proof of Theorem 3 reduces the learning problem to a multiple hypothesis testing problem by constructing a suitable packing set for the set of hypotheses.

Consider an agnostic PAC learning instance where the hypothesis class $\mathcal{H}$ has VC dimension d_{VC}. Let $\mathcal{X}_0 = \{\mathbf{u}_1, \mathbf{u}_2 \ldots, \mathbf{u}_{d_{\text{VC}}}\} \subseteq \mathcal{X}$ be a set of d_{VC} points shattered by $\mathcal{H}$. For any binary string $s \in \{0,1\}^{d_{\text{VC}}}$, there exists $h_s \in \mathcal{H}$ such that $h_s(\mathbf{u}_j) = s_j$ for all $j \in [d_{\text{VC}}]$. Let $\mathcal{P}_{\mathcal{X}}$ be a uniform distribution on $\mathcal{X}_0$. The $L_1(\mathcal{P}_{\mathcal{X}})$ distance between any two hypotheses $h_s, h_{s'}$, defined by $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{P}_{\mathcal{X}}}[h_s(\mathbf{x}) \neq h_{s'}(\mathbf{x})]$, is exactly the Hamming distance between s and s' divided by d_{VC}.

By Proposition 25, we can construct a packing set $\mathcal{W} = \{h_1, \ldots, h_M\} \subseteq \mathcal{H}$ such that:

1. **Separability:** For all $i \neq j$, the ℓ_1 distance satisfies $\rho(h_i, h_j) = \mathbf{Pr}_{\mathbf{x}\sim\mathcal{P}_{\mathcal{X}}}[h_i(\mathbf{x}) \neq h_j(\mathbf{x})] \geq \alpha$.
2. **Cardinality:** $\ln M \geq d_{\text{VC}}(\ln 2 - H_e(\alpha)) \geq d_{\text{VC}}(\ln 2 - \alpha \ln(1/\alpha) - O(\alpha))$ for $0 < \alpha < 1/2$.

To explicitly establish the lower bound for the agnostic setting, let J be a random variable distributed uniformly over $\{1, \ldots, M\}$, independently of $X_{1:n} \sim \mathcal{P}_{\mathcal{X}}^n$, so that $I(J; X_{1:n}) = 0$. Nature chooses the target hypothesis $h^* = h_J$. For each sample $\mathbf{x}_i$ ($i \in [n]$), the pre-privacy true label y_i is generated such that $\mathbf{Pr}[y_i = h_J(\mathbf{x}_i) \mid \mathbf{x}_i] = \frac{1}{2} + \gamma$ and $\mathbf{Pr}[y_i \neq h_J(\mathbf{x}_i) \mid \mathbf{x}_i] = \frac{1}{2} - \gamma$, where $\gamma \in (0, 1/2]$ is a margin parameter. The learner observes the complete transcript $T = (X_{1:n}, Z_{1:n})$, where $Z_{1:n} = (Z_1, \ldots, Z_n)$ and each $Z_i \leftarrow Q(\mathbf{x}_i, y_i)$.

By Fano's Inequality (Lemma 24), the probability of error $\mathbf{Pr}[\hat{J} \neq J]$ for any estimator $\hat{J}$ based on T is lower bounded by:

$$\mathbf{Pr}[\hat{J} \neq J] \geq 1 - \frac{I(J;T) + \ln 2}{\ln M}.$$

We contextualize local differential privacy by applying our derived approximate-DP contraction bound (Theorem 45). Let $X \sim \mathcal{P}_{\mathcal{X}}$ be the random feature, and let P_j be the joint distribution of the pre-privacy data pair (X, Y) given $h^* = h_j$. Note that for label-DP, the data domain is $\mathcal{R} = \mathcal{X} \times \mathcal{Y}$, and the maximum neighborhood size is $\Gamma(\mathcal{R}) = 1$. Following Theorem 45, for an (ε, δ)-local label-DP channel Q:

$$\begin{aligned} I(J;T) &= I(J; Z_{1:n} \mid X_{1:n}) \\ &\leq \frac{Cn(\varepsilon')^2}{M^2} \sum_{j,k=1}^{M} \mathop{\mathbb{E}}_{X\sim\mathcal{P}_{\mathcal{X}}} \left[D_{\text{TV}}^2(P_j(\cdot \mid X) \| P_k(\cdot \mid X))\right] + C'\delta' \ln\left(\frac{M}{\delta'}\right), \end{aligned} \tag{3}$$

where $\varepsilon' = 3\varepsilon + \ln(1/(1-2\delta'))$ and $\delta' = n\tilde{\delta} = n\delta/(1-e^{-\varepsilon})$. Here $P_j(\cdot \mid \mathbf{x}) = P_{Y|X}(\cdot \mid \mathbf{x}, h_j)$. Since the marginal distribution of the features X is identical (i.e., $\mathcal{P}_{\mathcal{X}}$) for all hypotheses, the bound averages the squared conditional label TV distance over this common marginal.

In our agnostic construction, if $h_j(\mathbf{x}) = h_k(\mathbf{x})$, the conditional pre-privacy label distributions are identical ($D_{\text{TV}} = 0$). If $h_j(\mathbf{x}) \neq h_k(\mathbf{x})$, the distributions are $\text{Bernoulli}(1/2+\gamma)$ and $\text{Bernoulli}(1/2-\gamma)$ up to permutation, yielding $D_{\text{TV}} = 2\gamma$. Thus:

$$D_{\text{TV}}^2(P_{Y|X}(\cdot \mid \mathbf{x}, h_j) \| P_{Y|X}(\cdot \mid \mathbf{x}, h_k)) = 4\gamma^2 \mathbf{1}\{h_j(\mathbf{x}) \neq h_k(\mathbf{x})\}$$

for a fixed $\mathbf{x}$. Taking the expectation over $\mathbf{x} \sim \mathcal{P}_{\mathcal{X}}$, we obtain:

$$\mathop{\mathbb{E}}_{X\sim\mathcal{P}_{\mathcal{X}}} \left[D_{\text{TV}}^2(P_j(\cdot \mid X) \| P_k(\cdot \mid X))\right] = 4\gamma^2 \mathbf{Pr}_{\mathbf{x}\sim\mathcal{P}_{\mathcal{X}}}[h_j(\mathbf{x}) \neq h_k(\mathbf{x})] \leq C''\gamma^2.$$

for some constant C''. Summing over the packing set and absorbing constants into C and C', the mutual information bound simplifies to:

$$I(J;T) = I(J; Z_{1:n} \mid X_{1:n}) \leq C \cdot n(\varepsilon')^2\gamma^2 + C'\delta' \ln\left(\frac{M}{\delta'}\right)$$

where C is another universal constant depending on C''. Let $\Phi : \mathcal{X}^n \times \mathcal{Z}^n \to [M]$ be any test function of the complete transcript. Substituting the bounds into Fano's Inequality and choosing $\alpha = \Theta(1)$ (so $\ln M = \Theta(d_{\text{VC}})$):

$$\mathbf{Pr}[\Phi(T) \neq J] \geq 1 - \frac{C \cdot n(\varepsilon')^2\gamma^2 + C'\delta' \ln\left(\frac{M}{\delta'}\right) + \ln 2}{C_0 \cdot d_{\text{VC}}}.$$

To ensure the error probability P_e is bounded away from zero (e.g., $P_e \geq 1/2$), we require:

$$C \cdot n(\varepsilon')^2\gamma^2 + C'\delta' \ln\left(\frac{M}{\delta'}\right) \leq \frac{1}{2}C_0 \cdot d_{\text{VC}} - \ln 2.$$

By our assumption that δ is sufficiently small, the term $C'\delta' \ln\left(\frac{M}{\delta'}\right)$ is at most $\frac{1}{4}C_0 d_{\text{VC}}$. This leaves:

$$C \cdot n(\varepsilon')^2\gamma^2 \leq \frac{1}{4}C_0 \cdot d_{\text{VC}} \implies \gamma \leq O\left(\frac{1}{\varepsilon'}\sqrt{\frac{d_{\text{VC}}}{n}}\right).$$

This implies that if we set the agnostic margin $\gamma \approx \frac{1}{\varepsilon'}\sqrt{\frac{d_{\text{VC}}}{n}}$, any estimator will have an error probability of at least $1/2$. In this agnostic PAC setting, let $c(\mathbf{x})$ be the random true label. The generalization error of h_k for $k \neq J$ corresponds to:

$$\begin{aligned}
R(h_k) - \min_{h^* \in \mathcal{H}} R(h^*) &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{P}_\mathcal{X}}[h_k(\mathbf{x}) \neq c(\mathbf{x})] - \min_{h^* \in \mathcal{H}} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{P}_\mathcal{X}}[h^*(\mathbf{x}) \neq c(\mathbf{x})] \\
&\geq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{P}_\mathcal{X}}[h_k(\mathbf{x}) \neq c(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{P}_\mathcal{X}}[h_J(\mathbf{x}) \neq c(\mathbf{x})] \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{P}_\mathcal{X}}\left[\mathbf{1}\{h_k(\mathbf{x}) \neq h_J(\mathbf{x})\} \cdot \left(\left(\frac{1}{2}+\gamma\right) - \left(\frac{1}{2}-\gamma\right)\right)\right] \\
&= 2\gamma \cdot \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{P}_\mathcal{X}}[\mathbf{1}\{h_k(\mathbf{x}) \neq h_J(\mathbf{x})\}] \\
&\geq 2\gamma\alpha,
\end{aligned}$$

from the separability of the packing set. Setting $\alpha = \Theta(1)$, we conclude that the minimax excess risk is at least:

$$R(\mathcal{H}, \varepsilon, \delta) = \Omega\left(\frac{1}{\varepsilon'}\sqrt{\frac{d_{\text{VC}}}{n}}\right).$$

This completes the proof of Theorem 3. □

E An Efficient Algorithm for Private Classification in Central Model

This section shows that central differential privacy enables a superior convergence rate of approximately $\tilde{O}(\sqrt{d_{\text{VC}}/(n\varepsilon)})$ for binary classification with a time-efficient algorithm. This performance surpasses the lower bound derived in Section D.2, even in the pure differential privacy setting.

We first recall that, simply using randomized response (Lemma 14) on the label of each data entry already gives a label DP algorithm in the local setting with a convergence rate of $\tilde{O}(\sqrt{\frac{d_{\text{VC}}}{\varepsilon^2 n}})$; this matches the lower bound (Theorem 3, Section D.2) up to logarithmic factors.

Theorem 46 (Esfandiari et al. [19])**.** *Fix $\varepsilon > 0$. Let $\mathcal{H}$ be a hypothesis class over an input space $\mathcal{X}$ with finite VC dimension $VC(\mathcal{H}) = d_{VC}$. Consider an agnostic binary classification task with an arbitrary data distribution $\mathcal{P}$ over $\mathcal{X} \times \{+1, -1\}$. Then, there exists an ε-local label differentially private algorithm $\mathcal{A}$ that, given n i.i.d. samples from $\mathcal{P}$, $\mathcal{A}$ outputs a hypothesis $\hat{h} \in \mathcal{H}$ such that with a constant probability (over the randomness of $\mathcal{A}$ and $\mathcal{P}$), the excess risk satisfies:*

$$R(\hat{h}) - \min_{h \in \mathcal{H}} R(h) = \tilde{O}\left(\frac{1}{\varepsilon}\sqrt{\frac{d_{VC}}{n}}\right),$$

where $R(h) = \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{P}}[h(\mathbf{x}) \neq y]$ and the $\tilde{O}(\cdot)$ notation hides logarithmic factors.

