RoLL: Robust Low-Rank Learning via Nesterov Momentum

Zhaojun Hu
Center for Applied Statistics
and School of Statistics,
Renmin University of China
Beijing, China
huzhaojun@ruc.edu.cn

Wenchen Liu
School of Statistics and Mathematics,
Shanghai Lixin University
of Accounting and Finance
Shanghai, China
liuwenchen@lixin.edu.cn

Ting Wei
Center for Applied Statistics
and School of Statistics,
Renmin University of China
Beijing, China
weiting1006@ruc.edu.cn

Biao Mei
Center for Applied Statistics
and School of Statistics,
Renmin University of China
Beijing, China
2022103705@ruc.edu.cn

Yifan Sun*
Center for Applied Statistics and School of Statistics,
Renmin University of China
Beijing, China
sunyifan@ruc.edu.cn

Abstract

In modern multivariate problems, exploiting the low-rank structure of the coefficient matrix serves as a powerful dimension reduction strategy to enhance statistical efficiency and interpretability, but its effectiveness can be severely compromised by outliers. In this work, we propose a robust low-rank learning (RoLL) framework for a broad range of loss functions, implemented via a fast algorithm that exploits local restricted strong convexity. Theoretically, we establish non-asymptotic error bounds for the obtained fixed-point estimator (not necessarily the global or local optimum), showing that it achieves the minimax optimal rate. Furthermore, we develop a novel information criterion with finite-sample theoretical guarantees for model selection. Extensive experiments demonstrate the effectiveness of our proposed framework in the presence of data anomalies.

1 Introduction

Multivariate response data is ubiquitous in modern statistical applications, ranging from gene expression networks [77] in genomics to asset return modeling [48] in finance. To capture the dependence of responses on 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

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

*Corresponding author.

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

Here, $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m]$ contains pm parameters. As p or m increases, the number of unknowns often far exceeds the sample size n, rendering the above problem ill-posed. Consequently, dimension reduction is essential. A powerful strategy for multivariate learning is to restrict the rank of $\boldsymbol{B}$, exploiting the assumption that the m responses are driven by a smaller number of shared latent factors. This low-rank paradigm has been widely used in multivariate modeling and related problems, including compressed sensing [15], image recovery [30], and sensor localization [39]. More broadly, low-rank parameterizations have also been widely used in parameter-efficient fine-tuning of language models [21, 28]. In this work, we focus specifically on low-rank learning for multivariate response models.

In real-data applications, outliers are bound to occur and may severely influence model fitting. Critically, such contamination often distorts the estimated low-rank structure, leading to wrong conclusions. For example, in the context of reduced-rank regression [24, 71], [51] demonstrated that even a single outlier can severely distort the low-rank matrix estimation. To enhance the robustness of the model fitting, one possible fix is to modify the loss in certain special cases [54, 10, 31]. For example, in multivariate regression, [55] proposed substituting the quadratic loss with Huber's loss. However, this remedy is difficult to extend to a loss function with more complicated expressions than the quadratic form [3], such as logistic deviance.

In addition to these methodological limitations, computation remains a significant bottleneck. Designing scalable algorithms that simultaneously achieve outlier robustness and enforce low-rank structures is a critical challenge. Traditional high-breakdown methods, such as Least Trimmed Squares [46] and Trimmed Likelihood Estimator [26], rely on subset sampling [47]. These approaches are both computationally prohibitive due to the high dimension of the coefficient matrix and are ill-equipped for low-rank learning tasks.

Moreover, existing studies of robust low-rank learning largely focus on regression [55]. Rigorous theoretical analysis is absent even for the logistic deviance, let alone for a broader class of loss functions. In particular, establishing statistical guarantees for practical algorithmic solutions in this context, which may only be *fixed* points, represents a critical void in the literature. Furthermore, how to derive a theoretically sound model selection criterion in this setting remains an open question.

This paper attempts to address the above challenges by proposing a unified framework for robust low-rank learning under a broad spectrum of loss functions. Our main **contributions** are as follows:

(1) A robust low-rank learning (RoLL) framework is introduced to robustify a wide class of loss functions. By utilizing a doubly penalized formulation, this approach simultaneously exploits the underlying low-rank structure and accommodates outliers.

(2) A scalable joint optimization algorithm is first developed. To address the high dimensionality inherent in matrix optimization, a novel momentum-based acceleration scheme is introduced, reducing iteration complexity by adapting to varying degrees of local restricted strong convexity.

(3) A rigorous non-asymptotic analysis is provided for the fixed points yielded by both the unaccelerated and the accelerated algorithms, demonstrating that they enjoy minimax rate optimality. Furthermore, a novel information criterion with theoretical support is proposed for model selection.

Notations. We use $a \wedge b$ to denote $\min(a, b)$ and e for the Euler constant. For a vector $\boldsymbol{v} \in \mathbb{R}^p$, $\|\boldsymbol{v}\|_2$ denotes its ℓ_2 norm. Let $\boldsymbol{I}$ denote the identity matrix. Given a matrix $\boldsymbol{A} = [a_{ij}] = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]^\top$, where a_{ij} denotes the (i, j)-th entry, we define its Frobenius norm as $\|\boldsymbol{A}\|_F$, its rank as $r(\boldsymbol{A})$, and the inner product with another matrix $\boldsymbol{B}$ as $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B})$. The standard vectorization operator is denoted by $\mathrm{vec}(\cdot)$. Let $P_H(t; \lambda) = (-t^2/2 + \lambda|t|)1_{|t|<\lambda} + (\lambda^2/2)1_{|t|\geq\lambda}$, where $1_{|t|<\lambda}$ is the indicator function, and define $P_{2,H}(\boldsymbol{A}; \lambda) = \sum_{i=1}^n P_H(\|\boldsymbol{a}_i\|_2; \lambda)$. The proximal operator associated with a function $h(\cdot)$ is defined as $\mathrm{prox}_{h(\cdot)}(\boldsymbol{A}) := \arg\min_{\boldsymbol{B}} \frac{1}{2}\|\boldsymbol{A} - \boldsymbol{B}\|_F^2 + h(\boldsymbol{B})$. The notation $a \lesssim b$ (resp. $a \gtrsim b$) indicates that $a \leq cb$ (resp. $a \geq cb$) for some universal positive constant c, while $a \asymp b$ means that $a \lesssim b$ and $a \gtrsim b$.

2 Related Work

Low-rank Modeling. The low-rank assumption is widely adopted in multivariate modeling to reduce the effective number of model parameters and enhance statistical performance [36, 9, 79]. Reduced-rank regression [1] achieves this by restricting the rank of the coefficient matrix; see [58] for a recent

overview. Beyond linear models, low-rank structures have been incorporated into generalized linear models, such as multivariate logistic regression [42]. The nuclear norm and Schatten p-norms are also commonly employed to promote low-rank structure in the coefficient matrix [40, 65]. However, outliers are common in real-world data and can significantly compromise the model fitting and mask its underlying low-rank structure [18].

Robust Matrix Learning. A standard method to gain robustness is to modify the loss. For example, in the context of multivariate regression, many studies proposed replacing the standard quadratic loss with robust loss functions [13, 55, 31], such as Huber's loss and Tukey's biweight loss [61]. However, this approach depends strongly on the specific form of the loss. Another important line of research focuses on the mean shift model [52], which handles outliers by introducing a sparse mean-shift parameter. Though this idea has proven effective in regression [51] and factor analysis [78], establishing a unified framework for robust matrix learning that offers both rigorous theoretical guarantees and efficient computation for diverse forms of the loss remains an open problem.

Noisy Multi-label Classification. As noisy multi-label classification [17] serves as a representative application for our proposed framework, we briefly review recent advancements in this domain. Recent advances in noisy multi-label classification have explored various strategies to mitigate label corruption. Some methods focus on correcting the labels, including label enhancement [8, 43], noise filtering [68], and extraction of label-specific representations [80]. In contrast, other approaches model the noise process, ranging from class-conditional noise [67] to instance-dependent [64, 66] and mixed noise settings [76]. However, these methods are typically tailored to specific noise mechanisms, limiting their applicability in real-world scenarios with unknown corruption patterns. A few studies have explored using the low-rank structure of the coefficient matrix to capture label correlations under data anomalies [12, 34], yet systematic investigations remain lacking.

3 RoLL: Robust Low-Rank Learning

To motivate our proposed approach, we revisit the mean shift model commonly employed in robust multivariate regression [51]: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B}^* + \boldsymbol{C}^* + \boldsymbol{E}$, where $\boldsymbol{B}^*$ represents the true coefficient matrix, $\boldsymbol{C}^* = [\boldsymbol{c}_1^*, \boldsymbol{c}_2^*, \ldots, \boldsymbol{c}_n^*]^\top \in \mathbb{R}^{n\times m}$ accounts for the outlying effects on $\boldsymbol{Y}$, and $\boldsymbol{E}$ denotes the random noise with rows independently and identically distributed following $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. The outlying effects can be element-wise in $\boldsymbol{C}$. However, outlying observations are more frequently encountered in robust statistics [27, 22]. Therefore, one can enforce the row-wise sparsity of $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n]^\top$ to propose the following problem:

$$\min_{(\boldsymbol{B},\boldsymbol{C})} \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{\eta}\|_F^2 + \sum_{i=1}^{n} P_{\boldsymbol{C}}(\|\boldsymbol{c}_i\|_2; \lambda_{\boldsymbol{C}}) \text{ s.t. } \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{C}, \tag{2}$$

where $P_{\boldsymbol{C}}(\cdot; \lambda_{\boldsymbol{C}})$ is a sparsity-inducing penalty with $\lambda_{\boldsymbol{C}}$ controlling the shrinkage.

This motivates us to incorporate $\boldsymbol{C}$ into (1) by redefining $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{C}$. To encourage the low-rank structure on $\boldsymbol{B}$, we further penalize the singular values of $\boldsymbol{B}$ and introduce the following *robust low-rank learning* (RoLL) framework:

$$\min_{(\boldsymbol{B},\boldsymbol{C})} l(\boldsymbol{\eta}; \boldsymbol{Y}) + \sum_{i=1}^{p\wedge m} P_{\boldsymbol{B}}(\sigma_i^{(\boldsymbol{B})}; \lambda_{\boldsymbol{B}}) + \sum_{i=1}^{n} P_{\boldsymbol{C}}(\|\boldsymbol{c}_i\|_2; \lambda_{\boldsymbol{C}}) \text{ s.t. } \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{C}, \tag{3}$$

where $\sigma_i^{(\boldsymbol{B})}$ is the i-th largest singular value of $\boldsymbol{B}$, and $P_{\boldsymbol{B}}, P_{\boldsymbol{C}}$ are sparsity-inducing regularizers.

To better illustrate the intuition of our robustification scheme, we consider the problem of multi-label classification as an example. In this setting, the optimization problem in (3) becomes $\min_{(\boldsymbol{B},\boldsymbol{C})} \sum_{i=1}^{n}\sum_{j=1}^{m}[\log\big(1+\exp(\boldsymbol{x}_i^\top \boldsymbol{b}_j + c_{ij})\big) - y_{ij}(\boldsymbol{x}_i^\top \boldsymbol{b}_j + c_{ij})] + \sum_{i=1}^{p\wedge m} P_{\boldsymbol{B}}\big(\sigma_i^{(\boldsymbol{B})}; \lambda_{\boldsymbol{B}}\big) + \sum_{i=1}^{n} P_{\boldsymbol{C}}\big(\|\boldsymbol{c}_i\|_2; \lambda_{\boldsymbol{C}}\big)$. Specifically, imposing a low-rank structure on $\boldsymbol{B}$ captures intrinsic label dependencies, such as drugs targeting identical pathways [42]. Simultaneously, $\boldsymbol{C}$ is introduced to explicitly accommodate outliers. By enforcing a row-sparse structure, $\boldsymbol{C}$ effectively absorbs large discrepancies arising from corrupted observations, thereby reducing their contribution to the loss. This helps prevent the outliers from distorting the approximation of the low-rank $\boldsymbol{B}^*$.

Direct minimization of (3) is challenging due to the potentially non-smooth penalties $P_{\boldsymbol{B}}$, $P_{\boldsymbol{C}}$. Hence, we develop a prototype algorithm based on the Majorization-minimization (MM) principle [23]. For

notational brevity, let $P^{\mathrm{s}}_{\boldsymbol{B}}(\boldsymbol{B};\lambda_{\boldsymbol{B}}) := \sum_{i=1}^{p\wedge m} P_{\boldsymbol{B}}(\sigma_i^{(\boldsymbol{B})};\lambda_{\boldsymbol{B}})$, $P^{\mathrm{s}}_{\boldsymbol{C}}(\boldsymbol{C};\lambda_{\boldsymbol{C}}) := \sum_{i=1}^{n} P_{\boldsymbol{C}}(\|\boldsymbol{c}_i\|_2;\lambda_{\boldsymbol{C}})$. By defining $\bar{\boldsymbol{X}} := [\boldsymbol{X}\ \boldsymbol{I}]$ and $\bar{\boldsymbol{B}} := [\boldsymbol{B}^\top\ \boldsymbol{C}^\top]^\top$, we reformulate (3) as the problem of minimizing $F(\bar{\boldsymbol{B}}) := f(\bar{\boldsymbol{B}}) + P(\bar{\boldsymbol{B}})$, where $f(\bar{\boldsymbol{B}}) := l(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}};\boldsymbol{Y})$ and $P(\bar{\boldsymbol{B}}) := P^{\mathrm{s}}_{\boldsymbol{B}}(\boldsymbol{B};\lambda_{\boldsymbol{B}}) + P^{\mathrm{s}}_{\boldsymbol{C}}(\boldsymbol{C};\lambda_{\boldsymbol{C}})$. Then, we construct a surrogate function by joint linearization

$$g(\bar{\boldsymbol{B}};\bar{\boldsymbol{B}}^-) := f(\bar{\boldsymbol{B}}^-) + \langle \nabla f(\bar{\boldsymbol{B}}^-), \bar{\boldsymbol{B}} - \bar{\boldsymbol{B}}^-\rangle + \frac{\rho}{2}\|\bar{\boldsymbol{B}} - \bar{\boldsymbol{B}}^-\|_F^2 + P(\bar{\boldsymbol{B}}), \tag{4}$$

where ρ is the inverse step size parameter to be chosen later. Then, starting at an arbitrary point $\bar{\boldsymbol{B}}_0 = [\boldsymbol{B}_0^\top\ \boldsymbol{C}_0^\top]^\top$, we generate a sequence of iterates according to

$$\bar{\boldsymbol{B}}_{t+1} = \arg\min_{\bar{\boldsymbol{B}}} g(\bar{\boldsymbol{B}};\bar{\boldsymbol{B}}_t), t \geq 0. \tag{5}$$

Noticing the separability in $\boldsymbol{B}$ and $\boldsymbol{C}$, the resultant update of the MM algorithm is given by

$$\boldsymbol{B}_{t+1} = \mathrm{prox}_{P^{\mathrm{s}}_{\boldsymbol{B}}(\cdot;\lambda_{\boldsymbol{B}})/\rho}(\boldsymbol{B}_t - \boldsymbol{X}^\top \nabla l(\boldsymbol{X}\boldsymbol{B}_t + \boldsymbol{C}_t;\boldsymbol{Y})/\rho), \tag{6a}$$

$$\boldsymbol{C}_{t+1} = \mathrm{prox}_{P^{\mathrm{s}}_{\boldsymbol{C}}(\cdot;\lambda_{\boldsymbol{C}})/\rho}(\boldsymbol{C}_t - \nabla l(\boldsymbol{X}\boldsymbol{B}_t + \boldsymbol{C}_t;\boldsymbol{Y})/\rho). \tag{6b}$$

Assuming that l is $\mathcal{L}$-smooth with respect to $\bar{\boldsymbol{B}}$ (cf. Definition A.1), the convergence of the prototype algorithm is guaranteed provided that ρ is chosen sufficiently large, e.g., $\rho \geq \mathcal{L}$. The Lipschitz continuous gradient condition, satisfied by many losses such as logistic loss and Huber's loss, is used only to determine a universal step size. In practice, a backtracking line search [6] is often employed to properly select a step size at iteration t, ensuring that the majorization condition $g(\bar{\boldsymbol{B}}_{t+1};\bar{\boldsymbol{B}}_t) \geq F(\bar{\boldsymbol{B}}_{t+1})$ is satisfied.

4 Theoretical Analysis

A major challenge in the theoretical analysis lies in the potential absence of global optimality guarantees for the estimator produced by the algorithm, which stands in sharp contrast to the analysis of traditional M-estimators [63]. Specifically, the obtained solutions from the MM algorithm (6) can be characterized as the fixed points of the surrogate function (4):

$$\hat{\bar{\boldsymbol{B}}} \in \arg\min_{\bar{\boldsymbol{B}}} g(\bar{\boldsymbol{B}};\hat{\bar{\boldsymbol{B}}}). \tag{7}$$

We term such estimators *F-estimators* due to their fixed nature.

4.1 Statistical Non-asymptotic Analysis

Our main tool to handle the diverse forms of the loss is the *Generalized Bregman Function* (GBF) [53], which is built upon the following notion of the one-sided directional derivative.

Definition 4.1. *(One-sided Directional Derivative)* The one-sided directional derivative of a function $\psi : D \subset \mathbb{R}^{n_1\times n_2} \to \mathbb{R}$ at $\boldsymbol{\beta} \in D$ with increment $\boldsymbol{h}$ is defined as

$$\delta\psi(\boldsymbol{\beta};\boldsymbol{h}) := \lim_{\epsilon\to 0+} \frac{\psi(\boldsymbol{\beta}+\epsilon\boldsymbol{h}) - \psi(\boldsymbol{\beta})}{\epsilon}, \tag{8}$$

provided $\boldsymbol{h}$ is admissible in the sense that $\boldsymbol{\beta}+\epsilon\boldsymbol{h} \in D$ for sufficiently small $\epsilon : 0 < \epsilon < \epsilon_0$. ψ is said to be one-sided directionally differentiable at $\boldsymbol{\beta}$ if $\delta\psi(\boldsymbol{\beta};\boldsymbol{h})$ exists and is finite for all admissible $\boldsymbol{h}$. Furthermore, if this holds for all $\boldsymbol{\beta} \in D$, ψ is called directionally differentiable.

Definition 4.2. *(Generalized Bregman Function)* The generalized Bregman function associated with a function ψ is defined by

$$\boldsymbol{\Delta}_\psi(\boldsymbol{\alpha},\boldsymbol{\beta}) := \psi(\boldsymbol{\alpha}) - \psi(\boldsymbol{\beta}) - \delta\psi(\boldsymbol{\beta};\boldsymbol{\alpha}-\boldsymbol{\beta}), \tag{9}$$

assuming $\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathrm{dom}(\psi)$ and $\delta\psi(\boldsymbol{\beta};\boldsymbol{\alpha}-\boldsymbol{\beta})$ is finite.

In particular, when ψ is differentiable and strictly convex, the generalized Bregman function $\boldsymbol{\Delta}_\psi$ becomes the standard Bregman divergence [7]: $\mathbf{D}_\psi(\boldsymbol{\alpha},\boldsymbol{\beta}) := \psi(\boldsymbol{\alpha}) - \psi(\boldsymbol{\beta}) - \langle\nabla\psi(\boldsymbol{\beta}), \boldsymbol{\alpha}-\boldsymbol{\beta}\rangle$. The Bregman divergence 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

$$\boldsymbol{\mathcal{E}} = -\nabla l(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*), \tag{10}$$

which does not depend on the regularizer. 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

$$\boldsymbol{\mathcal{E}} = \boldsymbol{Y} - g^{-1}(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*) = \boldsymbol{Y} - \mathbb{E}(\boldsymbol{Y}). \tag{11}$$

For 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 (cf. Definition A.3).

To facilitate the theoretical analysis, we first introduce some necessary notations. For any $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n]^\top$, define

$$\mathcal{J}(\boldsymbol{C}) = \{j : \boldsymbol{c}_j \neq \boldsymbol{0}\}, \tag{12}$$

and $J(\boldsymbol{C}) = |\mathcal{J}(\boldsymbol{C})|$ with $|\cdot|$ denoting the cardinality of the enclosed set. Let $q = r(\boldsymbol{X})$. Assume l is $\mathcal{L}$-smooth with respect to $\bar{\boldsymbol{B}}$. Without loss of generality, we take $\mathcal{L} = 1$ and assume $\|\boldsymbol{X}\|_2 \leq 1$. Furthermore, we assume that both $P_{\boldsymbol{B}}(\cdot;\lambda)$ and $P_{\boldsymbol{C}}(\cdot;\lambda)$ are instances of a generic penalty function $P_\Theta(\cdot;\lambda)$ with an associated critical number $\mathcal{L}_\Theta$ (see Appendix A.1 for details) satisfying

$$P_\Theta(0;\lambda) = 0, P_\Theta(t;\lambda) \geq P_H(t;\lambda). \tag{13}$$

This condition is mild and satisfied by several popular regularizers, such as the ℓ_1 penalty [57], SCAD [14], and MCP [73]. The following theorem establishes a non-asymptotic error bound for any F-estimator (7).

Theorem 4.3. *Consider* $\min_{(\boldsymbol{B},\boldsymbol{C})} l(\boldsymbol{X}\boldsymbol{B}+\boldsymbol{C};\boldsymbol{Y}) + \sum_{i=1}^{p\wedge m} P_{\boldsymbol{B}}(\sigma_i^{(\boldsymbol{B})};\lambda_{\boldsymbol{B}}) + \sum_{i=1}^n P_{\boldsymbol{C}}(\|\boldsymbol{c}_i\|_2;\lambda_{\boldsymbol{C}})$, *where* $\lambda_{\boldsymbol{B}} = A\sigma(q+m)^{1/2}$, $\lambda_{\boldsymbol{C}} = A\sigma(m+\log n)^{1/2}$ *with* A *a sufficiently large constant. Then the following inequality holds for any fixed point* $(\hat{\boldsymbol{B}}, \hat{\boldsymbol{C}})$ *defined in* (7)

$$\mathbb{E}[\mathbf{D}_2(\bar{\boldsymbol{X}}\hat{\bar{\boldsymbol{B}}}, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*)] \lesssim \sigma^2\{K_1^2(q+m)r^* + K_2^2(m+\log n)J^* + 1\} \tag{14}$$

under the assumption that there exists $K_1, K_2 \geq 0$ *such that* $\delta\mathbf{D}_2(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*) + \vartheta P_{2,H}(\boldsymbol{C}-\boldsymbol{C}^*;\lambda_{\boldsymbol{C}}) + \vartheta\sum_{i=1}^{p\wedge m} P_H(\sigma_i^{(\boldsymbol{B}-\boldsymbol{B}^*)};\lambda_{\boldsymbol{B}}) + \sum_{i=1}^{p\wedge m} P_{\boldsymbol{B}}(\sigma_i^{(\boldsymbol{B}^*)};\lambda_{\boldsymbol{B}}) + \mathcal{L}_\Theta\mathbf{D}_2(\bar{\boldsymbol{B}}, \bar{\boldsymbol{B}}^*) + \sum_{i=1}^n P_{\boldsymbol{C}}(\|\boldsymbol{c}_i^*\|_2;\lambda_{\boldsymbol{C}}) \leq 2\bar{\boldsymbol{\Delta}}_l(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*) + K_1^2\lambda_{\boldsymbol{B}}^2 r^* + K_2^2\lambda_{\boldsymbol{C}}^2 J^* + \sum_{i=1}^{p\wedge m} P_{\boldsymbol{B}}(\sigma_i^{(\boldsymbol{B})};\lambda_{\boldsymbol{B}}) + \sum_{i=1}^n P_{\boldsymbol{C}}(\|\boldsymbol{c}_i\|_2;\lambda_{\boldsymbol{C}})$ *holds for any* $\boldsymbol{B}, \boldsymbol{C}$, *where* $\mathcal{J} = \mathcal{J}(\boldsymbol{C}^*)$, $J^* = |\mathcal{J}|$, $r^* = r(\boldsymbol{B}^*)$, *and* $\vartheta > 0, \delta > 0$ *are constants.*

The regularity condition in Theorem 4.3 aligns with the standard assumptions widely adopted in Bregman-based statistical analysis [53] and is mild (see Remark A.5 for sufficient conditions and further discussion). In fact, the regularity condition can be relaxed to hold almost surely only at the F-estimator under consideration, since the proof invokes it only at this point. Importantly, if l is strongly convex with respect to $\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}$ (cf. Definition A.2) and $P_{\boldsymbol{B}}(\cdot;\lambda_{\boldsymbol{B}})$ and $P_{\boldsymbol{C}}(\cdot;\lambda_{\boldsymbol{C}})$ are bounded nonconvex penalties satisfying $P_\Theta(\cdot;\lambda) \lesssim \lambda^2$, the error rate in Theorem 4.3 can be established without any regularity condition (see Corollary A.7).

Under additional design conditions, estimation error bounds for both $\hat{\boldsymbol{B}}$ and the joint parameter $\hat{\bar{\boldsymbol{B}}}$ can be established. Specifically, a bound for $\|\hat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F^2$ follows from controlling $P_{2,H}(\hat{\boldsymbol{C}} - \boldsymbol{C}^*;\lambda_{\boldsymbol{C}})$ and imposing a subset strong convexity condition on $\boldsymbol{X}$. An additional upper spectral bound on $\boldsymbol{X}$ further yields a bound for $\|\hat{\bar{\boldsymbol{B}}} - \bar{\boldsymbol{B}}^*\|_F^2$. Formal derivations and sharper guarantees under more specialized design [35, 29] or contamination settings are left for future work.

4.2 Minimax Optimality and Model Selection via Information Criterion

In this part, we show that the error rate obtained in Theorem 4.3 is minimax optimal. We first derive a universal nonasymptotic lower bound satisfied by all estimators under GLMs. Let w be a nondecreasing loss function with $w(0) = 0$ and $w \not\equiv 0$.

