
The Gaussian Mixing Mechanism: Rényi Differential Privacy via Gaussian Sketches

Omri Lev^{1*} Vishwak Srinivasan^{1,†} Moshe Shenfeld^{2,†}

Katrina Ligett² Ayush Sekhari³ Ashia C. Wilson¹

¹ Massachusetts Institute of Technology

² The Hebrew University of Jerusalem ³ Boston University

Abstract

Gaussian sketching, which consists of pre-multiplying the data with a random Gaussian matrix, is a widely used technique in data science and machine learning. Beyond computational benefits, this operation also provides differential privacy guarantees due to its inherent randomness. In this work, we revisit this operation through the lens of Rényi Differential Privacy (RDP), providing a refined privacy analysis that yields significantly tighter bounds than prior results. We then demonstrate how this improved analysis leads to performance improvement in different linear regression settings, establishing theoretical utility guarantees. Empirically, our methods improve performance across multiple datasets and, in several cases, reduce runtime.

1 Introduction

Advances in data collection and storage have enabled the creation of massive datasets, which have necessitated the design of scalable algorithms for efficient processing, analysis, and inference. With the formation of such large datasets, there has also been an increased focus on privacy-preserving analyses in a bid to protect private attributes [Apple Research, 2017, Facebook Research, 2020, Snap Security, 2022, Ponomareva et al., 2023]. For instance, census data consists of a variety of private information about individuals recorded, and the United States Census Bureau has been taking various measures to ensure confidentiality in the data, and has adopted more modern methods since the late 1900s [Bureau, 2019].

A popular mathematical notion of privacy is given by *Differential Privacy* (DP) [Dwork, 2006, McSherry and Talwar, 2007, Dwork et al., 2014], which is currently the de facto standard for privacy-preserving mechanisms. A private mechanism is a (randomized) function $\mathcal{M}$ which takes as input a dataset X , and the key challenge in designing a “good” private mechanism is balancing the tradeoff between the loss in utility with the increase in privacy-preserving nature of the algorithm; it is possible to obtain complete privacy by simply returning noise regardless of the input, but this clearly does not have good utility.

With regard to designing scalable algorithms to handle massive datasets, *sketching* as a technique [Sarlos, 2006] has led to more computationally scalable algorithms relative to naively working with the full dataset; see Woodruff [2014] for a review. Given a data matrix $X \in \mathbb{R}^{n \times d}$, a *sketch* constructs a compressed representation $SX \in \mathbb{R}^{k \times d}$ with $k \ll n$, where $S \in \mathbb{R}^{k \times n}$ is a random matrix. For the case where S is comprised of i.i.d. Gaussian elements,

*Correspondence: omrilev@mit.edu. Code: <https://github.com/omrilev1/GaussMix>.

†Equal contribution.

for sufficiently large k it holds that $\frac{1}{k}S^\top S \approx \mathbb{I}_n$ with high probability. This property is particularly valuable in machine learning applications that rely on inner-products of the form $A^\top B$ for matrices A and B , as one could now apply a Gaussian sketching matrix S to both A and B and largely preserve this Hilbert-Schmidt inner product (up to scaling constants). In this work, we study a mechanism based on a “noisy” sketching operation, which we call the *Gaussian Mixing Mechanism* (abbrev. **GaussMix**) and is defined by

$$\mathcal{M}(X) := SX + \sigma\xi, \quad \text{with } S \sim \mathcal{N}(0, \mathbb{I}_{k \times n}), \xi \sim \mathcal{N}(0, \mathbb{I}_{k \times d}), S \perp \xi \quad (\text{GaussMix})$$

The additive Gaussian noise (hence the “noisy” sketching operation) essentially contributes a constant bias to the inner product between the outputs of **GaussMix** applied to A and B , as for sufficiently large k and appealing to standard concentration inequalities, we have

$$\frac{1}{k}(SA + \sigma\xi)^\top (SB + \sigma\xi) \approx A^\top B + \sigma^2 \mathbb{I}_d. \quad (1)$$

Thus, this noisy sketch is potentially well-suited for applications involving inner products. The intriguing aspect of **GaussMix** is that the Gaussian sketching operation contributes to stronger privacy guarantees when compared to the standard Gaussian mechanism.

The first evidence of sketching yielding a differentially private mechanism was given by [Blocki et al. \[2012\]](#), and later works [\[Sheffet, 2017, 2019\]](#) focused on establishing that **GaussMix** also provides DP. While it might seem surprising that **GaussMix** is a differentially private mechanism, the operation can be viewed as instantiating the classical *Gaussian mechanism* for DP [\[Dwork et al., 2014, Appendix A\]](#) for the Gaussian sketched data matrix SX .

Intuition for **GaussMix.** Interestingly, certain key operations in augmenting mechanisms for improved differential privacy can be expressed as left-multiplication operations. Two such instances are (1) permuting the rows of the data matrix X randomly [\[Erlingsson et al., 2019\]](#), which can be expressed as left-multiplication of X by a permutation matrix, and (2) subsampling the data matrix at random, which can be expressed as left-multiplication of X by a random matrix whose entries are either 0 or 1 [\[Kasiviswanathan et al., 2011, Balle et al., 2018\]](#). A rough intuition for why sketching is a reasonable operation for preserving privacy is that SX generates random linear combinations of rows, which can potentially hide the contribution of any one row. Differential privacy—a more rigorous notion—seeks to ensure that the presence of any one particular row in the data can not be guessed well, even by an adversary that has knowledge of the other rows. Note that sketching by itself is not impervious to such a setting: consider a data matrix X where all but one row contain 0s, and in this case, it is possible to make educated guesses for what the non-zero row might be from the sketched data matrix SX . This motivates the inclusion of the noise addition in **GaussMix**, which also covers this adversarial situation. Informally speaking, through our analysis, we observe that the “richness” of the data matrix (quantified by the minimum singular value) contributes to the success of the sketching mechanism as a differentially private mechanism.

Contributions. Expanding on the above, our primary contributions, which are both theoretical and empirical, are summarized below.

- **Tighter Privacy Analysis via Rényi Differential Privacy (RDP).** We present a new RDP analysis of **GaussMix**, which to the best of our knowledge, has not been previously explored. Our bounds are simpler to derive than existing analyses and in several cases, are tight with respect to RDP. Notably, our results show that **GaussMix** achieves stronger privacy guarantees than those established by [Sheffet \[2019\]](#) for the same mechanism parameters. Our improvement extend to settings that rely on **GaussMix**, such as [Prakash et al. \[2020\]](#), [Anand et al. \[2021\]](#), [Sun et al. \[2022\]](#), [Bartan and Pilanci \[2023\]](#).
- **Algorithm for Private Ordinary Least Squares.** We use **GaussMix** to develop a differentially private algorithm for linear regression under bounded data, extending the framework of [Sheffet \[2017\]](#). Leveraging our new RDP analysis, we derive bounds on the excess empirical risk that, in some cases, match those of the AdaSSP algorithm by [Wang \[2018\]](#), a standard benchmark under domain-bound assumptions. We validate these theoretical results through empirical evaluations, where our method consistently outperforms the baselines of [Sheffet \[2017\]](#) and [Wang \[2018\]](#) across several benchmark

datasets. Furthermore, we show that retrofitting the improved RDP analysis into the algorithm of Sheffet [2017] yields improved results, further demonstrating the broad applicability of our improved analysis.

- **Algorithm for Private Logistic Regression.** We adapt the algorithm for private linear regression to perform differentially private logistic regression. This specifically works by using a second-order approximation of the loss function [Huggins et al., 2017, Ferrando and Sheldon, 2025] that reduces the problem to a differentially private quadratic minimization, which makes it amenable to apply GaussMix. We derive theoretical guarantees for our proposed method and give numerical simulations over certain datasets that demonstrate improvements in both accuracy and computation time over the commonly used objective perturbation [Chaudhuri et al., 2011] and DP-SGD [Abadi et al., 2016] baselines.

Organization of the paper. In Section 2, we discuss related work. Section 3 introduces necessary preliminaries. We present the Gaussian Mixing Mechanism and our privacy analysis in Section 4, including comparisons with prior bounds. Applications of our mechanism to DP OLS (Section 5.1) and DP logistic regression (Section 5.2) are given in Section 5, supported by theoretical and empirical evaluations.

2 Related Work

A substantial body of research has investigated the use of random matrix projections for privacy, particularly through the Johnson–Lindenstrauss (JL) transform and its variants [Blocki et al., 2012, Kenthapadi et al., 2013, Sheffet, 2017, Showkatbakhsh et al., 2018, Sheffet, 2019]. Sheffet [2015, 2017, 2019] and Showkatbakhsh et al. [2018] are the most relevant to our work, as they propose using Gaussian sketches for private linear regression. Sheffet [2015, 2017, 2019] analyze the privacy-preserving characteristics of GaussMix for this problem, and show that it achieves (ϵ, δ) -differential privacy for certain settings of σ, k . Showkatbakhsh et al. [2018] study the same mechanism but under a modified notion of differential privacy known as MI-DP (defined by Cuff and Yu [2016]).

The mathematical intuition for why sketching is useful in privacy-sensitive optimization was studied by Pilanci and Wainwright [2015]. They observe that the mutual information between X and its sketched version SX can not be too large, thus providing a form of privacy. However, their analysis is centered around information-theoretic principles and quantities such as the mutual information, and does not actually provide guarantees of differential privacy. More recently, Bartan and Pilanci [2023] provides another application of Gaussian sketching in the specific context of DP distributed linear regression, based on the results obtained by Sheffet [2015].

In contrast to these prior approaches, our work uses the stronger RDP framework of Mironov [2017] in its privacy analysis. We derive a simple, closed-form expression for the RDP curve $\epsilon(\alpha)$ corresponding to GaussMix, which is tight and improves upon the bounds obtained in earlier works. These improvements yield practical gains in the performance of our algorithm for DP linear regression relative to both Sheffet [2019] and other common alternatives such as Wang [2018], and can be further used in other settings that currently use (GaussMix), such as Prakash et al. [2020], Anand et al. [2021], Sun et al. [2022], Bartan and Pilanci [2023]. Moreover, we further derive a computationally efficient algorithm for private logistic regression. Our numerical experiments demonstrate that these enhancements translate into improved accuracy over standard baselines.

3 Preliminaries

Basic Notation. We denote random variables in sans-serif (e.g., X, y), and their realizations in serif (e.g., x, y resp.). The set $\{1, \dots, n\}$ is denoted by $[n]$. For a vector $A \in \mathbb{R}^d$ its Euclidean norm is denoted by $\|A\|$. The all-zeros column vector of length d is denoted by $\vec{0}_d$. The $k \times k$ identity matrix is $\mathbb{I}_k$ and $\mathcal{N}(0, \mathbb{I}_{k_1 \times k_2})$ denotes a $k_1 \times k_2$ matrix of i.i.d. standard Gaussian entries. We give a more elaborate discussion of notation in Appendix A.

Differential Privacy. Differential privacy relies on the notion of a “neighboring” dataset, which we introduce first. Two datasets X, X' are called *neighbors* if X' is formed by removing an element from X ² or vice-versa, and we use $X \simeq X'$ to denote this relation. In this work, a dataset is regarded as a collection of n real-valued rows, each of length d for $n, d \geq 1$. For a dataset given to us, we assume knowledge of an upper bound C_X (called the *row bound*) where $\|x_i\| \leq C_X$ for all $i \in [n]$.

Intuitively, differential privacy, formalized in the next definition, requires that a randomized algorithm induce nearly identical output distributions given neighboring input datasets.

Definition 1 ((ϵ, δ) -Differential Privacy [Dwork et al., 2006]). A randomized mechanism $\mathcal{M}$ is said to satisfy (ϵ, δ) -differential privacy if for all X, X' such that $X' \simeq X$ and measurable subsets $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$,

$$\Pr(\mathcal{M}(X) \in \mathcal{S}) \leq e^\epsilon \cdot \Pr(\mathcal{M}(X') \in \mathcal{S}) + \delta.$$

A secondary, somewhat stronger notion of differential privacy that we adopt throughout this work is given by Rényi-DP, first introduced by Mironov [2017].

Definition 2 $(\alpha, \epsilon(\alpha))$ -RDP [Mironov, 2017]. A randomized mechanism $\mathcal{M}$ is said to satisfy $(\alpha, \epsilon(\alpha))$ -RDP for some $\alpha > 1$ if for all X, X' such that $X \simeq X'$,

$$D_\alpha(\mathcal{M}(X) \parallel \mathcal{M}(X')) \leq \epsilon(\alpha),$$

where $D_\alpha(P \parallel Q) := \frac{1}{\alpha-1} \log \left(\mathbb{E}_{x \sim Q} \left[\left(\frac{P(x)}{Q(x)} \right)^\alpha \right] \right)$ denotes the α -Rényi divergence [Rényi, 1961].

The notion of $(\epsilon, 0)$ -DP can be viewed as ensuring that the likelihood ratio of events induced by neighboring datasets are uniformly bounded, and δ in (ϵ, δ) -DP provides some additive slack on this condition. On the other hand, for any $\alpha > 1$, $(\alpha, \epsilon(\alpha))$ -Rényi-DP can be seen as another control that bounds the moments of this likelihood ratio. The latter can be translated into the former; this conversion is explicitly stated in the proposition below.

