DP-RoLL: Fortifying Private Low-Rank Learning against Data Contamination

Anonymous authors
Paper under double-blind review

Abstract

Multivariate response learning often involves sensitive individual-level data and corrupted observations. Low-rank coefficient structure can improve statistical efficiency under differential privacy by exploiting shared response factors, but its benefits may be compromised by outliers. In this work, we propose a differentially private and robust low-rank learning framework for a broad range of loss functions. Theoretically, we establish a non-asymptotic error bound that decomposes the error into optimization, statistical, and privacy-induced components and reveals an optimization-privacy trade-off in the number of private iterations. Extensive experiments demonstrate the effectiveness of our proposed framework in protecting sensitive data while maintaining strong performance in the presence of data anomalies.

1 Introduction

Multivariate response data arise in a broad range of modern applications, ranging from gene expression network analysis in genomics (Zhi et al., 2013) to asset return modeling in finance (Senneret et al., 2016). To characterize the joint dependence of multiple responses on a common set of predictors, we consider a general modeling framework. Given a loss function l, not necessarily a negative log-likelihood function, a response matrix $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]^\top \in \mathbb{R}^{n\times m}$ and a design matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times p}$, the coefficient matrix $\boldsymbol{B}$ can be estimated by solving

$$\min_{\boldsymbol{B}\in\mathbb{R}^{p\times m}} l(\boldsymbol{\eta}; \boldsymbol{Y}) \ \text{ s.t. } \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{B}. \tag{1}$$

In many real-world applications, multivariate data often contain sensitive individual-level information (Thomas et al., 2024) and are also frequently affected by outliers or corrupted observations (She & Chen, 2017). Examples include multivariate physiological measurements in smart-healthcare systems, which raise patient-privacy and data-poisoning concerns (Gupta et al., 2021), and multi-attribute financial transaction records containing personal information and fraudulent transactions (Kanamori et al., 2022). Differential privacy (Dwork et al., 2006) provides a rigorous framework for protecting individual-level information with formal guarantees. However, achieving such protection requires randomized perturbation, which may reduce statistical utility (Geng et al., 2020). This loss of utility can be further compounded by outlying observations, which may substantially distort the fitted relationship between predictors and responses.

A natural way to mitigate the statistical impact of privacy perturbation is to exploit low-dimensional structure in the model. In multivariate learning, this can be achieved by imposing a low-rank structure on the coefficient matrix, reflecting the assumption that multiple responses are driven by a small number of shared latent factors. This structural modeling reduces the effective degrees of freedom relative to an unrestricted coefficient matrix, thereby improving estimation efficiency (Izenman, 1975; Yuan et al., 2007). This reduction in model complexity is particularly valuable under differential privacy, as it can improve statistical utility under privacy constraints (Yu et al., 2021; Wang et al., 2023). However, this advantage can be undermined by data contamination, as even a few outlying observations may distort the underlying reduced-rank structure (Chen, 2016).

To address this vulnerability, a common strategy is to replace the conventional loss with a robust alternative in certain special cases (Cui et al., 2023; Ma et al., 2024). For example, in the multivariate regression, Tan et al. (2023) proposed substituting the quadratic loss with Huber's loss. However,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

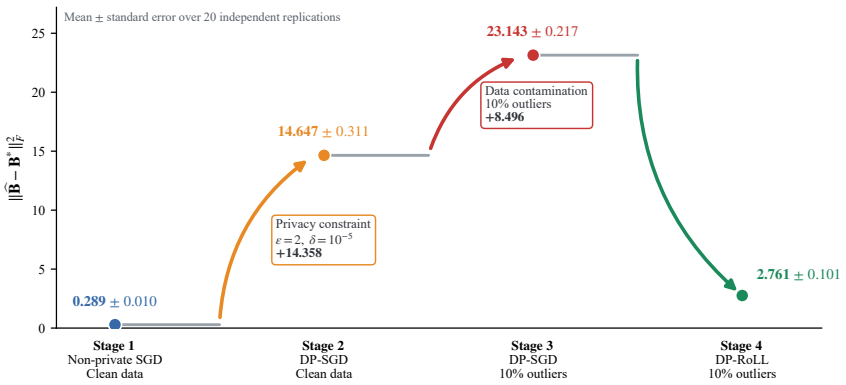

Figure 1: Motivating example in synthetic multi-label classification: moving from non-private SGD to DP-SGD increases estimation error, and 10% outliers further amplify this utility loss, whereas DP-RoLL markedly reduces the resulting error.

extending it beyond quadratic loss can be technically involved, since more complicated losses, such as logistic deviance, may require separately constructed robust criteria (Bianco & Yohai, 1996).

Given the practical importance of the challenges outlined above, existing methods have made progress but typically address only some of them. General-purpose differentially private empirical risk minimization applies to convex and smooth nonconvex objectives but does not exploit low-rank multivariate structure or provide robustness to outliers (Chaudhuri et al., 2011; Wang & Xu, 2019). The lack of robustness in general-purpose private learning is particularly concerning when sensitive data may also be corrupted (Asi et al., 2023). Yet private robust learning has received comparatively limited attention in multivariate settings, with existing work focusing mainly on mean estimation (Liu et al., 2021), single-response linear regression (Liu et al., 2023), and low-rank matrix approximation (Arora et al., 2018).

To address the above challenges, we propose DP-RoLL, a differentially private robust low-rank learning framework. Our main *contributions* are as follows: 1) A robust low-rank framework equipped with a privacy-preserving optimization algorithm is developed for multivariate response analysis. 2) We establish differential privacy and algorithmic convergence guarantees, derive a non-asymptotic error bound, and characterize the optimization–privacy trade-off. 3) Extensive experiments demonstrate the effectiveness of DP-RoLL.

Notations. We use $a \wedge b$ to denote $\min(a, b)$ and e for the Euler constant. For a set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinality, and $\mathbb{I}\{\mathcal{A}\}$ denotes the indicator of an event $\mathcal{A}$. For a vector $\boldsymbol{v}$, $\|\boldsymbol{v}\|_2$ denotes its Euclidean norm. We use $\boldsymbol{I}_d$ for the $d \times d$ identity matrix. For a matrix $\boldsymbol{A} = [a_{ij}] \in \mathbb{R}^{n \times m}$, $\boldsymbol{A}_{i\cdot}$ and $\boldsymbol{A}_{\cdot j}$ denote its i-th row and j-th column, respectively. We write $\|\boldsymbol{A}\|_F$ for its Frobenius norm, $\mathrm{rank}(\boldsymbol{A})$ for its rank, and $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B})$ for the matrix inner product. The standard vectorization operator is denoted by $\mathrm{vec}(\cdot)$, and $\|\boldsymbol{A}\|_{2,0} = \sum_{i=1}^n \mathbb{I}\{\|\boldsymbol{A}_{i\cdot}\|_2 \neq 0\}$. Let $P_H(t; \lambda) = (-t^2/2 + \lambda|t|)\mathbb{I}\{|t| < \lambda\} + (\lambda^2/2)\mathbb{I}\{|t| \geq \lambda\}$, and define $P_{2,H}(\boldsymbol{A}; \lambda) = \sum_{i=1}^n P_H(\|\boldsymbol{A}_{i\cdot}\|_2; \lambda)$. For nonnegative quantities a and b, we write $a \lesssim b$ and $a \gtrsim b$ if $a \leq cb$ and $a \geq cb$, respectively, for some universal constant $c > 0$, and write $a \asymp b$ if both relations hold. The notation $\mathcal{O}(b)$ denotes a quantity whose absolute value is bounded by cb for some universal constant $c > 0$.

2 Related Work

Robust Low-Rank Multivariate Learning. The low-rank assumption is widely used in multivariate modeling to capture shared response structure while reducing the effective number of parameters (Anderson, 1951; Cho & Park, 2022; Park et al., 2024); see Trendafilov et al. (2025) for a recent overview. Since outliers can mask this structure and significantly compromise model fitting (Guo & Lin, 2018), robust extensions have been developed using Huber- or Tukey-type losses for regression (Elsener & van de Geer, 2018; Tan et al., 2023; Ma et al., 2024) and generalized cross-entropy losses

for noisy-label classification (Zhang & Sabuncu, 2018). These methods, however, are typically tailored to specific losses or learning tasks and do not provide formal privacy guarantees.

Private Learning with Low-Dimensional Structure. Classical private empirical risk minimization relies on output perturbation, objective perturbation, and noisy-gradient methods (Chaudhuri et al., 2011; Abadi et al., 2016; Wang & Xu, 2019). Low-dimensional structure has subsequently been exploited to improve private learning through gradient projection, low-rank reparameterization, and fixed-rank matrix manifolds (Kairouz et al., 2021; Yu et al., 2021; Han et al., 2024). Related work has also studied private matrix and subspace estimation, including low-rank factorization (Upadhyay, 2018; Mangoubi et al., 2022), matrix completion (Chien et al., 2021; Wang et al., 2023), and principal component or subspace estimation (Liu et al., 2022; Tsfadia, 2024). However, their formulations do not cover low-rank multivariate response learning in the presence of data contamination.

Robust Learning under Differential Privacy. Recent studies have considered robustness and differential privacy simultaneously in several statistical problems. One line of work focuses on mean estimation under adversarial contamination (Liu et al., 2021), with subsequent extensions to robust multivariate mean estimation (Yu et al., 2024). Related work has explored regression problems, including scalar-response linear regression with corrupted responses (Liu et al., 2023) and private sparse least-absolute-deviation regression (Liu et al., 2024). A separate line of research has investigated robust low-rank matrix approximation (Arora et al., 2018) and robust private principal component analysis (Kim & Jung, 2025). Despite the practical importance of robust multivariate response learning, its development under differential privacy remains largely unexplored, particularly in the low-rank setting.

Appendix A provides a tabular comparison between DP-RoLL and representative related methods in terms of problem settings and theoretical guarantees.

3 DP-RoLL Framework

In this section, we introduce DP-RoLL for robust low-rank learning under privacy constraints. We begin by introducing the proposed formulation, followed by the development of its optimization algorithm.

Mean-shift modeling represents outlying effects through sparse observation-specific adjustments and has been used in robust regression (She & Owen, 2011; She & Chen, 2017). Motivated by this idea, we incorporate an adjustment matrix $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n]^\top \in \mathbb{R}^{n \times m}$ into equation 1 by redefining

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{B} + \sqrt{n}\boldsymbol{C} = [\boldsymbol{X}\ \sqrt{n}\boldsymbol{I}] \begin{pmatrix} \boldsymbol{B} \\ \boldsymbol{C} \end{pmatrix}, \tag{2}$$

where the scaling factor $\sqrt{n}$ is introduced as a normalization, ensuring that the columns of the augmented design matrix $[\boldsymbol{X}\ \sqrt{n}\boldsymbol{I}]$ are on a comparable scale. Although contamination may occur at the entry level, robust statistical analysis commonly focuses on observation-wise outliers (Liu et al., 2003; Huber, 2011). Accordingly, we impose row-wise sparsity on $\boldsymbol{C}$, with each nonzero row representing a contaminated observation in one or more responses. To capture the shared low-dimensional structure across responses, we further impose a rank constraint on $\boldsymbol{B}$ and propose the following robust low-rank learning framework:

$$\min_{\boldsymbol{B},\boldsymbol{C}} \quad l(\boldsymbol{\eta}; \boldsymbol{Y}) + \frac{\lambda_{\boldsymbol{C}}^2}{2}\|\boldsymbol{C}\|_{2,0} \quad \text{s.t.} \quad \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{B} + \sqrt{n}\boldsymbol{C}, \quad \operatorname{rank}(\boldsymbol{B}) \le r, \tag{3}$$

where $\lambda_{\boldsymbol{C}} > 0$ controls the row-wise sparsity of $\boldsymbol{C}$.

Our goal is to obtain a differentially private estimate of $\boldsymbol{B}$, with $\boldsymbol{C}$ used only as an internal robustification variable. Direct private optimization of equation 3 is nontrivial because the nonconvex rank constraint on $\boldsymbol{B}$ and the row-cardinality penalty on $\boldsymbol{C}$ must be handled simultaneously. Standard private gradient methods (Abadi et al., 2016) are not tailored to exploit the resulting low-rank and row-sparse structures. Moreover, since the data-dependent update of $\boldsymbol{C}$ affects subsequent gradient queries for $\boldsymbol{B}$, its influence must be accounted for when controlling the sensitivity of the private

updates (Dwork & Roth, 2014). To address these issues, we develop a privacy-preserving algorithm based on the majorization-minimization principle (Hunter & Lange, 2004).

Let $\bar{\boldsymbol{X}} = [\boldsymbol{X}\ \sqrt{n}\boldsymbol{I}]$ and $\bar{\boldsymbol{B}} = [\boldsymbol{B}^\top\ \boldsymbol{C}^\top]^\top$. Let $\vartheta \geq 1$ denote a rank-overspecification factor, so that the algorithm allows the rank of $\boldsymbol{B}$ to be bounded by ϑr. Starting from a quadratic surrogate of the loss, we replace the gradient block associated with $\boldsymbol{B}$ by its clipped counterpart and introduce Gaussian perturbation. The sequence of iterates is generated by

$$\bar{\boldsymbol{B}}^{(t+1)} \in \arg\min_{\mathrm{rank}(\boldsymbol{B})\leq\vartheta r} g(\bar{\boldsymbol{B}}; \bar{\boldsymbol{B}}^{(t)}), \tag{4}$$

where the privacy-perturbed surrogate is defined as

$$g(\bar{\boldsymbol{B}}; \bar{\boldsymbol{B}}^{(t)}) = l(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^{(t)}; \boldsymbol{Y}) + \left\langle \tilde{\nabla}_{\bar{\boldsymbol{B}}}^{(t)} + \boldsymbol{N}^{(t)}, \bar{\boldsymbol{B}} - \bar{\boldsymbol{B}}^{(t)} \right\rangle + \frac{\rho}{2}\|\bar{\boldsymbol{B}} - \bar{\boldsymbol{B}}^{(t)}\|_F^2 + \frac{\lambda_{\boldsymbol{C}}^2}{2}\|\boldsymbol{C}\|_{2,0}.$$

Here, ρ is the inverse step-size parameter, $\tilde{\nabla}_{\bar{\boldsymbol{B}}}^{(t)}$ denotes the clipped augmented gradient direction specified below, and

$$\boldsymbol{N}^{(t)} = \begin{pmatrix} \boldsymbol{N}_{\boldsymbol{B}}^{(t)} \\ \boldsymbol{0}_{n\times m} \end{pmatrix}, \tag{5}$$

where $\boldsymbol{N}_{\boldsymbol{B}}^{(t)} \in \mathbb{R}^{p\times m}$ has independent entries drawn from $\mathcal{N}(0, \sigma_{\mathrm{DP}}^2)$. Thus, Gaussian noise is added only to the gradient block associated with $\boldsymbol{B}$. The $\boldsymbol{C}$-block remains unperturbed and is retained as an internal variable rather than released. We next specify the clipped augmented gradient direction. At iteration t, let $\boldsymbol{Z}^{(t)} = \boldsymbol{X}\boldsymbol{B}^{(t)} + \sqrt{n}\boldsymbol{C}^{(t)}$, $\boldsymbol{G}^{(t)} = \nabla l(\boldsymbol{Z}^{(t)}; \boldsymbol{Y})$, and denote the i-th row of $\boldsymbol{G}^{(t)}$ by $\boldsymbol{G}_{i\cdot}^{(t)}$. Define the corresponding per-sample contribution to the $\boldsymbol{B}$-gradient as $\boldsymbol{U}_i^{(t)} = \boldsymbol{x}_i^\top \boldsymbol{G}_{i\cdot}^{(t)} \in \mathbb{R}^{p\times m}$. Given a public clipping threshold $C_B > 0$, each per-sample contribution is clipped as

$$\tilde{\boldsymbol{U}}_i^{(t)} = \frac{\boldsymbol{U}_i^{(t)}}{\max\left\{1, \|\boldsymbol{U}_i^{(t)}\|_F / C_B\right\}}, \qquad i = 1, \ldots, n, \tag{6}$$

so that $\|\tilde{\boldsymbol{U}}_i^{(t)}\|_F \leq C_B$. The clipped augmented gradient direction is then

$$\tilde{\nabla}_{\bar{\boldsymbol{B}}}^{(t)} = \begin{pmatrix} \sum_{i=1}^n \tilde{\boldsymbol{U}}_i^{(t)} \\ \sqrt{n}\boldsymbol{G}^{(t)} \end{pmatrix}. \tag{7}$$

Thus, clipping is applied only to the $\boldsymbol{B}$-gradient block, whereas the $\boldsymbol{C}$-gradient block remains unchanged. The privacy-perturbed gradient used in the $\boldsymbol{B}$-update is then $\hat{\nabla}_{\boldsymbol{B}}^{(t)} = \sum_{i=1}^n \tilde{\boldsymbol{U}}_i^{(t)} + \boldsymbol{N}_{\boldsymbol{B}}^{(t)}$. Noticing the separability in $\boldsymbol{B}$ and $\boldsymbol{C}$, the resulting updates are given by

$$\begin{aligned} \boldsymbol{B}^{(t+1)} &= \mathcal{P}_{\vartheta r}\left(\boldsymbol{B}^{(t)} - \frac{1}{\rho}\hat{\nabla}_{\boldsymbol{B}}^{(t)}\right), \\ \boldsymbol{C}_{i\cdot}^{(t+1)} &= \left(\boldsymbol{C}_{i\cdot}^{(t)} - \frac{\sqrt{n}}{\rho}\boldsymbol{G}_{i\cdot}^{(t)}\right)\mathbb{I}\left(\left\|\boldsymbol{C}_{i\cdot}^{(t)} - \frac{\sqrt{n}}{\rho}\boldsymbol{G}_{i\cdot}^{(t)}\right\|_2 \geq \frac{\lambda_{\boldsymbol{C}}}{\sqrt{\rho}}\right),\ i = 1, \ldots, n, \end{aligned} \tag{8}$$

where $\mathcal{P}_{\vartheta r}(\boldsymbol{A}) = \arg\min_{\mathrm{rank}(\boldsymbol{B})\leq\vartheta r} \|\boldsymbol{A} - \boldsymbol{B}\|_F^2$. The procedure is repeated for T iterations. When clipping is inactive, namely, when $\|\boldsymbol{U}_i^{(t)}\|_F \leq C_B$ for all i, we have $\tilde{\boldsymbol{U}}_i^{(t)} = \boldsymbol{U}_i^{(t)}$, and $\tilde{\nabla}_{\bar{\boldsymbol{B}}}^{(t)}$ coincides with the ordinary gradient $\bar{\boldsymbol{X}}^\top \nabla l(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^{(t)}; \boldsymbol{Y})$.

Remark 1 (Penalty versus Constraint on $\boldsymbol{C}$)**.** *A hard constraint $\|\boldsymbol{C}\|_{2,0} \leq s$ requires retaining the s rows with the largest norms. The resulting cutoff is the s-th largest row norm and therefore depends on the data. Replacing one observation may consequently change multiple rows of $\boldsymbol{C}$, complicating the sensitivity analysis of subsequent $\boldsymbol{B}$-updates. In contrast, the penalty $\lambda_{\boldsymbol{C}}^2\|\boldsymbol{C}\|_{2,0}/2$ yields row-wise hard thresholding at the fixed level $\lambda_{\boldsymbol{C}}/\sqrt{\rho}$. Under a sample-separable loss, the resulting row-wise locality makes the private sensitivity analysis tractable.*

4 Theoretical Analysis

In this section, we establish the theoretical properties of DP-RoLL. We first derive its differential privacy guarantee, then quantify the statistical error of its finite-iteration output, and finally characterize the convergence behavior of the privacy-perturbed algorithm.

4.1 Differential Privacy Guarantees

Let $D = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ denote the dataset. Two datasets D and D' are adjacent under replace-one adjacency, denoted by $D \simeq D'$, if they differ in exactly one sample. We assume that the loss is sample-wise separable, namely, $l(\boldsymbol{Z}; \boldsymbol{Y}) = \sum_{i=1}^n l_i(\boldsymbol{Z}_{i\cdot}; \boldsymbol{y}_i)$. Conditional on a fixed history of the previously released $\boldsymbol{B}$-iterates, $\boldsymbol{B}^{(t)}$ is fixed. We define the clipped pre-noise query at iteration t as $q_t(D) = \sum_{i=1}^n \tilde{\boldsymbol{U}}_i^{(t)}$. By the sample-wise separability of the loss and the row-wise separability of the $\boldsymbol{C}$-update, adjacent datasets differing only in the k-th sample have identical clipped gradient contributions for all $i \neq k$. Since $\|\tilde{\boldsymbol{U}}_i^{(t)}\|_F \leq C_B$, the global Frobenius sensitivity of the conditional query q_t satisfies

$$\Delta_B = \sup_{D \simeq D'} \|q_t(D) - q_t(D')\|_F \leq 2C_B, \tag{9}$$

where the supremum is taken conditional on the fixed released history. To account for the privacy loss accumulated over multiple iterations, we use zero-concentrated differential privacy (zCDP) (Bun & Steinke, 2016), which admits adaptive composition and can be converted to (ϵ, δ)-differential privacy (DP) (Dwork et al., 2006).

Theorem 2. *Let $\epsilon > 0$ and $\delta \in (0, 1)$. Suppose that the algorithm runs for T iterations and releases $\{\boldsymbol{B}^{(1)}, \ldots, \boldsymbol{B}^{(T)}\}$. At iteration t, let $\boldsymbol{N}_{\boldsymbol{B}}^{(t)}$ have independent entries drawn from $\mathcal{N}(0, \sigma_{\mathrm{DP}}^2)$, where*

$$\sigma_{\mathrm{DP}}^2 = \frac{T\Delta_B^2}{2\left(\sqrt{\epsilon + \log(1/\delta)} - \sqrt{\log(1/\delta)}\right)^2}. \tag{10}$$

The noise matrices are mutually independent across iterations. Then the released sequence satisfies (ϵ, δ)-DP. Consequently, any post-processing of this sequence, including the final output $\boldsymbol{B}^{(T)}$, is also (ϵ, δ)-differentially private.

