

## 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 Proof of Lemma 1

The proof relies on the next lemma, which establishes the  $\alpha$ -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,*

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

$$- \frac{1}{2(\alpha - 1)} \log \left( \frac{\det(\Sigma_1 + \alpha(\Sigma_2 - \Sigma_1))}{(\det(\Sigma_1))^{1-\alpha} (\det(\Sigma_2))^\alpha} \right)$$

for all  $\alpha$  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  $A$ , we have [Brookes, 2020, Section. 3.4]

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

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  $D_\alpha(\mathcal{M}(X) \parallel \mathcal{M}(X'))$ :

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

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:

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

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

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

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

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

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  $\sigma^2$  that makes (8a) equal to  $\varepsilon$  we guarantee that our mechanism is  $(\alpha, \varepsilon)$ -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  $\alpha$ .

□

## C 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  $\gamma > 2.5$ . 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  $\gamma > 1$ .  $\square$

## D 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.

## E Minimizing $\tilde{\varepsilon}(\eta, \gamma, k, \delta)$

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  $\varepsilon$  by increasing  $\eta$  and  $\gamma$ . 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  $\varepsilon$ , 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, note that by Corollary 1 we know 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  $\gamma$  and can be made arbitrarily small by increasing  $\gamma$ . The result then follows since the first term in (2) is monotonically decreasing in  $\eta$ , and holds further in the case where  $\eta = \frac{\gamma}{\sqrt{k}}$  by picking a sufficiently large  $\gamma$ .

## F Proof of Theorem 2

*Proof.* We first establish the performance of a method that adds noise with a general level  $\sigma$ , 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  $\theta_{\text{Lin}}$  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, Corrolary. 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  $\gamma$  is determined by the calculation done in step 1. Thus, since the bound is monotonically increasing in  $\sigma^2$  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  $\varepsilon$  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  $\chi$  under the constraints in the theorem.  $\square$

## G 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  $\|x_i\|_2^2 \leq C_X^2$ ,  $|y_i| \leq C_Y$  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  $\chi$  satisfying  $k\chi^2 > c_0 d$  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

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

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

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

where (11a) is by the optimality of  $\theta^*$ , (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.

## H Algorithms: Linear Regression

### H.1 AdaSSP

---

**Algorithm 3** AdaSSP [Wang, 2018]

---

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

- 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**  $\tilde{\theta} \leftarrow \left( \widetilde{X^\top X} + \lambda \mathbb{I}_d \right)^{-1} \widetilde{X^\top y}$

---

### H.2 Algorithm 1 from [Sheffet, 2017]

---

**Algorithm 4** Sheffet's Algorithm [Sheffet, 2017, Algorithm 1]

---

**Require:** Dataset  $(X, Y)$ ; Privacy parameters  $\varepsilon, \delta$ ; 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  $\mathbf{S} \sim \mathcal{N}(0, \mathbb{I}_{k \times n})$ .
- 4: **if**  $\lambda_{\min} > \gamma + \mathbf{z} + \frac{4(C_X^2 + C_Y^2) \log(1/\delta)}{\varepsilon}$  for  $\mathbf{z} \sim \text{Lap}\left(\frac{4(C_X^2 + C_Y^2)}{\varepsilon}\right)$  **then**
- 5:     **return**  $\tilde{\theta} \leftarrow ((\mathbf{S}X)^\top (\mathbf{S}X))^{-1} (\mathbf{S}X)^\top (\mathbf{S}Y)$
- 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**  $\tilde{\theta} \leftarrow ((\mathbf{S}X + \gamma \xi_1)^\top (\mathbf{S}X + \gamma \xi_1))^{-1} (\mathbf{S}X + \gamma \xi_1)^\top (\mathbf{S}Y + \gamma \xi_2)$

---



---

**Algorithm 5** Sheffet's Algorithm with Our Analysis

---

**Require:** Dataset  $(X, Y)$ ; Privacy parameters  $\varepsilon, \delta$ ; 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  $\gamma$  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  $\mathbf{S} \sim \mathcal{N}(0, \mathbb{I}_{k \times n})$ .
- 4: **if**  $\lambda_{\min} > \gamma + \mathbf{z} + \frac{4(C_X^2 + C_Y^2) \log(1/\delta)}{\varepsilon}$  for  $\mathbf{z} \sim \text{Lap}\left(\frac{4(C_X^2 + C_Y^2)}{\varepsilon}\right)$  **then**
- 5:     **return**  $\tilde{\theta} \leftarrow ((\mathbf{S}X)^\top (\mathbf{S}X))^{-1} (\mathbf{S}X)^\top (\mathbf{S}Y)$
- 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**  $\tilde{\theta} \leftarrow ((\mathbf{S}X + \gamma \xi_1)^\top (\mathbf{S}X + \gamma \xi_1))^{-1} (\mathbf{S}X + \gamma \xi_1)^\top (\mathbf{S}Y + \gamma \xi_2)$

---



## I Algorithms: Logistic Regression

### I.1 Objective Perturbation

---

**Algorithm 6** Objective Perturbation [Kifer et al., 2012]

---

**Require:** Dataset  $(X, Y)$ ; privacy parameters  $\varepsilon$  and  $\delta$ ; 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  $\mathbf{b} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$
  - 3: **return**  $\tilde{\theta} \leftarrow \underset{\theta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n -\frac{1}{n} \log(1 + \exp\{-y_i x_i^\top \theta\}) + \frac{\mathbf{b}^\top \theta}{n} + \frac{\Delta}{2n} \|\theta\|^2 \right\}$ .
- 

## J Experimental Details

All the experiments were run on an NVIDIA A100 GPU.