Proposition 1 (Canonne et al. [2020, Proposition 12]). If $\mathcal{M}$ satisfies $(\alpha, \epsilon(\alpha))$ -RDP, then it also satisfies $(\epsilon_{\text{DP}}, \delta)$ -DP for any $0 < \delta < 1$, where $\epsilon_{\text{DP}} = \epsilon(\alpha) + \log \left(1 - \frac{1}{\alpha} \right) - \frac{\log(\alpha\delta)}{(\alpha-1)}$.

Both $(\alpha, \epsilon(\alpha))$ -RDP and (ϵ, δ) -DP satisfy key properties such as graceful degradation under composition and post-processing. In particular, the post-processing property ensures that if a mechanism f satisfies either privacy definition, then so does $g \circ f$ for any (possibly randomized) function g [Dwork et al., 2014, Mironov, 2017].

We highlight a special family of mechanisms that satisfy $(\alpha, \epsilon(\alpha))$ -RDP for a range of values of α wherein $\epsilon(\alpha)$ grows at most linearly in α within this range. Such mechanisms are said to satisfy *truncated concentrated DP* (tCDP) which is defined formally below.

Definition 3 (tCDP [Bun et al., 2018]). Let $\rho > 0$ and $w > 1$. A mechanism $\mathcal{M}$ satisfies (ρ, w) -tCDP if $D_\alpha(\mathcal{M}(X) \parallel \mathcal{M}(X')) \leq \rho \cdot \alpha$ for all neighboring X, X' and for all $\alpha \in (1, w)$.

The (ρ, w) -tCDP property lends to a tighter translation to $(\epsilon_{\text{DP}}, \delta)$ -DP in comparison to simply picking $\alpha \in (1, w)$ and instantiating Proposition 1. This fact leads to tighter privacy guarantees for GaussMix as we present in the latter parts of this paper.

Proposition 2 (Bun et al. [2018, Lemma 6]). If $\mathcal{M}$ satisfies (ρ, w) -tCDP, then its also satisfies $(\epsilon_{\text{DP}}, \delta)$ -DP for all $\delta > 0$ where

$$\epsilon_{\text{DP}} = \begin{cases} \rho + 2\sqrt{\rho \cdot \log(1/\delta)} & \text{if } \log(1/\delta) \leq (w-1)^2 \cdot \rho \\ \rho \cdot w + \frac{\log(1/\delta)}{w-1} & \text{if } \log(1/\delta) > (w-1)^2 \cdot \rho. \end{cases}$$

Gaussian mechanism. The *Gaussian mechanism* is a standard baseline for achieving (ϵ, δ) -DP by simply adding Gaussian noise to (some function of) the data before releasing it. In our notation it amounts to $\mathcal{M}_G(X) = X + \sigma\xi$ with $\xi \sim \mathcal{N}(0, \mathbb{I}_{n \times d})$. The Gaussian mechanism is $\left(\frac{C_X^2}{2\sigma^2}, \infty \right)$ -tCDP (also known as zCDP [Bun and Steinke, 2016]), and (ϵ, δ) -DP

where $\epsilon = \frac{\sqrt{2 \log(1.25/\delta)}}{\sigma}$ for any $\delta \in (0, 1)$ [Dwork et al., 2014, Appendix A].

²For simplicity, we identify a removal of a row with its replacement by $\vec{0}_d$, so the dimension remains constant. This notion is sometimes referred to as *zero-out* neighboring.

4 The Gaussian Mixing Mechanism

We start by providing a new privacy analysis of [GaussMix](#), under the assumption that we have a lower bound $\bar{\lambda}_{\min}$ for the minimum eigenvalue of $X^\top X$ (called the *scale bound*).

Lemma 1. *For any data matrix $X \in \mathbb{R}^{n \times d}$ that satisfies row bound C_X and scale bound $\bar{\lambda}_{\min}$, [GaussMix](#) with parameters k and σ such that $\gamma := C_X^{-2} \cdot (\sigma^2 + \bar{\lambda}_{\min}) > 1$ satisfies $(\alpha, \varphi(\alpha; k, \gamma))$ -RDP for any $\alpha \in (1, \gamma)$, where $\varphi(\alpha; k, \zeta) := \frac{k\alpha}{2(\alpha-1)} \log\left(1 - \frac{1}{\zeta}\right) - \frac{k}{2(\alpha-1)} \log\left(1 - \frac{\alpha}{\zeta}\right)$.*

To help understand the role of γ , this can be viewed as a lower bound on the minimal eigenvalue of the matrix $\tilde{X}^\top \tilde{X}$ where $\tilde{X} := C_X^{-1} \cdot [X^\top, \sigma \mathbb{I}_d]^\top \in \mathbb{R}^{(n+d) \times d}$. Using this perspective, the noise addition can be seen as a way to artificially raise the minimum singular value of the matrix X to a predetermined threshold γ , after which a standard Gaussian sketching step is applied. This reinforces the prior intuition that privacy arises from applying Gaussian sketching to a matrix with a sufficiently large minimum singular value.

Proof sketch of Lemma 1. For a fixed X , we first deduce that every row of $\mathcal{M}(X)$ is distributed according to a multivariate Gaussian distribution with zero mean and covariance $\Sigma := X^\top X + \sigma^2 \mathbb{I}_d$. Using the closed form expression of the Rényi divergence between multivariate Gaussian distributions, the quantity $D_\alpha(\mathcal{M}(X) \| \mathcal{M}(X'))$ is a monotonic function of $x^\top \Sigma^{-1} x$ where x is the row in X that is zeroed out in X' . This quantity can be bounded as $x^\top \Sigma^{-1} x \leq (\lambda_{\min}(\Sigma))^{-1} \cdot \|x\|^2$, and $\lambda_{\min}(\Sigma) \geq \bar{\lambda}_{\min} + \sigma^2$ by the scale bound. We defer the details of this proof to Appendix C.1.

Two key observations from the analysis is that (1) the function φ is non-negative for every $\alpha \in (1, \gamma)$ and further bounds the RDP curve of $\mathcal{M}(X)$ from above, and (2) for sufficiently large γ , the function φ in Lemma 1 grows at most linearly in α in a certain range. The latter property is precisely the definition of tCDP (Definition 3).

Corollary 1. *Consider the setup of Lemma 1. If $\gamma > 5/2$, then [GaussMix](#) satisfies $(k/2\gamma^2, 2\gamma/5)$ -tCDP.*

Tightness of Bound. In our computations, the only inequality that occurs is the one stated in the proof sketch. This inequality is tight when the row in which X differs from X' happens to be the eigenvector corresponding to the minimal eigenvalue of $X^\top X$. This implies that the lemma is tight for specific inputs, ensuring the bound is optimal under certain input assumptions. More precisely, for the set of matrices $X \in \mathbb{R}^{n \times d}$ such that $n \geq d$ with norm bounds C_X and such that $\lambda_{\min}(X^\top X) \geq C_X^2$, our bound is achieved with equality for any X such that $X^\top X = C_X^2 \cdot \mathbb{I}_d$, corresponding to the case where $\frac{1}{C_X} X$ is a semi-orthogonal matrix.

Comparison to Existing Literature. Below we summarise the comparison between our guarantees for the (ε, δ) -DP of [GaussMix](#) and a prior result by [Sheffet \[2019\]](#).

	Our result (Corollary 1 + Proposition 2)	Sheffet [2019] (Theorem 2)
ε	$\frac{k}{2\gamma^2} + \frac{\sqrt{2k \log(1/\delta)}}{\gamma}$	$\frac{2\sqrt{2k \log(4/\delta)}}{\gamma} + \frac{2 \log(4/\delta)}{\gamma}$

For a fixed k , the dependence on γ in our guarantee is strictly better than the result of [Sheffet \[2019\]](#); specifically the bound from [Sheffet \[2019\]](#) suffers from the additional $\frac{2 \log(4/\delta)}{\gamma}$ term. Numerical analysis based on the exact RDP guarantee (derived by using the conversion from RDP to DP of Proposition 1) provides a larger improvement of Lemma 1 over the bound of [Sheffet \[2019\]](#); see Figure 1 for plots to showcasing this comparison.

Moreover, we note that with [GaussMix](#) the entire dataset contributes to the privacy protection of a single element via mixing. Since $\bar{\lambda}_{\min}$ is at most $\frac{nC_X^2}{d}$ (see discussion in [\[Wang, 2018\]](#)), this resembles privacy amplification by shuffling or subsampling, whereby the added noise to other elements contributes to the privacy protection of any element. The dependence of ε on the parameters is not trivially comparable since k affects only [GaussMix](#), and the utility of the two outputs might be arbitrarily different.

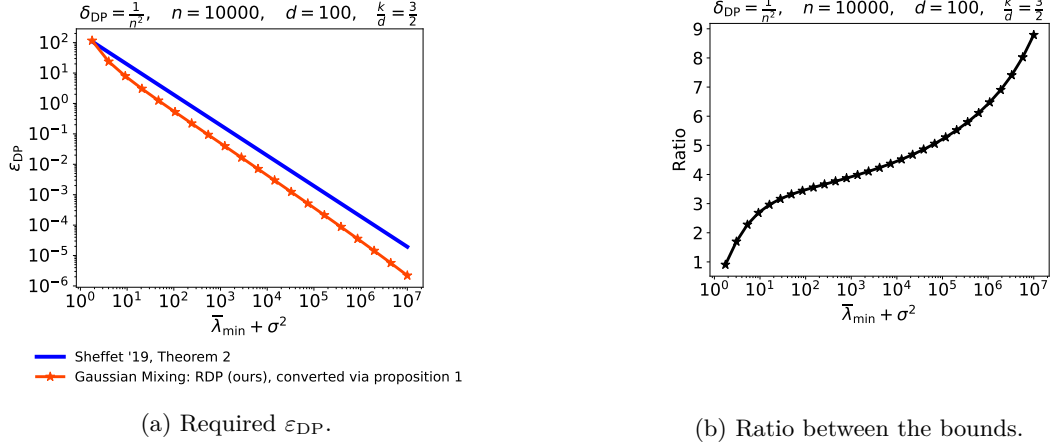


Figure 1: $\bar{\lambda}_{\min} + \sigma^2$ required for attaining ε_{DP} , using Lemma 1 converted to $(\varepsilon_{\text{DP}}, \delta)$ -DP via Proposition 1, and compared to the bound from Sheffet [2019, Theorem 2]. We plot (a) ε_{DP} as a function of $\bar{\lambda}_{\min} + \sigma^2$ and (b) the ratio between the ε_{DP} bounds. Our improved RDP analysis reduces the final ε_{DP} , illustrating the benefit of our new approach.

Algorithm 1 ModifiedGaussMix

Require: Dataset $X \in \mathbb{R}^{n \times d}$, row bound C_X , parameters k, γ, τ, η .

- 1: **if** $\gamma \leq \tau$ **then**:
 - 2: **Output:** $SX + \gamma C_X \xi_1$ with $S \sim \mathcal{N}(0, \mathbb{I}_{k \times n})$ and $\xi_1 \sim \mathcal{N}(0, \mathbb{I}_{k \times d})$.
 - 3: **else**:
 - 4: Set $\tilde{\lambda} = \max \{ \lambda_{\min}(X^\top X) - \eta C_X^2 (\tau - z), 0 \}$, where $z \sim \mathcal{N}(0, 1)$.
 - 5: Set $\tilde{\eta} = \sqrt{\max \{ \gamma - \tilde{\lambda}, 0 \}}$.
 - 6: **Output:** $SX + \tilde{\eta} C_X \xi_1$ with $S \sim \mathcal{N}(0, \mathbb{I}_{k \times n})$ and $\xi_1 \sim \mathcal{N}(0, \mathbb{I}_{k \times d})$.
-

Usage Without $\bar{\lambda}_{\min}$. Since the bounded scale assumption cannot be enforced in the general case, one way to utilize Lemma 1 is by simply using the fact that $\bar{\lambda}_{\min} \geq 0$ and relying solely on the added noise. In this case, the GaussMix mechanism can be thought of as a complicated counterpart of the Gaussian mechanism. Comparing their privacy guarantees reveals a clear similarity, with $\sqrt{k}/\gamma$ taking the role of $1/\sigma$ in the Gaussian mechanism. In this case $\gamma = \sigma^2$ which amounts to a quadratic dependence of ε on σ in the leading term (rather than the linear one in the Gaussian mechanism), balanced by $\sqrt{k}$, which implies improved privacy results when $\sigma > \sqrt{k}$. Since the output of these two mechanisms follows different distributions, the privacy-utility tradeoff depends on the use case.

Instance Specific Bound. While setting $\bar{\lambda}_{\min} = 0$ provides a privacy guarantee for any matrix, one might wish to reduce the added noise in a data-dependent manner based on $\lambda_{\min}(X^\top X)$. Since this quantity is data-dependent, it must be used in a privacy-preserving manner, which is achieved using Algorithm 1. A similar algorithm was proposed by Sheffet [2017, Algorithm 1], with z sampled from a Laplace distribution, and η set either to 0 or to $\sqrt{\gamma}$ based on $\tilde{\lambda}$, rather than interpolating between the two as we do.

Theorem 1. Let $\delta \in (0, 1)$, $k \geq 1$, $\eta > 0$ and $\gamma > 5/2$, and set $\tau \geq \sqrt{2 \log(3/\delta)}$. Then, the output of Algorithm 1 is $(\tilde{\varepsilon}(\eta, \gamma, k, \delta), \delta)$ -DP for

$$\tilde{\varepsilon}(\eta, \gamma, k, \delta) = \frac{\sqrt{2 \log(3.75/\delta)}}{\eta} + \min_{1 < \alpha < \gamma} \left\{ \varphi(\alpha; k, \gamma) + \frac{\log(3/\delta) + (\alpha - 1) \log(1 - 1/\alpha) - \log(\alpha)}{\alpha - 1} \right\} \quad (2)$$

where $\varphi(\alpha; k, \gamma)$ was previously defined in Lemma 1.