In the central model, Beimel et al. [7] achieve a better convergence rate $\tilde{O}(\sqrt{d_{\text{VC}}/(\varepsilon n)})$ by using the exponential mechanism on a subset of the dataset, which satisfies label DP in the central setting. However, their mechanism is inefficient. Interestingly, we show that there is an efficient algorithm that achieves the same convergence rate $\tilde{O}(\sqrt{d_{\text{VC}}/(\varepsilon n)})$ for the central model in the high privacy regime of a constant ε, which is a significant improvement over the local DP setting.

Algorithm 1: PAC Learning Under Label DP in Central Model

Input: A dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, privacy budget $\varepsilon \in (0, 1]$.

Output: A hypothesis $\hat{h} \in \mathcal{H}$.

Initialize a subset $S = \varnothing$;

for *each* $(\mathbf{x}_i, y_i) \in D$ **do**

  Add $(\mathbf{x}_i, y_i)$ to S independently with probability $\gamma = \frac{e^\varepsilon - 1}{e - 1}$;

Initialize a private dataset $\tilde{D} = \varnothing$;

for *each* $(\mathbf{x}, y) \in S$ **do**

  Set $\tilde{y} = y$ with probability $\frac{e}{e+1}$, and $\tilde{y} = -y$ with probability $\frac{1}{e+1}$;

  Add $(\mathbf{x}, \tilde{y})$ to $\tilde{D}$;

Select $\hat{h} \in \mathcal{H}$ that minimizes the empirical error on $\tilde{D}$;

return $\hat{h}$.

Intuition behind Algorithm 1. The simplicity of Algorithm 1 relies on the synergistic combination of two theoretical pillars: privacy amplification by subsampling and binary classification with noisy labels. The core mechanism applied to the labels is the standard Randomized Response (Lemma 14) with a fixed privacy parameter $\varepsilon_0 = 1$, which satisfies 1-DP. Crucially, instead of applying this to the entire dataset, we independently subsample each data point with probability $\gamma = \frac{e^\varepsilon - 1}{e - 1}$. According to the classic privacy amplification by subsampling theorem [4], running an ε_0-DP mechanism on a random γ-fraction of the dataset amplifies the privacy guarantee to exactly $\ln(1 + \gamma(e^{\varepsilon_0} - 1))$-DP. Since $\varepsilon_0 = 1$, substituting our choice of γ yields exactly ε-central label DP. Finally, we apply Theorem 46 on the smaller dataset which gives the desired error bound.

Remark 47. We note that privacy amplification by subsampling is inherently exclusive to the central DP model, as the aggregator perfectly observes whether and how each user transmits their local message, the uncertainty of participation is exposed, rendering local-DP impossible.

Theorem 4. *Let $\mathcal{H}$ be a hypothesis class over an input space $\mathcal{X}$ with finite VC dimension $\text{VC}(\mathcal{H}) = d_{\text{VC}}$. Consider an agnostic binary classification task with a data distribution $\mathcal{P}$ over $\mathcal{X} \times \{+1, -1\}$. For any $\varepsilon \in (10/n, 1]$, there exists a polynomial-time ε-central label differentially private algorithm (Algorithm 1) that, given n i.i.d. samples from $\mathcal{P}$, outputs a hypothesis $\hat{h} \in \mathcal{H}$ such that with a constant probability, the excess risk satisfies:*

$$R(\hat{h}) - \min_{h \in \mathcal{H}} R(h) = \tilde{O}\left(\sqrt{\frac{d_{\text{VC}}}{n\varepsilon}}\right), \quad \textit{where } R(h) = \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{P}}[h(\mathbf{x}) \neq y].$$

Proof. We divide the proof into two parts: the privacy guarantee and bounding the excess risk.

Privacy Analysis. In Algorithm 1, the sub-routine applied to the subsampled dataset S is the Randomized Response mechanism. Specifically, for each retained sample $(\mathbf{x}, y) \in S$, we flip its label y with probability $\eta = 1/(e + 1)$. By Lemma 14, this label-flipping mechanism $\mathcal{M}$ satisfies ε_0-local label differential privacy with $\varepsilon_0 = \ln\left(\frac{e/(e+1)}{1/(e+1)}\right) = 1$.

The overall algorithm can be represented as the composition $\mathcal{M} \circ \mathcal{S}_\gamma$, where $\mathcal{S}_\gamma$ is the Poisson subsampling step with probability $\gamma = \frac{e^\varepsilon - 1}{e - 1}$. According to Theorem 15, the composed mechanism is ε-central label DP, as

$$\ln(1 + \gamma(e^{\varepsilon_0} - 1)) = \ln\left(1 + \frac{e^\varepsilon - 1}{e - 1} \cdot (e - 1)\right) = \varepsilon.$$

Utility Analysis. Let $m = |\tilde{D}|$ be the random variable denoting the size of the privatized dataset. Since each of the n original samples is included independently with probability γ, m follows the distribution Binomial(n, γ). Note that for all $\varepsilon > 0$, we have $e^{\varepsilon} - 1 \geq \varepsilon$, which implies $\gamma \geq \frac{\varepsilon}{e-1} = \Theta(\varepsilon)$. Assuming $\varepsilon = \Omega(1/n)$ such that $n\gamma$ is sufficiently large, the Chernoff bound (Lemma 28) ensures that $m \geq \Omega(n\varepsilon)$ with high probability $1 - o(1)$.

Crucially, conditioned on the subsampled set S, the process of generating $\tilde{D}$ and selecting $\hat{h}$ is equivalent to a PAC learning task under 1-local label differential privacy, executed on a dataset of size m. By directly applying the upper bound established in Theorem 46 with effective sample size m and privacy budget $\varepsilon_{local} = 1$, we obtain that with a constant probability:

$$R(\hat{h}) - \min_{h \in \mathcal{H}} R(h) = \tilde{O}\left(\frac{1}{\varepsilon_{local}}\sqrt{\frac{d_{\text{VC}}}{m}}\right) = \tilde{O}\left(\sqrt{\frac{d_{\text{VC}}}{n\varepsilon}}\right).$$

This matches the stated convergence rate and completes the proof of Theorem 4. □

Remark 48. The improvement from $\tilde{O}(\sqrt{d_{\text{VC}}/(\varepsilon^2 n)})$ to $\tilde{O}(\sqrt{d_{\text{VC}}/(\varepsilon n)})$ formally shows that the interaction in the central model can significantly reduce the sample complexity by a factor of $1/\sqrt{\varepsilon}$ compared to the local model. From a high level perspective, it implies the central DP *contracts* the information of the dataset by a factor of ε while the local DP contracts the information by a factor of ε^2. The latter is consistent with the contraction results in Duchi et al [14].

F Label-Private Regression: Lower and Upper Bound

In this section, we consider the regression problem under label differential privacy. We start by giving the formal problem statement and auxiliary lemma used in the proof of our algorithm. We then give our algorithm for local label-DP in Appendix F.1, then give our algorithm that uses output perturbation for regression in Appendix F.2 followed by our lower bound stating that this result is tight for isotropic Gaussian setting in Appendix F.3 . We end this section by giving our main algorithmic result in Appendix F.4.

Problem Statement. As an *extension* of the standard setting in Cai et al. [10], we assume that each public data entry $\mathbf{x}_i$ is a random d-dimensional vector that is drawn i.i.d. from some (not necessarily isotropic) sub-Gaussian distribution with an arbitrary covariance matrix $\Sigma_x \in \mathbb{R}^{d \times d}$. In particular, we assume that the distribution of each $\mathbf{x}_i$ (for $i \in [n]$) satisfies that

(1) (Zero mean and known variance.) $\mathbb{E}[\mathbf{x}_i] = \mathbf{0}$; $\quad \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\top] = \Sigma_x$.

(2) (Sub-Gaussianity.) There exists a constant $K > 0$ such that for any vector $\mathbf{v} \in \mathbb{R}^d$, we have $\|\langle \mathbf{x}_i, \mathbf{v} \rangle\|_{\psi_2} \leq K \|\Sigma_x^{1/2} \mathbf{v}\|_2$.

We remark that from Proposition 36, by letting $K = 2\sqrt{2}/3$[10], this setting recovers that in Cai et al. [10] where each data entry x_i is assumed to identically follow a Gaussian distribution $\mathcal{N}(0, \Sigma_x)$. More broadly, this setting also allows our analysis to apply to a wider range of practical distributions, including uniform and discrete distributions such as Rademacher vectors.

For the distribution of the private labels, we consider the Gaussian linear model [10] where the private label y_i is generated by

$$y_i = \mathbf{x}_i^\top \theta^* + \zeta_i.$$

Here, the **unit** vector $\theta^* \in \mathbb{R}^d$ is the unknown parameter and $\zeta_i \sim \mathcal{N}(0, \sigma_\zeta^2)$ is the noise random variable that is drawn i.i.d. for some $\sigma_\zeta > 0$.

The following lemma is for upper bounding the bias caused by the regularization term λ in the regression algorithm for both our algorithms.

[10]Note that if $x_i \sim \mathcal{N}(0, \Sigma_x)$, then Proposition 36 implies that $\|\langle x_i, v \rangle\|_{\psi_2} \leq \|z\|_{\psi_2} \|\Sigma_x^{1/2} v\|_2$ holds for any $v \in \mathbb{R}^d$, where $z \sim \mathcal{N}(0, 1)$. It is a well-known result (e.g., Vershynin [35]) that a standard Gaussian random vector satisfies $\|z\|_{\psi_2} \leq 2\sqrt{2}/3$.

