Supplementary Materials: Differentially Private Truncation of Unbounded Data via Public Second Moments

Zilong Cao[1], Xuan Bi[2] and Hai Zhang[⋆,1]

[1]*School of Mathematics, Northwest University, China*

[2]*Department of Information and Decision Sciences, University of Minnesota, USA*

Overall Organization of Appendix: This is the appendix of the paper as the supplementary materials, *Differentially Private Truncation of Unbounded Data via Public Second Moments*. Appendix 1 lists the key notation used throughout the paper to check. Appendix 2 provides more detailed theoretical results on the comparison of the inverse bounds, the DP ridge, and DP logistic regression, and extends to the generalized linear regression based on PMT. And the appendix contains all technical proofs for the paper. Appendix 3 lists the useful tools in the main theorems. Appendix 4 provides the proofs of the main theorems in this paper. Appendix 5 contains the proofs of lemmas in the context. Appendix 6 provides extra experiments.

1 Appendix: Key Notation

Unless otherwise specified, bold uppercase letters are used for matrices, bold lowercase letters for vectors, and regular non-bold letters for scalar quantities. Table 1 summarizes the notation specific to the private data, public data, and the public-moment transformation (PMT).

⋆Corresponding author (zhanghai@nwu.edu.cn)

Table 1: Key notation for private, public, and transformed data.

Symbol	Meaning
Private and public samples	
$\{(\boldsymbol{\xi}_i, y_{\xi_i})\}_{i=1}^{n_\xi}$	Private sample. Here $\boldsymbol{\xi}_i \in \mathbb{R}^d$ is the private covariate vector and y_{ξ_i} is its response.
$\{(\boldsymbol{\upsilon}_i, y_{\upsilon_i})\}_{i=1}^{n_\upsilon}$	Public sample used to estimate second-moment and response-scale information.
n_ξ, n_υ	Private and public sample sizes, respectively.
$\boldsymbol{\Xi} \in \mathbb{R}^{n_\xi \times d}$	Private design matrix with rows $\boldsymbol{\xi}_i^T$.
$\boldsymbol{\Upsilon} \in \mathbb{R}^{n_\upsilon \times d}$	Public design matrix with rows $\boldsymbol{\upsilon}_i^T$.
$\mathbf{y}_\xi$, $\mathbf{y}_\upsilon$	Private and public response vectors.
$\hat{\boldsymbol{\beta}}$	Original estimator from private samples.
Public moments and scales	
$\hat{\Sigma}_\upsilon$	Public covariate second moment, $\hat{\Sigma}_\upsilon = n_\upsilon^{-1}\boldsymbol{\Upsilon}^T\boldsymbol{\Upsilon}$.
$\hat{\sigma}_\upsilon^2$	Public response second moment, $\hat{\sigma}_\upsilon^2 = n_\upsilon^{-1}\sum_{i=1}^{n_\upsilon} y_{\upsilon_i}^2$.
Σ	Population covariate second moment shared by the private and public covariates under the benchmark matched-distribution setting.
PMT-transformed private data	
$\tilde{\boldsymbol{\xi}}_i$	PMT-transformed private covariate, $\tilde{\boldsymbol{\xi}}_i = \hat{\Sigma}_\upsilon^{-1/2}\boldsymbol{\xi}_i$, followed by truncation when required.
$\tilde{\boldsymbol{\Xi}}$	Transformed private design matrix with rows $\tilde{\boldsymbol{\xi}}_i^T$.
$\tilde{y}_{\xi_i}$	Transformed private response, $\tilde{y}_{\xi_i} = \hat{\sigma}_\upsilon^{-1} y_{\xi_i}$ in DP-PMTRR.
$\tilde{\mathbf{y}}_\xi$	Transformed private response vector with entries $\tilde{y}_{\xi_i}$.
$\tilde{\Sigma}_\xi$	Transformed private covariate second moment, $\tilde{\Sigma}_\xi = n_\xi^{-1}\tilde{\boldsymbol{\Xi}}^T\tilde{\boldsymbol{\Xi}}$.
$\tilde{\Sigma}_{\xi y}$	Transformed covariate–response moment, $\tilde{\Sigma}_{\xi y} = n_\xi^{-1}\tilde{\boldsymbol{\Xi}}^T\tilde{\mathbf{y}}_\xi$.
$\tilde{\boldsymbol{\beta}}$	Estimator from PMT-transformed samples.
DP quantities after transformation	
$\tilde{\Sigma}_{DP}$	Privatized transformed second moment, usually $\tilde{\Sigma}_{DP} = \tilde{\Sigma}_\xi + \mathbf{G}$.
$\mathbf{G}$	Symmetric Gaussian matrix noise added to the transformed second moment with the scale σ_1.
$\mathbf{g}$	Gaussian vector noise added to the covariate–response moment with the scale σ_2.
σ_1, σ_2	Noise scales for matrix-valued and vector-valued Gaussian mechanisms, respectively.
$\tilde{\boldsymbol{\beta}}^{DP}$	DP estimator computed by PMT-transformed data.
$\bar{\boldsymbol{\beta}}^{DP}$	Original-domain DP estimator recovered from the transformed estimator; for DP-PMTRR, $\bar{\boldsymbol{\beta}}^{DP} = \hat{\sigma}_\upsilon\hat{\Sigma}_\upsilon^{-1/2}\tilde{\boldsymbol{\beta}}^{DP}$.

2 Additional Results

In this section, we first give the detailed comparison about the inverse error bounds. Second, we propose the results about private-data-only methods, including differentially private ridge regression (DP-RR) and differently private logistic regression (DP-LR). Third, we provide the more detailed theoretical analysis about our method (DP-PMTLR). Finally, we extend to the generalized linear models (GLM) and provide the algorithm and theory.

2.1 Comparison Detail of the Inverse Bounds

We compare the bounds in **Theorem 5** and **Theorem 6**. For convenience, we can consider that the amount of private data n_ξ is enough and makes $\lambda_{min}(\Sigma)(1-O(\sqrt{\frac{d}{n_\xi}}+\sqrt{\frac{\log(1/\eta)}{n_\xi}}))^2 = \lambda_{min}(\Sigma)(1-o(1))^2$. Actually, the appropriate regularization parameter λ is very small so that it doesn't produce large bias to the original matrix; especially, it may be $o(1)$ compared to large $\lambda_{min}(\Sigma)$, or approximate to small $\lambda_{min}(\Sigma)$. Hence, in **Theorem 6**, the necessary n_ξ to hold the stable condition is simplified as

$$\frac{\sqrt{d^3\log(\frac{2d}{\eta})}}{\mu n_\xi}\cdot\frac{d^{-1}\mathrm{Tr}(\Sigma)+\log(\frac{n_\xi}{\eta})}{\lambda_{min}(\Sigma)\big(1-o(1)\big)^2+o(1)} \simeq \frac{\sqrt{d^3\log(\frac{2d}{\eta})}}{\mu n_\xi}\cdot(\bar{\kappa}(\Sigma)+\|\Sigma^{-1}\|\log(\frac{n_\xi}{\eta})) \leqslant \frac{1}{2},$$

and the error bound is simplified as

$$\frac{\sqrt{d^3\log(\frac{2d}{\eta})}}{\mu n_\xi}\cdot\frac{d^{-1}\mathrm{Tr}(\Sigma)+\log(\frac{n_\xi}{\eta})}{\lambda^2_{min}(\Sigma)\big(1-o(1)\big)^4+o(1)} \simeq \frac{\sqrt{d^3\log(\frac{2d}{\eta})}}{\mu n_\xi}\cdot\|\Sigma^{-1}\|\cdot(\bar{\kappa}(\Sigma)+\|\Sigma^{-1}\|\log(\frac{n_\xi}{\eta})),$$

where $\bar{\kappa}(\Sigma) = d^{-1}\sum_{i=1}^{d}\kappa_i(\Sigma) \geqslant 1$, $\kappa_i(\Sigma) = \frac{\lambda_i(\Sigma)}{\lambda_{\min}(\Sigma)} \geqslant 1$ and $\lambda_i(\cdot)$ is the i-th eigenvalue. It is worth noting that an ill-conditioned matrix can significantly inflate the first some condition numbers, thereby making $\bar{\kappa}(\Sigma)$ disproportionately large. This behavior is commonly encountered in practical scenarios. Then, we analyze the advantages of **Theorem 5**.

- **Few necessary public data.** U and L tend to 1 fast as the increase of the amount of public data n_υ. $U = \left(1-O\left(\sqrt{\frac{d}{n_\upsilon}}+\sqrt{\frac{\log(1/\eta)}{n_\upsilon}}\right)\right)^{-2}$ and $L = \left(1+O\left(\sqrt{\frac{d}{n_\upsilon}}+\sqrt{\frac{\log(1/\eta)}{n_\upsilon}}\right)\right)^{-2}$ show that the necessary amount of public data n_υ controlling the ill-condition just needs $O(d+\log(1/\eta))$ to make U and L tend to 1.

- **Weakly impacted by Σ.** The bounds and the condition about the necessary amount of private data are independent of the average condition number $\bar{\kappa}(\Sigma)$ and eliminate a factor $\|\Sigma^{-1}\|$.

  1. **Better robustness condition.** That is a significant improvement for resisting the impact of DP. Namely, it makes the stable condition,

  $$\frac{\sqrt{d^3\log(\frac{2d}{\eta})}(1+\log(\frac{n_\xi}{\eta}))}{\mu n_\xi\cdot(L(1-o(1))^2+\lambda\|\hat{\Sigma}_\upsilon\|^{-1})} \simeq \frac{\sqrt{d^3\log(\frac{2d}{\eta})}}{\mu n_\xi}\cdot\log(\frac{n_\xi}{\eta}) \leqslant \frac{1}{2},$$

  easier to satisfy, meaning better robustness and stable efficiency. Particularly, the term $\lambda\|\hat{\Sigma}_\upsilon\|^{-1}$ degrades the impact of the regularization parameter.

  2. **Tighter error bound.** The error bound of **Theorem 5** is simplified as

  $$\frac{\sqrt{d^3\log(\frac{2d}{\eta})}}{\mu n_\xi}\cdot\frac{\|\hat{\Sigma}_\upsilon^{-1}\|(1+\log(\frac{2n_\xi}{\eta}))}{L^2(1-o(1))^4+o(1)} \simeq \frac{\sqrt{d^3\log(\frac{2d}{\eta})}}{\mu n_\xi}\cdot\log(\frac{n_\xi}{\eta})\cdot\|\Sigma^{-1}\|,$$

  where the public data n_υ makes L tend to 1 and $\|\hat{\Sigma}_\upsilon^{-1}\|$ tend to $\|\Sigma^{-1}\|$. A smaller

error bound represents improved utility of the DP inverse second-moment matrix, making the DP estimation substantially more reliable.

- **Weakly depending on the regularization parameter.** The robustness and utility are dependent on the regularization parameter λ weakly. The λ term is always divided by the largest eigenvalue, $\|\hat{\Sigma}_v\|$. That avoids the underlying bias $\|(\hat{\Sigma}+\lambda\mathbf{I})^{-1}-\hat{\Sigma}^{-1}\|$ resulting from the large regularization λ, where the large regularization λ is generally for the inverse stability.

2.2 Ridge Regression Only Using Private Data

The following theorem will give the DP ridge regression of private data only (DP-RR). Without loss of generality, we give the result under the assumption that the private data are bounded by $\sqrt{\mathrm{Tr}(\Sigma)+d\log(\frac{n_\xi}{\eta})}$ and the private responses are bounded by $\sqrt{\mathrm{Tr}(\Sigma)+d\log(\frac{n_\xi}{\eta})}\|\boldsymbol{\beta}\|+o(1)$, with at least probability $1-\eta$.

Theorem 12 (DP-RR)**.** *Assume that truncated data* $\|\boldsymbol{\xi}_i\| \leqslant \sqrt{\mathrm{Tr}(\Sigma)+d\log(\frac{n_\xi}{\eta})}$ *and* $y_{\boldsymbol{\xi}_i} \leqslant \sqrt{\mathrm{Tr}(\Sigma)+d\log(\frac{n_\xi}{\eta})}\|\boldsymbol{\beta}\|$ *w.p.* $1-\eta$, $\forall i \in [n_\xi]$. *Denote* $\hat{\Sigma}_\xi = \frac{\boldsymbol{\Xi}^T\boldsymbol{\Xi}}{n_\xi}$ *and* $\hat{\Sigma}_{\xi y} = \frac{\boldsymbol{\Xi}^T\mathbf{y}_\xi}{n_\xi}$. *Then the* $\sqrt{2}\mu$*-GDP ridge regression is*

$$\hat{\boldsymbol{\beta}}_{DP} = (\hat{\Sigma}_\xi + \lambda\mathbf{I} + \mathbf{G})^{-1}(\hat{\Sigma}_{\xi y} + \mathbf{g}),$$

where $\mathbf{G} \sim SG_d(\sigma_1^2)$, $\sigma_1 = \frac{2(\mathrm{Tr}(\Sigma)+d\log(\frac{n_\xi}{\eta}))}{\mu\cdot n_\xi}$ *and* $\mathbf{g} \sim \mathcal{N}(0,\sigma_2^2\mathbf{I})$, $\sigma_2 = \frac{2(\mathrm{Tr}(\Sigma)+d\log(\frac{n_\xi}{\eta}))\|\boldsymbol{\beta}\|}{\mu\cdot n_\xi}$.

If n_ξ *makes* $\frac{\sqrt{d^3\log(\frac{2d}{\eta})}(d^{-1}\mathrm{Tr}(\Sigma)+\log(\frac{n_\xi}{\eta}))}{\mu\cdot n_\xi\left(\lambda_{min}(\Sigma)\left(1-O(\sqrt{\frac{d}{n_\xi}}+\sqrt{\frac{\log(1/\eta)}{n_\xi}})\right)^2+\lambda\right)} \leqslant \frac{1}{2}$, *with at least probability* $1-O(\eta)$,

we have

$$\|\hat{\boldsymbol{\beta}}_{DP} - \hat{\boldsymbol{\beta}}\| \leqslant O\Big(\frac{\sqrt{d^3 \log(\frac{2d}{\eta})}}{\mu n_\xi} \cdot \frac{\|\Sigma\|\|\boldsymbol{\beta}\|(d^{-1}\mathrm{Tr}(\Sigma) + \log(\frac{n_\xi}{\eta}))}{\lambda_{min}^2(\Sigma)\big(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}})\big)^4 + \lambda^2}\Big).$$

Remark 1. *The population quantities Σ and $\boldsymbol{\beta}$ in **Theorem 12** are used only to specify the theoretical truncation radii and noise scales. In an implementable DP procedure, any calibration quantity computed from the private sample, such as an empirical covariance matrix $\hat{\Sigma}$ or an empirical response second moment $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n y_i^2$, must either be released through a differentially private mechanism or replaced by quantities estimated from public data. Otherwise, using such private empirical quantities directly would leak information and invalidate the claimed privacy guarantee.*

2.3 Generalized Linear Models

This section provides more detailed theoretical results about DP-PMTLR and the differentially private logistic regression (DP-LR, **Algorithm 4**). Then, we extend to the generalized linear regressions based on PMT.

2.3.1 Logistic Regression

We list the gradients and Hessian matrices about the two loss functions of the logistic regression, seeing **Theorem 9**, so as to discuss them in the following. The gradients are

$$\nabla\mathcal{L}(\boldsymbol{\beta}; \boldsymbol{\Xi}) = -\frac{1}{n_\xi}\boldsymbol{\Xi}^T(\mathbf{y} - \mathbf{p}) + \lambda\boldsymbol{\beta},$$

and

$$\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\Xi}}) = -\frac{1}{n_\xi}\tilde{\boldsymbol{\Xi}}^T(\mathbf{y}-\tilde{\mathbf{p}}) + \lambda\hat{\Sigma}_v^{-1}\boldsymbol{\beta}$$

where $\mathbf{p} = (p_1, ..., p_{n_\xi})^\top \in [0,1]^{n_\xi}$ and $\tilde{\mathbf{p}} = (\tilde{p}_1, ..., \tilde{p}_{n_\xi})^\top \in [0,1]^{n_\xi}$. Note that $p_i = \frac{1}{1+e^{-\boldsymbol{\xi}_i^T\boldsymbol{\beta}}}$ and $\tilde{p}_i = \frac{1}{1+e^{-\tilde{\boldsymbol{\xi}}_i^T\boldsymbol{\beta}}}$ are different due to different samples. The Hessian matrices are

$$\nabla^2\mathcal{L}(\boldsymbol{\beta};\boldsymbol{\Xi}) = \mathbf{H}_\beta + \lambda\mathbf{I} = \frac{1}{n_\xi}\boldsymbol{\Xi}^T\mathbf{W}_\beta\boldsymbol{\Xi} + \lambda\mathbf{I},$$

and

$$\nabla^2\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\Xi}}) = \tilde{\mathbf{H}}_\beta + \lambda\hat{\Sigma}_v^{-1} = \frac{1}{n_\xi}\tilde{\boldsymbol{\Xi}}^T\tilde{\mathbf{W}}_\beta\tilde{\boldsymbol{\Xi}} + \lambda\hat{\Sigma}_v^{-1}, \tag{1}$$

where $\mathbf{W}_\beta$ is a diagonal matrix with $(\mathbf{W}_\beta)_{ii} = p_i(1-p_i)$, similarly, $(\tilde{\mathbf{W}}_\beta)_{ii} = \tilde{p}_i(1-\tilde{p}_i)$, $i = 1, ..., n_\xi$.