4.2 Non-asymptotic Error Analysis

Having established the privacy guarantee, we next analyze the finite-iteration output of DP-RoLL by deriving a non-asymptotic error bound.

Our main tool to deal with the diverse form of the loss is the *generalized Bregman function* (GBF) (She et al., 2021): given any differentiable ψ

$$\boldsymbol{\Delta}_\psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \psi(\boldsymbol{\alpha}) - \psi(\boldsymbol{\beta}) - \langle \nabla\psi(\boldsymbol{\beta}), \boldsymbol{\alpha} - \boldsymbol{\beta}\rangle. \tag{11}$$

In particular, when ψ is differentiable and strictly convex, the generalized Bregman function $\boldsymbol{\Delta}_\psi$ becomes the standard Bregman divergence $\mathbf{D}_\psi(\boldsymbol{\alpha}, \boldsymbol{\beta})$. The Bregman associated with half of the squared Frobenius norm is denoted by $\mathbf{D}_2$: $\mathbf{D}_2(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbf{D}_{\|\cdot\|_F^2/2}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_F^2/2$. In general, $\boldsymbol{\Delta}_\psi(\boldsymbol{\alpha}, \boldsymbol{\beta})$ may not be symmetric, and we define its symmetrized version by $\bar{\boldsymbol{\Delta}}_\psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\boldsymbol{\Delta}_\psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\alpha}))/2$. To simplify the notation, we use $\boldsymbol{\Delta}_\psi \geq \boldsymbol{\Delta}_\phi$ to denote $\boldsymbol{\Delta}_\psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) \geq \boldsymbol{\Delta}_\phi(\boldsymbol{\alpha}, \boldsymbol{\beta})$ for all $\boldsymbol{\alpha}, \boldsymbol{\beta}$.

Introducing the notion of noise in the non-likelihood setting is another essential component. We define the effective noise associated with the statistical truth $\bar{\boldsymbol{B}}^* = [\boldsymbol{B}^{*\top}\ \boldsymbol{C}^{*\top}]^\top$ by $\boldsymbol{\mathcal{E}} = -\nabla l(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*)$. In a generalized linear model (GLM) with cumulant function $b(\cdot)$ and canonical link function $g = (\nabla b)^{-1}$, the loss is given by $-\langle \boldsymbol{Y}, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}\rangle + b(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}})$. By definition of $\boldsymbol{\mathcal{E}}$, we have $\boldsymbol{\mathcal{E}} = \boldsymbol{Y} - g^{-1}(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*) = \boldsymbol{Y} - \mathbb{E}(\boldsymbol{Y})$. For the multivariate regression, the effective noise $\boldsymbol{\mathcal{E}} = \boldsymbol{Y} - \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*$ corresponds to the raw noise, which is conventionally assumed to be Gaussian or sub-Gaussian. In fact, any loss with a bounded derivative, such as Huber's loss or Hampel's loss, results in a sub-Gaussian effective noise. Accordingly, we assume that $\mathrm{vec}(\boldsymbol{\mathcal{E}})$ is *sub-Gaussian* with mean zero and ψ_2-norm bounded by σ.

We next introduce the notation needed for the subsequent analysis. For any $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n]^\top$, define $\mathcal{J}(\boldsymbol{C}) = \{j : \boldsymbol{c}_j \neq \boldsymbol{0}\}, J(\boldsymbol{C}) = |\mathcal{J}(\boldsymbol{C})|$. Let $r = \mathrm{rank}(\boldsymbol{B}^*)$, $J = J(\boldsymbol{C}^*) \geq 1$ and $q = \mathrm{rank}(\boldsymbol{X})$. In the analysis below, we consider the regime in which the public clipping threshold upper-bounds the per-sample gradient norms along the iterates. The surrogate can then be rewritten using the GBF. Let $\psi(\bar{\boldsymbol{B}}) = \frac{\rho}{2}\|\bar{\boldsymbol{B}}\|_F^2 - l(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}; \boldsymbol{Y})$. Then, up to terms independent of $\bar{\boldsymbol{B}}$,

$$g(\bar{\boldsymbol{B}}; \bar{\boldsymbol{B}}^{(t)}) = l(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}; \boldsymbol{Y}) + \boldsymbol{\Delta}_\psi(\bar{\boldsymbol{B}}, \bar{\boldsymbol{B}}^{(t)}) + \left\langle \boldsymbol{N}^{(t)}, \bar{\boldsymbol{B}} - \bar{\boldsymbol{B}}^{(t)} \right\rangle + \frac{\lambda_{\boldsymbol{C}}^2}{2}\|\boldsymbol{C}\|_{2,0}. \tag{12}$$

As emphasized in She et al. (2021), in nonconvex surrogate algorithms, the choice of an appropriate error measure is often problem-dependent and can be naturally constructed through generalized Bregman functions. Following this perspective, we evaluate the statistical accuracy of the algorithmic iterate using the surrogate-induced discrepancy $\boldsymbol{\Delta}_\psi(\bar{\boldsymbol{B}}^*, \bar{\boldsymbol{B}}^{(T)})$. Define $P_{\mathrm{stat}}(r, J) = (1+\vartheta)r(q+m) + J(m+\log n)$ and $P_{\mathrm{DP}}(r) = (1+\vartheta)r(p+m) + \log n$.

Theorem 3. *Let $\lambda_{\boldsymbol{C}} = A\sigma(m+\log n)^{1/2}$ with $A > 0$ sufficiently large. Assume that the regularity condition*

$$
\begin{aligned}
&(2\bar{\boldsymbol{\Delta}}_l - K_1\mathbf{D}_2)(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*) + \tau\boldsymbol{\Delta}_l(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}) + \frac{\lambda_{\boldsymbol{C}}^2}{2}(\|\boldsymbol{C}\|_{2,0} - \|\boldsymbol{C}^*\|_{2,0}) + K\lambda_{\boldsymbol{C}}^2 J \\
&\geq \left(K_2 + \frac{\rho}{\sqrt{\vartheta}} + \tau\rho\right)\mathbf{D}_2(\boldsymbol{B}, \boldsymbol{B}^*) + (1+\tau)\rho\mathbf{D}_2(\boldsymbol{C}, \boldsymbol{C}^*) + K' P_{2,H}(\sqrt{n}(\boldsymbol{C} - \boldsymbol{C}^*); \lambda_{\boldsymbol{C}}).
\end{aligned} \tag{13}
$$

with $\mathrm{rank}(\boldsymbol{B}) \leq \vartheta r, \|\boldsymbol{C}\|_{2,0} \leq C_s J$ holds for some constants $K, K', K_1, K_2, \tau, C_s > 0$. Then, with probability at least $1 - \mathcal{O}(n^{-c_0})$, the T-th iterate generated by the algorithm satisfies

$$
\begin{aligned}
&\boldsymbol{\Delta}_\psi(\bar{\boldsymbol{B}}^*, \bar{\boldsymbol{B}}^{(T)}) \leq (1+\tau)^{-T}\,\boldsymbol{\Delta}_\psi(\bar{\boldsymbol{B}}^*, \bar{\boldsymbol{B}}^{(0)}) \\
&+C\{\sigma^2 P_{\mathrm{stat}}(r, J) + \sigma_{\mathrm{DP}}^2 P_{\mathrm{DP}}(r) + \sigma_{\mathrm{DP}}^2 \log T\}/\tau - \frac{1-(1+\tau)^{-T}}{\tau}\min_{1\leq t\leq T}\boldsymbol{\Delta}_\psi(\bar{\boldsymbol{B}}^{(t)}, \bar{\boldsymbol{B}}^{(t-1)}).
\end{aligned} \tag{14}
$$

where $c_0, C > 0$ are universal absolute constants.

The regularity condition equation 13 in Theorem 3 can be viewed as a composite restricted strong convexity (RSC) condition, which is standard in Bregman-based analysis (She et al., 2021; 2022). Compared to the classical RSC condition in Loh & Wainwright (2015), which requires a unit factor for the dominant curvature term $\bar{\boldsymbol{\Delta}}_l$, our condition equation 13 allows this factor to be 2. We also note that, according to the proof, equation 13 only needs to hold for $\bar{\boldsymbol{B}}^{(t)}$ ($0 \leq t \leq T$), and so different starting values may give different values of τ. Further discussion is provided in the Appendix B.4.

Theorem 3 yields an explicit estimation error bound for the low-rank coefficient matrix. Define $M_{\bar{\boldsymbol{X}}} = \sup_{\bar{\boldsymbol{B}}_1 \neq \bar{\boldsymbol{B}}_2} \|\bar{\boldsymbol{X}}(\bar{\boldsymbol{B}}_1 - \bar{\boldsymbol{B}}_2)\|_F^2 / \|\bar{\boldsymbol{B}}_1 - \bar{\boldsymbol{B}}_2\|_F^2$, where $\bar{\boldsymbol{B}}_k = [\boldsymbol{B}_k^\top\ \boldsymbol{C}_k^\top]^\top$ and $\mathrm{rank}(\boldsymbol{B}_k) \leq \vartheta r$ for $k = 1, 2$. Suppose that $l(\cdot; \boldsymbol{Y})$ is $\mathcal{L}$-smooth with respect to its first argument. Combining this with Theorem 3 and choosing ρ such that $\rho - \mathcal{L}M_{\bar{\boldsymbol{X}}} \asymp n$, we obtain the following rate

$$
\|\boldsymbol{B}^{(T)} - \boldsymbol{B}^*\|_F^2 \lesssim \frac{(1+\tau)^{-T}}{n}\boldsymbol{\Delta}_\psi(\bar{\boldsymbol{B}}^*, \bar{\boldsymbol{B}}^{(0)}) + \frac{\sigma^2 P_{\mathrm{stat}}(r, J) + \sigma_{\mathrm{DP}}^2 P_{\mathrm{DP}}(r) + \sigma_{\mathrm{DP}}^2 \log T}{\tau n}. \tag{15}
$$

Specifically, the error rate grows linearly with the rank r and contamination proportion J/n. This scaling is consistent with the empirical results in Appendix C.9 and C.10.

Remark 4. *The bound exhibits an explicit optimization–privacy trade-off. The term $(1+\tau)^{-T}\boldsymbol{\Delta}_\psi(\bar{\boldsymbol{B}}^*, \bar{\boldsymbol{B}}^{(0)})$ decays geometrically with T, whereas the privacy term $\sigma_{\mathrm{DP}}^2(P_{\mathrm{DP}}(r) + \log T)$ increases with the number of private iterations. Hence, an appropriate stopping time is needed to balance optimization error and privacy-induced error. Through the dependence of σ_{DP}^2 on (ϵ, δ), the bound also reflects the standard privacy–utility trade-off.*

4.3 Convergence Analysis

In this subsection, we characterize the convergence of DP-RoLL using successive-iterate differences, which provide a natural stability measure for the nonconvex rank-constrained updates in equation 8. Define $F(\bar{\boldsymbol{B}}) = l(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}; \boldsymbol{Y}) + \frac{\lambda_{\boldsymbol{C}}^2}{2}\|\boldsymbol{C}\|_{2,0}$.

Theorem 5. *Suppose that $l(\cdot; \boldsymbol{Y})$ has $\mathcal{L}$-Lipschitz continuous gradients with respect to its first argument and that $F_{\inf} = \inf_{\mathrm{rank}(\boldsymbol{B})\leq\vartheta r,\, \boldsymbol{C}\in\mathbb{R}^{n\times m}} F(\bar{\boldsymbol{B}}) > -\infty$. For any fixed $a > 0$, choose the inverse step size ρ such that $\rho > \mathcal{L}M_{\bar{\boldsymbol{X}}} + a$ and let $\omega = \rho - \mathcal{L}M_{\bar{\boldsymbol{X}}} - a > 0$. Define $\Phi_{\mathrm{DP}} = C\sigma_{\mathrm{DP}}^2[2\vartheta r(p+m) + \log T + c_0 \log n]/a$, where C, c_0 are universal positive constants. Then, with probability at least $1 - \mathcal{O}(n^{-c_0})$, we have*

$$
\min_{0\leq t\leq T-1}\|\boldsymbol{B}^{(t+1)} - \boldsymbol{B}^{(t)}\|_F^2 \leq \frac{2\big(F(\bar{\boldsymbol{B}}^{(0)}) - F_{\inf}\big)}{\omega T} + \frac{2\Phi_{\mathrm{DP}}}{\omega}.
$$

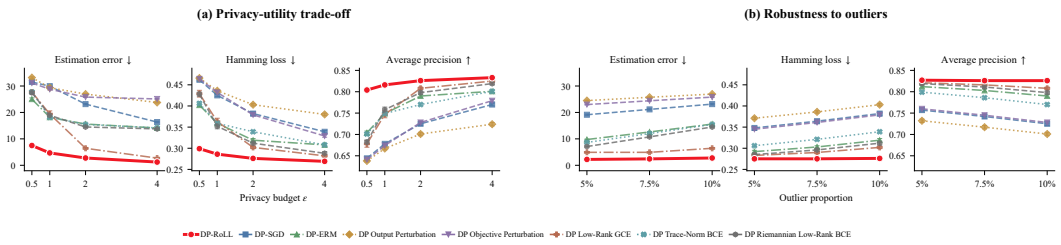

Figure 2: Privacy–utility performance and robustness to contamination in synthetic multi-label classification. Panel (a) varies the privacy budget ϵ, and panel (b) varies the outlier proportion at $\epsilon = 2$. Points and error bars denote means and standard errors over 20 independent replications.

The first term in Theorem 5 decreases at the rate $\mathcal{O}(1/T)$ and represents the optimization error determined by the initial objective gap. The second term is a privacy-induced error. In the non-private case, $\sigma_{\mathrm{DP}} = 0$ and hence $\Phi_{\mathrm{DP}} = 0$, recovering the standard $\mathcal{O}(1/T)$ rate for the minimum squared successive-iterate difference. Under a fixed total privacy budget, increasing T reduces the optimization error but amplifies the privacy-induced error through composition.

5 Numerical Experiments

In this part, we evaluate DP-RoLL primarily on multi-label classification tasks, where label dependence induces a low-rank coefficient structure (He & Yang, 2019) and corrupted observations commonly arise from label noise (Zhao et al., 2024). Additional experiments under multivariate regression are provided in Appendix D.

5.1 Performance on Synthetic Data

Synthetic data are generated from a multivariate logistic model with a low-rank coefficient matrix $\boldsymbol{B}^*$ and a bounded predictor design. To mimic real-world data contamination, 10% of the training samples are corrupted (see Appendix C.1 for details). The bounded design provides a public per-record gradient bound, so DP-RoLL requires no additional clipping threshold.

We compare DP-RoLL with standard private learning methods, including DP-SGD (Abadi et al., 2016), DP-ERM (Wang & Xu, 2019), DP Output Perturbation, and DP Objective Perturbation (Chaudhuri et al., 2011). We also consider private adaptations of two non-private robust or low-rank methods, namely DP Low-Rank GCE (Zhang & Sabuncu, 2018) and DP Trace-Norm BCE (Pontil & Maurer, 2013), together with DP Riemannian low-rank BCE (Han et al., 2024). Detailed implementations are provided in Appendix C.2.

To ensure a fair comparison, method-specific hyperparameters are selected from prespecified grids using the public-calibration procedure described in the Appendix C.12, with independently generated training and validation data following the same data-generating mechanism as the private experiment. For iterative methods, the number of iterations T is selected before tuning the remaining hyperparameters. The selected configuration is then fixed over 20 independent replications. Consistent with previous studies on multi-label learning (Chen et al., 2024; Wu et al., 2025), we measure estimation error by $\|\hat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F^2$ and predictive performance by Hamming loss (HL), average precision (AP), one-error (OE), ranking loss (RL), and coverage (Cov.).

For the privacy–utility comparison, we set $\delta = 10^{-5}$ and vary $\epsilon \in \{0.5, 1, 2, 4\}$ while keeping the contamination level and all other settings fixed. For the robustness comparison, we fix $\epsilon = 2$ and vary the outlier proportion over $\{5\%, 7.5\%, 10\%\}$. Figure 2 compares the methods in terms of estimation error, Hamming loss, and average precision. As shown in panel (a), utility improves as ϵ increases, and DP-RoLL achieves the best performance on all three metrics across privacy budgets. Panel (b) shows that this advantage persists as the outlier proportion increases. Appendix C provides complete results for all evaluation metrics, together with additional analyses of the true-rank scaling, sensitivity to the working rank and $\lambda_{\boldsymbol{C}}$.

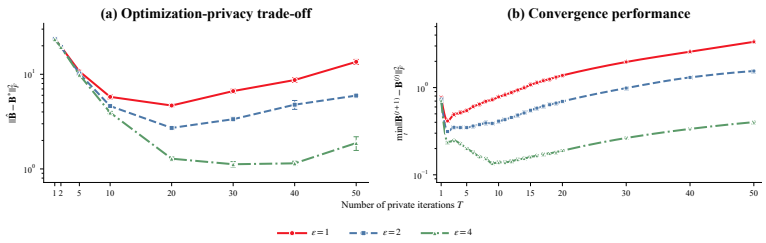

Figure 3: Optimization–privacy trade-off (panel (a)) and convergence performance (panel (b)) of DP-RoLL across different privacy budgets. Points and error bars denote means and standard errors over 20 independent replications.

5.2 Optimization–Privacy Trade-off and Convergence Analysis

For each $\epsilon \in \{1, 2, 4\}$, we vary the number of private iterations over $T \in \{1, 2, 5, 10, 20, 30, 40, 50\}$ while keeping the total privacy budget and all other classification settings fixed. For the convergence analysis, we evaluate the finer grid $T \in \{1, 2, \ldots, 19, 20, 30, 40, 50\}$ and report $\min_{0 \le t \le T-1} \|\boldsymbol{B}^{(t+1)} - \boldsymbol{B}^{(t)}\|_F^2$, the convergence measure appearing in Theorem 5.

In panel (a) of Figure 3, the estimation error initially decreases as additional iterations improve optimization, but eventually increases because privacy composition requires stronger per-iteration perturbation. This non-monotone pattern reflects the optimization–privacy trade-off, with moderate iteration counts yielding the lowest estimation error. The convergence curves in panel (b) show that the minimum squared iterate change initially decreases but subsequently increases, as the stronger per-iteration perturbation associated with larger T prevents further convergence.

5.3 Performance on Real-World Data

In this section, we evaluate the eight methods considered in the synthetic experiments under the same privacy budgets on 11 datasets, comprising three noisy datasets and eight clean benchmarks. Dataset characteristics are provided in the Appendix C.4. Following common cardinality-preserving label noise protocols (Song et al., 2024; Requena et al., 2026), for each clean dataset, we uniformly sample without replacement 10% of the training samples from those containing both positive and negative labels. For each selected sample, we randomly flip half of its positive labels to negative and the same number of negative labels to positive, thereby preserving its label cardinality. No corruption is applied to the test set. For noisy datasets, the provided candidate or partial labels are used for training and the ground-truth labels for testing, without additional corruption.

We use the same five evaluation metrics as in the synthetic experiments and report mean results over 20 independent 80/20 train–test splits shared across methods. To ensure a fair comparison, method-specific hyperparameters, including the number of iterations T for iterative methods, are selected separately for each dataset by minimizing a held-out validation criterion over two prespecified pilot splits, following the public-calibration principle described in Appendix C.12. Treating these pilot splits as public separates hyperparameter selection from private training, so no additional privacy budget is incurred.

To provide a comprehensive view of the performance across different evaluation metrics, we visualize the results on four of the eleven benchmark datasets using radar charts in Figure 4. Each metric is normalized to $[0, 1]$, with 1 and 0 denoting the best and worst results, respectively. As observed, DP-RoLL consistently encompasses the competitive baselines, forming the largest enclosed area across all datasets. Similar behaviors are also observed across the remaining datasets. Appendix C.5 provides additional results and average method ranks across all eleven datasets and five evaluation metrics for each privacy budget, further demonstrating the overall advantage of DP-RoLL.

5.4 Empirical Privacy Evaluation

We assess privacy leakage through membership inference (Shokri et al., 2017) using the loss-based attack of Yeom et al. (2018). For each candidate record, the attack score is the negative binary

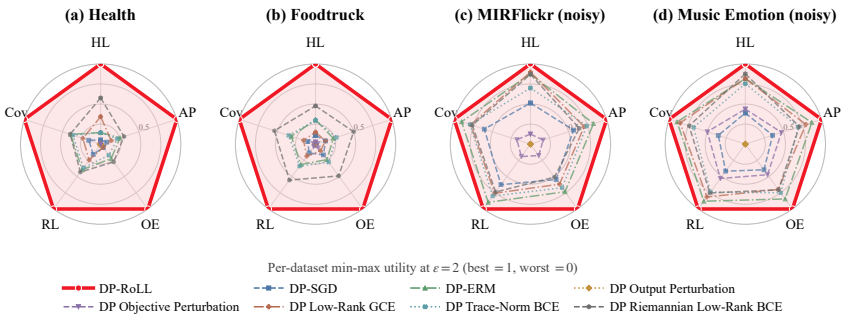

Figure 4: Comparative performance on four real-world multi-label datasets using radar charts at $\epsilon = 2$. The axes are normalized so that larger values indicate better performance, and a larger enclosed area represents better overall performance. DP-RoLL consistently encloses the largest area among the eight methods.

Table 1: Ablation results on Foodtruck at $\epsilon = 2$. Results are reported as mean $\pm$ standard error over 20 independent replications. Bold values denote the best results.

