Robust Learning under Hybrid Noise

YANG WEI, Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Jiangsu Key Laboratory of Image and Video Understanding for Social Security, and School of Computer Science and Engineering, Nanjing University of Science and Technology, China
SHUO CHEN, RIKEN Center for Advanced Intelligence Project, Japan
SHANSHAN YE, Australian Artificial Intelligence Institute, University of Technology Sydney, Australia
BO HAN, Department of Computer Science, Hong Kong Baptist University, China
CHEN GONG, Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Jiangsu Key Laboratory of Image and Video Understanding for Social Security, and School of Computer Science and Engineering, Nanjing University of Science and Technology, China

Feature noise and label noise are ubiquitous in practical scenarios, which pose great challenges for training a robust machine learning model. Most previous approaches usually deal with only a single problem of either feature noise or label noise. However, in real-world applications, hybrid noise, which contains both feature noise and label noise, is very common due to the unreliable data collection and annotation processes. Although some results have been achieved by a few representation learning based attempts, this issue is still far from being addressed with promising performance and guaranteed theoretical analyses. To address the challenge, we propose a novel unified learning framework called "**F**eature and **L**abel **R**ecovery" (FLR) to combat the hybrid noise from the perspective of data recovery, where we concurrently reconstruct both the feature matrix and the label matrix of input data. Specifically, the clean feature matrix is discovered by the low-rank approximation, and the ground-truth label matrix is embedded based on the recovered features with a nuclear norm regularization. Meanwhile, the feature noise and label noise are characterized by their respective adaptive matrix norms to satisfy the corresponding maximum likelihood. As this framework leads to a non-convex optimization problem, we develop the non-convex Alternating Direction Method of Multipliers (ADMM) with the convergence guarantee to solve our learning objective. We also provide the theoretical analysis to show that the generalization error of FLR can be upper-bounded in the presence of hybrid noise. Experimental results on several typical benchmark datasets clearly demonstrate the superiority of our proposed method over the state-of-the-art robust learning approaches for various noises.

CCS Concepts: • **Computing methodologies** → **Machine learning**; **Machine learning approaches**.

Corresponding authors: Chen Gong and Shuo Chen.
This research is supported by NSF of China (Nos: 62336003, 12371510), NSF for Distinguished Young Scholar of Jiangsu Province (No: BK20220080), and China Scholar Council.
Authors' Contact Information: Yang Wei, Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Jiangsu Key Laboratory of Image and Video Understanding for Social Security, and School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, China, csywei@njust.edu.cn; Shuo Chen, RIKEN Center for Advanced Intelligence Project, Tokyo, Japan, shuo.chen.ya@riken.jp; Shanshan Ye, Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney, Australia, shanshan.ye@student.uts.edu.au; Bo Han, Department of Computer Science, Hong Kong Baptist University, Hong Kong, China, bhanml@comp.hkbu.edu.hk; Chen Gong, Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Jiangsu Key Laboratory of Image and Video Understanding for Social Security, and School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, China, chen.gong@njust.edu.cn.

ACM 1557-735X/2024/11-ART*
https://doi.org/XXXXXXX.XXXXXXX

Additional Key Words and Phrases: Hybrid noise, Matrix recovery, Generalization bound

ACM Reference Format:
Yang Wei, Shuo Chen, Shanshan Ye, Bo Han, and Chen Gong. 2024. Robust Learning under Hybrid Noise. *J. ACM* *, *, Article * (November 2024), 27 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The successes of most existing machine learning algorithms usually depend on the availability of high-quality data with clean features and accurate annotated labels. However, such high-quality data can hardly be guaranteed in lots of real-world scenarios due to various subjective and objective factors such as the insufficient expert knowledge [11] and uncontrollable measurement error [12].

Although many previous methods have designed robust classifiers to solely tackle the feature noise or label noise, they are still in the mess of performance degradation when the hybrid noise occurs in training data. Hybrid noise is usually induced by simultaneously corrupted features and labels, and it is very common in many real-world applications. For instance, if we have numerous image examples that are collected from visual sensors and are annotated manually, those images may be contaminated under non-ideal illumination or occlusion [3]; also, labeling errors are inevitably introduced due to human fatigue [27, 35]. In this case, the hybrid noise will significantly damage the model training and severely decline the classification accuracy because the input data and output prediction are both misled by the feature noise and label noise, respectively. Therefore, a new unified learning algorithm is desired to deal with the hybrid noise in practical applications.

Most existing methods for tackling label noise mainly focus on picking up clean labeled examples from the raw training data, such as MentorNet [17] and Co-teaching [15]. Yet these kinds of works cannot theoretically guarantee the label correctness of the selected clean examples, so they can hardly obtain stable performance in practical uses. Therefore, some previous works, including μ Stochastic Gradient Descent (μSGD) [30] and Symmetric Cross Entropy (SCE) [38], further consider directly designing the different robust loss functions, avoiding the above sample selection. Nevertheless, they inevitably become weak when learning with some complicated noise, because they do not explicitly characterize the generation process of the label noise. In addition, there are some other approaches which propose to correct the data distribution based on the estimated statistics, such as transition matrix [14] and dataset centroid [10]. However, all the above methods share an implicit deficiency, where they assume the features of all examples are completely clean. It means that these label-noise-specific learning algorithms cannot effectively handle the hybrid noise.

On the other hand, dealing with feature noise is also a classical topic in the robust learning area. To be specific, reconstruction based methods and regression based methods have shown the great effectiveness in some common noise cases, including the Gaussian noise and Laplacian noise [7]. For example, Robust Principal Component Analysis (RPCA) [1] and Low-Rank Representation (LRR) [23] employ ℓ_1-norm or $\ell_{2,1}$-norm to characterize the reconstruction residual which follows the Laplacian distribution. Moreover, the Frobenius norm $\|\cdot\|_F$ can be utilized to characterize the Gaussian distributed noise [2] in some popular models such as denoising Autoencoder [36]. Note that all the above reconstruction based algorithms usually need an additional classifier to perform classification tasks. In contrast, some regression based methods including Sparse Representation Classifier (SRC) [41] and Nuclear-Norm based Matrix Regression (NMR) [44] propose to directly classify examples based on the reconstruction error, but they require fully correct labels as supervision, so they are still unable to handle the hybrid noise. Note that although Robust Representation Learning (RRL) [20] can address label noise, out-of-distribution input, and input corruption simultaneously, it is still short of the explicit module to handle feature noise and

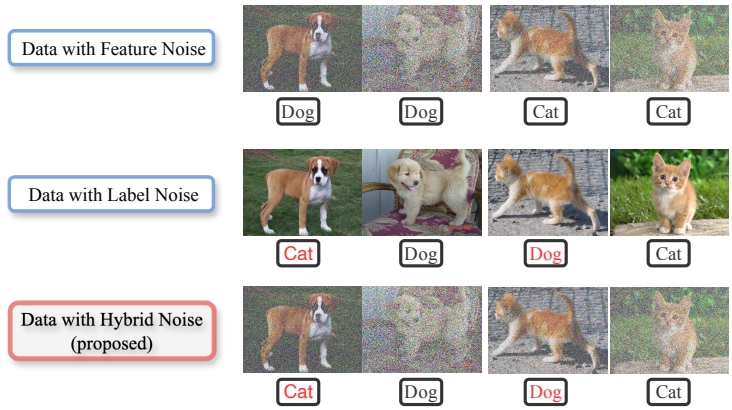


Fig. 1. The problem setting of our proposed hybrid noise learning. Data with Feature Noise: all labels are correct, yet the features of examples are corrupted. Data with Label Noise: the features of all examples are clean, while some labels are incorrect. We consider a more challenging case, namely Data with Hybrid Noise: both features and labels of training examples are noisy. The noisy label is indicated in red.

Fig. 2. The proposed unified learning paradigm for the hybrid noise removal. The clean feature matrix $\boldsymbol{X}$ and correct label matrix $\boldsymbol{Y}$ are recovered from noisy data $\tilde{\boldsymbol{X}}$, $\tilde{\boldsymbol{Y}}$, respectively. Meanwhile, the true label matrix $\boldsymbol{Y}$ is embedded by $\boldsymbol{X}$ via a low-rank projection $\boldsymbol{Z}$, where the feature error matrix is $\boldsymbol{E}_f$ and the label error matrix is $\boldsymbol{E}_l$, respectively.

the theoretical guarantee of denoising capability. In addition, RRL is limited to dealing with noisy data in visual tasks, while hybrid noise is also widespread in various non-visual real-world tasks.

To overcome the above drawbacks of existing robust learning approaches, we propose a novel unified learning paradigm called "**F**eature and **L**abel **R**ecovery" (FLR), which allows us to train a robust classifier with the heavy hybrid noise. The setting of hybrid noise is shown in Figure 1, which describes the main difference among the hybrid noise, label noise, and feature noise. Specifically, the red box shows the hybrid noise case where four corrupted animal images belong to two classes (here the first and third examples are mislabeled). To address this challenging setting, we construct a new data recovery framework to simultaneously learn both the clean feature matrix and the true label matrix in a single optimization objective. Given the observed noisy feature matrix $\tilde{\boldsymbol{X}}$ and the corresponding noisy label matrix $\tilde{\boldsymbol{Y}}$, the clean feature matrix $\boldsymbol{X}$ is a low-rank approximation to its noisy matrix $\tilde{\boldsymbol{X}}$ (as shown in Figure 2), and the feature error $\boldsymbol{E}_f$ is characterized by adaptive matrix norms under different types of noise assumption (*i.e.*, $\tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{E}_f$) to satisfy the corresponding

maximum likelihood. Furthermore, the correct label matrix $\boldsymbol{Y}$ is embedded based on the recovered feature matrix $\boldsymbol{X}$ with a low-rank projection $\boldsymbol{Z}$ (*i.e.*, $\boldsymbol{Y} = \boldsymbol{XZ}$), so that the label noise is captured by a row-sparse matrix $\boldsymbol{E}_l$.

FLR seamlessly integrates the hybrid noise removal and classifier learning (as the projection matrix $\boldsymbol{Z}$ predicts class labels for all examples) into a unified optimization model, which is solved by our designed non-convex Alternating Direction Method of Multipliers (ADMM). We prove that our iteration algorithm can converge to a stationary point of the learning objective. Meanwhile, we also provide theoretical analyses to reveal that both the clean feature matrix and the true label matrix can be correctly recovered with increasing example size. Thanks to the integrated consideration of the feature noise and label noise, as well as the effectiveness of the low-rank matrix recovery technique, our proposed method achieves better results than existing robust learning approaches. The main contributions of our paper are summarized below:

- We propose the hybrid noise problem formally (*i.e.*, the integration of both the feature noise and label noise), and we mathematically define this practical problem setting to a general data recovery objective.
- We propose a unified learning framework to deal with the hybrid noise problem, where we provide comprehensive theoretical analyses to guarantee the algorithm convergence as well as the generalization ability of our method.
- We conduct intensive experiments on popular benchmark datasets to demonstrate the superiority of our method over the state-of-the-art robust learning approaches.