The assumption about the Hessian matrix is necessary to ensure the convergence of logistic regression and show the advantage of our method. It's well known the Hessian matrix of $-\frac{1}{n_\xi}\sum_{i=1}^{n_\xi} l_i(\boldsymbol{\xi}_i^T\boldsymbol{\beta})$ is

$$\mathbf{H}_\beta = \frac{\partial^2(-\frac{1}{n_\xi}\sum_{i=1}^{n_\xi} l_i(\boldsymbol{\xi}_i^T\boldsymbol{\beta}))}{\partial\boldsymbol{\beta}^2} = \frac{1}{n_\xi}\boldsymbol{\Xi}^T\mathbf{W}_\beta\boldsymbol{\Xi},$$

where $\mathbf{W}_\beta$ is a diagonal matrix with $(\mathbf{W}_\beta)_{ii} = p_i(1-p_i)$, similarly, $(\tilde{\mathbf{W}}_\beta)_{ii} = \tilde{p}_i(1-\tilde{p}_i)$, $i = 1, ..., n_\xi$. Then, we assume that

$$\textbf{Hessian assumption:}\quad \tau_0\lambda_{\min}(\hat{\Sigma}_\xi) \preccurlyeq \tau_\beta\lambda_{\min}(\hat{\Sigma}_\xi) \preccurlyeq \mathbf{H}_\beta \preccurlyeq \frac{1}{4}\lambda_{\max}(\hat{\Sigma}_\xi), \tag{2}$$

where $\tau_\beta \in (0, 1/4]$ depends on the classified probabilities p_is which are the function of $\boldsymbol{\beta}$ and $\tau_0 = \inf_\beta \tau_\beta$. The lower bound τ_β tends to be small easily at the latter stages of iteration, because the classified accuracy becomes high and leads to $p_i(1-p_i) \to 0$. The condition number of the

Hessian matrix $\mathbf{H}_\beta$ is approximately

$$\kappa(\mathbf{H}_\beta) = O(\frac{\lambda_{\max}(\hat{\Sigma}_\xi)}{\tau_0 \lambda_{\min}(\hat{\Sigma}_\xi)})^{n_\xi \overset{\text{large}}{=}} O(\frac{\kappa(\Sigma)}{\tau_0}).$$

This suggests that the weight matrix $\mathbf{W}_\beta$ leads to a more ill-conditioned Hessian, thereby making Newton's method more vulnerable to disruption from DP noise. Introducing a large regularization parameter λ can alleviate this issue by decreasing the condition number, thus enhancing resistance to noise. However, this comes at the cost of increased bias in the estimation, which may lead to an underfitted model. The proposed transformation substantially reduces the condition number of the Hessian tending to $O(\tau_0^{-1})$ and also enhances the following favorable properties of the associated loss function.

Lemma 2 (Strong convexity)**.** *Assume that the subG(Σ) data $\boldsymbol{\Xi} \in \mathbb{R}^{n_\xi \times d}$, $\mathbf{y} = \{y_1, ..., y_{n_\xi}\}$, $y_i \in \{0, 1\}$ and $\hat{\Sigma}_\upsilon$ is the second-moment estimation from other n_υ samples. Considering the convexity of two loss functions $\mathcal{L}(\boldsymbol{\beta}; \boldsymbol{\Xi})$ and $\tilde{\mathcal{L}}(\boldsymbol{\beta}; \tilde{\boldsymbol{\Xi}})$, with at least probability $1 - O(\eta)$, we have*

$$\nabla^2 \mathcal{L}(\boldsymbol{\beta}; \boldsymbol{\Xi}) \succcurlyeq \gamma_\Sigma \mathbf{I},$$

where $\gamma_\Sigma = \tau_0 \lambda_{\min}(\Sigma)\big(1 - O\big(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\eta)}{n}}\big)\big)^2 + \lambda$. And, with at least probability $1 - 2\eta$, we have

$$\nabla^2 \tilde{\mathcal{L}}(\boldsymbol{\beta}; \tilde{\boldsymbol{\Xi}}) \succcurlyeq \gamma_L \mathbf{I},$$

where $\gamma_L = \tau_0 L\big(1 - O\big(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}}\big)\big)^2 + \frac{\lambda}{\lambda_{\max}(\hat{\Sigma}_\upsilon)}$ and $L = \frac{n_\upsilon}{(\sqrt{n_\upsilon} + O(\sqrt{d} + \sqrt{\log(\frac{1}{\eta})}))^2}$.

Remark 2. *That shows the better convexity of the transformation loss $\tilde{\mathcal{L}}(\boldsymbol{\beta}; \tilde{\boldsymbol{\Xi}})$. Because $L \to 1$ is greater than $\lambda_{\min}(\Sigma)$, and the regularization parameter $\frac{\lambda}{\lambda_{\max}(\hat{\Sigma}_\upsilon)}$ weaken the dependence on*

the regularization parameter.

Lemma 3 (Lipschitz continuity of Hessian)**.** *The Hessian matrix* $\mathbf{H}_\beta$ *is Lipschitz continuous with respect to the* ℓ_2 *norm, i.e., there exists a constant* $C_\xi > 0$ *such that*

$$\|\mathbf{H}_\beta - \mathbf{H}_{\beta'}\|_2 \leqslant C_\xi \|\boldsymbol{\beta} - \boldsymbol{\beta'}\|_2,$$

for all $\boldsymbol{\beta}, \boldsymbol{\beta'} \in \mathbb{R}^d$*. In particular, we have* $C_\xi = \frac{\sup_i \|\boldsymbol{\xi}_i\|^3}{6\sqrt{3}}$*. Moreover,*

$$\|\nabla^2 \mathcal{L}(\boldsymbol{\beta}; \boldsymbol{\Xi}) - \nabla^2 \mathcal{L}(\boldsymbol{\beta'}; \boldsymbol{\Xi})\|_2 \leqslant C_\xi \|\boldsymbol{\beta} - \boldsymbol{\beta'}\|_2,$$

and

$$\|\nabla^2 \tilde{\mathcal{L}}(\boldsymbol{\beta}; \tilde{\boldsymbol{\Xi}}) - \nabla^2 \tilde{\mathcal{L}}(\boldsymbol{\beta'}; \tilde{\boldsymbol{\Xi}})\|_2 \leqslant C_{\tilde{\boldsymbol{\xi}}} \|\boldsymbol{\beta} - \boldsymbol{\beta'}\|_2,$$

where $C_{\tilde{\boldsymbol{\xi}}} = \frac{(d(1+\log(n_\xi/\eta)))^{3/2}}{6\sqrt{3}}$ *and* $\eta > 0$.

Remark 3. ***Lemma 3*** *shows that the* l_2*-norm of sample points also impacts on the Lipschitz continuity of the Hessian matrix. The PMT method normalizes the the* l_2*-norm of sample points into a principled radius and reduces the Lipschitz continuity parameter, which improves the converged efficiency of Newton's method.*

These lemmas guarantee the convergence of DP-PMTLR and DP-LR, seeing the proof in Appendix 4.10.

Algorithm 4 Differentially Private Logistic Regression (DP-LR)

1: **Input:** Private dataset $\{(\boldsymbol{\xi}_i, y_i) \in \mathbb{R}^d \times \{0,1\}\}_{i=1}^{n_\xi}$. Parameters μ, λ, d, n_ξ, and η.
2: **Private parameter:**

$$\sigma_1 = \frac{\sqrt{T}(\mathrm{Tr}(\Sigma)+d\log(\frac{n_\xi}{\eta}))}{2\mu n_\xi},\ \sigma_2 = \frac{2\sqrt{T(\mathrm{Tr}(\Sigma)+d\log(\frac{n_\xi}{\eta}))}}{\mu n_\xi}.$$

3: **Intialization:** $\boldsymbol{\beta}^{(0)} = \mathbf{0}$
4: **for** $t = 0, ..., T$ **do**
5: **Gaussian mechanism and Newton update:** $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \Big(\nabla^2\mathcal{L}(\boldsymbol{\beta}^{(t)};\boldsymbol{\Xi}) + \mathbf{G}\Big)^{-1}\Big(\nabla\mathcal{L}(\boldsymbol{\beta}^{(t)};\boldsymbol{\Xi}) + \mathbf{g}\Big)$, where $\mathbf{G} \sim SG_d(\sigma_1^2)$ and $\mathbf{g} \sim \mathcal{N}(0, \sigma_2^2\mathbf{I})$.
6: **end for**
7: **Output:** DP estimator $\boldsymbol{\beta}^{(T)}$.

Theorem 13 (DP-LR). *Suppose every sub-Gaussian sample $\boldsymbol{\xi}_i$ is truncated as $\|\boldsymbol{\xi}_i\| \leqslant \sqrt{\mathrm{Tr}(\Sigma) + d\log(\frac{n_\xi}{\eta})}$, the minimizer $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}} \in \mathcal{B}_r(\hat{\boldsymbol{\beta}})$ and $\|\nabla\mathcal{L}(\boldsymbol{\beta}^{(0)};\boldsymbol{\Xi})\| \leqslant \min\{\gamma_\Sigma r, \frac{\gamma_\Sigma^2}{C_\xi}\}$, where $\gamma_\Sigma = \tau_0\lambda_{\min}(\Sigma)\big(1 - O\big(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\eta)}{n}}\big)\big)^2 + \lambda$. Let n_ξ makes $\Big(\frac{\sqrt{Td^3\log(\frac{2Td}{\eta})}(d^{-1}\mathrm{Tr}(\Sigma)+\log(\frac{n_\xi}{\eta}))}{\mu\cdot n_\xi\cdot\gamma_\Sigma}\Big)$ sufficiently small and $T = O(\log\log(n_\xi))$. About **Algorithm 4**, the T^{th} DP Newton's method iteration satisfies (i) $\boldsymbol{\beta}^{(T)}$ is $\sqrt{2}\mu$-GDP, and (ii) $\|\boldsymbol{\beta}^{(T)} - \hat{\boldsymbol{\beta}}\| \leqslant O\Big(\frac{\sqrt{Td^3\log(\frac{2Td}{\eta})}(d^{-1}\mathrm{Tr}(\Sigma)+\log(\frac{n_\xi}{\eta}))}{\mu\cdot n_\xi\gamma_\Sigma^2}\Big)$, w.p. $1-\eta$.*

2.3.2 Generalized Linear Model

In this section, the proposed method is applied to generalized linear models. Given the data $(\boldsymbol{\xi}_i, y_i) \in \mathbb{R}^{d+1}$, $i = 1, ..., n_\xi$ and y_i following the exponential family of distribution, the joint probability density function is

$$\prod_i^n f(y_i;\theta_i,\phi) = \prod_i^n \exp\big\{\frac{y_i\theta_i - b(\theta_i)}{r(\phi)} + c(y_i,\phi)\big\},$$

where θ_i is the natural parameter and ϕ is the dispersion parameter.

Given the bounded model parameters $\boldsymbol{\beta}$ and $\|\boldsymbol{\beta}\|_2 \leq R_\beta$, the generalized linear model (GLM) states that there exists a link function $g(\cdot)$ which satisfies

$$g(\mathbb{E}(y_i)) = g(\varpi_i) = \boldsymbol{\xi}_i^\top \boldsymbol{\beta} \quad i = 1, ..., n. \tag{3}$$

Here $\varpi_i = \mathbb{E}(y_i)$ and we consider the canonical link where $\theta_i = g(\varpi_i) = \boldsymbol{\xi}_i^\top \boldsymbol{\beta}$. Notice that the exponential family of distribution makes

$$\varpi_i = \mathbb{E}(y_i) = b'(\theta_i),$$

where $b'(\theta_i) = \frac{\partial b(\theta_i)}{\partial \theta_i}$ and it implies that $\theta_i = (b')^{-1}(\varpi_i)$ and the canonical link function $g(z) = (b')^{-1}(z)$. From Eq.(3), we have $\varpi_i = g^{-1}(\boldsymbol{\xi}_i^\top \boldsymbol{\beta}) = b'(\boldsymbol{\xi}_i^\top \boldsymbol{\beta})$ and $\theta_i = (b')^{-1}(\boldsymbol{\xi}_i^\top \boldsymbol{\beta})$. Considering the negative log-likelihood function with the regularization term, we have

$$\mathcal{L}(\boldsymbol{\beta}; \boldsymbol{\Xi}) = -\frac{1}{n_\xi} \sum_{i=1}^{n_\xi} \Big[y_i \theta_i - b(\theta_i) \Big] + \frac{\lambda}{2} \|\boldsymbol{\beta}\|_2^2, \tag{4}$$

where $\theta_i = \boldsymbol{\xi}_i^\top \boldsymbol{\beta}$ and we omit the scale parameter $r(\phi)$ unrelated to the optimization of $\boldsymbol{\beta}$. Denote the $l_i(\boldsymbol{\xi}_i^\top \boldsymbol{\beta}) = y_i \theta_i - b(\theta_i)$, we get the l_i's gradient and Hessian matrix as

$$\nabla_\beta l_i(\boldsymbol{\xi}_i^\top \boldsymbol{\beta}) = (y_i - b'(\boldsymbol{\xi}_i^\top \boldsymbol{\beta}))\boldsymbol{\xi}_i; \quad \nabla_\beta^2 l_i(\boldsymbol{\xi}_i^\top \boldsymbol{\beta}) = -b''(\boldsymbol{\xi}_i^\top \boldsymbol{\beta})\boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top.$$

Then, the Hessian of $\mathcal{L}(\boldsymbol{\beta}; \boldsymbol{\Xi})$ is

$$\nabla_\beta^2 \mathcal{L}(\boldsymbol{\beta}; \boldsymbol{\Xi}) = -\frac{1}{n_\xi} \nabla_\beta^2 l_i(\boldsymbol{\xi}_i^\top \boldsymbol{\beta}) + \lambda \mathbf{I} = \frac{1}{n_\xi} \boldsymbol{\Xi}^\top \mathbf{W}_\beta \boldsymbol{\Xi} + \lambda \mathbf{I}$$

with the weights $\mathbf{W}_\beta = diag(b''(\boldsymbol{\xi}_1^\top \boldsymbol{\beta}), \cdots, b''(\boldsymbol{\xi}_n^\top \boldsymbol{\beta}))$. That implies, in the generalized linear model, the Hessian matrix $\nabla_\beta \mathcal{L}(\boldsymbol{\beta}; \boldsymbol{\Xi})$ is also affected by the second-moment matrix $\hat{\Sigma}_\xi$ like the analysis of logistic regression. Namely,

$$\tau_0 \lambda_{\min}(\hat{\Sigma}_\xi) \preccurlyeq \mathbf{H}_\beta = \frac{1}{n_\xi} \boldsymbol{\Xi}^\top \mathbf{W}_\beta \boldsymbol{\Xi} \preccurlyeq \tau_1 \lambda_{\max}(\hat{\Sigma}_\xi),$$

where $\tau_0 = \inf_z b''(z) > 0$ and $\tau_1 = \sup_z b''(z) < \infty$. The condition number of $\mathbf{H}_\beta$ is

$$\kappa(\mathbf{H}_\beta) = O(\frac{\tau_1 \lambda_{\max}(\hat{\Sigma}_\xi)}{\tau_0 \lambda_{\min}(\hat{\Sigma}_\xi)}) \overset{n_\xi \text{ large}}{=} \frac{\tau_1}{\tau_0} \kappa(\Sigma).$$

The rate $\frac{\tau_1}{\tau_0}$ is fixed by the function $b(\cdot)$, so reducing the condition number $\kappa(\Sigma)$ is the only choice. Next, we illustrate that our method PMT can improve the GLM case via eliminating the $\kappa(\Sigma)$.

2.4 Public-moment-guided Generalized Linear Models

We consider the transformed loss function in GLM

$$\tilde{\mathcal{L}}(\boldsymbol{\beta}; \tilde{\boldsymbol{\Xi}}) = -\frac{1}{n_\xi} \sum_{i=1}^{n_\xi} \Big[\frac{y_i \tilde{\theta}_i - b(\tilde{\theta}_i)}{r(\phi)} \Big] + \frac{\lambda}{2} \|\hat{\Sigma}_\upsilon^{-1/2} \boldsymbol{\beta}\|_2^2, \tag{5}$$

where $\tilde{\theta}_i = \tilde{\boldsymbol{\xi}}_i^\top \boldsymbol{\beta}$ and $\tilde{\boldsymbol{\xi}}_i = \hat{\Sigma}_v^{-1/2}\boldsymbol{\xi}_i$ is the transformed data point from PMT. The gradient and Hessian matrix of $\tilde{\mathcal{L}}(\boldsymbol{\beta}; \tilde{\boldsymbol{\Xi}})$ are as follows

$$\begin{aligned} \nabla_\beta \tilde{\mathcal{L}}(\boldsymbol{\beta}; \tilde{\boldsymbol{\Xi}}) &= \frac{1}{n_\xi}\tilde{\boldsymbol{\Xi}}^\top(\mathbf{y} - \tilde{\mathbf{b}}') + \hat{\Sigma}_v^{-1}\boldsymbol{\beta}; \\ \nabla_\beta^2 \tilde{\mathcal{L}}(\boldsymbol{\beta}; \tilde{\boldsymbol{\Xi}}) &= \frac{1}{n_\xi}\tilde{\boldsymbol{\Xi}}^\top \tilde{\mathbf{W}}_\beta \tilde{\boldsymbol{\Xi}} + \lambda\hat{\Sigma}_v^{-1}, \end{aligned} \tag{6}$$

where $\mathbf{y} = (y_1, \cdots, y_{n_\xi})^\top$, $\tilde{\mathbf{b}}' = (b'(\tilde{\boldsymbol{\xi}}_1^\top\boldsymbol{\beta}), \cdots, b'(\tilde{\boldsymbol{\xi}}_{n_\xi}^\top\boldsymbol{\beta}))^\top$, and $\tilde{\mathbf{W}}_\beta = diag(b''(\tilde{\boldsymbol{\xi}}_1^\top\boldsymbol{\beta})), \cdots, b''(\tilde{\boldsymbol{\xi}}_{n_\xi}^\top\boldsymbol{\beta}))$.

Considering the convexity of GLM loss functions, we have the invariant iterations as the following corollary, which is similar to **Theorem 9**.

Corollary 2 (Equivalent GLM estimation)**.** *Considering GLM and loss functions* (4) *and* (5)*, their estimations is equivalent in each iteration of Newton's method. Moreover, their minimizers are equivalent*

$$\hat{\Sigma}_v^{-1/2}\tilde{\boldsymbol{\beta}} \;=\; \hat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}; \boldsymbol{\Xi}),$$

where $\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \tilde{\mathcal{L}}(\boldsymbol{\beta}; \tilde{\boldsymbol{\Xi}})$.