Theorem 4.4. *Assume $\boldsymbol{Y}|\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*$ follows a distribution in the regular exponential family with dispersion σ^2 with l the associated negative log-likelihood function* (A.33)*. Define a signal class*

$$(\boldsymbol{B}^*, \boldsymbol{C}^*) \in \mathcal{B}_{r^*,J^*} := \{(\boldsymbol{B}, \boldsymbol{C}) \in \mathbb{R}^{p\times m} \times \mathbb{R}^{n\times m} : r(\boldsymbol{B}) \le r^*, J(\boldsymbol{C}) \le J^*\}, \tag{15}$$

with $n \ge 2$, $1 \le J^* \le n/2$, $1 \le r^* \le q \wedge m$, $(q + m - r^*/2)r^*/2 \ge 8$, and a model complexity function

$$P_o(J^*, r^*) = \sigma^2\{J^*m + J^* \log(en/J^*) + (q + m - r^*)r^*\}. \tag{16}$$

Let the SVD of $\boldsymbol{X}$ be $\boldsymbol{UDV}^\top$ and define $\boldsymbol{Z} = \boldsymbol{UD} \in \mathbb{R}^{n\times q}$. Suppose the following restricted condition-number condition: $\underline{\kappa}\mathbf{D}_2(\boldsymbol{A}_1, \boldsymbol{A}_2) \le \mathbf{D}_2(\boldsymbol{ZA}_1, \boldsymbol{ZA}_2)$ and $\sigma^2\boldsymbol{\Delta}_l(\mathbf{0}, \boldsymbol{ZA}_1) \le \bar{\kappa}\mathbf{D}_2(\mathbf{0}, \boldsymbol{A}_1), \forall r(\boldsymbol{A}_i) \le r^*$, and $\sigma^2\boldsymbol{\Delta}_l(\mathbf{0}, \boldsymbol{C}) \le \kappa_{\boldsymbol{C}}\mathbf{D}_2(\mathbf{0}, \boldsymbol{C})$, where $\underline{\kappa}/\bar{\kappa}$ and $\kappa_{\boldsymbol{C}}$ are positive constants. Then, there exist positive constants c', c, depending on $w(\cdot)$, $\underline{\kappa}/\bar{\kappa}$, and $\kappa_{\boldsymbol{C}}$, such that

$$\inf_{(\hat{\boldsymbol{B}},\hat{\boldsymbol{C}})} \sup_{(\boldsymbol{B}^*,\boldsymbol{C}^*)\in\mathcal{B}_{r^*,J^*}} \mathbb{E}[w(\mathbf{D}_2(\bar{\boldsymbol{X}}\hat{\bar{\boldsymbol{B}}}, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*)/\{c' P_o(J^*, r^*)\})] \ge c > 0, \tag{17}$$

where $(\hat{\boldsymbol{B}}, \hat{\boldsymbol{C}})$ denotes any estimator of $(\boldsymbol{B}^*, \boldsymbol{C}^*)$.

The regularity conditions in Theorem 4.4 are easy to verify using the Hessian. For instance, in the multi-label classification setup, the logistic deviance has a $1/4$-Lipschitz continuous gradient. Consequently, the regularity conditions are satisfied with $\underline{\kappa} = \inf_{\boldsymbol{\Delta}\neq\mathbf{0}, r(\boldsymbol{\Delta})\le 2r^*} \|\boldsymbol{Z\Delta}\|_F^2/\|\boldsymbol{\Delta}\|_F^2$, $\bar{\kappa} = \sup_{\boldsymbol{A}\neq\mathbf{0}, r(\boldsymbol{A})\le r^*} \|\boldsymbol{ZA}\|_F^2/(4\|\boldsymbol{A}\|_F^2)$ and $\kappa_{\boldsymbol{C}} = 1/4$.

To provide a concrete understanding of the resulting bound, we consider a canonical choice of w. Specifically, setting $w(v) = v$ yields $\inf_{(\hat{\boldsymbol{B}},\hat{\boldsymbol{C}})} \sup_{(\boldsymbol{B}^*,\boldsymbol{C}^*)\in\mathcal{B}_{r^*,J^*}} \mathbb{E}[\mathbf{D}_2(\bar{\boldsymbol{X}}\hat{\bar{\boldsymbol{B}}}, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*)] \gtrsim \sigma^2\{J^*m + J^* \log(en/J^*) + (q + m - r^*)r^*\}$. Since $r^* \le q \wedge m$, we have $(q + m - r^*)r^* \asymp (q + m)r^*$, and hence $\inf_{(\hat{\boldsymbol{B}},\hat{\boldsymbol{C}})} \sup_{(\boldsymbol{B}^*,\boldsymbol{C}^*)\in\mathcal{B}_{r^*,J^*}} \mathbb{E}[\mathbf{D}_2(\bar{\boldsymbol{X}}\hat{\bar{\boldsymbol{B}}}, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*)] \gtrsim \sigma^2\{J^*m + J^* \log(en/J^*) + (q+m)r^*\}$. Comparing this with the upper bound in Theorem 4.3, and treating K_1, K_2 as constants, we conclude that our estimator is minimax optimal (up to a mild logarithm factor). Numerical results in Appendix C.3 further support the tightness of this rate.

Despite the above strong theoretical guarantees, selecting the optimal tuning parameters remains a practical challenge. Cross-validation is a standard method for model selection. However, its application in outlier-contaminated settings is non-trivial and lacks theoretical justification. Furthermore, the associated computational burden is often prohibitive.

To address these limitations, we develop an efficient criterion rooted in our theoretical findings. Motivated by the minimax lower bound (Theorem 4.4), we first construct a penalty term $P_m(\boldsymbol{B}, \boldsymbol{C}) = (m+q-r)r + Jm + J \log(en/J)$ (or $P_m(\bar{\boldsymbol{B}})$ for short), where $J = J(\boldsymbol{C})$, $r = r(\boldsymbol{B})$. We adopt the convention that $J \log(en/J) = 0$ when $J = 0$. Building on this, we propose the following *predictive information criterion* (PIC) for parameter tuning

$$l(\boldsymbol{XB} + \boldsymbol{C}; \boldsymbol{Y}) + A\sigma^2 P_m(\boldsymbol{B}, \boldsymbol{C}), \tag{18}$$

where A is a positive constant. Unlike cross-validation, the proposed PIC benefits from rigorous theoretical guarantees. Specifically, minimizing the PIC (18) achieves the minimax optimal error rate, as is shown below.

Theorem 4.5. *Assume $\bar{\boldsymbol{B}}^* \neq \mathbf{0}$, and let $\mathcal{M}$ be the collection of $\bar{\boldsymbol{B}}^*$ alongside all candidate models produced by RoLL over a comprehensive tuning parameter grid. If there exist constants $\delta > 0$ and $K \ge 0$ such that $(\boldsymbol{\Delta}_l - \delta\mathbf{D}_2)(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}') + K\sigma^2(P_m(\bar{\boldsymbol{B}}) + P_m(\bar{\boldsymbol{B}}')) \ge 0, \forall \bar{\boldsymbol{B}}, \bar{\boldsymbol{B}}' \in \mathcal{M}$, then any $\hat{\bar{\boldsymbol{B}}}$ that minimizes* (18) *over the set $\mathcal{M}$ must satisfy the following inequality*

$$\mathbb{E}[\mathbf{D}_2(\bar{\boldsymbol{X}}\hat{\bar{\boldsymbol{B}}}, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*)] \lesssim \sigma^2\{(m + q - r^*)r^* + J^*m + J^* \log(en/J^*)\}. \tag{19}$$

Regarding the practical implementation of PIC, the dispersion parameter σ^2 is typically known in classification setups. In scenarios where σ is unknown, it can be replaced by a preliminary estimate. As for the constant A, we recommend $A = 1$ in (18). Finally, empirical evaluations in Appendix B.7 further confirm the overall efficacy of PIC.

Overall, our unified analysis in this section accommodates a wide class of loss functions, such as those underlying multivariate regression, multi-label, and multi-class classification.

5 Fast Computation via Nesterov Momentum

The MM algorithm (6), though easy to implement, converges slowly when the problem dimension is high (cf. Section 6.2). In this section, we first introduce a new version of Nesterov's first acceleration [37] and establish its convergence theorem. Motivated by the theoretical insights, we then develop a line-search criterion to solve (3) efficiently. Finally, we detail the resulting algorithm, whose key innovation lies in *dynamically* leveraging the *varying* degrees of *local restricted strong convexity*.

Let $\{\rho_t\}_{t\geq 0}$ and $\{\mu_t\}_{t\geq 0}$ denote the positive and nonnegative sequences corresponding to the local restricted strong smoothness and convexity parameters. These sequences are designed to reduce *iteration complexity*. Then, we recursively define a sequence of relaxation parameters $\{\theta_t\}_{t\geq 0}$, initialized at $\theta_0 \in (0, 1]$, by:

$$\frac{\theta_t^2}{1-\theta_t} = \frac{\theta_{t-1}(\rho_{t-1}\theta_{t-1} + \mu_{t-1})}{\rho_t},\ t \geq 1. \tag{20}$$

For simplicity, we assume $\mathcal{L}$ is known, as is standard in our applications. In the standard convex optimization setting where $\rho_t \equiv \mathcal{L}$ and $\mu_t \equiv 0$, (20) simplifies to $\theta_t = (\sqrt{\theta_{t-1}^4 + 4\theta_{t-1}^2} - \theta_{t-1}^2)/2, \forall t \geq 1$, which aligns with the classical Nesterov's construction. However, (20) generalizes the traditional construction to accommodate *time-varying* $\{\mu_t\}$ and $\{\rho_t\}$, enabling the exploitation of *local restricted strong convexity*. This formulation is motivated by the observation that such local parameters often vary along the optimization trajectory [45].

We now introduce an iterative algorithm that generates two sequences of iterates, starting from $\bar{\boldsymbol{M}}_0 = \bar{\boldsymbol{B}}_0$:

$$\bar{\boldsymbol{M}}_t = \bar{\boldsymbol{B}}_t + \frac{\rho_{t-1}\theta_t(1-\theta_{t-1})}{\rho_{t-1}\theta_{t-1} + \mu_{t-1}}(\bar{\boldsymbol{B}}_t - \bar{\boldsymbol{B}}_{t-1}),\ \forall t \geq 1, \tag{21a}$$

$$\bar{\boldsymbol{B}}_{t+1} = \arg\min_{\bar{\boldsymbol{B}}} F(\bar{\boldsymbol{B}}) - \boldsymbol{\Delta}_f(\bar{\boldsymbol{B}}, \bar{\boldsymbol{M}}_t) + \mu_t \mathbf{D}_2(\bar{\boldsymbol{B}}, \bar{\boldsymbol{M}}_t) + \rho_t \mathbf{D}_2(\bar{\boldsymbol{B}}, \bar{\boldsymbol{M}}_t),\ \forall t \geq 0. \tag{21b}$$

Our approach introduces two novel elements. First, the update of $\bar{\boldsymbol{M}}_t = [\boldsymbol{M}_t^{\boldsymbol{B}\top}\ \boldsymbol{M}_t^{\boldsymbol{C}\top}]^\top$ (21a) incorporates ρ_{t-1}. The second one is the addition of $\mu_t \mathbf{D}_2(\bar{\boldsymbol{B}}, \bar{\boldsymbol{M}}_t)$ in (21b), which complements the traditional Bregman term $\rho_t \mathbf{D}_2(\bar{\boldsymbol{B}}, \bar{\boldsymbol{M}}_t)$ when $\mu_t \neq 0$. Since f is $\mathcal{L}$-smooth, $\boldsymbol{\Delta}_f(\bar{\boldsymbol{B}}, \bar{\boldsymbol{M}}_t) \leq \mathcal{L}\mathbf{D}_2(\bar{\boldsymbol{B}}, \bar{\boldsymbol{M}}_t)$. Thus, when $\mu_t + \rho_t = \mathcal{L}$, the objective in (21b) is an MM surrogate that majorizes $F(\bar{\boldsymbol{B}}) = f(\bar{\boldsymbol{B}}) + P(\bar{\boldsymbol{B}})$ and is tight at $\bar{\boldsymbol{M}}_t$. Consequently, if the algorithm terminates at $\bar{\boldsymbol{B}}_{t+1} = \bar{\boldsymbol{M}}_t$, the resulting point satisfies the F-estimator fixed-point condition in (7) and is therefore covered by the statistical guarantees in Theorem 4.3 under its stated conditions.

To provide the theoretical guarantee of the iterative algorithm (21), we present a convergence theorem. Given any function $\psi(\cdot)$, define $\mathbf{C}_\psi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \theta) := \theta\psi(\boldsymbol{\alpha}) + (\mathbf{1}-\theta)\psi(\boldsymbol{\beta}) - \psi(\theta\boldsymbol{\alpha} + (\mathbf{1}-\theta)\boldsymbol{\beta})$. Assume the loss function f is differentiable in $\bar{\boldsymbol{B}}$ (possibly nonconvex), and P is directionally differentiable in $\bar{\boldsymbol{B}}$. Let $\bar{\psi}_t(\cdot) := f(\cdot) - \mu_t\|\cdot\|_F^2/2$ and define

$$\mathcal{E}_t(\bar{\boldsymbol{B}}) = \boldsymbol{\Delta}_{\bar{\psi}_t}(\bar{\boldsymbol{B}}, \bar{\boldsymbol{M}}_t) + (\mathbf{C}_P(\bar{\boldsymbol{B}}, \bar{\boldsymbol{B}}_t, \theta_t) + \boldsymbol{\Delta}_P(\theta_t\bar{\boldsymbol{B}} + (1-\theta_t)\bar{\boldsymbol{B}}_t, \bar{\boldsymbol{B}}_{t+1}))/\theta_t.$$

Furthermore, let

$$R_t = (\rho_t \mathbf{D}_2 - \boldsymbol{\Delta}_{\bar{\psi}_t})(\bar{\boldsymbol{B}}_{t+1}, \bar{\boldsymbol{M}}_t) + (1-\theta_t)\boldsymbol{\Delta}_{\bar{\psi}_t}(\bar{\boldsymbol{B}}_t, \bar{\boldsymbol{M}}_t).$$

Here, R_t is determined by the iterates in (21), while $\mathcal{E}_t(\bar{\boldsymbol{B}})$ involves an additional dependency on $\bar{\boldsymbol{B}}$.

Theorem 5.1. *Suppose that f is differentiable and P is directionally differentiable in $\bar{\boldsymbol{B}}$. For arbitrary sequences $\rho_t > 0$ and $\mu_t \geq 0$, the iterates defined in* (21a)*,* (21b)*, satisfy the following bound for any $\bar{\boldsymbol{B}}$ and all $T \geq 0$:*

$$\begin{aligned} &F(\bar{\boldsymbol{B}}_{T+1}) - F(\bar{\boldsymbol{B}}) + \theta_T(\rho_T\theta_T + \mu_T)\mathbf{D}_2(\bar{\boldsymbol{B}}, \frac{\bar{\boldsymbol{M}}_{T+1} - (1-\theta_{T+1})\bar{\boldsymbol{B}}_{T+1}}{\theta_{T+1}}) \\ &+ \sum_{t=0}^{T}(\prod_{s=t+1}^{T}(1-\theta_s))(R_t + \theta_t\mathcal{E}_t(\bar{\boldsymbol{B}})) \\ \leq &(\prod_{t=1}^{T}(1-\theta_t))[(1-\theta_0)(F(\bar{\boldsymbol{B}}_0) - F(\bar{\boldsymbol{B}})) + \theta_0^2\rho_0\mathbf{D}_2(\bar{\boldsymbol{B}}, \bar{\boldsymbol{B}}_0)], \end{aligned} \tag{22}$$

where by convention, $\prod_{s=l}^{u} a_s = 1$ as $l > u$.

Algorithm 1 Adaptive Line Search for Selecting μ_t

Input: μ_{t-1}, $0 < \underline{c} < 1$, $\bar{c} > 1$, $0 < c < 1$, $M \in \mathbb{N}$
Function LINESEARCH(μ_{t-1})
$\mu_0 \leftarrow \underline{c}\,\mu_{t-1}$ if $\mu_{t-1} > 0$, otherwise $\mu_0 \leftarrow 10^{-3}$
$\mathcal{G} \leftarrow \{\mu_0, \bar{c}\mu_0, \ldots, \bar{c}^{M-1}\mu_0\}$
$W_t^{\max} \leftarrow \max_{\mu \in \mathcal{G}} W_t(\mu)$
$\mu_t \leftarrow \max\left\{\mu \in \mathcal{G} \;\middle|\; \mu < \mathcal{L},\ R_t(\mu) \geq 0,\ W_t(\mu) \in [(1-c)W_t^{\max}, W_t^{\max}]\right\}$ (set $\mu_t = 0$ if the set is empty)
Return: μ_t

The theoretical bound in Theorem 5.1 is tight. As a concrete illustration, under the standard optimization scenario where the loss f is strongly convex and smooth with a convex penalty P [16, 11], it achieves the optimal iteration complexity (cf. Remark A.14). It is worth noting that Theorem 5.1 imposes no conditions on μ_t or ρ_t, allowing these parameters to be *dynamically* varied to improve the algorithm's efficiency. Motivated by this, a line search strategy can be designed to adaptively select μ_t and ρ_t. Specifically, to satisfy the condition $\mu_t + \rho_t = \mathcal{L}$ required to achieve the desired statistical error rate, we enforce $\rho_t = \mathcal{L} - \mu_t$ and adjust μ_t based on μ_{t-1} to maximize $W_t(\mu)$, defined as

$$R_t(\mu) + \theta_t(\mu)\mathcal{E}_t(\bar{\boldsymbol{B}}_t^o; \mu). \tag{23}$$

Here, $\bar{\boldsymbol{B}}_t^o = \arg\min_{0 \leq s \leq t} F(\bar{\boldsymbol{B}}_s)$. To control the computational overhead, the number of such searches per iteration M is limited. Detailed implementation is summarized in Algorithm 1. Based on empirical evaluation, we recommend $\underline{c} = 0.4$, $\bar{c} = 2.8$, $c =$ 1e-3, and $M = 3$.

The adaptive line search is not required at every iteration. During the initial phase (first T_0 iterations, e.g., $T_0 = 50$), we employ the conservative parameters $\mu_t = 0$ and $\rho_t = \mathcal{L}$. Once the threshold is reached ($t \geq T_0$), we initialize μ_{T_0} with a small starting value (e.g., $\mu_{T_0} =$ 1e-3) and engage the line search strategy to find appropriate μ_t and ρ_t. The resulting **R**obust **A**ccelerated **P**roximal **I**terative **D**escent (RAPID) algorithm is provided in Appendix B.1. Furthermore, Appendix D details the key innovations and wider applicability of RAPID as a *generalized acceleration framework*.

6 Numerical Experiments

In this part, to validate the effectiveness of our proposed RoLL framework, we focus our empirical evaluation on noisy multi-label classification problems. This experimental setting is motivated by two considerations: (i) the label correlations are widely known to induce a low-rank structure in the coefficient matrix [20]; and (ii) practical multi-label data often contain corrupted observations due to label noise or systematic mislabeling [76]. To further demonstrate the *generality* of RoLL, we include additional experiments under regression settings in Appendix C. We perform all experiments using a personal computer (AMD Ryzen 7 5800U with integrated Radeon Graphics). [2]

6.1 Performance Comparison on Synthetic Data

Synthetic data are generated from a multivariate logistic model with a low-rank coefficient matrix $\boldsymbol{B}^*$. To mimic real-world data contamination, $O\%$ of the training samples are corrupted (see Appendix B.3 for details). We compare RoLL (implemented via the SCAD penalty for $P_{\boldsymbol{B}}(\cdot; \lambda_{\boldsymbol{B}})$ and $P_{\boldsymbol{C}}(\cdot; \lambda_{\boldsymbol{C}})$) with seven methods, comprising the classical Binary Relevance (BR) baseline (implemented via logistic regression) and recent state-of-the-art approaches: LENFN [8], NLR [68], LSNRLS [80], RIDE [66], LMAP [12], and AUC-RLR [34]. For these comparison methods, the tuning parameters were carefully chosen according to the suggestions in their papers.

We evaluate each method with *Hamming Loss*, *Subset Accuracy*, *Example F1*, and *Label F1* (see [74] for definitions). Figure 1 presents the performance of all methods as the outlier proportion $O\%$ increases from 5% to 10%, with results averaged over 20 independent replications. Overall, RoLL consistently achieves the best performance across all metrics. The performance of the most competitive method, AUC-RLR, degrades significantly as the outlier ratio rises, whereas RoLL

[2] Code is available at `https://github.com/Hypon-Jun/RoLL`.

maintains stable and superior results. For example, as the outlier proportion rises from 5% to 10%, AUC-RLR's subset accuracy drops from 0.67 to 0.42, whereas RoLL maintains accuracy around 0.67. For further evaluations using additional metrics and a more challenging setup, please refer to Appendices B.5 and B.6.

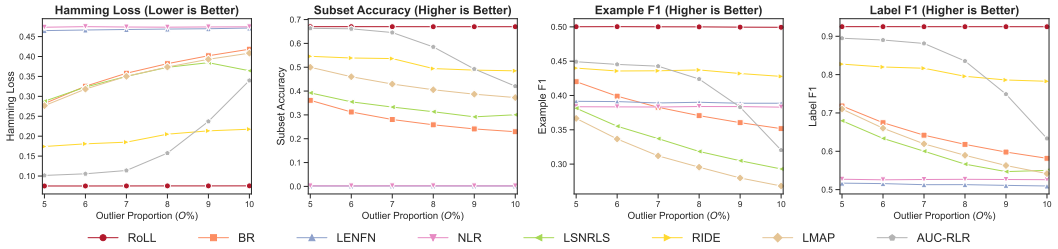


Figure 1: Performance comparison of different methods with outlier proportions ($O\%$) varying from 5% to 10%. RoLL (red line) consistently outperforms baselines across all metrics.

6.2 Power of Acceleration

In this part, we evaluate the computational efficiency of the proposed RAPID algorithm (Algorithm 2) against the MM baseline (6) under three outlier contamination levels: 5%, 8%, and 10%. Figure 2 (left) reports the average running time over 20 independent replications, where the reported time includes the computational overhead of model selection via a grid search over 25 candidate penalty parameters. RAPID consistently reduces the average running time by approximately 90% compared to the MM baseline across these setups. This substantial time reduction is attributed to RAPID's lower iteration complexity. As illustrated in Figure 2 (right), when evaluated at the optimal λ_B and λ_C (tuned by PIC) under 5% contamination, RAPID requires significantly fewer iterations to reach a 90% target accuracy compared to the MM baseline.

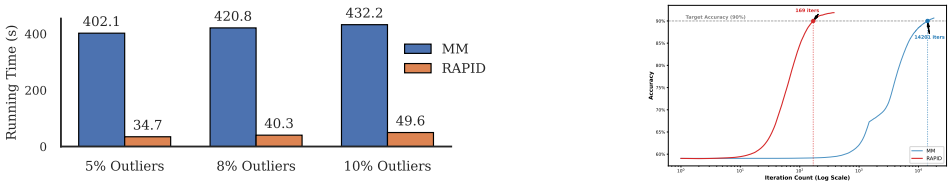


Figure 2: Computational efficiency comparison between the MM baseline and the proposed RAPID algorithm. **Left:** Average running time across different outlier contamination levels (inclusive of hyperparameter tuning overhead). **Right:** Accuracy versus iteration count (log scale) under 5% outlier contamination, illustrating the iterations required to reach a 90% target accuracy.

7 Real Data Application

7.1 Experimental Setting and Results

In this section, we evaluate the same methods as in Section 6. Each method is evaluated on 12 datasets, including 4 real-world noisy datasets and 8 clean benchmark datasets (characteristics summarized in Table B.1 in Appendix B.2). To evaluate the robustness of each method against extreme outliers on the clean datasets, we employ a random label flipping strategy [5]. Specifically, we randomly select 10% of the training samples and invert their labels.

While our simulation study primarily focuses on classification performance (e.g., Example F1), standard evaluation protocols for real-world multi-label benchmarks place greater emphasis on ranking quality. Accordingly, consistent with previous studies on noisy multi-label learning [8, 66], we evaluate all methods on real-world benchmark datasets using five standard metrics: *Hamming Loss*, *Average Precision*, *One Error*, *Ranking Loss*, and *Coverage*. We conduct 20 independent runs for each dataset using random 80/20 train-test splits and report the average results.

To provide a comprehensive view of the performance across different evaluation metrics, we visualize the results on four of the twelve benchmark datasets using radar charts in Figure 3. As observed, RoLL consistently encompasses the competitive baselines, forming the largest enclosed area across all datasets. Notably, on some datasets such as *emotions* and *mirflickr*, RoLL demonstrates significant margins over the second-best methods, highlighting its effectiveness in capturing latent label correlations and handling data anomalies. Similar behaviors are also observed across the other datasets, with more results provided in Appendix B.4.

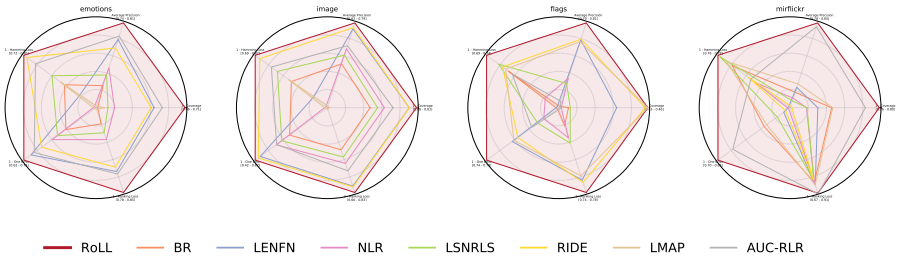


Figure 3: Comparative performance on four real-world multi-label datasets using radar charts. The axes are transformed such that larger values (outer positions) indicate better results and a larger enclosed area represents better overall performance. RoLL (red region) consistently yields the largest area against seven baselines.