2 Related Work

In this section, we briefly review representative works on label noise learning and feature noise learning.

2.1 Label Noise Learning

The main goal of label noise learning is to train a robust classifier with noisy supervision to avoid the degradation of classification performance on test data.

A straightforward idea is to improve the label reliabilities of a small part of examples. The early-stage approaches [28, 29] firstly identify correctly labeled examples, and then learn the corresponding classifier with selected clean examples. Recently, the deep neural network based method MentorNet [17] learns a data-driven curriculum to pick up the examples with potentially correct labels. Co-teaching [15] simultaneously trains two networks in a cooperative manner, where each network selects the small-loss examples for its peer network training. DivideMix [19] discards noisy labels but preserves the corresponding instances, and then trains DNNs in the manner of semi-supervised learning. RTLC [48] and CLC [46] aim to identify the noisy labels and correct them in Federated Learning. Nevertheless, it is hard to verify the label correctness of the selected examples, which makes these kinds of methods not always reliable in practical uses.

To avoid the unreliable process of sample selection, LICS [8] and μSGD [30] decompose the loss functions into a label-independent term and a label-dependent term, and then only correct the second term to reduce the negative effect caused by noisy labels. Ghosh et al. [9] further generalize the existing noise-tolerant loss functions to multi-class classification problems. Li et al. [21] propose a dynamics-aware loss to address the discrepancy between the static robust loss functions and the dynamics of DNNs in learning with noisy labels. WarPI [33] proposes to rectify the training process within the meta-learning scenario from the input of logits and labels.

Furthermore, some methods consider tackling the label noise based on the statistics estimation. Masking [14] conveys human cognition of invalid class transitions and naturally speculates the

structure of the noise transition matrix. PCSE [24] establishes the quantitative relationship between the clean (first-order and second-order) statistics and the corresponding noisy statistics for every class, inspired by the centroid estimation theory. Furthermore, the relationship is utilized to induce a generative classifier for model inference. However, existing estimators, especially those anchor points or cluster based methods, generally require informative representations. Therefore, Zhu et al. [50] build the transition matrix estimator using the distilled features, which can just handle the less informative features, but it does not take the feature noise into account.

2.2 Feature Noise Learning

The original idea for feature noise learning is to recover the latent clean examples from the corrupted data and feed the reconstructed examples to downstream tasks. Both RPCA [1] and LRR [23] aim to reconstruct corrupted examples with the low-rank constraint, where RPCA assumes that the clean data matrix is low-rank while LRR assumes the representation coefficient matrix is low-rank. Chen et al. [4] further propose the δ-norm to characterize the structural noise, so that the error matrix is structurally sparse during the training phase. Recently, several deep neural network based methods have been proposed to handle noisy examples. CFMNet [6] conducts the image noise removal with multi-layer conditional feature modulations. NIM-AdvDef [45] adopts denoising as the pre-text task, and reconstructs noisy images well despite severe corruption.

Note that all the above reconstruction based algorithms usually need an additional classifier to perform classification tasks. In contrast, some regression based methods such as SRC [41] and NMR [44] propose to directly classify examples based on the minimum reconstruction strategy. MSPM [26] proposes a novel spectral-spatial feature mining framework to handle robust feature extraction under feature noise and effective classification of HSI, which is highly noise-robust for deeply digging into complex features. However, all methods mentioned above still require fully correct labels as supervision. As a result, these approaches are unable to handle hybrid noise. RRL [20] can handle out-of-distribution input and input corruption in visual tasks, but no specific module is designed for tackling feature noise. This motivates us to propose a completely new learning framework to simultaneously recover clean data from noisy features and noisy labels (*i.e.*, the hybrid noise).

3 Our Proposed FLR

In this section, we firstly describe our new method FLR and then provide the optimization algorithm to solve it.

3.1 Model Formulation

We assume that the data matrix $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n\times d}$ is partially corrupted by noise $\boldsymbol{E}_f \in \mathbb{R}^{n\times d}$, and the latent clean training feature matrix $\boldsymbol{X} = \tilde{\boldsymbol{X}} - \boldsymbol{E}_f$, where n is the number of training examples, and d indicates the feature dimension. $\tilde{\boldsymbol{Y}} \in \mathbb{R}^{n\times c}$ is the matrix of corresponding observed noisy labels (c denotes the number of classes). The element $\tilde{\boldsymbol{Y}}_{i,j} = 1$ if the i-th example is labeled as j, and $\tilde{\boldsymbol{Y}}_{i,j} = 0$ otherwise, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, c$.

3.1.1 Low-Rank Projection from Feature to Label. To be more specific, here the noisy $\tilde{\boldsymbol{Y}}$ is also decomposed into two parts, *i.e.*, $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} + \boldsymbol{E}_l$. The first term is the correct label matrix $\boldsymbol{Y} = \boldsymbol{XZ}$, which is embedded based on the recovered feature matrix $\boldsymbol{X}$ with the projection matrix $\boldsymbol{Z}$. The second term $\boldsymbol{E}_l$ describes the error between the observed label and the accurate label. Moreover, it is reasonable to assume that the clean label matrix $\boldsymbol{Y} = \boldsymbol{XZ}$ lies in a low-dimensional subspace [42], as it is one-hot encoded with only c different classes. Therefore, we encourage the ground-truth label matrix $\boldsymbol{Y} = \boldsymbol{XZ}$ to be low-rank by constraining each factor matrix.

To be more specific, due to the high intra-class similarities of features, the latent clean feature matrix $\boldsymbol{X}$ is inherently low-rank, and it can be recovered from the two equality constraints regarding the feature matrix and label matrix. As a matter of course, the projection matrix $\boldsymbol{Z}$ is also low-rank, because we only need a small number of columns in $\boldsymbol{X}$ to construct the low-rank label space $\boldsymbol{Y} = \boldsymbol{XZ}$. Finally, here $\boldsymbol{Y} = \boldsymbol{XZ}$ is assumed as the true label matrix, so each element in it should be either 0 or 1.

3.1.2 Hybrid Noise Characterization. Considering the complexity and diversity of the feature noise, we have to leverage different norm metrics (*e.g.*, $\|\boldsymbol{E}_f\|_F^2$, $\|\boldsymbol{E}_f\|_1$, *etc.*) to constrain the feature noise matrix $\boldsymbol{E}_f$. Meanwhile, the number of the non-zero rows of label error matrix $\boldsymbol{E}_l$ is supposed to be small because those noisy labels are corrected while the remaining labels are still preserved. In this view, the label noise is characterized by the $\ell_{2,1}$-norm. By integrating the above regularization terms and contractions into a joint learning framework, we naturally have the following learning objective with double low-rank constraints:

$$\begin{aligned} &\min_{\substack{\boldsymbol{Z},\boldsymbol{X},\\ \boldsymbol{E}_f,\boldsymbol{E}_l}} \|\boldsymbol{X}\|_* + \lambda_1\|\boldsymbol{Z}\|_* + \lambda_2 \mathrm{R}(\boldsymbol{E}_f) + \lambda_3\|\boldsymbol{E}_l\|_{2,1} \\ &\text{s.t. } \tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{E}_f, \tilde{\boldsymbol{Y}} = \boldsymbol{XZ} + \boldsymbol{E}_l, \boldsymbol{XZ} \in \{0,1\}^{n\times c}, \end{aligned} \tag{1}$$

where $\mathrm{R}(\boldsymbol{E}_f)$ is the generalized regularization form of the feature noise $\boldsymbol{E_f}$. Here $\lambda_1, \lambda_2, \lambda_3 > 0$ are trade-off parameters. We note that Eq. (1) falls into an integer programming problem, which is generally NP-hard. To make Eq. (1) tractable, we relax the discrete constraint to a continues condition $\boldsymbol{XZ} \in [0,1]^{n\times c}$, which is a linear relaxation and has been used in several prior works [16, 40]. By solving the linear relaxation problem of Eq. (1), we can obtain the optimal projection matrix $\boldsymbol{Z}$, which can be subsequently applied as a projection based classifier. Since this classifier is learned from the denoising data, it is robust to hybrid noise, so that we can successfully predict the categories of the unseen test examples.

3.2 Optimization

We develop a unified optimization framework for both hybrid noise removal and classier learning. Our complex problem setting leads to a non-convex objective function. In this section, the non-convex ADMM [34, 37] is applied to address our proposed model with the ℓ_1-norm to constrain the feature noise $\boldsymbol{E}_f$.

By introducing three auxiliary variables and conducting linear relaxation, Eq. (1) is transformed to the following equivalent form:

$$\begin{aligned} &\min_{\substack{\boldsymbol{Z},\boldsymbol{X},\boldsymbol{B},\boldsymbol{J},\\ \boldsymbol{K},\boldsymbol{E}_f,\boldsymbol{E}_l}} \|\boldsymbol{X}\|_* + \lambda_1\|\boldsymbol{Z}\|_* + \lambda_2\|\boldsymbol{E}_f\|_1 + \lambda_3\|\boldsymbol{E}_l\|_{2,1}, \\ &\text{s.t. } \tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{E}_f, \tilde{\boldsymbol{Y}} = \boldsymbol{B} + \boldsymbol{E}_l, \boldsymbol{Z} = \boldsymbol{J}, \\ &\quad \boldsymbol{X} = \boldsymbol{K}, \boldsymbol{B} = \boldsymbol{KJ}, \boldsymbol{B} \in [0,1]^{n\times c}, \end{aligned} \tag{2}$$

and the augmented Lagrangian function of Eq. (2) with the continuous convex constraint is considered as:

$$\begin{aligned}
&\mathcal{L}\left(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{B}, \boldsymbol{J}, \boldsymbol{K}, \boldsymbol{E}_f, \boldsymbol{E}_l, \boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3, \boldsymbol{M}_4, \boldsymbol{M}_5\right) \\
&= \|\boldsymbol{X}\|_* + \lambda_1\|\boldsymbol{Z}\|_* + \lambda_2\|\boldsymbol{E}_f\|_1 + \lambda_3\|\boldsymbol{E}_l\|_{2,1} \\
&+ \operatorname{tr}(\boldsymbol{M}_1^\top(\tilde{\boldsymbol{X}} - \boldsymbol{X} - \boldsymbol{E}_f)) + \operatorname{tr}(\boldsymbol{M}_2^\top(\tilde{\boldsymbol{Y}} - \boldsymbol{B} - \boldsymbol{E}_l)) \\
&+ \operatorname{tr}(\boldsymbol{M}_3^\top(\boldsymbol{Z} - \boldsymbol{J})) + \operatorname{tr}(\boldsymbol{M}_4^\top(\boldsymbol{B} - \boldsymbol{K}\boldsymbol{J})) \\
&+ \operatorname{tr}(\boldsymbol{M}_5^\top(\boldsymbol{X} - \boldsymbol{K})) + \frac{\mu}{2}\left(\|\tilde{\boldsymbol{X}} - \boldsymbol{X} - \boldsymbol{E}_f\|_F^2 + \|\boldsymbol{Z} - \boldsymbol{J}\|_F^2\right. \\
&\left.+\|\tilde{\boldsymbol{Y}} - \boldsymbol{B} - \boldsymbol{E}_l\|_F^2 + \|\boldsymbol{B} - \boldsymbol{K}\boldsymbol{J}\|_F^2 + \|\boldsymbol{X} - \boldsymbol{K}\|_F^2\right),
\end{aligned} \tag{3}$$

where $\boldsymbol{M}_1$, $\boldsymbol{M}_2$, $\boldsymbol{M}_3$, $\boldsymbol{M}_4$, and $\boldsymbol{M}_5$ are Lagrangian multipliers, and $\mu > 0$ is the penalty coefficient.