Lemma 49. *Define $\hat{\theta}_{ridge} = (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top \mathbf{y}$. Then there is a universal constant $c > 0$ such that*

$$\mathbb{E}\left[\|\hat{\theta}_{ridge} - \theta^*\|_{\Sigma_x}^2\right] \leq \sigma_\zeta^2 \cdot \left(\frac{2d}{n} + \frac{e^{-cn/K^4} \cdot d \cdot \lambda_{\max}}{\lambda}\right) + \frac{4\lambda^2}{n^2 \lambda_{\min}} + \frac{e^{-cn/K^4} \lambda_{\max}^2}{\lambda_{\min}}.$$

Proof. We first note that,

$$\begin{aligned}
\hat{\theta}_{\text{ridge}} &= (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top \mathbf{y} = (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top (\mathbf{X}\theta^* + \zeta) \\
&= (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top \mathbf{X}\theta^* + (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top \zeta \\
&= (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} (\mathbf{X}^\top \mathbf{X} + \lambda I_d)\theta^* - (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} (\lambda I_d \theta^*) + (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top \zeta \\
&= \theta^* - \lambda (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \theta^* + (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top \zeta.
\end{aligned}$$

Therefore, we have

$$\hat{\theta}_{\text{ridge}} - \theta^* = (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top \zeta - \lambda (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \theta^*.$$

Thus,

$$\mathbb{E}\left[(\hat{\theta}_{\text{ridge}} - \theta^*)^\top \Sigma_x (\hat{\theta}_{\text{ridge}} - \theta^*)\right] \leq \mathbb{E}\left[\|(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top \zeta\|_{\Sigma_x}^2\right] + \lambda^2 \mathbb{E}\left[\|(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \theta^*\|_{\Sigma_x}^2\right]$$

as $\mathbb{E}[\zeta] = \mathbf{0}$. First, for the variance term, we have

$$\begin{aligned}
\mathbb{E}\left[\|(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top \zeta\|_{\Sigma_x}^2\right] &= \mathbb{E}\left[\zeta^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \Sigma_x (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top \zeta\right] \\
&= \mathbb{E}\left[\text{tr}\left(\zeta \zeta^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \Sigma_x (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top\right)\right] \\
&= \text{tr}\left(\mathbb{E}\left[\zeta \zeta^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \Sigma_x (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top\right]\right) \\
&= \sigma_\zeta^2 \cdot \text{tr}\left(\mathbb{E}\left[\mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \Sigma_x (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top\right]\right) \\
&= \sigma_\zeta^2 \cdot \mathbb{E}\left[\text{tr}\left(\mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \Sigma_x (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top\right)\right] \\
&= \sigma_\zeta^2 \cdot \mathbb{E}\left[\text{tr}\left((\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \Sigma_x (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} (\mathbf{X}^\top \mathbf{X})\right)\right] \\
&\leq \sigma_\zeta^2 \cdot \mathbb{E}\left[\text{tr}\left((\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \Sigma_x (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} (\mathbf{X}^\top \mathbf{X} + \lambda I_d)\right)\right] \\
&= \sigma_\zeta^2 \cdot \mathbb{E}\left[\text{tr}\left((\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \Sigma_x\right)\right] \\
&= \sigma_\zeta^2 \cdot \text{tr}\left(\Sigma_x \, \mathbb{E}[(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1}]\right).
\end{aligned}$$

Here, the fourth equality comes from the fact that ζ is independent of $\mathbf{X}$ and that $\mathbb{E}[\zeta \zeta^\top] = \sigma_\zeta^2 I_n$, and the inequality is because of Proposition 39 and Proposition 27. In particular, since both $(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1}$ and Σ_x are symmetric PSD matrices, Proposition 27 guarantees that $(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \Sigma_x (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1}$ is also PSD, which makes Proposition 39 applicable. Further, from Lemma 40 we have that

$$\mathbb{E}[(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1}] \preceq \frac{2}{n} \Sigma_x^{-1} + \frac{e^{-cn/K^4}}{\lambda} I_d.$$

Then, by applying Proposition 39 we have that

$$\begin{aligned}
\mathbb{E}\left[\|(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top \zeta\|_{\Sigma_x}^2\right] &\leq \sigma_\zeta^2 \cdot \text{tr}\left(\Sigma_x \left(\frac{2}{n} \Sigma_x^{-1} + \frac{e^{-cn/K^4}}{\lambda} I_d\right)\right) \\
&\leq \sigma_\zeta^2 \cdot \frac{2}{n} \text{tr}\left(I_d\right) + \sigma_\zeta^2 \cdot \frac{e^{-cn/K^4}}{\lambda} \text{tr}\left(\Sigma_x\right) \\
&\leq \sigma_\zeta^2 \cdot \left(\frac{2d}{n} + \frac{e^{-cn/K^4} \cdot d \cdot \lambda_{\max}}{\lambda}\right).
\end{aligned}$$

Next, for the bias term, we have

$$\mathbb{E}\left[\|(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \theta^*\|_{\Sigma_x}^2\right] = \mathbb{E}\left[(\theta^*)^\top (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \Sigma_x (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \theta^*\right]$$

Since the smallest eigenvalue of Σ_x^{-1} is $\lambda_{\max}^{-1}$, we have the following bound:

$$(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \preceq \frac{1}{\lambda} I_d \preceq \frac{\lambda_{\max}}{\lambda} \Sigma_x^{-1}. \tag{4}$$

Therefore, by Proposition 39 we conclude that:

$$(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \Sigma_x (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \preceq \frac{\lambda_{\max}^2}{\lambda^2} \Sigma_x^{-1}$$

Now, we define an event over the randomness of $\mathbf{X}$ to be $E = \{\mathbf{X} : (\mathbf{X}^\top \mathbf{X})^{-1} \preceq \frac{2}{n} \Sigma_x^{-1}\}$. By applying Proposition 43, we see that conditioned on E,

$$(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \Sigma_x (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \preceq \frac{4}{n^2} \Sigma_x^{-1}$$

and conditioned on $\overline{E}$, we have $(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \Sigma_x (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \preceq \frac{\lambda_{\max}^2}{\lambda^2} \Sigma_x^{-1}$ holds unconditionally. Combining these two facts, we have that:

$$\begin{aligned}
\mathbb{E}\left[\|(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \theta^*\|_{\Sigma_x}^2\right] &= \mathbb{E}\left[(\theta^*)^\top (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \Sigma_x (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \theta^*\right] \\
&\leq (1 - e^{-cn/K^4}) \, \mathbb{E}\left[\frac{4}{n^2} (\theta^*)^\top \Sigma_x^{-1} \theta^*\right] + e^{-cn/K^4} \, \mathbb{E}\left[\frac{\lambda_{\max}^2}{\lambda^2} (\theta^*)^\top \Sigma_x^{-1} \theta^*\right] \\
&\leq \left(\frac{4}{n^2} + e^{-cn/K^4} \cdot \frac{\lambda_{\max}^2}{\lambda^2}\right) \cdot \frac{1}{\lambda_{\min}} \\
&= \frac{4}{n^2 \lambda_{\min}} + \frac{e^{-cn/K^4} \lambda_{\max}^2}{\lambda^2 \lambda_{\min}}.
\end{aligned}$$

Putting everything together, we are able to conclude that:

$$\begin{aligned}
\mathbb{E}\left[(\hat{\theta}_{\text{ridge}} - \theta^*)^\top \Sigma_x (\hat{\theta}_{\text{ridge}} - \theta^*)\right] &\leq \mathbb{E}\left[\|(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top \zeta\|_{\Sigma_x}^2\right] + \lambda^2 \, \mathbb{E}\left[\|(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \theta^*\|_{\Sigma_x}^2\right] \\
&\leq \sigma_\zeta^2 \cdot \left(\frac{2d}{n} + \frac{e^{-cn/K^4} \cdot d \cdot \lambda_{\max}}{\lambda}\right) + \frac{4\lambda^2}{n^2 \lambda_{\min}} + \frac{e^{-cn/K^4} \lambda_{\max}^2}{\lambda_{\min}},
\end{aligned}$$

which completes the proof of Lemma 49. □

F.1 Local DP: A Trivial Randomized Response Algorithm

As a *warm-up* result, we first consider the following algorithm that uses local randomizer on each label. Here we write $\mathbf{X}$ to be the $n \times d$ data matrix where the i-th row is $\mathbf{x}_i$, and write $\mathbf{y}$ to be the vector of labels. The algorithm simply adds i.i.d. Gaussian noise to each label, which is a standard local randomizer for regression problems. Then, we compute the least squares estimate based on the privatized labels and the original data matrix. To enhance the stability for sub-Gaussian analysis we also introduce the ridge regularizer $\lambda > 0$.

Algorithm F.1: Linear Regression Under Local Label DP

1. **Input:** A dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, privacy parameters $\varepsilon, \delta \in (0, 1)$, and a ridge regularizer $\lambda > 0$.
2. Set the local noise scale $\sigma_{\text{ldp}} = \frac{\sqrt{2 \ln(1.25/\delta)}}{\varepsilon}$.
3. **For** each $i \in [n]$:
   - Sample local Gaussian noise $\eta_i \sim \mathcal{N}(0, \sigma_{\text{ldp}}^2)$.
   - Construct the privatized label $\tilde{y}_i = y_i + \eta_i$.
4. Let $\tilde{\mathbf{y}} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n)^\top$ be the vector of privatized labels, and $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the data matrix where the i-th row is $\mathbf{x}_i^\top$.
5. **Output:** The regularized least squares estimate $\hat{\theta}_{\text{ldp}} = (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top \tilde{\mathbf{y}}$.

We establish the following result:

Theorem 50. *Fix any $0 < \varepsilon, \delta < 1$. Let $\hat{\theta}_{ldp}$ be the output of Algorithm F.1 on input n i.i.d. samples from the Gaussian linear model with variance σ_ζ^2. Suppose that the feature distribution satisfies properties (1), (2). Then, there exists a choice of λ such that:*

$$R(\hat{\theta}_{ldp}) \leq M + e^{-2cn/(3K^4)} \cdot \left(2\sqrt[3]{M^2\Lambda} + \Lambda\right)$$

for $n \gg d$, where

$$M = \frac{4d\sigma_\zeta^2}{n} + \frac{4d\log(1.25/\delta)}{n\varepsilon^2} \quad \textit{and} \quad \Lambda = \frac{\lambda_{\max}^2(\Sigma_x)}{\lambda_{\min}(\Sigma_x)}.$$

Here, $\lambda_{\max}(\Sigma_x)$ and $\lambda_{\min}(\Sigma_x)$ are the largest and smallest eigenvalue of Σ_x, respectively.

Proof. Note that by the definition of the Gaussian linear model, we have that $\tilde{\mathbf{y}} = \mathbf{X}\theta^* + \zeta + \eta$, where ζ is the noise vector of the original labels and η is the noise vector added by the local randomizer. Therefore, we can write that

$$\tilde{\mathbf{y}} = \mathbf{X}\theta^* + \zeta + \eta = \mathbf{X}\theta^* + \mathbf{z} \in \mathbb{R}^n$$

where $\mathbf{z} = \zeta + \eta$ is the effective noise vector, and $\mathbf{z}_i$ is drawn i.i.d. from $\mathcal{N}(0, \sigma_\zeta^2 + \sigma_{\text{ldp}}^2)$. Then, we can write:

$$\begin{aligned}\hat{\theta}_{\text{ldp}} = (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top\tilde{\mathbf{y}} &= (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top(\mathbf{X}\theta^* + \mathbf{z}) \\ &= \underbrace{(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top\mathbf{y}}_{\hat{\theta}_{\text{ridge}}} + (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top\mathbf{z}.\end{aligned}$$

Therefore, the expected prediction risk can be written as

$$\begin{aligned}R(\hat{\theta}_{\text{ldp}}) &= \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{z}}[\|\hat{\theta}_{\text{ldp}} - \theta^*\|_{\Sigma_x}^2] \\ &= \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{z}}[\|\hat{\theta}_{\text{ridge}} - \theta^* + (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top\mathbf{z}\|_{\Sigma_x}^2] \\ &= \mathop{\mathbb{E}}_{\mathbf{X}}[\|\hat{\theta}_{\text{ridge}} - \theta^*\|_{\Sigma_x}^2] + \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{z}}[\|(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top\mathbf{z}\|_{\Sigma_x}^2].\end{aligned}$$

Here, the last equality is because by taking the expectation over $\mathbf{z}$ conditioned on $\mathbf{X}$, the cross term vanishes:

$$\mathop{\mathbb{E}}_{\mathbf{z}|\mathbf{X}}[(\hat{\theta}_{\text{ridge}} - \theta^*)^\top\Sigma_x(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top\mathbf{z}] = 0$$

as $\mathbf{z}$ is independent of $\mathbf{X}$ and $\mathbb{E}[\mathbf{z}] = \mathbf{0}$. The first term $\mathbb{E}_{\mathbf{X}}[\|\hat{\theta}_{\text{ridge}} - \theta^*\|_{\Sigma_x}^2]$ is bounded by Lemma 49.