7.2 Ablation Study

To validate the contribution of each component in RoLL, we conduct an ablation study by comparing RoLL against two variants: RoLL-R, which removes the singular value penalty on $\boldsymbol{B}$, and RoLL-C, which excludes the robustness component $\boldsymbol{C}$. Table 1 presents the results on the *emotions* dataset. Additional ablation results on the *foodtruck* and *yeast* datasets are provided in Appendix B.8. The ablation study shows that both components play a critical role in the performance of RoLL.

Table 1: Ablation study results on the emotions dataset (mean $\pm$ standard error). Metrics with $\downarrow$ indicate lower is better, and $\uparrow$ indicates higher is better. **Bold** values denote the best performance.

Metric	RoLL	RoLL-R	RoLL-C
Hamming Loss $\downarrow$	**0.2032±.0023**	0.2446±.0027	0.2225±.0025
Avg. Precision $\uparrow$	**0.8106±.0045**	0.7569±.0055	0.7777±.0055
One Error $\downarrow$	**0.2571±.0076**	0.3286±.0101	0.2996±.0095
Ranking Loss $\downarrow$	**0.1531±.0042**	0.2087±.0046	0.1864±.0048
Coverage $\downarrow$	**0.2870±.0044**	0.3389±.0043	0.3182±.0046

8 Conclusion

This paper proposes a unified framework for robust low-rank matrix learning under a wide class of losses. Additionally, we develop a novel predictive information criterion supported by rigorous theoretical analysis for model selection. For fast computation, a new version of Nesterov's first acceleration with established convergence guarantees is proposed. This method features a dynamic adjustment of local restricted strong convexity parameters to enhance convergence speed. Comprehensive experiments reveal that RoLL outperforms baselines under data contamination.

Acknowledgments and Disclosure of Funding

This work was supported by the NSFC Projects (No.12171479) and Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing. The authors would like to thank Shixiang Liu from Renmin University of China for insightful discussions. The authors declare no competing interests.

References

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

[2] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[3] 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*, pages 17–34. Springer, 1996.

[4] 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.

[5] Battista Biggio, Blaine Nelson, and Pavel Laskov. Support vector machines under adversarial label noise. In *Proceedings of the Asian Conference on Machine Learning*, volume 20, pages 97–112. PMLR, 2011.

[6] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[7] Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[8] 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*, pages 3722–3731, 2024.

[9] 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.

[10] 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.

[11] Alexandre d'Aspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. *Foundations and Trends in Optimization*, 5(1–2):1–245, 2021.

[12] Mark de Rooij, Dion Woestenburg, and Frank Busing. Supervised and unsupervised mapping of binary variables: a proximity perspective. *Behaviormetrika*, 52(2):461–503, 2025.

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

[14] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[15] Jianqing Fan, Wenyan Gong, and Ziwei Zhu. Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics*, 212(1):177–202, 2019.

[16] Mihai I. Florea and Sergiy A. Vorobyov. A generalized accelerated composite gradient method: Uniting Nesterov's fast gradient method and FISTA. *IEEE Transactions on Signal Processing*, 68:3033–3048, 2020.

[17] Amirmasoud Ghiassi, Cosmin Octavian Pene, Robert Birke, and Lydia Y. Chen. Trusted loss correction for noisy multi-label learning. In *Proceedings of the Asian Conference on Machine Learning*, volume 189, pages 343–358. PMLR, 2023.

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

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

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

[21] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[22] Peter J. Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

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

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

[25] Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5): 2302–2329, 2011.

[26] Fatma Sevinç Kurnaz, Irene Hoffmann, and Peter Filzmoser. Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemometrics and Intelligent Laboratory Systems*, 172:211–222, 2018.

[27] 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.

[28] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 32100–32121. PMLR, 2024.

[29] Shixiang Liu and Hanming Yang. Adversarial contamination meets hard thresholding: An iterative algorithm with signal adaptivity and minimax optimality. *Statistica Sinica*, 2026. Forthcoming.

[30] Canyi Lu, Jiashi Feng, Yudong Chen, Wei Liu, Zhouchen Lin, and Shuicheng Yan. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):925–938, 2020.

[31] 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.

[32] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah M. Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 6543–6553. PMLR, 2020.

[33] The Tien Mai. Robust reduced rank regression under heavy-tailed noise and missing data via non-convex penalization. *arXiv preprint arXiv:2512.24450*, 2025.

[34] The Tien Mai. Robust low-rank estimation with multiple binary responses using pairwise AUC loss. *arXiv preprint arXiv:2601.08618*, 2026.

[35] Stanislav Minsker, Mohamed Ndaoud, and Lang Wang. Robust and tuning-free sparse linear regression via square-root slope. *SIAM Journal on Mathematics of Data Science*, 6(2):428–453, 2024.

[36] Sahand Negahban and Martin J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(53): 1665–1697, 2012.

[37] Yurii E. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics. Doklady*, 27(2):372–376, 1983.

[38] Yurii E. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004.

[39] Luong Trung Nguyen, Junhan Kim, Sangtae Kim, and Byonghyo Shim. Localization of IoT networks via low-rank matrix completion. *IEEE Transactions on Communications*, 67(8): 5833–5847, 2019.

[40] Feiping Nie, Heng Huang, and Chris Ding. Low-rank matrix recovery via efficient Schatten p-norm minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 655–661, 2012.

[41] Brendan O'Donoghue and Emmanuel Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732, 2015.

[42] 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.

[43] 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.

[44] Philippe Rigollet and Alexandre B. Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

[45] Vincent Roulet and Alexandre d'Aspremont. Sharpness, restart and acceleration. In *Advances in Neural Information Processing Systems*, volume 30, pages 1119–1129, 2017.

[46] Peter J. Rousseeuw. Multivariate estimation with high breakdown point. In *Mathematical statistics and applications*, volume B, pages 283–297. D. Reidel Publishing Company, Dordrecht, 1985.

[47] Peter J. Rousseeuw and Katrien Van Driessen. Computing lts regression for large data sets. *Data Mining and Knowledge Discovery*, 12(1):29–45, 2006.

[48] 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.

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

[50] Yiyuan She. On the finite-sample analysis of θ-estimators. *Electronic Journal of Statistics*, 10 (2):1874–1895, 2016.

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

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

[53] 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.

[54] Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive Huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

[55] 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.

[56] Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. *SIAM Journal on Optimization*, 28(3):2274–2303, 2018.

[57] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[58] Nickolay Trendafilov, Michele Gallo, Violetta Simonacci, and Valentin Todorov. Overview of reduced-rank regression with dense and sparse coefficients, and a new estimation procedure. *Behaviormetrika*, 52(2):593–625, 2025.

[59] Paul Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming*, 125(2):263–295, 2010.

[60] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, NY, 2009.

[61] John Wilder Tukey. A survey of sampling from contaminated distributions. In *Contributions to probability and statistics*, pages 448–485. Stanford University Press, Stanford, CA, 1960.

[62] Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[63] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, New York, 1996.

[64] Yejiang Wang, Yuhai Zhao, Zhengkui Wang, Wen Shan, and Xingwei Wang. Limited-supervised multi-label learning with dependency noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15662–15670, 2024.

[65] Yue Wang, Wenqi Lu, Lei Wang, Zhongyi Zhu, Hongmei Lin, and Heng Lian. Regularized adaptive Huber matrix regression and distributed learning. *Statistica Sinica*, 35(2):919–937, 2025.

[66] 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*, pages 3458–3467, 2025.

[67] Ming-Kun Xie and Sheng-Jun Huang. CCMN: A general framework for learning with class-conditional multi-label noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):154–166, 2023.

[68] Fuchao Yang, Yuheng Jia, Hui Liu, Yongqiang Dong, and Junhui Hou. Noisy label removal for partial multi-label learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3724–3735, 2024.

[69] Thomas W. Yee and Trevor J. Hastie. Reduced-rank vector generalized linear models. *Statistical Modelling*, 3(1):15–41, 2003.

[70] 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)*, pages 1398–1403. IEEE, 2018.

[71] 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.

[72] Man-Chung Yue and Anthony Man-Cho So. A perturbation inequality for concave functions of singular values and its applications in low-rank matrix recovery. *Applied and Computational Harmonic Analysis*, 40(2):396–416, 2016.

[73] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[74] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

[75] Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in neural information processing systems*, volume 31, pages 8778–8788, 2018.

[76] 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*, pages 5581–5589, 2024.

[77] 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*, pages 121–139. Springer, 2013.

[78] Pengbiao Zhou, Jialiang Li, and Xiao Guo. Robust factor analysis for high-dimensional data with outliers. *Journal of Business & Economic Statistics*, 44(3):916–927, 2026.

[79] Ziwei Zhu, Tengyao Wang, and Richard J. Samworth. High-dimensional principal component analysis with heterogeneous missingness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):2000–2031, 2022.

[80] Yizhang Zou, Xuegang Hu, Peipei Li, and Yuhang Ge. Learning shared and non-redundant label-specific features for partial multi-label classification. *Information Sciences*, 656:119917, 2024.

A Technical Details

A.1 Definitions and Notations

Additional Notations. Let $[\cdot]_+ = \max(\cdot, 0)$. We denote the nuclear norm by $\|\cdot\|_*$, the $\ell_{2,1}$ norm by $\|\cdot\|_{2,1}$, and the ℓ_0 norm by $\|\cdot\|_0$. 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, for any matrix $\boldsymbol{A}$, we write $\boldsymbol{A}_{\mathcal{I}} = \boldsymbol{A}(\mathcal{I}, :)$ to denote the submatrix formed by the rows of $\boldsymbol{A}$ indexed by $\mathcal{I}$. Given a space $\mathcal{M}$, for any matrix $\boldsymbol{A}$, we denote by $\boldsymbol{A}_{\mathcal{M}}$ the orthogonal projection of $\boldsymbol{A}$ onto $\mathcal{M}$. Given $a \in \mathbb{R}$, $\lfloor a \rfloor$ and $\lceil a \rceil$ denote the floor and ceiling functions, respectively.

Definition A.1 ($\mathcal{L}$-smoothness)**.** The loss function $h(\boldsymbol{A})$ is said to be $\mathcal{L}$-smooth with respect to $\boldsymbol{A}$ if there exists $\mathcal{L} > 0$ such that

$$\|\nabla h(\boldsymbol{A}_1) - \nabla h(\boldsymbol{A}_2)\|_F \leq \mathcal{L}\|\boldsymbol{A}_1 - \boldsymbol{A}_2\|_F, \tag{A.1}$$

which implies $\boldsymbol{\Delta}_h(\boldsymbol{A}_1, \boldsymbol{A}_2) \leq \mathcal{L}\mathbf{D}_2(\boldsymbol{A}_1, \boldsymbol{A}_2), \forall \boldsymbol{A}_1, \boldsymbol{A}_2$.

Definition A.2 (μ-strong convexity)**.** The loss function $h(\boldsymbol{A})$ is said to be μ-strongly convex with respect to $\boldsymbol{A}$ if there exists $\mu > 0$ such that, for all $\boldsymbol{A}_1, \boldsymbol{A}_2$,

$$h(\boldsymbol{A}_2) \geq h(\boldsymbol{A}_1) + \langle \nabla h(\boldsymbol{A}_1), \boldsymbol{A}_2 - \boldsymbol{A}_1 \rangle + \frac{\mu}{2}\|\boldsymbol{A}_2 - \boldsymbol{A}_1\|_F^2, \tag{A.2}$$

or equivalently $\boldsymbol{\Delta}_h(\boldsymbol{A}_1, \boldsymbol{A}_2) \geq \mu \mathbf{D}_2(\boldsymbol{A}_1, \boldsymbol{A}_2)$.

Definition A.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 $\|\langle \boldsymbol{\xi}, \boldsymbol{\alpha} \rangle\|_{\psi_2} \leq \sigma\|\boldsymbol{\alpha}\|_2, \forall \boldsymbol{\alpha} \in \mathbb{R}^d$.

A thresholding function is a real-valued function $\Theta(t; \lambda)$ defined for $-\infty < t < \infty$ and $0 \leq \lambda < \infty$ such that (i) $\Theta(-t; \lambda) = -\Theta(t; \lambda)$; (ii) $\Theta(t; \lambda) \leq \Theta(t'; \lambda)$ for $t \leq t'$; (iii) $\lim_{t \to \infty} \Theta(t; \lambda) = \infty$; (iv) $0 \leq \Theta(t; \lambda) \leq t$ for $0 \leq t < \infty$. A vector version of Θ (still denoted by Θ) is defined componentwise if either t or λ is replaced by a vector. From the definition, $\Theta^{-1}(u; \lambda) := \sup\{t : \Theta(t; \lambda) \leq u\}, \forall u > 0$ must be monotonically non-decreasing and so its derivative is defined almost everywhere on $(0, \infty)$. Given Θ, a critical number $\mathcal{L}_\Theta \leq 1$ can be introduced such that $\mathrm{d}\Theta^{-1}(u; \lambda)/\mathrm{d}u \geq 1 - \mathcal{L}_\Theta$ for almost every $u \geq 0$, or $\mathcal{L}_\Theta := 1 - \operatorname{ess\,inf}\{\mathrm{d}\Theta^{-1}(u; \lambda)/\mathrm{d}u : u \geq 0\}$, where $\operatorname{ess\,inf}$ is the essential infimum. For any arbitrarily given Θ, we construct a penalty function $P_\Theta(t; \lambda)$ as follows $P_\Theta(t; \lambda) = \int_0^{|t|} (\Theta^{-1}(u; \lambda) - u)\,\mathrm{d}u = \int_0^{|t|} (\sup\{s : \Theta(s; \lambda) \leq u\} - u)\,\mathrm{d}u$ for any $t \in \mathbb{R}$.

A.2 Proof of Theorem 4.3

Proof. $(\hat{\boldsymbol{B}}, \hat{\boldsymbol{C}}) \in \arg\min_{\boldsymbol{B}, \boldsymbol{C}} g(\boldsymbol{B}, \boldsymbol{C}; \boldsymbol{B}^-, \boldsymbol{C}^-)|_{\boldsymbol{B}^- = \hat{\boldsymbol{B}}, \boldsymbol{C}^- = \hat{\boldsymbol{C}}}$, where $g(\boldsymbol{B}, \boldsymbol{C}; \boldsymbol{B}^-, \boldsymbol{C}^-) = l(\boldsymbol{X}\boldsymbol{B}^- + \boldsymbol{C}^-) + \langle \nabla l(\boldsymbol{X}\boldsymbol{B}^- + \boldsymbol{C}^-), \boldsymbol{X}\boldsymbol{B} - \boldsymbol{X}\boldsymbol{B}^- + \boldsymbol{C} - \boldsymbol{C}^- \rangle + \|\boldsymbol{B} - \boldsymbol{B}^-\|_F^2/2 + \|\boldsymbol{C} - \boldsymbol{C}^-\|_F^2/2 + \sum_{i=1}^{p \wedge m} P_{\boldsymbol{B}}(\sigma_i^{(\boldsymbol{B})}; \lambda_{\boldsymbol{B}}) + \sum_{i=1}^n P_{\boldsymbol{C}}(\|\boldsymbol{c}_i\|_2; \lambda_{\boldsymbol{C}})$. Then by Lemma 2 in [50], we have

$$\begin{aligned}
&\frac{1}{2}\|\boldsymbol{C}^* - \hat{\boldsymbol{C}} + \nabla l(\boldsymbol{X}\hat{\boldsymbol{B}} + \hat{\boldsymbol{C}})\|_F^2 - \frac{1}{2}\|\nabla l(\boldsymbol{X}\hat{\boldsymbol{B}} + \hat{\boldsymbol{C}})\|_F^2 + \sum_{i=1}^{p \wedge m} P_{\boldsymbol{B}}(\sigma_i^{(\boldsymbol{B}^*)}; \lambda_{\boldsymbol{B}}) \\
&- \sum_{i=1}^{p \wedge m} P_{\boldsymbol{B}}(\sigma_i^{(\hat{\boldsymbol{B}})}; \lambda_{\boldsymbol{B}}) + \sum_{i=1}^n P_{\boldsymbol{C}}(\|\boldsymbol{c}_i^*\|_2; \lambda_{\boldsymbol{C}}) - \sum_{i=1}^n P_{\boldsymbol{C}}(\|\hat{\boldsymbol{c}}_i\|_2; \lambda_{\boldsymbol{C}}) \\
&+ \frac{1}{2}\|\boldsymbol{B}^* - \hat{\boldsymbol{B}} + \boldsymbol{X}^\top \nabla l(\boldsymbol{X}\hat{\boldsymbol{B}} + \hat{\boldsymbol{C}})\|_F^2 - \frac{1}{2}\|\boldsymbol{X}^\top \nabla l(\boldsymbol{X}\hat{\boldsymbol{B}} + \hat{\boldsymbol{C}})\|_F^2 \geq \frac{1 - \mathcal{L}_\Theta}{2}\|\hat{\bar{\boldsymbol{B}}} - \bar{\boldsymbol{B}}^*\|_F^2,
\end{aligned} \tag{A.3}$$

or equivalently

$$
\begin{aligned}
&2\bar{\boldsymbol{\Delta}}_l(\boldsymbol{X}\hat{\boldsymbol{B}}+\hat{\boldsymbol{C}},\boldsymbol{X}\boldsymbol{B}^*+\boldsymbol{C}^*)\\
\le&\sum_{i=1}^{p\wedge m}P_{\boldsymbol{B}}(\sigma_i^{(\boldsymbol{B}^*)};\lambda_{\boldsymbol{B}})-\sum_{i=1}^{p\wedge m}P_{\boldsymbol{B}}(\sigma_i^{(\hat{\boldsymbol{B}})};\lambda_{\boldsymbol{B}})+\frac{\mathcal{L}_\Theta}{2}\|\hat{\bar{\boldsymbol{B}}}-\bar{\boldsymbol{B}}^*\|_F^2\\
&+\sum_{i=1}^{n}P_{\boldsymbol{C}}(\|\boldsymbol{c}_i^*\|_2;\lambda_{\boldsymbol{C}})-\sum_{i=1}^{n}P_{\boldsymbol{C}}(\|\hat{\boldsymbol{c}}_i\|_2;\lambda_{\boldsymbol{C}})+\langle\boldsymbol{\mathcal{E}},\bar{\boldsymbol{X}}\hat{\bar{\boldsymbol{B}}}-\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*\rangle. \qquad \text{(A.4)}
\end{aligned}
$$

Recall that $P_H(t;\lambda)=(-t^2/2+\lambda|t|)1_{|t|<\lambda}+(\lambda^2/2)1_{|t|\ge\lambda}$ and $P_{2,H}(\boldsymbol{C};\lambda_{\boldsymbol{C}})=\sum_{i=1}^n P_H(\|\boldsymbol{c}_i\|_2;\lambda_{\boldsymbol{C}})$. Then, we try to bound the stochastic term in a proper way.

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

$$
\sup_{(\boldsymbol{B},\boldsymbol{C})\in\mathcal{B}_{r,J}}\{2\langle\boldsymbol{\mathcal{E}},\boldsymbol{X}\boldsymbol{B}+\boldsymbol{C}\rangle-\frac{1}{a}\|\boldsymbol{X}\boldsymbol{B}+\boldsymbol{C}\|_F^2-\frac{1}{b}P_{2,H}(\boldsymbol{C};\lambda_{\boldsymbol{C}})-\frac{1}{b}\sum_{i=1}^{p\wedge m}P_H(\sigma_i^{(\boldsymbol{B})};\lambda_{\boldsymbol{B}})\}\ge a\sigma^2 t \qquad \text{(A.5)}
$$

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

The proof of this lemma is provided in Appendix A.4.

Let $l_H(\boldsymbol{B},\boldsymbol{C})=2\langle\boldsymbol{\mathcal{E}},\boldsymbol{X}\boldsymbol{B}+\boldsymbol{C}\rangle-\frac{1}{a}\|\boldsymbol{X}\boldsymbol{B}+\boldsymbol{C}\|_F^2-\frac{1}{b}P_{2,H}(\boldsymbol{C};\lambda_{\boldsymbol{C}})-\frac{1}{b}\sum_{i=1}^{p\wedge m}P_H(\sigma_i^{(\boldsymbol{B})};\lambda_{\boldsymbol{B}})$. Define $R=\sup_{1\le r\le m\wedge p,1\le J\le n}\sup_{(\boldsymbol{B},\boldsymbol{C})\in\mathcal{B}_{r,J}}l_H(\boldsymbol{B},\boldsymbol{C})$. From Lemma A.4, we have $\mathbb{E}(R)\le c'a\sigma^2/c$. Substituting the following bound into (A.4),

$$
\begin{aligned}
&2\langle\boldsymbol{\mathcal{E}},\boldsymbol{X}(\hat{\boldsymbol{B}}-\boldsymbol{B}^*)+(\hat{\boldsymbol{C}}-\boldsymbol{C}^*)\rangle\\
\le&\frac{1}{a}\|\boldsymbol{X}(\hat{\boldsymbol{B}}-\boldsymbol{B}^*)+(\hat{\boldsymbol{C}}-\boldsymbol{C}^*)\|_F^2+\frac{1}{b}P_{2,H}(\hat{\boldsymbol{C}}-\boldsymbol{C}^*;\lambda_{\boldsymbol{C}})\\
&+\frac{1}{b}\sum_{i=1}^{p\wedge m}P_H(\sigma_i^{(\hat{\boldsymbol{B}}-\boldsymbol{B}^*)};\lambda_{\boldsymbol{B}})+R,
\end{aligned}
$$

we have

$$
\begin{aligned}
2\bar{\boldsymbol{\Delta}}_l(\boldsymbol{X}\hat{\boldsymbol{B}}+\hat{\boldsymbol{C}},\boldsymbol{X}\boldsymbol{B}^*+\boldsymbol{C}^*)\le&\sum_{i=1}^{p\wedge m}P_{\boldsymbol{B}}(\sigma_i^{(\boldsymbol{B}^*)};\lambda_{\boldsymbol{B}})-\sum_{i=1}^{p\wedge m}P_{\boldsymbol{B}}(\sigma_i^{(\hat{\boldsymbol{B}})};\lambda_{\boldsymbol{B}})\\
&+\sum_{i=1}^{n}P_{\boldsymbol{C}}(\|\boldsymbol{c}_i^*\|_2;\lambda_{\boldsymbol{C}})-\sum_{i=1}^{n}P_{\boldsymbol{C}}(\|\hat{\boldsymbol{c}}_i\|_2;\lambda_{\boldsymbol{C}})+\frac{1}{2a}\|\boldsymbol{X}(\hat{\boldsymbol{B}}-\boldsymbol{B}^*)+(\hat{\boldsymbol{C}}-\boldsymbol{C}^*)\|_F^2\\
&+\frac{1}{2b}P_{2,H}(\hat{\boldsymbol{C}}-\boldsymbol{C}^*;\lambda_{\boldsymbol{C}})+\frac{1}{2b}\sum_{i=1}^{p\wedge m}P_H(\sigma_i^{(\hat{\boldsymbol{B}}-\boldsymbol{B}^*)};\lambda_{\boldsymbol{B}})+\frac{\mathcal{L}_\Theta}{2}\|\hat{\bar{\boldsymbol{B}}}-\bar{\boldsymbol{B}}^*\|_F^2+\frac{R}{2}. \qquad \text{(A.6)}
\end{aligned}
$$

Setting $b=1/(2\vartheta)$ with $\vartheta>0$ and using the regularity condition, we have

$$
(\frac{\delta}{2}-\frac{1}{2a})\|\boldsymbol{X}(\hat{\boldsymbol{B}}-\boldsymbol{B}^*)+(\hat{\boldsymbol{C}}-\boldsymbol{C}^*)\|_F^2\le K_1^2\lambda_{\boldsymbol{B}}^2r^*+K_2^2\lambda_{\boldsymbol{C}}^2J^*+\frac{R}{2}. \qquad \text{(A.7)}
$$

Taking $a=2/\delta+1/\vartheta$ and $A\ge(abA_1)^{1/2}$ gives the conclusion. □

Remark A.5. The regularity condition in Theorem 4.3 is mild. To demonstrate this, let the penalties $P_{\boldsymbol{B}}(\cdot;\lambda_{\boldsymbol{B}})$ and $P_{\boldsymbol{C}}(\cdot;\lambda_{\boldsymbol{C}})$ be the ℓ_1 penalty. Consequently, $\sum_{i=1}^{p\wedge m}P_{\boldsymbol{B}}(\sigma_i^{(\boldsymbol{B})};\lambda_{\boldsymbol{B}})$ and $\sum_{i=1}^n P_{\boldsymbol{C}}(\|\boldsymbol{c}_i\|_2;\lambda_{\boldsymbol{C}})$ correspond to the nuclear penalty ($\|\cdot\|_*$) and the $\ell_{2,1}$ penalty ($\|\cdot\|_{2,1}$), respectively. Therefore, we have $\mathcal{L}_\Theta=0$. Let $\boldsymbol{B}^*=\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$. Define the tangent space $\mathcal{M}=\{\boldsymbol{U}\boldsymbol{R}^\top+\boldsymbol{S}\boldsymbol{V}^\top:\boldsymbol{R}\in\mathbb{R}^{m\times r^*},\boldsymbol{S}\in\mathbb{R}^{p\times r^*}\}$ and its orthogonal complement $\mathcal{M}^\perp=$

$\{\boldsymbol{A} \in \mathbb{R}^{p\times m} : \boldsymbol{U}^\top \boldsymbol{A} = \boldsymbol{0},\ \boldsymbol{A}\boldsymbol{V} = \boldsymbol{0}\}$. A sufficient set of conditions for the regularity condition is $(\bar{\boldsymbol{\Delta}}_l - \frac{\delta}{2}\mathbf{D}_2)(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*) + \lambda_{\boldsymbol{C}}(1-\vartheta)\|(\boldsymbol{C}-\boldsymbol{C}^*)_{\mathcal{J}^c}\|_{2,1} + K_2^2\lambda_{\boldsymbol{C}}^2 J^* \geq \lambda_{\boldsymbol{C}}(1+\vartheta)\|(\boldsymbol{C}-\boldsymbol{C}^*)_{\mathcal{J}}\|_{2,1}$ and $(\bar{\boldsymbol{\Delta}}_l - \frac{\delta}{2}\mathbf{D}_2)(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*) + \lambda_{\boldsymbol{B}}(1-\vartheta)\|(\boldsymbol{B}-\boldsymbol{B}^*)_{\mathcal{M}^\perp}\|_* + K_1^2\lambda_{\boldsymbol{B}}^2 r^* \geq \lambda_{\boldsymbol{B}}(1+\vartheta)\|(\boldsymbol{B}-\boldsymbol{B}^*) - (\boldsymbol{B}-\boldsymbol{B}^*)_{\mathcal{M}^\perp}\|_*$.