Under the framework of ADMM, we can alternately minimize each of the variables $\boldsymbol{Z}$, $\boldsymbol{X}$, $\boldsymbol{B}$, $\boldsymbol{J}$, $\boldsymbol{K}$, $\boldsymbol{E}_f$ and $\boldsymbol{E}_l$ by keeping the others fixed in each iteration.

Update $\boldsymbol{X}$: The subproblem of $\boldsymbol{X}$ is

$$\begin{aligned}
&\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_* + \operatorname{tr}(\boldsymbol{M}_1^\top(\tilde{\boldsymbol{X}} - \boldsymbol{X} - \boldsymbol{E}_f)) + \operatorname{tr}(\boldsymbol{M}_5^\top(\boldsymbol{X} - \boldsymbol{K})) \\
&\quad + \frac{\mu}{2}(\|\tilde{\boldsymbol{X}} - \boldsymbol{X} - \boldsymbol{E}_f\|_F^2 + \|\boldsymbol{X} - \boldsymbol{K}\|_F^2) \\
\Rightarrow &\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_* + \mu\|\boldsymbol{X} - \frac{\boldsymbol{\mu}(\tilde{\boldsymbol{X}} - \boldsymbol{E}_f + \boldsymbol{K}) - \boldsymbol{M}_5 + \boldsymbol{M}_1}{2\boldsymbol{\mu}}\|_F^2 \\
\Rightarrow &\min_{\boldsymbol{X}} \frac{1}{2\mu}\|\boldsymbol{X}\|_* + \frac{1}{2}\|\boldsymbol{X} - \frac{\boldsymbol{\mu}(\tilde{\boldsymbol{X}} - \boldsymbol{E}_f + \boldsymbol{K}) - \boldsymbol{M}_5 + \boldsymbol{M}_1}{2\boldsymbol{\mu}}\|_F^2 \\
\Rightarrow &\min_{\boldsymbol{X}} \tau\|\boldsymbol{X}\|_* + \frac{1}{2}\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_F^2,
\end{aligned} \tag{4}$$

in which $\tau = \frac{1}{2\mu}$ and $\hat{\boldsymbol{X}} = \frac{\mu(\tilde{\boldsymbol{X}} - \boldsymbol{E}_f + \boldsymbol{K}) - \boldsymbol{M}_5 + \boldsymbol{M}_1}{2\mu}$. The closed-form solution to Eq. (4) is

$$\boldsymbol{X} = \boldsymbol{U}_x diag\left(\max\left\{\Sigma_{x(ii)} - \tau, 0\right\}\right)\boldsymbol{V}_x^\top, \tag{5}$$

where $\forall i = 1, 2, \ldots, \min(n, d)$, $\boldsymbol{U}_x$ and $\boldsymbol{V}_x$ are obtained by conducting the Singular Value Decomposition (SVD) on $\hat{\boldsymbol{X}}$ (*i.e.*, $\hat{\boldsymbol{X}} = \boldsymbol{U}_x\Sigma_x\boldsymbol{V}_x^\top$), and $\Sigma_{x(ii)}$ is the i-th diagonal element of the singular value matrix Σ_x.

Update $\boldsymbol{Z}$: The subproblem related to the variable $\boldsymbol{Z}$ is

$$\begin{aligned}
&\min_{\boldsymbol{Z}} \lambda_1\|\boldsymbol{Z}\|_* + \operatorname{tr}(\boldsymbol{M}_3^\top(\boldsymbol{Z} - \boldsymbol{J})) + \frac{\mu}{2}\|\boldsymbol{Z} - \boldsymbol{J}\|_F^2 \\
\Rightarrow &\min_{\boldsymbol{Z}} \frac{\lambda_1}{\mu}\|\boldsymbol{Z}\|_* + \frac{1}{2}\|\boldsymbol{Z} - \boldsymbol{J} + \frac{\boldsymbol{M}_3}{\mu}\|_F^2 \\
\Rightarrow &\min_{\boldsymbol{Z}} \eta\|\boldsymbol{Z}\|_* + \frac{1}{2}\|\boldsymbol{Z} - \hat{\boldsymbol{Z}}\|_F^2,
\end{aligned} \tag{6}$$

where $\eta = \frac{\lambda_1}{\mu}$ and $\hat{\boldsymbol{Z}} = \boldsymbol{J} - \frac{\boldsymbol{M}_3}{\mu}$. Similarly to the closed-form solution Eq. (5), we have

$$\boldsymbol{Z} = \boldsymbol{U}_z diag\left(\max\left\{\Sigma_{z(ii)} - \eta, 0\right\}\right)\boldsymbol{V}_z^\top, \tag{7}$$

in which $\forall i = 1, 2, \ldots, \min(d, c)$.

Update $\boldsymbol{B}$: The subproblem on $\boldsymbol{B}$ with the continuous convex set $\boldsymbol{B} \in [0, 1]^{n\times c}$ is

$$\begin{aligned}
&\min_{\boldsymbol{B}} \operatorname{tr}(\boldsymbol{M}_2^\top(\tilde{\boldsymbol{Y}} - \boldsymbol{B} - \boldsymbol{E}_l)) + \operatorname{tr}(\boldsymbol{M}_4^\top(\boldsymbol{B} - \boldsymbol{K}\boldsymbol{J})) \\
&\quad + \frac{\mu}{2}\left(\|\tilde{\boldsymbol{Y}} - \boldsymbol{B} - \boldsymbol{E}_l\|_F^2 + \|\boldsymbol{B} - \boldsymbol{K}\boldsymbol{J}\|_F^2\right).
\end{aligned} \tag{8}$$

Obviously, the optimal $\boldsymbol{B}^*$ to Eq. (8) can be expressed as

$$\boldsymbol{B}^* = \frac{\mu(\tilde{\boldsymbol{Y}} - \boldsymbol{E}_l + \boldsymbol{K}\boldsymbol{J}) + \boldsymbol{M}_2 - \boldsymbol{M}_4}{2\mu}. \tag{9}$$

To restrict $\boldsymbol{B}^*$ to the feasible region, all its elements can be further projected to $[0, 1]$ as

$$\boldsymbol{B}_{ij} = \Pi(\boldsymbol{B}^*_{ij}), \tag{10}$$

where the projection function $\Pi(x) = 1$ and $= 0$ for $x > 1$ and $x < 0$, respectively, and $\Pi(x) = x$ for any $x \in [0, 1]$.

Update $\boldsymbol{J}$: The subproblem regarding $\boldsymbol{J}$ is

$$\begin{aligned} \min_{\boldsymbol{J}} \operatorname{tr}\left(\boldsymbol{M}_3^\top (\boldsymbol{Z} - \boldsymbol{J})\right) + \operatorname{tr}\left(\boldsymbol{M}_4^\top (\boldsymbol{B} - \boldsymbol{K}\boldsymbol{J})\right) \\ + \frac{\mu}{2}\left(\|\boldsymbol{Z} - \boldsymbol{J}\|_F^2 + \|\boldsymbol{B} - \boldsymbol{K}\boldsymbol{J}\|_F^2\right). \end{aligned} \tag{11}$$

By computing the derivation of Eq. (11) *w.r.t.* $\boldsymbol{J}$ and then setting it as zero, the closed-form solution is

$$\boldsymbol{J} = \left(\boldsymbol{I} + \boldsymbol{K}^\top \boldsymbol{K}\right)^{-1} \left(\boldsymbol{Z} + \boldsymbol{K}^\top \boldsymbol{B} + \frac{\boldsymbol{M}_3 + \boldsymbol{K}^\top \boldsymbol{M}_4}{\mu}\right). \tag{12}$$

Update $\boldsymbol{K}$: By dropping the unrelated terms to $\boldsymbol{K}$, we have

$$\begin{aligned} \min_{\boldsymbol{K}} \operatorname{tr}(\boldsymbol{M}_4^\top (\boldsymbol{B} - \boldsymbol{K}\boldsymbol{J})) + \operatorname{tr}(\boldsymbol{M}_5^\top (\boldsymbol{X} - \boldsymbol{K})) \\ + \frac{\mu}{2}(\|\boldsymbol{B} - \boldsymbol{K}\boldsymbol{J}\|_F^2 + \|\boldsymbol{X} - \boldsymbol{K}\|_F^2). \end{aligned} \tag{13}$$

Similarly, we compute the derivation of Eq. (13) *w.r.t.* $\boldsymbol{K}$ and then set it as zero, so we have

$$\boldsymbol{K} = \left(\frac{\boldsymbol{M}_4 \boldsymbol{J}^\top + \boldsymbol{M}_5}{\mu} + \boldsymbol{B}\boldsymbol{J}^\top + \boldsymbol{X}\right)\left(\boldsymbol{J}\boldsymbol{J}^\top + \boldsymbol{I}\right)^{-1}. \tag{14}$$

Update $\boldsymbol{E}_f$ constrained by ℓ_1-norm:

$$\begin{aligned} &\min_{\boldsymbol{E}_f} \lambda_2 \|\boldsymbol{E}_f\|_1 + \operatorname{tr}(\boldsymbol{M}_1^\top (\tilde{\boldsymbol{X}} - \boldsymbol{X} - \boldsymbol{E}_f)) + \frac{\mu}{2}\|\tilde{\boldsymbol{X}} - \boldsymbol{X} - \boldsymbol{E}_f\|_F^2 \\ \Rightarrow &\min_{\boldsymbol{E}_f} \frac{\lambda_2}{\mu}\|\boldsymbol{E}_f\|_1 + \frac{1}{2}\|\boldsymbol{E}_f - \left(\tilde{\boldsymbol{X}} - \boldsymbol{X} + \frac{\boldsymbol{M}_1}{\mu}\right)\|_F^2 \\ \Rightarrow &\min_{\boldsymbol{E}_f} \omega\|\boldsymbol{E}_f\|_1 + \frac{1}{2}\|\boldsymbol{E}_f - \hat{\boldsymbol{E}}_f\|_F^2, \end{aligned} \tag{15}$$

where $\omega = \frac{\lambda_2}{\mu}$ and $\hat{\boldsymbol{E}}_f = \tilde{\boldsymbol{X}} - \boldsymbol{X} + \frac{\boldsymbol{M}_1}{\mu}$. We can extend the soft-thresholding (shrinkage) operator introduced in [22, 32] to the matrix by applying it in an element-wise way, so all its elements can be updated as

$$[\boldsymbol{E}_f]_{ij} = \begin{cases} [\hat{\boldsymbol{E}}_f]_{ij} - \omega, & if \ \ [\hat{\boldsymbol{E}}_f]_{ij} > \omega, \\ [\hat{\boldsymbol{E}}_f]_{ij} + \omega, & if \ \ [\hat{\boldsymbol{E}}_f]_{ij} < -\omega, \\ 0, & otherwise, \end{cases} \tag{16}$$

where $[\cdot]_{ij}$ represents the (i, j)-th element of the corresponding matrix.