Method	HL↓	AP↑	OE↓	RL↓	Cov.↓
DP-RoLL	$\mathbf{0.2001 \pm 0.0035}$	$\mathbf{0.6251 \pm 0.0109}$	$\mathbf{0.4018 \pm 0.0188}$	$\mathbf{0.2520 \pm 0.0088}$	$\mathbf{0.4066 \pm 0.0087}$
w/o $\boldsymbol{C}$	0.2256 ± 0.0040	0.5914 ± 0.0098	0.4622 ± 0.0165	0.2704 ± 0.0078	0.4190 ± 0.0084
w/o Low Rank	0.2602 ± 0.0044	0.5530 ± 0.0085	0.5299 ± 0.0127	0.2889 ± 0.0072	0.4325 ± 0.0080

cross-entropy loss averaged over labels, with larger values indicating greater membership likelihood. The attacker uses only the candidate record and the released $\boldsymbol{B}$ and intercept. We conduct the attack on the Image benchmark, treating uncorrupted training records as members and test records as non-members. Restricting the member set to uncorrupted records avoids confounding membership leakage with the artificial label-corruption mechanism.

The resulting AUCs are 0.5073 ± 0.0044, 0.5106 ± 0.0046, 0.5151 ± 0.0045, and 0.5223 ± 0.0047 at $\epsilon = 0.5, 1, 2$, and 4, respectively, reported as mean $\pm$ standard error over 20 independent replications. These values remain close to the random-guessing level of 0.5 across all privacy budgets, indicating limited empirical membership leakage under this attack. We additionally conduct a stronger first-gradient inversion stress test (Geiping et al., 2020) that exposes information beyond DP-RoLL's formal final-model release. No successful joint reconstruction is observed across the evaluated privacy budgets, with the threat model and complete results reported in Appendix C.7.

5.5 Ablation Analysis

At $\epsilon = 2$, we compare DP-RoLL with two structural ablations obtained by removing either $\boldsymbol{C}$ or the low-rank constraint. Table 1 reports the results on Foodtruck, where DP-RoLL significantly outperforms both ablations on all five evaluation metrics. Additional results on Foodtruck at $\epsilon = 0.5$, and on Education and Health at $\epsilon \in \{0.5, 2\}$, are reported in the Appendix C.6. DP-RoLL is also stable to the choices of hyperparameters. Detailed sensitivity results are provided in Appendix C.8.

6 Conclusion

In this paper, we proposed DP-RoLL for differentially private robust low-rank learning with multivariate responses. Theoretically, we established differential privacy, a non-asymptotic error bound, and convergence guarantees. Empirical results demonstrated the favorable estimation and predictive performance of DP-RoLL across different privacy budgets and contamination levels. A primary limitation of our proposed framework is its restriction to linear systematic components. Extending DP-RoLL to structured nonlinear models, including neural networks, is a promising direction for future research.

AI USE STATEMENT

Generative AI tools were used only for language polishing and to improve the clarity of author-written text. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content.

ETHICS STATEMENT

This work is methodological and uses only synthetic and existing datasets. It involves no human-subject data collection, and we are not aware of substantial ethical concerns.

REPRODUCIBILITY STATEMENT

The proposed formulation and optimization procedure are described in Section 3. The experimental setup and protocols, including synthetic and real-world evaluations, baselines, evaluation metrics, privacy budgets, and calibration procedures, are reported in Section 5, with complete data-generation, implementation, and additional experimental details provided in Appendix C and D. The assumptions and theoretical guarantees are stated in Section 4, with complete proofs provided in Appendix B.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Theodore Wilbur Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, 22(3):327–351, 1951.

Raman Arora, Vladimir Braverman, and Jalaj Upadhyay. Differentially private robust low-rank approximation. *Advances in neural information processing systems*, 31, 2018.

Hilal Asi, Jonathan Ullman, and Lydia Zakynthinou. From robustness to privacy and back. In *International Conference on Machine Learning*, pp. 1121–1146. PMLR, 2023.

Ana M Bianco and Victor J Yohai. Robust estimation in the logistic regression model. In *Robust Statistics, Data Analysis, and Computer Intensive Methods: In Honor of Peter Huber's 60th Birthday*, pp. 17–34. Springer, 1996.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of cryptography conference*, pp. 635–658. Springer, 2016.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11:1–11:37, 2011.

Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(29):1069–1109, 2011.

Kun Chen. Model diagnostics in reduced-rank estimation. *Statistics and its interface*, 9(4):469–484, 2016.

Yu Chen, Yanan Wu, Na Han, Xiaozhao Fang, Bingzhi Chen, and Jie Wen. Partial multi-label learning based on near-far neighborhood label enhancement and nonlinear guidance. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3722–3731, 2024.

Steve Chien, Prateek Jain, Walid Krichene, Steffen Rendle, Shuang Song, Abhradeep Thakurta, and Li Zhang. Private alternating least squares: Practical private matrix completion with tighter rates. In *International Conference on Machine Learning*, pp. 1877–1887. PMLR, 2021.

Youngjin Cho and Seyoung Park. Multivariate response regression with low-rank and generalized sparsity. *Journal of the Korean Statistical Society*, 51(3):847–867, 2022.

Xiaolong Cui, Lei Shi, Wei Zhong, and Changliang Zou. Robust high-dimensional low-rank matrix estimation: Optimal rate and data-adaptive tuning. *Journal of Machine Learning Research*, 24 (350):1–57, 2023.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3–4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Andreas Elsener and Sara van de Geer. Robust low-rank matrix estimation. *The Annals of Statistics*, 46(6B):3481–3509, 2018.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients - how easy is it to break privacy in federated learning? *Advances in neural information processing systems*, 33:16937–16947, 2020.

Quan Geng, Wei Ding, Ruiqi Guo, and Sanjiv Kumar. Tight analysis of privacy and utility tradeoff in approximate differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pp. 89–99. PMLR, 2020.

Xiaojie Guo and Zhouchen Lin. Low-rank matrix recovery via robust outlier estimation. *IEEE Transactions on Image Processing*, 27(11):5316–5327, 2018.

Deepti Gupta, Olumide Kayode, Smriti Bhatt, Maanak Gupta, and Ali Saman Tosun. Hierarchical federated learning based anomaly detection using digital twins for smart healthcare. In *2021 IEEE 7th international conference on collaboration and internet computing (CIC)*, pp. 16–25. IEEE, 2021.

Andi Han, Bamdev Mishra, Pratik Jawanpuria, and Junbin Gao. Differentially private riemannian optimization. *Machine Learning*, 113(3):1133–1161, 2024.

Jun-Yi Hang and Min-Ling Zhang. Partial multi-label learning with probabilistic graphical disambiguation. *Advances in Neural Information Processing Systems*, 36:1339–1351, 2023.

Zhi-Fen He and Ming Yang. Sparse and low-rank representation for multi-label classification. *Applied Intelligence*, 49(5):1708–1723, 2019.

Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. Evaluating gradient inversion attacks and defenses in federated learning. *Advances in neural information processing systems*, 34:7232–7241, 2021.

Peter J Huber. Robust statistics. In *International encyclopedia of statistical science*, pp. 1248–1251. Springer, 2011.

David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58 (1):30–37, 2004.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

Peter Kairouz, Monica Ribero Diaz, Keith Rush, and Abhradeep Thakurta. (nearly) dimension independent private ERM with AdaGrad rates via publicly estimated subspaces. In *Proceedings of the Thirty-Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2717–2746. PMLR, 2021.

Sachiko Kanamori, Taeko Abe, Takuma Ito, Keita Emura, Lihua Wang, Shuntaro Yamamoto, Le Trieu Phong, Kaien Abe, Sangwook Kim, Ryo Nojima, Seiichi Ozawa, and Shiho Moriai. Privacy-preserving federated learning for detecting fraudulent financial transactions in japanese banks. *Journal of Information Processing*, 30:789–795, 2022.

Minwoo Kim and Sungkyu Jung. Robust and differentially private principal component analysis. *Statistical Analysis and Data Mining: An ASA Data Science Journal*, 18(6):e70053, 2025.

Li Liu, Douglas M Hawkins, Sujoy Ghosh, and S Stanley Young. Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences*, 100(23): 13167–13172, 2003.

Weidong Liu, Xiaojun Mao, Xiaofei Zhang, and Xin Zhang. Efficient sparse least absolute deviation regression with differential privacy. *IEEE Transactions on Information Forensics and Security*, 19:2328–2339, 2024.

Xiyang Liu, Weihao Kong, Sham Kakade, and Sewoong Oh. Robust and differentially private mean estimation. *Advances in neural information processing systems*, 34:3887–3901, 2021.

Xiyang Liu, Weihao Kong, Prateek Jain, and Sewoong Oh. Dp-pca: Statistically optimal and differentially private pca. *Advances in neural information processing systems*, 35:29929–29943, 2022.

Xiyang Liu, Prateek Jain, Weihao Kong, Sewoong Oh, and Arun Suggala. Label robust and differentially private linear regression: Computational and statistical efficiency. *Advances in Neural Information Processing Systems*, 36:23019–23033, 2023.

Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(19):559–616, 2015.

Xiaoyan Ma, Lili Wei, and Wanfeng Liang. D4r: Doubly robust reduced rank regression in high dimension. *Journal of Statistical Planning and Inference*, 232:106162, 2024.

Oren Mangoubi, Yikai Wu, Satyen Kale, Abhradeep Thakurta, and Nisheeth K Vishnoi. Private matrix approximation and geometry of unitary orbits. In *Conference on Learning Theory*, pp. 3547–3588. PMLR, 2022.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275. IEEE, 2017.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Seyoung Park, Eun Ryung Lee, and Hongyu Zhao. Low-rank regression models for multiple binary responses and their applications to cancer cell-line encyclopedia data. *Journal of the American Statistical Association*, 119(545):202–216, 2024.

Massimiliano Pontil and Andreas Maurer. Excess risk bounds for multitask learning with trace norm regularization. In *Conference on Learning Theory*, pp. 55–76. PMLR, 2013.

Wenbin Qian, Yanqiang Tu, Jintao Huang, Wenhao Shu, and Yiu-Ming Cheung. Partial multilabel learning using noise-tolerant broad learning system with label enhancement and dimensionality reduction. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):3758–3772, 2025.

Antonio Requena, Alejandro Galan-Cuenca, Antonio Javier Gallego, and Jose J Valero-Mas. Exploring the impact of label-level noise on multi-label k-nearest neighbor classification. *Pattern Analysis and Applications*, 29(3):130, 2026.

Marc Senneret, Yannick Malevergne, Patrice Abry, Gerald Perrin, and Laurent Jaffrès. Covariance versus precision matrix estimation for efficient asset allocation. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):982–993, 2016.

Yiyuan She. An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics & Data Analysis*, 56(10):2976–2990, 2012.

Yiyuan She. Reduced rank vector generalized linear models for feature extraction. *Statistics and Its Interface*, 6(2):197–209, 2013.

Yiyuan She and Kun Chen. Robust reduced-rank regression. *Biometrika*, 104(3):633–647, 2017.

Yiyuan She and Art B Owen. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639, 2011.

Yiyuan She, Zhifeng Wang, and Jiuwu Jin. Analysis of generalized bregman surrogate algorithms for nonsmooth nonconvex statistical learning. *The Annals of Statistics*, 49(6):3434–3459, 2021.

Yiyuan She, Zhifeng Wang, and Jiahui Shen. Gaining outlier resistance with progressive quantiles: Fast algorithms and theoretical studies. *Journal of the American Statistical Association*, 117(539): 1282–1295, 2022.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Toward robustness in multi-label classification: A data augmentation strategy against imbalance and noise. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 21592–21601, 2024.

Kean Ming Tan, Qiang Sun, and Daniela Witten. Sparse reduced rank huber regression in high dimensions. *Journal of the American Statistical Association*, 118(544):2383–2393, 2023.

Mara Thomas, Nuria Mackes, Asad Preuss-Dodhy, Thomas Wieland, and Markus Bundschus. Assessing privacy vulnerabilities in genetic data sets: scoping review. *JMIR Bioinformatics and Biotechnology*, 5(1):e54332, 2024.

N Trendafilov, M Gallo, V Simonacci, and V Todorov. Overview of reduced-rank regression with dense and sparse coefficients, and a new estimation procedure. *Behaviormetrika*, 52(2):593–625, 2025.

Eliad Tsfadia. On differentially private subspace estimation in a distribution-free setting. *Advances in Neural Information Processing Systems*, 37:39923–39962, 2024.

Jalaj Upadhyay. The price of privacy for low-rank factorization. *Advances in Neural Information Processing Systems*, 31, 2018.

Di Wang and Jinhui Xu. Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1182–1189, 2019.

Lingxiao Wang, Boxin Zhao, and Mladen Kolar. Differentially private matrix completion through low-rank matrix factorization. In *International Conference on Artificial Intelligence and Statistics*, pp. 5731–5748. PMLR, 2023.

Yu-Xiang Wang. Revisiting differentially private linear regression: Optimal and adaptive prediction & estimation in unbounded domain, 2018.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1226–1235. PMLR, 2019.

You Wu, Yabo Shi, Yizhang Zou, and Peipei Li. Robust multi-label learning with instance-dependent label noise. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 3458–3467, 2025.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282. IEEE, 2018.

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Large scale private learning via low-rank reparametrization. In *International Conference on Machine Learning*, pp. 12208–12218. PMLR, 2021.

Guoxian Yu, Xia Chen, Carlotta Domeniconi, Jun Wang, Zhao Li, Zili Zhang, and Xindong Wu. Feature-induced partial multi-label learning. In *2018 IEEE international conference on data mining (ICDM)*, pp. 1398–1403. IEEE, 2018.

Myeonghun Yu, Zhao Ren, and Wen-xin Zhou. Gaussian differentially private robust mean estimation and inference. *Bernoulli*, 30(4):3059–3088, 2024.

Ming Yuan, Ali Ekici, Zhaosong Lu, and Renato Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(3):329–346, 2007.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Yuhai Zhao, Yejiang Wang, Zhengkui Wang, Wen Shan, Miaomiao Huang, Meixia Wang, Min Huang, and Xingwei Wang. Towards robust multi-label learning against dirty label noise. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 5581–5589, 2024.

Wei Zhi, Jane Minturn, Eric Rappaport, Garrett Brodeur, and Hongzhe Li. Network-based analysis of multivariate gene expression data. In *Statistical Methods for Microarray Data Analysis: Methods and Protocols*, pp. 121–139. Springer, 2013.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

A Comparison with Representative Related Work

Table A.1: Comparison with representative related work. The columns indicate whether each work applies to multivariate responses (MR), exploits low-rank structure (LR), addresses robustness (Rob.), provides a formal differential privacy guarantee (DP), and establishes statistical (Stat.) or optimization (Opt.) theory under differential privacy.

	Problem setting			Privacy and theory		
Related Work	MR	LR	Rob.	DP	Stat.	Opt.
Pontil & Maurer (2013); Park et al. (2024)	✓	✓	×	×	×	×
She & Chen (2017); Tan et al. (2023)	✓	✓	✓	×	×	×
Zhang & Sabuncu (2018)	✓	×	✓	×	×	×
Abadi et al. (2016)	✓	×	×	✓	×	×
Chaudhuri et al. (2011); Wang (2018)	✓	×	×	✓	✓	×
Wang & Xu (2019)	✓	×	×	✓	✓	✓
Han et al. (2024)	✓	✓	×	✓	×	✓
Arora et al. (2018)	×	✓	✓	✓	×	×
Yu et al. (2024)	×	×	✓	✓	✓	✓
DP-RoLL (ours)	✓	✓	✓	✓	✓	✓

B Technical Details

B.1 Preliminaries

We first recall the definitions of differential privacy and zero-concentrated differential privacy. Two datasets D and D' are adjacent, denoted by $D \simeq D'$, if they differ in one individual record.

Definition B.1 ((ϵ, δ)-Differential Privacy (Dwork et al., 2006)). *A randomized mechanism $\mathcal{M}$ satisfies (ϵ, δ)-differential privacy if, for any adjacent datasets $D \simeq D'$ and any measurable set $\mathcal{S}$, $\Pr\{\mathcal{M}(D) \in \mathcal{S}\} \leq e^{\epsilon} \Pr\{\mathcal{M}(D') \in \mathcal{S}\} + \delta$.*

Definition B.2 (Zero-Concentrated Differential Privacy (Bun & Steinke, 2016)). *A randomized mechanism $\mathcal{M}$ satisfies ξ-zero-concentrated differential privacy (ξ-zCDP) if, for any adjacent datasets $D \simeq D'$ and every $\alpha > 1$, $D_\alpha(\mathcal{M}(D) \| \mathcal{M}(D')) \leq \xi\alpha$, where $D_\alpha(\cdot\|\cdot)$ denotes the Rényi divergence of order α.*

A mechanism satisfying ξ-zCDP also satisfies $(\xi + 2\sqrt{\xi \log(1/\delta)}, \delta)$-differential privacy for any $\delta \in (0, 1)$. Moreover, zCDP composes additively: the adaptive composition of mechanisms satisfying $\xi_1, \ldots, \xi_T$-zCDP satisfies $\sum_{t=1}^{T} \xi_t$-zCDP.

Definition B.3 (Sub-Gaussian random variable/vector). *ξ is called a sub-Gaussian random variable if and only if there exist constants $C, c > 0$ such that $\mathbb{P}\{|\xi| \geq t\} \leq Ce^{-ct^2}, \forall t > 0$. The scale (or ψ_2-norm) of ξ is defined as $\sigma(\xi) = \inf\{\sigma > 0 : \mathbb{E}\exp(\xi^2/\sigma^2) \leq 2\}$. More generally, we say $\boldsymbol{\xi}$ is a sub-Gaussian random vector with scale bounded by σ if all one-dimensional marginals $\|\langle \boldsymbol{\xi}, \boldsymbol{\alpha} \rangle\|_{\psi_2} \leq \sigma \|\boldsymbol{\alpha}\|_2, \forall \boldsymbol{\alpha} \in \mathbb{R}^d$.*

Additional Notations. Let $[\cdot]_+ = \max(\cdot, 0)$. For a matrix $\boldsymbol{A}$, $\mathrm{CS}(\boldsymbol{A})$ denotes its column space. Furthermore, $\mathcal{P}_{\boldsymbol{A}}$ and $\mathcal{P}_{\boldsymbol{A}}^{\perp}$ represent the orthogonal projection matrices onto $\mathrm{CS}(\boldsymbol{A})$ and its orthogonal complement, respectively. Let $[n] = \{1, \ldots, n\}$. Given $\mathcal{I} \subset [n]$ and $\mathcal{J} \subset [p]$, let $\boldsymbol{X}(\mathcal{I}, \mathcal{J})$ denote the submatrix of $\boldsymbol{X}$ obtained by extracting the rows indexed by $\mathcal{I}$ and the columns indexed by $\mathcal{J}$. For convenience, we write $\boldsymbol{X}_{\mathcal{J}} = \boldsymbol{X}(\mathcal{J}, :)$ to denote the submatrix formed by the columns of $\boldsymbol{X}$ indexed by $\mathcal{J}$. Given $a \in \mathbb{R}$, $\lfloor a \rfloor$ denotes ceiling function.

B.2 Proof of Theorem 2

Proof. We analyze the privacy loss conditionally on the previously released iterates. Assume that $\boldsymbol{B}^{(0)}$ and $\boldsymbol{C}^{(0)}$ are fixed and public. For $t = 0, \ldots, T-1$, let $\mathcal{H}_t^B = (\boldsymbol{B}^{(0)}, \boldsymbol{B}^{(1)}, \ldots, \boldsymbol{B}^{(t)})$ denote the transcript available before iteration t. Fix an arbitrary transcript value h_t. Conditional on $\mathcal{H}_t^B = h_t$, the $\boldsymbol{B}$-iterates up to iteration t are fixed. For $t \geq 1$, the internal contamination iterates are then recursively determined from the data and h_t for $s = 0, \ldots, t-1$ by

$$\begin{aligned} \widetilde{\boldsymbol{C}}_{i\cdot}^{(s+1)} &= \boldsymbol{C}_{i\cdot}^{(s)} - \frac{\sqrt{n}}{\rho} \boldsymbol{G}_{i\cdot}^{(s)}, \\ \boldsymbol{C}_{i\cdot}^{(s+1)} &= \widetilde{\boldsymbol{C}}_{i\cdot}^{(s+1)} \mathbb{I}\left\{ \left\| \widetilde{\boldsymbol{C}}_{i\cdot}^{(s+1)} \right\|_2 \geq \frac{\lambda_{\boldsymbol{C}}}{\sqrt{\rho}} \right\}. \end{aligned} \tag{B.1}$$

Let D and D' be adjacent datasets differing only in the k-th observation. We first show that, under the same fixed transcript h_t,

$$\boldsymbol{C}_{i\cdot}^{(s)}(D; h_t) = \boldsymbol{C}_{i\cdot}^{(s)}(D'; h_t), \qquad i \neq k, \quad s = 0, \ldots, t. \tag{B.2}$$

The claim holds for $s = 0$ because $\boldsymbol{C}^{(0)}$ is fixed and public. Assume that it holds for some $s < t$. For any $i \neq k$, the corresponding observations are identical under D and D', that is, $(\boldsymbol{x}_i, \boldsymbol{y}_i) = (\boldsymbol{x}_i', \boldsymbol{y}_i')$. Moreover, conditional on h_t, $\boldsymbol{B}^{(s)}$ is fixed, and the induction hypothesis gives

$$\boldsymbol{C}_{i\cdot}^{(s)}(D; h_t) = \boldsymbol{C}_{i\cdot}^{(s)}(D'; h_t). \tag{B.3}$$

Hence, the corresponding linear predictors are identical:

$$\boldsymbol{x}_i^{\top} \boldsymbol{B}^{(s)} + \sqrt{n}\, \boldsymbol{C}_{i\cdot}^{(s)}(D; h_t) = \boldsymbol{x}_i'^{\top} \boldsymbol{B}^{(s)} + \sqrt{n}\, \boldsymbol{C}_{i\cdot}^{(s)}(D'; h_t). \tag{B.4}$$

By the sample-wise separability of the loss, it follows that $\boldsymbol{G}_{i\cdot}^{(s)}(D; h_t) = \boldsymbol{G}_{i\cdot}^{(s)}(D'; h_t)$. Applying the same row-wise update equation B.1 then gives $\boldsymbol{C}_{i\cdot}^{(s+1)}(D; h_t) = \boldsymbol{C}_{i\cdot}^{(s+1)}(D'; h_t)$. This completes the induction and proves equation B.2.