Under the invariant iterations case, the transformed loss function holds Hessian matrices with a smaller condition number

$$\kappa(\tilde{\mathbf{H}}_\beta) = O(\frac{\tau_1}{\tau_0}\kappa(\tilde{\Sigma}_\xi))^{\,n_\xi \overset{\text{large}}{=}} O(\frac{\tau_1}{\tau_0})$$

where $\tilde{\mathbf{H}}_\beta = \frac{1}{n_\xi}\tilde{\boldsymbol{\Xi}}^\top \mathbf{W}_\beta \tilde{\boldsymbol{\Xi}}$ and $\tilde{\Sigma}_\xi = \frac{1}{n_\xi}\tilde{\boldsymbol{\Xi}}^\top\tilde{\boldsymbol{\Xi}}$, and the second equation follows from **Theorem 3**. That illustrates that our method is also able to improve the GLM estimations in DP Newton's methods. We propose **Algorithm 5** that illustrates how PMT is used in generalized linear models.

Algorithm 5 Differentially Private PMT Generalized Linear Models(DP-PMTGLM)

1: **Input:** Private dataset $\{(\boldsymbol{\xi}_i, y_i) \in \mathbb{R}^d \times (-R_y, R_y)\}_{i=1}^{n_\xi}$, public dataset $\{\boldsymbol{\upsilon}_i \in \mathbb{R}^d\}_{i=1}^{n_\upsilon}$. Parameters μ, λ, d, n_ξ, n_υ and η. Constraint $\|\boldsymbol{\beta}\|_2 \leq R_\beta$, $M_{b'} = max_z b'(z)$ and $M_{b''} = max_z b''(z)$, $z \in [-R_\beta\sqrt{d(1+\log(\frac{2n_\xi}{\eta}))}, R_\beta\sqrt{d(1+\log(\frac{2n_\xi}{\eta}))}]$.

2: **Transform covariates:**
$(\{\tilde{\boldsymbol{\Xi}}_i\}_{i=1}^{n_\xi}, \hat{\Sigma}_\upsilon) = PMT(\{\boldsymbol{\xi}_i\}_{i=1}^{n_\xi}, \{\boldsymbol{\upsilon}_i\}_{i=1}^{n_\upsilon}, d, n_\xi, n_\upsilon, \eta)$.

3: **Private parameter:**

$$\sigma_1 = \frac{2\sqrt{T}d(1+\log(\frac{2n_\xi}{\eta}))M_{b''}}{\mu n_\xi},\ \sigma_2 = \frac{2\sqrt{Td(1+\log(\frac{2n_\xi}{\eta}))}(R_y+M_{b'})}{\mu n_\xi}.$$

4: $\boldsymbol{\beta}^{(0)} = \mathbf{0}$

5: **for** $t = 0, ..., T$ **do**

6: **Gaussian mechanism and Newton update:**

$$\tilde{\boldsymbol{\beta}}^{(t+1)} = \tilde{\boldsymbol{\beta}}^{(t)} - \Big(\nabla^2\tilde{\mathcal{L}}(\tilde{\boldsymbol{\beta}}^{(t)}; \tilde{\boldsymbol{\Xi}}) + \mathbf{G}\Big)^{-1}\Big(\nabla\tilde{\mathcal{L}}(\tilde{\boldsymbol{\beta}}^{(t)}; \tilde{\boldsymbol{\Xi}}) + \mathbf{g}\Big),$$

where $\mathbf{G} \sim SG_d(\sigma_1^2)$ and $\mathbf{g} \sim \mathcal{N}(0, \sigma_2^2\mathbf{I})$.

7: **end for**

8: **Recover:** $\bar{\boldsymbol{\beta}}^{DP} \leftarrow \hat{\Sigma}_\upsilon^{-1/2} \cdot \tilde{\boldsymbol{\beta}}^{(T)}$.

9: **Output:** DP estimator $\bar{\boldsymbol{\beta}}^{DP}$.

Theorem 14 (DP-PMTGLM). *Suppose $y_i \in [-R_y, R_y]$ with a constant R_y, $\boldsymbol{\xi}_i \in subG(\Sigma)$ and $b(\theta)$ is locally τ_0-strong convexity. Constraint $\forall\ \|\boldsymbol{\beta}\|_2 \leq R_\beta$, the minimizer $\|\tilde{\boldsymbol{\beta}}\|_2 \leq R_\beta$, $\boldsymbol{\beta}^{(0)} \in \mathcal{B}_r(\tilde{\boldsymbol{\beta}})$, the Hessian matrix $\nabla^2\tilde{\mathcal{L}}(\boldsymbol{\beta}; \tilde{\boldsymbol{\Xi}})$ is locally C_l-Lipschitz continuous when $\boldsymbol{\beta} \in \mathcal{B}_r(\tilde{\boldsymbol{\beta}})$, and $\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta}^{(0)}; \tilde{\boldsymbol{\Xi}})\| \leqslant \min\{\gamma_L r, \frac{\gamma_L^2}{C_l}\}$, where $\gamma_L = \tau_0 L(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}}))^2 + \lambda\|\hat{\Sigma}_\upsilon\|^{-1}$. Let $\sqrt{n_\upsilon} \geqslant O(\sqrt{d} + \sqrt{\log(1/\eta)})$, n_ξ makes $\Big(\frac{\sqrt{Td^3\log(Td/\eta)}(1+\log(\frac{2n_\xi}{\eta}))M_{b''}}{\mu\cdot n_\xi\cdot\gamma_L}\Big)$ small enough and $T = O(\log\log(n_\xi))$. The outout of **Algorithm 5** satisfies $\sqrt{2}\mu$-GDP, and we have $\|\boldsymbol{\beta}^{(T)} - \tilde{\boldsymbol{\beta}}\| \leqslant O\Big(\frac{\sqrt{Td^3\log(\frac{2Td}{\eta})}(1+\log(\frac{2n_\xi}{\eta}))}{\mu\cdot n_\xi\cdot\gamma_L^2}\Big)$, w.p. $1-\eta$.*

3 Useful Tools

There are some facts about Gaussian vector and Gaussian matrix, which help us to control the DP noise. We denote the $d \times d$ symmetric Gaussian random matrix $\mathbf{G} \sim SG_d(\sigma^2)$ with elements $G_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

Lemma 4 (Symmetric Gaussian matrix bound Avella-Medina et al. (2023))**.** *For a $d \times d$ symmetric matrix $\mathbf{G} \sim SG_d(\sigma^2)$, with probability $1 - \eta$,*

$$\|\mathbf{G}\|_2 \leq \sigma\sqrt{2d\log(2d/\eta)}. \tag{7}$$

Lemma 5 (Gaussian vector bound Laurent and Massart (2000))**.** *For a gaussian vector $\mathbf{g} = (g_1, ..., g_d)$, $g_i \sim \mathcal{N}(0, \sigma^2)$, the $\|\mathbf{g}\|_2^2$ satisfies*

$$\mathbb{P}[\|\mathbf{g}\|_2^2 \geq \sigma^2(2\sqrt{d\log(1/\eta)} + 2\log(1/\eta) + d)] \leq \eta.$$

The next lemma gives the largest and smallest singular values of the sub-Gaussian matrix. It helps us to analyze the second-moment matrix estimation and control its condition number.

Lemma 6 (Singular values bound Vershynin (2010))**.** *Let $\mathbf{X}$ be a $n \times d$ random matrix whose each row $\mathbf{x}_i$ is independently non-isotropic sub-gaussian random vectors in $\mathbb{R}^d$ with the second-moment matrix Σ. Then for every $\eta > 0$ and $\sqrt{n} \geqslant O(\sqrt{d} + \sqrt{\log(1/\eta)})$, with at least $1 - 2\eta$, one has*

$$\sqrt{\lambda_{min}(\Sigma)}(\sqrt{n} - C\sqrt{d} - \sqrt{\frac{1}{c}\log(1/\eta)}) \leq \lambda_{min}(\mathbf{X}),$$
$$\sqrt{\lambda_{max}(\Sigma)}(\sqrt{n} + C\sqrt{d} + \sqrt{\frac{1}{c}\log(1/\eta)}) \geq \lambda_{max}(\mathbf{X}),$$

where $\lambda(\cdot)$ means singular value, and $C = C_K$, $c = c_K > 0$ depend on the sub-gaussian norm

$K = max_i \|\mathbf{x}_i\|_{\psi_2}$. *Typically, for the isotropic situation* $\Sigma = \mathbf{I}$, *one has*

$$\sqrt{n} - C\sqrt{d} - \sqrt{\frac{1}{c}\log(1/\eta)} \leq \lambda_{min}(\mathbf{X}),$$
$$\sqrt{n} + C\sqrt{d} + \sqrt{\frac{1}{c}\log(1/\eta)} \geq \lambda_{max}(\mathbf{X}).$$

Corollary 3. *Let* $\mathbf{X}$ *be a* $n \times d$ *random matrix whose each row* $\mathbf{x}_i$ *is independently isotropic sub-gaussian random vector in* $\mathbb{R}^d$ *with the second-moment matrix* $\Sigma = \mathbf{I}$. *Then for every* $\eta > 0$ *and* $\sqrt{n} \geqslant O(\sqrt{d} + \sqrt{\log(1/\eta)})$, *one has*

$$\mathbb{P}\big[\frac{\lambda_{max}(\mathbf{X}^T\mathbf{X})}{n} = \frac{\lambda^2_{max}(\mathbf{X})}{n} \geq \Big(1 + O\Big(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\eta)}{n}}\Big)\Big)^2\big] \leq \eta$$

and

$$\mathbb{P}\big[\frac{\lambda_{min}(\mathbf{X}^T\mathbf{X})}{n} = \frac{\lambda^2_{min}(\mathbf{X})}{n} \leq \Big(1 - O\Big(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\eta)}{n}}\Big)\Big)^2\big] \leq \eta.$$

The following lemma on the $l_2 - norm$ of a sub-gaussian vector illustrates that the sample's length depends on the second-moment matrix. It provides the theoretical guarantee about a principled truncating radius.

Lemma 7 (Bounded l_2-norm of Sub-gaussian Vector)**.** *Let* $\mathbf{x} = (x_1, ...x_d)^T \in \mathbb{R}^d$ *is a non-isotropic sub-gaussian random vector with* $\mathbb{E}\mathbf{x}\mathbf{x}^T = \Sigma$. *Then* $\|\mathbf{x}\|_2^2$ *is sub-exponential and*

$$\mathbb{P}\Big[\big|\|\mathbf{x}\|_2^2 - \sum_{i=1}^{d}\lambda_i\big| \geq t\Big] \leq 2\exp\big(-\frac{ct}{dK^2}\big),$$

where $K = max_i\|x_i\|_{\psi_2}$, c *is an absolute constant on* K *and* $\lambda_1 > ... > \lambda_d$ *are the eigenvalues*

of Σ*. One more thing, with at least probability* $1-\eta$*,*

$$\|\mathbf{x}\|_2^2 \leq \mathrm{Tr}(\Sigma) + \big(\frac{dK^2}{c}\log(\frac{2}{\eta})\big) \leq O\big(d(\overline{\mathrm{Tr}}(\Sigma) + \log(\frac{2}{\eta}))\big),$$

where $\overline{\mathrm{Tr}}(\Sigma) = \frac{1}{d}\mathrm{Tr}(\Sigma)$ *is the average trace of* Σ*.*

The following lemmas are used to analyze the noisy inverse matrix and provide theoretical tools to quantify the robustness condition.

Lemma 8. *Denote a square matrix* Σ *and a disturb matrix* $\mathbf{G}$*, the condition number of* $\kappa(\Sigma) = \|\Sigma\|\|\Sigma^{-1}\|$*. Then,*

$$\frac{\|(\Sigma+\mathbf{G})^{-1} - \Sigma^{-1}\|}{\|(\Sigma+\mathbf{G})^{-1}\|} \leq \kappa(\Sigma)\frac{\|\mathbf{G}\|}{\|\Sigma\|}.$$

Moreover, if $\kappa(\Sigma)\frac{\|\mathbf{G}\|}{\|\Sigma\|} = \|\Sigma^{-1}\|\|\mathbf{G}\| \leq 1$*, then*

$$\|(\Sigma+\mathbf{G})^{-1}\| \leq \frac{\|\Sigma^{-1}\|}{1-\kappa(\Sigma)\frac{\|\mathbf{G}\|}{\|\Sigma\|}}.$$

Moreover

$$\frac{\|(\Sigma+\mathbf{G})^{-1} - \Sigma^{-1}\|}{\|\Sigma^{-1}\|} \leq \frac{\kappa(\Sigma)\frac{\|\mathbf{G}\|}{\|\Sigma\|}}{1-\kappa(\Sigma)\frac{\|\mathbf{G}\|}{\|\Sigma\|}}.$$

Lemma 9 (Weyl's Inequality)**.** *Denote a matrix* $\mathbf{M}$ *and a* $\mathbf{H}$ *are real symmetric matrices , then*

$$\lambda_{max}(\mathbf{M}+\mathbf{H}) \leqslant \lambda_{max}(\mathbf{M}) + \lambda_{max}(\mathbf{H}),$$

$$\lambda_{min}(\mathbf{M}+\mathbf{H}) \geqslant \lambda_{min}(\mathbf{M}) + \lambda_{min}(\mathbf{H}),$$

where $\lambda_{max}(\cdot)$ *and* $\lambda_{min}(\cdot)$ *are the largest and smallest eigenvalues of the matrix, respectively.*

The following lemma (**Lemma 21** in Avella-Medina et al. (2023)) is the main tool to guarantee the converge of generalized linear regressions via GDP Newton's method.

Lemma 10 (Avella-Medina et al. (2023))**.** *A loss function* $\mathcal{L}(\boldsymbol{\beta};\boldsymbol{\Xi}) = \frac{1}{n_\xi}\sum_{i=1}^{n} l(\boldsymbol{\beta};\boldsymbol{\xi}_i) + \lambda\|\boldsymbol{\beta}\|^2$ *is the locally* γ*-strong convexity. For any* $\boldsymbol{\beta}$ *in some neighboring area of the minimizer* $\hat{\boldsymbol{\beta}}$*, its Hessian matrix* $\|\nabla^2\mathcal{L}(\boldsymbol{\beta};\boldsymbol{\Xi})\|_2 \leq B_h$ *exhibits local* C_h*-Lipschitz continuity and the gradient* $\|\nabla\mathcal{L}(\boldsymbol{\beta};\boldsymbol{\Xi})\|_2 \leq B_g$*. For the* μ*-GDP Newton's method, when the private data* n_ξ *makes* $\frac{\Delta_h\sqrt{Td\log(Td/\eta)}}{\mu\cdot n_\xi}$ *small enough and* $\Delta_h = \max_{\boldsymbol{\beta},A\sim\boldsymbol{\xi'}} \|\nabla^2 l(\boldsymbol{\beta};A) - \nabla^2 l(\boldsymbol{\beta};\boldsymbol{\xi'})\|_F$ *is the sensitivity of the function* $\nabla^2 l(\boldsymbol{\beta};A)$*, with at least probability* $1-O(\eta)$*, we have the following equivalent cases:*

1. $\|\boldsymbol{\beta}^{(t)} - \hat{\boldsymbol{\beta}}\|_2 \leq r$*, where* $\boldsymbol{\beta}^{(t)}$ *is the iteration of the t-th GDP Newton's method and* $\hat{\boldsymbol{\beta}}$ *is the non-DP exact solution of the loss.*

2. $\|\nabla\mathcal{L}(\boldsymbol{\beta}^{(t)};\boldsymbol{\Xi})\|_2 \leq \min\{\gamma r, \frac{\gamma^2}{C_h}\}$*, and*

3. $$\frac{C_h}{2\gamma^2}\|\nabla\mathcal{L}(\boldsymbol{\beta}^{(t+1)};\boldsymbol{\Xi})\|_2 \leq \big(\frac{C_h}{2\gamma^2}\|\nabla\mathcal{L}(\boldsymbol{\beta}^{(t)};\boldsymbol{\Xi})\|_2\big)^2 + O\Big(\frac{C_h B_h \sigma_1\sqrt{d\log(Td/\eta)}}{\gamma^3}\Big), \quad (8)$$

   where the second term is the GDP noisy bound and $\sigma_1 = \frac{2\sqrt{T}\Delta_h}{\mu n_\xi}$ *is the parameter of* $SG_d(\sigma_1^2)$ *adding in the iterative Hessian matrices.*

4 Proofs of the Main Theorems

4.1 Proof of Theorem 3

Proof 1. *The equation is equal to*

$$L\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} \preceq \mathbf{I} \preceq U\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}. \tag{9}$$

Then,

$$\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} = \frac{1}{n}\sum_{i=1}^{n}(\Sigma^{-1/2}\boldsymbol{v}_i)(\Sigma^{-1/2}\boldsymbol{v}_i)^T = \frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{v}}_i\tilde{\boldsymbol{v}}_i{}^T = \frac{1}{n}\tilde{\boldsymbol{\Upsilon}}^T\tilde{\boldsymbol{\Upsilon}} = \tilde{\Sigma},$$

where $\tilde{\boldsymbol{v}}_i \sim subG(\mathbf{I})$*, so* $\tilde{\Sigma} = \frac{1}{n}\tilde{\boldsymbol{\Upsilon}}^T\tilde{\boldsymbol{\Upsilon}}$ *is an estimation of* $\mathbf{I}$*. That means that*

$$\mathbf{I} \preceq U\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} \Longleftrightarrow 1 \leq U\lambda_{min}(\tilde{\Sigma}),$$

$$L\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} \preceq \mathbf{I} \Longleftrightarrow 1 \geq L\lambda_{max}(\tilde{\Sigma}).$$