Update $\boldsymbol{E}_l$:

$$\begin{aligned} &\min_{\boldsymbol{E}_l} \lambda_3 \|\boldsymbol{E}_l\|_{2,1} + \operatorname{tr}(\boldsymbol{M}_2^\top (\tilde{\boldsymbol{Y}} - \boldsymbol{B} - \boldsymbol{E}_l)) + \frac{\mu}{2}(\|\tilde{\boldsymbol{Y}} - \boldsymbol{B} - \boldsymbol{E}_l\|_F^2) \\ \Rightarrow &\min_{\boldsymbol{E}_l} \frac{\lambda_3}{\mu}\|\boldsymbol{E}_l\|_{2,1} + \frac{1}{2}\|\boldsymbol{E}_l - \left(\tilde{\boldsymbol{Y}} - \boldsymbol{B} + \frac{\boldsymbol{M}_2}{\mu}\right)\|_F^2 \\ \Rightarrow &\min_{\boldsymbol{E}_l} \xi\|\boldsymbol{E}_l\|_{2,1} + \frac{1}{2}\|\boldsymbol{E}_l - \hat{\boldsymbol{E}}_l\|_F^2, \end{aligned} \tag{17}$$

Algorithm 1: The algorithm for solving the proposed FLR

Input: noisy feature matrix $\tilde{\boldsymbol{X}}$, noisy label matrix $\tilde{\boldsymbol{Y}}$,
trade-off parameters: $\lambda_1 > 0$, $\lambda_2 > 0$, $\lambda_3 > 0$.
Output: optimized $\boldsymbol{X}^*$ and $\boldsymbol{Z}^*$.

1 Let $\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{J}$, $\boldsymbol{E}_f$, $\boldsymbol{E}_l$, $\boldsymbol{B}$, $\boldsymbol{K}$, $\boldsymbol{M}_1$, $\boldsymbol{M}_2$, $\boldsymbol{M}_3$, $\boldsymbol{M}_4$, and $\boldsymbol{M}_5$ as zero matrices, $\mu = 10^{-3}$, $\rho = 1.2$, $\epsilon = 10^{-6}$, $iter\_max = 1000$, $iter = 0$;
2 **while** *not converge* **do**
3 Update $\boldsymbol{X}$ via Eq. (5), $\boldsymbol{Z}$ via Eq. (7), $\boldsymbol{B}$ via Eq. (10), $\boldsymbol{J}$ via Eq. (12), $\boldsymbol{K}$ via Eq. (14), $\boldsymbol{E}_f$ via Eq. (16), $\boldsymbol{E}_l$ via Eq. (18), respectively.
4 Update the multipliers
5 $\boldsymbol{M}_1 := \boldsymbol{M}_1 + \mu(\tilde{\boldsymbol{X}} - \boldsymbol{X} - \boldsymbol{E}_f)$,
6 $\boldsymbol{M}_2 := \boldsymbol{M}_2 + \mu(\tilde{\boldsymbol{Y}} - \boldsymbol{B} - \boldsymbol{E}_l)$,
7 $\boldsymbol{M}_3 := \boldsymbol{M}_3 + \mu\,(\boldsymbol{Z} - \boldsymbol{J})$,
8 $\boldsymbol{M}_4 := \boldsymbol{M}_4 + \mu\,(\boldsymbol{B} - \boldsymbol{K}\boldsymbol{J})$,
9 $\boldsymbol{M}_5 := \boldsymbol{M}_5 + \mu\,(\boldsymbol{X} - \boldsymbol{K})$.
10 Update the parameter μ by $\mu := \rho\mu$.
11 $iter := iter + 1$.
12 Check the convergence conditions:
13 $\left\|\tilde{\boldsymbol{X}} - \boldsymbol{X} - \boldsymbol{E}_f\right\|_F \leq \epsilon$ and $\left\|\tilde{\boldsymbol{Y}} - \boldsymbol{B} - \boldsymbol{E}_l\right\|_F \leq \epsilon$ and $\|\boldsymbol{Z} - \boldsymbol{J}\|_F \leq \epsilon$ and $\|\boldsymbol{B} - \boldsymbol{K}\boldsymbol{J}\|_F \leq \epsilon$ and $\|\boldsymbol{X} - \boldsymbol{K}\|_F \leq \epsilon$; or $iter > iter\_max$.
14 **end**

in which $\xi = \frac{\lambda_3}{\mu}$ and $\hat{\boldsymbol{E}}_l = \tilde{\boldsymbol{Y}} - \boldsymbol{B} + \frac{\boldsymbol{M}_2}{\mu}$.

As provided in [40], the closed-form solution to the general optimization problem related to $\ell_{2,1}$-norm is:

$$[\boldsymbol{E}_l]_i = \begin{cases} \dfrac{\|[\hat{\boldsymbol{E}}_l]_i\|_2 - \xi}{\|[\hat{\boldsymbol{E}}_l]_i\|_2}[\hat{\boldsymbol{E}}_l]_i, & if \;\; \|[\hat{\boldsymbol{E}}_l]_i\|_2 > \xi, \\ 0, & otherwise, \end{cases} \tag{18}$$

where $[\cdot]_i$ represents the i-th row of the related matrix.

The entire optimization process for our model is summarized in Algorithm 1. These Lagrangian multipliers (*i.e.*, $\{\boldsymbol{M}_i\}_{i=1}^5$) are introduced to enforce the constraints of the augmented Lagrangian, ensuring that the feature and label reconstruction process remains consistent throughout the iterations. The rule utilized to update the penalty parameter μ (a positive scalar) is commonly used in ADMM. Different values of penalty parameter μ are employed for each iteration, with the goal of improving the convergence in practice, as well as making performance less dependent on the initial choice of the penalty parameter.

3.2.1 Computational Complexity. Eq. (5) and Eq. (7) in Algorithm 1 are both accomplished by SVD, and their complexities are $O(\min(n^2, nd^2))$ and $O(\min(n^2d, nd^2))$, respectively. Furthermore, two $d \times d$ matrices are inverted in Eq. (12) and Eq. (14), and the complexity of these two steps is $O(2d^3)$. In Eq. (16), one should compute the ℓ_1-norm of a $n \times d$ matrix $\boldsymbol{E}_f$, so the complexity is $O(nd)$. In Eq. (18), one should compute the $\ell_{2,1}$-norm of each row of a $n \times c$ matrix $\boldsymbol{E}_l$, so the complexity is $O(nc)$. Thereby, the total computational complexity of Algorithm 1 is $O\left(\left(\min(n^2, nd^2) + \min(n^2d, nd^2) + 2d^3 + nd + nc\right)k\right)$ by assuming that all equations in Line 3 are

iterated k times. Note that the complexity of Algorithm 1 is squared to the number of training examples n at most (as lots of other robust learning approaches), so the complexity is reasonably acceptable in practical uses.

4 Theoretical Analyses

This section provides the theoretical analyses on FLR. Firstly, we prove that the optimization process in Algorithm 1 will converge to a stationary point and then theoretically reveal that the generalization risk of FLR is upper bounded.

4.1 Proof of Convergence

Note that our unified optimization objective is non-convex with multi-variables. It is difficult to guarantee the convergence of the standard ADMM optimal framework for solving multi-block problems, especially with non-convex constraints, so it is unrealistic to analyze the global optimality of our iteration algorithm. Therefore, here we investigate the local convergence of Algorithm 1, which reveals that the iteration points converge to a stationary point under some mild conditions.

Theorem 4.1. *Let* $\{\Gamma_t = (\boldsymbol{X}_t, \boldsymbol{Z}_t, \boldsymbol{B}_t, \boldsymbol{J}_t, \boldsymbol{K}_t, \boldsymbol{E}_{l,t}, \boldsymbol{E}_{f,t}, \{\boldsymbol{M}_{i,t}\}_{i=1}^5)\}_{t=1}^\infty$ *be the sequence generated by Algorithm 1. Assume that* $\lim_{t\to+\infty} \mu_t(\boldsymbol{K}_{t+1} - \boldsymbol{K}_t) = 0$, $\lim_{t\to+\infty} \mu_t(\boldsymbol{J}_{t+1} - \boldsymbol{J}_t) = 0$, $\lim_{t\to+\infty} \mu_t(\boldsymbol{E}_{f,t+1} - \boldsymbol{E}_{f,t}) = 0$, $\lim_{t\to+\infty} \mu_t(\boldsymbol{E}_{l,t+1} - \boldsymbol{E}_{l,t}) = 0$, *and then we have*

1. *The sequence* $\{\Gamma_t\}_{t=1}^\infty$ *is bounded;*
2. *The sequence* $\{\Gamma_t\}_{t=1}^\infty$ *has at least one accumulation point. For any accumulation point* $\Gamma^* = (\boldsymbol{X}^*, \boldsymbol{Z}^*, \boldsymbol{B}^*, \boldsymbol{J}^*, \boldsymbol{K}^*, \boldsymbol{E}_f^*, \boldsymbol{E}_l^*, \{\boldsymbol{M}_i^*\}_{i=1}^5)$, $(\boldsymbol{X}^*, \boldsymbol{Z}^*, \boldsymbol{B}^*, \boldsymbol{J}^*, \boldsymbol{K}^*, \boldsymbol{E}_f^*, \boldsymbol{E}_l^*)$ *is a stationary point of the optimization Eq.* (2).

Note that the above limiting equations are actually mild conditions and are also used in [13, 47]. The detailed proof is included in Appendix. Moreover, we empirically verify the convergence of our proposed algorithm by experimental results.

4.2 Generalization Bound

This subsection analyses the property of generalizability of our proposed FLR.