For the fixed transcript h_t, define the clipped query at iteration t as

$$q_t(D; h_t) = \sum_{i=1}^{n} \widetilde{\boldsymbol{U}}_i^{(t)}(D; h_t). \tag{B.5}$$

By equation B.2, sample-wise separability, and the equality of the corresponding observations, we have $\boldsymbol{U}_i^{(t)}(D; h_t) = \boldsymbol{U}_i^{(t)}(D'; h_t)$ for all $i \neq k$. Since clipping is deterministic and C_B is public, it follows that

$$\widetilde{\boldsymbol{U}}_i^{(t)}(D; h_t) = \widetilde{\boldsymbol{U}}_i^{(t)}(D'; h_t), \qquad i \neq k. \tag{B.6}$$

Consequently,

$$q_t(D; h_t) - q_t(D'; h_t) = \widetilde{\boldsymbol{U}}_k^{(t)}(D; h_t) - \widetilde{\boldsymbol{U}}_k^{(t)}(D'; h_t). \tag{B.7}$$

Since clipping ensures that $\|\widetilde{\boldsymbol{U}}_i^{(t)}(D; h_t)\|_F \leq C_B$ for every dataset, transcript, sample, and iteration, we have

$$\begin{aligned} \|q_t(D; h_t) - q_t(D'; h_t)\|_F &\leq \|\widetilde{\boldsymbol{U}}_k^{(t)}(D; h_t)\|_F + \|\widetilde{\boldsymbol{U}}_k^{(t)}(D'; h_t)\|_F \\ &\leq 2C_B. \end{aligned} \tag{B.8}$$

Thus, we take $\Delta_B = 2C_B$ as a uniform upper bound on the conditional sensitivity and calibrate the Gaussian noise accordingly.

At iteration t, the algorithm forms

$$\hat{q}_t = q_t(D; h_t) + \boldsymbol{N}_{\boldsymbol{B}}^{(t)}, \tag{B.9}$$

where $\boldsymbol{N}_{\boldsymbol{B}}^{(t)}$ has i.i.d. $\mathcal{N}(0, \sigma_{\mathrm{DP}}^2)$ entries and is independent of the data, the preceding transcript, and the noise added at other iterations. Since $\|\operatorname{vec}(\boldsymbol{A})\|_2 = \|\boldsymbol{A}\|_F$, the vectorized clipped query has Euclidean sensitivity at most Δ_B. By the Gaussian mechanism for zCDP (Bun & Steinke, 2016, Proposition 1.6), $\hat{q}_t$ conditionally satisfies ξ_0-zCDP, where

$$\xi_0 = \frac{\Delta_B^2}{2\sigma_{\mathrm{DP}}^2}. \tag{B.10}$$

Conditional on the fixed transcript h_t, $\boldsymbol{B}^{(t)}$ is fixed, and $\boldsymbol{B}^{(t+1)}$ is obtained by deterministic post-processing of $\hat{q}_t$:

$$\boldsymbol{B}^{(t+1)} = \mathcal{P}_{\vartheta r}\left(\boldsymbol{B}^{(t)} - \frac{1}{\rho}\hat{q}_t\right). \tag{B.11}$$

Therefore, releasing $\boldsymbol{B}^{(t+1)}$ incurs no additional privacy loss by the post-processing property of zCDP (Bun & Steinke, 2016, Lemma 1.8).

Because the conditional mechanism is ξ_0-zCDP for every fixed transcript, adaptive composition over the T iterations gives

$$\xi_{\mathrm{tot}} = T\xi_0 = \frac{T\Delta_B^2}{2\sigma_{\mathrm{DP}}^2} \tag{B.12}$$

for the entire transcript $(\boldsymbol{B}^{(1)}, \ldots, \boldsymbol{B}^{(T)})$ (Bun & Steinke, 2016; Mironov, 2017).

A ξ_{tot}-zCDP mechanism is (ϵ, δ)-differentially private for any $\delta \in (0, 1)$ whenever $\epsilon = \xi_{\mathrm{tot}} + 2\sqrt{\xi_{\mathrm{tot}} \log(1/\delta)}$ (Bun & Steinke, 2016, Proposition 1.3). Solving for ξ_{tot} gives

$$\xi_{\mathrm{tot}} = \left(\sqrt{\epsilon + \log(1/\delta)} - \sqrt{\log(1/\delta)}\right)^2. \tag{B.13}$$

Then, we have

$$\sigma_{\mathrm{DP}}^2 = \frac{T\Delta_B^2}{2\left(\sqrt{\epsilon + \log(1/\delta)} - \sqrt{\log(1/\delta)}\right)^2}. \tag{B.14}$$

Therefore, the released sequence $(\boldsymbol{B}^{(1)}, \ldots, \boldsymbol{B}^{(T)})$ satisfies (ϵ, δ)-differential privacy. □

B.3 Proof of Theorem 3

By the optimality condition and Lemma B.7, we have

$$\begin{aligned}
&\frac{\rho}{2}\|\boldsymbol{B}^* - \boldsymbol{B}^{(t)} + (\boldsymbol{X}^\top\nabla l(\boldsymbol{X}\boldsymbol{B}^{(t)} + \sqrt{n}\boldsymbol{C}^{(t)}) + \boldsymbol{N}_{\boldsymbol{B}}^{(t)})/\rho\|_F^2 \\
&- \frac{\rho}{2}\|\boldsymbol{B}^{(t+1)} - \boldsymbol{B}^{(t)} + (\boldsymbol{X}^\top\nabla l(\boldsymbol{X}\boldsymbol{B}^{(t)} + \sqrt{n}\boldsymbol{C}^{(t)}) + \boldsymbol{N}_{\boldsymbol{B}}^{(t)})/\rho\|_F^2 \\
&+ \frac{\rho}{2}\|\boldsymbol{C}^* - \boldsymbol{C}^{(t)} + \sqrt{n}\nabla l(\boldsymbol{X}\boldsymbol{B}^{(t)} + \sqrt{n}\boldsymbol{C}^{(t)})/\rho\|_F^2 \\
&- \frac{\rho}{2}\|\boldsymbol{C}^{(t+1)} - \boldsymbol{C}^{(t)} + \sqrt{n}\nabla l(\boldsymbol{X}\boldsymbol{B}^{(t)} + \sqrt{n}\boldsymbol{C}^{(t)})/\rho\|_F^2 \\
&+ \frac{\lambda_{\boldsymbol{C}}^2}{2}(\|\boldsymbol{C}^*\|_{2,0} - \|\boldsymbol{C}^{(t+1)}\|_{2,0}) \\
\geq&\frac{\rho}{2}(1 - 1/\sqrt{\vartheta})\|\boldsymbol{B}^{(t+1)} - \boldsymbol{B}^*\|_F^2.
\end{aligned} \tag{B.15}$$

Then, we have

$$\begin{aligned}
&\boldsymbol{\Delta}_\psi(\bar{\boldsymbol{B}}^*, \bar{\boldsymbol{B}}^{(t+1)}) + 2\bar{\boldsymbol{\Delta}}_l(\boldsymbol{X}\boldsymbol{B}^{(t+1)} + \sqrt{n}\boldsymbol{C}^{(t+1)}, \boldsymbol{X}\boldsymbol{B}^* + \sqrt{n}\boldsymbol{C}^*) + \boldsymbol{\Delta}_\psi(\bar{\boldsymbol{B}}^{(t+1)}, \bar{\boldsymbol{B}}^{(t)}) \\
&+ \frac{\lambda_{\boldsymbol{C}}^2}{2}(\|\boldsymbol{C}^{(t+1)}\|_{2,0} - \|\boldsymbol{C}^*\|_{2,0}) \\
\leq&\boldsymbol{\Delta}_\psi(\bar{\boldsymbol{B}}^*, \bar{\boldsymbol{B}}^{(t)}) + \frac{\rho}{2\sqrt{\vartheta}}\|\boldsymbol{B}^{(t+1)} - \boldsymbol{B}^*\|_F^2 + \frac{\rho}{2}\|\boldsymbol{C}^{(t+1)} - \boldsymbol{C}^*\|_F^2 \\
&+ \langle\boldsymbol{\mathcal{E}}, \boldsymbol{X}\boldsymbol{B}^{(t+1)} + \sqrt{n}\boldsymbol{C}^{(t+1)} - \boldsymbol{X}\boldsymbol{B}^* - \sqrt{n}\boldsymbol{C}^*\rangle \\
&+ \langle\boldsymbol{N}_{\boldsymbol{B}}^{(t)}, \boldsymbol{B}^* - \boldsymbol{B}^{(t+1)}\rangle.
\end{aligned} \tag{B.16}$$

Next, by Lemma B.4, with probability at least $1 - c'n^{-c_0}$, we have

$$\begin{aligned}
&\langle\boldsymbol{\mathcal{E}}, \boldsymbol{X}\boldsymbol{B}^{(t+1)} + \sqrt{n}\boldsymbol{C}^{(t+1)} - \boldsymbol{X}\boldsymbol{B}^* - \sqrt{n}\boldsymbol{C}^*\rangle \\
\leq&\frac{1}{2a_{\text{stat}}}\|\boldsymbol{X}\boldsymbol{B}^{(t+1)} + \sqrt{n}\boldsymbol{C}^{(t+1)} - \boldsymbol{X}\boldsymbol{B}^* - \sqrt{n}\boldsymbol{C}^*\|_F^2 \\
&+ \frac{1}{2b_{\text{stat}}}P_{2,H}\Big(\sqrt{n}(\boldsymbol{C}^{(t+1)} - \boldsymbol{C}^*); \lambda_{\boldsymbol{C}}\Big) \\
&+ \frac{a_{\text{stat}}A_0}{2}\sigma^2(q+m)(1+\vartheta)r + \frac{c_0}{2c}a_{\text{stat}}\sigma^2\log n,
\end{aligned} \tag{B.17}$$

where $a_{\text{stat}} \geq 2b_{\text{stat}} > 0$ and $A_0, c_0, c, c' > 0$ are universal constants.

By applying Lemma B.4 with $\boldsymbol{X} = \boldsymbol{I}_p$ and $\boldsymbol{C} = \boldsymbol{0}$, followed by a union bound over T iterations, with probability at least $1 - c'n^{-c_0}$, we have

$$\begin{aligned}
&\max_{0\leq t\leq T-1}\left\{\langle\boldsymbol{N}_{\boldsymbol{B}}^{(t)}, \boldsymbol{B}^* - \boldsymbol{B}^{(t+1)}\rangle - \frac{1}{2a_{\text{DP}}}\|\boldsymbol{B}^* - \boldsymbol{B}^{(t+1)}\|_F^2\right\} \\
\leq&\frac{a_{\text{DP}}A_0}{2}\sigma_{\text{DP}}^2(p+m)(1+\vartheta)r + \frac{1}{2c}a_{\text{DP}}\sigma_{\text{DP}}^2\{\log T + c_0\log n\},
\end{aligned} \tag{B.18}$$

where $a_{\text{DP}}, c_0, c, c' > 0$ are universal constants.

Let

$$\begin{aligned}
e(T) :=& \frac{a_{\text{stat}}A_0}{2}\sigma^2(1+\vartheta)r(q+m) + Ka_{\text{stat}}b_{\text{stat}}A_1\sigma^2 J(m+\log n) \\
&+ \frac{a_{\text{DP}}A_0}{2}\sigma_{\text{DP}}^2(1+\vartheta)r(p+m) \\
&+ \frac{c_0}{2c}a_{\text{stat}}\sigma^2\log n + \frac{1}{2c}a_{\text{DP}}\sigma_{\text{DP}}^2(\log T + c_0\log n).
\end{aligned} \tag{B.19}$$

We define the structural complexity for the statistical and privacy errors respectively as

$$\begin{aligned} P_{\mathrm{stat}}(r, J) &:= (1+\vartheta)r(q+m) + J(m+\log n), \\ P_{\mathrm{DP}}(r) &:= (1+\vartheta)r(p+m) + \log n. \end{aligned} \tag{B.20}$$

Then, with probability at least $1 - \mathcal{O}(n^{-c_0})$, we obtain

$$e(T) \le C_1\sigma^2 P_{\mathrm{stat}}(r, J) + C_2\sigma_{\mathrm{DP}}^2 P_{\mathrm{DP}}(r) + C_3\sigma_{\mathrm{DP}}^2 \log T, \tag{B.21}$$

where

$$C_1 := a_{\mathrm{stat}} \max\left\{\frac{A_0}{2}, Kb_{\mathrm{stat}}A_1 + \frac{c_0}{2c}\right\}, \quad C_2 := a_{\mathrm{DP}} \max\left\{\frac{A_0}{2}, \frac{c_0}{2c}\right\}, \quad C_3 := \frac{a_{\mathrm{DP}}}{2c}$$

are universal absolute constants.

Applying the regularity condition with $K_1 = 1/a_{\mathrm{stat}}$, $K_2 = 1/a_{\mathrm{DP}}$, and $K' = 1/(2b_{\mathrm{stat}})$ to bound the surrogate descent yields the one-step contraction:

$$(1+\tau)\boldsymbol{\Delta}_\psi(\bar{\boldsymbol{B}}^*, \bar{\boldsymbol{B}}^{(t+1)}) + \boldsymbol{\Delta}_\psi(\bar{\boldsymbol{B}}^{(t+1)}, \bar{\boldsymbol{B}}^{(t)}) \le \boldsymbol{\Delta}_\psi(\bar{\boldsymbol{B}}^*, \bar{\boldsymbol{B}}^{(t)}) + e(T). \tag{B.22}$$

By iteratively applying equation B.22 from $t = 0$ to $T-1$, we obtain the final geometric convergence rate:

$$\begin{aligned} \boldsymbol{\Delta}_\psi(\bar{\boldsymbol{B}}^*, \bar{\boldsymbol{B}}^{(T)}) \le{}& (1+\tau)^{-T}\boldsymbol{\Delta}_\psi(\bar{\boldsymbol{B}}^*, \bar{\boldsymbol{B}}^{(0)}) \\ &+ \frac{1-(1+\tau)^{-T}}{\tau}\left\{C_1\sigma^2 P_{\mathrm{stat}}(r, J) + C_2\sigma_{\mathrm{DP}}^2 P_{\mathrm{DP}}(r) + C_3\sigma_{\mathrm{DP}}^2 \log T\right\} \\ &- \frac{1-(1+\tau)^{-T}}{\tau} \min_{1\le t\le T} \boldsymbol{\Delta}_\psi(\bar{\boldsymbol{B}}^{(t)}, \bar{\boldsymbol{B}}^{(t-1)}). \end{aligned} \tag{B.23}$$

With $C = \max(C_1, C_2, C_3)$, the proof is complete.

B.4 Discussion of the Regularity Condition in Theorem 3

The regularity condition in Theorem 3 consists of two parts: a curvature part and a penalty-tolerance part. A sufficient condition of the regularity condition in Theorem 3 is that

$$\frac{\lambda_{\boldsymbol{C}}^2}{2}(\|\boldsymbol{C}\|_{2,0} - \|\boldsymbol{C}^*\|_{2,0}) + K\lambda_{\boldsymbol{C}}^2 J \ge K' P_{2,H}(\sqrt{n}(\boldsymbol{C} - \boldsymbol{C}^*); \lambda_{\boldsymbol{C}}) \tag{B.24}$$

and

$$\begin{aligned} &(2\bar{\boldsymbol{\Delta}}_l - K_1\mathbf{D}_2)\big(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*\big) + \tau\boldsymbol{\Delta}_l\big(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}\big) \\ &\qquad \ge \left(K_2 + \frac{\rho}{\sqrt{\vartheta}} + \tau\rho\right)\mathbf{D}_2(\boldsymbol{B}, \boldsymbol{B}^*) + (1+\tau)\rho\,\mathbf{D}_2(\boldsymbol{C}, \boldsymbol{C}^*). \end{aligned} \tag{B.25}$$

for all $\mathrm{rank}(\boldsymbol{B}) \le \vartheta r$ and $\|\boldsymbol{C}\|_{2,0} \le C_s J$.

To verify the condition equation B.24, we utilize the sub-additivity of $P_{2,H}$ and $P_{2,H}(\boldsymbol{A}; \lambda_{\boldsymbol{C}}) \le \frac{\lambda_{\boldsymbol{C}}^2}{2}\|\boldsymbol{A}\|_{2,0}$. These properties yield

$$\begin{aligned} &\frac{\lambda_{\boldsymbol{C}}^2}{2}\|\boldsymbol{C}\|_{2,0} - \frac{\lambda_{\boldsymbol{C}}^2}{2}\|\boldsymbol{C}^*\|_{2,0} + K\lambda_{\boldsymbol{C}}^2 J - K' P_{2,H}(\sqrt{n}(\boldsymbol{C} - \boldsymbol{C}^*)) \\ &\ge \frac{\lambda_{\boldsymbol{C}}^2}{2}(1-K')\|\boldsymbol{C}\|_{2,0} - \frac{\lambda_{\boldsymbol{C}}^2}{2}(1+K')\|\boldsymbol{C}^*\|_{2,0} + K\lambda_{\boldsymbol{C}}^2 J \\ &\ge \lambda_{\boldsymbol{C}}^2 J\left(K - \frac{1+K'}{2}\right) \ge 0. \end{aligned} \tag{B.26}$$

Thus, choosing $K \ge \frac{2b_{\mathrm{stat}}+1}{4b_{\mathrm{stat}}}$ ensures equation B.24.

Now ,we discuss the condition equation B.25. To give readers more intuition, let $l(\bar{\boldsymbol{B}}) = \|\bar{\boldsymbol{X}}\bar{\boldsymbol{B}} - \boldsymbol{Y}\|_F^2/2$. In this case, the generalized Bregman divergences $\bar{\boldsymbol{\Delta}}_l$ reduce to the Frobenius distance $\mathbf{D}_2$.

Let $\Delta_{\boldsymbol{B}} = \boldsymbol{B} - \boldsymbol{B}^*$ and $\Delta_{\boldsymbol{C}} = \boldsymbol{C} - \boldsymbol{C}^*$. Then the regularity condition in Theorem 3 becomes a joint restricted lower isometry, or restricted eigenvalue, condition (Bickel et al., 2009; Negahban et al., 2012) for the augmented design $\bar{\boldsymbol{X}} = [\boldsymbol{X}\ \sqrt{n}\boldsymbol{I}]$. Specifically, it can be written as

$$c_1\|\boldsymbol{X}\Delta_{\boldsymbol{B}} + \sqrt{n}\Delta_{\boldsymbol{C}}\|_F^2 \geq c_3\|\Delta_{\boldsymbol{B}}\|_F^2 + c_2\|\Delta_{\boldsymbol{C}}\|_F^2, \tag{B.27}$$

for all restricted perturbation directions satisfying $\mathrm{rank}(\Delta_{\boldsymbol{B}}) \leq (1+\vartheta)r, \|\Delta_{\boldsymbol{C}}\|_{2,0} \leq (1+C_s)J$, where the constants are defined as $c_1 = \frac{2-K_1+\tau}{2}, c_2 = \frac{(1+\tau)\rho}{2}, c_3 = \frac{1}{2}\left(K_2 + \frac{\rho}{\sqrt{\vartheta}} + \tau\rho\right)$, provided that $c_1 > 0$. Geometrically, equation B.27 ensures that the joint prediction discrepancy controls the errors in both $\boldsymbol{B}$ and $\boldsymbol{C}$, thereby precluding pathological cancellation between $\boldsymbol{X}\Delta_{\boldsymbol{B}}$ and $\sqrt{n}\Delta_{\boldsymbol{C}}$.

A sufficient condition for equation B.27 is that $\boldsymbol{X}$ has restricted curvature over the relevant low-rank directions (Negahban et al., 2012)

$$\|\boldsymbol{X}\Delta_{\boldsymbol{B}}\|_F^2 \geq \kappa_B n\|\Delta_{\boldsymbol{B}}\|_F^2 \tag{B.28}$$

for some $\kappa_B > 0$, and that the low-rank and row-sparse structured components are sufficiently incoherent (Candès et al., 2011; Chandrasekaran et al., 2011)

$$\left|\left\langle \boldsymbol{X}\Delta_{\boldsymbol{B}}, \sqrt{n}\Delta_{\boldsymbol{C}}\right\rangle\right| \leq \mu\|\boldsymbol{X}\Delta_{\boldsymbol{B}}\|_F\|\sqrt{n}\Delta_{\boldsymbol{C}}\|_F \tag{B.29}$$

for some $\mu \in [0,1)$. These conditions imply

$$\begin{aligned}\|\boldsymbol{X}\Delta_{\boldsymbol{B}} + \sqrt{n}\Delta_{\boldsymbol{C}}\|_F^2 &\geq (1-\mu)\left(\|\boldsymbol{X}\Delta_{\boldsymbol{B}}\|_F^2 + n\|\Delta_{\boldsymbol{C}}\|_F^2\right)\\ &\geq (1-\mu)n\left(\kappa_B\|\Delta_{\boldsymbol{B}}\|_F^2 + \|\Delta_{\boldsymbol{C}}\|_F^2\right).\end{aligned} \tag{B.30}$$

Hence, equation B.27 holds provided that $c_1(1-\mu)n\kappa_B \geq c_3, c_1(1-\mu)n \geq c_2$. Thus, the joint restricted lower-isometry condition follows from restricted curvature of the design and limited alignment between the two structured components.