Proof sketch. The case of $\gamma < \tau$ is the previously discussed $\bar{\lambda}_{\min} = 0$ setting. For the other case, we note that w.p. $\geq 1 - \delta/3$ we have $\tilde{\lambda} \geq \lambda_{\min}(X^\top X)$, so the conditions of Lemma 1 are met. Since $\bar{\lambda}_{\min}$ has sensitivity C_X^2 , the privacy properties of the Gaussian mechanism apply to the release of $\tilde{\lambda}$. Using simple composition, we combine it with the privacy of [GaussMix](#), which completes the proof. We give full details in Appendix C.3.

The quantity $\tilde{\varepsilon}(\eta, \gamma, k, \delta) \geq 0$ and further can be made arbitrarily small by increasing η and γ ; we show this formally in Appendix B. As mentioned previously, the parameter γ should be thought of as the target minimal eigenvalue, τ serves as the accuracy bound on the estimation of $\bar{\lambda}_{\min}$, and η controls the privacy loss of its private estimation. Setting $\eta = \gamma/\sqrt{k}$, and combining Corollary 1 with the DP implication of tCDP, we get that Algorithm 1 is (ε, δ) -DP where $\varepsilon = \frac{k}{2\gamma^2} + \frac{2\sqrt{2k \log(4/\delta)}}{\gamma}$, which matches the privacy bound on [GaussMix](#) up to a constant.

5 Applications

We now demonstrate applications of [GaussMix](#) for two private learning tasks that can be formalized as linear regression. In each case, we instantiate the mechanism on a concatenated dataset (X_1, X_2) where $X_1 \in \mathbb{R}^{n \times d_1}$ and $X_2 \in \mathbb{R}^{n \times d_2}$. The application of [GaussMix](#) to the concatenation (X_1, X_2) is given by

$$\mathcal{M}(X_1, X_2) = \mathbf{S}(X_1, X_2) + \sigma(\xi_1, \xi_2) = (\mathbf{S}X_1 + \sigma\xi_1, \mathbf{S}X_2 + \sigma\xi_2)$$

where $\mathbf{S} \sim \mathcal{N}(0, \mathbb{I}_{k \times n})$, $\xi_1 \sim \mathcal{N}(0, \mathbb{I}_{k \times d_1})$, and $\xi_2 \sim \mathcal{N}(0, \mathbb{I}_{k \times d_2})$. If $\mathcal{M}(X_1, X_2)$ satisfies RDP w.r.t. (X_1, X_2) , then by post-processing the inner-product $\mathcal{K}$ defined as

$$\mathcal{K}(\mathcal{M}(X_1, X_2)) := (\mathbf{S}X_1 + \sigma\xi_1)^\top (\mathbf{S}X_2 + \sigma\xi_2) \quad (3)$$

also satisfies RDP. The inner-product $\mathcal{K}$ will form the core component of our algorithms.

5.1 Differentially Private Ordinary Least Squares

We begin with the problem of DP linear regression. Let the dataset (X, Y) such that the design matrix $X \in \mathbb{R}^{n \times d}$ and the response vector $Y \in \mathbb{R}^n$. Throughout, we make the following assumptions:

- (A₁) *Bounded domain:* $\|x_i\| \leq C_X$ and $|y_i| \leq C_Y$ for all $i \in [n]$ ³.
- (A₂) *Overspecified system:* $n \geq d$.

Our goal is to estimate a linear predictor under (ε, δ) -DP, while preserving the privacy of each individual data pair $(x_i^\top, y_i)$. Our non-private baseline is the standard Ridge regression estimator [[Tikhonov, 1963](#), [Hoerl and Kennard, 1970](#)]:

$$\theta^*(\nu) := \underset{\theta}{\operatorname{argmin}} \{ \|Y - X\theta\|_2^2 + \nu \|\theta\|_2^2 \} = (X^\top X + \nu \mathbb{I}_d)^{-1} X^\top Y,$$

where $\nu \geq 0$ is a regularization parameter. The unregularized least-squares solution is denoted $\theta^* := \theta^*(0)$, and we define the empirical loss as $L_{X,Y}(\theta) := \|Y - X\theta\|_2^2$. Our algorithm, summarized in Algorithm 2, uses Algorithm 1 for obtaining a privatized version of the pair (X, Y) (interpreted as halves of one joint matrix) by setting a large enough γ and then solving the resulting least-squares problem. The existence of a γ that satisfies the conditions in Line 1 in Algorithm 2 is ensured by the analysis presented in Appendix B. While the structure of our algorithm resembles earlier proposal by [Sheffet \[2017\]](#), our refined privacy analysis under Rényi-DP improves the overall privacy-utility trade-off. Moreover, our algorithm exploits $\lambda_{\min}$ in a modified way, ensuring that the variance of the additive noise component is always reduced by utilizing the private estimation of $\lambda_{\min}(X^\top X)$.

Corollary 2. *For any $k \geq 1$ the output θ_{Lin} of Algorithm 2 is (ε, δ) -DP.*

³These domain bounds appear in many standard DP linear regression settings, such as [[Sheffet, 2017](#), [Wang, 2018](#)].

Algorithm 2 LinearMixing

Require: Dataset (X, Y) satisfying $(\mathbf{A}_1, \mathbf{A}_2)$, privacy parameters (ε, δ) , parameter k .

1: Find smallest $\gamma > 5/2$ such that $\tilde{\varepsilon}(\eta, \gamma, k, \delta) \leq \varepsilon$, while setting $\eta = \gamma/\sqrt{k}$.

2: Calculate $[\tilde{X}, \tilde{Y}] = \text{ModifiedGaussMix}([X, Y], \sqrt{C_X^2 + C_Y^2}, k, \gamma, \sqrt{2 \log(3/\delta)}, \eta)$.

3: **Output:** $\theta_{\text{Lin}} := (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{Y}$.

Proof. We first note that for any $k \geq 1$, we are guaranteed to find a γ in Line 1 such that the target $\tilde{\varepsilon}(\eta, \gamma, k, \delta) \leq \varepsilon$ where $\eta = \gamma/\sqrt{k}$ (see Appendix B). Since we apply **ModifiedGaussMix** (Line 2) with the appropriate domain bounds, the differential privacy guarantee follows from Theorem 1. Finally, since θ_{Lin} is a function of $(\tilde{X}, \tilde{Y})$, which is (ε, δ) -DP, we have by post-processing that θ_{Lin} also satisfies (ε, δ) -DP. $\square$

Theorem 2. *There exist universal constants c_0, c_1, c_2 such that for any $\chi \in (0, 1]$ satisfying $k\chi^2 \geq c_0 d$, the following holds with probability at least $1 - c_1 \cdot \exp\{-c_2 k\chi^2\}$:*

$$L_{X,Y}(\theta_{\text{Lin}}) - (1 + \chi)^2 \cdot L_{X,Y}(\theta^*) \leq O\left((1 + \chi)^2 \cdot \frac{\sqrt{k \log(1/\delta)} \cdot (C_X^2 + C_Y^2)}{\varepsilon} \cdot (1 + \|\theta^*\|^2)\right).$$

The proof is deferred to Appendix C.4. Notably, in the regime where $C_Y \approx C_X \|\theta^*\|$ and $d \gg 1$, for a target failure probability ϱ , setting $\chi = \sqrt{k^{-1} \cdot \max\{c_0 d, (1/c_2) \log(c_1/\varrho)\}}$ and $k = (1 + \beta) \max\{c_0 d, 1/c_2 \log(c_1/\varrho)\}$ for some finite $\beta > 0$ yields excess empirical risk of $O\left(\varepsilon^{-1} \sqrt{\max\{d, \log(1/\varrho)\} \log(1/\delta)} C_X^2 (1 + \|\theta^*\|^2) + \chi L_{X,Y}(\theta^*)\right)$. In particular, we note that the first term avoids the multiplicative $\log(d^2/\varrho)$ factor that exists in the guarantees from Wang [2018]. Thus, whenever it holds that $\lambda_{\min}(X^\top X) \ll 1$ and furthermore $L_{X,Y}(\theta^*) \ll \varepsilon^{-1} \sqrt{\max\{d, \log(1/\varrho)\} \log(1/\delta)} C_X^2 (1 + \|\theta^*\|^2)$ we expect our method to outperform the AdaSSP baseline. Conversely, in high-residual regimes—where our bound incurs the term $\chi L_{X,Y}(\theta^*)$ —or when $\lambda_{\min}(X^\top X)$ is large—so that AdaSSP’s guarantee enjoy from this minimal eigenvalue in the denominator of its guarantees, AdaSSP is likely to perform better.

To demonstrate the usefulness of Algorithm 2, we have simulated its performance on four different datasets: the Communities & Crime dataset [Redmond and Baveja, 2002], the Tecator dataset [Thodberg, 2015], a synthetic dataset comprised of Gaussian features transformed via an MLP and another dataset comprised of Gaussian features. Full experimental details are provided in Appendix F. We use algorithm. 1 from [Sheffet, 2017] as well as the AdaSSP algorithm of [Wang, 2018] as our private baselines. We further simulated a variant of algorithm. 1 from [Sheffet, 2017] that uses the analysis established in Lemma 1. The AdaSSP algorithm is considered the leading baseline for DP linear regression under bounded domain assumptions [Liu et al., 2022, Brown et al., 2024]. While some recent works, such as [Ferrando and Sheldon, 2025], have proposed practical enhancements over AdaSSP, this method involves significantly higher computational cost and lacks theoretical utility guarantees, as opposed to our method which runs essentially within the same time constants. As Figure 2 shows, our method outperforms the AdaSSP method and the method of [Sheffet, 2017], achieving lower or equal test MSE across all privacy levels on every dataset. Furthermore, incorporating the analysis of Lemma 1 into the algorithm of [Sheffet, 2017] improves its performance, hinting on the further usefulness of our new analysis technique.

5.2 Differentially Private Logistic Regression

Our second application considers the problem of DP logistic regression, with the goal of developing DP solutions to the ERM problem:

$$\theta^* = \operatorname{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^n -\log(1 + \exp\{-y_i \theta^\top x_i\}) := \operatorname{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^n \ell_{\theta}(s_i, y_i)$$

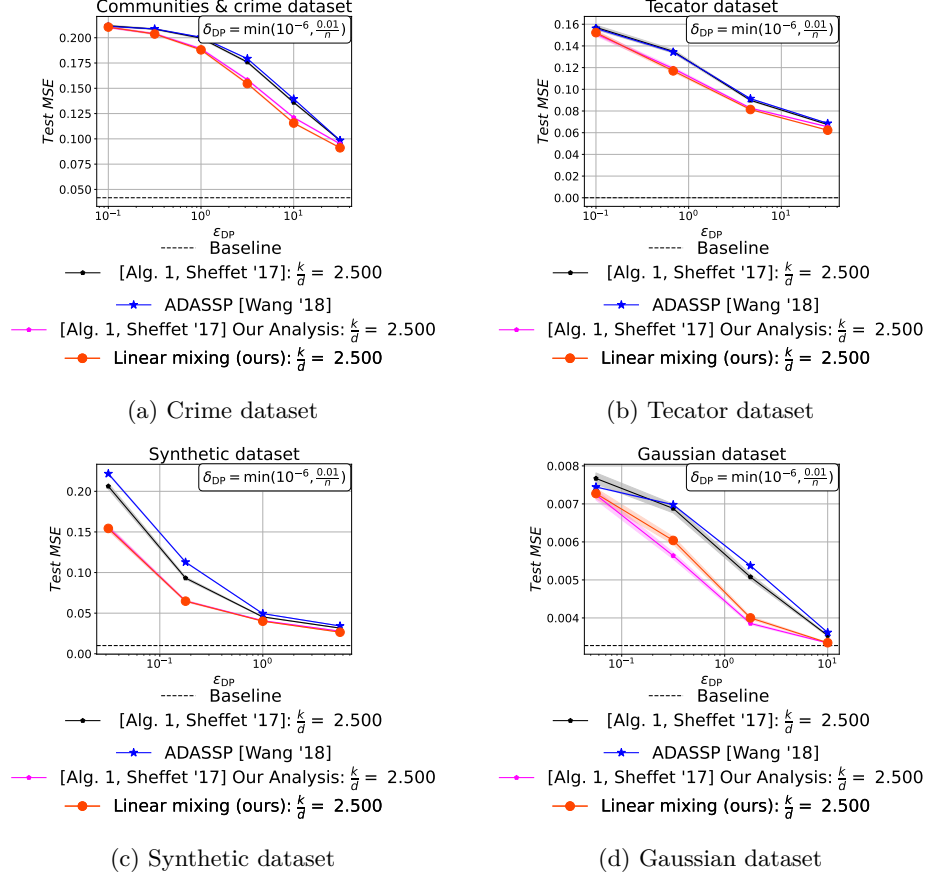


Figure 2: Linear mixing performance on four linear regression tasks. The parameter k was chosen to ensure negligible approximation error in the sketched solution.

and where $y_i \in \{-1, +1\}$ for all $i \in [n]$. Following [Huggins et al., 2017] (see also [Ferrando and Sheldon, 2025]), this problem can be solved by approximating $\ell_\theta(s, y)$ with a second-order polynomial $q(s) = b_0 + b_1 s + b_2 s^2$, i.e. $-\log(1 + \exp\{-s\}) \approx q(s)$, for $s \in \mathcal{J}$, where $\mathcal{J}$ satisfies $y\theta^\top x \in \mathcal{J}$ for all datapoints (x, y) , ensuring the surrogate is valid over the dataset. Substituting $s_i = y_i \theta^\top x_i$ and discarding the constant b_0 , minimising the surrogate objective reduces to ordinary least squares with a response vector $\tilde{Y} = -\frac{b_1}{2b_2} Y$, with $Y = (y_1, \dots, y_n)^\top$. Therefore, we can invoke Algorithm 2, originally designed for linear regression, to obtain a (ϵ, δ) -DP estimate for logistic regression. The utility of this solution follows via similar arguments to those of Theorem 2, with the replacement of C_Y^2 with $\left(\frac{b_1}{2b_2}\right)^2$ and with an extra factor due to the loss approximation and is presented in Appendix D. However, in this case, the complexity of sketching the data and then solving the linear system (for example, using a QR decomposition) will be $O(nkd + kd^2)$. In many cases, this one-shot approach is more computationally efficient than classical approaches that use iterative optimization techniques.