From ***Corollary 3****, we know*

$$U = \Big(1 - O\Big(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\eta)}{n}}\Big)\Big)^{-2} \Longrightarrow 1 \leq U\lambda_{min}(\tilde{\Sigma}),$$
$$L = \Big(1 + O\Big(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\eta)}{n}}\Big)\Big)^{-2} \Longrightarrow 1 \geq L\lambda_{max}(\tilde{\Sigma}).$$

□

4.2 Proof of Theorem 4

Proof 2. ***1. Privacy.*** *Given two neighboring data sets with the different samples* $\tilde{\boldsymbol{\xi}}$ *and* $\tilde{\boldsymbol{\xi}}'$*, the* l_2*-sensitivity of the truncation second-moment matrix is*

$$\left\| \frac{1}{n_\xi} \left(\tilde{\boldsymbol{\xi}} \tilde{\boldsymbol{\xi}}^T - \tilde{\boldsymbol{\xi}}' \tilde{\boldsymbol{\xi}}'^T \right) \right\|_F \leq \frac{1}{n_\xi} \|\tilde{\boldsymbol{\xi}}\|_2^2 + \frac{1}{n_\xi} \|\tilde{\boldsymbol{\xi}}'\|_2^2 \leq \frac{2d(1 + \log(\frac{2n_\xi}{\eta}))}{n_\xi},$$

where $\tilde{\boldsymbol{\xi}}$ *and* $\tilde{\boldsymbol{\xi}}'$ *are the different samples in the neighboring data sets. From the Gaussian mechanism (**Theorem 2**), we get the privacy guarantee.*

2. Noisy matrix bound. *It's a direct conclusion from the **Lemma 4**.* □

4.3 Proof of Theorem 5

Proof 3. ***1. Recovery.*** *From **Corollary 1**, we know with at least probability* $1 - O(\eta)$*, the transformed data* $\tilde{\boldsymbol{\xi}}_i s$ *are not truncated, namely,* $\tilde{\boldsymbol{\xi}}_i = \hat{\Sigma}_\upsilon^{-1/2} \boldsymbol{\xi}_i, \quad \forall i \in [n_\xi]$*. That means*

$$\hat{\Sigma}_\upsilon^{1/2} (\tilde{\Sigma}_\xi + \lambda \hat{\Sigma}_\upsilon^{-1}) \hat{\Sigma}_\upsilon^{1/2} = \hat{\Sigma}_\upsilon^{1/2} (\frac{1}{n_\xi} \sum_{i=1}^{n_\xi} \tilde{\boldsymbol{\xi}}_i \tilde{\boldsymbol{\xi}}_i^T + \lambda \hat{\Sigma}_\upsilon^{-1}) \hat{\Sigma}_\upsilon^{1/2} = \hat{\Sigma}_\xi + \lambda \mathbf{I}.$$

2. Inverse error. *We discuss the following analysis under the transformed data* $\tilde{\boldsymbol{\xi}}_i s$ *are not truncated, namely, the **Recovery** case holds. Especially, we will omit the probability* $1 - O(\eta)$ *from tools in the following analysis for simplicity.*

2.1 Singular values bound. *From **Theorem 3**, the transformed data* $\tilde{\boldsymbol{\xi}}_i \sim subG(\tilde{\Sigma})$ *with* $\tilde{\Sigma} = \hat{\Sigma}_\upsilon^{-1/2} \Sigma \hat{\Sigma}_\upsilon^{-1/2}$ *and* $\tilde{\Sigma}$ *has the smallest and largest eigenvalues related with the number of public*

data n_υ,

$$\begin{aligned}\lambda_{min}(\tilde{\Sigma}) &\geqslant L = \frac{n_\upsilon}{(\sqrt{n_\upsilon} + O(\sqrt{d} + \sqrt{\log(\frac{1}{\eta})}))^2},\\ \lambda_{max}(\tilde{\Sigma}) &\leqslant U = \frac{n_\upsilon}{(\sqrt{n_\upsilon} - O(\sqrt{d} + \sqrt{\log(\frac{1}{\eta})}))^2}.\end{aligned} \tag{10}$$

The $\tilde{\Sigma}_\xi = \frac{\tilde{\Xi}^T\tilde{\Xi}}{n_\xi}$ *is the estimation of* $\tilde{\Sigma}$ *and* ***Lemma 6*** *shows*

$$\begin{aligned}\lambda_{min}(\tilde{\Sigma}_\xi) &\geqslant L(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}}))^2,\\ \lambda_{max}(\tilde{\Sigma}_\xi) &\leqslant U(1 + O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}}))^2,\end{aligned}$$

where L *and* U *are from the Eq.*(10)*. Moreover, from* ***Lemma 9*** *(Wely's inequality), we have*

$$\begin{aligned}\lambda_{min}(\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1}) &\geqslant \lambda_{min}(\tilde{\Sigma}_\xi) + \lambda_{min}(\lambda\hat{\Sigma}_\upsilon^{-1}) \geqslant L(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}}))^2 + \lambda\|\hat{\Sigma}_\upsilon\|^{-1},\\ \lambda_{max}(\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1}) &\leqslant \lambda_{max}(\tilde{\Sigma}_\xi) + \lambda_{max}(\lambda\hat{\Sigma}_\upsilon^{-1}) \leqslant U(1 + O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}}))^2 + \lambda\|\hat{\Sigma}_\upsilon^{-1}\|,\end{aligned}$$

where λ *is the regularization parameter.*

So we have the bound of $\|(\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1})^{-1}\|_2 = \frac{1}{\lambda_{min}(\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1})}$ *and*

$$\frac{1}{L(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}}))^2 + \lambda\|\hat{\Sigma}_\upsilon\|^{-1}} \geqslant \frac{1}{\lambda_{min}(\tilde{\Sigma}_\xi) + \lambda\|\hat{\Sigma}_\upsilon\|^{-1}} \geqslant \|(\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1})^{-1}\|_2. \tag{11}$$

2.2 Get Eq.(3.2). *If* $\|(\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1})^{-1}\|\|\mathbf{G}\| \leqslant \frac{\|\mathbf{G}\|}{L(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}}))^2 + \lambda\|\hat{\Sigma}_\upsilon\|^{-1}} \leqslant \frac{1}{2}$,

Lemma 8 *gives the following bound*

$$\begin{aligned}\|(\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1} + \mathbf{G})^{-1} - (\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1})^{-1}\| &\leq \frac{\|\mathbf{G}\|\|(\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1})^{-1}\|^2}{1 - \|(\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1})^{-1}\|\|\mathbf{G}\|} \\ &\leq 2\|\mathbf{G}\|\|(\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1})^{-1}\|^2 \\ &\overset{(i)}{\leq} \frac{2\|\mathbf{G}\|}{L^2(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}}))^4 + \lambda^2\|\hat{\Sigma}_\upsilon\|^{-2}},\end{aligned}$$

where (i) *is from the Eq.*(11) *and* $\frac{1}{(a+b)^2} \leqslant \frac{1}{a^2+b^2}, a, b > 0$. *So we need to bound* $\|\mathbf{G}\|$ *so that the condition* $\frac{\|\mathbf{G}\|}{L(1-O(\sqrt{\frac{d}{n_\xi}}+\sqrt{\frac{\log(1/\eta)}{n_\xi}}))^2+\lambda\|\hat{\Sigma}_\upsilon\|^{-1}} \leqslant \frac{1}{2}$ *holds. From* ***Theorem 4****, we know that the sufficient condition is the enough number of private data* n_ξ *such that*

$$\frac{\sqrt{d^3\log(\frac{2d}{\eta})}(1 + \log(\frac{2n_\xi}{\eta}))}{(L(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}}))^2 + \lambda\|\hat{\Sigma}_\upsilon\|^{-1}) \cdot \mu \cdot n_\xi} \leqslant \frac{1}{2},$$

where $L = \frac{n_\upsilon}{(\sqrt{n_\upsilon}+O(\sqrt{d}+\sqrt{\log(\frac{1}{\eta})}))^2}$. *Then, using* ***Theorem 4*** *again, we get*

$$\|(\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1} + \mathbf{G})^{-1} - (\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1})^{-1}\| \leqslant \frac{\sqrt{d^3\log(\frac{2d}{\eta})}}{\mu n_\xi} \cdot \frac{2(1 + \log(\frac{2n_\xi}{\eta}))}{L^2(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}}))^4 + \lambda^2\|\hat{\Sigma}_\upsilon\|^{-2}}.$$

2.3 Get Eq.(3.3). *According to the* ***Recover*** *case, we have*

$$\hat{\Sigma}_\upsilon^{-1/2}((\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1} + \mathbf{G})^{-1} - (\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1})^{-1})\hat{\Sigma}_\upsilon^{-1/2} = \hat{\Sigma}_\upsilon^{-1/2}(\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1} + \mathbf{G})^{-1}\hat{\Sigma}_\upsilon^{-1/2} - (\hat{\Sigma}_\xi + \lambda\mathbf{I})^{-1}.$$

So we have

$$\begin{aligned}\|\hat{\Sigma}_v^{-1/2}(\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_v^{-1} + \mathbf{G})^{-1}\hat{\Sigma}_v^{-1/2} - (\hat{\Sigma}_\xi + \lambda\mathbf{I})^{-1}\| &\leqslant \|\hat{\Sigma}_v^{-1/2}\|\|(\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_v^{-1} + \mathbf{G})^{-1} - (\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_v^{-1})^{-1}\|\|\hat{\Sigma}_v^{-1/2}\| \\ &= \|\hat{\Sigma}_v^{-1}\|\|(\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_v^{-1} + \mathbf{G})^{-1} - (\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_v^{-1})^{-1}\| \\ &\leqslant \frac{\sqrt{d^3\log(\frac{2d}{\eta})}}{\mu n_\xi} \cdot \frac{\|\hat{\Sigma}_v^{-1}\|(1+\log(\frac{2n_\xi}{\eta}))}{L^2(1-O(\sqrt{\frac{d}{n_\xi}}+\sqrt{\frac{\log(1/\eta)}{n_\xi}}))^4 + \lambda^2\|\hat{\Sigma}_v\|^{-2}}.\end{aligned}$$

□

4.4 Proof of Theorem 6

Proof 4. ***1. Privacy.*** *Given two neighboring data sets with the different samples* $\boldsymbol{\xi}$ *and* $\boldsymbol{\xi'}$*, the* l_2*-sensitivity of the second-moment matrix is*

$$\Big\|\frac{1}{n_\xi}\Big(\boldsymbol{\xi}\boldsymbol{\xi}^T - \boldsymbol{\xi'}\boldsymbol{\xi'}^T\Big)\Big\|_F \leq \frac{1}{n_\xi}\|\boldsymbol{\xi}\|_2^2 + \frac{1}{n_\xi}\|\boldsymbol{\xi'}\|_2^2 \leq \frac{2(\mathrm{Tr}(\Sigma) + d\log(\frac{n_\xi}{\eta}))}{n_\xi},$$

where $\boldsymbol{\xi}$ *and* $\boldsymbol{\xi'}$ *are the different samples in the neighboring data sets. From the Gaussian mechanism (**Theorem 2**), we get the privacy guarantee.*

2. Bound error. Lemma 6 *shows*

$$\hat{\Sigma}_\xi \geqslant \lambda_{min}(\Sigma)\big(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}})\big)^2.$$

*According to **Lemma 9**, we have*

$$\|(\hat{\Sigma}_\xi + \lambda\mathbf{I})^{-1}\| \leqslant \frac{1}{\lambda_{min}(\hat{\Sigma}_\xi) + \lambda} \leqslant \frac{1}{\lambda_{min}(\Sigma)\big(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}})\big)^2 + \lambda}.$$

If $\|(\hat{\Sigma}_\xi + \lambda\mathbf{I})^{-1}\|\|\mathbf{G}\| \leqslant \frac{\|\mathbf{G}\|}{\lambda_{min}(\Sigma)\big(1-O(\sqrt{\frac{d}{n_\xi}}+\sqrt{\frac{\log(1/\eta)}{n_\xi}})\big)^2+\lambda} \leqslant \frac{1}{2}$, ***Lemma 8*** *gives the following bound*

$$\begin{aligned}\|(\hat{\Sigma}_\xi + \lambda\mathbf{I} + \mathbf{G})^{-1} - (\hat{\Sigma}_\xi + \lambda\mathbf{I})^{-1}\| &\leqslant \frac{\|\mathbf{G}\|\|(\hat{\Sigma}_\xi + \lambda\mathbf{I})^{-1}\|^2}{1 - \|(\hat{\Sigma}_\xi + \lambda\mathbf{I})^{-1}\|\|\mathbf{G}\|} \\ &\leqslant 2\|\mathbf{G}\|\|(\hat{\Sigma}_\xi + \lambda\mathbf{I})^{-1}\|^2 \\ &\overset{(i)}{\leqslant} \frac{2\|\mathbf{G}\|}{\lambda_{min}^2(\Sigma)\big(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}})\big)^4 + \lambda^2},\end{aligned}$$

where $\frac{1}{(a+b)^2} \leqslant \frac{1}{a^2+b^2}, a, b > 0$. *So we need to bound* $\|\mathbf{G}\|$ *so that the condition* $\frac{\|\mathbf{G}\|}{\lambda_{min}(\Sigma)\big(1-O(\sqrt{\frac{d}{n_\xi}}+\sqrt{\frac{\log(1/\eta)}{n_\xi}})\big)^2+\lambda} \leqslant \frac{1}{2}$ *holds. From* ***Lemma 4****, we know that the sufficient condition is the enough number of private data* n_ξ *such that*

$$\frac{\sqrt{d^3\log(\frac{2d}{\eta})}(d^{-1}\mathrm{Tr}(\Sigma) + \log(\frac{n_\xi}{\eta}))}{\mu \cdot n_\xi\big(\lambda_{min}(\Sigma)\big(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}})\big)^2 + \lambda\big)} \leqslant \frac{1}{2}.$$

Then, using ***Lemma 4*** *again, we get*

$$\begin{aligned}\|(\hat{\Sigma}_\xi + \lambda\mathbf{I} + \mathbf{G})^{-1} - (\hat{\Sigma}_\xi + \lambda\mathbf{I})^{-1}\| &\leqslant \frac{2\|\mathbf{G}\|}{\lambda_{min}^2(\Sigma)\big(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}})\big)^4 + \lambda^2} \\ &\leqslant \frac{\sqrt{d^3\log(\frac{2d}{\eta})}}{\mu n_\xi} \cdot \frac{2(d^{-1}\mathrm{Tr}(\Sigma) + \log(\frac{n_\xi}{\eta}))}{\lambda_{min}^2(\Sigma)\big(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}})\big)^4 + \lambda^2}.\end{aligned}$$

□

4.5 Proof of Theorem 7

Proof 5. *Consider the event*

$$\mathcal{E} = \Big\{ \tilde{\boldsymbol{\Xi}} \ \ \textit{and} \ \ \tilde{\mathbf{y}}_\xi \ \ \textit{are not truncated.} \Big\}.$$

Due to the assumption of linear model, $\mathbf{y}_\xi$ *is the sum of the sub-Gaussian randoms, so it is a sub-Gaussian random and* $\tilde{y}_{\boldsymbol{\xi}_i} \leqslant \sqrt{1 + \log(\frac{2n_\xi}{\eta})}$, *w.p.* $1 - O(\eta)$, *from* ***Corollary 1****. We get the conclusion*

$$\mathbb{P}(\mathcal{E}) \geq 1 - O(\eta).$$

And, we will discuss the following analysis under the event. We know the original ridge regression is solved by the loss function

$$\mathcal{L}(\boldsymbol{\beta}; \boldsymbol{\Xi}, \mathbf{y}_\xi) = \frac{1}{2n} ||\mathbf{y}_\xi - \boldsymbol{\Xi}\boldsymbol{\beta}||_2^2 + \frac{\lambda}{2} ||\boldsymbol{\beta}||_2^2.$$

Taking the $\hat{\Sigma}_\upsilon^{-1/2}$ *into the loss function and replace* $\mathbf{y}_\xi$ *as* $\tilde{\mathbf{y}}_\xi$ *, we have*

$$\mathcal{L}(\hat{\Sigma}_\upsilon^{-1/2}\boldsymbol{\beta}; \boldsymbol{\Xi}, \tilde{\mathbf{y}}_\xi) = \frac{1}{2n} ||\tilde{\mathbf{y}}_\xi - \boldsymbol{\Xi}\hat{\Sigma}_\upsilon^{-1/2}\boldsymbol{\beta}||_2^2 + \frac{\lambda}{2} ||\hat{\Sigma}_\upsilon^{-1/2}\boldsymbol{\beta}||_2^2 = \frac{1}{2n} ||\tilde{\mathbf{y}}_\xi - \tilde{\boldsymbol{\Xi}}\boldsymbol{\beta}||_2^2 + \frac{\lambda}{2} ||\hat{\Sigma}_\upsilon^{-1/2}\boldsymbol{\beta}||_2^2,$$

where $\tilde{\boldsymbol{\Xi}}$ *is the transformed data. Then, the parameter* $\tilde{\boldsymbol{\beta}}$ *minimizes this formula, namely,*

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \mathcal{L}(\hat{\Sigma}_\upsilon^{-1/2}\boldsymbol{\beta}; \boldsymbol{\Xi}, \tilde{\mathbf{y}}_\xi) \\ &\overset{(i)}{=} \Big(\frac{\tilde{\boldsymbol{\Xi}}^T \tilde{\boldsymbol{\Xi}}}{n_\xi} + \lambda \hat{\Sigma}_\upsilon^{-1} \Big)^{-1} \Big(\frac{\tilde{\boldsymbol{\Xi}}^T \tilde{\mathbf{y}}_\xi}{n_\xi} \Big). \end{aligned}$$