To give readers more intuition, let $l(\bar{\boldsymbol{B}}) = \|\bar{\boldsymbol{X}}\bar{\boldsymbol{B}} - \boldsymbol{Y}\|_F^2/2$. In this scenario, the conditions above simplify to $(1-\frac{\delta}{2})\|\bar{\boldsymbol{X}}\bar{\boldsymbol{B}} - \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*\|_F^2/2 + \lambda_{\boldsymbol{C}}(1-\vartheta)\|(\boldsymbol{C}-\boldsymbol{C}^*)_{\mathcal{J}^c}\|_{2,1} + K_2^2\lambda_{\boldsymbol{C}}^2 J^* \geq \lambda_{\boldsymbol{C}}(1+\vartheta)\|(\boldsymbol{C}-\boldsymbol{C}^*)_{\mathcal{J}}\|_{2,1}$ and $(1-\frac{\delta}{2})\|\bar{\boldsymbol{X}}\bar{\boldsymbol{B}} - \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*\|_F^2/2 + \lambda_{\boldsymbol{B}}(1-\vartheta)\|(\boldsymbol{B}-\boldsymbol{B}^*)_{\mathcal{M}^\perp}\|_* + K_1^2\lambda_{\boldsymbol{B}}^2 r^* \geq \lambda_{\boldsymbol{B}}(1+\vartheta)\|(\boldsymbol{B}-\boldsymbol{B}^*) - (\boldsymbol{B}-\boldsymbol{B}^*)_{\mathcal{M}^\perp}\|_*$. Let $\boldsymbol{Q}$ denote the orthogonal projector onto the orthogonal complement of the column space of $\boldsymbol{X}$. Taking $0 < \delta < 2$ and $0 < \vartheta < 1$, and redefining K_1, K_2, ϑ if necessary, we obtain the following sufficient conditions:

$$K_1\sqrt{r^*}\|\bar{\boldsymbol{X}}\bar{\boldsymbol{B}} - \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*\|_F + \|(\boldsymbol{B}-\boldsymbol{B}^*)_{\mathcal{M}^\perp}\|_* \geq (1+\vartheta)\|(\boldsymbol{B}-\boldsymbol{B}^*) - (\boldsymbol{B}-\boldsymbol{B}^*)_{\mathcal{M}^\perp}\|_*, \quad \text{(A.8)}$$

and

$$K_2\sqrt{J^*}\|\boldsymbol{Q}\boldsymbol{C} - \boldsymbol{Q}\boldsymbol{C}^*\|_F + \|(\boldsymbol{C}-\boldsymbol{C}^*)_{\mathcal{J}^c}\|_{2,1} \geq (1+\vartheta)\|(\boldsymbol{C}-\boldsymbol{C}^*)_{\mathcal{J}}\|_{2,1}. \quad \text{(A.9)}$$

Next, we compare (A.9) with the restricted eigenvalue (RE) and compatibility conditions [62, 4]. Let $\boldsymbol{\Delta}^{\boldsymbol{C}} = \boldsymbol{C} - \boldsymbol{C}^*$. These standard conditions typically require that there exist positive numbers $\varkappa, \vartheta_{\mathrm{RE}}$ such that $J^*\|\boldsymbol{Q}\boldsymbol{\Delta}^{\boldsymbol{C}}\|_F^2 \geq \varkappa\|\boldsymbol{\Delta}^{\boldsymbol{C}}_{\mathcal{J}}\|_{2,1}^2$ (compatibility) or more restrictively, $\|\boldsymbol{Q}\boldsymbol{\Delta}^{\boldsymbol{C}}\|_F^2 \geq \varkappa\|\boldsymbol{\Delta}^{\boldsymbol{C}}_{\mathcal{J}}\|_F^2$ (RE) for all $\boldsymbol{C}, \boldsymbol{C}^*$: $(1+\vartheta_{\mathrm{RE}})\|\boldsymbol{\Delta}^{\boldsymbol{C}}_{\mathcal{J}}\|_{2,1} \geq \|\boldsymbol{\Delta}^{\boldsymbol{C}}_{\mathcal{J}^c}\|_{2,1}$. Choose the redefined constants such that $\vartheta \leq \vartheta_{\mathrm{RE}}$ and $K_2\sqrt{\varkappa} \geq 1+\vartheta$. Under these conditions, $K_2\sqrt{J^*}\|\boldsymbol{Q}\boldsymbol{\Delta}^{\boldsymbol{C}}\|_F \vee \|\boldsymbol{\Delta}^{\boldsymbol{C}}_{\mathcal{J}^c}\|_{2,1} \geq (1+\vartheta)\|\boldsymbol{\Delta}^{\boldsymbol{C}}_{\mathcal{J}}\|_{2,1}$ holds, implying that (A.9) is less restrictive than both the RE and compatibility conditions. Analogously, for heuristic comparison, under a suitable incoherence condition controlling the interaction between the coefficient and contamination components, a matrix RE-type condition [25] can imply (A.8) over the corresponding restricted perturbation set. Hence, conditional on this incoherence condition, (A.8) is less restrictive than the matrix RE-type condition on this set.

For completeness, in the quadratic-loss case, the regularity condition in Theorem 4.3 also follows from a joint restricted identifiability assumption. Specifically, suppose that, for some $\kappa > 0$,

$$\|\bar{\boldsymbol{X}}\bar{\boldsymbol{B}} - \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*\|_F^2 \geq \kappa(\|(\boldsymbol{B}-\boldsymbol{B}^*)_{\mathcal{M}}\|_F^2 + \|(\boldsymbol{\Delta}^{\boldsymbol{C}})_{\mathcal{J}}\|_F^2)$$

whenever

$$\begin{aligned}&(1-\vartheta)\left[\lambda_{\boldsymbol{B}}\|(\boldsymbol{B}-\boldsymbol{B}^*)_{\mathcal{M}^\perp}\|_* + \lambda_{\boldsymbol{C}}\|(\boldsymbol{\Delta}^{\boldsymbol{C}})_{\mathcal{J}^c}\|_{2,1}\right]\\ &\leq (1+\vartheta)\left[\lambda_{\boldsymbol{B}}\|(\boldsymbol{B}-\boldsymbol{B}^*)_{\mathcal{M}}\|_* + \lambda_{\boldsymbol{C}}\|(\boldsymbol{\Delta}^{\boldsymbol{C}})_{\mathcal{J}}\|_{2,1}\right].\end{aligned}$$

Then, for $0 < \delta < 2$ and $0 < \vartheta < 1$, the regularity condition holds for all $(\boldsymbol{B}, \boldsymbol{C})$ provided that K_1 and K_2 are sufficiently large, as follows from decomposability and Young's inequality.

Remark A.6. Theorem 4.3 generalizes Theorem 6 in [51] by extending the analysis from the quadratic loss (used in reduced rank regression) to a broader class of loss functions, recovering their bound as a special case. Furthermore, while their analysis relies on the M-estimator framework (requiring a global optimum rarely guaranteed in practice), our analysis only requires the estimator to be a fixed point of the algorithm.

A.3 Corollary A.7

Corollary A.7. *Consider* $\min_{(\boldsymbol{B},\boldsymbol{C})} l(\boldsymbol{X}\boldsymbol{B}+\boldsymbol{C};\boldsymbol{Y}) + \sum_{i=1}^{p\wedge m} P_{\boldsymbol{B}}(\sigma_i^{(\boldsymbol{B})};\lambda_{\boldsymbol{B}}) + \sum_{i=1}^{n} P_{\boldsymbol{C}}(\|\boldsymbol{c}_i\|_2;\lambda_{\boldsymbol{C}})$, *where* $\lambda_{\boldsymbol{B}} = A\sigma(q+m)^{1/2}$, $\lambda_{\boldsymbol{C}} = A\sigma(m+\log n)^{1/2}$ *with* A *a sufficiently large constant. Suppose* $P_{\boldsymbol{B}}(\cdot;\lambda_{\boldsymbol{B}})$ *and* $P_{\boldsymbol{C}}(\cdot;\lambda_{\boldsymbol{C}})$ *are bounded nonconvex penalties satisfying* $P_\Theta(\cdot;\lambda) \lesssim \lambda^2$. *Then, for any global optimum* $(\hat{\boldsymbol{B}}, \hat{\boldsymbol{C}})$, *we have*

$$\mathbb{E}[\mathbf{D}_2(\bar{\boldsymbol{X}}\hat{\bar{\boldsymbol{B}}}, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*)] \lesssim \sigma^2\{(q+m)r^* + (m+\log n)J^* + 1\} \quad \text{(A.10)}$$

under the assumption that l *is strongly convex with respect to* $\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}$, *where* $\mathcal{J} = \mathcal{J}(\boldsymbol{C}^*)$, $J^* = |\mathcal{J}|$, $r^* = r(\boldsymbol{B}^*)$.

Proof. By the same stochastic bound as in the proof of Theorem 4.3, we have

$$\begin{aligned}\boldsymbol{\Delta}_l(\boldsymbol{X}\hat{\boldsymbol{B}}+\hat{\boldsymbol{C}},\boldsymbol{X}\boldsymbol{B}^*+\boldsymbol{C}^*) \leq & \sum_{i=1}^{p\wedge m} P_{\boldsymbol{B}}(\sigma_i^{(\boldsymbol{B}^*)};\lambda_{\boldsymbol{B}}) - \sum_{i=1}^{p\wedge m} P_{\boldsymbol{B}}(\sigma_i^{(\hat{\boldsymbol{B}})};\lambda_{\boldsymbol{B}}) \\ & + \sum_{i=1}^{n} P_{\boldsymbol{C}}(\|\boldsymbol{c}_i^*\|_2;\lambda_{\boldsymbol{C}}) - \sum_{i=1}^{n} P_{\boldsymbol{C}}(\|\hat{\boldsymbol{c}}_i\|_2;\lambda_{\boldsymbol{C}}) + \frac{1}{2a}\|\boldsymbol{X}(\hat{\boldsymbol{B}}-\boldsymbol{B}^*)+(\hat{\boldsymbol{C}}-\boldsymbol{C}^*)\|_F^2 \\ & + \frac{1}{2b}P_{2,H}(\hat{\boldsymbol{C}}-\boldsymbol{C}^*;\lambda_{\boldsymbol{C}}) + \frac{1}{2b}\sum_{i=1}^{p\wedge m} P_H(\sigma_i^{(\hat{\boldsymbol{B}}-\boldsymbol{B}^*)};\lambda_{\boldsymbol{B}}) + \frac{R}{2}. \end{aligned} \tag{A.11}$$

It remains to deal with

$$I_1 := \sum_{i=1}^{p\wedge m} P_{\boldsymbol{B}}(\sigma_i^{(\boldsymbol{B}^*)};\lambda_{\boldsymbol{B}}) - \sum_{i=1}^{p\wedge m} P_{\boldsymbol{B}}(\sigma_i^{(\hat{\boldsymbol{B}})};\lambda_{\boldsymbol{B}}) + \frac{1}{2b}\sum_{i=1}^{p\wedge m} P_H(\sigma_i^{(\hat{\boldsymbol{B}}-\boldsymbol{B}^*)};\lambda_{\boldsymbol{B}}), \tag{A.12}$$

and

$$I_2 := \sum_{i=1}^{n} P_{\boldsymbol{C}}(\|\boldsymbol{c}_i^*\|_2;\lambda_{\boldsymbol{C}}) - \sum_{i=1}^{n} P_{\boldsymbol{C}}(\|\hat{\boldsymbol{c}}_i\|_2;\lambda_{\boldsymbol{C}}) + \frac{1}{2b}P_{2,H}(\hat{\boldsymbol{C}}-\boldsymbol{C}^*;\lambda_{\boldsymbol{C}}). \tag{A.13}$$

According to Theorem 1 in [72], as P_H is concave in $[0,+\infty)$ and $P_H(0;\lambda_{\boldsymbol{B}})=0$, we have

$$\begin{aligned} I_1 \leq & \sum_{i=1}^{p\wedge m} P_{\boldsymbol{B}}(\sigma_i^{(\boldsymbol{B}^*)};\lambda_{\boldsymbol{B}}) - \sum_{i=1}^{p\wedge m} P_H(\sigma_i^{(\hat{\boldsymbol{B}})};\lambda_{\boldsymbol{B}}) + \frac{1}{2b}\sum_{i=1}^{p\wedge m} P_H(\sigma_i^{(\hat{\boldsymbol{B}}-\boldsymbol{B}^*)};\lambda_{\boldsymbol{B}}) \\ \leq & \sum_{i=1}^{p\wedge m} P_{\boldsymbol{B}}(\sigma_i^{(\boldsymbol{B}^*)};\lambda_{\boldsymbol{B}}) + \frac{1}{2b}\sum_{i=1}^{p\wedge m} P_H(\sigma_i^{(\boldsymbol{B}^*)};\lambda_{\boldsymbol{B}}) + \frac{1}{2b}\sum_{i=1}^{p\wedge m} P_H(\sigma_i^{(\hat{\boldsymbol{B}})};\lambda_{\boldsymbol{B}}) - \sum_{i=1}^{p\wedge m} P_H(\sigma_i^{(\hat{\boldsymbol{B}})};\lambda_{\boldsymbol{B}}) \\ \leq & (1+\frac{1}{2b})\sum_{i=1}^{p\wedge m} P_{\boldsymbol{B}}(\sigma_i^{(\boldsymbol{B}^*)};\lambda_{\boldsymbol{B}}), \end{aligned} \tag{A.14}$$

if $b \geq 1/2$.

Due to the sub-additivity of P_H, we have

$$\begin{aligned} I_2 \leq & \sum_{i=1}^{n} P_{\boldsymbol{C}}(\|\boldsymbol{c}_i^*\|_2;\lambda_{\boldsymbol{C}}) - P_{2,H}(\hat{\boldsymbol{C}};\lambda_{\boldsymbol{C}}) + \frac{1}{2b}P_{2,H}(\hat{\boldsymbol{C}}-\boldsymbol{C}^*;\lambda_{\boldsymbol{C}}) \\ \leq & \sum_{i=1}^{n} P_{\boldsymbol{C}}(\|\boldsymbol{c}_i^*\|_2;\lambda_{\boldsymbol{C}}) + \frac{1}{2b}P_{2,H}(\boldsymbol{C}^*;\lambda_{\boldsymbol{C}}) + \frac{1}{2b}P_{2,H}(\hat{\boldsymbol{C}};\lambda_{\boldsymbol{C}}) - P_{2,H}(\hat{\boldsymbol{C}};\lambda_{\boldsymbol{C}}) \\ \leq & (1+\frac{1}{2b})\sum_{i=1}^{n} P_{\boldsymbol{C}}(\|\boldsymbol{c}_i^*\|_2;\lambda_{\boldsymbol{C}}), \end{aligned} \tag{A.15}$$

if $b \geq 1/2$.

Combining (A.11), (A.14) and (A.15), it follows that

$$\begin{aligned} & \boldsymbol{\Delta}_l(\boldsymbol{X}\hat{\boldsymbol{B}}+\hat{\boldsymbol{C}},\boldsymbol{X}\boldsymbol{B}^*+\boldsymbol{C}^*) \\ \leq & (1+\frac{1}{2b})\sum_{i=1}^{p\wedge m} P_{\boldsymbol{B}}(\sigma_i^{(\boldsymbol{B}^*)};\lambda_{\boldsymbol{B}}) + (1+\frac{1}{2b})\sum_{i=1}^{n} P_{\boldsymbol{C}}(\|\boldsymbol{c}_i^*\|_2;\lambda_{\boldsymbol{C}}) \\ & + \frac{1}{2a}\|\boldsymbol{X}(\hat{\boldsymbol{B}}-\boldsymbol{B}^*)+(\hat{\boldsymbol{C}}-\boldsymbol{C}^*)\|_F^2 + \frac{R}{2}. \end{aligned}$$

Since l is strongly convex with respect to $\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}$, choosing a sufficiently large such that $1/a$ is smaller than the strong convexity constant, and using $P_{\boldsymbol{B}}(\cdot;\lambda_{\boldsymbol{B}}) \lesssim \lambda_{\boldsymbol{B}}^2$, $P_{\boldsymbol{C}}(\cdot;\lambda_{\boldsymbol{C}}) \lesssim \lambda_{\boldsymbol{C}}^2$ and $\mathbb{E}(R) \lesssim \sigma^2$, we have

$$\mathbb{E}[\mathbf{D}_2(\bar{\boldsymbol{X}}\hat{\bar{\boldsymbol{B}}},\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*)] \lesssim \lambda_{\boldsymbol{B}}^2 r^* + \lambda_{\boldsymbol{C}}^2 J^* + \sigma^2. \tag{A.16}$$

Therefore,

$$\mathbb{E}[\mathbf{D}_2(\bar{\boldsymbol{X}}\hat{\bar{\boldsymbol{B}}},\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*)] \lesssim \sigma^2(q+m)r^* + \sigma^2(m+\log n)J^* + \sigma^2. \tag{A.17}$$

The proof is complete. □

A.4 Proof of Lemma A.4

Proof. Let $P_0(t;\lambda) = (\lambda^2/2)1_{t\neq 0}$ and $P_{2,0}(\boldsymbol{C};\lambda_{\boldsymbol{C}}) = \sum_{i=1}^n P_0(\|\boldsymbol{c}_i\|_2;\lambda_{\boldsymbol{C}})$. Define $l_H(\boldsymbol{B},\boldsymbol{C}) = 2\langle \boldsymbol{\mathcal{E}}, \boldsymbol{X}\boldsymbol{B}+\boldsymbol{C}\rangle - \frac{1}{a}\|\boldsymbol{X}\boldsymbol{B}+\boldsymbol{C}\|_F^2 - \frac{1}{b}P_{2,H}(\boldsymbol{C};\lambda_{\boldsymbol{C}}) - \frac{1}{b}\sum_{i=1}^{p\wedge m} P_H(\sigma_i^{(\boldsymbol{B})};\lambda_{\boldsymbol{B}})$. Similarly, we define $l_0(\boldsymbol{B},\boldsymbol{C}) = 2\langle \boldsymbol{\mathcal{E}}, \boldsymbol{X}\boldsymbol{B}+\boldsymbol{C}\rangle - \frac{1}{a}\|\boldsymbol{X}\boldsymbol{B}+\boldsymbol{C}\|_F^2 - \frac{1}{b}P_{2,0}(\boldsymbol{C};\lambda_{\boldsymbol{C}}) - \frac{1}{b}\sum_{i=1}^{p\wedge m} P_0(\sigma_i^{(\boldsymbol{B})};\lambda_{\boldsymbol{B}})$. Let $\mathcal{A}_H = \{\sup_{(\boldsymbol{B},\boldsymbol{C})\in\mathcal{B}_{r,J}} l_H(\boldsymbol{B},\boldsymbol{C}) \geq a\sigma^2 t\}$, $\mathcal{A}_0 = \{\sup_{(\boldsymbol{B},\boldsymbol{C})\in\mathcal{B}_{r,J}} l_0(\boldsymbol{B},\boldsymbol{C}) \geq a\sigma^2 t\}$. The occurrence of $\mathcal{A}_H$ implies that

$$l_H(\boldsymbol{B}^o,\boldsymbol{C}^o) \geq a\sigma^2 t, \tag{A.18}$$

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

$$\min_{(\boldsymbol{B},\boldsymbol{C})\in\mathcal{B}_{r,J}} -2\langle \boldsymbol{\mathcal{E}}, \boldsymbol{X}\boldsymbol{B}+\boldsymbol{C}\rangle + \frac{1}{a}\|\boldsymbol{X}\boldsymbol{B}+\boldsymbol{C}\|_F^2 + \frac{1}{b}P_{2,H}(\boldsymbol{C};\lambda_{\boldsymbol{C}}) + \frac{1}{b}\sum_{i=1}^{p\wedge m} P_H(\sigma_i^{(\boldsymbol{B})};\lambda_{\boldsymbol{B}}). \tag{A.19}$$

Lemma A.8 (Lemma 1 in [49])**.** *Given any $\tau \geq 1$, 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 $\|\boldsymbol{c}_i^o\|_2 \geq \lambda\tau^{1/2} \geq \lambda$.*

Lemma A.9. *Given any $\tau \geq \|\boldsymbol{X}\|_2^2$, there always exists a globally optimal $\boldsymbol{B}^o$ to $\min_{\boldsymbol{B}} \|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\|_F^2/2 + \tau\sum_{i=1}^{p\wedge m} P_H(\sigma_i^{(\boldsymbol{B})};\lambda)$ such that for any i either $\sigma_i^{(\boldsymbol{B}^o)} = 0$ or $\sigma_i^{(\boldsymbol{B}^o)} \geq \lambda\sqrt{\tau}/\|\boldsymbol{X}\|_2 \geq \lambda$.*

The proof arguments of Lemmas A.8 and A.9 continue to apply under the constraints $J(\boldsymbol{C}) \leq J$ and $r(\boldsymbol{B}) \leq r$, since the rowwise and singular-value perturbations used therein do not increase the support size or rank, respectively. Hence, with $a \geq 2b$, whenever $\mathcal{A}_H$ occurs and the supremum is attained, an optimal solution $(\boldsymbol{B}^o,\boldsymbol{C}^o)$ to (A.19) can be chosen such that $l_0(\boldsymbol{B}^o,\boldsymbol{C}^o) = l_H(\boldsymbol{B}^o,\boldsymbol{C}^o) \geq a\sigma^2 t$. Alternatively, the same perturbations can be applied to any feasible pair to obtain a feasible pair with no smaller l_H value and with $l_0 = l_H$, so the event inclusion also follows without requiring the supremum to be attained. Therefore, $\mathcal{A}_H \subseteq \mathcal{A}_0$ 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

$$2\langle \boldsymbol{\mathcal{E}}, \boldsymbol{X}\boldsymbol{B}+\boldsymbol{C}\rangle = 2\langle \boldsymbol{\mathcal{E}}, \mathcal{P}_{I_{\mathcal{J}}}^{\perp}\boldsymbol{X}\boldsymbol{B}\rangle + 2\langle \boldsymbol{\mathcal{E}}, \mathcal{P}_{I_{\mathcal{J}}}(\boldsymbol{X}\boldsymbol{B}+\boldsymbol{C})\rangle \equiv 2\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}\rangle + 2\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}\boldsymbol{I}\rangle. \tag{A.20}$$

Lemma A.10. (Lemma 5 in [51]) *Suppose* $\mathrm{vec}(\boldsymbol{\mathcal{E}})$ *is sub-Gaussian with mean zero and ψ_2-norm bounded by σ. Given $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $1 \leq r \leq m\wedge p$, $1 \leq J \leq n$, define $\Gamma_{r,J}^1 = \{\boldsymbol{A} \in \mathbb{R}^{n\times m} : \|\boldsymbol{A}\|_F \leq 1, r(\boldsymbol{A}) \leq r, \mathrm{CS}(\boldsymbol{A}) \subset \mathrm{CS}\{\mathcal{P}_{I_{\mathcal{J}}}^{\perp}\boldsymbol{X}\}$ for some $\mathcal{J} : |\mathcal{J}| = J\}$. Define*

$$P_o^1(J,r) = \sigma^2\left[\{q\wedge(n-J)\}r + (m-r)r + \log\binom{n}{J}\right]. \tag{A.21}$$

Then for any $t \geq 0$,

$$\mathbb{P}[\sup_{\boldsymbol{A}\in\Gamma_{r,J}^1} \langle \boldsymbol{\mathcal{E}}, \boldsymbol{A}\rangle \geq t\sigma + \{LP_o^1(J,r)\}^{1/2}] \leq c'\exp(-ct^2), \tag{A.22}$$

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

For $\boldsymbol{I} \neq \boldsymbol{0}$ (the case $\boldsymbol{I} = \boldsymbol{0}$ being trivial), we have

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

where $[x]_+ = \max(x,0)$. Applying Lemma A.10, it then follows that for L sufficiently large

$$\mathbb{P}\{2\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}\rangle - 2aLP_o^1(J,r) - \frac{1}{a}\|\boldsymbol{I}\|_F^2 \geq \frac{1}{2}a\sigma^2 t\} \leq \frac{1}{2}c'\exp(-ct). \tag{A.23}$$

Similarly, with L large enough, we have

$$\mathbb{P}\{2\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}\boldsymbol{I}\rangle - 2aLP_o^2(J) - \frac{1}{a}\|\boldsymbol{I}\boldsymbol{I}\|_F^2 \geq \frac{1}{2}a\sigma^2 t\} \leq \frac{1}{2}c'\exp(-ct). \tag{A.24}$$

with

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

Applying the union bound to (A.23) and (A.24), and enlarging the universal constant L if necessary, yields

$$\mathbb{P}[\sup_{(\boldsymbol{B},\boldsymbol{C})\in\mathcal{B}_{r,J}} 2\langle \boldsymbol{\mathcal{E}}, \boldsymbol{X}\boldsymbol{B}+\boldsymbol{C}\rangle - \frac{1}{a}\|\boldsymbol{X}\boldsymbol{B}+\boldsymbol{C}\|_F^2 - 2aL\sigma^2\{(q+m-r)r + Jm + J\log(en/J)\} \geq a\sigma^2 t] \leq c'\exp(-ct).$$

Then, applying the preceding uniform bound to each possible value of $r(\boldsymbol{B}) \leq r$ and $J(\boldsymbol{C}) \leq J$, and taking a union bound, we have, by enlarging the universal constant L if necessary,

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

With $\lambda_{\boldsymbol{C}} = A\lambda^o_{\boldsymbol{C}}$, $\lambda_{\boldsymbol{B}} = A\lambda^o_{\boldsymbol{B}}$, $\lambda^o_{\boldsymbol{C}} = \sigma(m+\log n)^{1/2}$, $\lambda^o_{\boldsymbol{B}} = \sigma(q+m)^{1/2}$, and $A = (abA_1)^{1/2}$, it then follows that $\frac{1}{b}P_{2,0}(\boldsymbol{C};\lambda_{\boldsymbol{C}}) = \frac{1}{2}aA_1\sigma^2 J(\boldsymbol{C})(m+\log n)$, $\frac{1}{b}\sum_{i=1}^{p\wedge m} P_0(\sigma_i^{(\boldsymbol{B})};\lambda_{\boldsymbol{B}}) = \frac{1}{2}aA_1\sigma^2(q+m)r(\boldsymbol{B})$. Taking $A_1 = 4L$, we obtain $\mathbb{P}(\mathcal{A}_0) \leq c'\exp(-ct)$, which immediately yields the desired conclusion. □

A.5 Proof of Lemma A.9

Proof. Denote $L(\boldsymbol{B}) := \|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\|_F^2/2 + \tau\sum_{i=1}^{p\wedge m} P_H(\sigma_i^{(\boldsymbol{B})};\lambda)$. Let the singular value decomposition (SVD) of $\boldsymbol{B}^o \in \mathbb{R}^{p\times m}$ be $\boldsymbol{B}^o = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$, where $\boldsymbol{U} = [\boldsymbol{u}_1,\ldots,\boldsymbol{u}_p] \in \mathbb{R}^{p\times p}$ and $\boldsymbol{V} = [\boldsymbol{v}_1,\ldots,\boldsymbol{v}_m] \in \mathbb{R}^{m\times m}$ are orthogonal matrices, and $\boldsymbol{D} \in \mathbb{R}^{p\times m}$ contains the singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{p\wedge m} \geq 0$ on the main diagonal and zeros elsewhere. To characterize the i-th singular value σ_i, we define the univariate function

$$\begin{aligned} g(s) =& L(\boldsymbol{B}^o + (s-\sigma_i)\boldsymbol{u}_i\boldsymbol{v}_i^\top) \\ =& \|\boldsymbol{R}_{-i} - s\boldsymbol{X}\boldsymbol{u}_i\boldsymbol{v}_i^\top\|_F^2/2 + \tau P_H(s;\lambda) + \text{const}, \end{aligned} \tag{A.25}$$

where $\boldsymbol{R}_{-i} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{B}^o - \sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^\top)$, and const denotes terms that do not depend on s.