4.2.1 Preliminaries. Note that the goal of FLR is to find a suitable project matrix $\boldsymbol{Z}$, given the corrupted example features $\tilde{\boldsymbol{X}}$ and noisy labels $\tilde{\boldsymbol{Y}}$.

Eq. (1) can be rewritten to the following expression with several hard constraints, and that is

$$\begin{aligned} &\min_{\substack{\boldsymbol{Z},\boldsymbol{X},\\ \boldsymbol{E}_f,\boldsymbol{E}_l}} \sum_{(i,j)} \ell\left((\boldsymbol{X}\boldsymbol{Z} + \boldsymbol{E})_{i,j}, \tilde{\boldsymbol{Y}}_{i,j}\right) \\ &\text{s.t. } \|\boldsymbol{X}\|_* \le \mathcal{X}_*, \|\boldsymbol{Z}\|_* \le \mathcal{Z}_*, \|\boldsymbol{E}_l\|_{2,1} \le \mathcal{E}_{l,21}, \\ &\qquad \|\tilde{\boldsymbol{X}} - \boldsymbol{X}\|_1 \le \mathcal{E}_{f,1}, \boldsymbol{X}\boldsymbol{Z} \in \{0,1\}^{n\times c}, \end{aligned} \tag{19}$$

where $\boldsymbol{E}_f = \tilde{\boldsymbol{X}} - \boldsymbol{X}$ and $(i, j) \in \{1, \ldots, n\} \times \{1, \ldots, c\}$. The decision function $f_\theta : \tilde{\mathcal{X}} \to \tilde{\mathcal{Y}}$, and $f_\theta(i, j) = (\tilde{\boldsymbol{X}}_i - \boldsymbol{E}_{fi})\boldsymbol{Z}\boldsymbol{I}^j + \boldsymbol{E}_{ij}$, which is controlled by the feasible solution $\theta = (\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{E}_f, \boldsymbol{E}_l)$. The feasible solution set $\Theta = \left\{(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{E}_f,\boldsymbol{E}_l) \middle| \|\boldsymbol{X}\|_* \le \mathcal{X}_*, \|\boldsymbol{Z}\|_* \le \mathcal{Z}_*, \|\boldsymbol{E}_l\|_{2,1} \le \mathcal{E}_{l,21}, \|\tilde{\boldsymbol{X}} - \boldsymbol{X}\|_F \le \mathcal{E}_{f,1}, \boldsymbol{X}\boldsymbol{Z} \in \{0,1\}^{n\times c}\right\}$ and the set of feasible functions $\mathcal{F}_\Theta = \{f_\theta | \theta \in \Theta\}$. $\boldsymbol{I}^j$ is the j-th column of identity matrix $\boldsymbol{I} \in \mathbb{R}^{c\times c}$, and $\boldsymbol{X}_i$ is the i-th row of matrix $\boldsymbol{X}$ in general.

Furthermore, we introduce the following two "ℓ-risk" quantities:

- Expected "ℓ-risk": $R_\ell(f) = \mathbb{E}_{i,j}\left[\ell\left(f(i, j), \tilde{\boldsymbol{Y}}_{i,j}\right)\right]$;

- Empirical "ℓ-risk": $\hat{R}_\ell(f) = \frac{1}{n_r}\sum_{(i,j)} \ell\left(f(i,j), \tilde{Y}_{i,j}\right)$,

where n_r is the number of observed entries. Therefore, the proposed FLR is designed to obtain a proper θ^* that parameterizes the optimal $f_\theta^* = \arg\min_{f\in\mathcal{F}_\Theta} \hat{R}_\ell(f)$.

4.2.2 Generalization Bound of FLR. The Rademacher complexity quantitatively measures the diversity of a function class, which is a useful tool for analyzing the bound of the generalization error of the learning algorithm. Specifically, we link the quality of training features and labels to Rademacher complexity, and show that the high-quality features and labels will result in the lower model complexity and thus a smaller error bound. To this end, we denote that $\mathcal{R}_n(\mathcal{F}) := \mathbb{E}[\mathcal{R}(\mathcal{F})]$ as the Rademacher complexity of the function class $\mathcal{F}$, and $\mathcal{R}(\mathcal{F}) := \mathbb{E}_\sigma[\sup_{f\in\mathcal{F}} \frac{1}{n_r}\sum_{\alpha=1}^{n_r} \sigma_\alpha f(\alpha)]$ as the empirical Rademacher complexity on the training set, where $\sigma_i \in \{-1, 1\}$ $(i = 1, 2, \ldots, n_r)$ are *i.i.d.* Rademacher random variables.

Theorem 4.2. *Let ℓ be the loss function bounded by $\mathcal{B}$ with Lipschitz constant L_ℓ, and δ be a constant where $0 \le \delta \le 1$. Then with probability at least $1-\delta$, we have*

$$\max_{f\in\mathcal{F}_\Theta} |R_l(f) - \hat{R}_l(f)| \le 2L_\ell \mathcal{R}_n(\mathcal{F}_\Theta) + \mathcal{B}\sqrt{\frac{\ln 1/\delta}{2nc}}, \tag{20}$$

in which

$$\mathcal{R}_n(\mathcal{F}_\Theta) \le \mathcal{E}_{l,21}C_1 + \min\left\{\mathcal{X}_*\mathcal{Z}_*C_2, \mathcal{Z}_*(\tilde{\mathcal{X}}_F + \sqrt{d}\mathcal{E}_{f,1})C_3, C_4\right\}, \tag{21}$$

with $C_1 = \sqrt{\frac{3\ln c}{nc}}$, $C_2 = \sqrt{\frac{\ln(2n_c)}{nc}}$, $C_3 = \frac{1}{\sqrt{nc}}$ *and* $C_4 = \sqrt{\frac{2}{c}}$.

The proof of Theorem 4.2 is provided in Appendix. As mentioned before, the matrix $\boldsymbol{E}_f$ captures the feature noise, and its ℓ_1-norm value is upper bounded by $\mathcal{E}_{f,1}$. Also, the label noise matrix $\boldsymbol{E}_l$ model is upper bounded by $\mathcal{E}_{l,21}$. The recovered clean feature matrix $\boldsymbol{X}$ is assumed to be sparse. Specifically, the $\mathcal{E}_{f,1}$ and $\mathcal{E}_{l,21}$ are governed by the severity of feature noise and label noise, respectively. The light noise will lead to small $\mathcal{E}_{f,1}$ and $\mathcal{E}_{l,21}$, which can further reduce the upper bound of the expected ℓ-risk in the right-hand side of Eq. (21). Conclusively, the high-quality features and labels of training examples will contribute to a lower model complexity and thus ensure a smaller error bound.

Table 1. The comparison of mean test accuracy (%) of various methods on four UCI benchmark datasets. The highest and second highest records are **bolded** and underlined, respectively.

Gaussian Noise	Glass ($n=214, d=9, c=6$)			Wine ($n=178, d=13, c=3$)			CNAE9 ($n=1080, d=856, c=9$)			Waveform($n=5000, d=21, c=3$)		
(σ_f, η_l)	(0.2, 0.3)	(0.2, 0.6)	(0.5, 0.6)	(0.2, 0.3)	(0.2, 0.6)	(0.5, 0.6)	(0.2, 0.3)	(0.2, 0.6)	(0.5, 0.6)	(0.2, 0.3)	(0.2, 0.6)	(0.5, 0.6)
RPCA [1]	45.25±3.12	35.45±1.32	31.47±2.43	85.65±2.11	73.41±1.78	67.41±2.13	65.25±1.21	50.12±1.76	28.45±1.33	55.15±1.12	50.12±1.76	40.40±1.24
GCE [49]	24.55±13.09	26.37±17.72	23.64±13.78	40.00±17.30	40.00±16.85	36.67±21.37	49.26±25.75	29.07±13.30	18.52±4.49	62.00±9.73	56.04±8.67	55.08±8.45
Co-teaching [15]	57.67±4.26	38.53±9.18	48.22±6.12	89.99±5.05	57.78±14.62	48.33±21.57	76.60±2.80	50.19±2.39	30.62±6.93	81.59±0.55	71.55±1.31	70.26±2.76
JoCoR [39]	37.36±2.45	26.36±4.21	27.67±5.57	83.33±3.93	44.44±3.40	43.89±10.28	43.30±0.70	30.02±2.15	21.40±2.77	56.57±2.00	40.19±1.33	39.43±3.35
$\mathcal{L}_{DMI}$ [43]	54.55±3.21	39.09±6.10	24.55±13.09	90.00±8.24	76.67±17.74	76.67±7.24	88.33±2.82	69.63±8.37	60.00±4.87	**84.96±2.23**	69.68±3.99	57.36±14.21
LQF [50]	43.64±10.47	37.27±6.74	34.54±5.18	89.82±7.86	74.44±19.48	68.89±12.79	49.63±14.07	30.92±10.85	13.15±2.65	80.36±3.69	67.92±1.61	61.48±2.15
FLR (ours)	**68.37±4.82**	**66.98±6.45**	**51.43±2.13**	**96.57±1.27**	**84.57±6.26**	**78.28±1.57**	**89.72±1.10**	**76.57±3.58**	**67.49±1.60**	84.48±1.05	**76.08±1.13**	**71.98±1.06**
Laplacian Noise												
RPCA [1]	45.25±3.12	39.21±2.56	32.74±3.65	92.25±0.12	75.20±2.50	62.04±3.55	62.52±3.11	45.52±2.05	29.24±1.10	52.25±0.11	55.25±5.02	47.42±2.30
GCE [49]	23.64±12.61	25.45±12.28	21.82±14.15	40.00±22.01	28.89±19.40	31.11±21.37	47.59±24.04	24.44±9.32	17.41±6.33	59.68±7.06	55.84±8.39	51.28±7.71
Co-teaching [15]	58.60±4.50	38.53±3.49	35.63±2.99	91.67±1.96	45.00±18.15	44.45±5.20	71.98±0.86	44.54±6.12	20.06±5.79	80.96±0.48	72.64±1.34	69.15±2.85
JoCoR [39]	38.68±4.16	28.22±4.70	26.12±5.75	83.33±4.39	52.78±14.30	38.33±11.69	38.86±1.16	27.95±3.44	16.37±1.68	55.93±2.86	40.89±3.13	35.68±4.01
$\mathcal{L}_{DMI}$ [43]	44.55±13.41	24.55±19.97	28.18±10.85	92.22±8.42	76.66±13.26	61.11±13.03	**85.37±2.48**	67.59±5.36	51.67±4.87	81.72±2.33	65.76±3.93	60.44±3.12
LQF [50]	38.18±8.26	29.38±5.71	29.09±11.85	**93.33±4.65**	70.00±15.52	67.78±14.38	53.15±6.27	26.85±9.49	12.22±2.49	78.48±4.56	60.92±18.63	52.04±11.26
FLR (ours)	**68.37±4.82**	**50.48±1.06**	**47.14±1.99**	92.56±2.56	**81.71±3.26**	**74.86±4.70**	84.91±2.31	**68.60±1.31**	**53.33±3.50**	**82.16±2.10**	**74.60±0.10**	**74.68±1.04**