To upper bound the second term $\mathbb{E}_{\mathbf{X},\mathbf{z}}[\|(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top\mathbf{z}\|_{\Sigma_x}^2]$, we note that:

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{X},\mathbf{z}}\left[\|(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top\mathbf{z}\|_{\Sigma_x}^2\right] &= \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{z}}[\mathbf{z}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\Sigma_x(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top\mathbf{z}] \\ &= \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{z}}\left[\text{tr}\left(\mathbf{z}\mathbf{z}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\Sigma_x(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top\right)\right] \\ &= (\sigma_\zeta^2 + \sigma_{\text{ldp}}^2) \cdot \mathop{\mathbb{E}}_{\mathbf{X}}[\text{tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\Sigma_x(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top)] \\ &= (\sigma_\zeta^2 + \sigma_{\text{ldp}}^2) \cdot \mathop{\mathbb{E}}_{\mathbf{X}}[\text{tr}((\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\Sigma_x(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}(\mathbf{X}^\top\mathbf{X}))] \\ &\leq (\sigma_\zeta^2 + \sigma_{\text{ldp}}^2) \cdot \mathop{\mathbb{E}}_{\mathbf{X}}[\text{tr}((\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\Sigma_x(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}(\mathbf{X}^\top\mathbf{X} + \lambda I_d))] \\ &= (\sigma_\zeta^2 + \sigma_{\text{ldp}}^2) \cdot \mathop{\mathbb{E}}_{\mathbf{X}}[\text{tr}(\Sigma_x(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1})] \\ &= (\sigma_\zeta^2 + \sigma_{\text{ldp}}^2) \cdot \text{tr}\left(\Sigma_x \mathop{\mathbb{E}}_{\mathbf{X}}[(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}]\right).\end{aligned}$$

Here, the second equality follows from the cyclic property of the trace, and the third equality holds because $\mathbf{z}$ is independent of $\mathbf{X}$ and that $\mathbb{E}[\mathbf{z}\mathbf{z}^\top] = (\sigma_\zeta^2 + \sigma_{\text{ldp}}^2)I_n$. Further, the inequality follows from Proposition 27 and Proposition 39. Followed by Lemma 40 we have that

$$\mathop{\mathbb{E}}_{\mathbf{X}}[(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}] \preceq \frac{2}{n}\Sigma_x^{-1} + \frac{e^{-cn/K^4}}{\lambda} I_d.$$

Then, by applying Proposition 39 we have that

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{X},\mathbf{z}}\left[\|(\mathbf{X}^\top\mathbf{X}+\lambda I_d)^{-1}\mathbf{X}^\top\mathbf{z}\|_{\Sigma_x}^2\right] &\le (\sigma_\zeta^2+\sigma_{\text{ldp}}^2)\cdot\text{tr}\left(\Sigma_x\cdot\left(\frac{2}{n}\Sigma_x^{-1}+\frac{e^{-cn/K^4}}{\lambda}I_d\right)\right)\\ &\le (\sigma_\zeta^2+\sigma_{\text{ldp}}^2)\cdot\frac{2}{n}\text{tr}\left(I_d\right)+(\sigma_\zeta^2+\sigma_{\text{ldp}}^2)\cdot\frac{e^{-cn/K^4}}{\lambda}\text{tr}\left(\Sigma_x\right)\\ &\le (\sigma_\zeta^2+\sigma_{\text{ldp}}^2)\cdot\left(\frac{2d}{n}+\frac{e^{-cn/K^4}\cdot d\cdot\lambda_{\max}}{\lambda}\right).\end{aligned}$$

Combining this with Lemma 49, setting

$$\lambda=\frac{1}{2}\left[n^2\cdot d\cdot\lambda_{\min}\cdot\lambda_{\max}\cdot(2\sigma_\zeta^2+\sigma_{\text{ldp}}^2)\cdot e^{-cn/K^4}\right]^{\frac{1}{3}}$$

we conclude that

$$\begin{aligned}R(\hat{\theta}_{\text{ldp}}) &= \mathop{\mathbb{E}}_{\mathbf{X}}[\|\hat{\theta}_{\text{ridge}}-\theta^*\|_{\Sigma_x}^2]+\mathop{\mathbb{E}}_{\mathbf{X},\mathbf{z}}[\|(\mathbf{X}^\top\mathbf{X}+\lambda I_d)^{-1}\mathbf{X}^\top\mathbf{z}\|_{\Sigma_x}^2]\\ &\le (2\sigma_\zeta^2+\sigma_{\text{ldp}}^2)\cdot\left(\frac{2d}{n}+\frac{e^{-cn/K^4}\cdot d\cdot\lambda_{\max}}{\lambda}\right)+\frac{4\lambda^2}{n^2\lambda_{\min}}+\frac{e^{-cn/K^4}\lambda_{\max}^2}{\lambda_{\min}}\\ &= 3\left[\frac{(2\sigma_\zeta^2+\sigma_{\text{ldp}}^2)^2\cdot d^2\cdot\lambda_{\max}^2\cdot e^{-2cn/K^4}}{n^2\lambda_{\min}}\right]^{\frac{1}{3}}+(2\sigma_\zeta^2+\sigma_{\text{ldp}}^2)\cdot\frac{2d}{n}+\frac{\lambda_{\max}^2\cdot e^{-cn/K^4}}{\lambda_{\min}}\end{aligned}$$

Replacing σ_{ldp}^2 with $\frac{2\log(1.25/\delta)}{\varepsilon^2}$, write $\Lambda=\lambda_{\max}^2/\lambda_{\min}$, we have that

$$R(\hat{\theta}_{\text{ldp}})\le\frac{4d\sigma_\zeta^2}{n}+\frac{4d\log(1.25/\delta)}{n\varepsilon^2}+e^{-2cn/(3K^4)}\cdot\left(2\sqrt[3]{\left(\frac{4d\sigma_\zeta^2}{n}+\frac{4d\log(1.25/\delta)}{n\varepsilon^2}\right)^2\Lambda}+\Lambda\right),$$

which completes the proof of Theorem 50. □

F.2 Central DP I: Output Perturbation Algorithm

As our second *warm-up* result, we shift our focus to the central model of label differential privacy. A standard and straightforward approach to achieving central DP for regression problems is *output (objective) perturbation*. Specifically, the algorithm first computes the exact non-private least squares estimate based on the original data matrix $\mathbf{X}$ and the clean label vector $\mathbf{y}$. Crucially, to ensure that the output sensitivity is bounded (which is a prerequisite for adding calibrated noise), we also introduce the ridge regularizer $\lambda>0$ as in Section F.1.

Since the data matrix $\mathbf{X}$ is considered public in the label-DP setting, it is straightforward to explicitly calculate the exact ℓ_2-sensitivity of the output with respect to a single label change:

$$\begin{aligned}\Delta_2(\mathbf{X}) &= \max_{\mathbf{y}'\sim\mathbf{y}}\|(\mathbf{X}^\top\mathbf{X}+\lambda I_d)^{-1}\mathbf{X}^\top(\mathbf{y}-\mathbf{y}')\|_2\\ &\le \max_{\|\mathbf{y}'-\mathbf{y}\|_1\le 1}\|(\mathbf{X}^\top\mathbf{X}+\lambda I_d)^{-1}\mathbf{X}^\top(\mathbf{y}-\mathbf{y}')\|_2\\ &= \max_{\|\mathbf{v}\|_1\le 1}\|(\mathbf{X}^\top\mathbf{X}+\lambda I_d)^{-1}\mathbf{X}^\top\mathbf{v}\|_2\\ &= \|(\mathbf{X}^\top\mathbf{X}+\lambda I_d)^{-1}\mathbf{X}^\top\|_{1\to 2}=\max_{i\in[n]}\|(\mathbf{X}^\top\mathbf{X}+\lambda I_d)^{-1}\mathbf{x}_i\|_2,\end{aligned}\tag{5}$$

assuming the label difference is bounded by $\|\mathbf{y}-\mathbf{y}'\|_1\le 1$. Here, the $1\to 2$ operator norm evaluates to the maximum ℓ_2-norm among all columns of the matrix $(\mathbf{X}^\top\mathbf{X}+\lambda I_d)^{-1}\mathbf{X}^\top$. After obtaining the exact ridge regression weights, we inject a scaled d-dimensional Gaussian noise vector η calibrated to $\Delta_2(\mathbf{X})$ directly into the output. The complete procedure is detailed in Algorithm F.2.

Algorithm F.2: Label-DP Regression via Output Perturbation

1. **Input:** A dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, privacy parameters $\varepsilon, \delta \in (0,1)$, and a ridge regularizer $\lambda > 0$.
2. Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ be the vector of clean labels, and $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the data matrix where the i-th row is $\mathbf{x}_i^\top$.
3. Compute the exact regularized least squares estimate:
$$\hat{\theta}_{\text{ridge}} = (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top \mathbf{y}.$$
4. Compute the data-dependent ℓ_2-sensitivity of the output:
$$\Delta_2(\mathbf{X}) = \max_{i \in [n]} \|(\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{x}_i\|_2.$$
5. Set the central noise scale $\sigma_{\text{cdp}} = \frac{\Delta_2(\mathbf{X})\sqrt{2\ln(1.25/\delta)}}{\varepsilon}$.
6. Sample a d-dimensional Gaussian noise vector $\eta \sim \mathcal{N}(0, \sigma_{\text{cdp}}^2 I_d)$.
7. **Output:** The privatized estimate $\hat{\theta}_{\text{cdp}} = \hat{\theta}_{\text{ridge}} + \eta$.

In the analysis of both Section F.2 and Section F.4, recall that we make the following assumption:

Assumption 1. (Linear sub-Gaussian Design) Let $\mathbf{x}$ be the random vector sampled from subgaussian data distribution with property (1) and (2). We assume that the coordinates of the random vector $\Sigma^{-1/2}\mathbf{x}$ are mutually independent.

Remark 51. This assumption is equivalent to the setting where each data entry $\mathbf{x}_i$ in $\mathbf{X}$ is generated by a *linear process* $\mathbf{x} = \Sigma^{1/2}\mathbf{z}$, where $\mathbf{z} \in \mathbb{R}^d$ is a random vector with independent, zero-mean sub-gaussian coordinates. We note that under this assumption, our setting still strictly generalize the case in Cai et al. [10] where $\mathbf{x}_i$'s are Gaussian data since isotropic Gaussian vector automatically satisfies the independent component condition. This is a standard and widely adopted structural condition in high-dimensional probability and statistics for random matrix theory [35] and generative statistical models [36]. Furthermore, it acts as a foundational assumption in seminal works on covariance matrix estimation [25] and compressed sensing [11] to establish optimal non-asymptotic minimax rates and isometry properties.