Since $\boldsymbol{B}^o$ is a global minimizer of $L(B)$, σ_i must be a global minimizer of $g(s)$ on $[0,+\infty)$. Let $a_i = \|\boldsymbol{X}\boldsymbol{u}_i\|_2^2$, $b_i = \langle \boldsymbol{R}_{-i}, \boldsymbol{X}\boldsymbol{u}_i\boldsymbol{v}_i^\top\rangle$. Then, (A.25) simplifies to

$$g(s) = \frac{1}{2}a_i s^2 - b_i s + \tau P_H(s;\lambda) + \text{const}. \tag{A.26}$$

Consider the case where $a_i = \|\boldsymbol{X}\boldsymbol{u}_i\|_2^2 = 0$. This implies $\boldsymbol{X}\boldsymbol{u}_i = 0$, $b_i = 0$, and $g(s) = \tau P_H(s;\lambda) + \text{const}$ which is minimized at $s = 0$. Since this satisfies the lemma, we restrict our attention to the non-trivial case $a_i > 0$ for the remainder of the proof. It is easy to show that $a_i \leq \|\boldsymbol{X}\|_2^2\|\boldsymbol{u}_i\|_2^2 = \|\boldsymbol{X}\|_2^2$. By the assumption $\tau \geq \|\boldsymbol{X}\|_2^2$, we have $a_i \leq \tau$.

To show that $\sigma_i \notin (0,\lambda)$, consider the objective function on this interval. Substituting $P_H(s;\lambda) = -s^2/2 + \lambda s$ into $g(s)$ yields

$$g(s) = \frac{1}{2}(a_i - \tau)s^2 + (\tau\lambda - b_i)s + \text{const}. \tag{A.27}$$

Since $a_i \leq \tau$, $g(s)$ is a concave function on $(0,\lambda)$. Thus, a global minimizer can always be chosen outside $(0,\lambda)$. Repeating this argument for all singular values, there exists a global minimizer such that either $\sigma_i = 0$ or $\sigma_i \geq \lambda$ for every i.

Next, we focus on the case where $\sigma_i \geq \lambda$. When $s \in [\lambda,+\infty)$, $P_H(s;\lambda) = \lambda^2/2$. Thus,

$$g(s) = \frac{1}{2}a_i s^2 - b_i s + \frac{\tau\lambda^2}{2} + \text{const}. \tag{A.28}$$

Let $t^* = b_i/a_i$ denote the unconstrained minimizer of this convex quadratic function. We now show that t^* must lie in $[\lambda,\infty)$. Indeed, if $t^* < \lambda$, then $b_i < a_i\lambda$ and $g(s)$ would be monotonically

increasing on $[\lambda, \infty)$, forcing the minimizer to be $\sigma_i = \lambda$. Since $\sigma_i = \lambda$ is the global minimizer of $g(s)$, we have $g(\lambda) \leq g(0)$. However, using $b_i < a_i\lambda$, we obtain

$$\frac{1}{2}a_i\lambda^2 - b_i\lambda + \frac{\tau\lambda^2}{2} \leq 0, \tag{A.29}$$

$$\frac{1}{2}a_i\lambda^2 - (a_i\lambda)\lambda + \frac{\tau\lambda^2}{2} < 0. \tag{A.30}$$

This implies $\tau < a_i$. It contradicts $\tau \geq a_i$. Thus, we conclude that $t^* \geq \lambda$ and $\sigma_i = t^* = b_i/a_i$. Using the global optimality condition $g(\sigma_i) \leq g(0)$ and substituting $b_i = a_i\sigma_i$, we have

$$\frac{1}{2}a_i\sigma_i^2 - (a_i\sigma_i)\sigma_i + \frac{\tau\lambda^2}{2} \leq 0. \tag{A.31}$$

It follows that $\sigma_i \geq \frac{\lambda\sqrt{\tau}}{\sqrt{a_i}} \geq \frac{\lambda\sqrt{\tau}}{\|\boldsymbol{X}\|_2}$. This completes the proof. □

A.6 Proof of Theorem 4.4

Assume the density of $\boldsymbol{Y} \in \boldsymbol{\mathcal{Y}}^{n\times m} \subset \mathbb{R}^{n\times m}$ given $\boldsymbol{\eta}^* = \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*$ is given by

$$p_{\boldsymbol{\eta}^*}(\boldsymbol{Y}) = \exp\{(\langle \boldsymbol{Y}, \boldsymbol{\eta}^*\rangle - b(\boldsymbol{\eta}^*))/\sigma^2\} \tag{A.32}$$

with respect to the measure ν defined on $\boldsymbol{\mathcal{Y}}^{n\times m}$. The corresponding loss for $\boldsymbol{\eta}$ is

$$l(\boldsymbol{\eta}; \boldsymbol{Y}) = \{-\langle \boldsymbol{Y}, \boldsymbol{\eta}\rangle + b(\boldsymbol{\eta})\}/\sigma^2. \tag{A.33}$$

Assume the natural parameter space $\Omega = \{\boldsymbol{\eta} \in \mathbb{R}^{n\times m} : b(\boldsymbol{\eta}) < +\infty\}$ is open. Then l corresponds to a distribution in the regular exponential dispersion family with dispersion σ^2, natural parameter $\boldsymbol{\eta}$ and cumulant function $b(\cdot)$. We next turn to the proof of Theorem 4.4.

Proof. We first introduce a useful lemma (Lemma 3(iii) in [53]).

Lemma A.11. *For $p_{\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}}$ in the regular exponential family, the Kullback-Leibler divergence of $p_{\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}_2}$ from $p_{\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}_1}$ satisfies*

$$\mathrm{KL}(p_{\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}_1} \| p_{\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}_2}) = \boldsymbol{\Delta}_l(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}_2, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}_1). \tag{A.34}$$

To prove the desired minimax rate, we consider two cases.

Case (i) $J^*m + J^*\log(en/J^*) > (q + m - r^*)r^*$. Construct a signal subclass

$$\begin{aligned}\mathcal{B}^1 = \{\boldsymbol{B} = \boldsymbol{0}, \boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n]^\top : \boldsymbol{c}_i = \boldsymbol{0} \text{ or } c_{ij} \in \{0, \gamma R\}, 1 \leq j \leq \lceil m/2\rceil,\\ c_{ij} = \gamma R, \lceil m/2\rceil < j \leq m, J(\boldsymbol{C}) \leq J^*\},\end{aligned}$$

where $R = \sigma\{1 + \log(en/J^*)/m\}^{1/2}/\kappa_{\boldsymbol{C}}^{1/2}$ and γ is a small positive constant. It is easy to see that $\mathcal{B}^1 \subset \mathcal{B}_{r^*,J^*}$. Let $d(\bar{\boldsymbol{B}}_1, \bar{\boldsymbol{B}}_2) = \|\mathrm{vec}(\bar{\boldsymbol{B}}_1) - \mathrm{vec}(\bar{\boldsymbol{B}}_2)\|_0$, the Hamming distance between vectorized $\bar{\boldsymbol{B}}_1$ and $\bar{\boldsymbol{B}}_2$. Applying Lemma A.3 in [44] and Lemma 2.9 in [60], there exists a subset $\mathcal{B}^{10} \subset \mathcal{B}^1$ such that $\bar{\boldsymbol{B}}_0 = \boldsymbol{0} \in \mathcal{B}^{10}$ and

$$\log|\mathcal{B}^{10}| \geq c_1(J^*m + J^*\log(en/J^*)) \text{ and } d(\bar{\boldsymbol{B}}_1, \bar{\boldsymbol{B}}_2) \geq c_2J^*m, \forall \bar{\boldsymbol{B}}_1, \bar{\boldsymbol{B}}_2 \in \mathcal{B}^{10}, \bar{\boldsymbol{B}}_1 \neq \bar{\boldsymbol{B}}_2, \tag{A.35}$$

with some universal constants $c_1, c_2 > 0$. Then, we have

$$\begin{aligned}&\|\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}_1 - \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}_2\|_F^2 = \|\boldsymbol{C}_1 - \boldsymbol{C}_2\|_F^2\\ =&\gamma^2R^2d(\boldsymbol{C}_1, \boldsymbol{C}_2) \geq \gamma^2c_2J^*m\sigma^2\{1 + \log(en/J^*)/m\}/\kappa_{\boldsymbol{C}}\\ =&c_2\gamma^2\sigma^2(J^*m + J^*\log(en/J^*))/\kappa_{\boldsymbol{C}},\end{aligned} \tag{A.36}$$

for any $\bar{\boldsymbol{B}}_1, \bar{\boldsymbol{B}}_2 \in \mathcal{B}^{10}, \bar{\boldsymbol{B}}_1 \neq \bar{\boldsymbol{B}}_2$.

From Lemma A.11, and the regularity condition $\sigma^2\boldsymbol{\Delta}_l(\boldsymbol{0}, \boldsymbol{C}) \leq \kappa_{\boldsymbol{C}}\mathbf{D}_2(\boldsymbol{0}, \boldsymbol{C})$, we have

$$\begin{aligned}&\mathrm{KL}(p_{\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}}, p_{\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}_0}) = \boldsymbol{\Delta}_l(\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}_0, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}) = \boldsymbol{\Delta}_l(\boldsymbol{0}, \boldsymbol{C}) \leq \kappa_{\boldsymbol{C}}\mathbf{D}_2(\boldsymbol{0}, \boldsymbol{C})/\sigma^2\\ =&\kappa_{\boldsymbol{C}}\gamma^2R^2d(\boldsymbol{0}, \boldsymbol{C})/(2\sigma^2) \leq \kappa_{\boldsymbol{C}}\gamma^2R^2J^*m/(2\sigma^2) = \gamma^2(J^*m + J^*\log(en/J^*))/2,\end{aligned}$$

for any $\bar{\boldsymbol{B}} \in \mathcal{B}^{10}$. Therefore,

$$\frac{1}{|\mathcal{B}^{10}|-1} \sum_{\bar{\boldsymbol{B}} \in \mathcal{B}^{10} \setminus \{\mathbf{0}\}} \mathrm{KL}(p_{\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}}, p_{\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}_0}) \leq \gamma^2 (J^* m + J^* \log(en/J^*))/2. \tag{A.37}$$

Combining (A.36) and (A.37) and choosing a sufficiently small γ, Theorem 2.7 in [60] can be applied to obtain the lower bound.

Case (ii) $(q+m-r^*)r^* \geq J^* m + J^* \log(en/J^*)$. Define $\tilde{\mathcal{B}}_{r^*,J^*} := \{(\boldsymbol{A}, \boldsymbol{C}) \in \mathbb{R}^{q\times m} \times \mathbb{R}^{n\times m} : r(\boldsymbol{A}) \leq r^*, J(\boldsymbol{C}) \leq J^*\}$. For any estimator $(\hat{\boldsymbol{B}}, \hat{\boldsymbol{C}})$, let $\hat{\boldsymbol{A}} = \boldsymbol{V}^\top \hat{\boldsymbol{B}}$. Then, we have

$$\begin{aligned}
&\sup_{(\boldsymbol{B}^*, \boldsymbol{C}^*) \in \mathcal{B}_{r^*,J^*}} \mathbb{P}(\|\boldsymbol{X}\boldsymbol{B}^* + \boldsymbol{C}^* - \boldsymbol{X}\hat{\boldsymbol{B}} - \hat{\boldsymbol{C}}\|_F^2 \geq c P_o(J^*, r^*)) \\
\geq &\sup_{(\boldsymbol{A}^*, \boldsymbol{C}^*) \in \tilde{\mathcal{B}}_{r^*,J^*}} \mathbb{P}(\|\boldsymbol{Z}\boldsymbol{A}^* + \boldsymbol{C}^* - \boldsymbol{Z}\hat{\boldsymbol{A}} - \hat{\boldsymbol{C}}\|_F^2 \geq c P_o(J^*, r^*)),
\end{aligned}$$

where $c > 0$ is a universal constant. Consequently, we can restrict our attention to the new design matrix $\boldsymbol{Z}$ and coefficient matrix $\boldsymbol{A}$ in the following analysis. Denote $\bar{\boldsymbol{Z}} := [\boldsymbol{Z}\ \boldsymbol{I}]$, $\bar{\boldsymbol{A}} := [\boldsymbol{A}^\top\ \boldsymbol{C}^\top]^\top$. Let $s = \lceil r^*/2 \rceil$ and $t = \lfloor r^*/2 \rfloor$ if $m \geq q$, and let $s = \lfloor r^*/2 \rfloor$ and $t = \lceil r^*/2 \rceil$ otherwise. Consider a signal subclass

$$\begin{aligned}
\mathcal{B}^2 = \{\boldsymbol{A} = (a_{ij}), \boldsymbol{C} = \mathbf{0} : a_{ij} \in \{0, \gamma R\} \text{ if } 1 \leq i \leq q,\ 1 \leq j \leq t \\
\text{or } 1 \leq i \leq s,\ 1 \leq j \leq m, \text{ and } a_{ij} = 0 \text{ otherwise}\},
\end{aligned}$$

where $R = \sigma/\bar{\kappa}^{1/2}$ and γ is a positive constant. Since $r(\boldsymbol{A}) \leq s + t = r^*$, we have $\mathcal{B}^2 \subset \tilde{\mathcal{B}}_{r^*,J^*}$. The number of free entries is $d = sm + tq - st$, which satisfies $d \geq \frac{r^*}{2}\left(q + m - \frac{r^*}{2}\right) \geq 8$ and $d \asymp r^*(q+m-r^*)$. By Lemma 2.9 in [60], there exists a subset $\mathcal{B}^{20} \subset \mathcal{B}^2$ such that $\bar{\boldsymbol{A}}_0 = \mathbf{0} \in \mathcal{B}^{20}$ and

$$\log|\mathcal{B}^{20}| \geq c_1 r^*(q+m-r^*) \text{ and } d(\bar{\boldsymbol{A}}_1, \bar{\boldsymbol{A}}_2) \geq c_2 r^*(q+m-r^*), \forall \bar{\boldsymbol{A}}_1, \bar{\boldsymbol{A}}_2 \in \mathcal{B}^{20}, \bar{\boldsymbol{A}}_1 \neq \bar{\boldsymbol{A}}_2, \tag{A.38}$$

where $c_1, c_2 > 0$ are some universal constants. Then, with the regularity condition, we obtain

$$\begin{aligned}
&\|\bar{\boldsymbol{Z}}\bar{\boldsymbol{A}}_1 - \bar{\boldsymbol{Z}}\bar{\boldsymbol{A}}_2\|_F^2 = \|\boldsymbol{Z}\boldsymbol{A}_1 - \boldsymbol{Z}\boldsymbol{A}_2\|_F^2 \geq \underline{\kappa}\|\boldsymbol{A}_1 - \boldsymbol{A}_2\|_F^2 \\
=&\underline{\kappa}\gamma^2 R^2 d(\bar{\boldsymbol{A}}_1, \bar{\boldsymbol{A}}_2) \geq \underline{\kappa}\gamma^2 R^2 c_2 r^*(q+m-r^*) \\
=&\underline{\kappa}\gamma^2\sigma^2 c_2 r^*(q+m-r^*)/\bar{\kappa},
\end{aligned} \tag{A.39}$$

for any $\bar{\boldsymbol{A}}_1, \bar{\boldsymbol{A}}_2 \in \mathcal{B}^{20}, \bar{\boldsymbol{A}}_1 \neq \bar{\boldsymbol{A}}_2$.

According to Lemma A.11 and regularity conditions $\underline{\kappa}\mathbf{D}_2(\boldsymbol{A}_1, \boldsymbol{A}_2) \leq \mathbf{D}_2(\boldsymbol{Z}\boldsymbol{A}_1, \boldsymbol{Z}\boldsymbol{A}_2)$ and $\sigma^2\boldsymbol{\Delta}_l(\mathbf{0}, \boldsymbol{Z}\boldsymbol{A}_1) \leq \bar{\kappa}\mathbf{D}_2(\mathbf{0}, \boldsymbol{A}_1), \forall r(\boldsymbol{A}_i) \leq r^*$, we have

$$\begin{aligned}
&\mathrm{KL}(p_{\bar{\boldsymbol{Z}}\bar{\boldsymbol{A}}}, p_{\bar{\boldsymbol{Z}}\bar{\boldsymbol{A}}_0}) = \boldsymbol{\Delta}_l(\bar{\boldsymbol{Z}}\bar{\boldsymbol{A}}_0, \bar{\boldsymbol{Z}}\bar{\boldsymbol{A}}) = \boldsymbol{\Delta}_l(\mathbf{0}, \boldsymbol{Z}\boldsymbol{A}) \leq \bar{\kappa}\mathbf{D}_2(\mathbf{0}, \boldsymbol{A})/\sigma^2 \\
=&\bar{\kappa}\gamma^2 R^2 d(\mathbf{0}, \boldsymbol{A})/(2\sigma^2) \leq \gamma^2(q+m-r^*)r^*,
\end{aligned} \tag{A.40}$$

for any $\bar{\boldsymbol{A}} \in \mathcal{B}^{20}$. Therefore,

$$\frac{1}{|\mathcal{B}^{20}|-1} \sum_{\bar{\boldsymbol{A}} \in \mathcal{B}^{20} \setminus \{\mathbf{0}\}} \mathrm{KL}(p_{\bar{\boldsymbol{Z}}\bar{\boldsymbol{A}}}, p_{\bar{\boldsymbol{Z}}\bar{\boldsymbol{A}}_0}) \leq \gamma^2(q+m-r^*)r^*. \tag{A.41}$$

Provided that γ is chosen sufficiently small, we can combine (A.39) and (A.41) and invoke Theorem 2.7 of [60] to establish the desired lower bound. The proof is complete. □

A.7 Proof of Theorem 4.5

Proof. From the optimality of $\hat{\bar{\boldsymbol{B}}}$, we have

$$\boldsymbol{\Delta}_l(\bar{\boldsymbol{X}}\hat{\bar{\boldsymbol{B}}}, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*) + A\sigma^2 P_m(\hat{\bar{\boldsymbol{B}}}) \leq A\sigma^2 P_m(\bar{\boldsymbol{B}}^*) + \langle \boldsymbol{\mathcal{E}}, \bar{\boldsymbol{X}}\hat{\bar{\boldsymbol{B}}} - \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^* \rangle. \tag{A.42}$$

We will show that the stochastic term $\langle \boldsymbol{\mathcal{E}}, \bar{\boldsymbol{X}}\hat{\bar{\boldsymbol{B}}} - \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^* \rangle$ can be upper bounded by $\mathbf{D}_2(\bar{\boldsymbol{X}}\hat{\bar{\boldsymbol{B}}}, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*)$ and $\sigma^2(P_m(\bar{\boldsymbol{B}}^*)+P_m(\hat{\bar{\boldsymbol{B}}}))$ up to some multiplicative constants. Let $\boldsymbol{\Delta_B} = \hat{\boldsymbol{B}}-\boldsymbol{B}^*$, $\boldsymbol{\Delta_C} = \hat{\boldsymbol{C}}-\boldsymbol{C}^*$, $\hat{r} = r(\hat{\boldsymbol{B}})$, $r^* = r(\boldsymbol{B}^*)$, $\hat{\mathcal{J}} = \mathcal{J}(\hat{\boldsymbol{C}})$, $\mathcal{J}^* = \mathcal{J}(\boldsymbol{C}^*)$, $\hat{J} = |\hat{\mathcal{J}}|$, and $J^* = |\mathcal{J}^*|$. Let $\mathcal{P}_r$ be the orthogonal projection onto the row space of $\boldsymbol{X}\boldsymbol{B}^*$ and $\mathcal{P}_r^\perp$ be its orthogonal complement.

Decompose $\boldsymbol{X}\boldsymbol{\Delta_B} + \boldsymbol{\Delta_C}$ as follows

$$
\begin{aligned}
&\boldsymbol{X}\boldsymbol{\Delta_B} + \boldsymbol{\Delta_C} \\
=&\mathcal{P}_{I_{\mathcal{J}^*}}(\boldsymbol{X}\boldsymbol{\Delta_B} + \boldsymbol{\Delta_C}) + \mathcal{P}_{I_{(\mathcal{J}^*)^c\cap\hat{\mathcal{J}}}}(\boldsymbol{X}\boldsymbol{\Delta_B} + \boldsymbol{\Delta_C}) + \mathcal{P}_{I_{(\mathcal{J}^*\cup\hat{\mathcal{J}})^c}}(\boldsymbol{X}\boldsymbol{\Delta_B} + \boldsymbol{\Delta_C}) \\
=&\mathcal{P}_{I_{\mathcal{J}^*}}(\boldsymbol{X}\boldsymbol{\Delta_B} + \boldsymbol{\Delta_C}) + \mathcal{P}_{I_{(\mathcal{J}^*)^c\cap\hat{\mathcal{J}}}}(\boldsymbol{X}\boldsymbol{\Delta_B} + \boldsymbol{\Delta_C}) + \mathcal{P}_{I_{(\mathcal{J}^*\cup\hat{\mathcal{J}})^c}}(\boldsymbol{X}\boldsymbol{\Delta_B} + \boldsymbol{\Delta_C})\mathcal{P}_r \\
&+ \mathcal{P}_{I_{(\mathcal{J}^*\cup\hat{\mathcal{J}})^c}}(\boldsymbol{X}\boldsymbol{\Delta_B} + \boldsymbol{\Delta_C})\mathcal{P}_r^\perp \\
\equiv&\boldsymbol{I}_1 + \boldsymbol{I}_2 + \boldsymbol{I}_3 + \boldsymbol{I}_4. \qquad \text{(A.43)}
\end{aligned}
$$

Clearly, $\|\boldsymbol{I}_1\|_F^2 + \|\boldsymbol{I}_2\|_F^2 + \|\boldsymbol{I}_3\|_F^2 + \|\boldsymbol{I}_4\|_F^2 = \|\boldsymbol{X}\boldsymbol{\Delta_B} + \boldsymbol{\Delta_C}\|_F^2$.

Next, we bound the four stochastic terms $\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}_1\rangle$, $\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}_2\rangle$, $\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}_3\rangle$ and $\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}_4\rangle$ one by one.

Since $\mathcal{J}(\boldsymbol{I}_1) \subseteq \mathcal{J}^*$, $\boldsymbol{I}_1$ is row-sparse with at most J^* nonzero rows. Following the row-sparse argument used in the proof of Lemma A.4, define

$$\Gamma_J^2 = \left\{\boldsymbol{A} \in \mathbb{R}^{n\times m} : \|\boldsymbol{A}\|_F \le 1,\ \mathcal{J}(\boldsymbol{A}) \subseteq \mathcal{J} \text{ for some } \mathcal{J} \subset [n],\ |\mathcal{J}| = J\right\},$$

and

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

For any $a, a', b > 0$, when $\boldsymbol{I}_1 \neq \boldsymbol{0}$,

$$
\begin{aligned}
&\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}_1\rangle - \frac{1}{a}\|\boldsymbol{I}_1\|_F^2 - bLP_o^2(J^*) \\
\le& \|\boldsymbol{I}_1\|_F[\langle \boldsymbol{\mathcal{E}}, \frac{\boldsymbol{I}_1}{\|\boldsymbol{I}_1\|_F}\rangle - 2(\frac{b}{a})^{1/2}\{LP_o^2(J^*)\}^{1/2}] \qquad \text{(A.44)} \\
\le& \frac{1}{a'}\|\boldsymbol{I}_1\|_F^2 + \frac{a'}{4}R',
\end{aligned}
$$

where

$$R' = \sup_{1\le J\le n}\sup_{\boldsymbol{A}\in\Gamma_J^2}[\langle \boldsymbol{\mathcal{E}}, \boldsymbol{A}\rangle - 2(\frac{b}{a})^{1/2}\{LP_o^2(J)\}^{1/2}]_+^2.$$

The same inequality trivially holds when $\boldsymbol{I}_1 = \boldsymbol{0}$. Therefore,

$$\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}_1\rangle - \frac{1}{a}\|\boldsymbol{I}_1\|_F^2 - bLP_o^2(J^*) \le \frac{1}{a'}\|\boldsymbol{I}_1\|_F^2 + \frac{a'}{4}R'. \qquad \text{(A.45)}$$

By the same row-sparse concentration argument as in the proof of Lemma A.4, for sufficiently large L and $4b > a$, we have $\mathbb{E}(R') \le C\sigma^2$, where $C > 0$ is a universal constant.