To demonstrate this point, we have tested our approach in a similar setting to that presented by Guo et al. [2020], where we (i) trained a CNN with DP stochastic gradient descent (DP-SGD) implemented using Opacus [Yousefpour et al., 2021] and (ii) used the pre-trained private embeddings for DP fine-tuning of a logistic head. The CNN architecture and training hyperparameters mirror those in [Guo et al., 2020] (see Appendix F for full details). We train the head using one of three methods: DP-SGD [Abadi et al., 2016, Yousefpour et al., 2021], objective perturbation as in [Guo et al., 2020], or with our approach (Linear Mixing as depicted in Algorithm 2). As shown in Figure 3, our approach delivers lower runtime than

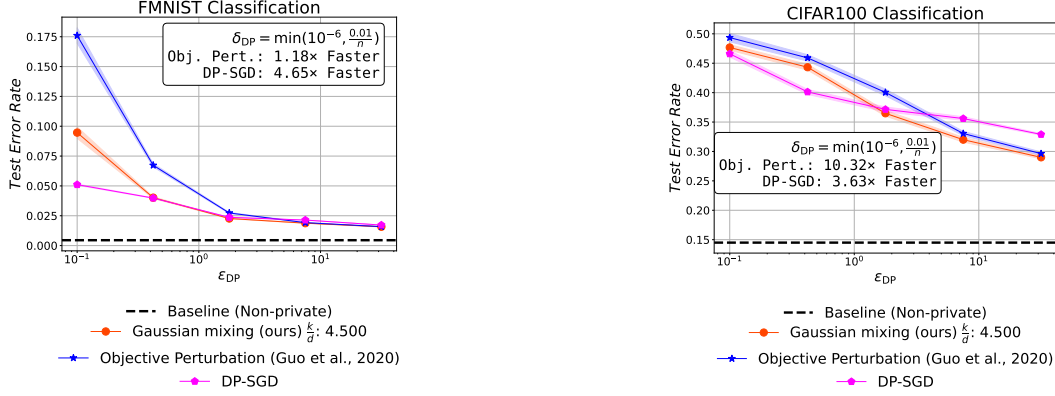


Figure 3: DP logistic regression using features from a privately trained CNN on binary subsets of Fashion-MNIST and CIFAR100. We fixed k on $k = 4.5d$.

both baselines and achieves consistent accuracy gains over objective perturbation. It further exceeds DP-SGD accuracy for larger values of the privacy parameter ϵ . Full experimental details are provided in Appendix F.2.

6 Discussion and Conclusion

In this work, we revisited the Gaussian mixing mechanism originally introduced by Blocki et al. [2012] and later studied by Sheffet [2017, 2019]. We derived its RDP curve, which yields tighter bounds on the relationship between the noise parameter σ and (ϵ, δ) , thereby strengthening the privacy analysis of this mechanism. We further demonstrated the practical usefulness of this improved analysis by applying (GaussMix) to two distinct machine learning tasks and providing: (i) an improved algorithm for DP linear regression, (ii) an algorithm for DP logistic regression. The analysis used in the proof of Lemma 1 is tighter than that presented in [Sheffet, 2019]. Thus, it also offers performance improvement in other settings currently invoking the results of [Sheffet, 2019], such as [Bartan and Pilanci, 2023].

A key technical property that underpins the usefulness of (GaussMix) is its compatibility with formulations in which terms involving the random projection matrix S cancel in expectation, as described intuitively in (1). Identifying additional machine learning problems and formulations that naturally admit this structure and which facilitate an implementation using the Gaussian mixing mechanism is an interesting direction for future work, open the possibility for applying (GaussMix) beyond the currently studied linear regression settings. Moreover, this work focuses on the case where S is Gaussian. Extending the framework to other classes of projections, such as those studied by Woodruff [2014], Pilanci and Wainwright [2015, 2016], could enable computationally efficient implementation of (3).

Broader Impact The paper develops the DP guarantees for the Gaussian mixing mechanism. However, when used in practice, one needs to verify all the prior mathematical assumptions we made throughout in order to guarantee data privacy in practice.

Acknowledgements

The authors thank the reviewers at NeurIPS 2025 for their valuable feedback on a preliminary version of this work. AS thanks Gautam Kamath for helpful discussions. The work of MS and KL was supported in part by ERC grant 101125913, Simons Foundation Collaboration 733792, Israel Science Foundation (ISF) grant 2861/20, Apple, and a grant from the Israeli Council of Higher Education. AW was supported in part by Simons Foundation Collaboration on Algorithmic Fairness. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

- Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2016.*, pages 308–318, 2016.
- Arjun Anand, Sagar Dhakal, Mustafa Akdeniz, Brandon Edwards, and Nageen Himayat. Differentially private coded federated linear regression. In *IEEE Data Science and Learning Workshop*, pages 1–6. IEEE, 2021.
- Apple Research. Learning with Privacy at Scale. *Apple Machine Learning Research*, 2017. URL <https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf>.
- Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in Neural Information Processing Systems*, 31, 2018.
- Burak Bartan and Mert Pilanci. Distributed sketching for randomized optimization: Exact characterization, concentration, and lower bounds. *IEEE Transactions on Information Theory*, 69(6):3850–3879, 2023.
- bike. Bike sharing dataset. <https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset>, 2019.
- Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. The Johnson-Lindenstrauss transform itself preserves differential privacy. In *IEEE Symposium on Foundations of Computer Science*, pages 410–419. IEEE, 2012.
- Gavin R Brown, Krishnamurthy Dj Dvijotham, Georgina Evans, Daogao Liu, Adam Smith, and Abhradeep Guha Thakurta. Private gradient descent for linear regression: Tighter error bounds and instance-specific uncertainty estimation. In *Proceedings of the International Conference on Machine Learning*, pages 4561–4584. PMLR, 2024.
- Mark Bun and Thomas Steinke. Concentrated differential privacy: simplifications, extensions, and lower bounds. In *Theory of cryptography conference*, pages 635–658. Springer, 2016.
- Mark Bun, Cynthia Dwork, Guy N Rothblum, and Thomas Steinke. Composable and versatile privacy via truncated CDP. In *Proceedings of the 50th ACM SIGACT Symposium on Theory of Computing*, pages 74–86, 2018.
- United States Census Bureau. A History of Census Privacy Protections. <https://www.census.gov/library/visualizations/2019/comm/history-privacy-protection.html>, 2019.
- Clément L Canonne, Gautam Kamath, and Thomas Steinke. The discrete Gaussian for differential privacy. *Advances in Neural Information Processing Systems*, 33:15676–15688, 2020.
- Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.
- Paul Cuff and Lanqing Yu. Differential privacy as a mutual information constraint. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 43–54, 2016.
- Cynthia Dwork. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, pages 1–12. Springer, 2006.
- Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

- Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.
- Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM, 2019.
- Facebook Research. Protecting privacy in facebook mobility data during the covid-19 response, June 2020. URL <https://research.facebook.com/blog/2020/6/protecting-privacy-in-facebook-mobility-data-during-the-covid-19-response/>.
- Cecilia Ferrando and Daniel Sheldon. Private regression via data-dependent sufficient statistic perturbation. *Transactions on Machine Learning Research*, 2025.
- Manuel Gil, Fady Alajaji, and Tamas Linder. Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 249:124–131, 2013.
- Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. In *Proceedings of the International Conference on Machine Learning*, pages 3832–3842. PMLR, 2020.
- David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978.
- Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.
- Jonathan Huggins, Ryan P Adams, and Tamara Broderick. PASS-GLM: polynomial approximate sufficient statistics for scalable Bayesian GLM inference. *Advances in Neural Information Processing Systems*, 30, 2017.
- Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.
- Krishnaram Kenthapadi, Aleksandra Korolova, Ilya Mironov, and Nina Mishra. Privacy via the Johnson-Lindenstrauss transform. *Journal of Privacy and Confidentiality*, 5(1), July 2013.
- Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1. PMLR, 2012.
- Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, ON, Canada, 2009. URL <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>.
- Yann LeCun and Corinna Cortes. MNIST handwritten digit database, 2010. URL <http://yann.lecun.com/exdb/mnist/>.
- Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Conference on Learning Theory*, pages 1167–1246. PMLR, 2022.
- Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *IEEE Symposium on Foundations of Computer Science*, pages 94–103. IEEE, 2007.
- Ilya Mironov. Rényi differential privacy. In *IEEE 30th Computer Security Foundations Symposium*, pages 263–275. IEEE, 2017.
- Mert Pilanci and Martin J Wainwright. Randomized sketches of convex programs with sharp guarantees. *IEEE Transactions on Information Theory*, 61(9):5096–5115, 2015.

- Mert Pilanci and Martin J Wainwright. Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares. *Journal of Machine Learning Research*, 17(53):1–38, 2016.
- Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. How to dp-fy ml: a practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77:1113–1201, 2023.
- Saurav Prakash, Sagar Dhakal, Mustafa Riza Akdeniz, Yair Yona, Shilpa Talwar, Salman Avestimehr, and Nageen Himayat. Coded computing for low-latency federated learning over wireless edge networks. *IEEE Journal on Selected Areas in Communications*, 39(1): 233–250, 2020.
- Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.
- Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press, 1961.
- Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *2006 47th annual IEEE symposium on foundations of computer science*, pages 143–152. IEEE, 2006.
- Or Sheffet. Private approximations of the 2nd-moment matrix using existing techniques in linear regression. *arXiv preprint arXiv:1507.00056*, 2015.
- Or Sheffet. Differentially private ordinary least squares. In *Proceedings of the International Conference on Machine Learning*, pages 3105–3114. PMLR, 2017.
- Or Sheffet. Old techniques in differentially private linear regression. In *Algorithmic Learning Theory*, pages 789–827. PMLR, 2019.
- Mehrdad Showkatbakhsh, Can Karakus, and Suhas Diggavi. Privacy-utility trade-off of linear regression under random projections and additive noise. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 186–190. IEEE, 2018.
- Snap Security. Differential privacy at snapchat, December 2022. URL <https://eng.snap.com/differential-privacy-at-snap>.
- Yuchang Sun, Jiawei Shao, Songze Li, Yuyi Mao, and Jun Zhang. Stochastic coded federated learning with convergence and privacy guarantees. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 2028–2033. IEEE, 2022.
- H. H. Thodberg. Tecator meat sample dataset. StatLib Datasets Archive, Carnegie Mellon University, 2015.
- Andrei N Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Sov Dok*, 4:1035–1038, 1963.
- Yu-Xiang Wang. Revisiting differentially private linear regression: Optimal and adaptive prediction & estimation in unbounded domain. In *Conference on Uncertainty in Artificial Intelligence*, pages 93–103. AUAI Press, 2018.
- wine. Wine quality data set. <https://archive.ics.uci.edu/dataset/186/wine+quality>, 2009.
- David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.
- Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, et al. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

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: The paper deals with differential privacy via a mechanism we dubbed as Gaussian mixing. This is reflected in the paper title.

Guidelines:

- The answer NA 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 NA 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: The paper describes the application of the mixing mechanism to linear regression. Furthermore, we demonstrate results on a finite number of datasets which we describe in the paper body.

Guidelines:

- The answer NA 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: The assumptions made in each theorem are described in the theorem body.

Guidelines:

- The answer NA 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: The methods and the datasets used are detailed in the paper, along with experimental details section in the paper appendix.

Guidelines:

- The answer NA 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: [No]

Justification: Code will be released upon publication.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.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://nips.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, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper contain an experimental section with details of all of the experiments.

Guidelines:

- The answer NA 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: All the plots in the paper contain corresponding confidence intervals, claiming the statistical significance of the results.

Guidelines:

- The answer NA 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: The computer resources of all the experiments are detailed in the paper body.

Guidelines:

- The answer NA 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 have verified that the research performed in this paper conformed with the NeurIPS code of ethics.

Guidelines:

- The answer NA 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: [\[Yes\]](#)

Justification: We have included a broader impact subsection in our paper.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

Answer: [\[NA\]](#)

Justification:

Guidelines:

- The answer NA 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: [Yes]

Justification: All the datasets are publicly available and their sources are cited in the paper main body.

Guidelines:

- The answer NA 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: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA 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: [NA]

Justification:

Guidelines:

- The answer NA 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: [NA]

Justification:

Guidelines:

- The answer NA 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 rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA 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 (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

Appendix to the paper

The Gaussian Mixing Mechanism:

Rényi Differential Privacy via Gaussian Sketches

Organization In Section A, we elaborate on notation used throughout this work. In Section B, we discuss a salient property of $\tilde{\varepsilon}$ that appears in Algorithm 2. In Section C, we give proofs for the main theoretical results presented in the main text. Section D contains the utility guarantees for our differentially private logistic regression method in Section 5.2. We then discuss details of the various DP algorithms considered as baselines in Section E. Finally, we give details of the experiments in Section F, and additional numerical experiments in Section G.