B.5 Proof of Theorem 5

Proof. Since $\bar{\boldsymbol{B}}^{(t+1)}$ minimizes the surrogate function $g(\bar{\boldsymbol{B}}; \bar{\boldsymbol{B}}^{(t)})$ over the feasible set $\mathcal{F} = \{[\boldsymbol{B}^\top, \boldsymbol{C}^\top]^\top : \mathrm{rank}(\boldsymbol{B}) \leq \vartheta r\}$, and the previous iterate $\bar{\boldsymbol{B}}^{(t)} \in \mathcal{F}$, the optimality condition $g(\bar{\boldsymbol{B}}^{(t+1)}; \bar{\boldsymbol{B}}^{(t)}) \leq g(\bar{\boldsymbol{B}}^{(t)}; \bar{\boldsymbol{B}}^{(t)})$ implies

$$\langle \bar{\boldsymbol{X}}^\top\nabla l(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^{(t)}) + \boldsymbol{N}^{(t)}, \bar{\boldsymbol{B}}^{(t+1)} - \bar{\boldsymbol{B}}^{(t)}\rangle + \frac{\rho}{2}\|\bar{\boldsymbol{B}}^{(t+1)} - \bar{\boldsymbol{B}}^{(t)}\|_F^2 + \frac{\lambda_{\boldsymbol{C}}^2}{2}\|\boldsymbol{C}^{(t+1)}\|_{2,0} \leq \frac{\lambda_{\boldsymbol{C}}^2}{2}\|\boldsymbol{C}^{(t)}\|_{2,0}.$$

Rearranging this inequality, we have

$$\begin{aligned}\left\langle \bar{\boldsymbol{X}}^\top\nabla l\left(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^{(t)}\right), \bar{\boldsymbol{B}}^{(t+1)} - \bar{\boldsymbol{B}}^{(t)}\right\rangle \leq{}& \frac{\lambda_{\boldsymbol{C}}^2}{2}\|\boldsymbol{C}^{(t)}\|_{2,0} - \frac{\lambda_{\boldsymbol{C}}^2}{2}\|\boldsymbol{C}^{(t+1)}\|_{2,0}\\ &- \frac{\rho}{2}\|\bar{\boldsymbol{B}}^{(t+1)} - \bar{\boldsymbol{B}}^{(t)}\|_F^2 - \left\langle \boldsymbol{N}^{(t)}, \bar{\boldsymbol{B}}^{(t+1)} - \bar{\boldsymbol{B}}^{(t)}\right\rangle.\end{aligned} \tag{B.31}$$

On the other hand, the $\mathcal{L}$-smoothness of $l(\cdot; \boldsymbol{Y})$ and the definition of $M_{\bar{\boldsymbol{X}}}$ yield

$$l(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^{(t+1)}) \leq l(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^{(t)}) + \langle \bar{\boldsymbol{X}}^\top\nabla l(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^{(t)}), \bar{\boldsymbol{B}}^{(t+1)} - \bar{\boldsymbol{B}}^{(t)}\rangle + \frac{\mathcal{L}M_{\bar{\boldsymbol{X}}}}{2}\|\bar{\boldsymbol{B}}^{(t+1)} - \bar{\boldsymbol{B}}^{(t)}\|_F^2.$$

Combining the above two inequalities and add $\frac{\lambda_{\boldsymbol{C}}^2}{2}\|\boldsymbol{C}^{(t+1)}\|_{2,0}$ to both sides, we obtain the following inequality

$$F(\bar{\boldsymbol{B}}^{(t+1)}) \leq F(\bar{\boldsymbol{B}}^{(t)}) - \frac{\rho - \mathcal{L}M_{\bar{\boldsymbol{X}}}}{2}\|\bar{\boldsymbol{B}}^{(t+1)} - \bar{\boldsymbol{B}}^{(t)}\|_F^2 - \langle \boldsymbol{N}^{(t)}, \bar{\boldsymbol{B}}^{(t+1)} - \bar{\boldsymbol{B}}^{(t)}\rangle. \tag{B.32}$$

Since the DP noise is only injected into $\boldsymbol{B}$, the noise inner product precisely reduces to $-\langle \boldsymbol{N}_{\boldsymbol{B}}^{(t)}, \boldsymbol{B}^{(t+1)} - \boldsymbol{B}^{(t)}\rangle$. By applying Lemma B.4 and a union bound over T steps, with probability at least $1 - \mathcal{O}(n^{-c_0})$, we have

$$\begin{aligned}-\left\langle \boldsymbol{N}_{\boldsymbol{B}}^{(t)}, \boldsymbol{B}^{(t+1)} - \boldsymbol{B}^{(t)}\right\rangle &\leq \frac{1}{2a_{\mathrm{DP}}}\|\boldsymbol{B}^{(t+1)} - \boldsymbol{B}^{(t)}\|_F^2 + \Phi_{\mathrm{DP}}\\ &\leq \frac{1}{2a_{\mathrm{DP}}}\|\bar{\boldsymbol{B}}^{(t+1)} - \bar{\boldsymbol{B}}^{(t)}\|_F^2 + \Phi_{\mathrm{DP}}, \qquad \forall\, t \in [0, T-1].\end{aligned}$$

Substitute this into equation B.32 and using $\omega = \rho - \mathcal{L}M_{\bar{\boldsymbol{X}}} - 1/a_{\mathrm{DP}} > 0$ yields

$$F(\bar{\boldsymbol{B}}^{(t+1)}) \le F(\bar{\boldsymbol{B}}^{(t)}) - \frac{\omega}{2}\|\bar{\boldsymbol{B}}^{(t+1)} - \bar{\boldsymbol{B}}^{(t)}\|_F^2 + \Phi_{\mathrm{DP}}.$$

Summing over $t = 0, \ldots, T-1$ and applying $F(\bar{\boldsymbol{B}}^{(T)}) \ge F_{\inf}$ gives

$$\frac{\omega}{2}\sum_{t=0}^{T-1}\|\bar{\boldsymbol{B}}^{(t+1)} - \bar{\boldsymbol{B}}^{(t)}\|_F^2 \le F(\bar{\boldsymbol{B}}^{(0)}) - F_{\inf} + T\Phi_{\mathrm{DP}}.$$

This completes the proof. □

B.6 Technical Lemmas

Lemma B.4. *For each $0 \le J \le n$, $1 \le r \le p \wedge m$, define $\mathcal{B}_{r,J} := \{(\boldsymbol{B}, \boldsymbol{C}) \in \mathbb{R}^{p\times m} \times \mathbb{R}^{n\times m} : r(\boldsymbol{B}) \le r, J(\boldsymbol{C}) = J\}$. Then there exists universal constants $A_0, A_1, c, c' > 0$ such that for any $a \ge 2b > 0$, the following event*

$$\sup_{0\le J\le n}\sup_{(\boldsymbol{B},\boldsymbol{C})\in\mathcal{B}_{r,J}}\{2\langle\boldsymbol{\mathcal{E}}, \boldsymbol{XB}+\boldsymbol{C}\rangle - \frac{1}{a}\|\boldsymbol{XB}+\boldsymbol{C}\|_F^2 - \frac{1}{b}P_{2,H}(\boldsymbol{C};\lambda_{\boldsymbol{C}}) - aA_0\sigma^2 r(q+m)\} \ge a\sigma^2 t \tag{B.33}$$

occurs with probability at most $c'\exp(-ct)$, where $\lambda_{\boldsymbol{C}} = A\lambda^o_{\boldsymbol{C}}$, $\lambda^o_{\boldsymbol{C}} = \sigma(m + \log n)^{1/2}$, $A = (abA_1)^{1/2}$, and $t \ge 0$.

Proof. Let $P_0(t;\lambda) = (\lambda^2/2)1_{t\ne 0}$ and $P_{2,0}(\boldsymbol{C};\lambda_{\boldsymbol{C}}) = \sum_{i=1}^n P_0(\|\boldsymbol{c}_i\|_2;\lambda_{\boldsymbol{C}})$. Define

$$\begin{aligned} l_H(\boldsymbol{B},\boldsymbol{C}) :=& 2\langle\boldsymbol{\mathcal{E}}, \boldsymbol{XB}+\boldsymbol{C}\rangle - \frac{1}{a}\|\boldsymbol{XB}+\boldsymbol{C}\|_F^2 - \frac{1}{b}P_{2,H}(\boldsymbol{C};\lambda_{\boldsymbol{C}}) \\ &- aA_0\sigma^2 r(q+m), \end{aligned} \tag{B.34}$$

and similarly

$$\begin{aligned} l_0(\boldsymbol{B},\boldsymbol{C}) :=& 2\langle\boldsymbol{\mathcal{E}}, \boldsymbol{XB}+\boldsymbol{C}\rangle - \frac{1}{a}\|\boldsymbol{XB}+\boldsymbol{C}\|_F^2 - \frac{1}{b}P_{2,0}(\boldsymbol{C};\lambda_{\boldsymbol{C}}) \\ &- aA_0\sigma^2 r(q+m). \end{aligned} \tag{B.35}$$

Let

$$\mathcal{A}_H = \left\{\sup_{0\le J\le n}\sup_{(\boldsymbol{B},\boldsymbol{C})\in\mathcal{B}_{r,J}} l_H(\boldsymbol{B},\boldsymbol{C}) \ge a\sigma^2 t\right\}, \tag{B.36}$$

and define

$$\mathcal{A}_0 = \left\{\sup_{0\le J\le n}\sup_{(\boldsymbol{B},\boldsymbol{C})\in\mathcal{B}_{r,J}} l_0(\boldsymbol{B},\boldsymbol{C}) \ge a\sigma^2 t\right\}. \tag{B.37}$$

The occurrence of $\mathcal{A}_H$ implies that

$$l_H(\boldsymbol{B}^o,\boldsymbol{C}^o) \ge a\sigma^2 t, \tag{B.38}$$

for any $(\boldsymbol{B}^o,\boldsymbol{C}^o)$ that solves

$$\min_{\boldsymbol{B}:r(\boldsymbol{B})\le r,\boldsymbol{C}} -2\langle\boldsymbol{\mathcal{E}}, \boldsymbol{XB}+\boldsymbol{C}\rangle + \frac{1}{a}\|\boldsymbol{XB}+\boldsymbol{C}\|_F^2 + \frac{1}{b}P_{2,H}(\boldsymbol{C};\lambda_{\boldsymbol{C}}). \tag{B.39}$$

Lemma B.5 (Lemma 1 in She (2012))**.** *Given any $\tau \ge 1$, there exists a globally optimal solution $\boldsymbol{C}^o$ to $\min_{\boldsymbol{C}}\|\boldsymbol{Y}-\boldsymbol{C}\|_F^2/2 + \tau P_{2,H}(\boldsymbol{C};\lambda)$ such that for any i, either $\boldsymbol{c}_i^o = \boldsymbol{0}$ or $\|\boldsymbol{c}_i^o\|_2 \ge \lambda\tau^{1/2} \ge \lambda$.*

From Lemma B.5 and $a \ge 2b$, equation B.38 implies the existence of an optimal solution $(\boldsymbol{B}^o,\boldsymbol{C}^o)$ to equation B.39 such that $l_0(\boldsymbol{B}^o,\boldsymbol{C}^o) = l_H(\boldsymbol{B}^o,\boldsymbol{C}^o) \ge a\sigma^2 t$. Therefore, $\mathcal{A}_H \subset \mathcal{A}_0$ and it suffices to study $\mathbb{P}(\mathcal{A}_0)$.

Let $\mathcal{J} = \mathcal{J}(\boldsymbol{C})$ for short and $I_{\mathcal{J}}$ be the submatrix of $\boldsymbol{I}_{n\times n}$ formed by the columns indexed by $\mathcal{J}$. We decompose the stochastic term into

$$2\langle\boldsymbol{\mathcal{E}}, \boldsymbol{XB}+\boldsymbol{C}\rangle = 2\langle\boldsymbol{\mathcal{E}}, \mathcal{P}^{\perp}_{I_{\mathcal{J}}}\boldsymbol{XB}\rangle + 2\langle\boldsymbol{\mathcal{E}}, \mathcal{P}_{I_{\mathcal{J}}}(\boldsymbol{XB}+\boldsymbol{C})\rangle \equiv 2\langle\boldsymbol{\mathcal{E}}, \boldsymbol{I}\rangle + 2\langle\boldsymbol{\mathcal{E}}, \boldsymbol{II}\rangle. \tag{B.40}$$

Lemma B.6. (Lemma 5 in She & Chen (2017)) *Suppose* $\mathrm{vec}(\boldsymbol{\mathcal{E}})$ *is sub-Gaussian with mean zero and* ψ_2*-norm bounded by* σ*. Given* $\boldsymbol{X} \in \mathbb{R}^{n\times p}$*,* $1 \le r \le m \wedge p$*,* $0 \le J \le n$*, define* $\Gamma^1_{r,J} = \{\boldsymbol{A} \in \mathbb{R}^{n\times m} : \|\boldsymbol{A}\|_F \le 1, r(\boldsymbol{A}) \le r, \mathrm{CS}(\boldsymbol{A}) \subset \mathrm{CS}\{\boldsymbol{X}(\mathcal{J}^c, :)\}$ *for some* $\mathcal{J} : |\mathcal{J}| = J\}$*. Define*

$$P^1_o(J, r) = \sigma^2 \left[\{q \wedge (n - J)\}r + (m - r)r + \log \binom{n}{J}\right]. \tag{B.41}$$

Then for any $t \ge 0$*,*

$$\mathbb{P}[\sup_{\boldsymbol{A}\in\Gamma^1_{r,J}} \langle \boldsymbol{\mathcal{E}}, \boldsymbol{A}\rangle \ge t\sigma + \{LP^1_o(J, r)\}^{1/2}] \le c' \exp(-ct^2), \tag{B.42}$$

where L, c, c' *are universal positive constants.*

Observe that

$$\begin{aligned}
&2\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}\rangle - 2aLP^1_o(J, r) - \frac{1}{a}\|\boldsymbol{I}\|^2_F \\
\le&2\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}/\|\boldsymbol{I}\|_F\rangle\|\boldsymbol{I}\|_F - \frac{1}{2a}\|\boldsymbol{I}\|^2_F - 2\|\boldsymbol{I}\|_F\{LP^1_o(J, r)\}^{1/2} \\
\le&2a[\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}/\|\boldsymbol{I}\|_F\rangle - \{LP^1_o(J, r)\}^{1/2}]^2_+,
\end{aligned}$$

where $[x]_+ = \max(x, 0)$. Applying Lemma B.6, it then follows that for L sufficiently large

$$\mathbb{P}\{\sup_{(\boldsymbol{B},\boldsymbol{C})\in\mathcal{B}_{r,J}} \{2\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}\rangle - 2aLP^1_o(J, r) - \frac{1}{a}\|\boldsymbol{I}\|^2_F\} \ge \frac{1}{2}a\sigma^2 t\} \le \frac{1}{2}c' \exp(-ct). \tag{B.43}$$

Similarly, with L large enough, we have

$$\mathbb{P}\{\sup_{(\boldsymbol{B},\boldsymbol{C})\in\mathcal{B}_{r,J}} \{2\langle \boldsymbol{\mathcal{E}}, \boldsymbol{II}\rangle - 2aLP^2_o(J) - \frac{1}{a}\|\boldsymbol{II}\|^2_F\} \ge \frac{1}{2}a\sigma^2 t\} \le \frac{1}{2}c' \exp(-ct). \tag{B.44}$$

with

$$P^2_o(J) = \sigma^2 \left[Jm + \log \binom{n}{J}\right].$$

Applying the union bound to equation B.43 and equation B.44 then yields

$$\begin{aligned}
\mathbb{P}[\sup_{(\boldsymbol{B},\boldsymbol{C})\in\mathcal{B}_{r,J}} \{2\langle \boldsymbol{\mathcal{E}}, \boldsymbol{XB} + \boldsymbol{C}\rangle - \frac{1}{a}\|\boldsymbol{XB} + \boldsymbol{C}\|^2_F - 2aL\sigma^2\{(q + m)r \\
+ Jm + 2\log \binom{n}{J}\}\} \ge a\sigma^2 t] \le c' \exp(-ct).
\end{aligned}$$

With $\lambda^2_{\boldsymbol{C}} = abA_1\sigma^2(m + \log n)$, for $(\boldsymbol{B}, \boldsymbol{C}) \in \mathcal{B}_{r,J}$ we have

$$\frac{1}{b}P_{2,0}(\boldsymbol{C}; \lambda_{\boldsymbol{C}}) = \frac{aA_1}{2}\sigma^2 J(m + \log n). \tag{B.45}$$

For $J \ge 1$, since $\log \binom{n}{J} \le J \log(en/J)$, we have

$$Jm + 2\log \binom{n}{J} \le Jm + 2J \log(en/J) \lesssim J(m + \log n). \tag{B.46}$$

Hence, by choosing A_1 sufficiently large, there exists a universal constant $c_0 > 0$ such that

$$\frac{1}{b}P_{2,0}(\boldsymbol{C}; \lambda_{\boldsymbol{C}}) \ge 2aL\sigma^2\Big\{Jm + 2\log \binom{n}{J}\Big\} + ac_0\sigma^2 J(m + \log n). \tag{B.47}$$

Similarly, choosing A_0 sufficiently large gives

$$aA_0\sigma^2 r(q + m) \ge 2aL\sigma^2 r(q + m). \tag{B.48}$$

Consequently, for any fixed $J \ge 1$,

$$\begin{aligned}
\mathbb{P}\left[\sup_{(\boldsymbol{B},\boldsymbol{C})\in\mathcal{B}_{r,J}} \Big\{l_0(\boldsymbol{B}, \boldsymbol{C})\Big\} \ge a\sigma^2 t\right] &\le c' \exp\{-c[t + c_0 J(m + \log n)]\} \\
&\le c' \exp(-ct) \exp\{-cc_0 J(m + \log n)\}.
\end{aligned} \tag{B.49}$$

When $J = 0$, we have $\boldsymbol{C} = \mathbf{0}$, $P_{2,0}(\boldsymbol{C}; \lambda_{\boldsymbol{C}}) = 0$, and $\log \binom{n}{0} = 0$. Thus, we have

$$\mathbb{P}\left[\sup_{(\boldsymbol{B},\boldsymbol{C})\in\mathcal{B}_{r,0}} \left\{ l_0(\boldsymbol{B},\boldsymbol{C}) \right\} \geq a\sigma^2 t \right] \leq c' \exp(-ct). \tag{B.50}$$

Finally, applying the union bound over $J = 0, \ldots, n$ gives

$$\mathbb{P}(\mathcal{A}_0) \leq c' \exp(-ct) + c' \exp(-ct) \sum_{J=1}^{n} \exp\{-cc_0 J(m + \log n)\}. \tag{B.51}$$

Since $\mathcal{A}_H \subset \mathcal{A}_0$, with c' redefined, we conclude that

$$\mathbb{P}(\mathcal{A}_H) \leq c' \exp(-ct). \tag{B.52}$$

The proof is complete.

□

Lemma B.7. *The optimization problem $Q(\boldsymbol{B}) = \|\boldsymbol{B} - \boldsymbol{Y}\|_F^2/2$ s.t. $\operatorname{rank}(\boldsymbol{B}) \leq r$ has $\boldsymbol{B}^o$ as a globally optimal solution. Then for any $\boldsymbol{B}$ satisfying $\operatorname{rank}(\boldsymbol{B}) \leq r/\vartheta$ with $\vartheta \geq 1$, we have*

$$Q(\boldsymbol{B}) - Q(\boldsymbol{B}^o) \geq \{1 - \vartheta^{-1/2}\}\|\boldsymbol{B} - \boldsymbol{B}^o\|_F^2/2. \tag{B.53}$$

Proof. The result follows directly from Proposition 2.2 of She (2013) combined with Lemma B.8. For brevity, the detailed proof is omitted. □

Lemma B.8 (Lemma 9 in She (2013))**.** *The optimization problem $\min_{\boldsymbol{C}} l(\boldsymbol{C}) = \|\boldsymbol{Y} - \boldsymbol{C}\|_F^2/2$ s.t. $\|\boldsymbol{C}\|_{2,0} \leq q$ has $\hat{\boldsymbol{C}}$ as a globally optimal solution. Assume that $J(\hat{\boldsymbol{C}}) = q$, where $J(\cdot) = \|\cdot\|_{2,0}$. Then for any $\boldsymbol{C}$ with $J(\boldsymbol{C}) \leq s = q/\vartheta$ and $\vartheta \geq 1$, we have*

$$l(\boldsymbol{C}) - l(\hat{\boldsymbol{C}}) \geq \{1 - \mathcal{L}(\mathcal{J}, \hat{\mathcal{J}})\}\|\hat{\boldsymbol{C}} - \boldsymbol{C}\|_F^2/2, \tag{B.54}$$

where $\mathcal{L}(\mathcal{J}, \hat{\mathcal{J}}) = (|\mathcal{J} \setminus \hat{\mathcal{J}}|/|\hat{\mathcal{J}} \setminus \mathcal{J}|)^{1/2} \leq (s/q)^{1/2} = \vartheta^{-1/2}$, with $\mathcal{J} = \mathcal{J}(\boldsymbol{C})$ and $\hat{\mathcal{J}} = \mathcal{J}(\hat{\boldsymbol{C}})$ denoting their respective row support sets.

C Additional Multi-Label Classification Experiments

This section provides complete results for the synthetic multi-label classification experiments. All results are reported as mean ± standard error over 20 independent replications.