Similarly, since $\mathcal{J}(\boldsymbol{I}_2) \subseteq (\mathcal{J}^*)^c \cap \hat{\mathcal{J}} \subseteq \hat{\mathcal{J}}$, $\boldsymbol{I}_2$ has at most $\hat{J}$ nonzero rows. Applying the same argument gives

$$\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}_2\rangle - \frac{1}{a}\|\boldsymbol{I}_2\|_F^2 - bLP_o^2(\hat{J}) \le \frac{1}{a'}\|\boldsymbol{I}_2\|_F^2 + \frac{a'}{4}R'. \qquad \text{(A.46)}$$

To handle $\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}_3\rangle$, we first introduce Lemma 10 in [51].

Lemma A.12. *Suppose* $\mathrm{vec}(\boldsymbol{\mathcal{E}})$ *is sub-Gaussian with mean zero and* ψ_2*-norm bounded by* σ*. Given* $\boldsymbol{X} \in \mathbb{R}^{n\times p}$*,* $1 \le r \le m \wedge p$*,* $1 \le J_1 \le n$*,* $1 \le J_2 \le n$*, define* $\Gamma_{r,J_1,J_2} := \{\boldsymbol{A} \in \mathbb{R}^{n\times m} : \|\boldsymbol{A}\|_F \le 1, r(\boldsymbol{A}) \le r, \mathrm{CS}(\boldsymbol{A}) \subseteq \mathrm{CS}\{\mathcal{P}_{I_{(\mathcal{J}_1\cup\mathcal{J}_2)^c}}\boldsymbol{X}\}$ *for some* $\mathcal{J}_1, \mathcal{J}_2 : |\mathcal{J}_1| = J_1, |\mathcal{J}_2| = J_2\}$*. Define*

$$P_o(J_1, J_2, r) = \sigma^2\left[(q+m-r)r + \log\binom{n}{J_1} + \log\binom{n}{J_2}\right]. \qquad \text{(A.47)}$$

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

$$\mathbb{P}[\sup_{\boldsymbol{A}\in\Gamma_{r,J_1,J_2}}\langle \boldsymbol{\mathcal{E}}, \boldsymbol{A}\rangle \ge t\sigma + \{LP_o(J_1, J_2, r)\}^{1/2}] \le c'\exp(-ct^2), \qquad \text{(A.48)}$$

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

Let $\mathcal{S} = \mathcal{J}^* \cup \hat{\mathcal{J}}$. Since $\mathcal{P}_{I_{\mathcal{S}^c}} \boldsymbol{\Delta_C} = \mathbf{0}$, we have $\boldsymbol{I}_3 = \mathcal{P}_{I_{\mathcal{S}^c}} \boldsymbol{X} \boldsymbol{\Delta_B} \mathcal{P}_r$. Therefore, $\mathrm{CS}(\boldsymbol{I}_3) \subseteq \mathrm{CS}(\mathcal{P}_{I_{\mathcal{S}^c}} \boldsymbol{X}), r(\boldsymbol{I}_3) \leq r(\mathcal{P}_r) = r(\boldsymbol{X}\boldsymbol{B}^*) \leq r^*$. Hence, Lemma A.12 can be applied to bound $\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}_3 \rangle$. For any $a, a', b > 0$,

$$
\begin{aligned}
&\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}_3 \rangle - \frac{1}{a}\|\boldsymbol{I}_3\|_F^2 - bLP_o(J^*, \hat{J}, r^*) \\
\leq &\|\boldsymbol{I}_3\|_F \langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}_3/\|\boldsymbol{I}_3\|_F \rangle - 2(\frac{b}{a})^{1/2}\|\boldsymbol{I}_3\|_F \{LP_o(J^*, \hat{J}, r^*)\}^{1/2} \qquad \text{(A.49)} \\
\leq &\frac{1}{a'}\|\boldsymbol{I}_3\|_F^2 + \frac{a'}{4} \sup_{1\leq J_1\leq n, 1\leq J_2\leq n, 1\leq r\leq m\wedge p} \sup_{\boldsymbol{A}\in\Gamma_{r,J_1,J_2}} [\langle \boldsymbol{\mathcal{E}}, \boldsymbol{A} \rangle - 2(\frac{b}{a})^{1/2}\{LP_o(J_1, J_2, r)\}^{1/2}]_+^2.
\end{aligned}
$$

Let $R'' = \sup_{1\leq J_1\leq n, 1\leq J_2\leq n, 1\leq r\leq m\wedge p} \sup_{\boldsymbol{A}\in\Gamma_{r,J_1,J_2}} [\langle \boldsymbol{\mathcal{E}}, \boldsymbol{A} \rangle - 2(\frac{b}{a})^{1/2}\{LP_o(J_1, J_2, r)\}^{1/2}]_+^2$. With the definition of R'', we have

$$
\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}_3 \rangle - \frac{1}{a}\|\boldsymbol{I}_3\|_F^2 - bLP_o(J^*, \hat{J}, r^*) \leq \frac{1}{a'}\|\boldsymbol{I}_3\|_F^2 + \frac{a'}{4}R''. \qquad \text{(A.50)}
$$

With $4b > a$, from Lemma A.12, we have $\mathbb{E}(R'') \leq C\sigma^2$ where C is a universal constant.

Moreover, since $\boldsymbol{X}\boldsymbol{B}^*\mathcal{P}_r^\perp = \mathbf{0}$ and $\mathcal{P}_{I_{\mathcal{S}^c}} \boldsymbol{\Delta_C} = \mathbf{0}$, we have $\boldsymbol{I}_4 = \mathcal{P}_{I_{\mathcal{S}^c}} \boldsymbol{X}\hat{\boldsymbol{B}}\mathcal{P}_r^\perp$. Thus, $\mathrm{CS}(\boldsymbol{I}_4) \subseteq \mathrm{CS}(\mathcal{P}_{I_{\mathcal{S}^c}} \boldsymbol{X}), r(\boldsymbol{I}_4) \leq r(\hat{\boldsymbol{B}}) = \hat{r}$. Therefore, the same argument can be applied to $\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}_4 \rangle$ to obtain

$$
\langle \boldsymbol{\mathcal{E}}, \boldsymbol{I}_4 \rangle - \frac{1}{a}\|\boldsymbol{I}_4\|_F^2 - bLP_o(J^*, \hat{J}, \hat{r}) \leq \frac{1}{a'}\|\boldsymbol{I}_4\|_F^2 + \frac{a'}{4}R''. \qquad \text{(A.51)}
$$

The boundary cases with zero support sizes are handled by extending the corresponding suprema and concentration bounds to include empty support sets, with $\binom{n}{0} = 1$ and $J\log(en/J) = 0$ at $J = 0$. When $r^* = 0$ or $\hat{r} = 0$, the corresponding terms $\boldsymbol{I}_3$ or $\boldsymbol{I}_4$ vanish, respectively. In summary, we have

$$
\begin{aligned}
&\langle \boldsymbol{\mathcal{E}}, \boldsymbol{X}\boldsymbol{\Delta_B} + \boldsymbol{\Delta_C} \rangle \\
\leq &(\frac{1}{a} + \frac{1}{a'})\|\boldsymbol{X}\boldsymbol{\Delta_B} + \boldsymbol{\Delta_C}\|_F^2 + \frac{1}{2}a'(R' + R'') \\
&+ bL(P_o^2(J^*) + P_o^2(\hat{J}) + P_o(J^*, \hat{J}, r^*) + P_o(J^*, \hat{J}, \hat{r})) \\
\leq &(\frac{1}{a} + \frac{1}{a'})\|\boldsymbol{X}\boldsymbol{\Delta_B} + \boldsymbol{\Delta_C}\|_F^2 + \frac{1}{2}a'(R' + R'') + 4bL\sigma^2(P_m(\hat{\boldsymbol{B}}, \hat{\boldsymbol{C}}) + P_m(\boldsymbol{B}^*, \boldsymbol{C}^*)). \qquad \text{(A.52)}
\end{aligned}
$$

Combining the above inequality with (A.42) and using the regularity condition, we obtain

$$
\begin{aligned}
&[\delta - 2(\frac{1}{a} + \frac{1}{a'})]\mathbf{D}_2(\bar{\boldsymbol{X}}\hat{\bar{\boldsymbol{B}}}, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*) \\
&\leq \sigma^2(K + A + 4bL)P_m(\bar{\boldsymbol{B}}^*) + \sigma^2(K - A + 4bL)P_m(\hat{\bar{\boldsymbol{B}}}) + \frac{a'}{2}(R' + R'').
\end{aligned} \qquad \text{(A.53)}
$$

Choose $a, a', b > 0$ such that $\frac{1}{a} + \frac{1}{a'} = \frac{\delta}{4}$, $4b > a$, and take $A \geq K + 4bL$. Then, we obtain

$$
\frac{\delta}{2}\mathbf{D}_2(\bar{\boldsymbol{X}}\hat{\bar{\boldsymbol{B}}}, \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}^*) \leq 2A\sigma^2 P_m(\bar{\boldsymbol{B}}^*) + \frac{a'}{2}(R' + R''). \qquad \text{(A.54)}
$$

Taking expectations on both sides, we obtain the conclusion. □

Remark A.13. If $\bar{\boldsymbol{B}}^*$ is selected as a minimizer of (18) over the candidate set $\mathcal{M}$, the prediction error is trivially zero. Otherwise, the selected minimizer is a candidate model produced by RoLL. Theorem 4.5 establishes a prediction error bound for PIC selection over $\mathcal{M}$, covering both cases.

A.8 Proof of Theorem 5.1

Proof. The desired result can be obtained from the proof of part (ii) of Theorem 7 in [53]. The primary modification involves replacing a constant μ with an arbitrary nonnegative sequence $\{\mu_t\}_{t\geq 0}$.

Define $h_t(\cdot) = F(\cdot) - \boldsymbol{\Delta}_{\bar{\psi}_t}(\cdot, \bar{\boldsymbol{M}}_t)$. Let $\bar{\boldsymbol{B}}_t' = \theta_t\bar{\boldsymbol{B}} + (1 - \theta_t)\bar{\boldsymbol{B}}_t$ with θ_t to be determined. First, with Lemma A.2 in [53], we have

$$
\begin{aligned}
&\boldsymbol{\Delta}_{h_t}(\bar{\boldsymbol{B}}_t', \bar{\boldsymbol{B}}_{t+1}) + \rho_t\mathbf{D}_2(\bar{\boldsymbol{B}}_t', \bar{\boldsymbol{B}}_{t+1}) \\
\leq &h_t(\bar{\boldsymbol{B}}_t') + \rho_t\mathbf{D}_2(\bar{\boldsymbol{B}}_t', \bar{\boldsymbol{M}}_t) - h_t(\bar{\boldsymbol{B}}_{t+1}) - \rho_t\mathbf{D}_2(\bar{\boldsymbol{B}}_{t+1}, \bar{\boldsymbol{M}}_t).
\end{aligned} \qquad \text{(A.55)}
$$

By the definition of $\mathbf{C}_{h_t}$,

$$\begin{aligned}-F(\bar{\boldsymbol{B}}_t') = &-\theta_t F(\bar{\boldsymbol{B}}) + \theta_t \boldsymbol{\Delta}_{\bar{\psi}_t}(\bar{\boldsymbol{B}}, \bar{\boldsymbol{M}}_t) - (1-\theta_t)F(\bar{\boldsymbol{B}}_t) + (1-\theta_t)\boldsymbol{\Delta}_{\bar{\psi}_t}(\bar{\boldsymbol{B}}_t, \bar{\boldsymbol{M}}_t) \\ &- \boldsymbol{\Delta}_{\bar{\psi}_t}(\bar{\boldsymbol{B}}_t', \bar{\boldsymbol{M}}_t) + \mathbf{C}_{h_t}(\bar{\boldsymbol{B}}, \bar{\boldsymbol{B}}_t, \theta_t).\end{aligned} \tag{A.56}$$

Plugging the last equality into (A.55) yields

$$\begin{aligned}&F(\bar{\boldsymbol{B}}_{t+1}) - F(\bar{\boldsymbol{B}}) + (\mu_t + \rho_t)\mathbf{D}_2(\bar{\boldsymbol{B}}_t', \bar{\boldsymbol{B}}_{t+1}) + \mu_t\theta_t(1-\theta_t)\mathbf{D}_2(\bar{\boldsymbol{B}}, \bar{\boldsymbol{B}}_t) + R_t + \theta_t\mathcal{E}_t(\bar{\boldsymbol{B}}) \\ \leq &(1-\theta_t)(F(\bar{\boldsymbol{B}}_t) - F(\bar{\boldsymbol{B}})) + \rho_t\mathbf{D}_2(\bar{\boldsymbol{B}}_t', \bar{\boldsymbol{M}}_t),\end{aligned} \tag{A.57}$$

where R_t is given by

$$R_t = (\rho_t\mathbf{D}_2 - \boldsymbol{\Delta}_{\bar{\psi}_t})(\bar{\boldsymbol{B}}_{t+1}, \bar{\boldsymbol{M}}_t) + (1-\theta_t)\boldsymbol{\Delta}_{\bar{\psi}_t}(\bar{\boldsymbol{B}}_t, \bar{\boldsymbol{M}}_t), \tag{A.58}$$

and $\mathcal{E}_t(\bar{\boldsymbol{B}}) = \boldsymbol{\Delta}_{\bar{\psi}_t'}(\bar{\boldsymbol{B}}, \bar{\boldsymbol{M}}_t) + (\mathbf{C}_P(\bar{\boldsymbol{B}}, \bar{\boldsymbol{B}}_t, \theta_t) + \boldsymbol{\Delta}_P(\theta_t\bar{\boldsymbol{B}} + (1-\theta_t)\bar{\boldsymbol{B}}_t, \bar{\boldsymbol{B}}_{t+1}))/\theta_t$, using the idempotence and linear properties of $\mathbf{C}$ and $\boldsymbol{\Delta}$ as derived in [53] and the differentiability of f and $\|\cdot\|_F^2/2$.

We would like to relate $(\mu_t+\rho_t)\mathbf{D}_2(\bar{\boldsymbol{B}}_t', \bar{\boldsymbol{B}}_{t+1}) + \mu_t\theta_t(1-\theta_t)\mathbf{D}_2(\bar{\boldsymbol{B}}, \bar{\boldsymbol{B}}_t)$, $\rho_t\mathbf{D}_2(\bar{\boldsymbol{B}}_t', \bar{\boldsymbol{M}}_t)$ to terms of the form $c_{t+1}\mathbf{D}_2(\bar{\boldsymbol{B}}, \boldsymbol{v}_{t+1})$, $d_t\mathbf{D}_2(\bar{\boldsymbol{B}}, \boldsymbol{v}_t)$ for some $\boldsymbol{v}_{t+1}$, $\boldsymbol{v}_t$ respectively with a vector sequence $\{\boldsymbol{v}_t\}_{t=0}^{\infty}$ and scalar sequences $\{c_t\}_{t=1}^{\infty}$, $\{d_t\}_{t=0}^{\infty}$. Such sequences can be found by solving the gradient equations with respect to $\bar{\boldsymbol{B}}$:

$$\begin{cases} (\mu_t+\rho_t)\theta_t[\theta_t\bar{\boldsymbol{B}} + (1-\theta_t)\bar{\boldsymbol{B}}_t - \bar{\boldsymbol{B}}_{t+1}] + \mu_t\theta_t(1-\theta_t)(\bar{\boldsymbol{B}} - \bar{\boldsymbol{B}}_t) = c_{t+1}(\bar{\boldsymbol{B}} - \boldsymbol{v}_{t+1}), \\ \rho_t\theta_t[\theta_t\bar{\boldsymbol{B}} + (1-\theta_t)\bar{\boldsymbol{B}}_t - \bar{\boldsymbol{M}}_t] = d_t(\bar{\boldsymbol{B}} - \boldsymbol{v}_t). \end{cases} \tag{A.59}$$

By matching the terms, we have $c_{t+1} = \theta_t(\rho_t\theta_t + \mu_t)$, $d_t = \rho_t\theta_t^2$, and

$$\begin{cases} \boldsymbol{v}_{t+1} = \frac{(\rho_t+\mu_t)\bar{\boldsymbol{B}}_{t+1} - \rho_t(1-\theta_t)\bar{\boldsymbol{B}}_t}{\rho_t\theta_t+\mu_t}, \\ \boldsymbol{v}_t = \frac{\bar{\boldsymbol{M}}_t - (1-\theta_t)\bar{\boldsymbol{B}}_t}{\theta_t}. \end{cases} \tag{A.60}$$

Note that the first quadratic expression is bounded below by $c_{t+1}\mathbf{D}_2(\bar{\boldsymbol{B}}, \boldsymbol{v}_{t+1})$, with a nonnegative remainder $\frac{\mu_t(\rho_t+\mu_t)(1-\theta_t)}{\rho_t\theta_t+\mu_t}\mathbf{D}_2(\bar{\boldsymbol{B}}_{t+1}, \bar{\boldsymbol{B}}_t)$, while the second quadratic expression equals $d_t\mathbf{D}_2(\bar{\boldsymbol{B}}, \boldsymbol{v}_t)$.

Combining the equalities in (A.60) results in

$$\bar{\boldsymbol{M}}_t = \bar{\boldsymbol{B}}_t + \frac{\rho_{t-1}\theta_t(1-\theta_{t-1})}{\rho_{t-1}\theta_{t-1} + \mu_{t-1}}(\bar{\boldsymbol{B}}_t - \bar{\boldsymbol{B}}_{t-1}), \tag{A.61}$$

as in (21a). Therefore, (A.57) implies

$$\begin{aligned}&F(\bar{\boldsymbol{B}}_{t+1}) - F(\bar{\boldsymbol{B}}) + \theta_t(\rho_t\theta_t + \mu_t)\mathbf{D}_2(\bar{\boldsymbol{B}}, \boldsymbol{v}_{t+1}) + R_t + \theta_t\mathcal{E}_t(\bar{\boldsymbol{B}}) \\ \leq &(1-\theta_t)[F(\bar{\boldsymbol{B}}_t) - F(\bar{\boldsymbol{B}})] + \rho_t\theta_t^2\mathbf{D}_2(\bar{\boldsymbol{B}}, \boldsymbol{v}_t), \forall t \geq 0.\end{aligned} \tag{A.62}$$

From (20), we have

$$\begin{aligned}&F(\bar{\boldsymbol{B}}_{t+1}) - F(\bar{\boldsymbol{B}}) + (\rho_t\theta_t^2 + \mu_t\theta_t)\mathbf{D}_2(\bar{\boldsymbol{B}}, \boldsymbol{v}_{t+1}) + R_t + \theta_t\mathcal{E}_t(\bar{\boldsymbol{B}}) \\ \leq &(1-\theta_t)[F(\bar{\boldsymbol{B}}_t) - F(\bar{\boldsymbol{B}}) + (\rho_{t-1}\theta_{t-1}^2 + \mu_{t-1}\theta_{t-1})\mathbf{D}_2(\bar{\boldsymbol{B}}, \boldsymbol{v}_t)], \forall t \geq 1.\end{aligned} \tag{A.63}$$

Using a recursive argument on (A.63) from $t = T$ to $t = 1$ and applying (A.62) for $t = 0$, we obtain the conclusion in the theorem. □

Remark A.14. When f satisfies $\mu\mathbf{D}_2 \leq \boldsymbol{\Delta}_f \leq \mathcal{L}\mathbf{D}_2$ with $0 < \mu \leq \mathcal{L}$, and P satisfies $\boldsymbol{\Delta}_P \geq 0$, taking $\mu_t \leq \mu$ and $\rho_t \geq \mathcal{L} - \mu_t$ ensures $\mathcal{E}_t(\bar{\boldsymbol{B}}) \geq (\boldsymbol{\Delta}_f - \mu\mathbf{D}_2)(\bar{\boldsymbol{B}}, \bar{\boldsymbol{M}}_t) \geq 0$ and $R_t \geq (\rho_t\mathbf{D}_2 - \boldsymbol{\Delta}_f + \mu_t\mathbf{D}_2)(\bar{\boldsymbol{B}}_{t+1}, \bar{\boldsymbol{M}}_t) \geq (\rho_t + \mu_t - \mathcal{L})\mathbf{D}_2(\bar{\boldsymbol{B}}_{t+1}, \bar{\boldsymbol{M}}_t) \geq 0$.

On the other hand, with $\theta_0 = 1$, under the conditions $\frac{\mu_{t-1}}{\rho_t} \geq \frac{1}{r}$, $\frac{\rho_{t-1}}{\rho_t} \geq 1 - \frac{c}{\sqrt{r}}$, for some $r \geq 1$ and a constant $0 \leq c \leq 1$, the relaxation parameters determined by (20) satisfy $\frac{1}{\theta_t} \leq$

$\frac{1}{2}(\sqrt{(c\sqrt{r}-1)^2+4r}+4\sqrt{r}+1)$. Consequently, the cumulative factor $\prod_{t=1}^{T}(1-\theta_t)$ in (22) leads to an iteration complexity of $\mathcal{O}(\sqrt{r}\log\frac{1}{\epsilon})$ where ϵ denotes the target accuracy level. Specifically, for $0<\mu<\mathcal{L}$, setting $\mu_t\equiv\mu$ and $\rho_t\equiv\mathcal{L}-\mu$ allows us to take $r=\max\{1,\kappa-1\}$ and $\theta_t\geq 2/(\sqrt{4\kappa-3}+1)$, where $\kappa=\mathcal{L}/\mu$ is the condition number. When $\bar{\boldsymbol{B}}$ is chosen as a minimizer of F in (22), both $F(\bar{\boldsymbol{B}}_{T+1})-F(\bar{\boldsymbol{B}})$ and $\mathbf{D}_2(\bar{\boldsymbol{B}},\frac{\bar{\boldsymbol{M}}_{T+1}-(1-\theta_{T+1})\bar{\boldsymbol{B}}_{T+1}}{\theta_{T+1}})$ converge to zero geometrically fast with an iteration complexity of

$$\mathcal{O}(\sqrt{\kappa}\log\frac{1}{\epsilon}), \tag{A.64}$$

which matches the optimal convergence rate for first-order accelerated methods [59]. This is significantly faster than the linear convergence rate $\mathcal{O}(\kappa\log\frac{1}{\epsilon})$ of the standard proximal gradient algorithm [2], which corresponds to the update rule in (6).

Remark A.15 (Iteration complexity). Suppose that $R_t+\theta_t\mathcal{E}_t(\bar{\boldsymbol{B}})\geq 0$ for $t=0,\ldots,T$. Then Theorem 5.1 implies $F(\bar{\boldsymbol{B}}^{T+1})-F(\bar{\boldsymbol{B}})\leq\Gamma_0(\bar{\boldsymbol{B}})\prod_{t=1}^{T}(1-\theta_t)$, where $\Gamma_0(\bar{\boldsymbol{B}})=(1-\theta_0)\{F(\bar{\boldsymbol{B}}^0)-F(\bar{\boldsymbol{B}})\}+\theta_0^2\rho_0\mathbf{D}_2(\bar{\boldsymbol{B}},\bar{\boldsymbol{B}}^0)$. Assume that F attains its minimum, and let $\bar{\boldsymbol{B}}^{\mathrm{opt}}\in\arg\min F$ and $F^{\mathrm{opt}}=F(\bar{\boldsymbol{B}}^{\mathrm{opt}})$. For a prescribed accuracy $\epsilon>0$, define $T_\epsilon=\inf\{T\geq 0: F(\bar{\boldsymbol{B}}^{T+1})-F^{\mathrm{opt}}\leq\epsilon\}$. Then, for $0<\epsilon<\Gamma_0(\bar{\boldsymbol{B}}^{\mathrm{opt}})$, we have $T_\epsilon\leq\inf\{T\geq 0:\prod_{t=1}^{T}(1-\theta_t)\leq\epsilon/\Gamma_0(\bar{\boldsymbol{B}}^{\mathrm{opt}})\}$. When f is convex and $\mathcal{L}$-smooth and P is convex, taking $\mu_t=0$, $\rho_t=\mathcal{L}$, and $\theta_0=1$ gives $\prod_{t=1}^{T}(1-\theta_t)=\theta_T^2\leq 4/(T+2)^2$. Consequently, $T_\epsilon\leq\max\{0,\lceil 2\sqrt{\mathcal{L}\mathbf{D}_2(\bar{\boldsymbol{B}}^{\mathrm{opt}},\bar{\boldsymbol{B}}^0)/\epsilon}-2\rceil\}=\mathcal{O}(\epsilon^{-1/2})$. The strongly convex case achieves an iteration complexity of $\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon))$ (cf. Remark A.14).