Concluding from the equation (i)*, we have*

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &\overset{(i)}{=} \Big(\frac{\tilde{\boldsymbol{\Xi}}^T\tilde{\boldsymbol{\Xi}}}{n_\xi} + \lambda\hat{\Sigma}_\upsilon^{-1}\Big)^{-1}\Big(\frac{\tilde{\boldsymbol{\Xi}}^T\tilde{\mathbf{y}}_\xi}{n_\xi}\Big)\\ &= \Big(\frac{\hat{\Sigma}_\upsilon^{-1/2}\boldsymbol{\Xi}^T\boldsymbol{\Xi}\hat{\Sigma}_\upsilon^{-1/2}}{n_\xi} + \lambda\hat{\Sigma}_\upsilon^{-1}\Big)^{-1}\Big(\frac{\hat{\Sigma}_\upsilon^{-1/2}}{\hat{\sigma}_\upsilon}\frac{\boldsymbol{\Xi}^T\mathbf{y}_\xi}{n_\xi}\Big)\\ &= \frac{\hat{\Sigma}_\upsilon^{1/2}}{\hat{\sigma}_\upsilon}\Big(\frac{\boldsymbol{\Xi}^T\boldsymbol{\Xi}}{n_\xi} + \lambda\mathbf{I}\Big)^{-1}\Big(\frac{\boldsymbol{\Xi}^T\mathbf{y}_\xi}{n_\xi}\Big)\\ &\overset{(ii)}{=} \frac{\hat{\Sigma}_\upsilon^{1/2}}{\hat{\sigma}_\upsilon}\hat{\boldsymbol{\beta}},\end{aligned}$$

where $\hat{\sigma}_\upsilon = \sqrt{\frac{1}{n_\upsilon}\sum_{i=1}^{n_\upsilon} y_{\upsilon i}^2}$*. Combining the event* $\mathcal{E}$ *and the equation* (ii)*, we get the conclusion*

$$\hat{\boldsymbol{\beta}} = \hat{\sigma}_\upsilon\hat{\Sigma}_\upsilon^{-1/2}\tilde{\boldsymbol{\beta}},$$

with at least probability $1 - O(\eta)$*.* □

4.6 Proof of Theorem 8

Proof 6. ***1. Privacy.*** *For convenience, we denote* $\tilde{\Sigma}_\xi = \frac{\tilde{\boldsymbol{\Xi}}^T\tilde{\boldsymbol{\Xi}}}{n_\xi}$*,* $\tilde{\Sigma}_{\xi y} = \frac{\tilde{\boldsymbol{\Xi}}^T\tilde{\mathbf{y}}_\xi}{n_\xi}$ *and* $\hat{\Sigma}_\upsilon = \frac{\boldsymbol{\Upsilon}^T\boldsymbol{\Upsilon}}{n_\upsilon}$*. Then,* $\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1} + \mathbf{G}$ *satisfies* μ*-GDP from* ***Theorem 4****, where* $\hat{\Sigma}_\upsilon^{-1}$ *is unrelated with privacy.*

We mainly discuss the second term $\tilde{\Sigma}_{\xi y}+\mathbf{g}$. *The sensitivity of* $\tilde{\Sigma}_{\xi y}$ *is*

$$\begin{aligned}
\Delta_{\tilde{\Sigma}_{\xi y}} &= \max_{(\tilde{\boldsymbol{\Xi}},\tilde{\mathbf{y}}_\xi)\sim(\tilde{\boldsymbol{\Xi}}',\tilde{\mathbf{y}}'_{\xi'})} \frac{1}{n_\xi}\|\tilde{\boldsymbol{\Xi}}^T\tilde{\mathbf{y}}_\xi-\tilde{\boldsymbol{\Xi}}'^T\tilde{\mathbf{y}}'_{\xi'}\|_2 \\
&\overset{(i)}{=} \max_{\substack{(\tilde{\boldsymbol{\xi}},\tilde{y}_\xi)\\(\tilde{\boldsymbol{\xi}}',\tilde{y}'_{\xi'})}} \frac{1}{n_\xi}\|\tilde{\boldsymbol{\xi}}^T\tilde{y}_\xi-\tilde{\boldsymbol{\xi}}'^T\tilde{y}'_{\xi'}\|_2 \\
&\leq \max_{(\tilde{\boldsymbol{\xi}},\tilde{y}_\xi)} \frac{2}{n_\xi}\|\tilde{\boldsymbol{\xi}}^T\tilde{y}_\xi\|_2 \\
&\leq \max_{(\tilde{\boldsymbol{\xi}},\tilde{y}_\xi)} \frac{2}{n_\xi}\|\tilde{\boldsymbol{\xi}}^T\|_2\|\tilde{y}_\xi\|_2 \\
&\leq \frac{2\sqrt{d}}{n_\xi}(1+\log(\frac{2n_\xi}{\eta})),
\end{aligned}$$

where $(\tilde{\boldsymbol{\xi}},\tilde{y}_\xi)$ *and* $(\tilde{\boldsymbol{\xi}}',\tilde{y}'_{\xi'})$ *are the different samples in* (i). ***Theorem 1*** *shows that* $\tilde{\boldsymbol{\beta}}^{DP}$ *satisfies* $\sqrt{2}\mu$*-GDP.*

2. Accuracy. *We discuss the accuracy of the DP-PDTMRR under* ***Theorem 7****. We firstly analyze the error between* $\tilde{\boldsymbol{\beta}}^{DP}$ *and* $\tilde{\boldsymbol{\beta}}$,

$$\begin{aligned}
\|\tilde{\boldsymbol{\beta}}^{DP}-\tilde{\boldsymbol{\beta}}\|_2 &= \|(\tilde{\Sigma}_\xi+\lambda\hat{\Sigma}_\upsilon^{-1}+\mathbf{G})^{-1}(\tilde{\Sigma}_{\xi y}+\mathbf{g})-(\tilde{\Sigma}_\xi+\lambda\hat{\Sigma}_\upsilon^{-1})^{-1}\tilde{\Sigma}_{\xi y}\|_2 \\
&\leqslant \underbrace{\|(\tilde{\Sigma}_\xi+\lambda\hat{\Sigma}_\upsilon^{-1}+\mathbf{G})^{-1}-(\tilde{\Sigma}_\xi+\lambda\hat{\Sigma}_\upsilon^{-1})^{-1}\|_2}_{(*)}\|\tilde{\Sigma}_{\xi y}\|_2+\underbrace{\|(\tilde{\Sigma}_\xi+\lambda\hat{\Sigma}_\upsilon^{-1}+\mathbf{G})^{-1}\|_2\|\mathbf{g}\|_2}_{(\diamond)},
\end{aligned}$$

where $\tilde{\Sigma}_\xi=\frac{\tilde{\boldsymbol{\Xi}}^T\tilde{\boldsymbol{\Xi}}}{n_\xi}$ *and* $\tilde{\Sigma}_{\xi y}=\frac{\tilde{\boldsymbol{\Xi}}^T\tilde{\mathbf{y}}_\xi}{n_\xi}$.

2.1 Bound $(*)$***.*** *Using* ***Theorem 5****, we get*

$$(*)\leqslant O\Big(\frac{\sqrt{d^3\log(\frac{2d}{\eta})}}{\mu n_\xi}\cdot\frac{(1+\log(\frac{2n_\xi}{\eta}))}{L^2(1-O(\sqrt{\frac{d}{n_\xi}}+\sqrt{\frac{\log(1/\eta)}{n_\xi}}))^4+\lambda^2\|\hat{\Sigma}_\upsilon\|^{-2}}\Big). \tag{12}$$

2.2 Bound $(\diamond)$***.*** *From* ***Lemma 8****, we have*

$$(\diamond) \leqslant 2\|(\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1})^{-1}\|\|\mathbf{g}\|.$$

Since $\mathbf{g} \sim \mathcal{N}(0, \sigma_2^2\mathbf{I})$ *with* $\sigma_2 = \frac{2\sqrt{d}\Big(1+\log\Big(\frac{2n_\xi}{\eta}\Big)\Big)}{\mu\cdot n_\xi}$*, and by applying* ***Lemma 5****, we obtain*

$$\begin{aligned}\|\mathbf{g}\| &\leqslant \sigma_2(\sqrt{d} + \sqrt{\log(\frac{1}{\eta})}) \\ &\leqslant O\Big(\frac{(d + \sqrt{d\log(\frac{1}{\eta})})(1+\log(\frac{2n_\xi}{\eta}))}{\mu\cdot n_\xi}\Big).\end{aligned}$$

As the proof in ***Theorem 5****, we have*

$$\frac{1}{L(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}}))^2 + \lambda\|\hat{\Sigma}_\upsilon\|^{-1}} \geqslant \|(\tilde{\Sigma}_\xi + \lambda\hat{\Sigma}_\upsilon^{-1})^{-1}\|,$$

where $L = \frac{n_\upsilon}{(\sqrt{n_\upsilon}+O(\sqrt{d}+\sqrt{\log(\frac{1}{\eta})}))^2}$*. So, we get*

$$(\diamond) \leqslant O\Big(\frac{(d + \sqrt{d}\log(\frac{1}{\eta}))}{\mu\cdot n_\xi} \cdot \frac{(1+\log(\frac{2n_\xi}{\eta}))}{L(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}}))^2 + \lambda\|\hat{\Sigma}_\upsilon\|^{-1}}\Big). \tag{13}$$

Note that the order $(\diamond)$ *is smaller than* $(*)$*, so we can ignore the order of* $(\diamond)$ *in the final bound.*

2.3 Bound $\tilde{\Sigma}_{\xi\upsilon}$***.*** *Because of the compatibility of the matrix norm, we have*

$$\|\tilde{\Sigma}_{\xi\upsilon}\| \leqslant \hat{\sigma}_\upsilon^{-1}\|\hat{\Sigma}_\upsilon^{-1/2}\|\|\hat{\Sigma}_{\xi\upsilon}\| \overset{(i)}{\leqslant} \hat{\sigma}_\upsilon^{-1}\|\hat{\Sigma}_\upsilon^{-1/2}\| \cdot O(\|\Sigma\|\|\boldsymbol{\beta}\|), \tag{14}$$

where (i) *is due to* $\|\frac{\boldsymbol{\Xi}^T\boldsymbol{\Xi}}{n_\xi}\| \to \|\Sigma\|$ *and* $\|\frac{\boldsymbol{\Xi}^T\boldsymbol{\epsilon}}{n_\xi}\| \to o(1)$*, as* $n_\xi \to \infty$*.*

2.4 Final bound. *Combining the equations* (12), (13) *and* (14), *we have*

$$\|\tilde{\boldsymbol{\beta}}^{DP} - \tilde{\boldsymbol{\beta}}\| \leqslant O\Big(\frac{\sqrt{d^3\log(\frac{2d}{\eta})}}{\mu n_\xi} \cdot \frac{\hat{\sigma}_\upsilon^{-1}\|\hat{\Sigma}_\upsilon^{-1/2}\|\|\Sigma\|\|\boldsymbol{\beta}\|(1+\log(\frac{2n_\xi}{\eta}))}{L^2(1-O(\sqrt{\frac{d}{n_\xi}}+\sqrt{\frac{\log(1/\eta)}{n_\xi}}))^4+\lambda^2\|\hat{\Sigma}_\upsilon\|^{-2}}\Big),$$

with at least probability $1-O(\eta)$.

3. Original Bound. *From* ***Theorem 7****, the output* $\bar{\boldsymbol{\beta}}^{DP} = \hat{\sigma}_\upsilon \cdot \hat{\Sigma}_\upsilon^{-1/2} \cdot \tilde{\boldsymbol{\beta}}^{DP}$ *is bounded by, with at least probability* $1-O(\eta)$,

$$\|\hat{\sigma}_\upsilon\hat{\Sigma}_\upsilon^{-1/2}\tilde{\boldsymbol{\beta}}^{DP} - \hat{\boldsymbol{\beta}}\| \leqslant O\Big(\frac{\sqrt{d^3\log(\frac{2d}{\eta})}}{\mu n_\xi} \cdot \frac{\|\hat{\Sigma}_\upsilon^{-1}\|\|\Sigma\|\|\boldsymbol{\beta}\|(1+\log(\frac{2n_\xi}{\eta}))}{L^2(1-O(\sqrt{\frac{d}{n_\xi}}+\sqrt{\frac{\log(1/\eta)}{n_\xi}}))^4+\lambda^2\|\hat{\Sigma}_\upsilon\|^{-2}}\Big).$$

□

4.7 Proof of Theorem 12

Proof 7. ***1. Privacy.*** *From* ***Theorem 6****,* $\hat{\Sigma}_\xi + \lambda\mathbf{I} + \mathbf{G}$ *satisfies* μ*-GDP. Then, we consider the sensitivity of the term* $\hat{\Sigma}_{\xi y}$

$$\begin{aligned}
\Delta_{\hat{\Sigma}_{\xi y}} &= \max_{(\boldsymbol{\Xi},\mathbf{y}_\xi)\sim(\boldsymbol{\Xi}',\mathbf{y}'_{\xi'})} \frac{1}{n_\xi}\|\boldsymbol{\Xi}^T\mathbf{y}_\xi - \boldsymbol{\Xi}'^T\mathbf{y}'_{\xi'}\|_2 \\
&\overset{(i)}{=} \max_{\substack{(A,y_\xi)\\(A,y'_{\xi'})}} \frac{1}{n_\xi}\|\hat{A}^T y_\xi - \boldsymbol{\xi}'^T y'_{\xi'}\|_2 \\
&\leq \max_{(A,y_\xi)} \frac{2}{n_\xi}\|A^T\|_2\|y_\xi\|_2 \\
&\leq \frac{2(\mathrm{Tr}(\Sigma)+d\log(\frac{n_\xi}{\eta}))\|\boldsymbol{\beta}\|}{n_\xi},
\end{aligned}$$

so we have $\hat{\Sigma}_{\xi y}+\mathbf{g}$ *is* μ*-GDP. Through the composition property, we get* $\hat{\boldsymbol{\beta}}_{DP}$ *satisfies* $\sqrt{2}\mu$*-GDP.*

2. Bound. *We decompose the* $\|\hat{\boldsymbol{\beta}}_{DP} - \hat{\boldsymbol{\beta}}\|$ *as two terms* $(*)$ *and* $(\diamond)$*,*

$$\|\hat{\boldsymbol{\beta}}_{DP} - \hat{\boldsymbol{\beta}}\| \leqslant \underbrace{\|(\hat{\Sigma}_\xi + \lambda\mathbf{I} + \mathbf{G})^{-1} - (\hat{\Sigma}_\xi + \lambda\mathbf{I})^{-1}\|_2}_{(*)} \|\hat{\Sigma}_{\xi y}\| + \underbrace{\|(\hat{\Sigma}_\xi + \lambda\mathbf{I} + \mathbf{G})^{-1}\|_2 \|\mathbf{g}\|_2}_{(\diamond)}.$$

From ***Theorem 6****, we have*

$$(*) \leqslant \frac{\sqrt{d^3 \log(\frac{2d}{\eta})}}{\mu n_\xi} \cdot \frac{d^{-1}\mathrm{Tr}(\Sigma) + \log(\frac{n_\xi}{\eta})}{\lambda_{min}^2(\Sigma)\big(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}})\big)^4 + \lambda^2}.$$

From ***Lemma 8****, we have*

$$(\diamond) \leqslant 2\|(\hat{\Sigma}_\xi + \lambda\mathbf{I})^{-1}\|\|\mathbf{g}\|.$$

Since $\mathbf{g} \sim \mathcal{N}(0, \sigma_2^2)$*,* $\sigma_2 = \frac{2(\mathrm{Tr}(\Sigma) + d\log(\frac{n_\xi}{\eta}))\|\boldsymbol{\beta}\|}{\mu \cdot n_\xi}$ *and, by applying* ***Lemma 5****, we have*

$$\begin{aligned}\|\mathbf{g}\| &\leqslant \sigma_2(\sqrt{d} + \sqrt{\log(\frac{1}{\eta})}) \\ &\leqslant O\Big(\frac{(\sqrt{d^3} + d\sqrt{\log(\frac{1}{\eta})})(d^{-1}\mathrm{Tr}(\Sigma) + \log(\frac{n_\xi}{\eta}))\|\boldsymbol{\beta}\|}{\mu \cdot n_\xi}\Big).\end{aligned}$$

And the term $\|\hat{\Sigma}_{\xi y}\| \leqslant O(\|\Sigma\|\|\boldsymbol{\beta}\|)$*.*

3. Conclusion. *So combining these inequalities, we have*

$$\|\hat{\boldsymbol{\beta}}_{DP} - \hat{\boldsymbol{\beta}}\| \leqslant O\Big(\frac{\sqrt{d^3 \log(\frac{2d}{\eta})}}{\mu n_\xi} \cdot \frac{\|\Sigma\|\|\boldsymbol{\beta}\|(d^{-1}\mathrm{Tr}(\Sigma) + \log(\frac{n_\xi}{\eta}))}{\lambda_{min}^2(\Sigma)\big(1 - O(\sqrt{\frac{d}{n_\xi}} + \sqrt{\frac{\log(1/\eta)}{n_\xi}})\big)^4 + \lambda^2}\Big).$$

4.8 Proof of Theorem 9

Proof 8. ***1. Equivalent Estimation.*** *We get the equivalent conclusion by the following steps:*

$$\begin{aligned}\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\Xi}}) &= -\frac{1}{n_\xi}\sum_{i=1}^{n_\xi} l_i(\tilde{\boldsymbol{\xi}}_i^T\boldsymbol{\beta}) + \frac{\lambda}{2}\|\hat{\Sigma}_\upsilon^{-1/2}\boldsymbol{\beta}\|_2^2 \\ &= -\frac{1}{n_\xi}\sum_{i=1}^{n_\xi} l_i(\boldsymbol{\xi}_i^T\hat{\Sigma}_\upsilon^{-1/2}\boldsymbol{\beta}) + \frac{\lambda}{2}\|\hat{\Sigma}_\upsilon^{-1/2}\boldsymbol{\beta}\|_2^2 \\ &= \mathcal{L}(\hat{\Sigma}_\upsilon^{-1/2}\boldsymbol{\beta};\boldsymbol{\Xi}).\end{aligned}$$

Because of the convexity of the loss function $\mathcal{L}$ on the $\boldsymbol{\beta}$ and $\hat{\Sigma}_\upsilon^{1/2}$ is a defined transformation holding the convexity, we know

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \mathcal{L}(\hat{\Sigma}_\upsilon^{-1/2}\boldsymbol{\beta};\boldsymbol{\Xi}) = \arg\min_{\hat{\Sigma}_\upsilon^{1/2}\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta};\boldsymbol{\Xi}) = \hat{\Sigma}_\upsilon^{1/2}\arg\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta};\boldsymbol{\Xi}) = \hat{\Sigma}_\upsilon^{1/2}\hat{\boldsymbol{\beta}}.$$

Moreover, we can guarantee the iterative invariance at each iteration in the Newton's method.