For the output perturbation algorithm (Algorithm F.2), we show the following:

Theorem 52. *Fix any $0 < \varepsilon, \delta < 1$. Let $\hat{\theta}_{cdp}$ be the output of Algorithm F.2 on input n i.i.d. samples from the Gaussian linear model with variance σ_ζ^2 and subGaussian parameter K (Definition 29). Suppose that the feature distribution satisfies properties (1), (2). Then, there is a choice of $\lambda > 0$ and a universal constant $c > 0$ such that:*
$$R(\hat{\theta}_{cdp}) \leq O\left(\frac{d\sigma_\zeta^2}{n}\right) + O\left(\frac{d^2}{n^2} \cdot \kappa\left(\Sigma_x\right) \cdot \frac{\ln^2(n/\delta)}{\varepsilon^2}\right) + e^{-\frac{2cn}{3K^4}} \cdot 3\sqrt[3]{M_\zeta^2 \Lambda}.$$
for $n \gg d$, where $M_\zeta = \frac{\sigma_\zeta^2 d}{n}$ and $\Lambda = \frac{\lambda_{\max}^2}{\lambda_{\min}}$.

Proof. Let $\hat{\theta}_{\text{ridge}} = (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top \mathbf{y}$ be the least squares estimate with ridge regularizer. Then we have
$$\begin{aligned}
R(\hat{\theta}_{\text{cdp}}) &= \mathbb{E}[\|\hat{\theta}_{\text{cdp}} - \theta^*\|_{\Sigma_x}^2] = \mathbb{E}\left[(\hat{\theta}_{\text{cdp}} - \theta^*)^\top \Sigma_x (\hat{\theta}_{\text{cdp}} - \theta^*)\right] \\
&= \mathbb{E}\left[(\hat{\theta}_{\text{ridge}} - \theta^* + \boldsymbol{\eta})^\top \Sigma_x (\hat{\theta}_{\text{ridge}} - \theta^* + \boldsymbol{\eta})\right] \\
&= \mathbb{E}\left[(\hat{\theta}_{\text{ridge}} - \theta^*)^\top \Sigma_x (\hat{\theta}_{\text{ridge}} - \theta^*)\right] + \mathbb{E}\left[\boldsymbol{\eta}^\top \Sigma_x \boldsymbol{\eta}\right] + 2\,\mathbb{E}\left[(\hat{\theta}_{\text{ridge}} - \theta^*)^\top \Sigma_x \boldsymbol{\eta}\right] \\
&= \mathbb{E}\left[(\hat{\theta}_{\text{ridge}} - \theta^*)^\top \Sigma_x (\hat{\theta}_{\text{ridge}} - \theta^*)\right] + \mathbb{E}\left[\boldsymbol{\eta}^\top \Sigma_x \boldsymbol{\eta}\right]. \qquad (6)
\end{aligned}$$

Here, the last equality is because $\mathbb{E}[\boldsymbol{\eta}] = \mathbf{0}$ and $\boldsymbol{\eta}$ is independent of $\mathbf{X}$ and $\mathbf{y}$. We bound each of the terms in eq. (6) separately. First, choosing

$$\lambda = \frac{n}{2}\sqrt[3]{\sigma_\zeta^2 d\lambda_{\max}\lambda_{\min}e^{-cn/K^4}}$$

in Lemma 49, we have that

$$\begin{aligned}\mathbb{E}\left[\|\hat{\theta}_{\text{ridge}} - \theta^*\|_{\Sigma_x}^2\right] &\leq \sigma_\zeta^2 \cdot \left(\frac{2d}{n} + \frac{e^{-cn/K^4}\cdot d\cdot\lambda_{\max}}{\lambda}\right) + \frac{4\lambda^2}{n^2\lambda_{\min}} + \frac{e^{-cn/K^4}\lambda_{\max}^2}{\lambda_{\min}} \\ &= \frac{2d\sigma_\zeta^2}{n} + \frac{e^{-cn/K^4}\lambda_{\max}^2}{\lambda_{\min}} + 3\left(\frac{\sigma_\zeta^4 d^2\lambda_{\max}^2 e^{-2cn/K^4}}{n^2\lambda_{\min}}\right)^{1/3} \end{aligned} \tag{7}$$

for $n \gg d$. Next we focus on the term of privacy error $\mathbb{E}\left[\boldsymbol{\eta}^\top\Sigma_x\boldsymbol{\eta}\right]$. First we see that:

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{X},\boldsymbol{\eta}}\left[\boldsymbol{\eta}^\top\Sigma_x\boldsymbol{\eta}\right] &= \mathop{\mathbb{E}}_{\mathbf{X}}\left[\mathop{\mathbb{E}}_{\boldsymbol{\eta}}\left[\text{tr}(\boldsymbol{\eta}^\top\Sigma_x\boldsymbol{\eta}) \mid \mathbf{X}\right]\right] = \mathop{\mathbb{E}}_{\mathbf{X}}\left[\mathop{\mathbb{E}}_{\boldsymbol{\eta}}\left[\text{tr}(\Sigma_x\boldsymbol{\eta}\boldsymbol{\eta}^\top) \mid \mathbf{X}\right]\right] \\ &= \mathop{\mathbb{E}}_{\mathbf{X}}\left[\text{tr}\left(\Sigma_x \mathop{\mathbb{E}}_{\boldsymbol{\eta}}\left[\boldsymbol{\eta}\boldsymbol{\eta}^\top \mid \mathbf{X}\right]\right)\right].\end{aligned}$$

Let $\sigma_{\mathbf{X}} = \sigma_{\text{cdp}}$. We note that given any $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\boldsymbol{\eta}$ is a random Gaussian vector with each entry i.i.d. sampled from $\mathcal{N}(0, \sigma_{\mathbf{X}}^2)$. Therefore, its conditional covariance is $\mathbb{E}_{\boldsymbol{\eta}}\left[\boldsymbol{\eta}\boldsymbol{\eta}^\top \mid \mathbf{X}\right] = \sigma_{\mathbf{X}}^2 I_d$. Then, we have that

$$\mathop{\mathbb{E}}_{\mathbf{X},\boldsymbol{\eta}}\left[\boldsymbol{\eta}^\top\Sigma_x\boldsymbol{\eta}\right] = \mathop{\mathbb{E}}_{\mathbf{X}}[\text{tr}(\Sigma_x\cdot I_d\cdot\sigma_{\mathbf{X}}^2)] = \text{tr}(\Sigma_x)\cdot\mathop{\mathbb{E}}_{\mathbf{X}}[\sigma_{\mathbf{X}}^2]. \tag{8}$$

Next, we analyze the scale of $\sigma_{\mathbf{X}}$, and compute $\mathbb{E}_{\mathbf{X}}[\sigma_{\mathbf{X}}^2]$. Recall that from eq. (5), the ℓ_2-sensitivity of the output is

$$\Delta_2(\mathbf{X}) = \max_{i\in[n]}\|(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{x}_i\|_2.$$

This implies that

$$\Delta_2^2(\mathbf{X}) = \max_{i\in[n]}\mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-2}\mathbf{x}_i. \tag{9}$$

We write $\mathbf{x}_i = \Sigma_x^{1/2}\mathbf{u}_i$ where $\mathbf{u}_i \in \mathbb{R}^d$ is a random sub-Gaussian vector whose entries $\{u_{ij}\}_{j\in[d]}$ are i.i.d. sampled from a zero-mean, unit variance sub-Gaussian distribution with $\|\mathbf{u}_i\|_{\psi_2} \leq K$. Then, for each $i \in [n]$,

$$\mathbf{x}_i^\top\Sigma_x^{-2}\mathbf{x}_i = (\Sigma_x^{-1/2}\mathbf{x}_i)^\top\Sigma_x^{-1}(\Sigma_x^{-1/2}\mathbf{x}_i) = \mathbf{u}_i^\top\Sigma_x^{-1}\mathbf{u}_i \leq \frac{1}{\lambda_{\min}(\Sigma_x)}\cdot\|\mathbf{u}_i\|_2^2.$$

Let $Z = \max_{i\in[n]}\|\mathbf{u}_i\|_2^2$. Plugging into Lemma 38 and taking the union bound over all $i \in [n]$, we have for any $p \in (0,1)$ that with probability at least $1-p$

$$Z \leq d + CK^2\sqrt{d\ln(n/p)} + CK^2\ln(n/p)$$

for an absolute constant $C > 0$. By setting $t = \ln(n/p)$ (so $p = ne^{-t}$) and loosening the bound unconditionally via $\sqrt{dt} \leq d + t$, we express the upper tail of Z for any $t > 0$ as:

$$\mathbf{Pr}[Z \geq (1+CK^2)d + 2CK^2t] \leq ne^{-t}. \tag{10}$$

To compute $\mathbb{E}_{\mathbf{X}}[\Delta_2^2(\mathbf{X})]$, we cannot simply bound it unconditionally because the worst-case sensitivity scales with $1/\lambda^2$, which is exponentially large. Instead, we rigorously decompose the expectation by conditioning on the high-probability event $E = \{\frac{1}{n}(\mathbf{X}^\top\mathbf{X}) \succeq \frac{1}{2}\Sigma_x\}$. Recall that by Claim 37, $\mathbf{Pr}[\overline{E}] \leq e^{-cn/K^4}$. We note that:

$$\mathop{\mathbb{E}}_{\mathbf{X}}[\Delta_2^2(\mathbf{X})] = \mathop{\mathbb{E}}_{\mathbf{X}}[\Delta_2^2(\mathbf{X})\mathbf{1}_E] + \mathop{\mathbb{E}}_{\mathbf{X}}[\Delta_2^2(\mathbf{X})\mathbf{1}_{\overline{E}}].$$

First, conditioned on E, $(\mathbf{X}^\top\mathbf{X})^{-2} \preceq \frac{4}{n^2}\Sigma_x^{-2}$, which combined with eq. (9) guarantees $\Delta_2^2(\mathbf{X}) \leq \frac{4}{n^2\lambda_{\min}(\Sigma_x)}Z$. We compute $\mathbb{E}[Z]$ by formally integrating its tail probability from eq. (10):

$$\mathbb{E}[Z] = \int_0^\infty \mathbf{Pr}[Z \geq z]dz \leq ((1+CK^2)d + 2CK^2\ln n) + \int_{(1+CK^2)d+2CK^2\ln n}^\infty \mathbf{Pr}[Z \geq z]dz.$$

Using the change of variables $z = (1+CK^2)d + 2CK^2t$, the integral becomes:

$$\begin{aligned}\int_{\ln n}^\infty \mathbf{Pr}[Z \geq (1+CK^2)d + 2CK^2t] \cdot (2CK^2)dt &\leq \int_{\ln n}^\infty ne^{-t}(2CK^2)dt = 2CK^2n\left[-e^{-t}\right]_{\ln n}^\infty \\ &= 2CK^2.\end{aligned}$$

Thus, $\mathbb{E}[Z] \leq (1+CK^2)d + 2CK^2\ln n + 2CK^2 = O(K^2d + K^2\ln n)$. Consequently,

$$\mathop{\mathbb{E}}_{\mathbf{X}}[\Delta_2^2(\mathbf{X})\mathbf{1}_E] \leq \frac{4}{n^2\lambda_{\min}(\Sigma_x)}\mathbb{E}[Z] \leq O\left(\frac{K^2d + K^2\ln n}{n^2\lambda_{\min}(\Sigma_x)}\right).$$