C.1 Classification Data Generation

The predictor matrix $\boldsymbol{X} \in \mathbb{R}^{n_{\text{train}} \times p}$ is constructed using a bounded spherical design, where each row is generated by first drawing $\boldsymbol{z}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$ and then setting $\boldsymbol{x}_i = \sqrt{p}\boldsymbol{z}_i/\|\boldsymbol{z}_i\|_2$, such that every predictor row satisfies $\|\boldsymbol{x}_i\|_2 = \sqrt{p}$. The true coefficient matrix $\boldsymbol{B}^* \in \mathbb{R}^{p \times m}$ is generated as $\boldsymbol{B}^* = 4\boldsymbol{U}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{p \times r^*}$ and $\boldsymbol{V} \in \mathbb{R}^{m \times r^*}$ are obtained by orthonormalizing independent standard Gaussian matrices. Outliers are then added exclusively to the training set. Specifically, let $\boldsymbol{Z}^* = \boldsymbol{X}\boldsymbol{B}^*$ denote the clean logit matrix. For the $\lfloor 10\% \times n_{\text{train}} \rfloor$ samples yielding the largest average absolute clean margins $\frac{1}{m}\sum_{j=1}^{m} |z_{ij}^*|$, we apply an adversarial logit contamination by setting the corresponding entries of the contamination matrix $\boldsymbol{C}^* = (c_{ij}^*)_{1 \leq i \leq n_{\text{train}}, 1 \leq j \leq m} \in \mathbb{R}^{n_{\text{train}} \times m}$ to $c_{ij}^* = -(\alpha/\sqrt{n_{\text{train}}})\operatorname{sign}(z_{ij}^*)$. The remaining rows of $\boldsymbol{C}^*$ are set to zero. The binary training response is generated entrywise as $y_{ij} \sim \text{Bernoulli}(\sigma(z_{ij}^* + \sqrt{n_{\text{train}}}c_{ij}^*))$, where $\sigma(t) = 1/(1 + \exp(-t))$. An independent test set is generated from the same bounded spherical design and the clean logistic model without $\boldsymbol{C}^*$, ensuring evaluation on uncontaminated data. Unless otherwise specified, we set $n_{\text{train}} = 5000$, $n_{\text{test}} = 2000$, $p = 20$, $m = 10$, $r^* = 2$, the contamination proportion to 10%, and $\alpha = 8$.

C.2 Implementation Details and Privacy Accounting for Baselines

We abbreviate binary cross-entropy, generalized cross-entropy, differentially private stochastic gradient descent, and differentially private empirical risk minimization as BCE, GCE, DP-SGD, and DP-ERM, respectively.

Privacy accounting. We adopt replace-one adjacency, treating each complete multi-label observation as one private record. All iterative full-batch baselines based on Gaussian gradient perturbation use public analytical per-record gradient bounds and the same zCDP composition rule as DP-RoLL. DP-SGD is accounted for using the fixed-size subsampled RDP bound of Wang et al. (2019), followed by conversion to (ϵ, δ)-DP. DP Output Perturbation and DP Objective Perturbation allocate ϵ/m to each label. Sequential composition over the m label-wise models yields pure ϵ-DP and hence (ϵ, δ)-DP. Only the final privately trained model parameters are released.

Gradient bounding. All methods use labels encoded as $\{0, 1\}$ and sum the per-label losses over the m labels. Under the bounded synthetic design, the Frobenius norm of each per-record coefficient gradient is bounded by $\sqrt{pm}$. Hence, no explicit gradient clipping is required. For Gaussian gradient-perturbation methods, noise is calibrated using the public analytical bound $C_B = \sqrt{pm}$. Hard-rank projection, singular-value soft thresholding, and SVD retraction are applied after privatization and therefore constitute post-processing.

DP-SGD. DP-SGD uses fixed-size mini-batches sampled without replacement. Per-record gradients are summed over the mini-batch, perturbed with Gaussian noise calibrated to the public analytical bound $C_B = \sqrt{pm}$, and divided by the mini-batch size (Abadi et al., 2016). Privacy amplification and composition are accounted for using the fixed-size subsampled RDP bound of Wang et al. (2019).

DP-ERM. DP-ERM uses the full batch and adds Gaussian noise, calibrated to the public analytical bound $C_B = \sqrt{pm}$, to the aggregated gradient at each iteration. It is implemented as unconstrained noisy gradient descent without a low-rank constraint, following the general DP-ERM framework of Wang & Xu (2019).

DP Output Perturbation. For each label, output perturbation fits ridge-regularized logistic regression and adds spherical Laplace noise to the fitted coefficients (Chaudhuri et al., 2011). The sensitivity bound for the exact minimizer is conservatively enlarged to account for the fixed Newton-solver tolerance. Each label is allocated privacy budget ϵ/m, and composition over the m label-wise models gives pure ϵ-DP.

DP Objective Perturbation. For each label, objective perturbation normalizes the features to unit norm and adds a spherical random linear term to the ridge-regularized logistic objective (Chaudhuri et al., 2011). We allocate 99% of the per-label privacy budget to objective perturbation and reserve the remaining 1% for a spherical Laplace correction that accounts for numerical optimization error. The regularization strength is increased when necessary to satisfy the objective-perturbation privacy condition. Composition over labels gives pure ϵ-DP.

DP Low-Rank GCE. DP Low-Rank GCE combines the bounded robust GCE score of Zhang & Sabuncu (2018) with Gaussian noisy-gradient updates and a hard rank projection. Gaussian noise is calibrated using the public analytical bound $C_B = \sqrt{pm}$. It is a private multi-label adaptation of the original non-private GCE method.

DP trace-norm BCE. DP trace-norm BCE combines BCE with trace-norm regularization and uses Gaussian noisy proximal-gradient updates followed by singular-value soft thresholding (Pontil & Maurer, 2013). The private optimizer is an adaptation introduced for the present comparison.

DP Riemannian low-rank BCE. DP Riemannian low-rank BCE projects each analytically bounded per-observation gradient onto the tangent space, averages the resulting tangent gradients, adds tangent-space Gaussian noise, and applies a truncated SVD retraction to maintain the working

Table C.1: Complete comparison on synthetic multi-label classification data under different privacy budgets. Results are reported as mean ± standard error over 20 independent replications. Bold values denote the best results.

Method	$\|\hat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F^2 \downarrow$	HL↓	AP↑	OE↓	RL↓	Cov.↓
$\epsilon = 0.5$						
DP-RoLL	$\mathbf{7.485 \pm 0.317}$	$\mathbf{0.299 \pm 0.003}$	$\mathbf{0.804 \pm 0.003}$	$\mathbf{0.179 \pm 0.005}$	$\mathbf{0.233 \pm 0.003}$	$\mathbf{0.645 \pm 0.003}$
DP-SGD	31.509 ± 0.042	0.462 ± 0.003	0.644 ± 0.003	0.425 ± 0.006	0.445 ± 0.004	0.779 ± 0.002
DP-ERM	25.030 ± 0.481	0.402 ± 0.003	0.704 ± 0.004	0.314 ± 0.006	0.363 ± 0.005	0.734 ± 0.003
DP Output Perturbation	33.182 ± 0.306	0.466 ± 0.003	0.638 ± 0.003	0.438 ± 0.006	0.454 ± 0.004	0.782 ± 0.002
DP Objective Perturbation	31.145 ± 0.162	0.464 ± 0.003	0.641 ± 0.003	0.433 ± 0.006	0.450 ± 0.004	0.781 ± 0.002
DP Low-Rank GCE	27.720 ± 0.636	0.429 ± 0.008	0.678 ± 0.008	0.369 ± 0.014	0.399 ± 0.011	0.750 ± 0.006
DP trace-norm BCE	27.725 ± 0.487	0.406 ± 0.004	0.701 ± 0.003	0.321 ± 0.006	0.366 ± 0.005	0.737 ± 0.003
DP Riemannian low-rank BCE	27.487 ± 0.534	0.428 ± 0.005	0.682 ± 0.005	0.349 ± 0.012	0.395 ± 0.007	0.750 ± 0.005
$\epsilon = 1$						
DP-RoLL	$\mathbf{4.678 \pm 0.153}$	$\mathbf{0.286 \pm 0.003}$	$\mathbf{0.816 \pm 0.003}$	$\mathbf{0.159 \pm 0.005}$	$\mathbf{0.217 \pm 0.003}$	$\mathbf{0.633 \pm 0.002}$
DP-SGD	29.938 ± 0.072	0.425 ± 0.003	0.678 ± 0.003	0.360 ± 0.006	0.397 ± 0.004	0.754 ± 0.002
DP-ERM	18.174 ± 0.256	0.359 ± 0.003	0.749 ± 0.003	0.244 ± 0.005	0.303 ± 0.004	0.696 ± 0.003
DP Output Perturbation	29.363 ± 0.146	0.436 ± 0.003	0.667 ± 0.003	0.385 ± 0.007	0.413 ± 0.005	0.761 ± 0.002
DP Objective Perturbation	28.778 ± 0.296	0.432 ± 0.003	0.674 ± 0.003	0.371 ± 0.005	0.404 ± 0.004	0.759 ± 0.002
DP Low-Rank GCE	19.388 ± 1.206	0.363 ± 0.008	0.747 ± 0.008	0.256 ± 0.014	0.309 ± 0.010	0.696 ± 0.006
DP trace-norm BCE	18.175 ± 0.256	0.359 ± 0.003	0.749 ± 0.003	0.244 ± 0.005	0.303 ± 0.004	0.696 ± 0.003
DP Riemannian low-rank BCE	18.917 ± 0.568	0.352 ± 0.006	0.758 ± 0.006	0.235 ± 0.009	0.294 ± 0.007	0.686 ± 0.005
$\epsilon = 2$						
DP-RoLL	$\mathbf{2.761 \pm 0.101}$	$\mathbf{0.276 \pm 0.002}$	$\mathbf{0.826 \pm 0.003}$	$\mathbf{0.147 \pm 0.005}$	$\mathbf{0.205 \pm 0.003}$	$\mathbf{0.624 \pm 0.002}$
DP-SGD	23.143 ± 0.217	0.382 ± 0.004	0.725 ± 0.004	0.282 ± 0.007	0.336 ± 0.006	0.716 ± 0.003
DP-ERM	15.593 ± 0.154	0.319 ± 0.004	0.790 ± 0.003	0.185 ± 0.005	0.251 ± 0.004	0.659 ± 0.003
DP Output Perturbation	26.944 ± 0.146	0.403 ± 0.003	0.701 ± 0.004	0.329 ± 0.007	0.367 ± 0.005	0.736 ± 0.003
DP Objective Perturbation	25.800 ± 0.147	0.379 ± 0.003	0.728 ± 0.003	0.277 ± 0.006	0.332 ± 0.004	0.716 ± 0.003
DP Low-Rank GCE	6.433 ± 0.280	0.302 ± 0.003	0.808 ± 0.003	0.169 ± 0.007	0.229 ± 0.004	0.641 ± 0.003
DP trace-norm BCE	15.468 ± 0.237	0.339 ± 0.003	0.770 ± 0.003	0.216 ± 0.006	0.276 ± 0.003	0.677 ± 0.002
DP Riemannian low-rank BCE	14.492 ± 0.271	0.312 ± 0.005	0.798 ± 0.004	0.177 ± 0.006	0.241 ± 0.004	0.650 ± 0.003
$\epsilon = 4$						
DP-RoLL	$\mathbf{1.227 \pm 0.055}$	$\mathbf{0.269 \pm 0.002}$	$\mathbf{0.833 \pm 0.002}$	$\mathbf{0.137 \pm 0.004}$	$\mathbf{0.196 \pm 0.003}$	$\mathbf{0.616 \pm 0.002}$
DP-SGD	16.358 ± 0.218	0.339 ± 0.003	0.770 ± 0.003	0.212 ± 0.005	0.277 ± 0.004	0.678 ± 0.003
DP-ERM	14.125 ± 0.128	0.308 ± 0.003	0.801 ± 0.003	0.174 ± 0.005	0.236 ± 0.003	0.648 ± 0.002
DP Output Perturbation	23.762 ± 0.176	0.380 ± 0.003	0.724 ± 0.004	0.288 ± 0.007	0.335 ± 0.005	0.716 ± 0.003
DP Objective Perturbation	25.104 ± 0.088	0.329 ± 0.003	0.779 ± 0.003	0.203 ± 0.006	0.265 ± 0.004	0.670 ± 0.003
DP Low-Rank GCE	2.714 ± 0.130	0.282 ± 0.003	0.825 ± 0.003	0.144 ± 0.005	0.207 ± 0.003	0.625 ± 0.002
DP trace-norm BCE	14.134 ± 0.128	0.308 ± 0.003	0.801 ± 0.003	0.174 ± 0.005	0.236 ± 0.003	0.648 ± 0.002
DP Riemannian low-rank BCE	13.768 ± 0.135	0.288 ± 0.004	0.819 ± 0.003	0.149 ± 0.005	0.214 ± 0.003	0.630 ± 0.003

rank (Han et al., 2024). The multi-label objective, retraction, and zCDP accounting are adaptations to the present setting.

Further implementation details, including solver tolerances, noise calibration, and privacy-accounting parameters, are provided in the supplementary code.

C.3 Complete Results under Different Privacy Budgets

We vary the privacy budget over $\epsilon \in \{0.5, 1, 2, 4\}$ while keeping all other classification settings fixed. As shown in Table C.1, DP-RoLL achieves the best performance across all six metrics at every privacy budget.

C.4 Real-world Datasets Description

We select these datasets to cover diverse application domains and a broad range of sample sizes, feature dimensions, and label cardinalities. The standard benchmark datasets enable evaluation under controlled contamination, whereas the datasets marked with † contain intrinsic label noise and provide a complementary assessment of robustness in naturally noisy settings.

TABLE C.2: CHARACTERISTICS OF THE DATASETS USED IN THE EXPERIMENTS. DATASETS MARKED WITH † ARE REAL-WORLD NOISY DATASETS. ALL DATASETS ARE WIDELY USED IN PRIOR WORKS (YU ET AL., 2018; HANG & ZHANG, 2023; CHEN ET AL., 2024; QIAN ET AL., 2025).

DATASET	DOMAIN	#INSTANCES	#FEATURES	#LABELS
CAL500	AUDIO	502	68	174
EDUCATION	TEXT	5,000	550	33
EMOTIONS	MUSIC	593	72	6
FLAGS	IMAGE	194	10	7
FOODTRUCK	FOOD SERVICE	407	17	12
IMAGE	IMAGE	2,000	294	5
SCENE	IMAGE	2,407	294	6
HEALTH	TEXT	5,000	612	32
MIRFLICKR†	IMAGE	10,433	100	7
MUSIC EMOTION†	AUDIO	6,833	98	11
MUSIC STYLE†	AUDIO	6,839	98	10

C.5 ADDITIONAL REAL DATA RESULTS

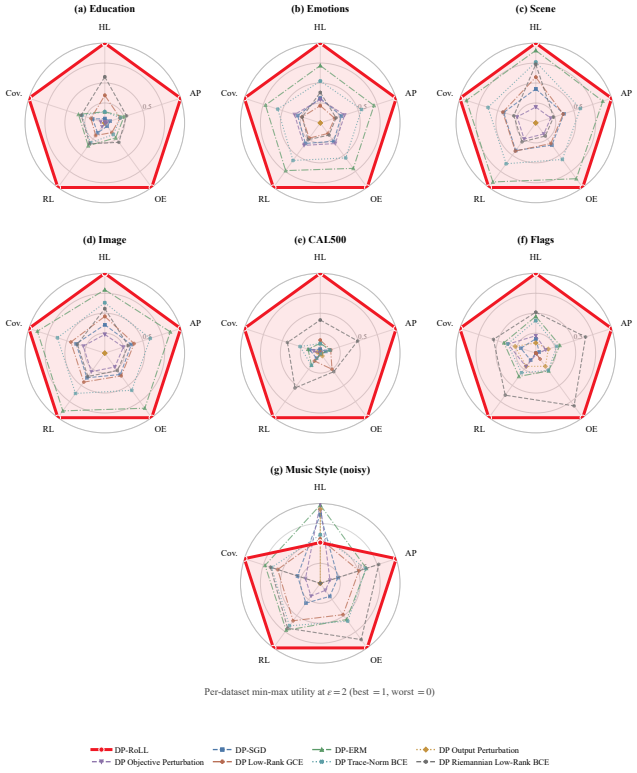

Figure C.1: Radar charts for the remaining benchmark datasets at $\epsilon = 2$. The axes are normalized so that larger values indicate better performance, and a larger enclosed area represents better overall performance.

To summarize performance across all real-world datasets, we rank the eight methods separately for each dataset, evaluation metric, and privacy budget, using the appropriate direction for each metric and assigning average ranks to ties. For each method and privacy budget, the resulting ranks are averaged over the eleven datasets and five metrics, yielding 55 comparisons per entry. Lower average ranks indicate better overall performance.

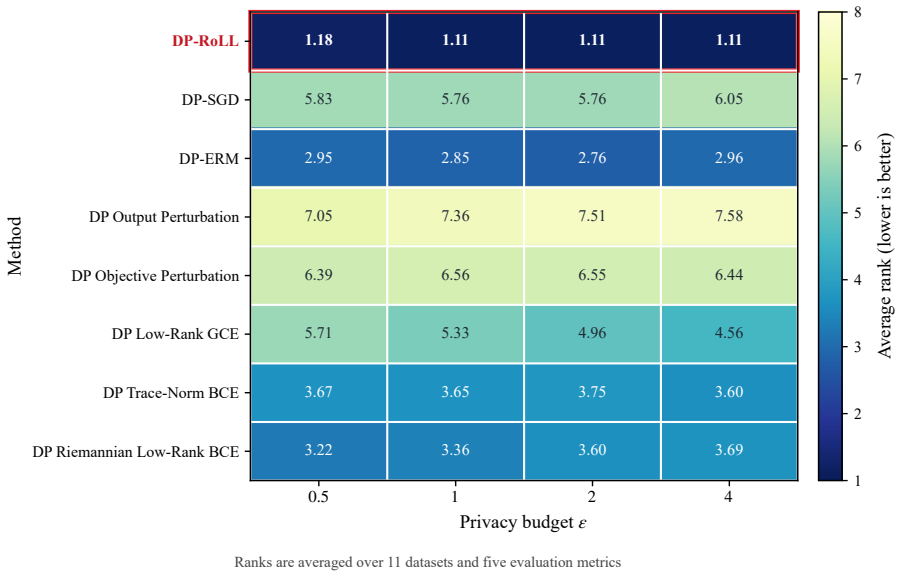

Figure C.2: Average ranks over 11 real-world datasets and five evaluation metrics under different privacy budgets. Lower values indicate better performance.

As shown in Figure C.2, DP-RoLL achieves average ranks of 1.18, 1.11, 1.11, and 1.11 at $\epsilon = 0.5$, 1, 2, and 4, respectively, consistently ranking first across all four privacy budgets.

C.6 Additional Ablation Results

We report additional ablation results on Foodtruck at $\epsilon = 0.5$ and on Health and Education at $\epsilon \in \{0.5, 2\}$. DP-RoLL achieves the best mean performance across all five metrics on Foodtruck at $\epsilon = 0.5$, on Health at $\epsilon = 2$, and on Education under both privacy budgets. On Health at $\epsilon = 0.5$, DP-RoLL achieves the best Hamming loss, ranking loss, and coverage, while the variant without the low-rank constraint is slightly better in average precision and one-error.

Table C.3: Ablation results on Foodtruck at $\epsilon = 0.5$, reported as mean $\pm$ standard error over 20 independent train–test splits. Bold values denote the best results.

Method	HL↓	AP↑	OE↓	RL↓	Cov.↓
DP-RoLL	**0.3854±0.0090**	**0.4014±0.0126**	**0.7470±0.0207**	**0.4180±0.0117**	**0.5371±0.0114**
w/o $\boldsymbol{C}$	0.4048±0.0073	0.3951±0.0120	0.7488±0.0160	0.4291±0.0107	0.5485±0.0103
w/o low rank	0.4259±0.0066	0.3821±0.0090	0.7665±0.0123	0.4446±0.0100	0.5596±0.0088

Table C.4: Ablation results on Education at $\epsilon = 0.5$, reported as mean $\pm$ standard error over 20 independent train–test splits. Bold values denote the best results.

Method	HL↓	AP↑	OE↓	RL↓	Cov.↓
DP-RoLL	**0.3713±0.0029**	**0.1620±0.0045**	**0.9395±0.0067**	**0.4637±0.0041**	**0.5183±0.0042**
w/o $\boldsymbol{C}$	0.4216±0.0020	0.1557±0.0040	0.9412±0.0043	0.4753±0.0031	0.5283±0.0028
w/o low rank	0.4667±0.0009	0.1503±0.0016	0.9456±0.0017	0.4717±0.0034	0.5267±0.0034

C.7 Gradient Inversion Attack under Hypothetical First-Gradient Leakage

Following gradient-matching attacks (Zhu et al., 2019; Geiping et al., 2020), we consider a hypothetical leakage of the first privatized summed gradients with respect to the coefficient matrix and intercept. The per-record gradients are computed at zero initialization, summed, and perturbed with Gaussian noise calibrated using the public analytical gradient bound. Under a strong leave-one-out

Table C.5: Ablation results on Education at $\epsilon = 2$, reported as mean $\pm$ standard error over 20 independent train–test splits. Bold values denote the best results.

Method	HL↓	AP↑	OE↓	RL↓	Cov.↓
DP-RoLL	$\mathbf{0.1226 \pm 0.0020}$	$\mathbf{0.2411 \pm 0.0043}$	$\mathbf{0.8672 \pm 0.0055}$	$\mathbf{0.3672 \pm 0.0055}$	$\mathbf{0.4246 \pm 0.0053}$
w/o $\boldsymbol{C}$	0.1946 ± 0.0030	0.2039 ± 0.0055	0.8997 ± 0.0068	0.4081 ± 0.0046	0.4641 ± 0.0046
w/o low rank	0.3878 ± 0.0013	0.1881 ± 0.0019	0.9116 ± 0.0021	0.4217 ± 0.0025	0.4768 ± 0.0024

Table C.6: Ablation results on Health at $\epsilon = 0.5$, reported as mean $\pm$ standard error over 20 independent train–test splits. Bold values denote the best results.

Method	HL↓	AP↑	OE↓	RL↓	Cov.↓
DP-RoLL	$\mathbf{0.3764 \pm 0.0042}$	0.1724 ± 0.0107	0.9299 ± 0.0129	$\mathbf{0.4497 \pm 0.0071}$	$\mathbf{0.5318 \pm 0.0062}$
w/o $\boldsymbol{C}$	0.4231 ± 0.0027	0.1676 ± 0.0060	0.9315 ± 0.0073	0.4655 ± 0.0047	0.5483 ± 0.0050
w/o low rank	0.4684 ± 0.0011	$\mathbf{0.1731 \pm 0.0028}$	$\mathbf{0.9241 \pm 0.0028}$	0.4531 ± 0.0043	0.5360 ± 0.0044

Table C.7: Ablation results on Health at $\epsilon = 2$, reported as mean $\pm$ standard error over 20 independent train–test splits. Bold values denote the best results.