2. Affine invariance of Newton's method. *We simplify the $\mathcal{L}(\boldsymbol{\beta};\boldsymbol{\Xi})$ as $\mathcal{L}(\boldsymbol{\beta})$ and let $\hat{\boldsymbol{\beta}} = \hat{\Sigma}_\upsilon^{-1/2}\tilde{\boldsymbol{\beta}}$ and $\mathcal{H}(\tilde{\boldsymbol{\beta}}) = \mathcal{L}(\hat{\Sigma}_\upsilon^{-1/2}\tilde{\boldsymbol{\beta}})$. The Newton step is given by*

$$\begin{aligned}\tilde{\boldsymbol{\beta}}^{(t+1)} &= \tilde{\boldsymbol{\beta}}^{(t)} - (\nabla^2_{\tilde{\boldsymbol{\beta}}}\mathcal{H}(\tilde{\boldsymbol{\beta}}^{(t)}))^{-1}(\nabla_{\tilde{\boldsymbol{\beta}}}\mathcal{H}(\tilde{\boldsymbol{\beta}}^{(t)})) \\ &= \tilde{\boldsymbol{\beta}}^{(t)} - (\hat{\Sigma}_\upsilon^{-1/2}\nabla^2_{\hat{\boldsymbol{\beta}}}\mathcal{L}(\hat{\boldsymbol{\beta}}^{(t)})\hat{\Sigma}_\upsilon^{-1/2})^{-1}(\hat{\Sigma}_\upsilon^{-1/2}\nabla_{\hat{\boldsymbol{\beta}}}\mathcal{L}(\hat{\boldsymbol{\beta}}^{(t)})) \\ &= \tilde{\boldsymbol{\beta}}^{(t)} - \hat{\Sigma}_\upsilon^{1/2}(\nabla^2_{\hat{\boldsymbol{\beta}}}\mathcal{L}(\hat{\boldsymbol{\beta}}^{(t)}))^{-1}(\nabla_{\hat{\boldsymbol{\beta}}}\mathcal{L}(\hat{\boldsymbol{\beta}}^{(t)})).\end{aligned}$$

Multiplying both sides by $\hat{\Sigma}_\upsilon^{-1/2}$, we have

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} - (\nabla^2_{\hat{\boldsymbol{\beta}}}\mathcal{L}(\hat{\boldsymbol{\beta}}^{(t)}))^{-1}(\nabla_{\hat{\boldsymbol{\beta}}}\mathcal{L}(\hat{\boldsymbol{\beta}}^{(t)})). \qquad \square$$

4.9 Proof of Theorem 10

Proof 9. *We analyze every iteration t. The sensitivity of $\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\Xi}})$ is*

$$\begin{aligned}\max_{(\tilde{\boldsymbol{\Xi}},\mathbf{y})\sim(\tilde{\boldsymbol{\xi}},\mathbf{y}')}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\Xi}})-\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\xi}})\| &= \max_{(\tilde{\boldsymbol{\Xi}},\mathbf{y})\sim(\tilde{\boldsymbol{\xi}},\mathbf{y}')}\|-\frac{1}{n_\xi}\tilde{\boldsymbol{\Xi}}^T(\mathbf{y}-\tilde{\mathbf{p}})+\frac{1}{n_\xi}\tilde{\boldsymbol{\Xi}}'^T(\mathbf{y}'-\tilde{\mathbf{p}})\|\\ &\leqslant \max_{\tilde{\boldsymbol{\xi}}}\frac{2}{n_\xi}\|\tilde{\boldsymbol{\xi}}\|\\ &\leqslant \frac{2\sqrt{d(1+\log(\frac{2n_\xi}{\eta}))}}{n_\xi}.\end{aligned}$$

The sensitivity of $\nabla^2\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\Xi}})$ is

$$\begin{aligned}\max_{\tilde{\boldsymbol{\Xi}},\tilde{\boldsymbol{\xi}}'}\|\nabla^2\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\Xi}})-\nabla^2\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\mathbf{A}}')\|_F &= \max_{\tilde{\boldsymbol{\Xi}},\tilde{\boldsymbol{\xi}}'}\|\frac{1}{n_\xi}\tilde{\boldsymbol{\xi}}^T\tilde{\boldsymbol{W}}_\beta\tilde{\boldsymbol{\xi}}-\frac{1}{n_\xi}\tilde{\boldsymbol{\xi}}'^T\tilde{\boldsymbol{W}}_\beta\tilde{\boldsymbol{\xi}}'\|_F\\ &\leqslant \max_{\tilde{\boldsymbol{\xi}}}\frac{2}{n_\xi}\tilde{p}(1-\tilde{p})\|\tilde{\boldsymbol{\xi}}\|_2^2\\ &\leqslant \frac{d(1+\log(\frac{2n_\xi}{\eta}))}{2n_\xi}.\end{aligned}$$

So every iteration satisfies $\sqrt{\frac{2}{T}}\mu$-GDP and the final output satisfies $\sqrt{2}\mu$-GDP. □

4.10 Proof of Theorem 11

Proof 10. ***Lemma 10*** *illustrates that the starting value condition $\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta}^{(0)};\tilde{\boldsymbol{\Xi}})\|_2\leq\frac{\gamma_L^2}{C_{\tilde{\boldsymbol{\xi}}}}$ induces to the following inequality*

$$\frac{C_{\tilde{\boldsymbol{\xi}}}}{2\gamma_L^2}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta}^{(t+1)};\boldsymbol{\Xi})\|_2\leq\big(\frac{C_{\tilde{\boldsymbol{\xi}}}}{2\gamma_L^2}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta}^{(0)};\tilde{\boldsymbol{\Xi}})\|_2\big)^2+C_{pri},\tag{15}$$

where C_{pri} is a constant about $O(\frac{C_{\tilde{\boldsymbol{\xi}}}\sigma_1\sqrt{d\log(Td/\eta)}}{\gamma_L^3})$, $\sigma_1=\frac{\sqrt{T}d(1+\log(\frac{2n_\xi}{\eta}))}{2\mu n_\xi}$, $\gamma_L=\tau_0L\big(1-O\big(\sqrt{\frac{d}{n_\xi}}+\sqrt{\frac{\log(1/\eta)}{n_\xi}}\big)\big)^2+\frac{\lambda}{\lambda_{\max}(\hat{\Sigma}_v)}$ and $L=\frac{n_v}{(\sqrt{n_v}+O(\sqrt{d}+\sqrt{\log(\frac{1}{\eta})}))^2}$, $C_{\tilde{\boldsymbol{\xi}}}=\frac{(d(1+\log(n_\xi/\eta)))^{3/2}}{6\sqrt{3}}$,

and $\eta > 0$. *Moreover, we simplify* $C_{pri} = O(C_{\tilde{\boldsymbol{\xi}}} \frac{\sqrt{Td^3 \log(Td/\eta)}(1+\log(\frac{2n_\xi}{\eta}))}{\mu \cdot n_\xi \cdot \gamma_L^3})$. *We prove Eq.*(15) *by mathematical induction: we can get the inequality when* $t = 1$ *by Eq.*(8), *and then we assume the inequality holds for some* $t > 1$. *Then, we get*

$$\begin{aligned}
\frac{C_{\tilde{\boldsymbol{\xi}}}}{2\gamma_L^2}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta}^{(t+1)};\boldsymbol{\Xi})\|_2 &\overset{(i)}{\leq} \big(\frac{C_{\tilde{\boldsymbol{\xi}}}}{2\gamma_L^2}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta}^{(t)};\tilde{\boldsymbol{\Xi}})\|_2\big)^2 + C_{pri} \\
&\overset{(ii)}{\leq} \Big(\big(\frac{C_{\tilde{\boldsymbol{\xi}}}}{2\gamma_L^2}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta}^{(0)};\tilde{\boldsymbol{\Xi}})\|_2\big)^{2^t} + 3C_{pri}\Big)^2 + C_{pri} \\
&\overset{(iii)}{\leq} \Big(\frac{C_{\tilde{\boldsymbol{\xi}}}}{2\gamma_L^2}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta}^{(0)};\tilde{\boldsymbol{\Xi}})\|_2\Big)^{2^{(t+1)}} + C_{pri}(\frac{3}{2} + 9C_{pri}) + C_{pri} \\
&\overset{(iv)}{\leq} \Big(\frac{C_{\tilde{\boldsymbol{\xi}}}}{2\gamma_L^2}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta}^{(0)};\tilde{\boldsymbol{\Xi}})\|_2\Big)^{2^{(t+1)}} + 3C_{pri},
\end{aligned}$$

where the inequality (i) *is from* (8); *the inequality* (ii) *comes from the induction; the inequality* (iii) *follows by* $\frac{C_{\tilde{\boldsymbol{\xi}}}}{2\gamma_L^2}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta}^{(t)};\tilde{\boldsymbol{\Xi}})\|_2 \leq \frac{1}{2}$; *and we can guarantee the inequality* (iv) *through controlling the private sample size* n_ξ, *making* $C_{priv} \leq \frac{1}{18}$. *That completes the conclusion Eq.*(15).

The inequality (15) *illustrates that*

$$\frac{C_{\tilde{\boldsymbol{\xi}}}}{2\gamma_L^2}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta}^{(T)};\boldsymbol{\Xi})\|_2 \leq \big(\frac{1}{2}\big)^{2^T} + 3C_{pri},$$

and when $T > \frac{1}{\log 2}\log\big(\frac{\log 2}{C_{pri}}\big) = O(\log\log(n_\xi))$, *we get*

$$4C_{pri} \geq \frac{C_{\tilde{\boldsymbol{\xi}}}}{2\gamma_L^2}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\beta}^{(T)};\boldsymbol{\Xi})\|_2 \overset{(i)}{\geq} \frac{C_{\tilde{\boldsymbol{\xi}}}}{2\gamma_L}\|\boldsymbol{\beta}^{(T)} - \tilde{\boldsymbol{\beta}}\|_2,$$

where the inequality (i) *follows for the* γ_L*-strong convexity. Moreover, we have*

$$\|\boldsymbol{\beta}^{(T)} - \tilde{\boldsymbol{\beta}}\|_2 \leq \frac{8\gamma_L}{C_{\tilde{\boldsymbol{\xi}}}}C_{pri} \leq O(\frac{\gamma_L}{C_{\tilde{\boldsymbol{\xi}}}}\frac{C_{\tilde{\boldsymbol{\xi}}}\sigma_1\sqrt{d\log(Td/\eta)}}{\gamma_L^3}) \leq O(\frac{\sigma_1\sqrt{d\log(Td/\eta)}}{\gamma_L^2}),$$

$\sigma_1 = \frac{\sqrt{T}d(1+\log(\frac{2n_\xi}{\eta}))}{2\mu n_\xi}$, $\gamma_L = \tau_0 L\big(1-O\big(\sqrt{\frac{d}{n_\xi}}+\sqrt{\frac{\log(1/\eta)}{n_\xi}}\big)\big)^2+\frac{\lambda}{\lambda_{\max}(\hat{\Sigma}_\upsilon)}$ *and* $L = \frac{n_\upsilon}{(\sqrt{n_\upsilon}+O(\sqrt{d}+\sqrt{\log(\frac{1}{\eta})}))^2}$.

□.

Remark 4. *We remark the proof of* ***Theorem 13*** *here. Considering* $\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\Xi}}) = \mathcal{L}(\boldsymbol{\beta};\boldsymbol{\Xi})$ *and its properties' parameters in* ***Lemma 2*** *and* ***Lemma 10****, then the rest of the proof follows the proof of* ***Theorem 11*** *in* ***Appendix 4.10****.*

4.11 Proof of Theorem 14

Proof 11. ***1. Privacy.*** *The sensitivity of gradient* $\nabla_\beta\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\Xi}})$ *is*

$$\begin{aligned}
\max_{(\tilde{\boldsymbol{\Xi}},\mathbf{y})\sim(\tilde{\boldsymbol{\Xi}}',\mathbf{y}')} \|\nabla_\beta\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\Xi}}) - \nabla_\beta\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\Xi}}')\| &= \max_{(\tilde{\boldsymbol{\xi}}_i,y_i)\sim(\tilde{\boldsymbol{\xi}}_i',y_i')} \frac{1}{n_\xi}\|\nabla_\beta l_i(\tilde{\boldsymbol{\xi}}_i^\top\boldsymbol{\beta}) - \nabla_\beta l_i(\tilde{\boldsymbol{\xi}'}_i^\top\boldsymbol{\beta})\| \\
&= \max_{(\tilde{\boldsymbol{\xi}}_i,y_i)\sim(\tilde{\boldsymbol{\xi}}_i',y_i')} \frac{1}{n_\xi}\|(y_i - b'(\tilde{\boldsymbol{\xi}}_i^\top\boldsymbol{\beta}))\tilde{\boldsymbol{\xi}}_i - (y_i' - b'(\tilde{\boldsymbol{\xi}'}_i^\top\boldsymbol{\beta}))\tilde{\boldsymbol{\xi}}_i'\| \\
&\le \frac{2}{n_\xi}\max_{\tilde{\boldsymbol{\xi}}_i,y_i}\{|y_i|\|\tilde{\boldsymbol{\xi}}_i\| + |b'(\tilde{\boldsymbol{\xi}'}_i^\top\boldsymbol{\beta})|\|\tilde{\boldsymbol{\xi}}_i'\|\} \\
&\le \frac{2\sqrt{d(1+\log(\frac{2n_\xi}{\eta}))}}{n_\xi}(R_y + M_{b'}),
\end{aligned}$$

where $M_{b'} = max_z b'(z), z \in [-R_\beta\sqrt{d(1+\log(\frac{2n_\xi}{\eta}))}, R_\beta\sqrt{d(1+\log(\frac{2n_\xi}{\eta}))}]$ *and* $\|\boldsymbol{\beta}\|_2 \le R_\beta$*. The sensitivity of gradient* $\nabla_\beta^2\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\Xi}})$ *is*

$$\begin{aligned}
\max_{(\tilde{\boldsymbol{\Xi}},\mathbf{y})\sim(\tilde{\boldsymbol{\Xi}}',\mathbf{y}')} \|\nabla_\beta^2\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\Xi}}) - \nabla_\beta^2\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\Xi}}')\|_F &= \max_{(\tilde{\boldsymbol{\xi}}_i,y_i)\sim(\tilde{\boldsymbol{\xi}}_i',y_i')} \frac{1}{n_\xi}\|\nabla_\beta^2 l_i(\tilde{\boldsymbol{\xi}}_i^\top\boldsymbol{\beta}) - \nabla_\beta^2 l_i(\tilde{\boldsymbol{\xi}'}_i^\top\boldsymbol{\beta})\|_F \\
&\le \max_{\tilde{\boldsymbol{\xi}}_i,\tilde{\boldsymbol{\xi}}_i'} \frac{2}{n_\xi}\|b''(\tilde{\boldsymbol{\xi}}_i^\top\boldsymbol{\beta})\tilde{\boldsymbol{\xi}}_i\tilde{\boldsymbol{\xi}}_i^\top\|_F \\
&\le \max_{\tilde{\boldsymbol{\xi}}_i} \frac{2}{n_\xi}|b''(\tilde{\boldsymbol{\xi}}_i^\top\boldsymbol{\beta})|\|\tilde{\boldsymbol{\xi}}_i\|_2^2 \\
&\le \frac{2}{n_\xi}d(1+\log(\frac{2n_\xi}{\eta}))M_{b''},
\end{aligned}$$

where $M_{b''} = max_z b''(z), z \in [-R_\beta\sqrt{d(1+\log(\frac{2n_\xi}{\eta}))}, R_\beta\sqrt{d(1+\log(\frac{2n_\xi}{\eta}))}]$ *and* $\|\boldsymbol{\beta}\|_2 \leq R_\beta$.

Hence, let $\sigma_1 = \frac{2\sqrt{T}d(1+\log(\frac{2n_\xi}{\eta}))M_{b''}}{\mu n_\xi}$ *and* $\sigma_2 = \frac{2\sqrt{Td(1+\log(\frac{2n_\xi}{\eta}))}(R_y+M_{b'})}{\mu n_\xi}$. *Considering the* T *times compositions of GDP (**Theorem 1**), the final output satisfy* $\sqrt{2}\mu-GDP$.