### J.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 [Redmond 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  $\theta_0$  sampled as a unit vector uniformly from the  $(d-1)$ -dimensional sphere,  $\xi_i \sim \operatorname{Unif}(-1, 1)$ , 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  $\theta_0$  but with independent covariates and additive noise.

In all cases, we normalized the training data so that the maximum  $\ell_2$ -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.

#### J.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 H.1. Our second baseline, from [Sheffet, 2017, Alg. 1], was implemented according to the description in Appendix H.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  $(\epsilon, \delta)$ -DP using the conversion provided in Proposition 1 (see also Appendix H.2).

## J.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 I), 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 none; 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 K.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  $8 \times 8$ . 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  $\epsilon$ .

## K Additional Experiments

### K.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.

### K.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

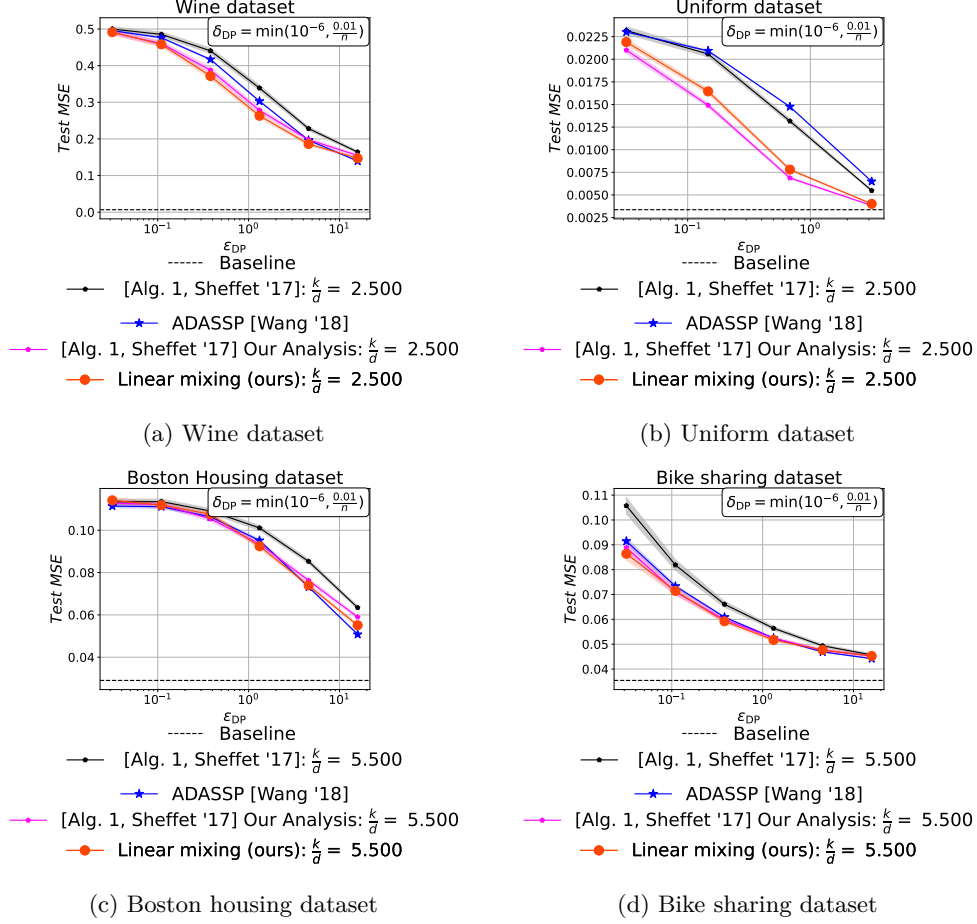


Figure 4: Linear mixing performance on the additional four linear regression tasks.

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  $\varepsilon$ .

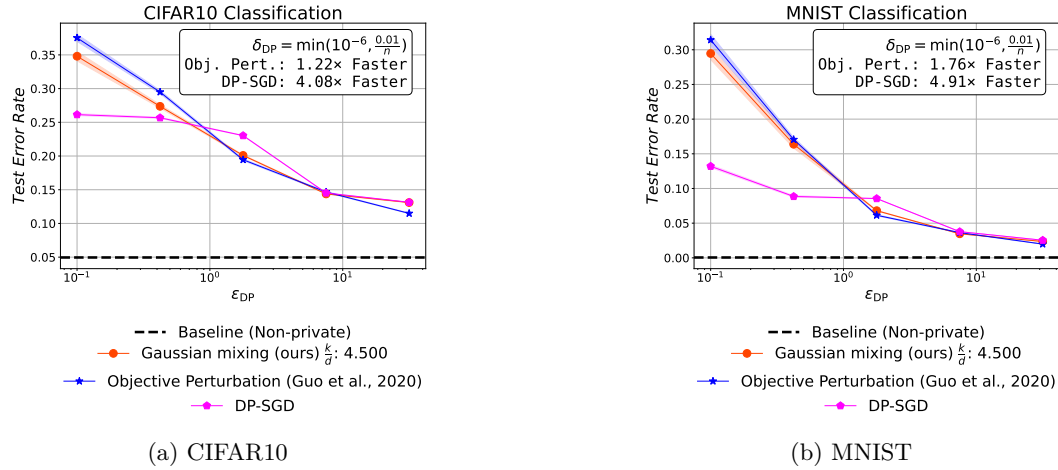


Figure 5: DP logistic regression using a privately trained CNN feature extractor on binary subsets of CIFAR10 and MNIST.