B Algorithm and Additional Experimental Details

B.1 RAPID algorithm

Algorithm 2 RAPID algorithm

Input: $\bar{\boldsymbol{B}}_0$, $\mathcal{L}$, T_0, μ_{T_0}, $0<\underline{c}<1$, $\bar{c}>1$, $0<c<1$, $M\in\mathbb{N}$ ($T_0=50$, $\mu_{T_0}=$ 1e-3, $\underline{c}=0.4$, $\bar{c}=2.8$, $c=$ 1e-3, $M=3$)
Initialize $\theta_0\leftarrow 1$, $t\leftarrow 0$
while not converged **do**
  if $t<T_0$ **then**
    $\mu_t\leftarrow 0$
    $\rho_t\leftarrow\mathcal{L}$
  else
    if $t=T_0$ **then**
      Initialize the line-search trial value with μ_{T_0}
    end if
    $\mu_t\leftarrow$ LINESEARCH(μ_{t-1}) if $(t-T_0)\bmod 10=0$, otherwise $\mu_t\leftarrow\mu_{t-1}$
    $\rho_t\leftarrow\mathcal{L}-\mu_t$
  end if
  if $t\geq 1$ **then**
    $\theta_t\leftarrow(\sqrt{r^2+4r}-r)/2$, where $r=(\rho_{t-1}\theta_{t-1}+\mu_{t-1})\theta_{t-1}/\rho_t$
  end if
  if $t=0$ **then**
    $\bar{\boldsymbol{M}}_t\leftarrow\bar{\boldsymbol{B}}_t$
  else
    $\bar{\boldsymbol{M}}_t\leftarrow\bar{\boldsymbol{B}}_t+\frac{\rho_{t-1}\theta_t(1-\theta_{t-1})}{\rho_{t-1}\theta_{t-1}+\mu_{t-1}}(\bar{\boldsymbol{B}}_t-\bar{\boldsymbol{B}}_{t-1})$
  end if
  $\boldsymbol{B}_{t+1}\leftarrow\mathrm{prox}_{P_{\boldsymbol{B}}^{s}(\cdot;\lambda_{\boldsymbol{B}})/\mathcal{L}}\big(\boldsymbol{M}_t^{\boldsymbol{B}}-\boldsymbol{X}^\top\nabla l(\boldsymbol{X}\boldsymbol{M}_t^{\boldsymbol{B}}+\boldsymbol{M}_t^{\boldsymbol{C}};\boldsymbol{Y})/\mathcal{L}\big)$
  $\boldsymbol{C}_{t+1}\leftarrow\mathrm{prox}_{P_{\boldsymbol{C}}^{s}(\cdot;\lambda_{\boldsymbol{C}})/\mathcal{L}}\big(\boldsymbol{M}_t^{\boldsymbol{C}}-\nabla l(\boldsymbol{X}\boldsymbol{M}_t^{\boldsymbol{B}}+\boldsymbol{M}_t^{\boldsymbol{C}};\boldsymbol{Y})/\mathcal{L}\big)$
  $\bar{\boldsymbol{B}}_{t+1}\leftarrow[\boldsymbol{B}_{t+1}^\top\ \boldsymbol{C}_{t+1}^\top]^\top$
  $t\leftarrow t+1$
end while
Return: $\bar{\boldsymbol{B}}_t$

B.2 Real-world Datasets Description

Table B.1: Characteristics of the datasets used in the experiments. Datasets marked with † are real-world noisy datasets. All datasets are widely used in prior works [70, 19, 8, 43].

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
YEAST	BIOLOGY	2,417	103	14
MIRFLICKR†	IMAGE	10,433	100	7
MUSIC EMOTION†	AUDIO	6,833	98	11
YEASTBP†	BIOLOGY	6,139	6,139	217
MUSIC STYLE†	AUDIO	6,839	98	10

B.3 Simulation Data Generation

The predictor matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n-1}, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n \times p}$ is constructed with rows independently drawn from a multivariate normal distribution with mean zero and a Toeplitz covariance matrix $\boldsymbol{\Sigma} = [\tau^{|i-j|}]$. $\boldsymbol{B}^*$ is randomly generated as $\boldsymbol{B}^* = \boldsymbol{B}_1 \boldsymbol{B}_2^\top$, where $\boldsymbol{B}_1 \in \mathbb{R}^{p \times r^*}$ and $\boldsymbol{B}_2 \in \mathbb{R}^{m \times r^*}$ have entries independently drawn from a standard half-normal distribution. Outliers are then added exclusively to the training set by replacing the first $\lfloor O\% \times n_{\text{train}} \rfloor$ rows of $\boldsymbol{X}$ with $[3, \ldots, 3]$, and setting the corresponding rows of $\boldsymbol{C}^* = [\boldsymbol{c}_1^*, \ldots, \boldsymbol{c}_n^*]^\top = [\tilde{\boldsymbol{c}}_1^*, \ldots, \tilde{\boldsymbol{c}}_m^*]$ to be $[-180, \ldots, -180]$ (the remaining rows of $\boldsymbol{C}^*$ are set to zero), where $\lfloor a \rfloor$ denotes the largest integer not exceeding the given scalar a. The response $\boldsymbol{y}_j$ is a Bernoulli random vector with mean $1/(1 + \exp(-\boldsymbol{X}\boldsymbol{b}_j^* - \tilde{\boldsymbol{c}}_j^*))$ (with the division computed componentwise). Notably, the test data is generated without $\boldsymbol{C}^*$ to evaluate the performance on the clean dataset. For the experiments in Section 6, we set $n_{\text{train}} = 200$, $n_{\text{test}} = 10,000$, $p = m = 10$, $\tau = 0.5$, and $r^* = 3$.

B.4 Additional Real Data Experiments

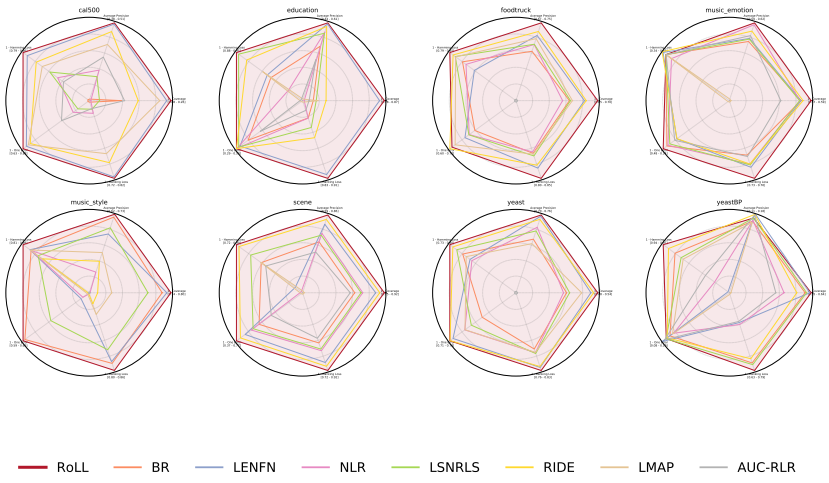


Figure B.1: Radar charts for the remaining benchmark datasets, complementing the results in Section 7.1. Consistent with the findings in Section 7.1, RoLL (red region) encloses the largest area, indicating better overall results across datasets.

B.5 Detailed Simulation Results

Table B.2: Performance comparison (mean ± standard error) across different outlier proportions ($O\%$). Metrics with ↓ indicate lower is better, and ↑ indicates higher is better. **Bold** values denote the best performance under each contamination ratio. Estimation errors ($\|\hat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F^2/(pm)$) are not reported for baselines that do not model the coefficient matrix $\boldsymbol{B}$.

$O\%$	Metric	RoLL (Ours)	BR	LENFN	NLR	LSNRLS	RIDE	LMAP	AUC-RLR
5%	Hamming Loss (↓)	**0.0750±.0005**	0.2819±.0075	0.4650±.0028	0.4735±.0017	0.2877±.0038	0.1741±.0056	0.2766±.0071	0.1015±.0021
	Example Acc (↑)	**0.4785±.0011**	0.3762±.0047	0.2993±.0038	0.2926±.0014	0.3399±.0110	0.4147±.0036	0.3444±.0072	0.4348±.0028
	Example F1 (↑)	**0.5002±.0012**	0.4201±.0048	0.3915±.0033	0.3836±.0015	0.3814±.0115	0.4399±.0036	0.3666±.0076	0.4492±.0031
	Subset Acc (↑)	**0.6701±.0012**	0.3611±.0074	0.0018±.0002	0.0021±.0002	0.3923±.0080	0.5458±.0104	0.5000±.0067	0.6632±.0015
	Label Acc (↑)	**0.8617±.0008**	0.5620±.0096	0.3565±.0044	0.3584±.0017	0.5169±.0098	0.7068±.0080	0.5525±.0102	0.8109±.0040
	Label F1 (↑)	**0.9251±.0005**	0.7182±.0077	0.5169±.0052	0.5271±.0018	0.6796±.0085	0.8274±.0055	0.7101±.0084	0.8949±.0024
	Micro-F1 (↑)	**0.9251±.0005**	0.7182±.0077	0.5369±.0043	0.5271±.0018	0.6815±.0084	0.8276±.0055	0.7110±.0084	0.8950±.0024
	Est. Error (↓)	**0.2987±.0161**	3.2709±.0114	—	—	—	—	3.2463±.0146	2.1823±.0101
6%	Hamming Loss (↓)	**0.0750±.0004**	0.3254±.0068	0.4665±.0025	0.4749±.0016	0.3233±.0031	0.1808±.0059	0.3182±.0062	0.1056±.0026
	Example Acc (↑)	**0.4785±.0011**	0.3534±.0042	0.2988±.0036	0.2917±.0013	0.3147±.0117	0.4101±.0038	0.3145±.0068	0.4310±.0031
	Example F1 (↑)	**0.5003±.0012**	0.3990±.0046	0.3911±.0031	0.3834±.0013	0.3552±.0126	0.4357±.0038	0.3366±.0073	0.4453±.0034
	Subset Acc (↑)	**0.6702±.0011**	0.3122±.0062	0.0017±.0001	0.0020±.0002	0.3548±.0083	0.5387±.0094	0.4601±.0054	0.6612±.0017
	Label Acc (↑)	**0.8617±.0008**	0.5101±.0081	0.3552±.0041	0.3570±.0015	0.4667±.0107	0.6962±.0085	0.4945±.0084	0.8034±.0047
	Label F1 (↑)	**0.9251±.0005**	0.6746±.0070	0.5157±.0050	0.5256±.0017	0.6336±.0102	0.8199±.0059	0.6602±.0075	0.8903±.0029
	Micro-F1 (↑)	**0.9251±.0005**	0.6746±.0070	0.5356±.0040	0.5256±.0017	0.6364±.0100	0.8202±.0059	0.6614±.0075	0.8904±.0029
	Est. Error (↓)	**0.3057±.0158**	3.3519±.0105	—	—	—	—	3.3272±.0140	2.2150±.0117
7%	Hamming Loss (↓)	**0.0751±.0004**	0.3580±.0059	0.4678±.0022	0.4740±.0016	0.3500±.0035	0.1850±.0066	0.3502±.0052	0.1139±.0048
	Example Acc (↑)	**0.4783±.0011**	0.3365±.0038	0.2968±.0039	0.2920±.0013	0.2965±.0114	0.4103±.0041	0.2902±.0065	0.4268±.0032
	Example F1 (↑)	**0.5001±.0012**	0.3828±.0044	0.3891±.0034	0.3832±.0013	0.3371±.0120	0.4360±.0042	0.3120±.0069	0.4427±.0036
	Subset Acc (↑)	**0.6701±.0011**	0.2806±.0058	0.0019±.0002	0.0022±.0002	0.3328±.0084	0.5363±.0098	0.4294±.0044	0.6456±.0091
	Label Acc (↑)	**0.8616±.0008**	0.4736±.0068	0.3526±.0042	0.3577±.0015	0.4316±.0107	0.6916±.0095	0.4501±.0069	0.7895±.0076
	Label F1 (↑)	**0.9250±.0005**	0.6420±.0061	0.5129±.0052	0.5264±.0016	0.6002±.0107	0.8166±.0066	0.6194±.0066	0.8813±.0049
	Micro-F1 (↑)	**0.9250±.0005**	0.6420±.0273	0.5330±.0041	0.5264±.0016	0.6027±.0104	0.8168±.0066	0.6207±.0066	0.8814±.0049
	Est. Error (↓)	**0.3017±.0156**	3.4060±.0102	—	—	—	—	3.3830±.0140	2.2615±.0216
8%	Hamming Loss (↓)	**0.0752±.0005**	0.3820±.0052	0.4689±.0021	0.4736±.0013	0.3729±.0040	0.2050±.0057	0.3736±.0047	0.1575±.0200
	Example Acc (↑)	**0.4783±.0011**	0.3239±.0034	0.2979±.0038	0.2925±.0011	0.2786±.0128	0.4065±.0075	0.2738±.0065	0.4045±.0097
	Example F1 (↑)	**0.5000±.0012**	0.3706±.0041	0.3901±.0034	0.3839±.0012	0.3184±.0137	0.4372±.0079	0.2955±.0068	0.4240±.0098
	Subset Acc (↑)	**0.6700±.0011**	0.2584±.0061	0.0019±.0002	0.0022±.0002	0.3133±.0098	0.4943±.0113	0.4055±.0041	0.5853±.0223
	Label Acc (↑)	**0.8614±.0008**	0.4478±.0057	0.3527±.0040	0.3582±.0013	0.3998±.0126	0.6626±.0093	0.4195±.0065	0.7278±.0257
	Label F1 (↑)	**0.9250±.0005**	0.6180±.0054	0.5130±.0049	0.5269±.0014	0.5665±.0138	0.7956±.0069	0.5895±.0065	0.8355±.0201
	Micro-F1 (↑)	**0.9249±.0005**	0.6180±.0054	0.5336±.0039	0.5269±.0014	0.5709±.0127	0.7972±.0067	0.5911±.0064	0.8356±.0201
	Est. Error (↓)	**0.3053±.0157**	3.4428±.0102	—	—	—	—	3.4223±.0142	2.4683±.0827
9%	Hamming Loss (↓)	**0.0752±.0004**	0.4019±.0046	0.4697±.0020	0.4742±.0011	0.3845±.0060	0.2134±.0065	0.3930±.0043	0.2371±.0272
	Example Acc (↑)	**0.4780±.0011**	0.3136±.0030	0.2965±.0041	0.2920±.0011	0.2660±.0128	0.4007±.0077	0.2585±.0064	0.3591±.0134
	Example F1 (↑)	**0.4996±.0012**	0.3604±.0036	0.3887±.0036	0.3838±.0012	0.3049±.0137	0.4320±.0079	0.2798±.0066	0.3830±.0140
	Subset Acc (↑)	**0.6700±.0011**	0.2411±.0062	0.0019±.0002	0.0020±.0002	0.2918±.0091	0.4879±.0121	0.3869±.0042	0.4927±.0263
	Label Acc (↑)	**0.8614±.0008**	0.4271±.0049	0.3510±.0042	0.3576±.0012	0.3814±.0149	0.6498±.0105	0.3932±.0063	0.6163±.0344
	Label F1 (↑)	**0.9250±.0005**	0.5980±.0048	0.5111±.0051	0.5264±.0013	0.5470±.0159	0.7860±.0079	0.5629±.0066	0.7491±.0280
	Micro-F1 (↑)	**0.9249±.0005**	0.5980±.0048	0.5317±.0042	0.5264±.0013	0.5499±.0153	0.7878±.0075	0.5646±.0064	0.7494±.0280
	Est. Error (↓)	**0.3031±.0157**	3.4735±.0102	—	—	—	—	3.4567±.0146	2.7978±.1091
10%	Hamming Loss (↓)	**0.0753±.0004**	0.4183±.0038	0.4715±.0021	0.4745±.0011	0.3642±.0110	0.2175±.0063	0.4084±.0038	0.3398±.0258
	Example Acc (↑)	**0.4778±.0011**	0.3050±.0025	0.2964±.0040	0.2915±.0010	0.2542±.0125	0.3980±.0078	0.2470±.0063	0.2981±.0138
	Example F1 (↑)	**0.4994±.0012**	0.3517±.0030	0.3887±.0035	0.3830±.0011	0.2929±.0133	0.4278±.0083	0.2681±.0065	0.3204±.0145
	Subset Acc (↑)	**0.6700±.0011**	0.2294±.0063	0.0019±.0002	0.0019±.0002	0.3005±.0136	0.4849±.0113	0.3727±.0043	0.4205±.0199
	Label Acc (↑)	**0.8613±.0008**	0.4106±.0040	0.3498±.0042	0.3571±.0012	0.3855±.0161	0.6447±.0096	0.3731±.0063	0.4780±.0328
	Label F1 (↑)	**0.9249±.0005**	0.5817±.0040	0.5095±.0051	0.5259±.0013	0.5502±.0172	0.7826±.0071	0.5418±.0068	0.6334±.0274
	Micro-F1 (↑)	**0.9249±.0005**	0.5817±.0181	0.5309±.0041	0.5258±.0013	0.5534±.0170	0.7836±.0071	0.5436±.0066	0.6337±.0274
	Est. Error (↓)	**0.3120±.0154**	3.4978±.0099	—	—	—	—	3.4852±.0148	3.1876±.1027

B.6 Performance Comparison under Unstructured Outlier Setting

To evaluate robustness under a more challenging and less structured contamination regime, an additional experiment with *unstructured outliers* is conducted. We adopt a data generation procedure similar to the one detailed in Appendix B.3, with the key distinction lying in the label generation. Here, labels are generated directly from the logistic model with linear predictor $\boldsymbol{X}\boldsymbol{B}^*$ (i.e., the response $\boldsymbol{y}_j$ is a Bernoulli random vector with mean $1/(1+\exp(-\boldsymbol{X}\boldsymbol{b}_j^*))$). To introduce unstructured contamination, we randomly select 10% of the observations and independently flip each of their labels with probability 50%. As reported in Table B.3, RoLL demonstrates competitive performance in this harder setting.

Table B.3: Comprehensive performance comparison across different metrics (mean ± standard error) under the unstructured outlier setting. Metrics with ↓ indicate lower is better, while ↑ indicates higher is better. **Bold** values denote the best performance.

METRIC	ROLL	BR	LENFN	NLR	LSNRLS	RIDE	LMAP	AUC-RLR
HAMMING LOSS (↓)	**0.0748±.0004**	0.2636±.0073	0.4702±.0051	0.4690±.0030	0.2986±.0046	0.2200±.0079	0.2490±.0087	0.1032±.0026
EXAMPLE ACC (↑)	**0.4780±.0009**	0.3853±.0044	0.2961±.0051	0.2931±.0018	0.3109±.0076	0.3926±.0089	0.3888±.0067	0.4453±.0032
EXAMPLE F1 (↑)	**0.4998±.0010**	0.4300±.0045	0.3885±.0047	0.3834±.0017	0.3542±.0078	0.4280±.0088	0.4136±.0071	0.4688±.0034
SUBSET ACC (↑)	**0.6700±.0012**	0.3679±.0075	0.0018±.0001	0.0020±.0003	0.3520±.0059	0.4482±.0113	0.5215±.0075	0.6399±.0045
LABEL ACC (↑)	**0.8619±.0006**	0.5860±.0095	0.3577±.0064	0.3632±.0028	0.4964±.0080	0.6351±.0128	0.6086±.0110	0.8076±.0046
LABEL F1 (↑)	**0.9253±.0004**	0.7362±.0074	0.5104±.0068	0.5306±.0030	0.6615±.0072	0.7742±.0097	0.7552±.0084	0.8929±.0028
MICRO-F1 (↑)	**0.9253±.0004**	0.7362±.0074	0.5313±.0068	0.5306±.0030	0.6629±.0071	0.7760±.0094	0.7558±.0084	0.8930±.0028
EST. ERROR (↓)	**0.3394±.0645**	3.2175±.0140	—	—	—	—	3.1512±.0187	2.1856±.0102

B.7 Efficacy of PIC

To evaluate the efficacy of our proposed PIC, we compare it against an ideal *Oracle tuning* benchmark that selects hyperparameters by directly minimizing the *Oracle value* $\|\boldsymbol{X}\hat{\boldsymbol{B}} + \hat{\boldsymbol{C}} - \boldsymbol{X}\boldsymbol{B}^* - \boldsymbol{C}^*\|_F^2$. Figure B.2 displays the heatmaps over the grid of regularization parameters $(\lambda_{\boldsymbol{B}}, \lambda_{\boldsymbol{C}})$ under 5% data contamination as described in Appendix B.3. We observe that the best model selected by PIC coincides with the Oracle tuning benchmark. This empirical evidence confirms that PIC effectively selects the optimal model without knowing the ground truth.

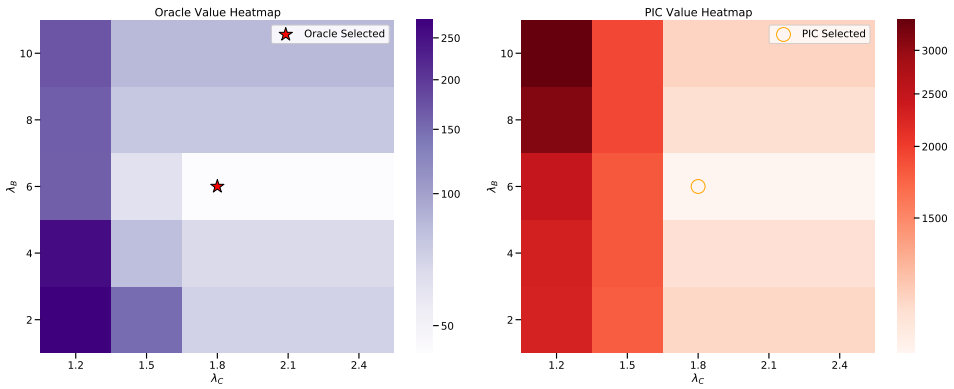


Figure B.2: Comparison of hyperparameter selection landscapes under a 5% outlier contamination rate. The heatmaps show the Oracle values (left) and PIC values (right) with varying $\lambda_{\boldsymbol{B}}$ and $\lambda_{\boldsymbol{C}}$. The optimal hyperparameters selected by PIC ($\circ$) and the Oracle ($\star$) coincide.

To further demonstrate the effectiveness of our proposed PIC, we compare it against a robust variant of 5-fold cross-validation (CV). In the presence of data contamination, standard CV often fails because outliers in the validation folds severely distort the evaluation metrics. To construct a competitive baseline, we implement a *trimmed* CV approach specifically tailored for the 10% contamination setting. During the validation phase of each fold, we compute the losses for all validation observations and discard the 10% of observations with the largest losses. The hyperparameter configuration is then selected based on the lowest average Hamming loss evaluated exclusively on the remaining 90% of the validation data.

As shown in Table B.4, despite this *ad-hoc* robustness correction in the CV baseline, our proposed PIC consistently and significantly outperforms the trimmed CV across all evaluation metrics. Notably, PIC achieves better performance, highlighting its superior advantage for model selection under data contamination.

Table B.4: Performance comparison between the proposed PIC and a trimmed 5-fold cross-validation (CV) baseline across different metrics (mean ± standard error) under 10% data contamination. Metrics with ↓ indicate lower is better, while ↑ indicates higher is better.

Metric	PIC	CV
Hamming Loss (↓)	**0.0753±.0004**	0.1281±.0281
Example Acc (↑)	**0.4778±.0011**	0.4501±.0150
Example F1 (↑)	**0.4994±.0012**	0.4730±.0143
Subset Acc (↑)	**0.6700±.0011**	0.6184±.0277
Label Acc (↑)	**0.8613±.0008**	0.7921±.0368
Label F1 (↑)	**0.9249±.0005**	0.8719±.0282
Micro-F1 (↑)	**0.9249±.0005**	0.8718±.0282
Est. Error (↓)	**0.3120±.0154**	0.8026±.2561

B.8 More Ablation Study Results

The following tables provide additional ablation study results on the foodtruck and yeast datasets.

Table B.5: Ablation study results on the foodtruck dataset (mean ± standard error). Metrics with ↓ indicate lower is better, and ↑ indicates higher is better. **Bold** values denote the best performance.

Metric	RoLL	RoLL-R	RoLL-C
Hamming Loss ↓	**0.1560±0.0030**	0.1632±0.0028	0.1566±0.0031
Avg. Precision ↑	**0.7455±0.0057**	0.7250±0.0053	0.7356±0.0065
One Error ↓	0.2866±0.0108	0.3128±0.0104	**0.2829±0.0106**
Ranking Loss ↓	**0.1473±0.0034**	0.1614±0.0036	0.1600±0.0043
Coverage ↓	**0.2961±0.0053**	0.3103±0.0053	0.3048±0.0055

Table B.6: Ablation study results on the yeast dataset (mean ± standard error). Metrics with ↓ indicate lower is better, and ↑ indicates higher is better. **Bold** values denote the best performance.

Metric	RoLL	RoLL-R	RoLL-C
Hamming Loss ↓	**0.2025±0.0014**	0.2139±0.0015	0.2137±0.0011
Avg. Precision ↑	**0.7596±0.0023**	0.7421±0.0026	0.7407±0.0017
One Error ↓	**0.2261±0.0043**	0.2575±0.0046	0.2442±0.0038
Ranking Loss ↓	**0.1717±0.0015**	0.1891±0.0020	0.1831±0.0015
Coverage ↓	0.4582±0.0022	0.4816±0.0027	**0.4565±0.0023**

B.9 Multi-class Classification with Multinomial Deviance

To further demonstrate the applicability of RoLL beyond multi-label logistic loss, we conduct an additional experiment on single-label multi-class classification using the multinomial deviance. We consider $K = 10$ classes with $n_{\text{train}} = 500$, $n_{\text{test}} = 10{,}000$, and $p = 10$. The covariate vectors are independently generated from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = [0.5^{|i-j|}]$. The true coefficient matrix is generated using a rank-3 Gaussian factorization, followed by centering and rescaling. Taking class K as the reference, we obtain $\boldsymbol{B}^\star \in \mathbb{R}^{10\times 9}$ with $\operatorname{rank}(\boldsymbol{B}^\star) \le 3$. Labels are generated from the corresponding multinomial model.

We contaminate 10% of the training observations by replacing all coordinates of their covariate vectors by 3 and reassigning each label to the least likely class other than the original class under the data-generating model. The test data remains uncontaminated. We compare RoLL with multinomial deviance (MD) [69], generalized cross entropy (GCE) [75], and normalized cross entropy combined with reverse cross entropy (NCE+RCE) [32]. Results are reported as means and standard errors over 20 independent replications.