2. Convergence. *Considering the conditions and **Lemma 10**, the proof is similar to the proof of **Theorem 11**.* □

5 Proofs of lemmas

5.1 Proof of Corollary 1

Proof 12. *Combining **Theorem 3** and the* n_ξ *union bound of **Lemma 7**, we get the proof. The second inequality is because* U *and* L *tend to* 1 *so that* $\mathrm{Tr}(\tilde{\Sigma}) \to d$. □

5.2 Proof of Lemma 2

Proof 13. ***1. Strong convexity of*** $\mathcal{L}(\boldsymbol{\beta};\boldsymbol{\Xi})$***.*** *From the assumption* (2) *and **Lemma 6**, with at least probability* $1-\eta$*, the* $\mathcal{L}(\boldsymbol{\beta};\boldsymbol{\Xi})$ *is* γ_Σ*-strong convexity. Namely, the Hessian matrix*

$$\begin{aligned}\nabla^2\mathcal{L}(\boldsymbol{\beta};\boldsymbol{\Xi}) = \mathbf{H}_\beta + \lambda\mathbf{I} &\succcurlyeq (\tau_0\lambda_{\min}(\hat{\Sigma}_\xi)+\lambda)\mathbf{I} \\ &\succcurlyeq \Big(\tau_0\lambda_{\min}(\Sigma)\big(1-O\big(\sqrt{\frac{d}{n_\xi}}+\sqrt{\frac{\log(1/\eta)}{n_\xi}}\big)\big)^2+\lambda\Big)\mathbf{I}.\end{aligned} \tag{16}$$

2. Strong convexity of $\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\Xi}})$***.*** *From the proof of **Theorem 9**, the loss function* $\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\Xi}})$ *is equivalent to* $\mathcal{L}(\hat{\Sigma}_\upsilon^{-1/2}\boldsymbol{\beta};\boldsymbol{\Xi})$*. So the Hessian matrix of* $\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\Xi}})$ *is*

$$\nabla^2\tilde{\mathcal{L}}(\boldsymbol{\beta};\tilde{\boldsymbol{\Xi}}) = \nabla^2\mathcal{L}(\hat{\Sigma}_\upsilon^{-1/2}\boldsymbol{\beta};\boldsymbol{\Xi}) = \hat{\Sigma}_\upsilon^{-1/2}(\mathbf{H}_{\hat{\Sigma}_\upsilon^{-1/2}\beta}+\lambda\mathbf{I})\hat{\Sigma}_\upsilon^{-1/2}.$$

Due to ***Theorem 3****, with at least probability* $1-2\eta$*, we have*

$$\begin{aligned}\hat{\Sigma}_v^{-1/2}(\mathbf{H}_{\hat{\Sigma}_v^{-1/2}\beta}+\lambda\mathbf{I})\hat{\Sigma}_v^{-1/2} &\succcurlyeq (\tau_0\lambda_{\min}(\hat{\Sigma}_v^{-1/2}\hat{\Sigma}_\xi\hat{\Sigma}_v^{-1/2})+\lambda\cdot\lambda_{\min}(\hat{\Sigma}_v^{-1}))\mathbf{I}\\ &\succcurlyeq (\tau_0 L\big(1-O\big(\sqrt{\frac{d}{n_\xi}}+\sqrt{\frac{\log(1/\eta)}{n_\xi}}\big)\big)^2+\lambda\cdot\lambda_{\min}(\hat{\Sigma}_v^{-1}))\mathbf{I},\end{aligned}$$

where $L=\frac{n_v}{(\sqrt{n_v}+O(\sqrt{d}+\sqrt{\log(\frac{1}{\eta})}))^2}$*. Here,* τ_0 *is still the lower bound, because the transformation doesn't change the classified accuracy of the final estimation.* □

5.3 Proof of Lemma 3

Proof 14. *We begin to get the Lipschitz constant of* $l_i(\boldsymbol{\xi}_i^T\boldsymbol{\beta})$*. We know the Hessian matrix of* $l_i(\boldsymbol{\xi}_i^T\boldsymbol{\beta})$ *is*

$$\nabla^2_{\boldsymbol{\beta}} l_i(\boldsymbol{\xi}_i^T\boldsymbol{\beta}) = p_i(1-p_i)\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T,$$

and its third-order derivative is

$$\nabla^3_{\boldsymbol{\beta}} l_i(\boldsymbol{\xi}_i^T\boldsymbol{\beta}) = p_i(1-p_i)(1-2p_i)\boldsymbol{\xi}_i\otimes\boldsymbol{\xi}_i\otimes\boldsymbol{\xi}_i.$$

So the Lipschitz constant of Hessian matrix is

$$\sup_{\boldsymbol{\beta}}\nabla^3_{\boldsymbol{\beta}} l_i(\boldsymbol{\xi}_i^T\boldsymbol{\beta}) \leqslant \sup_{0<p_i\leqslant\frac{1}{2}}\|p_i(1-p_i)(1-2p_i)\|\sup_{\boldsymbol{\xi}_i}\|\boldsymbol{\xi}_i\otimes\boldsymbol{\xi}_i\otimes\boldsymbol{\xi}_i\| \leqslant \frac{\sup_i\|\boldsymbol{\xi}_i\|^3}{6\sqrt{3}}.$$

That illustrates $C_\xi=\frac{\sup_i\|\boldsymbol{\xi}_i\|^3}{6\sqrt{3}}$ *and*

$$\|\nabla^2_{\boldsymbol{\beta}} l_i(\boldsymbol{\xi}_i^T\boldsymbol{\beta}) - \nabla^2_{\boldsymbol{\beta}} l_i(\boldsymbol{\xi}_i^T\boldsymbol{\beta}')\| \leqslant C_\xi\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|.$$

Then, we have the Lipschitz continuity of Hessian matrix $\mathbf{H}_\beta$ *is*

$$\begin{aligned}\|\mathbf{H}_\beta - \mathbf{H}_{\beta'}\|_2 &= \|\frac{1}{n_\xi}\sum_{i=1}^{n_\xi}\nabla^2_{\boldsymbol{\beta}} l_i(\boldsymbol{\xi}_i^T\boldsymbol{\beta}) - \frac{1}{n_\xi}\sum_{i=1}^{n_\xi}\nabla^2_{\boldsymbol{\beta}} l_i(\boldsymbol{\xi}_i^T\boldsymbol{\beta'})\|_2 \\ &\leqslant \frac{1}{n_\xi}\sum_{i=1}^{n_\xi}\|\nabla^2_{\boldsymbol{\beta}} l_i(\boldsymbol{\xi}_i^T\boldsymbol{\beta}) - \nabla^2_{\boldsymbol{\beta}} l_i(\boldsymbol{\xi}_i^T\boldsymbol{\beta'})\|_2 \\ &\leqslant C_\xi\|\boldsymbol{\beta} - \boldsymbol{\beta'}\|_2.\end{aligned} \tag{17}$$

The Lipschitz continuity of the Hessian matrix $\mathbf{H}_\beta$ *is proved.* □

5.4 Proof of Lemma 7

Proof 15. *For simplicity, we assume that* $K \geq 1$. *Since the random* x_i *is sub-gaussian,* x_i^2 *is sub-exponential, and more precisely*

$$\big\|\|\mathbf{x}\|_2^2\big\|_{\psi_1} = \|\sum_{i=1}^d x_i^2\|_{\psi_1} \leq \sum_{i=1}^d \|x_i^2\|_{\psi_1} \leq d\max_i \|x_i^2\|_{\psi_1} = d\max_i \|x_i\|_{\psi_2}^2.$$

Then, we compute the expectation of $\|\mathbf{x}\|_2^2$

$$\mathbb{E}\|\mathbf{x}\|_2^2 = \mathbb{E}\mathbf{x}^T\mathbf{x} = \mathbb{E}\,\mathrm{Tr}(\mathbf{x}^T\mathbf{x}) = \mathbb{E}\,\mathrm{Tr}(\mathbf{x}\mathbf{x}^T) = \mathrm{Tr}(\mathbb{E}\mathbf{x}\mathbf{x}^T) = \mathrm{Tr}(\Sigma) = \sum_{i=1}^d \lambda_i.$$

Reconsider the tail bound of the centered sub-exponential random, we have

$$\mathbb{P}\Big[\Big|\|\mathbf{x}\|_2^2 - \mathbb{E}\|\mathbf{x}\|_2^2\Big| > t\Big] \leq 2\exp\big(-\frac{ct}{dK^2}\big),$$

where $K = \max_i \|x_i\|_{\psi_2}$ *and* c *is an absolute constant.* □

5.5 Poof of Lemma 10

Proof 16. *Let* $\boldsymbol{s}_t = \nabla\mathcal{L}(\boldsymbol{\beta}^{(t)};\boldsymbol{\Xi})$ *and* $\boldsymbol{H}_t = \nabla^2\mathcal{L}(\boldsymbol{\beta}^{(t)};\boldsymbol{\Xi})$. *The corresponding noisy gradient and noisy Hessian used in the GDP Newton update are* $\widetilde{\boldsymbol{s}}_t = \boldsymbol{s}_t + \boldsymbol{g}_t$ *and* $\widetilde{\boldsymbol{H}}_t = \boldsymbol{H}_t + \boldsymbol{G}_t$, *where* $\boldsymbol{g}_t$ *and* $\boldsymbol{G}_t$ *denote the Gaussian perturbations added to the gradient and Hessian queries, respectively. The update can be written as*

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \widetilde{\boldsymbol{H}}_t^{-1}\widetilde{\boldsymbol{s}}_t.$$

We decompose it into the exact Newton step plus a perturbation:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \boldsymbol{\Delta}^{(t)} + \widetilde{\boldsymbol{N}}_t, \qquad \boldsymbol{\Delta}^{(t)} = -\boldsymbol{H}_t^{-1}\boldsymbol{s}_t,$$

where $\widetilde{\boldsymbol{N}}_t = -\widetilde{\boldsymbol{H}}_t^{-1}\widetilde{\boldsymbol{s}}_t + \boldsymbol{H}_t^{-1}\boldsymbol{s}_t$. *Equivalently,*

$$\widetilde{\boldsymbol{N}}_t = -\widetilde{\boldsymbol{H}}_t^{-1}(\boldsymbol{s}_t + \boldsymbol{g}_t) + \boldsymbol{H}_t^{-1}\boldsymbol{s}_t = \left(\boldsymbol{H}_t^{-1} - \widetilde{\boldsymbol{H}}_t^{-1}\right)\boldsymbol{s}_t - \widetilde{\boldsymbol{H}}_t^{-1}\boldsymbol{g}_t.$$

Thus $\widetilde{\boldsymbol{N}}_t$ *contains both the Hessian perturbation and the gradient perturbation.*

By local γ*-strong convexity,* $\|\boldsymbol{H}_t^{-1}\|_2 \le \gamma^{-1}$. *Based on the assumption of* n_ξ *small enough and on the high-probability event* $\|\boldsymbol{G}_t\|_2 \le \gamma/2$, *Weyl's inequality gives*

$$\lambda_{\min}(\widetilde{\boldsymbol{H}}_t) \ge \lambda_{\min}(\boldsymbol{H}_t) - \|\boldsymbol{G}_t\|_2 \ge \gamma/2.$$

Hence $\|\widetilde{\boldsymbol{H}}_t^{-1}\|_2 \le 2\gamma^{-1}$. *Using the inverse perturbation identity* $\boldsymbol{H}_t^{-1} - \widetilde{\boldsymbol{H}}_t^{-1} = \widetilde{\boldsymbol{H}}_t^{-1}\boldsymbol{G}_t\boldsymbol{H}_t^{-1}$,

we obtain

$$\begin{aligned}\|\widetilde{\boldsymbol{N}}_t\|_2 &\le \|\boldsymbol{H}_t^{-1} - \widetilde{\boldsymbol{H}}_t^{-1}\|_2\|\boldsymbol{s}_t\|_2 + \|\widetilde{\boldsymbol{H}}_t^{-1}\|_2\|\boldsymbol{g}_t\|_2 \\ &\le \frac{2}{\gamma^2}\|\boldsymbol{G}_t\|_2\|\boldsymbol{s}_t\|_2 + \frac{2}{\gamma}\|\boldsymbol{g}_t\|_2.\end{aligned}$$

Under the induction hypothesis $\|\boldsymbol{s}_t\|_2 \le \min\{\gamma r, \gamma^2/C_h\}$*, we use in particular* $\|\boldsymbol{s}_t\|_2 \le \gamma^2/C_h$*. Moreover, by Gaussian concentration for the noisy gradient and Hessian queries, with probability at least* $1 - O(\eta)$*, uniformly over* $0 \le t \le T$*,* $\|\boldsymbol{g}_t\|_2 \lesssim \|\boldsymbol{G}_t\|_2 \lesssim \sigma_1\sqrt{d\log(Td/\eta)}$*.*

Therefore

$$\|\widetilde{\boldsymbol{N}}_t\|_2 \lesssim \frac{1}{C_h}\sigma_1\sqrt{d\log(Td/\eta)} + \frac{1}{\gamma}\sigma_1\sqrt{d\log(Td/\eta)} \lesssim \left(\frac{1}{C_h} + \frac{1}{\gamma}\right)\sigma_1\sqrt{d\log(Td/\eta)}.$$

Define $r_N := C\left(\frac{1}{C_h} + \frac{1}{\gamma}\right)\sigma_1\sqrt{d\log(Td/\eta)}$*. Then, on the same high-probability event,*

$$\|\widetilde{\boldsymbol{N}}_t\|_2 \le r_N.$$

We now control the next gradient. By Taylor's theorem,

$$\boldsymbol{s}_{t+1} = \nabla\mathcal{L}(\boldsymbol{\beta}^{(t)} + \boldsymbol{\Delta}^{(t)} + \widetilde{\boldsymbol{N}}_t; \boldsymbol{\Xi}) = \boldsymbol{s}_t + \boldsymbol{H}_t\left(\boldsymbol{\Delta}^{(t)} + \widetilde{\boldsymbol{N}}_t\right) + \boldsymbol{R}_t,$$

where the local C_h*-Lipschitz continuity of the Hessian gives* $\|\boldsymbol{R}_t\|_2 \le \frac{C_h}{2}\left\|\boldsymbol{\Delta}^{(t)} + \widetilde{\boldsymbol{N}}_t\right\|_2^2$*.*

Since $\boldsymbol{\Delta}^{(t)} = -\boldsymbol{H}_t^{-1}\boldsymbol{s}_t$*, we have* $\boldsymbol{s}_t + \boldsymbol{H}_t\boldsymbol{\Delta}^{(t)} = \boldsymbol{0}$*. Thus*

$$\boldsymbol{s}_{t+1} = \boldsymbol{H}_t\widetilde{\boldsymbol{N}}_t + \boldsymbol{R}_t.$$

Using $\|\boldsymbol{H}_t\|_2 \le B_h$, $\|\widetilde{\boldsymbol{N}}_t\|_2 \le r_N$, *and* $\|\boldsymbol{\Delta}^{(t)}\|_2 \le \|\boldsymbol{H}_t^{-1}\|_2\|\boldsymbol{s}_t\|_2 \le \gamma^{-1}\|\boldsymbol{s}_t\|_2$, *we obtain*

$$\|\boldsymbol{s}_{t+1}\|_2 \le B_h r_N + \frac{C_h}{2}\left(\gamma^{-1}\|\boldsymbol{s}_t\|_2 + r_N\right)^2 \le \frac{C_h}{2\gamma^2}\|\boldsymbol{s}_t\|_2^2 + O\Big(B_h r_N + C_h r_N^2\Big).$$

Substituting the definition of r_N *yields*

$$\|\boldsymbol{s}_{t+1}\|_2 \le \frac{C_h}{2\gamma^2}\|\boldsymbol{s}_t\|_2^2 + O\left[B_h\left(\frac{1}{C_h}+\frac{1}{\gamma}\right)\sigma_1\sqrt{d\log(Td/\eta)}\right] + O\left[C_h\left(\frac{1}{C_h}+\frac{1}{\gamma}\right)^2\sigma_1^2 d\log(Td/\eta)\right].$$

Under the stated sample-size and privacy-noise conditions, the quadratic noise term is of smaller order than the linear privacy term. Hence

$$\|\boldsymbol{s}_{t+1}\|_2 \le \frac{C_h}{2\gamma^2}\|\boldsymbol{s}_t\|_2^2 + O\left[B_h\left(\frac{1}{C_h}+\frac{1}{\gamma}\right)\sigma_1\sqrt{d\log(Td/\eta)}\right].$$

Multiplying both sides by $C_h/(2\gamma^2)$, *we get*

$$\frac{C_h}{2\gamma^2}\|\boldsymbol{s}_{t+1}\|_2 \le \left(\frac{C_h}{2\gamma^2}\|\boldsymbol{s}_t\|_2\right)^2 + O\left[\frac{B_h}{\gamma^2}\left(1+\frac{C_h}{\gamma}\right)\sigma_1\sqrt{d\log(Td/\eta)}\right].$$

Equivalently, keeping only the dominant privacy order, this becomes

$$\frac{C_h}{2\gamma^2}\|\boldsymbol{s}_{t+1}\|_2 \le \left(\frac{C_h}{2\gamma^2}\|\boldsymbol{s}_t\|_2\right)^2 + O\left(\frac{C_h B_h \sigma_H\sqrt{d\log(Td/\eta)}}{\gamma^3}\right).$$

It remains to verify that the iterates stay in the local neighborhood. Suppose $\|\boldsymbol{s}_t\|_2 \le \min\{\gamma r, \gamma^2/C_h\}$. *By local* γ*-strong convexity around* $\widehat{\boldsymbol{\beta}}$, *this implies* $\|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}\|_2 \le r$. *Indeed, if* $\|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}\|_2 > r$, *let* $\bar{\boldsymbol{\beta}}$ *be the point on the line segment between* $\widehat{\boldsymbol{\beta}}$ *and* $\boldsymbol{\beta}^{(t)}$ *such that* $\|\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 =$

r. *Then strong convexity gives* $\left\langle \nabla \mathcal{L}(\bar{\boldsymbol{\beta}}; \boldsymbol{\Xi}), \frac{\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}}{\|\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2} \right\rangle \geq \gamma r$. *By monotonicity of the gradient along this ray, this would imply* $\|\boldsymbol{s}_t\|_2 = \|\nabla \mathcal{L}(\boldsymbol{\beta}^{(t)}; \boldsymbol{\Xi})\|_2 \geq \gamma r$, *contradicting the assumed strict inequality. Hence* $\boldsymbol{\beta}^{(t)} \in B_r(\widehat{\boldsymbol{\beta}})$.

Finally, by the sample-size condition, the privacy term satisfies

$$O\left(\frac{C_h B_h \sigma_1 \sqrt{d \log(Td/\eta)}}{\gamma^3}\right) \leq \frac{\gamma^2}{2C_h}.$$

Consequently, the recursion preserves

$$\|\boldsymbol{s}_{t+1}\|_2 \leq \min\{\gamma r, \gamma^2 / C_h\}.$$

Since the initialization satisfies the same bound by assumption, the claims follow by induction uniformly for $0 \leq t \leq T$.