Method	HL↓	AP↑	OE↓	RL↓	Cov.↓
DP-RoLL	$\mathbf{0.1375 \pm 0.0023}$	$\mathbf{0.3320 \pm 0.0078}$	$\mathbf{0.7592 \pm 0.0107}$	$\mathbf{0.3084 \pm 0.0063}$	$\mathbf{0.3964 \pm 0.0069}$
w/o $\boldsymbol{C}$	0.2156 ± 0.0021	0.2671 ± 0.0066	0.8355 ± 0.0088	0.3565 ± 0.0040	0.4446 ± 0.0044
w/o low rank	0.3964 ± 0.0012	0.2389 ± 0.0031	0.8625 ± 0.0034	0.3739 ± 0.0031	0.4586 ± 0.0034

assumption, the attacker knows and subtracts all non-target clipped contributions, leaving the noisy target contribution. Because `Image` has five labels, the attacker enumerates all binary label vectors, analytically reconstructs the features for each candidate, and selects the candidate with the smallest gradient-matching error. We set $\delta = 10^{-5}$ for all finite privacy budgets.

For each of 20 independent train–test splits, we attack 10 clean and 10 corrupted training records. Results are first averaged over the 20 targets within each split and then reported as the mean and standard error across splits. For corrupted records, label recovery is evaluated against the observed labels used for training. Since `Image` contains 294 precomputed visual features rather than raw pixels, reconstruction is assessed using feature cosine similarity (the cosine similarity between the reconstructed and target feature vectors), label F1 (the example-wise F1 score between the reconstructed and training label vectors), and joint success (the proportion of targets with feature cosine similarity of at least 0.9 and an exactly reconstructed label vector). Joint success requires feature cosine similarity of at least 0.9 and exact recovery of the complete five-label vector.

This threat model is stronger than DP-RoLL's formal release interface, which releases only the final $\boldsymbol{B}$ and privatized intercept. It therefore provides a conservative leakage stress test rather than a direct attack on the released model (Huang et al., 2021). Moreover, because the gradient is evaluated at $t = 0$ before the first parameter update, the experiment isolates the protection provided by Gaussian perturbation calibrated to the public analytical gradient bound, before the low-rank projection or the internal updates of $\boldsymbol{C}$ can affect subsequent gradients.

Table C.8: First-gradient reconstruction on `Image` under a strong leave-one-out threat model. Results are reported as mean $\pm$ standard error over 20 independent splits. Higher values indicate greater leakage. The non-private reference applies the same reconstruction procedure to an unclipped and unnoised gradient.

ϵ	Feature cosine	Label F1	Joint success
0.5	-0.001 ± 0.004	0.310 ± 0.028	0.000 ± 0.000
1	-0.001 ± 0.004	0.311 ± 0.029	0.000 ± 0.000
2	-0.001 ± 0.004	0.310 ± 0.028	0.000 ± 0.000
4	-0.001 ± 0.004	0.311 ± 0.029	0.000 ± 0.000
Non-private	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

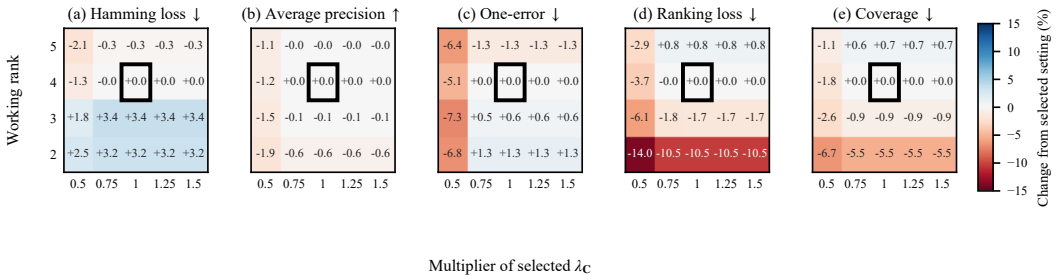

Figure C.3: Hyperparameter sensitivity of DP-RoLL to r and λ_C on MIRFlickr at $\epsilon = 2$. Cells show percentage changes from the selected configuration ($r = 4$, multiplier 1) over 20 independent replications. The black box marks the selected configuration.

Table C.8 shows near-zero feature cosine similarity, label F1 of approximately 0.31, and no successful joint reconstruction at any finite privacy budget. By contrast, the unclipped and unnoised non-private reference yields exact reconstruction.

C.8 Hyperparameter Sensitivity Analysis

We assess the sensitivity of DP-RoLL to the working rank and λ_C on the noisy MIRFlickr dataset at $(\epsilon, \delta) = (2, 10^{-5})$. We jointly vary $r \in \{2, 3, 4, 5\}$ and scale the selected λ_C by a factor in $\{0.5, 0.75, 1, 1.25, 1.5\}$, keeping the number of private iterations and all other settings fixed. Each configuration is evaluated on the same 20 independent splits.

Figure C.3 reports direction-adjusted percentage changes relative to the selected configuration, with positive values indicating improvement and negative values indicating degradation. Performance remains stable around the selected setting $(r, \text{multiplier}) = (4, 1)$. Across all combinations with $r \in \{3, 4, 5\}$ and multipliers in $\{0.75, 1, 1.25, 1.5\}$, the largest degradation across the five metrics is below 1.8%.

C.9 Performance Comparison and Error Scaling under Increasing Contamination

At $\epsilon = 2$, we first compare all methods at contamination proportions of 5%, 7.5%, and 10%, while keeping all other settings fixed. Table C.9 shows that DP-RoLL achieves the best performance across all six metrics at each contamination level. We then evaluate DP-RoLL over the finer grid $\{5\%, 6\%, 7\%, 8\%, 9\%, 10\%\}$, with the number of private iterations $T = 20$ and all other classification settings fixed. As shown in Figure C.4, its estimation error increases linearly with the contamination proportion, consistent with the theoretical error bound.

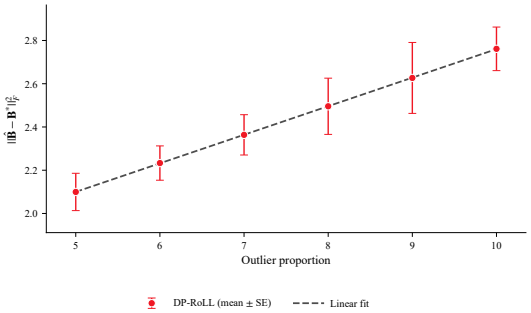

Figure C.4: Estimation error of DP-RoLL under increasing contamination proportions at $\epsilon = 2$. Points and error bars denote means and standard errors over 20 independent replications.

Table C.9: Classification performance under increasing contamination at $\epsilon = 2$, reported as mean $\pm$ standard error over 20 independent replications. Bold values denote the best results.

Method	$\|\hat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F^2 \downarrow$	HL↓	AP↑	OE↓	RL↓	Cov.↓
5% **contamination**						
DP-RoLL	$\mathbf{2.100 \pm 0.086}$	$\mathbf{0.275 \pm 0.002}$	$\mathbf{0.827 \pm 0.003}$	$\mathbf{0.147 \pm 0.005}$	$\mathbf{0.203 \pm 0.003}$	$\mathbf{0.622 \pm 0.002}$
DP-SGD	19.132 ± 0.196	0.348 ± 0.003	0.757 ± 0.004	0.236 ± 0.006	0.293 ± 0.005	0.688 ± 0.003
DP-ERM	9.751 ± 0.158	0.292 ± 0.003	0.812 ± 0.003	0.162 ± 0.005	0.223 ± 0.003	0.638 ± 0.002
DP Output Perturbation	24.487 ± 0.147	0.371 ± 0.003	0.732 ± 0.004	0.281 ± 0.006	0.325 ± 0.004	0.710 ± 0.003
DP Objective Perturbation	22.979 ± 0.157	0.345 ± 0.003	0.760 ± 0.003	0.235 ± 0.006	0.289 ± 0.004	0.686 ± 0.003
DP Low-Rank GCE	4.907 ± 0.176	0.283 ± 0.003	0.821 ± 0.003	0.154 ± 0.005	0.212 ± 0.003	0.628 ± 0.002
DP trace-norm BCE	8.655 ± 0.201	0.306 ± 0.002	0.799 ± 0.003	0.182 ± 0.005	0.238 ± 0.003	0.649 ± 0.002
DP Riemannian low-rank BCE	7.104 ± 0.180	0.285 ± 0.003	0.819 ± 0.003	0.152 ± 0.005	0.214 ± 0.003	0.630 ± 0.002
7.5% **contamination**						
DP-RoLL	$\mathbf{2.424 \pm 0.082}$	$\mathbf{0.275 \pm 0.002}$	$\mathbf{0.826 \pm 0.003}$	$\mathbf{0.147 \pm 0.005}$	$\mathbf{0.204 \pm 0.003}$	$\mathbf{0.623 \pm 0.002}$
DP-SGD	21.151 ± 0.213	0.364 ± 0.003	0.742 ± 0.004	0.257 ± 0.007	0.313 ± 0.005	0.701 ± 0.003
DP-ERM	12.604 ± 0.159	0.303 ± 0.003	0.803 ± 0.003	0.172 ± 0.005	0.234 ± 0.003	0.647 ± 0.003
DP Output Perturbation	25.737 ± 0.148	0.386 ± 0.003	0.717 ± 0.004	0.302 ± 0.007	0.345 ± 0.005	0.723 ± 0.003
DP Objective Perturbation	24.411 ± 0.153	0.361 ± 0.003	0.745 ± 0.003	0.253 ± 0.006	0.308 ± 0.004	0.700 ± 0.003
DP Low-Rank GCE	4.902 ± 0.174	0.290 ± 0.003	0.816 ± 0.003	0.159 ± 0.005	0.217 ± 0.003	0.632 ± 0.002
DP trace-norm BCE	11.999 ± 0.222	0.321 ± 0.003	0.786 ± 0.003	0.195 ± 0.005	0.255 ± 0.003	0.661 ± 0.002
DP Riemannian low-rank BCE	10.748 ± 0.213	0.296 ± 0.004	0.811 ± 0.003	0.162 ± 0.005	0.225 ± 0.004	0.638 ± 0.003
10% **contamination**						
DP-RoLL	$\mathbf{2.761 \pm 0.101}$	$\mathbf{0.276 \pm 0.002}$	$\mathbf{0.826 \pm 0.003}$	$\mathbf{0.147 \pm 0.005}$	$\mathbf{0.205 \pm 0.003}$	$\mathbf{0.624 \pm 0.002}$
DP-SGD	23.143 ± 0.217	0.382 ± 0.004	0.725 ± 0.004	0.282 ± 0.007	0.336 ± 0.006	0.716 ± 0.003
DP-ERM	15.593 ± 0.154	0.319 ± 0.004	0.790 ± 0.003	0.185 ± 0.005	0.251 ± 0.004	0.659 ± 0.003
DP Output Perturbation	26.944 ± 0.146	0.403 ± 0.003	0.701 ± 0.004	0.329 ± 0.007	0.367 ± 0.005	0.736 ± 0.003
DP Objective Perturbation	25.800 ± 0.147	0.379 ± 0.003	0.728 ± 0.003	0.277 ± 0.006	0.332 ± 0.004	0.716 ± 0.003
DP Low-Rank GCE	6.433 ± 0.280	0.302 ± 0.003	0.808 ± 0.003	0.169 ± 0.007	0.229 ± 0.004	0.641 ± 0.003
DP trace-norm BCE	15.468 ± 0.237	0.339 ± 0.003	0.770 ± 0.003	0.216 ± 0.006	0.276 ± 0.003	0.677 ± 0.002
DP Riemannian low-rank BCE	14.492 ± 0.271	0.312 ± 0.005	0.798 ± 0.004	0.177 ± 0.006	0.241 ± 0.004	0.650 ± 0.003

C.10 Performance Comparison and Error Scaling with the True Rank

At $\epsilon = 2$, we first compare all methods at true ranks $r^* \in \{2, 4, 6\}$, setting the working rank to the corresponding true value and keeping all other settings fixed. Table C.10 shows that DP-RoLL achieves the best performance across all six metrics at each rank. We then evaluate DP-RoLL over the finer grid $r^* \in \{2, 3, 4, 5, 6\}$, with the number of private iterations $T = 20$ and all other classification settings fixed. As shown in Figure C.5, its estimation error increases linearly with r^*, consistent with the rank dependence predicted by the theoretical error bound.

Table C.10: Classification performance under different true ranks at $\epsilon = 2$, reported as mean $\pm$ standard error over 20 independent replications. Bold values denote the best results.

Method	$\|\hat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F^2 \downarrow$	HL↓	AP↑	OE↓	RL↓	Cov.↓
Rank $r^* = 2$						
DP-RoLL	$\mathbf{2.761 \pm 0.101}$	$\mathbf{0.276 \pm 0.002}$	$\mathbf{0.826 \pm 0.003}$	$\mathbf{0.147 \pm 0.005}$	$\mathbf{0.205 \pm 0.003}$	$\mathbf{0.624 \pm 0.002}$
DP-SGD	23.143 ± 0.217	0.382 ± 0.004	0.725 ± 0.004	0.282 ± 0.007	0.336 ± 0.006	0.716 ± 0.003
DP-ERM	15.593 ± 0.154	0.319 ± 0.004	0.790 ± 0.003	0.185 ± 0.005	0.251 ± 0.004	0.659 ± 0.003
DP Output Perturbation	26.944 ± 0.146	0.403 ± 0.003	0.701 ± 0.004	0.329 ± 0.007	0.367 ± 0.005	0.736 ± 0.003
DP Objective Perturbation	25.800 ± 0.147	0.379 ± 0.003	0.728 ± 0.003	0.277 ± 0.006	0.332 ± 0.004	0.716 ± 0.003
DP Low-Rank GCE	6.433 ± 0.280	0.302 ± 0.003	0.808 ± 0.003	0.169 ± 0.007	0.229 ± 0.004	0.641 ± 0.003
DP trace-norm BCE	15.468 ± 0.237	0.339 ± 0.003	0.770 ± 0.003	0.216 ± 0.006	0.276 ± 0.003	0.677 ± 0.002
DP Riemannian low-rank BCE	14.492 ± 0.271	0.312 ± 0.005	0.798 ± 0.004	0.177 ± 0.006	0.241 ± 0.004	0.650 ± 0.003
Rank $r^* = 4$						
DP-RoLL	$\mathbf{5.948 \pm 0.210}$	$\mathbf{0.214 \pm 0.002}$	$\mathbf{0.884 \pm 0.002}$	$\mathbf{0.069 \pm 0.003}$	$\mathbf{0.131 \pm 0.002}$	$\mathbf{0.563 \pm 0.002}$
DP-SGD	42.299 ± 0.259	0.301 ± 0.002	0.804 ± 0.003	0.162 ± 0.004	0.230 ± 0.003	0.647 ± 0.002
DP-ERM	27.027 ± 0.217	0.229 ± 0.002	0.871 ± 0.002	0.084 ± 0.003	0.147 ± 0.002	0.578 ± 0.002
DP Output Perturbation	51.773 ± 0.218	0.333 ± 0.002	0.771 ± 0.003	0.211 ± 0.005	0.271 ± 0.004	0.677 ± 0.002
DP Objective Perturbation	48.762 ± 0.203	0.292 ± 0.003	0.811 ± 0.003	0.154 ± 0.005	0.222 ± 0.003	0.642 ± 0.002
DP Low-Rank GCE	11.812 ± 0.394	0.235 ± 0.002	0.867 ± 0.002	0.088 ± 0.003	0.152 ± 0.002	0.583 ± 0.002
DP trace-norm BCE	23.631 ± 0.298	0.246 ± 0.003	0.856 ± 0.002	0.098 ± 0.003	0.166 ± 0.003	0.595 ± 0.003
DP Riemannian low-rank BCE	22.756 ± 0.343	0.236 ± 0.002	0.866 ± 0.002	0.088 ± 0.003	0.154 ± 0.002	0.584 ± 0.002
Rank $r^* = 6$						
DP-RoLL	$\mathbf{9.970 \pm 0.247}$	$\mathbf{0.184 \pm 0.001}$	$\mathbf{0.911 \pm 0.001}$	$\mathbf{0.042 \pm 0.001}$	$\mathbf{0.100 \pm 0.001}$	$\mathbf{0.535 \pm 0.002}$
DP-SGD	62.783 ± 0.420	0.261 ± 0.003	0.845 ± 0.003	0.107 ± 0.004	0.180 ± 0.003	0.610 ± 0.003
DP-ERM	40.650 ± 0.230	0.187 ± 0.001	0.909 ± 0.001	0.044 ± 0.001	0.103 ± 0.001	0.538 ± 0.002
DP Output Perturbation	77.645 ± 0.262	0.299 ± 0.003	0.804 ± 0.003	0.160 ± 0.004	0.231 ± 0.003	0.651 ± 0.003
DP Objective Perturbation	74.155 ± 0.219	0.253 ± 0.002	0.849 ± 0.002	0.106 ± 0.003	0.173 ± 0.002	0.604 ± 0.002
DP Low-Rank GCE	17.882 ± 0.624	0.206 ± 0.002	0.893 ± 0.002	0.058 ± 0.002	0.122 ± 0.002	0.557 ± 0.002
DP trace-norm BCE	33.345 ± 0.329	0.202 ± 0.002	0.897 ± 0.001	0.055 ± 0.002	0.117 ± 0.001	0.552 ± 0.002
DP Riemannian low-rank BCE	30.616 ± 0.393	0.201 ± 0.002	0.897 ± 0.001	0.053 ± 0.002	0.116 ± 0.002	0.551 ± 0.002

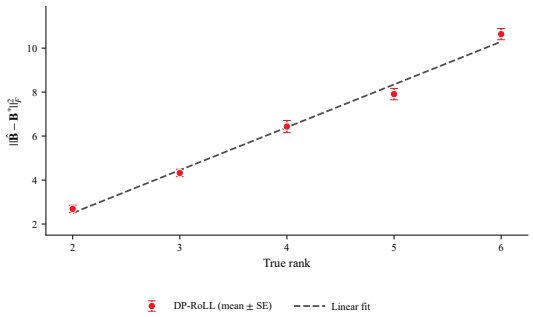

Figure C.5: Estimation error of DP-RoLL as the true rank increases at $\epsilon = 2$. Points and error bars denote means and standard errors over 20 independent replications.

C.11 Sensitivity to Working Rank r and λ_C

At $\epsilon = 2$, we jointly vary the working rank over $r \in \{2, 3, 4, 5\}$ and $\lambda_C \in \{2.4, 2.5, 2.6, 2.7\}$, while keeping the number of private iterations $T = 20$ and all other classification settings fixed. Figure C.6 shows that the estimation error remains stable around the selected configuration.

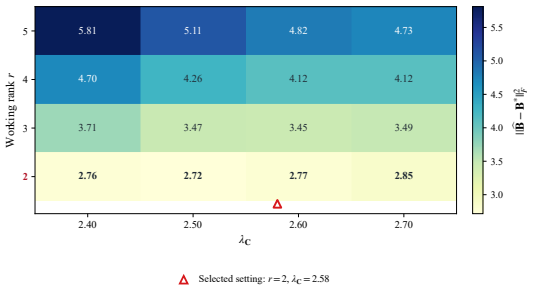

Figure C.6: Estimation error of DP-RoLL across working ranks r and λ_C at $\epsilon = 2$. Each cell reports the mean over 20 replications.

C.12 Public Hyperparameter Calibration

We fix $\epsilon = 2$, $\delta = 10^{-5}$, and $T = 20$, and tune only (r, λ_C). For each replication, we generate independent clean public training and validation sets of sizes 5,000 and 2,000, respectively, according to the mechanism described in Appendix C.1. From each public training set, we construct four contaminated versions with row-contamination proportions $\{2.5\%, 5\%, 7.5\%, 10\%\}$. The contaminated rows and directions are determined from the margins of a pilot logistic model fitted exclusively to the public training data, without using the private sample or B^*.

For each candidate (r, λ_C), we apply the same DP training procedure used in the private experiment to every contaminated public training set. Candidates are evaluated by clean public-validation BCE averaged over the four contamination levels. After selecting the minimizer, the private training data are accessed only once to fit the released model. As shown in Table C.11, public calibration selects the same working rank and a nearby λ_C as the oracle procedure, with negligible loss in estimation and prediction accuracy. The oracle selects hyperparameters by minimizing estimation error using B^* and is included only as a synthetic diagnostic.

Baseline hyperparameter calibration. The same public-calibration protocol is applied to all baseline methods using the same public training and validation sets and validation criterion. For each iterative method, we first select the number of iterations T over a prespecified grid while holding the remaining hyperparameters at their default values. We then fix the selected T and tune the remaining hyperparameters over method-specific candidate grids.

Table C.11: Classification results under different hyperparameter calibration protocols at $\epsilon = 2$. Results are reported as mean $\pm$ standard error over 20 independent replications.

Protocol	r	λ_C	$\|\hat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F^2 \downarrow$	HL$\downarrow$	AP$\uparrow$	OE$\downarrow$	RL$\downarrow$	Cov.$\downarrow$
Oracle	2	2.50	$\mathbf{2.694 \pm 0.098}$	0.276 ± 0.002	0.825 ± 0.003	0.148 ± 0.005	0.205 ± 0.003	0.624 ± 0.002
Public calibration	2	2.61	2.761 ± 0.101	0.276 ± 0.002	0.826 ± 0.003	0.147 ± 0.005	0.205 ± 0.003	0.624 ± 0.002

Real-data hyperparameter calibration implementation. In the real-data experiments, for each dataset, hyperparameters are selected using two prespecified public pilot train–validation splits whose random seeds are distinct from those used in the 20 reported evaluation splits. For clean benchmark datasets, the pilot training set is corrupted using the same protocol as the formal experiment: 10% of its rows undergo cardinality-preserving label mislabeling, while the pilot validation set remains clean. For intrinsically noisy datasets, the observed candidate or partial labels are used for pilot training and the ground-truth labels for validation, without additional corruption.