5 Experimental Results

In this section, we show the experimental results on various datasets to validate the effectiveness of our proposed FLR. In detail, we first compare our FLR with existing feature noise learning

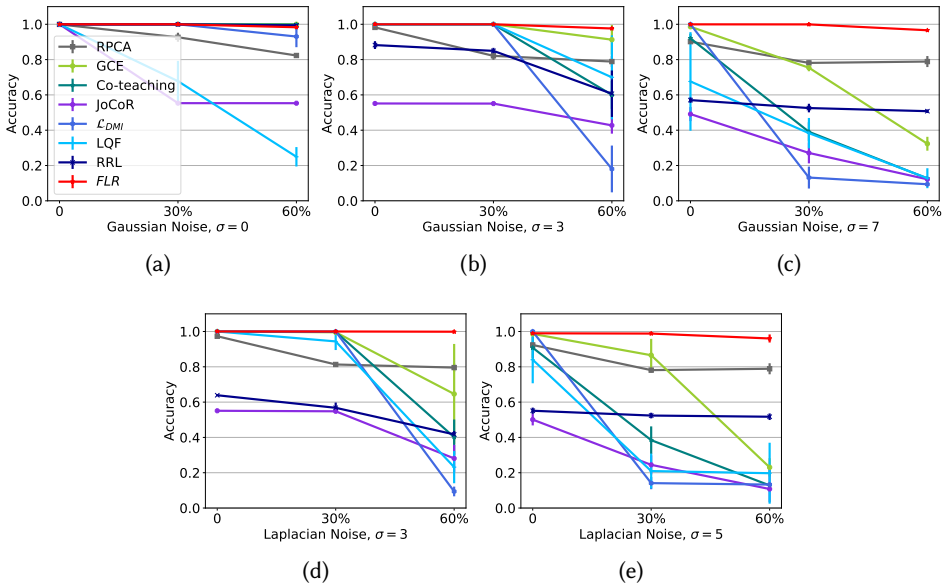


Fig. 3. The experimental results on CIFAR-10 dataset in various noise scenarios.

and label noise learning methods in different noisy setting cases. Our experimental data includes four classical UCI benchmark datasets and the popular CIFAR-10 and CIFAR-10N datasets. In our experiments, the training data is corrupted (with both the feature noise and label noise) while the test data is clean.

5.1 Experiments on UCI Benchmark Datasets

We compare FLR with six baseline algorithms on four UCI benchmark datasets, including Glass, Wine, CNAE9, and Waveform. Note that 80% of the examples in each dataset are randomly chosen to establish the training set, and the remaining 20% ones are severed as the test set. To incorporate different levels of hybrid noise into training sets, we add the Gaussian noise or Laplacian noise under different standard deviations σ_f with mean $\mu = 0$ on features of examples. Meanwhile, we randomly pick up 0%, 30%, 60% examples from the training sets and inject symmetric label noise [30] to these selected examples, and the noise rate is marked as η_l. Such contamination and partition are conducted five times, so the accuracies are the mean values of the outputs of five independent trials.

Our compared baseline methods include RPCA + a classifier [1], GCE [49], Co-teaching [15], JoCoR [39], $\mathcal{L}_{DMI}$ [43], LQF [50], and RRL [20]. Note that the first approach is designed for feature noise learning (it needs an additional classifier for the classification task), and the following six focus on label noise learning. Furthermore, RRL is designed for image tasks with data noise, so it is not suitable for UCI datasets.

The classification accuracies of all compared approaches on the test set are listed in Table 1. We observe that FLR yields better performance than other baseline methods in most cases or achieves the similar performance to them. Especially with increasing levels of hybrid noise, the accuracies of most compared methods decrease correspondingly, but FLR still performs robustly in those cases.

Table 2. The comparison of mean test accuracy (%) of various methods on CIFAR-10N dataset with the Gaussian noise. The highest and second highest records are **bolded** and underlined, respectively.

	Datasets	C-10N Aggre.	C-10N Random1	C-10N Worst
$\sigma_f = 3$	RPCA [1]	90.24±0.78	82.12±2.23	78.89±4.52
	GCE [49]	88.88±4.23	86.78±1.89	79.97±8.55
	Co-teaching [15]	94.03±0.06	94.14±0.05	90.16±1.83
	JoCoR [39]	52.94±0.42	52.94±0.38	49.04±0.33
	$\mathcal{L}_{DMI}$ [43]	**96.03±0.08**	95.75±0.23	47.85±22.49
	LQF [50]	94.23±2.78	95.82±0.23	92.71±4.40
	RRL [20]	89.70±1.44	85.00±1.18	75.62±0.89
	FLR (ours)	95.70±0.15	**96.05±0.05**	**94.92±0.06**
$\sigma_f = 7$	RPCA [1]	68.47±0.78	69.12±1.75	48.59±3.72
	GCE [49]	55.78±9.42	56.40±8.52	42.45±5.67
	Co-teaching [15]	78.68±5.32	70.12±5.41	19.43±4.31
	JoCoR [39]	52.91±0.50	52.49±0.21	42.08±0.76
	$\mathcal{L}_{DMI}$ [43]	91.93±6.74	42.05±22.82	34.43±11.29
	LQF [50]	82.42±6.74	67.30±13.17	8.97±1.97
	RRL [20]	61.63±1.08	55.93±0.80	50.98±1.12
	FLR (ours)	**95.93±0.06**	**93.75±0.10**	**95.08±0.06**

5.2 Experiments on CIFAR-10 Dataset

To further evaluate the effectiveness of FLR, we conduct experiments on the CIFAR-10 dataset, which contains 60,000 natural images across 10 classes. In our experiments, we follow the common practice [18, 25] to randomly pick up 6,000 image examples as our experimental data. The resolution of each image is 32×32. We extract the CNN features for each image, which are obtained from the fully connected layer of a pre-trained ResNet-18 with 512-dimensional features. The experimental settings are similar to those of UCI datasets.

From Figure 3, we can observe that FLR performs significantly better than baseline methods, especially when the training examples are heavily corrupted by hybrid noise. This is because FLR integrates the feature noise and label noise into a natural framework, so that we can successfully deal with the hybrid noise by merely using a single training process.

5.3 Experiments on CIFAR-10N Datasets

We compare FLR with the above baseline methods on the practical noisy dataset CIFAR-10N. Here CIFAR-10N consists of five inherently noisy label sets, among which the noise rate of "Random1", "Aggre.", and "Worst" are 17.23%, 9.03%, and 40.21%, respectively. We use three subsets of "Aggre.", "Random1" and "Worst" for our experiments, which contain 10,000 examples selected from their training sets and 2,000 examples from their test sets, respectively. We also extract the CNN features for each image.

All experimental settings are similar to those above. As labels in CIFAR-10N are inherently noisy, we need to artificially introduce the feature noise to training data. The average accuracy over three independent trials is reported. The experimental results under Gaussian noise are shown in Table 2 and experimental results under Laplacian noise are provided in Table 3.

From Table 2 and Table 3, we can clearly observe that when the feature noise is not heavy, FLR is comparable with other baseline methods. Furthermore, FLR can achieve much more robust and satisfactory performances when the σ_f is relatively large, while all baseline methods perform

Table 3. The comparison of mean test accuracy (%) of various methods on CIFAR-10N dataset with the Laplacian Noise. The highest and second highest records are **bolded** and underlined, respectively.

	Datasets	C-10N Aggre.	C-10N Random1	C-10N Worst
$\sigma_f = 3$	RPCA [1]	89.12±0.28	84.01±2.75	72.21±1.53
	GCE [49]	85.68±1.68	85.08±2.41	70.13±8.61
	Co-teaching [15]	95.62±0.32	95.33±0.17	77.42±5.57
	JoCoR [39]	52.91±0.50	52.49±0.21	42.08±0.76
	$\mathcal{L}_{DMI}$ [43]	92.35±6.11	**96.10±0.09**	19.12±9.70
	LQF [50]	95.66±0.45	95.70±0.64	66.02±15.33
	RRL [20]	65.40±0.85	65.37±1.32	59.78±1.60
	FLR (ours)	**95.78±0.05**	95.70±0.18	**94.12±0.10**
$\sigma_f = 5$	RPCA [1]	68.72±0.65	65.22±1.15	48.21±1.75
	GCE [49]	59.38±7.47	53.40±14.81	36.92±8.91
	Co-teaching [15]	46.17±10.73	30.75±8.15	13.38±1.63
	JoCoR [39]	25.77±7.52	22.94±4.14	12.81±1.53
	$\mathcal{L}_{DMI}$ [43]	72.70±24.25	52.40±22.15	21.07±18.04
	LQF [50]	52.98±8.71	51.83±16.16	12.70±5.17
	RRL [20]	51.87±1.62	50.37±1.11	50.05±0.44
	FLR (ours)	**76.00±0.31**	**78.85±0.39**	**61.23±4.00**