6 Supplementary Experiments

The supplementary experiments contain the comparisons with DP-PMTRR, DP-RR and differentially private gradient descent (DP-GD), and the other real-world experiment about DP logistic regression. In the simulation experiments, the truncation radii and noise scales are calibrated according to the population quantities specified by the theoretical model, and the resulting procedures follow the stated privacy requirements. In the real-data experiments, for the private-only baselines DP-RR and DP-LR, full-sample empirical moments are used only as oracle approximations to the corresponding population quantities. This oracle calibration gives these baselines favorable truncation radii and noise scales and is used solely for utility

benchmarking. It is not part of a claimed end-to-end differentially private implementation.

6.1 DP Ridge Regression

6.1.1 Simulations

We compare our method with DP-RR and DP-GD (DP gradient descent, Avella-Medina et al. (2023)) under the same privacy parameters and without regularization, varying the private data size n_ξ. The amount of public data is fixed: $n_\upsilon = 200$ in the simulation setting. For DP-GD, several tuning parameters must be specified to ensure convergence. We determine these values through preliminary experiments. In the simulation study, we set the number of iterations $T = 1000$, gradient clipping threshold $c = 1$, and step size $lr = 0.09$.

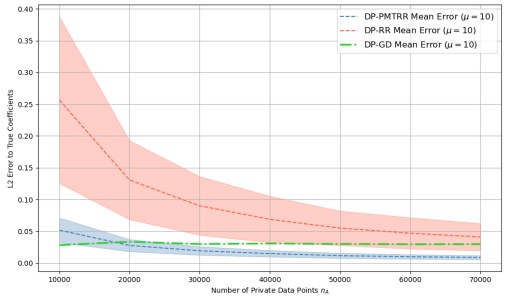


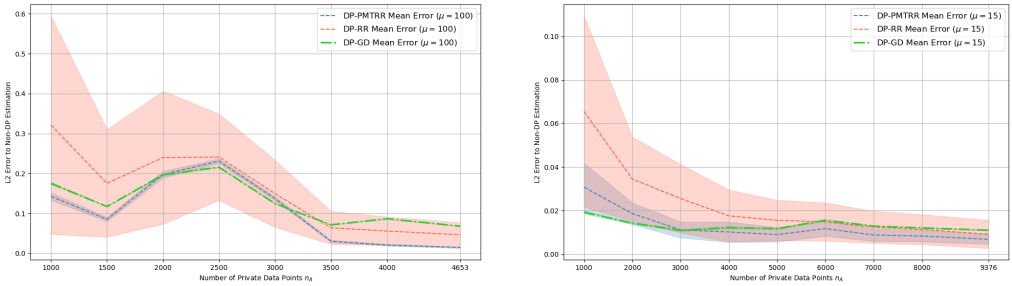
Figure 5: Simulations on DP-PMTRR, DP-RR and DP-GD.

Figure 5 presents the averaged errors (depicted as lines) along with their corresponding standard deviations (represented by shaded regions). The results indicate that DP-GD exhibits superior robustness compared to the Sufficient Statistics Perturbation (SSP) methods, where the iterative standard deviation of DP-GD is so small that the figures cannot be displayed. However, its performance shows limited sensitivity to the size of the private dataset. In contrast, both DP-RR and DP-PMTRR demonstrate a clear decreasing trend in error as the amount of

private data increases, with DP-PMTRR consistently outperforming DP-RR in terms of both robustness and accuracy. These differences can be attributed to the algorithmic characteristics of the respective methods. DP-GD is an iterative first-order method, where noise is injected into the gradients at every iteration. While this allows for general robustness, the convergence phase is particularly susceptible to noise disturb and the influence of hyperparameter tuning. As a result, the benefit from increasing data size is limited due to the persistent noise at each step. On the other hand, the SSP methods (including DP-RR and DP-PMTRR) are one-shot, tuning-free, and precise procedures based on second-order information. Although these methods introduce noise into the Hessian matrix (i.e., second-order information), whose inverse is inherently more sensitive to noise, they provide better data efficiency, and under favorable conditions—such as large private dataset sizes or highly ill-conditioned second-moment matrices (as in our method), they outperform DP-GD in terms of accuracy and stability. Furthermore, SSP methods generally involve lower tuning and computational costs compared to DP-GD. Overall, DP-PMTRR achieves better accuracy and robustness than others, even just leveraging a small amount of data.

6.2 Real-world Datasets

Firstly, we compare our method with DP-RR and DP-GD under the same privacy parameters and without regularization, varying the private data size n_ξ. In all experiments, the amount of public data is fixed: $n_\upsilon = 245$ for the White-wine Quality dataset and $n_\upsilon = 192$ for the Combined Cycle Power Plant dataset. For DP-GD, several tuning parameters must be specified to ensure convergence. We determine these values through preliminary experiments. In the White-wine Quality dataset, we use $T = 1000$, $c = 3.0$, and $lr = 0.5$. For the Combined Cycle Power Plant dataset, we use $T = 1000$, $c = 2.0$, and $lr = 0.04$.



(a) White-wine Quality (b) Combined Cycle Power Plant

Figure 6: Real-world experiments on DP-PMTRR, DP-RR and DP-GD.

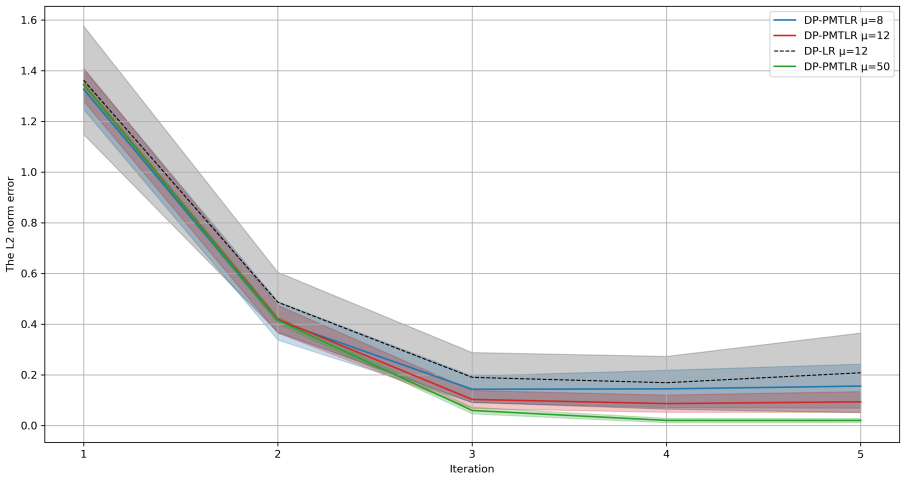
Figure 6 presents the averaged errors (depicted as lines) along with their corresponding standard deviations (represented by shaded regions). The results in the real-world datasets are similar to the first simulation and indicate DP-PMTRR achieves better accuracy and robustness than others for real-world data.

6.3 DP Logistic Regression

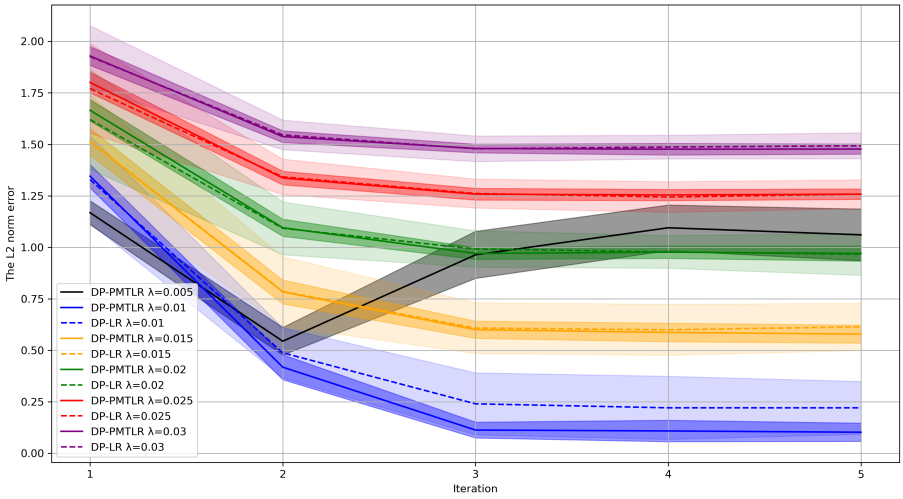
The Banknote Authentication dataset was extracted from images that were taken for the evaluation of an authentication procedure for banknotes, including 1372 samples and 4 features. we add a column of vectors containing all 1s, so the final feature dimension is $d = 5$. We set aside 10% for the public dataset. All experiments are repeated 100 times, and we show the averaged l_2-norm errors between the DP and the non-DP estimations at every iteration.

Firstly, we compare the DP-LR and DP-PMTLR (our method) with different privacy parameters μ, as shown in Figure 7(a). In the Banknote Authentication dataset, we fix the regularization parameter $\lambda = 0.01$ and the probability parameter $\eta = 1e - 3$. The privacy parameters is $\mu = 10$. Secondly, we explore the impact of the regularization parameter in the real-world datasets, seeing Figure 7(b). We fix the private data, the public data, and the

probability parameter $\eta = 1e-3$. The privacy parameter is $\mu = 10$.



(a) Different privacy parameters.



(b) Different regularization parameters.

Figure 7: Real experiments of DP-PMTLR and DP-LR.

Figure 7 displays the averaged errors (as lines) and the standard deviations of errors (as shaded areas). Overall, this result illustrates our method also outperforms standard DP

logistic regression in both utility and robustness in practice. Figure 7(b) shows that DP-LR fails to converge when $\lambda = 0.005$, and DP-PMTLR arrives at the best performance when $\lambda = 0.01$ (original regularization parameter). These results indicate that our method adequately addresses the trade-off concerning regularization, and its utility and resilience exhibit minimal dependence on the selection of λ.

6.4 Mismatch and High-dimension

This section explores the utility of our PMT method in DP-PMTRR when the distributions of public data and private data are mismatched. For the concise and effective measurement of the mismatch between the private data and public data, we adopt the eigenvalues mismatch of the second moment as follows. The private is

$$\Sigma_\xi = Q\,\mathrm{diag}(\lambda_1, \ldots, \lambda_d)Q^\top, \qquad \lambda_j = \kappa^{-(j-1)/(d-1)}.$$

The public follows the corrected mismatch path

$$\Sigma_v(\tau) = Q\,\mathrm{diag}(\lambda_1^{1-\tau}, \ldots, \lambda_d^{1-\tau})Q^\top, \qquad \tau \in [0, 1].$$

Hence, $\Sigma_v(\tau)^{-1/2}\Sigma_\xi\Sigma_v(\tau)^{-1/2} = Q\,\mathrm{diag}(\lambda_1^\tau, \ldots, \lambda_d^\tau)Q^\top$, and the transformed population condition number equals

$$\mathrm{cond}\{\Sigma_v(\tau)^{-1/2}\Sigma_\xi\Sigma_v(\tau)^{-1/2}\} = \kappa^\tau.$$

Thus, $\tau = 0$ corresponds to exact public/private second-moment matching, while $\tau = 1$ means the public covariance is isotropic and no longer captures the anisotropy of Σ_ξ. Hence, τ is a controllable and concise indicator used to describe the mismatches.

6.4.1 Mismatch experiments

In this section, we set the low dimension $d = 10$ and $\kappa = 5$ with different privacy parameters μ to evaluate our method in different mismatch levels. The private data $n_\xi = 2000$ and the public data $n_\upsilon = 100$. We repeated the experiment 300 times and show the averaged estimation errors $\|\hat{\boldsymbol{\beta}}_{DP} - \hat{\boldsymbol{\beta}}\|_2$. At each repetition, we only fix Σ_ξ, $Q, \lambda_j, \beta^\star, X_A, y_A$ for all mismatch levels τ_i. Figure 8 represents the results.

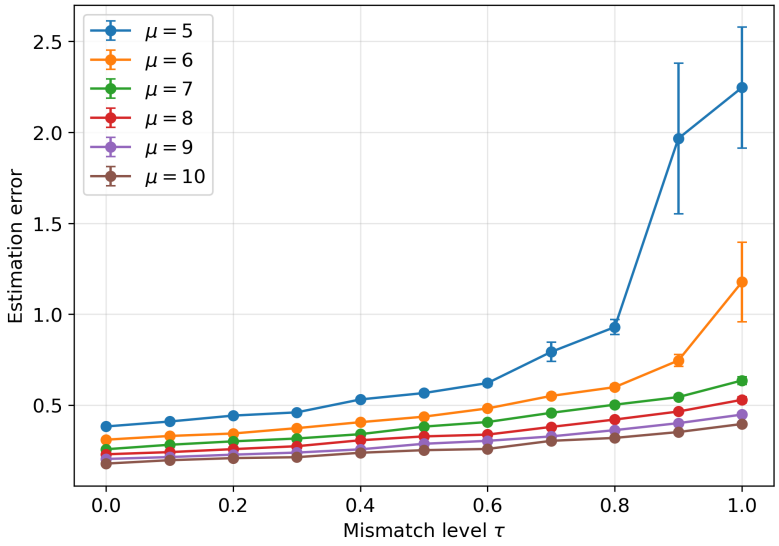


Figure 8: Mismatch levels τ_i with different parameters μ

The results clearly demonstrate the impact of public/private second-moment mismatch on the proposed PMT method. As expected, the estimation error increases when stronger privacy protection is imposed, that is, when the Gaussian DP parameter μ decreases. As the mismatch level increases, the performance of DP-PMTRR deteriorates, and the largest errors are observed at the most mismatched setting $\tau = 1$. In this case, the public second-moment matrix is close to the identity matrix, so the public preconditioner no longer captures the anisotropic structure of the private covariance. Consequently, the transformed private data retain nearly the same

second-moment structure as the original private data, and the benefit of PMT is substantially weakened.

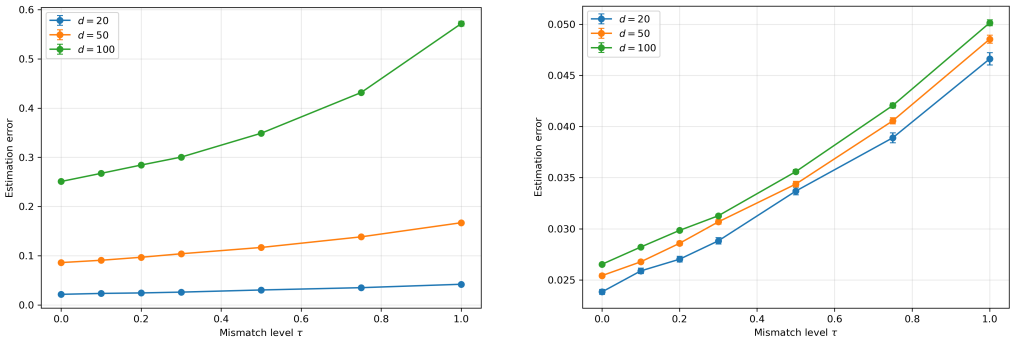
The negative effect of mismatch becomes more pronounced under stronger privacy protection. For relatively large privacy parameters, such as $\mu \in \{7, 8, 9, 10\}$, the estimation error increases only mildly as the mismatch level grows. In contrast, for smaller privacy parameters, such as $\mu \in \{5, 6\}$, the error increases sharply once the mismatch level exceeds approximately $\tau = 0.7$. This suggests that PMT is robust to mild or moderate public/private mismatch when the privacy noise is not too large, but its advantage becomes more sensitive to mismatch in the high-privacy regime.

6.4.2 High-dimension experiments

In this section, we evaluate the performance of the proposed method in high-dimensional settings. We consider covariate dimensions $d \in \{20, 50, 100\}$, set the condition number parameter to $\kappa = 5$, and use $n_v = 500$ public samples. The Gaussian differential privacy parameter is fixed at $\mu = 50$. Although this privacy level is relatively large, it provides a practical compromise for the high-dimensional experiments: substantially smaller values of μ lead to highly unstable estimates, especially when $d = 100$, and obscure the comparison of the underlying preconditioning effect. For the number of private samples n_ξ, we consider two sampling regimes in order to evaluate both the empirical performance of the proposed method and its consistency with the theoretical scaling,

$$n_\xi = 1e4, \qquad \text{and} \qquad n_\xi = 100 \times \sqrt{d^3}.$$

The first fixes the private data $n_\xi = 1e4$ so as to evaluate the impact of the mismatch level for the PMT method in a high-dimensional setting. The second $n_\xi = 100 \times \sqrt{d^3}$ is designed to verify our theory order $O(\frac{\sqrt{d^3}}{n_\xi})$.



(a) Fixed $n_\xi = 1e4$ (b) Dimension related $n_\xi = 100 \times \sqrt{d^3}$

Figure 9: High-dimensional experiments.

For the fixed-private-sample regime, Figure (9a) shows how the dimension affects the utility of the proposed method under different mismatch levels. The effect of mismatch becomes more pronounced as the dimension increases. In particular, the case $d = 100$ is more sensitive to the discrepancy between the public and private second moments, and the estimation error increases rapidly as the mismatch level becomes larger.

Figure (9b) provides two further observations. First, the empirical results are consistent with the theoretical scaling $O(\sqrt{d^3}/n_\xi)$. After normalizing the private sample size according to this scaling, the errors across different dimensions remain at the order of 10^{-2}. Second, the mismatch between the public and private second moments remains an intrinsic factor affecting the utility of PMT. Even after controlling for the dimensional effect through the sample-size scaling, larger mismatch levels still lead to higher estimation errors. This confirms that the effectiveness of PMT depends not only on the dimension and sample size, but also on the representativeness of the public second-moment matrix.

References

Avella-Medina, M., C. Bradshaw, and P.-L. Loh (2023). Differentially private inference via noisy optimization. *The Annals of Statistics 51*(5), 2067–2092.

Laurent, B. and P. Massart (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, 1302–1338.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Zilong Cao, School of Mathematics, Northwest University, Xi'an, China

E-mail: nwu_czl@stumail.nwu.edu.cn

Xuan Bi, Department of Information and Decision Sciences, University of Minnesota, USA

E-mail: xbi@umn.edu

Hai Zhang, School of Mathematics, Northwest University, Xi'an, China

E-mail:zhanghai@nwu.edu.cn