A Notation

Given a matrix $A \in \mathbb{R}^{m \times n}$ we denote its elements by A_{ij} and its column-stack representation by

$$\text{vec}(A) := (A_{11}, A_{21}, \dots, A_{m1}, A_{12}, A_{22}, \dots, A_{mn})^\top.$$

Random variables are denoted using sans-serif fonts (e.g., $\mathbf{X}, \mathbf{y}$), while their realizations are represented by regular italics (e.g., X, y). The L_2 norm of a vector $A := (a_1, \dots, a_d)$ is given by $\left(\sum_{i=1}^d a_i^2\right)^{\frac{1}{2}}$ and is denoted by $\|A\|$. We denote the minimal and the maximal eigenvalues of a matrix A by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$. We denote a PSD matrix A by $A \succeq 0$ and a PD matrix by $A \succ 0$. We usually denote our dataset $\{x_i\}_{i=1}^n$ where each $x_i \in \mathbb{R}^d$ in the matrix form $X = (x_1, \dots, x_n)^\top$. Then, we denote the j 'th entry of x_i by $x_i(j)$. The determinant of a matrix A is denoted by $\det(A)$. The set of integer numbers from 1 to n is denoted by $[n]$. The all zeros column vector of size d is denoted by $\vec{0}_d := (0, \dots, 0)^\top$. We denote by $\mathcal{N}(0, \mathbb{I}_{k_1 \times k_2})$ a $k_1 \times k_2$ matrix comprised of i.i.d. Gaussian elements with zero mean and unit variance. We denote by $\mathcal{N}_{\text{sym}}(0, \mathbb{I}_d)$ a $d \times d$ symmetric matrix whose elements on the upper triangular matrix are i.i.d. and distributed according to $\mathcal{N}(0, 1)$. The Kronecker product between matrices A and B is defined via

$$A \otimes B := \begin{pmatrix} A_{11}B & \dots & A_{1m}B \\ & \ddots & \\ A_{n1}B & \dots & A_{nm}B \end{pmatrix}.$$

The $k \times k$ identity matrix is denoted by $\mathbb{I}_k$.

B A central property of $\tilde{\varepsilon}(\eta, \gamma, k, \delta)$

Proposition 3. *The function $\tilde{\varepsilon}$ in (2) is non-negative. For fixed values of k, δ , $\tilde{\varepsilon}(\eta, \gamma, k, \delta)$ is a decreasing function in η , and satisfies $\tilde{\varepsilon}(\eta, \gamma, k, \delta) \leq h(\gamma)$ for a monotonically decreasing function h where $h(\gamma) \rightarrow 0$ as $\gamma \rightarrow \infty$.*

Proof. We now show that $\tilde{\varepsilon}(\eta, \gamma, k, \delta) \geq 0$ and further that one can make $\tilde{\varepsilon}(\eta, \gamma, k, \delta)$ as small as any desirable ε by increasing η and γ . We first note that $\varphi(\alpha; k, \gamma)$ is an upper bound on $D_\alpha(\mathcal{M}(X) \parallel \mathcal{M}(X'))$. Thus, following the validity of the conversion from RDP to DP of [Canonne et al. \[2020\]](#), the second term in $\tilde{\varepsilon}(\eta, \gamma, k, \delta)$ provides an upper bound on the privacy parameter ε , and thus is non-negative. Since the first term in $\tilde{\varepsilon}(\eta, \gamma, k, \delta)$ is non-negative we get that the entire expression is non-negative.

To prove that $\tilde{\varepsilon}(\eta, \gamma, k, \delta)$ can be made arbitrarily small, we use the result of Corollary 1 which tells us that $\varphi(\alpha; k, \gamma) \leq \frac{k\alpha}{2\gamma^2}$ for $1 < \alpha \leq \frac{2}{5}\gamma$ and provided that $\gamma > \frac{5}{2}$. However, we note that the minimum in $\tilde{\varepsilon}(\eta, \gamma, k, \delta)$ is upper bounded by

$$\min_{1 < \alpha < 2\gamma/5} \left\{ \frac{k\alpha}{2\gamma^2} + \frac{\log(3/\delta)}{\alpha - 1} \right\} \leq \frac{k}{5\gamma} + \frac{\log(3/\delta)}{\frac{2}{5}\gamma - 1}$$

which is derived by substituting $\alpha = 2\gamma/5$. Thus, this minimum is monotonically decreasing in γ and can be made arbitrarily small by increasing γ . The result then follows since the first term in (2) is monotonically decreasing in η , and holds further in the case where $\eta = \frac{\gamma}{\sqrt{k}}$ by picking a sufficiently large γ . $\square$

C Proofs

C.1 Proof of Lemma 1

The proof relies on the next lemma, which establishes the α -Rényi divergence between two multivariate Gaussian distributions.

Lemma 2 (Gil et al. [2013]). *Let $\mathbf{x}_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $\mathbf{x}_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$. Then,*

$$D_\alpha(\mathbf{x}_1 \parallel \mathbf{x}_2) = \frac{\alpha}{2} (\mu_1 - \mu_2)^\top (\Sigma_1 + \alpha(\Sigma_2 - \Sigma_1))^{-1} (\mu_1 - \mu_2) - \frac{1}{2(\alpha - 1)} \log \left(\frac{\det(\Sigma_1 + \alpha(\Sigma_2 - \Sigma_1))}{(\det(\Sigma_1))^{1-\alpha} (\det(\Sigma_2))^\alpha} \right) \quad (4)$$

for all α such that $\alpha \Sigma_1^{-1} + (1 - \alpha) \Sigma_2^{-1} \succ 0$.

Proof of Lemma 1. Instead of analyzing GaussMix, we will analyze the transformed mechanism

$$\widetilde{\mathcal{M}}(X) := (\mathcal{M}(X))^\top = X^\top \mathbf{S}^\top + \sigma \xi^\top,$$

which, in terms of privacy, is equivalent to $\mathcal{M}(X)$. We note that $\widetilde{\mathcal{M}}(X)$ is a matrix of Gaussian random variables, where its columns are i.i.d. and each column has a covariance of

$$\mathbb{E}[(X^\top \mathbf{s}_i + \sigma \xi_i)(X^\top \mathbf{s}_i + \sigma \xi_i)^\top] = X^\top X + \sigma^2 \mathbb{I}_d = \sum_{i=1}^n x_i x_i^\top + \sigma^2 \mathbb{I}_d$$

where we have denoted $\mathbf{S}^\top = (\mathbf{s}_1, \dots, \mathbf{s}_k)$ and $\xi^\top = (\xi_1, \dots, \xi_k)$. Thus, we first note that

$$\text{vec}(\widetilde{\mathcal{M}}(X)) \sim \mathcal{N}(0, \mathbb{I}_k \otimes (X^\top X + \sigma^2 \mathbb{I}_d)).$$

Let X' be our neighbor dataset that is different from X in a single row. Throughout we assume that X' is equivalent to X except for one row which is zeroed out (see Section 3). We then show that the proof also covers the inverse case where one row of X is zeroed out.

Without loss of generality, assume that the differing row is the first row of X . Thus, we first note that

$$\mathbb{I}_k \otimes (X^\top X + \sigma^2 \mathbb{I}_d) - \mathbb{I}_k \otimes (X'^\top X' + \sigma^2 \mathbb{I}_d) = \mathbb{I}_k \otimes (x_1 x_1^\top).$$

Let $\Sigma_1 := \mathbb{I}_k \otimes (X^\top X + \sigma^2 \mathbb{I}_d)$ and $\Sigma_2 := \mathbb{I}_k \otimes (X'^\top X' + \sigma^2 \mathbb{I}_d)$. Now,

$$\Sigma_1 + \alpha(\Sigma_2 - \Sigma_1) = \mathbb{I}_k \otimes (-\alpha x_1 x_1^\top + X^\top X + \sigma^2 \mathbb{I}_d)$$

and since $\mathbb{E}[\text{vec}(\widetilde{\mathcal{M}}(X))] = \mathbb{E}[\text{vec}(\widetilde{\mathcal{M}}(X'))] = 0$ by using the algebraic identity $\det(\mathbb{I}_k \otimes A) = (\det(A))^k$ and by using (4) we get

$$D_\alpha(\mathcal{M}(X) \parallel \mathcal{M}(X')) = -\frac{k}{2(\alpha - 1)} \log \left(\frac{\det(-\alpha x_1 x_1^\top + X^\top X + \sigma^2 \mathbb{I}_d)}{(\det(X^\top X + \sigma^2 \mathbb{I}_d))^{1-\alpha} (\det(X^\top X + \sigma^2 \mathbb{I}_d - x_1 x_1^\top))^\alpha} \right). \quad (5)$$

For an invertible matrix A , the matrix determinant lemma states that

$$\det(A + uv^\top) = \det(A) (1 + v^\top A^{-1} u).$$

Since the matrix $X^\top X + \sigma^2 \mathbb{I}_d$ is invertible whenever $\sigma^2 > 0$, this further tells us that the denominator of (5) can be simplified to

$$\begin{aligned} & (\det(X^\top X + \sigma^2 \mathbb{I}_d))^{1-\alpha} (\det(X^\top X + \sigma^2 \mathbb{I}_d - x_1 x_1^\top))^\alpha \\ &= (\det(X^\top X + \sigma^2 \mathbb{I}_d))^{1-\alpha} (\det(X^\top X + \sigma^2 \mathbb{I}_d))^\alpha (1 - x_1^\top (X^\top X + \sigma^2 \mathbb{I}_d)^{-1} x_1)^\alpha \\ &= \det(X^\top X + \sigma^2 \mathbb{I}_d) (1 - x_1^\top (X^\top X + \sigma^2 \mathbb{I}_d)^{-1} x_1)^\alpha. \end{aligned}$$

Thus, we further have

$$\begin{aligned} & \frac{\det(-\alpha x_1 x_1^\top + X^\top X + \sigma^2 \mathbb{I}_d)}{(\det(X^\top X + \sigma^2 \mathbb{I}_d))^{1-\alpha} (\det(X^\top X + \sigma^2 \mathbb{I}_d - x_1 x_1^\top))^\alpha} \\ &= \left(\frac{\det(-\alpha x_1 x_1^\top + X^\top X + \sigma^2 \mathbb{I}_d)}{\det(X^\top X + \sigma^2 \mathbb{I}_d)} \right) \cdot (1 - x_1^\top (X^\top X + \sigma^2 \mathbb{I}_d)^{-1} x_1)^{-\alpha}. \end{aligned}$$

Similarly, we can apply the same determinant identity to $\det(-\alpha x_1 x_1^\top + X^\top X + \sigma^2 \mathbb{I}_d)$ and get

$$\frac{\det(-\alpha x_1 x_1^\top + X^\top X + \sigma^2 \mathbb{I}_d)}{\det(X^\top X + \sigma^2 \mathbb{I}_d)} = 1 - \alpha x_1^\top (X^\top X + \sigma^2 \mathbb{I}_d)^{-1} x_1.$$

Next, we note that this yields the next simplified form for $D_\alpha(\mathcal{M}(X) \parallel \mathcal{M}(X'))$:

$$\begin{aligned} & D_\alpha(\mathcal{M}(X) \parallel \mathcal{M}(X')) \\ &= -\frac{k}{2(\alpha-1)} \log((1 - \alpha x_1^\top (X^\top X + \sigma^2 \mathbb{I}_d)^{-1} x_1)(1 - x_1^\top (X^\top X + \sigma^2 \mathbb{I}_d)^{-1} x_1)^{-\alpha}) \\ &= \frac{k}{2(\alpha-1)} \log\left(\frac{(1 - x_1^\top (X^\top X + \sigma^2 \mathbb{I}_d)^{-1} x_1)^\alpha}{1 - \alpha x_1^\top (X^\top X + \sigma^2 \mathbb{I}_d)^{-1} x_1}\right). \end{aligned} \tag{6}$$

We note that the function $\frac{(1-t)^\alpha}{1-\alpha t}$ is a monotonically non-decreasing function of t in the range $0 \leq t < \frac{1}{\alpha}$ for $\alpha > 1$. To see this, note that

$$\frac{\partial}{\partial t} \log\left(\frac{(1-t)^\alpha}{1-\alpha t}\right) = \frac{\partial}{\partial t} \{\alpha \log(1-t) - \log(1-\alpha t)\} = -\frac{\alpha}{1-t} + \frac{\alpha}{1-\alpha t} = \frac{\alpha(\alpha-1)t}{(1-t)(1-\alpha t)}$$

which is positive in the range $0 \leq t < \frac{1}{\alpha}$ (recall that $\alpha > 1$). Thus, to further simplify (6), we will try to find an upper bound on $x_1^\top (X^\top X + \sigma^2 \mathbb{I}_d)^{-1} x_1$. To that end, note that for a general symmetric positive-definite matrix A we have

$$x^\top A^{-1} x \leq \frac{\|x\|^2}{\lambda_{\min}(A)}$$

where equality is achieved whenever x is the eigenvector of A that correspond to $\lambda_{\min}(A)$. Then, using this relation with regard to $X^\top X + \sigma^2 \mathbb{I}_d \succ 0$ and using the identity

$$\lambda_{\min}(X^\top X + \sigma^2 \mathbb{I}_d) = \lambda_{\min}(X^\top X) + \sigma^2$$

yields

$$x_i^\top (X^\top X + \sigma^2 \mathbb{I}_d)^{-1} x_i \leq \frac{\|x_i\|^2}{\lambda_{\min}(X^\top X) + \sigma^2}, \quad \text{for all } i = 1, \dots, n.$$