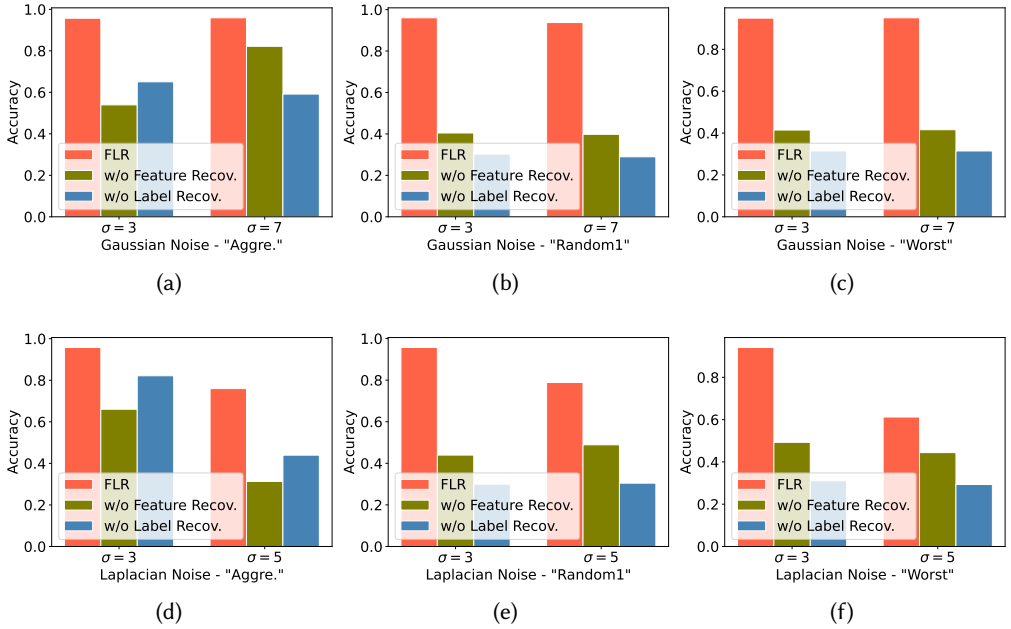
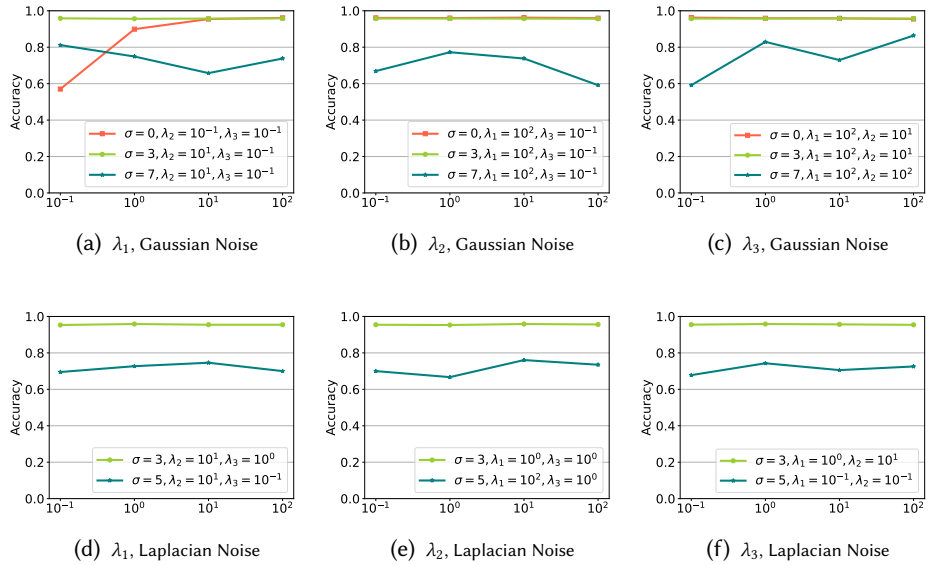


Fig. 4. Ablation of FLR on "Aggre.", "Random1", and "Worst" datasets.

unsatisfactorily. The relatively high classification accuracy of the FLR method indicates that the risk error on the test set is low and bounded. Consequently, the existing methods cannot handle hybrid noise well, yet FLR is indeed an effective solution to tackling the hybrid noise.



(a) λ_1, Gaussian Noise (b) λ_2, Gaussian Noise (c) λ_3, Gaussian Noise

(d) λ_1, Laplacian Noise (e) λ_2, Laplacian Noise (f) λ_3, Laplacian Noise

Fig. 5. Parametric sensitivity of FLR on "Aggre.".

5.4 Ablation Study

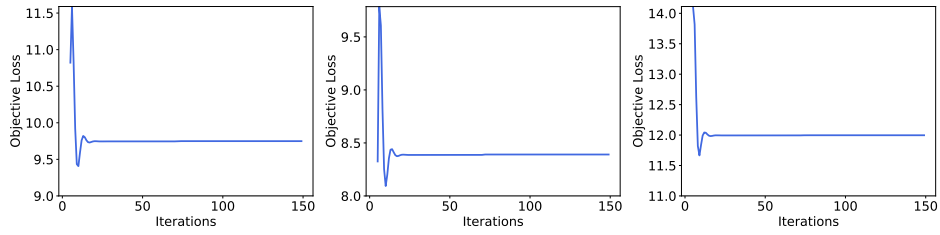
This section investigates the usefulness of each key component in FLR. We study the performances of three different settings on the CIFAR-10N "Aggre." dataset. First, both the two terms $\boldsymbol{E}_f$ and $\boldsymbol{E}_l$ are reserved to constitute the original model (abbreviated as "FLR"); second, the term $\boldsymbol{E}_f$ is removed, which means that the feature recovery process is ignored (abbreviated as "w/o feature recov."); third, we remove the term $\boldsymbol{E}_l$, so that the label recovery is abandoned (abbreviated as "w/o label recov.").

The experimental results of these models with different feature noise are illustrated in Figure 4. The results clearly reveal that the regular setting (namely, FLR) achieves the best performance on the "Aggre." dataset. By contrast, the accuracy will drop a lot without any of the two terms (*i.e.*, term $\boldsymbol{E}_f$ and term $\boldsymbol{E}_l$). Therefore, these two noise capture terms are essential to boost the performance of FLR.

5.5 Parametric Sensitivity

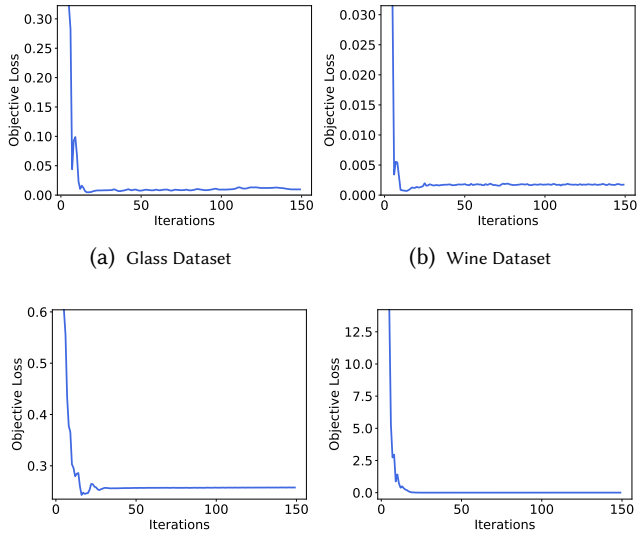
Note that Eq. (3) contains three trade-off parameters λ_1, λ_2, and λ_3 that need to be manually tuned. Therefore, we discuss whether the choices of them will significantly influence the performance of FLR. To this end, we examine the classification accuracy via changing one of λ_1, λ_2, and λ_3, and meanwhile fixing the others to the optimal constant values on "Aggre.". All experimental results are shown in Figure 5.

We can observe that the performance of FLR is relatively stable in various noise scenarios. Therefore, the performance of our proposed FLR is actually insensitive to the choice of hyperparameters, and thus the parameters of our method can be easily tuned in practical uses.



(a) "Aggre." (b) "Random1" (c) "Worst"

Fig. 6. The illustration of convergence process of the ADMM method adopted by FLR on "Aggre.", "Random1", and "Worst" datasets.



(a) Glass Dataset (b) Wine Dataset

(c) CNAE9 Dataset (d) Waveform Dataset

Fig. 7. The illustration of the convergence process of ADMM method adopted by FLR on four UCI benchmark datasets.

5.6 Illustration of Convergence

Note that we have theoretically proved that the optimization process in Algorithm 1 will converge to a stationary point. Here, we provide the convergence curves of FLR on the CIFAR-10N "Aggre.", "Random1" and "Worst" datasets in Figure 6. Furthermore, we present the convergence process on four UCI benchmark datasets in Figure 7. The curves justify our previous theoretical results and demonstrate that non-convex ADMM is effective (and also efficient) in solving the proposed model, where the loss function can successfully converge to a stable point after sufficient iterations.

6 Conclusion

In this paper, we proposed a problem setting of hybrid noise, where we aim to learn a classifier from corrupted examples with noisy features and labels. The advantages of this paper lie in three

aspects: 1) We built a novel unified robust learning paradigm to tackle the hybrid noise, where we jointly formulated the hybrid noise removal and classifier learning to a general data recovery objective; 2) We devised the optimization algorithm of non-convex ADMM for solving the proposed FLR, which is theoretically guaranteed to converge; 3) We also proved that the generalization error of FLR can be upper bounded, and it gradually shrinks even in the presence of the hybrid noise. Due to the aforementioned advantages of FLR, experimental results on several typical datasets demonstrated that FLR obtains higher classification accuracies than the existing feature noise learning and label noise learning methods in most cases. The learned classifier of FLR is limited by the fundamental assumption of linearity, which restricts its ability to effectively model and capture intricate, non-linear relationships inherent in many real-world datasets. Therefore, we intend to leverage the robust representational capabilities of deep learning to effectively manage hybrid noise within the aforementioned framework. In summary, this proposed FLR provides a heuristic framework to address hybrid noise by concurrently learning from both feature noise and label noise, which encourages us to address complex problems in a collaborative manner.

Acknowledgments

We would like to express our sincere gratitude to Professor Yuhong Guo at Carleton University, Canada, for her invaluable guidance and support throughout this research. Her expertise and insights have greatly contributed to the success of this work.

References

[1] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. 2011. Robust principal component analysis? *J. ACM* 58, 3 (2011), 1–37.

[2] Jinhui Chen and Jian Yang. 2013. Robust subspace segmentation via low-rank representation. *IEEE Transactions on Cybernetics* 44, 8 (2013), 1432–1445.

[3] Shuo Chen, Jian Yang, Lei Luo, Yang Wei, Kaihua Zhang, and Ying Tai. 2017. Low-rank latent pattern approximation with applications to robust image classification. *IEEE Transactions on Image Processing* 26, 11 (2017), 5519–5530.

[4] Shuo Chen, Jian Yang, Yang Wei, Lei Luo, Gui-Fu Lu, and Chen Gong. 2019. δ-norm-based robust regression with applications to image analysis. *IEEE Transactions on Cybernetics* 51, 6 (2019), 3371–3383.

[5] Kaiyang Chiang, Chojui Hsieh, and Inderjit S Dhillon. 2015. Matrix completion with noisy side information. In *Advances in Neural Information Processing Systems*. 3447–3455.

[6] Jiazhi Du, Xin Qiao, Zifei Yan, Hongzhi Zhang, and Wangmeng Zuo. 2024. Flexible image denoising model with multi-layer conditional feature modulation. *Pattern Recognition* (2024), 110372.

[7] Junbin Gao. 2008. Robust L1 principal component analysis and its Bayesian variational inference. *Neural Computation* 20, 2 (2008), 555–572.

[8] Wei Gao, Lu Wang, Zhi-Hua Zhou, et al. 2016. Risk minimization in the presence of label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[9] Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. 2017. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[10] Chen Gong, Hong Shi, Tongliang Liu, Chuang Zhang, Jian Yang, and Dacheng Tao. 2019. Loss decomposition and centroid estimation for positive and unlabeled learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 3 (2019), 918–932.

[11] Chen Gong, Hengmin Zhang, Jian Yang, and Dacheng Tao. 2017. Learning with inadequate and incorrect supervision. In *2017 IEEE International Conference on Data Mining*. 889–894.

[12] Frank E Grubbs. 1973. Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. *Technometrics* 15, 1 (1973), 53–66.