For the case where E does not hold, we still have unconditionally $(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-2} \preceq \frac{1}{\lambda^2}I_d$, yielding $\Delta_2^2(\mathbf{X}) \leq \frac{1}{\lambda^2}\max_i\|\mathbf{x}_i\|_2^2 \leq \frac{\lambda_{\max}(\Sigma_x)}{\lambda^2}Z$. We define $P_{\max} = e^{-cn/K^4}$. Since $\mathbf{Pr}[\overline{E}] \leq P_{\max}$, we can bound $\mathbb{E}[Z\mathbf{1}_{\overline{E}}]$ by integrating the minimum of the tail of Z and the maximal probability $P_{\max}$:

$$\mathbb{E}[Z\mathbf{1}_{\overline{E}}] = \int_0^\infty \mathbf{Pr}[Z\mathbf{1}_{\overline{E}} \geq z]dz \leq \int_0^\infty \min(\mathbf{Pr}[Z \geq z], P_{\max})dz.$$

The tail bound $ne^{-t} \leq P_{\max}$ intersects exactly at $t^* = \ln(n/P_{\max}) = \ln n + cn/K^4$. This corresponds to $z^* = (1+CK^2)d + 2CK^2t^* = O(d + n/K^2)$. We split the integral at z^*:

$$\begin{aligned}\int_0^\infty \min(\mathbf{Pr}[Z \geq z], P_{\max})dz &\leq \int_0^{z^*} P_{\max}dz + \int_{z^*}^\infty \mathbf{Pr}[Z \geq z]dz \\ &\leq z^*P_{\max} + \int_{t^*}^\infty ne^{-t}(2CK^2)dt \\ &\leq z^*P_{\max} + 2CK^2P_{\max} = O(n/K^2)e^{-cn/K^4}.\end{aligned}$$

Therefore, $\mathbb{E}_{\mathbf{X}}[\Delta_2^2(\mathbf{X})\mathbf{1}_{\overline{E}}] \leq \frac{\lambda_{\max}(\Sigma_x)}{\lambda^2}O(n/K^2)e^{-cn/K^4}$. Crucially, given our regularizer choice $\lambda = \frac{n}{2}\sqrt[3]{\sigma_\zeta^2 d\lambda_{\max}\lambda_{\min}e^{-cn/K^4}}$, we have $\lambda^2 = \Theta(n^2e^{-2cn/3K^4})$. Thus, $\frac{1}{\lambda^2}e^{-cn/K^4} = \Theta(n^{-2}e^{-cn/3K^4})$. This makes the second term $O(n^{-1}e^{-cn/3K^4})$, which decays exponentially to zero and is strictly dominated by the $O(1/n^2)$ term from the well-conditioned event E.

Putting everything together, we have established:

$$\mathop{\mathbb{E}}_{\mathbf{X}}[\sigma_{\mathbf{X}}^2] = \mathop{\mathbb{E}}_{\mathbf{X}}\left[\frac{2\ln(1.25/\delta)}{\varepsilon^2} \cdot \Delta_2^2(\mathbf{X})\right] \leq O\left(\frac{\ln(1/\delta)(K^2d + K^2\ln n)}{\varepsilon^2n^2\lambda_{\min}(\Sigma_x)}\right).$$

Plug this back into eq. (8), we have that

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{X},\boldsymbol{\eta}}\left[\boldsymbol{\eta}^\top\Sigma_x\boldsymbol{\eta}\right] &= \mathrm{tr}(\Sigma_x) \cdot \mathop{\mathbb{E}}_{\mathbf{X}}[\sigma_{\mathbf{X}}^2] \leq d \cdot \lambda_{\max}(\Sigma_x) \cdot O\left(\frac{\ln(1/\delta)(K^2d + K^2\ln n)}{\varepsilon^2n^2\lambda_{\min}(\Sigma_x)}\right) \\ &\leq O\left(\kappa(\Sigma_x)\frac{K^2d\ln(1/\delta)(d+\ln n)}{\varepsilon^2n^2}\right). \qquad (11)\end{aligned}$$

Plugging eq. (7) and eq. (11) into eq. (6), we have that

$$\begin{aligned}R(\hat{\theta}_{\text{cdp}}) &= \mathbb{E}\left[\|\hat{\theta}_{\text{ridge}} - \theta^*\|_{\Sigma_x}^2\right] + \mathbb{E}\left[\boldsymbol{\eta}^\top\Sigma_x\boldsymbol{\eta}\right] \\ &\leq O\left(\frac{d\sigma_\zeta^2}{n}\right) + O\left(\kappa(\Sigma_x)\frac{K^2d\ln(1/\delta)(d+\ln n)}{\varepsilon^2n^2}\right) + \frac{e^{-cn/K^4}\lambda_{\max}^2}{\lambda_{\min}} + 3\left(\frac{\sigma_\zeta^4d^2\lambda_{\max}^2e^{-2cn/K^4}}{n^2\lambda_{\min}}\right)^{1/3}.\end{aligned}$$

Setting $M_\zeta = \frac{\sigma_\zeta^2 d}{n}$ and $\Lambda = \frac{\lambda_{\max}^2}{\lambda_{\min}}$ concludes the proof of Theorem 52. □

F.3 Hardness of Approximating Least Squares Estimate Under Label Differential Privacy

In this section, we show that the $O(d^2/(n^2\varepsilon^2))$ privacy penalty achieved by the trivial output perturbation scheme (Algorithm F.2) is asymptotically optimal up to logarithmic factors when the data is sampled from an isotropic sub-Gaussian distribution.

Theorem 5. (Optimality of Output Perturbation for Isotropic sub-Gaussian). *Assume the public feature matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ has rows drawn i.i.d. from any zero-mean, isotropic sub-Gaussian distribution, where $n \gg d$. Let $\mathbf{y} \in \mathbb{R}^n$ be the private labels. For any fixed $\varepsilon \in (0,1)$ and sufficiently small $\delta > 0$, let $\mathcal{M}$ be any (ε,δ)-label DP mechanism outputting an estimate $\hat{\theta}$. Suppose $\max_{i\in[n]} \|\mathbf{x}_i\|_{\psi_2} \le K$. Then, with probability at least $1 - 2\exp(-cn/K^4)$ over the randomness of $\mathbf{X}$ for some constant $c > 0$, the expected privacy error (over the randomness of the mechanism $\mathcal{M}$) is lower bounded by:*

$$\max_{\mathbf{y}\in\mathbb{R}} \mathop{\mathbb{E}}_{\mathcal{M}} \left[\|\mathcal{M}(\mathbf{y}) - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}\|_2^2\right] = \widetilde{\Omega}\left(\frac{d^2}{n^2\varepsilon^2}\right),$$

Proof. For a fixed feature matrix $\mathbf{X}$, it is easy to verify that computing the least squares estimate $\hat{\theta}_{\text{lse}}$ is exactly equivalent to answering a set of linear queries (defined by the public workload matrix $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \in \mathbb{R}^{d\times n}$) on $\mathbf{y}$. Further, we recall that for any neighboring label lists $\mathbf{y}$ and $\mathbf{y}'$, $\|\mathbf{y} - \mathbf{y}'\|_1 \le 1$, which aligns with the privacy notion in [45].

By Theorem 20, for any (ε,δ)-DP mechanism $\mathcal{M}$, the worst-case expected ℓ_2^2 error is lower bounded by the Schatten-1 norm ($\|\mathbf{A}\|_1$) of the query matrix $\mathbf{A}$:

$$\max_{\mathbf{y}} \mathop{\mathbb{E}}_{\mathcal{M}} \left[\|\mathcal{M}(\mathbf{y}) - \mathbf{A}\mathbf{y}\|_2^2\right] = \widetilde{\Omega}\left(\frac{\|\mathbf{A}\|_1^2}{n\varepsilon^2}\right). \tag{12}$$

It remains to lower bound the Schatten-1 norm $\|\mathbf{A}\|_1 = \sum_{i=1}^d s_i(\mathbf{A})$, where $s_i(\mathbf{A})$ are the singular values of $\mathbf{A}$. Observe that $\mathbf{A}\mathbf{A}^\top = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = (\mathbf{X}^\top\mathbf{X})^{-1}$. Therefore, the singular values of $\mathbf{A}$ are exactly the square roots of the eigenvalues of $(\mathbf{X}^\top\mathbf{X})^{-1}$, which means:

$$s_i(\mathbf{A}) = \frac{1}{s_i(\mathbf{X})} \quad \text{for all } i \in [d],$$

where $s_i(\mathbf{X})$ are the singular values of the sub-Gaussian matrix $\mathbf{X}$. By the concentration inequality (Proposition 33) for the maximum singular value of a sub-Gaussian matrix $\mathbf{X}$, we have that for any $t > 0$:

$$\mathbf{Pr}\left[s_{\max}(\mathbf{X}) > \sqrt{n} + CK^2(\sqrt{d} + t)\right] \le 2\exp(-t^2).$$

Setting $t = \sqrt{cn}/K^2$ for some constant c, we have with probability at least $1 - 2\exp(-cn/K^4)$ that $s_{\max}(\mathbf{X}) \le 2\sqrt{n} + CK^2\sqrt{d}$. Since $n \gg d$, there exists a universal constant $c' > 0$ such that $s_{\max}(\mathbf{X}) \le c'\sqrt{n}$. Conditioned on this high-probability event, every singular value of $\mathbf{X}$ is bounded above by $c'\sqrt{n}$, which strictly bounds the singular values of $\mathbf{A}$ from below:

$$s_i(\mathbf{A}) \ge \frac{1}{c'\sqrt{n}} \quad \text{for all } i \in [d].$$

Consequently, the Schatten-1 norm of $\mathbf{A}$ is lower bounded by:

$$\|\mathbf{A}\|_1 = \sum_{i=1}^d s_i(\mathbf{A}) \ge \sum_{i=1}^d \frac{1}{c'\sqrt{n}} = \Omega\left(\frac{d}{\sqrt{n}}\right). \tag{13}$$

Substituting eq. (13) back into the lower bound in eq. (12), we obtain:

$$\max_{\mathbf{y}} \mathop{\mathbb{E}}_{\mathcal{M}} \left[\|\mathcal{M}(\mathbf{y}) - \hat{\theta}_{\text{lse}}\|_2^2\right] \ge \widetilde{\Omega}\left(\frac{(d/\sqrt{n})^2}{n\varepsilon^2}\right) = \widetilde{\Omega}\left(\frac{d^2}{n^2\varepsilon^2}\right).$$

This completes the proof of Theorem 5. □

F.4 Central DP II: An Encoder-Decoder Framework (Matrix Mechanism)

Building upon previous warm-up results, we now introduce a more sophisticated approach for the central label-DP model using the *Matrix Mechanism* framework [26, 18]. In standard output perturbation, calibrated noise is added directly to the final estimate. In contrast, the matrix mechanism factorizes the linear regression operation into an encoder-decoder pipeline, injecting noise into a carefully chosen latent space to optimize the overall error.