Since we know that $\max_{i \in [n]} \|x_i\|^2 \leq C_X^2$ we further have

$$x_i^\top (X^\top X + \sigma^2 \mathbb{I}_d)^{-1} x_i \leq \frac{C_X^2}{\lambda_{\min}(X^\top X) + \sigma^2}.$$

This further leads to the next final upper bound on the Rényi divergence:

$$\begin{aligned}
D_\alpha(\mathcal{M}(X) \parallel \mathcal{M}(X')) &= \frac{k}{2(\alpha-1)} \log \left(\frac{(1 - x_1^\top (X^\top X + \sigma^2 \mathbb{I}_d)^{-1} x_1)^\alpha}{1 - \alpha x_1^\top (X^\top X + \sigma^2 \mathbb{I}_d)^{-1} x_1} \right) \\
&\leq \frac{k}{2(\alpha-1)} \log \left(\frac{(1 - \frac{\|x_1\|^2}{\lambda_{\min}(X^\top X) + \sigma^2})^\alpha}{1 - \frac{\alpha \|x_1\|^2}{\lambda_{\min}(X^\top X) + \sigma^2}} \right) \\
&= \frac{k\alpha}{2(\alpha-1)} \log \left(1 - \frac{\|x_1\|^2}{\lambda_{\min}(X^\top X) + \sigma^2} \right) - \frac{k}{2(\alpha-1)} \log \left(1 - \frac{\alpha \|x_1\|^2}{\lambda_{\min}(X^\top X) + \sigma^2} \right),
\end{aligned} \tag{7a}$$

where (7a) requires that $\alpha < \frac{\lambda_{\min}(X^\top X) + \sigma^2}{\|x_1\|^2}$. Then, since a similar analysis holds when we replace x_1 with a general point x_i , the worst case divergence between X and an X'_i that is changed by zeroing out one entry x_i is

$$\begin{aligned}
&\sup_{i \in [n]} D_\alpha(\mathcal{M}(X) \parallel \mathcal{M}(X'_i)) \\
&\leq \sup_{i \in [n]} \left\{ \frac{k\alpha}{2(\alpha-1)} \log \left(1 - \frac{\|x_i\|^2}{\lambda_{\min}(X^\top X) + \sigma^2} \right) - \frac{k}{2(\alpha-1)} \log \left(1 - \frac{\alpha \|x_i\|^2}{\lambda_{\min}(X^\top X) + \sigma^2} \right) \right\} \\
&\leq \frac{k\alpha}{2(\alpha-1)} \log \left(1 - \frac{C_X^2}{\lambda_{\min}(X^\top X) + \sigma^2} \right) - \frac{k}{2(\alpha-1)} \log \left(1 - \frac{\alpha C_X^2}{\lambda_{\min}(X^\top X) + \sigma^2} \right) \\
&\leq \frac{k\alpha}{2(\alpha-1)} \log \left(1 - \frac{C_X^2}{\bar{\lambda}_{\min} + \sigma^2} \right) - \frac{k}{2(\alpha-1)} \log \left(1 - \frac{\alpha C_X^2}{\bar{\lambda}_{\min} + \sigma^2} \right),
\end{aligned} \tag{8a}$$

where (8a) is again by the monotonicity of $\frac{(1-t)^\alpha}{1-\alpha t}$ and since $\lambda_{\min}(X^\top X) \geq \bar{\lambda}_{\min}$, where $\alpha \leq \min_i \left\{ \frac{\bar{\lambda}_{\min} + \sigma^2}{\|x_i\|^2} \right\} = \frac{\bar{\lambda}_{\min} + \sigma^2}{C_X^2}$ and the bound holds whenever $\bar{\lambda}_{\min} + \sigma^2 > C_X^2$. Finally, note that since $\alpha - 1 > 0$ for all $\alpha > 1$ and since $\alpha \log \left(1 - \frac{C_X^2}{\bar{\lambda}_{\min} + \sigma^2} \right) - \log \left(1 - \frac{\alpha C_X^2}{\bar{\lambda}_{\min} + \sigma^2} \right) \geq 0$ for all $\alpha > 1$ (this follows since the function is 0 for $\alpha = 1$ and since its derivative is positive) this upper bound is non-negative and is a valid upper bound on this divergence.

For the case where one row of X is zeroed out, we note that we have $X'^\top X' + \sigma^2 \mathbb{I}_d = X^\top X + \sigma^2 \mathbb{I}_d + x_i x_i^\top$. Then, (6) is replaced with

$$D_\alpha(\mathcal{M}(X) \parallel \mathcal{M}(X')) = \frac{k}{2(\alpha-1)} \log \left(\frac{(1 + x_1^\top (X^\top X + \sigma^2 \mathbb{I}_d)^{-1} x_1)^\alpha}{1 + \alpha x_1^\top (X^\top X + \sigma^2 \mathbb{I}_d)^{-1} x_1} \right).$$

Now, we define the function $f(t; \alpha) = \log \left(\frac{(1-t)^\alpha}{1-\alpha t} \right) - \log \left(\frac{(1+t)^\alpha}{1+\alpha t} \right)$. Then, note that $f(0; \alpha) = 0$ and further since $\alpha > 1$ and $\alpha t < 1$ then

$$\frac{\partial}{\partial t} f(t; \alpha) = 2\alpha \left(\frac{1}{1 - (\alpha t)^2} - \frac{1}{1 - t^2} \right) \geq 0$$

and thus $f(t; \alpha) \geq 0$ for all $t < \frac{1}{\alpha}$, and we get that the maximum between the two divergences is always given by the case where X' contains a zero row. Thus, by finding the σ^2 that makes (8a) equal to ε we guarantee that our mechanism is (α, ε) -Rényi-DP.

It remains to validate that the condition $\alpha \Sigma_1^{-1} + (1 - \alpha) \Sigma_2^{-1} \succ 0$ holds. However, since throughout we have $\Sigma_2 = \Sigma_1 - x_i x_i^\top$ with $\Sigma_1 = X^\top X + \sigma^2 \mathbb{I}_d$, by using the formulas for the inverse of a rank-1 update we get

$$\alpha \Sigma_1^{-1} + (1 - \alpha) \Sigma_2^{-1} = \Sigma_1^{-1/2} \left(\mathbb{I}_d + (1 - \alpha) \cdot \frac{\Sigma_1^{-1/2} x_i x_i^\top \Sigma_1^{-1/2}}{1 - x_i^\top \Sigma_1^{-1} x_i} \right) \Sigma_1^{-1/2}.$$

We note that since $\Sigma_1 \succ 0$, for this term to be positive definite it suffices for the middle matrix to be positive definite. However, since this matrix is a rank-1 update of $\mathbb{I}_d$, its eigenvalues are 1's and an additional eigenvalue that is given by

$$1 + (1 - \alpha) \cdot \frac{\left\| \Sigma_1^{-1/2} x_i \right\|^2}{1 - x_i^\top \Sigma_1^{-1} x_i} = 1 + (1 - \alpha) \cdot \frac{x_i^\top \Sigma_1^{-1} x_i}{1 - x_i^\top \Sigma_1^{-1} x_i}.$$

We note that this term is positive whenever $\alpha \leq \frac{1}{x_i^\top \Sigma_1^{-1} x_i}$. However, since $x_i^\top \Sigma_1^{-1} x_i \leq \frac{\|x_i\|^2}{\lambda_{\min} + \sigma^2}$ this inequality is satisfied by the restrictions we have on the domain of α . $\square$

C.2 Proof of Corollary 1

Proof. We start by defining the difference function

$$\Delta(k, \alpha, \gamma) = \frac{k\alpha}{2\gamma^2} - \frac{k\alpha}{2(\alpha - 1)} \log \left(1 - \frac{1}{\gamma} \right) + \frac{k}{2(\alpha - 1)} \log \left(1 - \frac{\alpha}{\gamma} \right).$$

Our goal is to find when $\Delta(k, \alpha, \gamma) \geq 0$ for $1 < \alpha < \gamma$. We note multiplying by the positive factor $2\gamma^2(\alpha - 1)$ and cancelling the term $k > 0$ gives the equivalent condition

$$G(\alpha) := \alpha(\alpha - 1) - \alpha\gamma^2 \log \left(1 - \frac{1}{\gamma} \right) + \gamma^2 \log \left(1 - \frac{\alpha}{\gamma} \right) \geq 0, \quad 1 < \alpha < w,$$

where w will be specified shortly. On $\alpha = 1$ we further get $G(1) = 0$. Moreover,

$$G'(\alpha) = 2\alpha - 1 - \gamma^2 \log \left(1 - \frac{1}{\gamma} \right) - \frac{\gamma^2}{\gamma - \alpha},$$

and multiplying by $(\gamma - \alpha) > 0$ (recall that $\alpha < \gamma$) shows $G'(\alpha)$ has the same sign as the quadratic

$$\begin{aligned} H(\alpha) &:= (\gamma - \alpha)G'(\alpha) \\ &= (2\alpha - 1)(\gamma - \alpha) - \gamma^2(\gamma - \alpha) \log \left(1 - \frac{1}{\gamma} \right) - \gamma^2 \\ &= -2\alpha^2 + \left(1 + 2\gamma + \gamma^2 \log \left(1 - \frac{1}{\gamma} \right) \right) \alpha - \gamma \left(1 + \gamma + \gamma^2 \log \left(1 - \frac{1}{\gamma} \right) \right). \end{aligned}$$

We define the discriminant to be

$$\Delta_H = \left(1 + 2\gamma + \gamma^2 \log \left(1 - \frac{1}{\gamma} \right) \right)^2 - 8\gamma \left(1 + \gamma + \gamma^2 \log \left(1 - \frac{1}{\gamma} \right) \right)$$

which is non-negative. Thus, H has two real roots

$$\alpha_{\max/\min} = \frac{\gamma^2 \log(1 - 1/\gamma) + 2\gamma + 1 \pm \sqrt{\Delta_H}}{4},$$

and $H(\alpha) \geq 0$ for $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ since the coefficient of the quadratic term $H(\alpha)$ is negative. However, note that $\alpha_{\min} < 1/2$ for all $\gamma > 1$ and moreover $\alpha_{\max} > 1$ for $\gamma > 5/2$ and $\alpha_{\max} < \gamma$ for all $\gamma > 5/2$. Thus, since the derivative is positive and since $G(1) = 0$, setting $w := \alpha_{\max}$ yield $G'(\alpha) \geq 0$ for every $1 < \alpha < w$ and thus the inequality $G(\alpha) \geq 0$ holds throughout that interval, whenever $\gamma > 5/2$. The proof is completed since $\alpha_{\max} > \frac{2\gamma}{5}$ for all $\gamma > 1$. $\square$

C.3 Proof of Theorem 1

We recall that the sensitivity of the minimum eigenvalue $\lambda_{\min}(X^\top X)$ is C_X^2 (see, for example [Sheffet \[2017\]](#), [Wang \[2018\]](#)). Then, by using the standard formula of the Gaussian mechanism [[Dwork et al., 2014](#), Appendix A] we get that $\tilde{\lambda}$ is $(\sqrt{2 \log(3.75/\delta)}/\eta, \delta/3)$ release

of $\lambda_{\min}(X^\top X)$. Using Lemma 1 and Proposition 1, we note that whenever $\lambda_{\min}(X^\top X) + \tilde{\eta}^2 \geq \gamma$ the release of the output in both cases satisfies $(\tilde{\varepsilon}, \delta/3)$ -DP where

$$\tilde{\varepsilon} = \min_{1 < \alpha < \gamma} \left\{ \varphi(\alpha; k, \gamma) + \frac{\log(3/\delta) + (\alpha - 1) \log(1 - 1/\alpha) - \log(\alpha)}{\alpha - 1} \right\}.$$

The first case (when $\gamma \leq \tau$) trivially satisfies this. However, for the second case (whenever $\gamma > \tau$), this is satisfied only if $\tilde{\eta}^2 + \lambda_{\min}(X^\top X) \geq \gamma$, which by using the inequality

$$\tilde{\eta}^2 + \lambda_{\min}(X^\top X) = \gamma - \lambda_{\min}(X^\top X) + \eta C_X^2 \tau - \eta C_X^2 \mathbf{z} + \lambda_{\min}(X^\top X) = \gamma + \eta C_X^2 \tau - \eta C_X^2 \mathbf{z}$$

corresponds to having $\mathbf{z} \geq \tau$ (we note that the case $\tilde{\lambda} = 0$ immediately satisfies $\lambda_{\min}(X^\top X) + \tilde{\eta}^2 \geq \gamma$ since then we have $\tilde{\eta}^2 = \gamma$). Thus,

$$P(\tilde{\eta}^2 + \lambda_{\min}(X^\top X) \leq \gamma) = P(\mathbf{z} \geq \tau) \leq \exp\left\{-\frac{\tau^2}{2}\right\} \leq \frac{\delta}{3}.$$

Then, using simple composition [Dwork et al., 2014, Chapter. 3] and substituting $\tau \geq \sqrt{2 \log(3/\delta)}$ yields the desired result.