Table B.7: Performance comparison under multi-class classification. Results are reported as mean ± standard error over 20 independent replications. Lower values are better for all metrics. Bold values denote the best performance.

Method	Estimation Error	Test Logit Error	Predictive-distribution Error
MD	0.23 ± 0.01	30.49 ± 1.88	0.35 ± 0.01
GCE	0.22 ± 0.01	29.71 ± 1.90	0.35 ± 0.01
NCE+RCE	0.36 ± 0.01	42.71 ± 2.42	0.49 ± 0.01
RoLL	$\mathbf{0.19 \pm 0.01}$	$\mathbf{26.74 \pm 1.88}$	$\mathbf{0.34 \pm 0.01}$

Performance is evaluated using three metrics: coefficient estimation error $\|\hat{\boldsymbol{B}} - \boldsymbol{B}^\star\|_F^2/(pK)$, clean-test logit prediction error $\|\boldsymbol{X}_{\text{test}}(\hat{\boldsymbol{B}} - \boldsymbol{B}^\star)\|_F^2/n_{\text{test}}$, and predictive-distribution error $n_{\text{test}}^{-1}\sum_{i=1}^{n_{\text{test}}} \mathbf{D}_{\text{KL}}(\boldsymbol{\pi}_i^\star \| \hat{\boldsymbol{\pi}}_i)$, where $\boldsymbol{\pi}_i^\star$ and $\hat{\boldsymbol{\pi}}_i$ denote the oracle and estimated probability vectors over all K classes, respectively, and $\mathbf{D}_{\text{KL}}(\boldsymbol{\pi}\|\hat{\boldsymbol{\pi}}) = \sum_{k=1}^K \pi_k \log(\pi_k/\hat{\pi}_k)$. Lower values indicate better performance for all three metrics.

As shown in Table B.7, RoLL achieves the lowest estimation error, test logit prediction error, and predictive-distribution error, demonstrating its superior performance under data contamination.

C Supplementary Experiments on Regression Tasks

C.1 Data Generation

The predictor matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n-1}, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times p}$ is constructed with rows independently drawn from a multivariate normal distribution with mean zero and a Toeplitz covariance matrix $\boldsymbol{\Sigma} = [\tau^{|i-j|}]$. $\boldsymbol{B}^*$ is randomly generated as $\boldsymbol{B}^* = \boldsymbol{B}_1\boldsymbol{B}_2^\top$, where the entries of $\boldsymbol{B}_1 \in \mathbb{R}^{p\times r^*}$ and $\boldsymbol{B}_2 \in \mathbb{R}^{m\times r^*}$ are independently and identically distributed samples from a standard half-normal distribution. Outliers are then added exclusively to the training set by replacing the first $J^* = \lfloor 5\% \times n_{\text{train}} \rfloor$ rows of $\boldsymbol{X}$ with $[3, \ldots, 3]$, and setting the corresponding rows of $\boldsymbol{C}^* = [\boldsymbol{c}_1^*, \ldots, \boldsymbol{c}_n^*]^\top = [\tilde{\boldsymbol{c}}_1^*, \ldots, \tilde{\boldsymbol{c}}_m^*]$ to be $[10, \ldots, 10]$ (the remaining rows of $\boldsymbol{C}^*$ are set to zero), where $\lfloor a \rfloor$ denotes the largest integer not exceeding the given scalar a. The response $\boldsymbol{y}_j$ is generated as $\boldsymbol{y}_j = \boldsymbol{X}\boldsymbol{b}_j^* + \tilde{\boldsymbol{c}}_j^* + \boldsymbol{\epsilon}_j$, where the noise vectors $\boldsymbol{\epsilon}_j$ are independently drawn from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Notably, the test data is generated without $\boldsymbol{C}^*$ to evaluate the performance on the clean dataset. For the experiments in this section, we set $n_{\text{train}} = 200$, $n_{\text{test}} = 10,000$, $p = m = 10$, $\tau = 0.5$, and $r^* = 3$.

C.2 Performance Comparison

We compare RoLL with three recent robust low-rank regression approaches: RRHR [55], D4R [31], and R4-NCP [33]. We evaluate each method using *Prediction Error* ($\|\boldsymbol{Y}_{\text{test}} - \boldsymbol{X}_{\text{test}}\hat{\boldsymbol{B}}\|_F^2/(mn_{\text{test}})$), *In-sample MSE* ($\|\boldsymbol{X}_{\text{train}}(\hat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_F^2/(mn_{\text{train}})$), and *Estimation Error* defined as $\|\hat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F^2/(pm)$. Table C.1 presents the performance of all methods under a 5% outlier contamination setting, with results averaged over 20 independent replications. Overall, RoLL consistently achieves the best performance across all metrics.

Table C.1: Performance comparison on regression tasks under Gaussian noise (mean ± standard error). **Bold** values denote the best performance.

Metric	RoLL	RRHR	D4R	R4-NCP
Pre. Error	**1.2598±.0100**	2.3152±.0268	1.4705±.0168	2.3246±.0277
In-sample MSE	**0.2478±.0068**	3.3371±.0348	0.5269±.0168	3.3772±.0355
Est. Error	**0.0484±.0017**	0.0653±.0019	0.0784±.0027	0.0644±.0019

While our theoretical guarantees in Section 4 rely on the sub-Gaussian noise assumption, it is of practical interest to investigate the empirical robustness of our method when this assumption is

violated. To this end, we conduct an additional experiment evaluating the regression performance under *heavy-tailed* noise. Specifically, we modify the data generation process by sampling the entries of the noise vector $\boldsymbol{\epsilon}_j$ independently from a Student t-distribution ($df = 4$). As presented in Table C.2, our method consistently outperforms the baselines, demonstrating its strong empirical applicability even when the sub-Gaussian assumption is broken.

Table C.2: Performance comparison on regression tasks under Student t distribution noise ($df = 4$) (mean $\pm$ standard error). **Bold** values denote the best performance.

Metric	RoLL	RRHR	D4R	R4-NCP
Pre. Error	**2.2547±.0104**	3.3379±.0338	2.5141±.0210	3.3639±.0333
In-sample MSE	**0.2550±.0073**	3.3492±.0553	0.5970±.0224	3.4308±.0535
Est. Error	**0.0465±.0017**	0.0717±.0020	0.0838±.0033	0.0716±.0019

C.3 Minimax Tightness

To empirically validate our theoretical guarantees and investigate whether the minimax rates are tight in practice, we conducted targeted ablation studies isolating the key parameters of our theoretical analysis. Specifically, we evaluate the empirical error (i.e., $\|\boldsymbol{X}\hat{\boldsymbol{B}} + \hat{\boldsymbol{C}} - \boldsymbol{X}\boldsymbol{B}^* - \boldsymbol{C}^*\|_F^2$) under the regression setting by independently varying the true rank r^*, the outlier fraction J^*/n, and the ambient dimensions p and m (since $\boldsymbol{X}$ is full rank in our setup, the effective dimension q equals p). For each ablation, we vary a single parameter while holding the others constant at their default values as in Section C.1, with results averaged over 20 independent trials.

As illustrated in Figure C.1, the empirical error tightly tracks the theoretical minimax lower bound up to a universal constant across all examined parameters (to clearly visualize the identical scaling behavior, we plot the theoretical curve as $c \cdot (J^*m + J^* \log(en/J^*) + (p+m)r^*)$ with an empirical scaling constant $c = 0.76$). The fact that the empirical curves align perfectly with these theoretical predictions confirms that the minimax rates established in our analysis are indeed tight in practice.

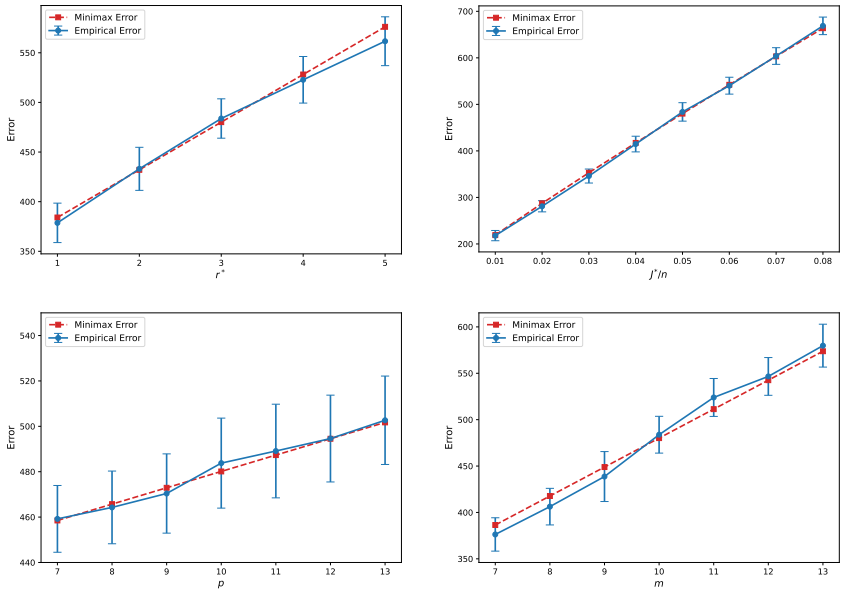


Figure C.1: Empirical error versus theoretical minimax error across the true rank r^*, the outlier fraction J^*/n, and the ambient dimensions p and m. Error bars represent the standard deviation over 20 independent trials.

Table C.3: Computational comparison in the large-scale regression setting with $n = 10{,}000$ and $p = m = 120$. Results are reported as mean $\pm$ standard deviation over 20 independent replications.

Method	Iterations	Time (s)
RAPID	$\mathbf{1666 \pm 11}$	$\mathbf{156.5 \pm 3.5}$
FISTA	12031 ± 618	848.0 ± 46.5
Gradient restart	2061 ± 5	182.9 ± 6.9
ZeroFPR-L-Broyden	5422 ± 1320	1538.3 ± 560.9

C.4 Power of Acceleration

In this part, we evaluate the computational efficiency of RAPID against the MM baseline under the regression setup. To achieve this, we compare the number of iterations each algorithm requires to reach a specific target objective value across varying condition numbers. Following the setup in Section C.1, except that $\boldsymbol{X}$ remains uncontaminated, we vary the Toeplitz parameter $\tau \in \{0, 0.25, 0.5, 0.75\}$, which yields condition numbers $\kappa \in \{292, 356, 544, 924\}$. The tuning parameters $\lambda_{\boldsymbol{B}}$ and $\lambda_{\boldsymbol{C}}$ are selected using PIC.

Figure C.2 illustrates the function value versus iteration count for both algorithms. The results demonstrate that the acceleration effect of RAPID becomes increasingly pronounced as the condition number grows. Specifically, the speedup ratio of RAPID over the MM baseline increases from approximately $9.7\times$ (134 vs. 1300 iterations at $\kappa = 292$) to $16.6\times$ (242 vs. 4008 iterations at $\kappa = 924$).

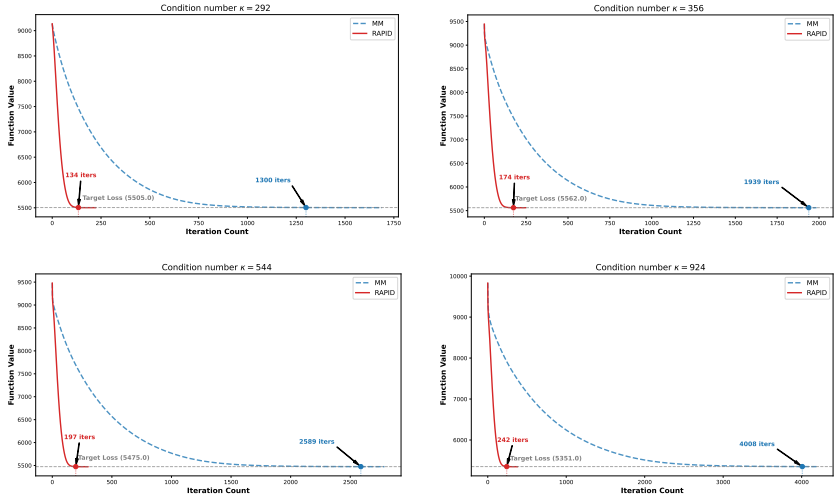


Figure C.2: Function value versus iteration count for both the MM baseline and the proposed RAPID algorithm across different condition numbers κ. The four subplots illustrate the number of iterations required to reach a target objective value for different condition numbers κ.

Large-scale solver comparison. To further evaluate the practical scalability of RAPID, we conduct a large-scale regression experiment with $n = 10{,}000$ and $p = m = 120$, corresponding to $1{,}214{,}400$ optimization variables in $(\boldsymbol{B}, \boldsymbol{C})$. We compare RAPID with FISTA [2], gradient restart [41], and ZeroFPR-L-Broyden [56]. All methods use the same initialization and are terminated upon reaching the same prescribed objective-function value. The results are averaged over 20 independent replications. As shown in Table C.3, RAPID reaches the prescribed objective value with the fewest iterations and the shortest runtime among the compared solvers.

Sensitivity to the switching threshold. We examine the sensitivity of RAPID to the switching threshold T_0, which controls when the adaptive line search is activated. Specifically, RAPID uses the conservative choice $\mu_t = 0$ and $\rho_t = \mathcal{L}$ for the first T_0 iterations and activates the adaptive line search thereafter. Using the same large-scale regression setting with $n = 10{,}000$ and $p = m = 120$, we vary

Table C.4: Sensitivity of RAPID to the switching threshold T_0 in the large-scale regression setting. Results are reported as mean $\pm$ standard deviation over 20 independent replications.

T_0	Iterations	Time (s)	T_0	Iterations	Time (s)
0	1930 ± 11	175.7 ± 1.7	60	1672 ± 13	178.0 ± 23.1
10	1939 ± 10	177.9 ± 3.2	70	1680 ± 13	176.7 ± 21.3
20	1947 ± 11	209.6 ± 18.4	80	1687 ± 14	169.8 ± 14.0
30	1956 ± 11	210.6 ± 16.2	90	1963 ± 18	194.9 ± 16.4
40	$\mathbf{1658 \pm 11}$	176.9 ± 21.4	100	1965 ± 20	176.7 ± 2.7
50	1666 ± 11	$\mathbf{156.5 \pm 3.5}$	∞ (FISTA)	12031 ± 618	848.0 ± 46.5

Table C.5: Performance across different contamination structures. Each entry reports estimation error ($\|\hat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F^2$) / test MSE ($\|\boldsymbol{Y}_{\text{test}} - \boldsymbol{X}_{\text{test}}\hat{\boldsymbol{B}}\|_F^2/(mn_{\text{test}})$) (mean $\pm$ standard error over 20 independent replications). Bold values denote the best performance.

k	RoLL	D4R [31]	RRR [24]
10	$\mathbf{0.930 \pm 0.240/1.065 \pm 0.017}$	$2.186 \pm 0.163/1.167 \pm 0.022$	$3.576 \pm 0.259/1.248 \pm 0.020$
5	$\mathbf{1.266 \pm 0.284/1.091 \pm 0.021}$	$2.096 \pm 0.140/1.157 \pm 0.019$	$3.052 \pm 0.221/1.221 \pm 0.016$
2	$\mathbf{1.453 \pm 0.247/1.102 \pm 0.018}$	$2.042 \pm 0.139/1.142 \pm 0.014$	$2.768 \pm 0.155/1.195 \pm 0.010$
1	$\mathbf{1.845 \pm 0.210/1.128 \pm 0.014}$	$2.052 \pm 0.129/1.142 \pm 0.014$	$2.570 \pm 0.120/1.180 \pm 0.008$

T_0 and terminate all runs upon reaching the same prescribed objective-function value. Results are averaged over 20 independent replications.

As shown in Table C.4, the number of iterations is stable for $T_0 \in [40, 80]$, differing by less than 2% from the smallest value over this range. In particular, $T_0 = 40$ yields the fewest iterations, whereas $T_0 = 50$ yields the shortest runtime. All finite values of T_0 considered here substantially outperform $T_0 = \infty$, under which the adaptive line search is never activated and RAPID reduces to FISTA. We therefore use $T_0 = 50$ as a practical default.

C.5 Robustness to Misspecified Contamination Structures

We investigate the robustness of RoLL when the contamination structure deviates from the assumed rowwise pattern. Following the regression setting in Section C.1, except that $\boldsymbol{X}$ remains uncontaminated, we set $n = 200$ and $p = m = 10$ and fix the total number of corrupted response entries at 100. For each $k \in \{10, 5, 2, 1\}$, the corrupted entries are distributed over $100/k$ observations, with k corrupted response coordinates per affected observation. Thus, $k = 10$ corresponds to fully rowwise contamination, whereas smaller values of k progressively disperse the contamination across more observations, with $k = 1$ representing an entrywise contamination pattern.

As shown in Table C.5, RoLL achieves the lowest estimation error and test MSE across all considered contamination structures. Although its estimation error increases as the contamination becomes more dispersed, RoLL remains more accurate than the competing methods even under strongly dispersed contamination. These results provide empirical evidence that RoLL retains robustness under misspecified contamination structures.

D Further Discussions on RAPID

RAPID (Algorithm 2) is distinct from standard Nesterov momentum [38] through three key innovations: 1) Rather than using separate updating schemes for convex and strongly convex losses [38], (21) provides a universal all-in-one form. 2) Instead of using a fixed extrapolation coefficient, RAPID dynamically exploits local geometry: (21a) incorporates ρ_{t-1} to adapt to local smoothness, and (21b) introduces μ_t to capture the degree of local strong convexity. 3) Local restricted strong convexity constants are notoriously difficult to determine in theory. RAPID bypasses this by initializing $\mu_0 = 0$ and utilizing a novel line search criterion to adaptively find suitable curvature values on the fly.

Importantly, the practical guarantee of RAPID relies on Theorem 5.1, which imposes no regularity conditions or restrictions on ρ_t and μ_t. This allows the algorithm to dynamically adapt to varying

local curvature, unifying locally non-strongly convex and strongly convex regimes. This is appealing in practice, since the local curvature may vary over iterations and is typically unknown in advance. Furthermore, the result in Theorem 5.1 is also established at the level of a general regularizer, so its scope is not tied to our specific problem. Notably, (21) provides a generalized framework that serves as a unified acceleration scheme for gradient descent type algorithms, such as standard gradient descent, proximal gradient descent, and projected gradient descent.

Finally, we discuss the computational complexity of RAPID. At each iteration, computing $\boldsymbol{X}\boldsymbol{B}$ and $\boldsymbol{X}^\top \nabla l$ requires $\mathcal{O}(npm)$ operations. The spectral proximal update of $\boldsymbol{B}$ requires a thin SVD, with complexity $\mathcal{O}\{pm\min(p,m)\}$, while the rowwise group proximal update of $\boldsymbol{C}$ requires $\mathcal{O}(nm)$ operations. The momentum and other elementwise operations in RAPID require $\mathcal{O}(pm+nm)$ operations. Moreover, the number of candidate curvature values examined by the line search is capped by a fixed constant M, and therefore does not change the asymptotic order. Consequently, RAPID has the same asymptotic per-iteration complexity as the MM algorithm, $\mathcal{O}\left[pm\{n+\min(p,m)\}\right]$, although the line search may introduce a constant-factor overhead. If T iterations are required to attain a prescribed optimization accuracy, the total computational complexity is $\mathcal{O}\left[T\{npm+pm\min(p,m)\}\right]$. Thus, reducing the number of iterations can improve the overall computational efficiency of RAPID.

E Discussions on Outlier Support and Identifiability

Exact outlier-support recovery. Suppose that the conditions of Theorem 4.3 hold. Choose $\nu > 0$ and $a \geq 2b > 0$ such that $a > 1/\delta$ and $b > 1/(2\vartheta)$. Define $\mathfrak{C}_\nu = K_1^2\lambda_{\boldsymbol{B}}^2 r^* + K_2^2\lambda_{\boldsymbol{C}}^2 J^* + a\sigma^2\nu/2$ and $s_\nu = \lceil 2\mathfrak{C}_\nu/[(\vartheta - 1/(2b))\lambda_{\boldsymbol{C}}^2]\rceil$. Suppose that $s_\nu \leq n-p$ and that there exist constants $\kappa_{\mathrm{ssc}} > 0$ and $\kappa_\Sigma > 0$, together with a positive definite matrix $\boldsymbol{\Sigma}$ satisfying $\lambda_{\min}(\boldsymbol{\Sigma}) \geq \kappa_\Sigma$, such that the subset strong convexity (SSC) condition $\boldsymbol{X}_{\mathcal{S}^c\cdot}^\top \boldsymbol{X}_{\mathcal{S}^c\cdot}/n \succeq \kappa_{\mathrm{ssc}}\boldsymbol{\Sigma}$ holds for every $\mathcal{S} \subset [n]$ with $|\mathcal{S}| \leq s_\nu$.

Consider the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B}^* + \boldsymbol{C}^* + \boldsymbol{E}$ with the squared loss, where $P_{\boldsymbol{C}}^s(\boldsymbol{C};\lambda_{\boldsymbol{C}})$ is the group SCAD penalty ($\gamma > 2$). Let $(\hat{\boldsymbol{B}}, \hat{\boldsymbol{C}})$ be an F-estimator defined in (7). Define $M_X = \max_{i\in[n]} \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i\|_2$ and, when $J^* > 0$, $c_{\min} = \min_{i\in\mathcal{J}(\boldsymbol{C}^*)} \|\boldsymbol{c}_i^*\|_2$. For $u \geq 1$, let $\varepsilon_{\nu,u} = C_E\sigma\sqrt{m+\log n+u} + M_X[\frac{4\mathfrak{C}_\nu}{n\kappa_{\mathrm{ssc}}}\{\frac{1}{\delta-1/a} + \frac{1}{\vartheta-1/(2b)}\}]^{1/2}$, where $C_E > 0$ is a universal constant. If $\varepsilon_{\nu,u} < \lambda_{\boldsymbol{C}}$ and, when $J^* > 0$, $c_{\min} > \lambda_{\boldsymbol{C}} + \varepsilon_{\nu,u}$, then $\Pr\{\mathcal{J}(\hat{\boldsymbol{C}}) = \mathcal{J}(\boldsymbol{C}^*)\} \geq 1 - c'\exp(-c\nu) - 2\exp(-c_E u)$, where $c, c', c_E > 0$ are universal constants.

Identifiability implication. The SSC condition also ensures identifiability of the systematic and contamination components. Consider the row-sparse class $\mathcal{C}(J_{\max}) = \{\boldsymbol{C} : J(\boldsymbol{C}) \leq J_{\max}\}$. Suppose $\boldsymbol{X}\boldsymbol{B}_1 + \boldsymbol{C}_1 = \boldsymbol{X}\boldsymbol{B}_2 + \boldsymbol{C}_2$ for $\boldsymbol{C}_1, \boldsymbol{C}_2 \in \mathcal{C}(J_{\max})$, and let $\mathcal{S} = \mathcal{J}(\boldsymbol{C}_1 - \boldsymbol{C}_2)$. Then $|\mathcal{S}| \leq J(\boldsymbol{C}_1) + J(\boldsymbol{C}_2) \leq 2J_{\max}$ and $\boldsymbol{X}_{\mathcal{S}^c\cdot}(\boldsymbol{B}_1 - \boldsymbol{B}_2) = \boldsymbol{0}$. If the SSC condition holds for all $\mathcal{S} \subset [n]$ with $|\mathcal{S}| \leq 2J_{\max}$, then $0 = \|\boldsymbol{X}_{\mathcal{S}^c\cdot}(\boldsymbol{B}_1 - \boldsymbol{B}_2)\|_F^2 \geq n\kappa_{\mathrm{ssc}}\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{B}_1 - \boldsymbol{B}_2)\|_F^2$. Since $\lambda_{\min}(\boldsymbol{\Sigma}) > 0$, we obtain $\boldsymbol{B}_1 = \boldsymbol{B}_2$ and hence $\boldsymbol{C}_1 = \boldsymbol{C}_2$.

PIC control of outlier-support complexity. The PIC penalty explicitly controls the complexity of the selected outlier support. For a fixed rank, increasing the support size from J to $J+1$ raises the penalty by at least $A\sigma^2\{m + \log(en/(J+1)) - 1\}$. Moreover, under the conditions of Theorem 4.5, with $A > K + 4bL$ as in its proof, the proof yields $\mathbb{E}[\hat{J}\{m + \log(en/\hat{J})\}] \lesssim P_m(\boldsymbol{B}^*, \boldsymbol{C}^*) + 1$, where $\hat{J} = J(\hat{\boldsymbol{C}})$ and the expression is taken as zero when $\hat{J} = 0$. Thus, the selected outlier-support complexity is controlled in expectation by the oracle model complexity.

F Limitations and Future Work

While the present paper focuses on linear systematic components, extending the proposed RoLL framework to encompass structured nonlinear function classes is a natural direction for future work. Furthermore, although our empirical evaluation in the appendix demonstrates the framework's practical robustness against heavy-tailed noise, formally establishing statistical guarantees under such distributions remains an important theoretical direction. Finally, as our current analysis focuses primarily on bounding the prediction error, extending the theory to guarantee exact structural recovery stands as a compelling open problem for future research.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope. See the Abstract and Section 1.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitations of the work is discussed in the Appendix F.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: For all theoretical results, the paper provides the corresponding proofs in Appendix A.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The paper fully discloses all the information needed to reproduce the main experimental results in the paper. See Sections 6 and 7.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: We provide the data and code in the supplemental material to reproduce the main experimental results.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: We specify all the training and test details necessary to understand the results. See Sections 6 and 7.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: In the tables reporting the experimental results, we provide the mean and standard deviation across multiple independent runs, with all results averaged over independent trials to ensure reliability and consistency.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: All experiments are conducted on a personal computer whose details are provided in the paper. See Section 6 .

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in the paper conform with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [No]

    Justification: This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [No]

    Justification: The paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All code, models, and datasets mentioned in this work are appropriately cited with their original papers.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: New assets introduced in the paper, such as code, are well documented. The documentation is provided alongside the assets in the supplementary material.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

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