Recall that the exact ridge regression estimate can be formulated as a linear query $\hat{\theta}_{\text{ridge}} = \mathbf{M}\mathbf{y}$, where the query matrix is $\mathbf{M} = (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top$. The matrix mechanism factorizes this operation into $\mathbf{M} = \mathbf{L}\mathbf{R}$. Here, $\mathbf{R}$ serves as an *encoder* that linearly pre-processes the original labels $\mathbf{y}$ into a spherical latent space. We then inject a calibrated Gaussian noise vector η directly into this latent space. Finally, the matrix $\mathbf{L}$ acts as a *decoder*, projecting the noisy latent vector back into the original d-dimensional parameter space. Specifically, to enhance numerical stability and optimally bound the sensitivity, we focus on the symmetric factorization based on the regularized covariance matrix:

- **Encoder R:** $\mathbf{R} = (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}\mathbf{X}^\top$.
- **Decoder L:** $\mathbf{L} = (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}$.

Similar to our analysis in Section F.2, since the data matrix $\mathbf{X}$ is public in the label-DP setting, we can exactly compute the ℓ_2-sensitivity of the encoded query $\mathbf{R}\mathbf{y}$ with respect to a single label change (assuming $\|\mathbf{y} - \mathbf{y}'\|_1 \le 1$):

$$\Delta_2(\mathbf{R}) = \max_{\|\mathbf{v}\|_1 \le 1} \|\mathbf{R}\mathbf{v}\|_2 = \|\mathbf{R}\|_{1\to 2} = \max_{i\in[n]} \|(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}\mathbf{x}_i\|_2.$$

Here, $\|\mathbf{R}\|_{1\to 2} = \max \frac{\|Rx\|_2}{\|x\|_1}$ is the maximum ℓ_2-norm among all columns of $\mathbf{R}$. Notice that the matrix mechanism effectively replaces the inverse matrix $(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}$ from the output perturbation sensitivity with its square root $(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}$. This structural shift significantly mitigates the noise inflation caused by the condition number of $\mathbf{X}^\top\mathbf{X}$. The full procedure is summarized in Algorithm F.4. We show the following result:

Theorem 53. *Fix any $0 < \varepsilon, \delta < 1$. Let $\hat{\theta}_{mm}$ be the output of Algorithm F.4 on input n i.i.d. samples from the Gaussian linear model with variance σ_ζ^2. Suppose that the feature distribution satisfies properties (1), (2). Then, we have that there exists a proper choice of $\lambda > 0$ and a universal constant $\tilde{c} > 0$ and that:*

$$R(\hat{\theta}_{mm}) \le O\left(\frac{d\sigma_\zeta^2}{n}\right) + O\left(\frac{d^2 \log n \log(1/\delta)}{n^2\varepsilon^2}\right) + e^{-\frac{2cn}{3K^4}}\left[3\left(M_\zeta^2\Lambda\right)^{1/3} + \Lambda\right].$$

for $n \gg d$, where $M_\zeta = \frac{\sigma_\zeta^2 d}{n}$ and $\Lambda = \frac{\lambda_{\max}^2}{\lambda_{\min}}$.

Algorithm F.4: Label-DP Regression via Matrix Mechanism

1. **Input:** A dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, privacy parameters $\varepsilon, \delta \in (0,1)$, and a ridge regularizer $\lambda > 0$.
2. Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ be the vector of clean labels, and $\mathbf{X} \in \mathbb{R}^{n\times d}$ be the data matrix where the i-th row is $\mathbf{x}_i^\top$.
3. Compute the encoder $\mathbf{R}$ and the decoder $\mathbf{L}$:
$$\mathbf{R} = (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}\mathbf{X}^\top, \quad \mathbf{L} = (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}.$$
4. Compute the exact data-dependent ℓ_2-sensitivity of the encoded labels:
$$\Delta_2(\mathbf{R}) = \max_{i\in[n]} \|(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}\mathbf{x}_i\|_2.$$
5. Set the latent noise scale $\sigma_{\text{matrix}} = \frac{\Delta_2(\mathbf{R})\sqrt{2\ln(1.25/\delta)}}{\varepsilon}$.
6. Sample a d-dimensional Gaussian noise vector $\eta \sim \mathcal{N}(0, \sigma_{\text{matrix}}^2 I_d)$.
7. **Output:** The privatized estimate decoded from the noisy latent space:
$$\hat{\theta}_{\text{matrix}} = \mathbf{L}(\mathbf{R}\mathbf{y} + \eta).$$

Proof. Let $\boldsymbol{\eta}_R \in \mathbb{R}^d$ be the random noise used to privatize $\mathbf{R}\mathbf{y}$. The final estimate is decoded as:
$$\begin{aligned}\hat{\theta}_{\text{mm}} = \mathbf{L}(\mathbf{R}\mathbf{y} + \boldsymbol{\eta}_R) &= (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}((\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}\mathbf{X}^\top\mathbf{y} + \boldsymbol{\eta}_R) \\ &= \underbrace{(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top\mathbf{y}}_{\hat{\theta}_{\text{ridge}}} + (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}\boldsymbol{\eta}_R.\end{aligned}$$
Let Δ_R be the ℓ_2-sensitivity of the encoded labels $\mathbf{R}\mathbf{y}$, then
$$\Delta_R = \max_{\mathbf{y}'\sim\mathbf{y}} \|(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}\mathbf{X}^\top(\mathbf{y} - \mathbf{y}')\|_2 = \max_{i\in[n]} \|(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}\mathbf{x}_i\|_2.$$
Write $\mathbf{r}_i = (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}\mathbf{x}_i$, and we have
$$\|\mathbf{r}_i\|_2^2 = \mathbf{r}_i^\top\mathbf{r}_i = \mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}\mathbf{x}_i = \mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{x}_i.$$
Next, we upper bound $h_i = \|\mathbf{r}_i\|_2^2$ for each $i \in [n]$. We observe from Proposition 41 that $\max_{i\in[n]} h_i \le 1$. While this unconditional upper bound is crucial for handling worst-case expectations, we first show that, with high probability, h_i is significantly smaller than $O(1)$.

Recall that each $\mathbf{x}_i$ ($i \in [n]$) is a sub-Gaussian vector such that for any $\mathbf{v} \in \mathbb{R}^d$, $\|\langle\mathbf{x}_i, \mathbf{v}\rangle\|_{\psi_2} \le K\|\Sigma_x^{1/2}\mathbf{v}\|_2$. Thus, we can parameterize $\mathbf{x}_i = \Sigma_x^{1/2}\mathbf{u}_i$ where $\mathbf{u}_i$ is isotropic (i.e., $\mathbb{E}[\mathbf{u}_i\mathbf{u}_i^\top] = I_d$) and $\|\mathbf{u}_i\|_{\psi_2} \le K$. From Claim 37, with probability at least $1 - e^{-cn/K^4}$ for some constant $c > 0$, we have
$$\frac{1}{n}\left(\mathbf{X}^\top\mathbf{X} + \lambda I_d\right) \succeq \frac{1}{n}\mathbf{X}^\top\mathbf{X} \succeq \frac{1}{2}\Sigma_x.$$
Conditioned on this event, we have:
$$\begin{aligned}h_i = \mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{x}_i &\le \mathbf{x}_i^\top\left(\frac{2}{n}\Sigma_x^{-1}\right)\mathbf{x}_i = (\Sigma_x^{1/2}\mathbf{u}_i)^\top\left(\frac{1}{n}\Sigma_x^{-1}\right)(\Sigma_x^{1/2}\mathbf{u}_i) \\ &= \frac{2}{n}\mathbf{u}_i^\top\Sigma_x^{1/2}\Sigma_x^{-1}\Sigma_x^{1/2}\mathbf{u}_i = \frac{2}{n}\mathbf{u}_i^\top I_d\mathbf{u}_i = \frac{2}{n}\|\mathbf{u}_i\|_2^2.\end{aligned}$$
By applying Lemma 38 and taking the union bound over all $i \in [n]$, we obtain that with probability at least $1 - p$ for any $0 < p < 1$ and some constant $C > 0$,
$$\Delta_R^2 = \max_{i\in[n]} h_i \le \frac{d + CK^2\sqrt{d\ln(n/p)} + CK^2\ln(n/p)}{n} = O\left(\frac{d\log n}{n}\right) \tag{14}$$

by setting $p = n^{-c'}$ for an arbitrarily large constant $c' > 0$. Following the Gaussian Mechanism, it is sufficient to sample each entry of $\boldsymbol{\eta}_R$ from $\mathcal{N}(0, \sigma_R^2)$ where

$$\sigma_R^2 = O\left(\frac{\Delta_R^2 \log(1/\delta)}{\varepsilon^2}\right) = O\left(\frac{d \log n \log(1/\delta)}{n\varepsilon^2}\right).$$

Finally, we decompose the expected prediction risk:

$$\begin{aligned}
R(\hat{\theta}_{\text{mm}}) &= \mathbb{E}[\|\hat{\theta}_{\text{mm}} - \theta^*\|_{\Sigma_x}^2] = \mathbb{E}\left[(\hat{\theta}_{\text{mm}} - \theta^*)^\top \Sigma_x (\hat{\theta}_{\text{mm}} - \theta^*)\right] \\
&= \mathbb{E}\left[(\hat{\theta}_{\text{ridge}} - \theta^* + \mathbf{L}\boldsymbol{\eta}_R)^\top \Sigma_x (\hat{\theta}_{\text{ridge}} - \theta^* + \mathbf{L}\boldsymbol{\eta}_R)\right] \\
&= \mathbb{E}\left[(\hat{\theta}_{\text{ridge}} - \theta^*)^\top \Sigma_x (\hat{\theta}_{\text{ridge}} - \theta^*)\right] + \mathbb{E}\left[(\mathbf{L}\boldsymbol{\eta}_R)^\top \Sigma_x (\mathbf{L}\boldsymbol{\eta}_R)\right] \\
&\quad + 2\,\mathbb{E}\left[(\hat{\theta}_{\text{ridge}} - \theta^*)^\top \Sigma_x (\mathbf{L}\boldsymbol{\eta}_R)\right] \\
&= \mathbb{E}\left[(\hat{\theta}_{\text{ridge}} - \theta^*)^\top \Sigma_x (\hat{\theta}_{\text{ridge}} - \theta^*)\right] + \mathbb{E}\left[(\mathbf{L}\boldsymbol{\eta}_R)^\top \Sigma_x (\mathbf{L}\boldsymbol{\eta}_R)\right].
\end{aligned}$$

The cross term vanishes because $\mathbb{E}[\boldsymbol{\eta}_R] = \mathbf{0}$ and $\boldsymbol{\eta}_R$ is independent of $\mathbf{X}$ and $\mathbf{y}$. The first term is bounded by Lemma 49. For the privacy error, we compute its expectation by conditioning on $\mathbf{X}$:

$$\begin{aligned}
\mathbb{E}\left[(\mathbf{L}\boldsymbol{\eta}_R)^\top \Sigma_x (\mathbf{L}\boldsymbol{\eta}_R)\right] &= \mathop{\mathbb{E}}_{\mathbf{X},\boldsymbol{\eta}_R}\left[\boldsymbol{\eta}_R^\top (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2} \Sigma_x (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2} \boldsymbol{\eta}_R\right] \\
&= \mathop{\mathbb{E}}_{\mathbf{X},\boldsymbol{\eta}_R}\left[\text{tr}\left(\boldsymbol{\eta}_R\boldsymbol{\eta}_R^\top (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2} \Sigma_x (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathbf{X}}\left[\sigma_R^2 \cdot \text{tr}\left((\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2} \Sigma_x (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1/2}\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathbf{X}}\left[\sigma_R^2 \cdot \text{tr}\left(\Sigma_x (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\right)\right] \\
&= \text{tr}\left(\Sigma_x \cdot \mathop{\mathbb{E}}_{\mathbf{X}}\left[\sigma_R^2 (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\right]\right).
\end{aligned}$$