C.4 Proof of Theorem 2

Proof. We first establish the performance of a method that adds noise with a general level σ , namely,

$$\theta_{\text{Lin}} := ((SX + \sigma \xi_1)^\top (SX + \sigma \xi_1))^{-1} (SX + \sigma \xi_1)^\top (SY + \sigma \xi_2).$$

Then, we can rewrite θ_{Lin} in the next form

$$\theta_{\text{Lin}} = \underset{\theta}{\operatorname{argmin}} \left\| (S, \xi_1, \xi_2) \left(\begin{pmatrix} Y \\ \vec{0}_d \\ \sigma \end{pmatrix} - \begin{pmatrix} X \\ \sigma \mathbb{I}_d \\ \vec{0}_d^\top \end{pmatrix} \theta \right) \right\|^2.$$

Now, since $\operatorname{rank}((X^\top, \sigma \mathbb{I}_d, \vec{0}_d)^\top) = d$ and since $\sigma^2 \geq 0$, by Pilanci and Wainwright [2015, Corollary. 2], whenever $k > \frac{c_0 d}{\chi^2}$ w.p. at least $1 - c_1 \cdot \exp\{-c_2 k \chi^2\}$ we have

$$L_{X,Y}(\theta_{\text{Lin}}) + \sigma^2 \|\theta_{\text{Lin}}\|^2 \leq (1 + \chi)^2 \left(\|Y - X\theta^*(\sigma^2)\|^2 + \sigma^2 \|\theta^*(\sigma^2)\|^2 + \sigma^2 \right).$$

We note that this further implies that

$$L_{X,Y}(\theta_{\text{Lin}}) \leq (1 + \chi)^2 \left(\|Y - X\theta^*(\sigma^2)\|^2 + \sigma^2 (1 + \|\theta^*\|^2) \right).$$

Thus, we can write

$$\begin{aligned} L_{X,Y}(\theta_{\text{Lin}}) - (1 + \chi)^2 L_{X,Y}(\theta^*) &\leq (1 + \chi)^2 \left(\|Y - X\theta^*(\sigma^2)\|^2 - \|Y - X\theta^*\|^2 + \sigma^2 (1 + \|\theta^*\|^2) \right) \\ &= O\left((1 + \chi)^2 \sigma^2 (1 + \|\theta^*\|^2)\right) \end{aligned}$$

where the last inequality is by Wang [2018, App. B.2]. Now, we note that in both of the cases of the algorithm the magnitude of the added noise is at most $\gamma(C_X^2 + C_Y^2)$, where γ is determined by the calculation done in step 1. Thus, since the bound is monotonically increasing in σ^2 we can further use the upper bound

$$L_{X,Y}(\theta_{\text{Lin}}) - (1 + \chi)^2 L_{X,Y}(\theta^*) = O\left((1 + \chi)^2 \gamma (C_X^2 + C_Y^2) (1 + \|\theta^*\|^2)\right). \quad (9)$$

We further note that

$$\begin{aligned} \varepsilon(\sigma, \gamma, k, \delta) &= \frac{\sqrt{2 \log(3.75/\delta)}}{\sigma} + \min_{1 < \alpha < \gamma} \left\{ \varphi(\alpha; k, \gamma) + \frac{\log(3/\delta) + (\alpha - 1) \log(1 - 1/\alpha) - \log(\alpha)}{\alpha - 1} \right\} \\ &\leq \frac{\sqrt{2k \log(3.75/\delta)}}{\gamma} + \min_{1 < \alpha < 2\gamma/5} \left\{ \frac{k\alpha}{2\gamma^2} + \frac{\log(3/\delta)}{\alpha - 1} \right\} \\ &= \frac{3\sqrt{2k \log(3.75/\delta)}}{\gamma}. \end{aligned}$$

Thus, equating this upper bound to ε suggests further that $\gamma = O\left(\frac{\sqrt{k \log(1/\delta)}}{\varepsilon}\right)$ and using this bound in (9) leads to

$$L_{X,Y}(\theta_{\text{Lin}}) - (1 + \chi)^2 L_{X,Y}(\theta^*) = O\left((1 + \chi)^2 \cdot \frac{\sqrt{k \log(1/\delta)}(C_X^2 + C_Y^2)}{\varepsilon} \cdot (1 + \|\theta^*\|^2)\right).$$

The proof is finished since this holds for any χ under the constraints in the theorem. $\square$

D Utility Guarantees for Logistic Regression

We now demonstrate utility guarantees on our method for DP logistic regression, presented in Section 5.2. Those derived similarly to Theorem 2, by considering both sources of errors: the error of approximating the objective with a polynomial and the empirical error of the linear regression solution. Throughout the proof, we denote by $\hat{\theta}$ the private solution obtained by scaling the output of Algorithm 2 by $-\frac{b_1}{2b_2}$. We also let $\tilde{\theta}^*$ denote the minimizer of the approximated loss, given explicitly by $-\frac{b_1}{2b_2}(X^\top X)^{-1}X^\top Y$. We further denote the empirical logistic loss via

$$L_{X,Y}(\theta) := -\frac{1}{n} \sum_{i=1}^n \log(1 + \exp\{-y_i x_i^\top \theta\})$$

and the approximated empirical logistic loss via

$$\begin{aligned} \tilde{L}_{X,Y}(\theta) &:= b_0 + b_1 \theta^\top \left(\frac{1}{n} X^\top Y\right) + b_2 \theta^\top \left(\frac{1}{n} X^\top X\right) \theta \\ &= b_0 - \frac{b_1^2}{4nb_2} \|Y\|^2 + \frac{b_2}{n} \left\| -\frac{b_1}{2b_2} Y - X\theta \right\|^2 \\ &= b_0 - \frac{b_1^2}{4nb_2} \|Y\|^2 + \frac{b_2}{n} F\left(X, -\frac{b_1}{2b_2} Y, \theta\right). \end{aligned}$$

We note that Corollary 2 guarantees that our logistic regression solution obtained by minimizing this surrogate is private. We now present the utility guarantees on this solution.

Corollary 3. *Assume that $\|x_i\|_2^2 \leq C_X^2$, $|y_i|^2 \leq C_Y^2$ and $|y_i x_i^\top \tilde{\theta}^*| \leq Q$ and $|y_i x_i^\top \hat{\theta}| \leq Q$ for all $i \in [n]$ and for some finite $Q > 0$. Let (b_0, b_1, b_2) chosen such that*

$$\max_{s \in [-Q, Q]} |-\log(1 + e^{-s}) - (b_0 + b_1 s + b_2 s^2)| \leq q. \quad (10)$$

Then, there exist universal constants c_0, c_1, c_2 such that for any χ satisfying $k\chi^2 > c_0$ and the following holds with probability at least $1 - c_1 \cdot \exp\{-c_2 k\chi^2\}$:

$$\begin{aligned} L_{X,Y}(\theta^*) - (1 + \chi)^2 L_{X,Y}(\hat{\theta}) &- (1 + (1 + \chi)^2)q + (1 - (1 + \chi)^2) \left(b_0 - \frac{b_1^2}{4b_2}\right) \\ &= O\left((1 + \chi)^2 \frac{\sqrt{k \log(1/\delta)} C_X^2}{n\varepsilon} \left(1 + \|\tilde{\theta}^*\|^2\right)\right). \end{aligned}$$

Proof. We first note that

$$L_{X,Y}(\theta^*) - (1 + \chi)^2 L_{X,Y}(\hat{\theta}) \leq L_{X,Y}(\tilde{\theta}^*) - (1 + \chi)^2 L_{X,Y}(\hat{\theta}) \quad (11a)$$

$$\begin{aligned} &= L_{X,Y}(\tilde{\theta}^*) - \tilde{L}_{X,Y}(\tilde{\theta}^*) + \tilde{L}_{X,Y}(\tilde{\theta}^*) - (1 + \chi)^2 \tilde{L}_{X,Y}(\hat{\theta}) \\ &\quad + (1 + \chi)^2 \tilde{L}_{X,Y}(\hat{\theta}) - (1 + \chi)^2 L_{X,Y}(\hat{\theta}) \\ &\leq (1 + (1 + \chi)^2)q + \tilde{L}_{X,Y}(\tilde{\theta}^*) - (1 + \chi)^2 \tilde{L}_{X,Y}(\hat{\theta}) \quad (11b) \end{aligned}$$

$$\begin{aligned} &= (1 + (1 + \chi)^2)q + (1 - (1 + \chi)^2) \left(b_0 - \frac{b_1^2}{4nb_2} \|Y\|^2 \right) \quad (11c) \\ &\quad + \frac{b_2}{n} \left(F\left(X, -\frac{b_1}{2b_2}Y, \tilde{\theta}^*\right) - (1 + \chi)^2 F\left(X, -\frac{b_1}{2b_2}Y, \hat{\theta}\right) \right) \end{aligned}$$

where (11a) is by the optimality of θ^* , (11b) is by (10) and (11c) is by the definition of $\tilde{L}_{X,Y}(\theta)$ and by the assumptions $|y_i x_i^\top \tilde{\theta}^*| \leq Q$ and $|y_i x_i^\top \hat{\theta}| \leq Q$. Then, the final result follows by using Theorem 2 and since in this case $|y_i| = 1$, thus $\|Y\| = n$ and $C_Y = 1$. $\square$

When we take $\chi \ll 1$, the bound acquires an extra $2q$ term in the excess risk, introduced by the polynomial approximation.

E Algorithms for DP regression

E.1 Linear regression

E.1.1 AdaSSP

Algorithm 3 AdaSSP [Wang, 2018]

Require: Dataset (X, Y) ; Privacy parameters ε, δ ; Bounds: $\max_{i \in [n]} \|x_i\|^2 \leq C_X^2, \max_{i \in [n]} |y_i| \leq C_Y$.

- 1: Calculate the minimum eigenvalue $\lambda_{\min}(X^\top X)$.
 - 2: Privately release $\tilde{\lambda}_{\min} = \max \left\{ \lambda_{\min} + \frac{\sqrt{\log(6/\delta)} C_X^2}{\varepsilon/3} \mathbf{z} - \frac{\log(6/\delta)}{\varepsilon/3} C_X^2, 0 \right\}$ where $\mathbf{z} \sim \mathcal{N}(0, 1)$.
 - 3: Set $\lambda = \max \left\{ 0, \frac{\sqrt{d \log(6/\delta) \log(2d^2/\varepsilon)} C_X^2}{\varepsilon/3} - \tilde{\lambda}_{\min} \right\}$.
 - 4: Privately release $\widetilde{X^\top X} = X^\top X + \frac{\sqrt{\log(6/\delta)} C_X^2}{\varepsilon/3} \xi_1$ for $\xi_1 \sim \mathcal{N}_{\text{sym}}(0, \mathbb{I}_d)$.
 - 5: Privately release $\widetilde{X^\top y} = X^\top y + \frac{\sqrt{\log(6/\delta)} C_X C_Y}{\varepsilon/3} \xi_2$ for $\xi_2 \sim \mathcal{N}(0, \mathbb{I}_d)$.
 - 6: **return** $\tilde{\theta} \leftarrow \left(\widetilde{X^\top X} + \lambda \mathbb{I}_d \right)^{-1} \widetilde{X^\top y}$
-

E.1.2 Algorithm 1 from [Sheffet \[2017\]](#) and our proposed modification

Algorithm 4 Sheffet’s Algorithm (Original)

Require: Dataset (X, Y) ; Privacy parameters ε, δ ; Bounds: $\max_{i \in [n]} \|x_i\|^2 \leq C_X^2, \max_{i \in [n]} |y_i|^2 \leq C_Y^2$; Hyperparameter k .

- 1: Compute $\lambda_{\min} := \lambda_{\min}((X, Y)^\top (X, Y))$.
- 2: Set $\gamma \leftarrow \frac{4(C_X^2 + C_Y^2)}{\varepsilon} \left(\sqrt{2k \log\left(\frac{8}{\delta}\right)} + 2 \log\left(\frac{8}{\delta}\right) \right)$.
- 3: Sample $S \sim \mathcal{N}(0, \mathbb{I}_{k \times n})$.
- 4: **if** $\lambda_{\min} > \gamma + \frac{4(C_X^2 + C_Y^2) \log(1/\delta)}{\varepsilon}$ **for** $z \sim \text{Lap}\left(\frac{4(C_X^2 + C_Y^2)}{\varepsilon}\right)$ **then**
- 5: **return** $\tilde{\theta} \leftarrow ((SX)^\top (SX))^{-1} (SX)^\top (SY)$
- 6: **else**
- 7: Sample noises $\xi_1 \sim \mathcal{N}(0, \mathbb{I}_{k \times d}), \xi_2 \sim \mathcal{N}(0, \mathbb{I}_k)$.
- 8: **return** $\tilde{\theta} \leftarrow ((SX + \gamma \xi_1)^\top (SX + \gamma \xi_1))^{-1} (SX + \gamma \xi_1)^\top (SY + \gamma \xi_2)$

Algorithm 5 Sheffet’s Algorithm with Our Analysis

Require: Dataset (X, Y) ; Privacy parameters ε, δ ; Bounds: $\max_{i \in [n]} \|x_i\|^2 \leq C_X^2, \max_{i \in [n]} |y_i|^2 \leq C_Y^2$; Hyperparameter k .