For DP-RoLL, we first fix

$$r_0 = \min\{4, p, m\}, \qquad \lambda_{\boldsymbol{C},0} = 0.05\sqrt{\frac{m\rho}{10}},$$

and select

$$T \in \{1, 2, 5, 10, 20, 30, 40, 50\}.$$

With T fixed, we then jointly search over

$$r \in \{2, 4, 6\}, \qquad \frac{\lambda_{\boldsymbol{C}}}{\lambda_{\boldsymbol{C},0}} \in \{0.25, 0.5, 1, 2\}.$$

Candidates minimize the average pilot-validation BCE. The selected configuration is subsequently fixed across all 20 reported splits. Each baseline is tuned analogously on the same two public pilot splits using the same validation criterion, with method-specific candidate grids based on literature-recommended defaults.

D Supplementary Regression Experiment

D.1 Regression Data Generation

The predictor matrix $\boldsymbol{X} \in \mathbb{R}^{n_{\text{train}} \times p}$ is constructed using a bounded spherical design, where each row is generated by first drawing $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ and then setting $\boldsymbol{x}_i = \sqrt{p}\boldsymbol{z}_i/\|\boldsymbol{z}_i\|_2$, such that every predictor row satisfies $\|\boldsymbol{x}_i\|_2 = \sqrt{p}$. The true coefficient matrix $\boldsymbol{B}^* \in \mathbb{R}^{p \times m}$ is generated as $\boldsymbol{B}^* = \boldsymbol{U}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{p \times r^*}$ and $\boldsymbol{V} \in \mathbb{R}^{m \times r^*}$ are obtained by orthonormalizing independent standard Gaussian matrices. Outliers are then added exclusively to the training set. Specifically, for the $\lfloor 10\% \times n_{\text{train}} \rfloor$ samples yielding the largest values in the first clean response coordinate z_{i1}^*, we set the corresponding rows of the contamination matrix $\boldsymbol{C}^* = [\boldsymbol{c}_1^*, \ldots, \boldsymbol{c}_{n_{\text{train}}}^*]^\top \in \mathbb{R}^{n_{\text{train}} \times m}$ to $[s_1, \ldots, s_m]$, where$s_j = \text{sd}(\boldsymbol{X}\boldsymbol{b}_j^*)$, where $\text{sd}(\cdot)$ denotes the empirical standard deviation of the entries of its vector argument. The remaining rows of $\boldsymbol{C}^*$ are set to zero. The training response is generated row-wise as $\boldsymbol{y}_i = \boldsymbol{B}^{*\top}\boldsymbol{x}_i + \sqrt{n_{\text{train}}}\boldsymbol{c}_i^* + \boldsymbol{\epsilon}_i$, with random noise $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$. An independent test set is generated from the same bounded spherical design and the same $\boldsymbol{B}^*$, but without contamination. Unless otherwise specified, we set $n_{\text{train}} = 5000$, $n_{\text{test}} = 2000$, $p = 20$, $m = 10$, $r^* = 2$, and the contamination proportion to 10%.

D.2 Regression Baselines

We compare DP-RoLL with standard private regression methods, including DP-SGD (Abadi et al., 2016), DP-ERM (Wang & Xu, 2019), and DP-AdaSSP (Wang, 2018). We also consider private adaptations of two non-private robust or low-rank regression methods, namely DP Huber nuclear–norm regression (Tan et al., 2023) and DP trace-norm regression (Pontil & Maurer, 2013), together with DP Riemannian low-rank regression based on the differentially private Riemannian optimization framework of Han et al. (2024).

Privacy accounting and gradient bounding. DP-SGD uses fixed-size subsampled Rényi differential privacy accounting (Wang et al., 2019). DP-ERM, DP-HNR, DP-TNR, and DP-RLR use Gaussian mechanisms composed over their fixed iteration counts through zCDP. DP-AdaSSP splits both ϵ and δ equally among three internal mechanisms using basic composition. Because squared-loss gradients are unbounded, DP-SGD, DP-ERM, DP-TNR, and DP-RLR clip per-observation gradients at prespecified public radii. DP-HNR requires no additional gradient-norm clipping because the bounded Huber score, together with the public feature-norm bound, controls per-observation gradient sensitivity. For all baselines, only the final coefficient matrix $\boldsymbol{B}$ is released. Rank projections, singular-value thresholding, and retractions applied after privatization constitute post-processing.

DP-SGD and DP-ERM. DP-SGD uses fixed-size mini-batches sampled without replacement and follows the clip–sum–noise–average update of Abadi et al. (2016). DP-ERM performs full-batch clipped noisy gradient descent on the unconstrained squared-loss objective, without a low-rank constraint, following the general DP-ERM framework of Wang & Xu (2019).

DP-AdaSSP. After projecting feature and response rows onto public norm bounds, DP-AdaSSP splits both ϵ and δ equally among three mechanisms: constructing a downward-biased private lower confidence bound for $\lambda_{\min}(\boldsymbol{X}^\top \boldsymbol{X})$ and perturbing the Gram matrix $\boldsymbol{X}^\top \boldsymbol{X}$ and the joint matrix query $\boldsymbol{X}^\top \boldsymbol{Y}$. The adaptive ridge parameter is determined by the private eigenvalue bound and a public upper bound on the spectral norm of the Gaussian Gram-matrix perturbation. Treating $\boldsymbol{X}^\top \boldsymbol{Y}$ as a single query yields the multivariate replace-one adaptation of Wang (2018) used here.

DP Huber nuclear-norm regression (DP-HNR). DP-HNR uses Huber loss with nuclear-norm regularization, corresponding to the dense-coefficient special case of Tan et al. (2023) with the entrywise ℓ_1 penalty set to zero. The bounded Huber score and public feature-norm bound control per-observation gradient sensitivity. Gaussian-perturbed proximal-gradient updates followed by singular-value soft thresholding provide a private adaptation of the source objective.

DP trace-norm regression (DP-TNR). DP-TNR privately adapts trace-norm regularized multitask regression (Pontil & Maurer, 2013). It clips each per-observation squared-loss gradient at a public bound, adds Gaussian noise to the aggregated gradient, and applies singular-value soft thresholding.

DP Riemannian low-rank regression (DP-RLR). DP-RLR adapts the tangent-space gradient-perturbation framework of Han et al. (2024). It projects each per-observation gradient onto the current tangent space, clips the tangent gradient at a public bound, adds Gaussian noise to the aggregate, and applies a truncated SVD retraction to maintain the working rank.

Further implementation details, including the clipping radii, solver tolerances, noise scales, and privacy-accounting parameters, are provided in the supplementary code.

D.3 Comparison with DP Baselines

We fix the regression data-generating process, $\delta = 10^{-5}$ and vary the privacy budget over $\epsilon \in \{0.5, 1, 2, 4\}$, keeping all other settings fixed. To ensure a fair comparison, method-specific hyperparameters are selected from prespecified grids using the public-calibration procedure described in Appendix D.8, with independently generated training and validation data following the same data-generating mechanism as the private experiment. The selected configuration is then fixed over 20 independent replications. The primary evaluation metrics are the estimation error $\|\hat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F^2$ and the test MSE

$$\text{Test MSE} = \frac{1}{n_{\text{test}} m} \left\| \boldsymbol{Y}_{\text{test}} - \boldsymbol{X}_{\text{test}} \hat{\boldsymbol{B}} \right\|_F^2 . \tag{D.1}$$

As shown in Table D.1, DP-RoLL achieves the lowest estimation error and test MSE for every $\epsilon \geq 1$. At $\epsilon = 0.5$, the Riemannian baseline attains slightly lower errors, whereas DP-RoLL remains competitive.

D.4 Performance Comparison and Error Scaling under Increasing Contamination

At $\epsilon = 2$, we first compare all methods at contamination proportions of 5%, 7.5%, and 10%, keeping all other regression settings fixed. Each method is evaluated over 20 independent replications. As shown in Table D.2, DP-RoLL achieves the lowest estimation error and test MSE at every contamination level. We then evaluate DP-RoLL over the finer grid $\{5\%, 6\%, 7\%, 8\%, 9\%, 10\%\}$, with the clipping norm fixed at $C_B = 5$ and the number of private iterations fixed at $T = 20$. As shown in Figure D.1, the estimation error of DP-RoLL increases linearly with the contamination proportion, consistent with the dependence predicted by the theoretical error bound.

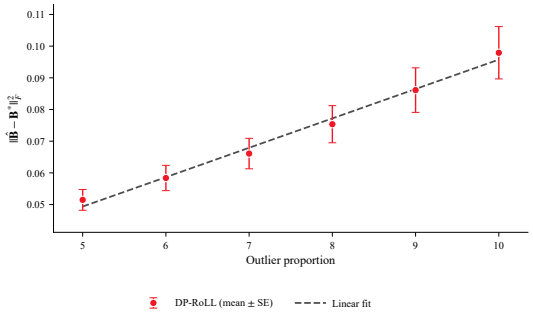

Figure D.1: Estimation error of DP-RoLL under increasing contamination proportions at $\epsilon = 2$. Points and error bars denote means and standard errors over 20 independent replications.

D.5 Performance Comparison and Error Scaling with the True Rank

At $\epsilon = 2$, we first compare all methods at true ranks $r^* \in \{2, 4, 6\}$, keeping all other regression settings fixed. Each method is evaluated over 20 independent replications. As shown in Table D.3, DP-RoLL achieves the lowest estimation error and test MSE at every considered rank. We then evaluate DP-RoLL over the finer grid $r^* \in \{2, 3, \ldots, 8\}$, with the clipping norm fixed at $C_B = 5$ and the number of private iterations fixed at $T = 20$. As shown in Figure D.2, the estimation error of DP-RoLL increases linearly with r^*, consistent with the rank dependence predicted by the theoretical error bound.

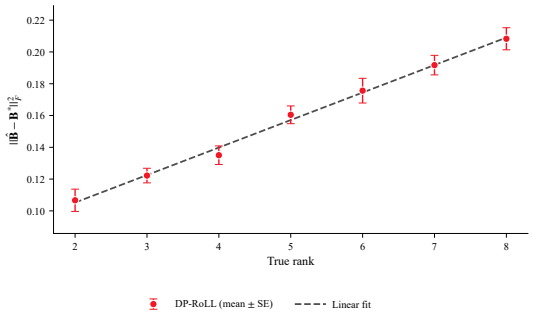

Figure D.2: Estimation error of DP-RoLL as the true rank increases at $\epsilon = 2$. Points and error bars denote means and standard errors over 20 independent replications. Bold values denote the best results.

D.6 Hyperparameter Sensitivity

At $\epsilon = 2$, we jointly vary the working rank and λ_C while fixing $C_B = 5$ and $T = 20$. All other regression settings are kept fixed. As shown in Figure D.3, the estimation error remains stable over a broad range of working ranks and λ_C values around the selected configuration.

Table D.1: Regression comparison with DP baselines across privacy budgets. Results are reported as mean ± standard error over 20 independent replications. Bold values denote the best results.

$\epsilon = 0.5$			$\epsilon = 1$		
Method	Estimation error	Test MSE	Method	Estimation error	Test MSE
DP-RoLL	0.468±0.036	1.042±0.004	DP-RoLL	**0.134±0.006**	**1.010±0.002**
DP-SGD	1.775±0.021	1.174±0.003	DP-SGD	1.346±0.031	1.131±0.004
DP-ERM	0.920±0.049	1.089±0.006	DP-ERM	0.608±0.041	1.057±0.005
DP-AdaSSP	1.624±0.018	1.159±0.003	DP-AdaSSP	1.201±0.030	1.117±0.004
DP-HNR	1.054±0.037	1.102±0.005	DP-HNR	0.834±0.037	1.080±0.005
DP-TNR	0.857±0.042	1.082±0.005	DP-TNR	0.582±0.039	1.055±0.005
DP-RLR	**0.405±0.033**	**1.036±0.004**	DP-RLR	0.328±0.032	1.029±0.004
$\epsilon = 2$			$\epsilon = 4$		
Method	Estimation error	Test MSE	Method	Estimation error	Test MSE
DP-RoLL	**0.049±0.002**	**1.001±0.002**	DP-RoLL	**0.025±0.001**	**0.999±0.002**
DP-SGD	0.750±0.040	1.071±0.005	DP-SGD	0.562±0.038	1.053±0.005
DP-ERM	0.522±0.038	1.049±0.005	DP-ERM	0.487±0.037	1.045±0.005
DP-AdaSSP	0.803±0.039	1.077±0.005	DP-AdaSSP	0.733±0.038	1.070±0.005
DP-HNR	0.709±0.037	1.068±0.005	DP-HNR	0.668±0.037	1.064±0.005
DP-TNR	0.494±0.039	1.046±0.005	DP-TNR	0.475±0.039	1.044±0.005
DP-RLR	0.309±0.032	1.027±0.004	DP-RLR	0.304±0.032	1.027±0.004

Table D.2: Regression performance under different contamination proportions at $\epsilon = 2$. Results are reported as mean ± standard error over 20 independent replications. Bold values denote the best results.

5% contamination			**7.5% contamination**			**10% contamination**		
Method	Estimation error	Test MSE	Method	Estimation error	Test MSE	Method	Estimation error	Test MSE
DP-RoLL	**0.042±0.002**	**1.001±0.002**	DP-RoLL	**0.045±0.002**	**1.001±0.002**	DP-RoLL	**0.049±0.002**	**1.001±0.002**
DP-SGD	0.521±0.018	1.048±0.003	DP-SGD	0.625±0.028	1.059±0.004	DP-SGD	0.750±0.040	1.071±0.005
DP-ERM	0.281±0.017	1.025±0.003	DP-ERM	0.389±0.027	1.035±0.004	DP-ERM	0.522±0.038	1.049±0.005
DP-AdaSSP	0.651±0.020	1.062±0.003	DP-AdaSSP	0.726±0.029	1.069±0.004	DP-AdaSSP	0.803±0.039	1.077±0.005
DP-HNR	0.339±0.018	1.031±0.003	DP-HNR	0.501±0.027	1.047±0.004	DP-HNR	0.709±0.037	1.068±0.005
DP-TNR	0.257±0.018	1.022±0.003	DP-TNR	0.363±0.028	1.033±0.004	DP-TNR	0.494±0.039	1.046±0.005
DP-RLR	0.148±0.014	1.011±0.002	DP-RLR	0.223±0.023	1.019±0.003	DP-RLR	0.309±0.032	1.027±0.004

Table D.3: Regression performance under different true ranks at $\epsilon = 2$. Results are reported as mean ± standard error over 20 independent replications.

$r^* = 2$			$r^* = 4$			$r^* = 6$		
Method	Estimation error	Test MSE	Method	Estimation error	Test MSE	Method	Estimation error	Test MSE
DP-RoLL	**0.049±0.002**	**1.001±0.002**	DP-RoLL	**0.085±0.003**	**1.005±0.002**	DP-RoLL	**0.120±0.003**	**1.009±0.002**
DP-SGD	0.750±0.040	1.071±0.005	DP-SGD	0.861±0.031	1.083±0.003	DP-SGD	1.154±0.026	1.111±0.004
DP-ERM	0.522±0.038	1.049±0.005	DP-ERM	0.597±0.027	1.056±0.003	DP-ERM	0.805±0.022	1.076±0.003
DP-AdaSSP	0.803±0.039	1.077±0.005	DP-AdaSSP	1.292±0.025	1.125±0.003	DP-AdaSSP	2.017±0.027	1.196±0.004
DP-HNR	0.709±0.037	1.068±0.005	DP-HNR	0.835±0.032	1.079±0.003	DP-HNR	1.166±0.029	1.111±0.004
DP-TNR	0.494±0.039	1.046±0.005	DP-TNR	0.600±0.027	1.056±0.003	DP-TNR	0.859±0.022	1.081±0.003
DP-RLR	0.309±0.032	1.027±0.004	DP-RLR	0.498±0.028	1.046±0.003	DP-RLR	0.868±0.022	1.082±0.003

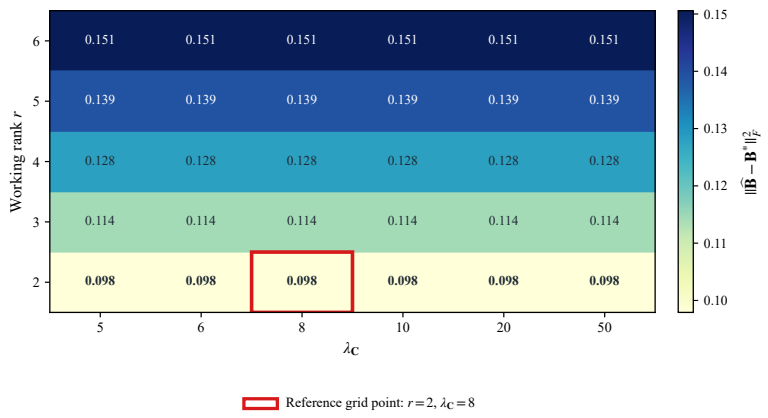

Figure D.3: Sensitivity of DP-RoLL to the working rank and λ_C at $\epsilon = 2$. The red box marks the selected configuration. Each cell reports the mean over 20 replications.

D.7 Sensitivity to Private Iterations and Clipping Radius

For each $\epsilon \in \{1, 2, 4\}$, we first vary the number of private iterations with the clipping radius fixed at $C_B = 5$, and then vary the clipping radius with $T = 20$, keeping all other regression settings fixed. Results are shown in Figure D.4. Moderate iteration counts yield the lowest estimation error, reflecting the trade-off between optimization progress and stronger per-iteration perturbation under privacy composition. Performance is also stable over a practical range of clipping radius, whereas overly small values introduce clipping bias and overly large values increase privacy noise.

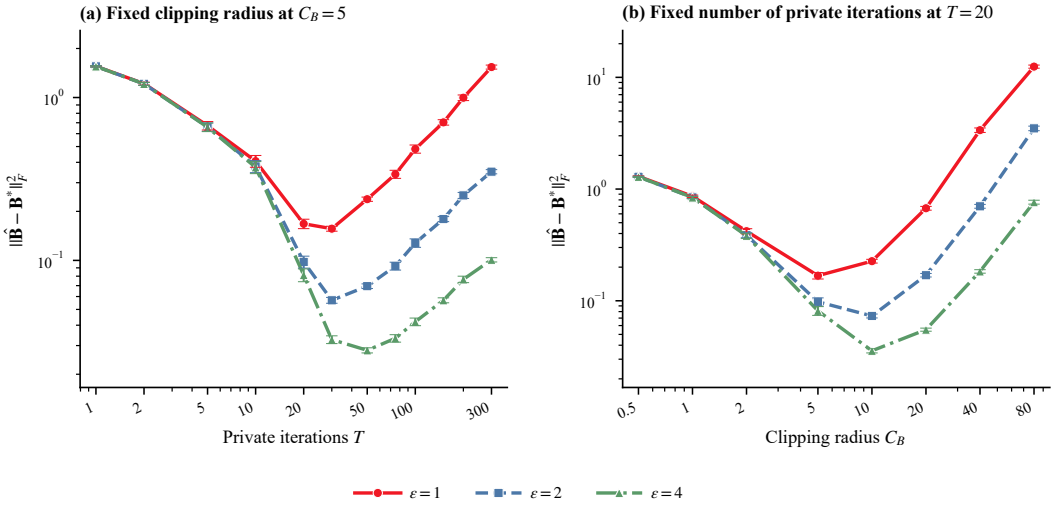

Figure D.4: Estimation error of DP-RoLL under varying numbers of private iterations and clipping radius across different privacy budgets. Points and error bars denote means and standard errors over 20 independent replications.

D.8 Public Hyperparameter Calibration

We fix $\epsilon = 2$, $\delta = 10^{-5}$, and tune only (r, λ_C) while keeping all other regression settings fixed. For each replication, independent clean public training and validation sets of sizes 5,000 and 2,000, respectively, are generated according to the mechanism described in Appendix D.1. From each public training set, we construct four contaminated versions with contamination proportions $\{2.5\%, 5\%, 7.5\%, 10\%\}$ by selecting rows with the largest standardized response norms and adding

response-scale contamination. This procedure uses only public observations and does not require $\boldsymbol{B}^*$.

Each candidate $(r, \lambda_{\boldsymbol{C}})$ is trained using the same DP procedure as in the private experiment and evaluated by clean public-validation MSE averaged over the four contamination levels. The selected parameters are then fixed, and the private training data are accessed only once.

Table D.4: Regression performance under oracle and public hyperparameter calibration at $\epsilon = 2$. Results are reported as mean $\pm$ standard error over 20 independent replications.

Protocol	r	$\lambda_{\boldsymbol{C}}$	$\|\hat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F^2 \downarrow$	Test MSE$\downarrow$
Oracle	2	5.00	0.049 ± 0.002	1.001 ± 0.002
Public calibration	2	5.00	0.049 ± 0.002	1.001 ± 0.002

As shown in Table D.4, public calibration selects $(r, \lambda_{\boldsymbol{C}}) = (2, 5)$, matching the oracle choice, and achieves the same estimation error and test MSE. The oracle minimizes estimation error using $\boldsymbol{B}^*$ and is included only as a synthetic diagnostic.

Baseline hyperparameter calibration. The same public-calibration protocol is applied to all baseline methods using the same public training and validation sets and validation criterion. For each iterative method, we first select the number of iterations T over a prespecified grid while holding the remaining hyperparameters at their default values. We then fix the selected T and tune the remaining hyperparameters over method-specific candidate grids.

E Computing environment

All experiments were conducted on a personal computer running 64-bit Windows 11 Home (Version 21H2), equipped with an AMD Ryzen 7 5800U CPU (8 cores and 16 threads) and 16 GB of memory. All computations were performed on the CPU. The implementation was developed in Python 3.13 using NumPy (≥ 1.26), SciPy (≥ 1.11), and Matplotlib (≥ 3.8).