Let $E_1 = \left\{\mathbf{X} : \frac{1}{n}(\mathbf{X}^\top\mathbf{X} + \lambda I_d) \succeq \frac{1}{2}\Sigma_x\right\}$ be the covariance concentration event, and E_2 be the event that eq.(14) holds. By Claim 37 we have $\mathbf{Pr}[\overline{E_1}] \leq e^{-cn/K^4}$ and by Lemma 38 $\mathbf{Pr}[\overline{E_2} \mid E_1] \leq n^{-c'}$. Let $E' = E_1 \cap E_2$ be the joint good event over the randomness of $\mathbf{X}$. By the definition of conditional probability, we have:

$$\begin{aligned}
\mathbf{Pr}[E'] &= \mathbf{Pr}[E_1] \cdot \mathbf{Pr}[E_2 \mid E_1] \\
&\geq (1 - e^{-cn/K^4})(1 - n^{-c'}) \\
&> 1 - e^{-cn/K^4} - n^{-c'}.
\end{aligned} \tag{15}$$

Assuming K is a constant and n is sufficiently large, we can choose the constant c' to be sufficiently large such that $\mathbf{Pr}[E'] \geq 1 - n^{-100}$. Under this explicitly defined high-probability event E', we have established that the required noise scale is upper bounded by:

$$\sigma_R^2 = O\left(\frac{d \log n \log(1/\delta)}{n\varepsilon^2}\right).$$

Crucially, due to Proposition 41 we have $\max_{i\in[n]} h_i \leq 1$, providing the unconditional bound:

$$\sigma_R^2 = O\left(\frac{\Delta_R^2 \log(1/\delta)}{\varepsilon^2}\right) = O\left(\frac{(\max_{i\in[n]} h_i) \log(1/\delta)}{\varepsilon^2}\right) = O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right).$$

Using the indicator function decomposition $\mathbf{1}_{E'} + \mathbf{1}_{\overline{E'}} = 1$, we bound the expectation:

$$\begin{aligned}
\mathbb{E}\left[(\mathbf{L}\boldsymbol{\eta}_R)^\top \Sigma_x (\mathbf{L}\boldsymbol{\eta}_R)\right] &= \text{tr}\left(\Sigma_x \cdot \mathop{\mathbb{E}}_{\mathbf{X}}\left[\sigma_R^2 (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}(\mathbf{1}_{E'} + \mathbf{1}_{\overline{E'}})\right]\right) \\
&\le \text{tr}\left(\Sigma_x \cdot \left(O\left(\frac{d\log n \log(1/\delta)}{n\varepsilon^2}\right)\mathop{\mathbb{E}}_{\mathbf{X}}\left[(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\right]\right.\right. \\
&\quad \left.\left. + \mathbf{Pr}[\overline{E'}] \cdot O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)\frac{1}{\lambda} I_d\right)\right) \\
&= O\left(\frac{d\log n\log(1/\delta)}{n\varepsilon^2}\right) \cdot \text{tr}\left(\Sigma_x \cdot \mathop{\mathbb{E}}_{\mathbf{X}}\left[(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\right]\right) + O\left(\frac{d\log(1/\delta)}{n^{100}\lambda\varepsilon^2}\right) \\
&= O\left(\frac{d\log n\log(1/\delta)}{n\varepsilon^2}\right) \cdot \text{tr}\left(\Sigma_x \cdot \left(\frac{2}{n}\Sigma_x^{-1} + \frac{e^{-cn/K^4}}{\lambda} I_d\right)\right) \\
&= O\left(\frac{d^2\log n\log(1/\delta)}{n^2\varepsilon^2}\right) + e^{-cn/K^4} \cdot O\left(\frac{d^2\lambda_{\max}\log n\log(1/\delta)}{\lambda n\varepsilon^2}\right)
\end{aligned} \tag{16}$$

for $n \gg d$. Here, the extremely small probability $\mathbf{Pr}[\overline{E'}] \le n^{-100}$ strictly dominates the unconditional $1/\lambda$ bound, allowing the $\overline{E'}$ term to be absorbed. The penultimate inequality relies on Proposition 39 and Lemma 40. Finally, combining Lemma 49 and eq. (16), we have:

$$\begin{aligned}
R(\hat{\theta}_{\text{mm}}) &= \mathbb{E}\left[(\hat{\theta}_{\text{ridge}} - \theta^*)^\top \Sigma_x (\hat{\theta}_{\text{ridge}} - \theta^*)\right] + \mathbb{E}\left[(\mathbf{L}\boldsymbol{\eta}_R)^\top \Sigma_x (\mathbf{L}\boldsymbol{\eta}_R)\right] \\
&\le \frac{2d\sigma_\zeta^2}{n} + \frac{4\lambda^2}{n^2\lambda_{\min}} + e^{-cn/K^4}\left(\frac{\sigma_\zeta^2 d\lambda_{\max}}{\lambda} + \frac{\lambda_{\max}^2}{\lambda_{\min}}\right) + O\left(\frac{d^2\log n\log(1/\delta)}{n^2\varepsilon^2}\right).
\end{aligned}$$

Setting

$$\lambda = \frac{n^{2/3}}{2} e^{-cn/(3K^4)}\left(\lambda_{\min}\lambda_{\max}\sigma_\zeta^2 d\right)^{1/3},$$

we obtain:

$$R(\hat{\theta}_{\text{mm}}) \le \frac{2d\sigma_\zeta^2}{n} + O\left(\frac{d^2\log n\log(1/\delta)}{n^2\varepsilon^2}\right) + e^{-\frac{2cn}{3K^4}}\left[3\left(\frac{\sigma_\zeta^4 d^2\lambda_{\max}^2}{n^2\lambda_{\min}}\right)^{1/3} + \frac{\lambda_{\max}^2}{\lambda_{\min}}\right],$$

which completes the proof of Theorem 53 by substituting $M_\zeta = \frac{\sigma_\zeta^2 d}{n}$ and $\Lambda = \frac{\lambda_{\max}^2}{\lambda_{\min}}$. □

G Empirical Studies for Local versus Central DP in Classification Tasks

Here we conduct an empirical simulation on a binary classification task with synthetic datasets. The synthetic data is generated from a d-dimensional Gaussian distribution where each data entry is sampled from $\mathbf{x}_i \sim \mathcal{N}(0, I_d)$, with labels perfectly separated by a randomly initialized target hyperplane, $y_i = \text{sign}(\mathbf{x}_i^\top \theta^*)$, where $\theta^* \sim \mathcal{N}(0, I_d)$. We set the feature dimension to $d = 100$ and utilize a large training set of $50,000$ samples to ensure stable convergence. To accurately measure the excess risk, we calculate the zero-one classification error on a fixed, independent test set consisting of $5,000$ samples. Recall that we define the empirical risk as $\hat{R}(h) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(h(\mathbf{x}_i) \neq y_i)$. To minimize the empirical risk, we train an unregularized Logistic Regression classifier using the cross-entropy loss optimized via the L-BFGS algorithm [43].

G.1 Empirical Risk with Increasing Privacy Budget

The first experiment visualizes in Figure 3 the performance of the standard local label-DP (randomized response) versus the central label-DP (Algorithm 1) across varying privacy budgets $\varepsilon \in [0.05, 0.5]$. To provide a theoretical baseline, reference curves representing the corresponding asymptotic error bounds (given in Theorem 46 for local-DP and Theorem 4 for central-DP, respectively) are plotted alongside the empirical data with the leading constant picked so that the reference curve and empirical curve intersect at some point to be aligned with the definition of asymptotic notation (here the point is

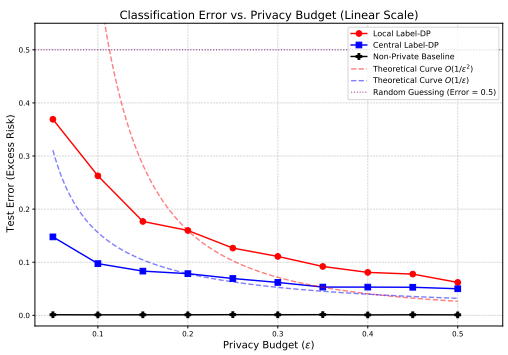

Figure 3: Empirical risk with different privacy budgets.

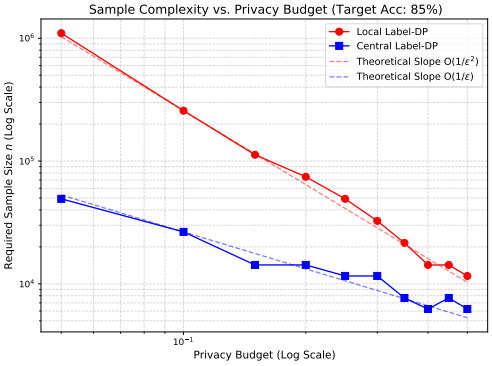

Figure 4: Sample complexity with different privacy budgets.

$\varepsilon = 0.2$). The results align with the theoretical implications of Theorem 4: as the privacy constraints tighten (i.e., smaller ε), the error of the local DP mechanism degrades rapidly, approaching random guessing. In contrast, the central DP mechanism exhibits a much more slower degradation and maintains a lower error rate. This empirical gap validates that by leveraging interaction and privacy amplification via subsampling, the central model significantly mitigates the privacy penalty, reflecting the improved theoretical dependence on ε established in our analysis.

G.2 Sample Complexity with Increasing Privacy Budget

To more explicitly illustrate the accuracy gap between the two settings (local and central label-DP) with respect to the privacy budget ε, we next focus on sample complexity. Instead of fixing the sample size, we established a target test accuracy of 85% and utilized a logarithmic grid search (ranging from $n = 10^2$ to $10^{5.5}$) to determine the minimum number of training samples required for each $\varepsilon \in [0.05, 1.0]$. The results are visualized in Figure 4 using a log-log plot to show the polynomial scaling rates of the two algorithms. We plot two theoretical reference lines: a slope of -2 representing the $\tilde{O}(\sqrt{d/(\varepsilon^2 n)})$ bounds of the local DP mechanism, and a slope of -1 corresponding to the improved $\tilde{O}(\sqrt{d/(\varepsilon n)})$ convergence rate of the central DP approach. The empirical trajectories tightly align with these theoretical slopes. As the privacy budget ε decreases, the required sample size for local DP grows quadratically while that of the central DP shows linear increase.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract summarizes the main contributions and their scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We make clear assumption for each theorem especially for the analysis in the regression sections.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide a full proof with clear stated assumptions of each result.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We give detailed instructions on the implementation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Provided in the supplemental material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We stated details in the setting of experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The error metric corresponds to the one used in theoretical results, and we do repetitive tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We stated the computer resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: We present a theory work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We present a theory work.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: We present a theory work.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We present a theory work.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We present a theory work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We present a theory work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: For full disclaimer we use LLM to check on typos and improve presentation.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.