- 1: Compute $\lambda_{\min} := \lambda_{\min}((X, Y)^\top (X, Y))$.
- 2: Set γ s.t. $\min_{1 < \alpha < \gamma} \left\{ \varphi(\alpha; k, \gamma) + \log\left(1 - \frac{1}{\alpha}\right) - \frac{\log(\alpha\delta)}{\alpha-1} \right\} \leq \varepsilon/2$.
- 3: Sample $S \sim \mathcal{N}(0, \mathbb{I}_{k \times n})$.
- 4: **if** $\lambda_{\min} > \gamma + \frac{4(C_X^2 + C_Y^2) \log(1/\delta)}{\varepsilon}$ **for** $z \sim \text{Lap}\left(\frac{4(C_X^2 + C_Y^2)}{\varepsilon}\right)$ **then**
- 5: **return** $\tilde{\theta} \leftarrow ((SX)^\top (SX))^{-1} (SX)^\top (SY)$
- 6: **else**
- 7: Sample noises $\xi_1 \sim \mathcal{N}(0, \mathbb{I}_{k \times d}), \xi_2 \sim \mathcal{N}(0, \mathbb{I}_k)$.
- 8: **return** $\tilde{\theta} \leftarrow ((SX + \gamma \xi_1)^\top (SX + \gamma \xi_1))^{-1} (SX + \gamma \xi_1)^\top (SY + \gamma \xi_2)$

E.2 Logistic Regression

E.2.1 Objective Perturbation

Algorithm 6 Objective Perturbation [\[Kifer et al., 2012\]](#)

Require: Dataset (X, Y) ; privacy parameters ε and δ ; Bound $\|x_i\| \leq C_X$ for all $i \in [n]$;

- 1: Set $\sigma = \frac{\sqrt{4\varepsilon + 8 \log(2/\delta)}}{\varepsilon} C_X$ and $\Delta = \frac{C_X^2}{2\varepsilon}$.
- 2: Sample $b \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$
- 3: **return** $\tilde{\theta} \leftarrow \underset{\theta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n -\frac{1}{n} \log(1 + \exp\{-y_i x_i^\top \theta\}) + \frac{b^\top \theta}{n} + \frac{\Delta}{2n} \|\theta\|^2 \right\}$.

F Experimental Details

All the linear regression experiments were run on 12th Gen Intel(R) Core(TM) i7-1255U, and all the logistic experiments were run on an NVIDIA A100 GPU.

F.1 Linear Regression

For the linear regression experiments, we used four datasets. The first two are real-world datasets: the Tecator dataset [\[Thodberg, 2015\]](#) and the Communities and Crime dataset [\[Red-](#)

mond and Baveja, 2002]. We have used a random train-test split of 80%/20% for generating a train and a test set.

The other two are synthetic datasets where the responses were generated via the linear model $y_i = x_i^\top \theta_0 + \sigma \xi_i$, with θ_0 sampled as a unit vector uniformly from the $(d-1)$ -dimensional sphere, $\xi_i \sim \text{Unif}(-1, 1)$, and $\sigma = 0.1$.

In the first synthetic dataset (termed *Gaussian dataset*), the parameters were $n = 8192$, $d = 512$, and the covariates were sampled as $x_i \sim \mathcal{N}(0, \mathbf{Q}\mathbf{Q}^\top)$, where $\mathbf{Q} \in \mathbb{R}^{d \times q}$ is a random orthogonal matrix with $q = 4$, ensuring the data lies on a 4-dimensional subspace. The matrix $\mathbf{Q}$ was generated via QR decomposition of a random matrix with i.i.d. standard Gaussian entries.

The second synthetic dataset (termed the *synthetic dataset*) was constructed as follows. First, we sampled latent covariates $\tilde{x}_i \sim \mathcal{N}(0, \mathbb{I}_2)$. Then, we generated final covariates using a two-layer neural network:

$$x_i = \phi(\mathbf{W}_2 \phi(\mathbf{W}_1 \tilde{x}_i + \mathbf{b}_1) + \mathbf{b}_2),$$

where $\phi(\cdot)$ is the element-wise sigmoid function, $\mathbf{W}_1 \sim \mathcal{N}(0, \mathbb{I}_{100 \times 2})$, $\mathbf{W}_2 \sim \mathcal{N}(0, \mathbb{I}_{d \times 100})$, $\mathbf{b}_1 \sim \mathcal{N}(0, 10^{-6} \cdot \mathbb{I}_{100})$, and $\mathbf{b}_2 \sim \mathcal{N}(0, 10^{-6} \cdot \mathbb{I}_d)$. In our experiments, we have fixed $d = 2^9$.

For both synthetic datasets, the train and test sets were generated independently, using the same fixed θ_0 but with independent covariates and additive noise.

In all cases, we normalized the training data so that the maximum ℓ_2 -norm of any training sample was 1. The test data was scaled using the same normalization factor as the training data.

The baseline (non-private) estimator was computed as $\hat{\theta} = (X^\top X + \lambda \mathbb{I}_d)^{-1} X^\top Y$ for $\lambda = 10^{-6}$, ensuring invertibility in all cases. We report the mean squared error (MSE) for both the train and the test set, computed as the squared error in predicting y_i via $x_i^\top \hat{\theta}$. All results are averaged over 250 independent trials, and we report both the empirical means and confidence intervals.

F.1.1 Algorithms

Our algorithm was implemented as described in Algorithm 2. The AdaSSP algorithm was implemented based on [Wang, 2018, Alg. 2], following the procedure detailed in Appendix E.1.1. Our second baseline, from [Sheffet, 2017, Alg. 1], was implemented according to the description in Appendix E.1.2. This implementation matches that of [Sheffet, 2017, Alg. 1], except for an adjustment to account for a factor of 2 in the parameter w , which arises due to using the zero-out neighboring definition rather than the replacement definition. In the variant of this baseline that incorporates our improved privacy analysis, we replaced the original noise calibration with bounds derived from Lemma 1, translated to (ε, δ) -DP using the conversion provided in Proposition 1 (see also Appendix E.1.2).

F.2 Logistic Regression

In this set of experiments, we trained a logistic regression classifier for a binary classification task without applying any regularization. Our non-private baseline is the standard `LogisticRegression` solver from the `sklearn.linear_model` library. The private baselines are the objective perturbation method (described in Appendix E.2), where the minimization is carried out using `torch.optim.LBFGS` with a maximum of 500 iterations and a tolerance of 10^{-6} , following the setup of [Guo et al., 2020], and DP-SGD [Abadi et al., 2016] as implemented in Opacus [Yousefpour et al., 2021] with a batch size of 1024, 10 epochs, and a learning rate of 0.5. While DP-SGD may benefit from hyperparameter tuning, our method requires only one; to avoid spending additional privacy budget on tuning, we use a fixed, reasonable configuration. We also fixed the parameter k on $4.5d$.

We conducted experiments on the Fashion-MNIST [Xiao et al., 2017] and the CIFAR100 [Krizhevsky and Hinton, 2009] datasets, using the implementations provided in `torchvision.datasets`. From each dataset, we selected only the samples corresponding to classes 3 and 8, and relabeled them as $-1, 1$ to fit the binary classification setting. We used

the standard PyTorch train/test splits and normalized the training data by the maximum L_2 norm across all training samples, ensuring that each training sample has a norm of at most 1. The same normalization factor was then applied to the test set. The train and test loaders were generated using `torch.utils.data.DataLoader` with shuffling enabled. In Appendix G.2 we present additional simulations with the CIFAR10 [Krizhevsky and Hinton, 2009] and the MNIST [LeCun and Cortes, 2010] datasets.

The network architecture used is a compact convolutional neural network for RGB image classification. It consists of two convolutional layers with ReLU activations and max pooling, reducing the input to a 64-channel feature map of size 8×8 . The flattened features are passed through a fully connected layer with 128 hidden units and ReLU, followed by a final linear layer that outputs class logits. In both of the experiments, we have first trained this network end-to-end using the DP-SGD primitive implemented in Opacus [Yousefpour et al., 2021], where we have set the clipping parameter to 4.0, learning rate to 0.001, the number of epochs to 20, and the batch size to 500.

Performance metrics are averaged over 50 independent runs, and as before, we report test accuracy along with confidence intervals. Runtime comparisons show the ratio of execution times for the largest simulated ε .

G Additional Experiments

G.1 Linear Regression

We have simulated additional four datasets: the Boston housing dataset [Harrison Jr and Rubinfeld, 1978] that contains 506 measurements of 13-dimensional features with the goal of predicting house prices in the Boston area, the Wine quality dataset [wine, 2009] which contains 1359 measurements of 11-dimensional features, with the goal of predicting wine quality, the Bike sharing dataset [bike, 2019] with the goal of predicting the count of rental bikes, and another artificial dataset that follows the same description as that of the Gaussian dataset but now with i.i.d. features where the distribution of each entry is $\text{Unif}([-1, 1])$. The additional results are presented in Figure 4.

G.2 Logistic Regression

We have simulated two additional datasets: the CIFAR10 [Krizhevsky and Hinton, 2009] and the MNIST [LeCun and Cortes, 2010] datasets, using the same logistic regression setting. The additional results are presented in Figure 5. Both settings demonstrate the computational improvement of our method, as well as utility improvement for a range of the simulated values of ε .

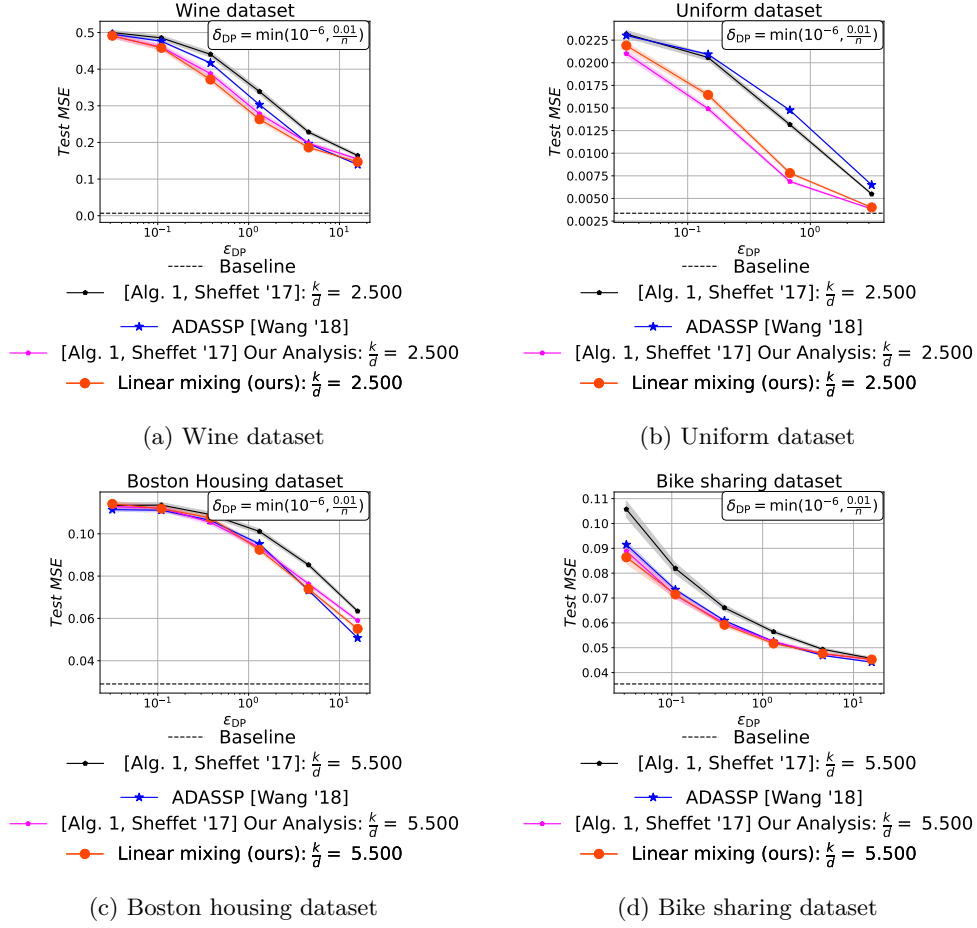


Figure 4: Linear mixing performance on the additional four linear regression tasks.

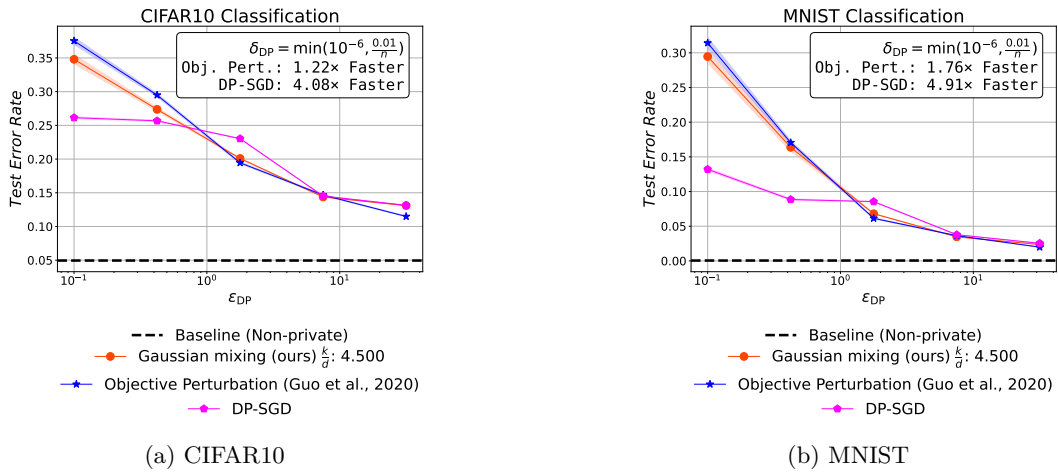


Figure 5: DP logistic regression using a privately trained CNN feature extractor on binary subsets of CIFAR10 and MNIST.