[13] Jipeng Guo, Yanfeng Sun, Junbin Gao, Yongli Hu, and Baocai Yin. 2021. Rank consistency induced multiview subspace clustering via low-rank matrix factorization. *IEEE Transactions on Neural Networks and Learning Systems* 33, 7 (2021), 3157–3170.

[14] Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama. 2018. Masking: A new perspective of noisy supervision. In *Advances in Neural Information Processing Systems*, Vol. 31.

[15] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information*

Processing Systems, Vol. 31.
[16] Cho-Jui Hsieh, Nagarajan Natarajan, and Inderjit Dhillon. 2015. PU learning for matrix completion. In *Proceedings of International Conference on Machine Learning*. 2445–2453.
[17] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of International Conference on Machine Learning*. 2304–2313.
[18] Jingchen Ke, Chen Gong, Tongliang Liu, Lin Zhao, Jian Yang, and Dacheng Tao. 2020. Laplacian Welsch regularization for robust semisupervised learning. *IEEE Transactions on Cybernetics* 52, 1 (2020), 164–177.
[19] Junnan Li, Richard Socher, and Steven CH Hoi. 2019. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *Proceedings of International Conference on Learning Representations*. 1–14.
[20] Junnan Li, Caiming Xiong, and Steven CH Hoi. 2021. Learning from noisy data with robust representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9485–9494.
[21] Xiuchuan Li, Xiaobo Xia, Fei Zhu, Tongliang Liu, Xuyao Zhang, and Chenglin Liu. 2023. Dynamics-aware loss for learning with label noise. *Pattern Recognition* 144 (2023), 109835.
[22] Zhouchen Lin, Minming Chen, and Yi Ma. 2010. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055* (2010).
[23] Guangcan Liu, Zhouchen Lin, and Yong Yu. 2010. Robust subspace segmentation by low-rank representation. In *Proceedings of International Conference on Machine Learning*. 663–670.
[24] Wenshui Luo, Shuo Chen, Tongliang Liu, Bo Han, Gang Niu, Masashi Sugiyama, Dacheng Tao, and Chen Gong. 2024. Estimating per-class statistics for label noise learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024), 1–17.
[25] Yijing Luo, Bo Han, and Chen Gong. 2021. A bi-level formulation for label noise learning with spectral cluster discovery. In *Proceedings of International Conference on International Joint Conferences on Artificial Intelligence*. 2605–2611.
[26] Ping Ma, Jinchang Ren, Genyun Sun, Huimin Zhao, Xiuping Jia, Yijun Yan, and Jaime Zabalza. 2023. Multiscale superpixelwise prophet model for noise-robust feature extraction in hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–12.
[27] George D. Magoulas and Andriana Prentza. 2001. *Machine learning in medical applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, 300–307.
[28] André LB Miranda, Luís Paulo F Garcia, André CPLF Carvalho, and Ana C Lorena. 2009. Use of classification algorithms in noise detection and elimination. In *Hybrid Artificial Intelligence Systems*. Springer, 417–424.
[29] Fabrice Muhlenbach, Stéphane Lallich, and Djamel A Zighed. 2004. Identifying and handling mislabelled instances. *Journal of Intelligent Information Systems* 22, 1 (2004), 89–109.
[30] Giorgio Patrini, Frank Nielsen, Richard Nock, and Marcello Carioni. 2016. Loss factorization, weakly supervised learning and label noise robustness. In *Proceedings of International Conference on Machine Learning*. 708–717.
[31] Halsey Royden and Patrick Michael Fitzpatrick. 2010. *Real analysis*. China Machine Press.
[32] Ziqiang Shi, Jiqing Han, and Tieran Zheng. 2014. Audio classification with low-rank matrix representation features. *ACM Transactions on Intelligent Systems and Technology* 5, 1 (2014), 1–17.
[33] Haoliang Sun, Chenhui Guo, Qi Wei, Zhongyi Han, and Yilong Yin. 2022. Learning to rectify for robust learning with noisy labels. *Pattern Recognition* 124 (2022), 108467.
[34] Le Sun, Zebin Wu, Jianjun Liu, and Zhihui Wei. 2014. A Novel Supervised Method for Hyperspectral Image Classification with Spectral-Spatial Constraints. *Chinese Journal of Electronics* 23, EN20140125 (2014), 135.
[35] Vasileios Tsouvalas, Aaqib Saeed, Tanir Ozcelebi, and Nirvana Meratnia. 2024. Labeling chaos to learning harmony: Federated learning with noisy labels. *ACM Transactions on Intelligent Systems and Technology* 15, 2 (2024), 1–26.
[36] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research* 11, 12 (2010).
[37] Deqing Wang and Guoqiang Hu. 2024. Efficient nonnegative tensor decomposition using alternating direction proximal method of multipliers. *Chinese Journal of Electronics* 33, 5 (2024), 1308–1316.
[38] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. 2019. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 322–330.
[39] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13726–13735.
[40] Yang Wei, Chen Gong, Shuo Chen, Tongliang Liu, Jian Yang, and Dacheng Tao. 2019. Harnessing side information for classification under label noise. *IEEE Transactions on Neural Networks and Learning Systems* 31, 9 (2019), 3178–3192.
[41] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. 2008. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 2 (2008), 210–227.

[42] Chang Xu, Dacheng Tao, and Chao Xu. 2016. Robust extreme multi-label learning. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1275–1284.
[43] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. 2019. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In *Advances in Neural Information Processing Systems*, Vol. 32.
[44] Jian Yang, Lei Luo, Jianjun Qian, Ying Tai, Fanlong Zhang, and Yong Xu. 2016. Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 1 (2016), 156–171.
[45] Zunzhi You, Daochang Liu, Bohyung Han, and Chang Xu. 2024. Beyond pretrained features: Noisy image modeling provides adversarial defense. In *Advances in Neural Information Processing Systems*, Vol. 36.
[46] Bixiao Zeng, Xiaodong Yang, Yiqiang Chen, Hanchao Yu, and Yingwei Zhang. 2022. CLC: A consensus-based label correction approach in federated learning. *ACM Transactions on Intelligent Systems and Technology* 13, 5 (2022), 1–23.
[47] Hengmin Zhang, Jian Yang, Fanhua Shang, Chen Gong, and Zhenyu Zhang. 2018. LRR for subspace segmentation via tractable schatten-p norm minimization and factorization. *IEEE Transactions on Cybernetics* 49, 5 (2018), 1722–1734.
[48] Jinghui Zhang, Dingyang Lv, Qiangsheng Dai, Fa Xin, and Fang Dong. 2023. Noise-aware local model training mechanism for federated learning. *ACM Transactions on Intelligent Systems and Technology* 14, 4 (2023), 1–22.
[49] Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, Vol. 31.
[50] Zhaowei Zhu, Jialu Wang, and Yang Liu. 2022. Beyond images: Label noise transition matrix estimation for tasks with lower-quality features. In *Proceedings of International Conference on Machine Learning*. 27633–27653.

A Supplementary Material

In this supplementary material, we provide the proofs for the theorems we proposed in the manuscript. The notations and numbering of equations follow those of the main paper.

A.1 The Proof of Convergence Property (Theorem 1)

Firstly, we provide the following lemma as a preliminary.

Lemma A.1. *[22] Let $\mathcal{H}$ be a real Hilbert space endowed with an inner product $\langle\cdot,\cdot\rangle$, and a corresponding norm be $\|\cdot\|$. Considering $y \in \partial\|x\|$, it holds that $\|y\|^* = 1$, if $x \neq 0$, and $\|y\|^* \leq 1$, if $x = 0$, where $\partial f(\cdot)$ is the sub-gradient of $f(\cdot)$, and $\|\cdot\|^*$ is the dual norm of $\|\cdot\|$.*

Next, we provide the proof of convergence property exhaustively.

Proof. (1) Here, we will prove that the generated sequences $\{\{\boldsymbol{M}_{i,t}\}_{i=1}^5\}$, $\{\boldsymbol{X}_t\}$, $\{\boldsymbol{Z}_t\}$, $\{\boldsymbol{B}_t\}$, $\{\boldsymbol{J}_t\}$, $\{\boldsymbol{K}_t\}$, $\{\boldsymbol{E}_{f,t}\}$, and $\{\boldsymbol{E}_{l,t}\}$ are all bounded.
According to the updating rule of $\boldsymbol{B}$ in Eq. (10), $\boldsymbol{B} \in [0,1]^{n\times c}$ always holds, and thus the sequence $\{\boldsymbol{B}_t\}$ is bounded.
As stated in the definition of the augmented Lagrangian function $\mathcal{L}$, *i.e.*, Eq. (3), and the first-order optimality condition of $\boldsymbol{E}_{f,t+1}$ and $\boldsymbol{E}_{l,t+1}$, we have

$$\begin{aligned}\mathbf{0} &\in \partial\mathcal{L}_{\boldsymbol{E}_f}\left(\boldsymbol{X}_{t+1}, \boldsymbol{Z}_{t+1}, \boldsymbol{B}_{t+1}, \boldsymbol{J}_{t+1}, \boldsymbol{K}_{t+1}, \boldsymbol{E}_{f,t+1}, \boldsymbol{E}_{l,t}, \{\boldsymbol{M}_{i,t}\}_{i=1}^5\right)\\ &= \partial(\lambda_2\|\boldsymbol{E}_{f,t+1}\|_1) - \boldsymbol{M}_{1,t} - \mu_t(\tilde{\boldsymbol{X}} - \boldsymbol{X}_{t+1} - \boldsymbol{E}_{f,t+1}),\end{aligned} \tag{22}$$

as

$$\boldsymbol{M}_{1,t+1} = \boldsymbol{M}_{1,t} + \mu_t(\tilde{\boldsymbol{X}} - \boldsymbol{X}_{t+1} - \boldsymbol{E}_{f,t+1}), \tag{23}$$

namely

$$\mathbf{0} \in \partial(\lambda_2\|\boldsymbol{E}_{f,t+1}\|_1) - \boldsymbol{M}_{1,t+1}. \tag{24}$$

So we have that

$$\boldsymbol{M}_{1,t+1} \in \partial(\lambda_2\|\boldsymbol{E}_{f,t+1}\|_1). \tag{25}$$

Then by Lemma A.1, it holds that

$$\|\boldsymbol{M}_{1,t+1}\|_1^* \leq 1. \tag{26}$$

Therefore, the sequence $\{\boldsymbol{M}_{1,t}\}$ is bounded. Similarly, the sequence $\{\boldsymbol{M}_{2,t}\}$ can be proved to be bounded by calculating the partial derivative of $\mathcal{L}$ *w.r.t.* $\boldsymbol{E}_{l,t+1}$.

Then by the optimality of $\boldsymbol{Z}_{t+1}$,

$$\mathbf{0} \in \partial \mathcal{L}_Z \left(\boldsymbol{X}_{t+1}, \boldsymbol{Z}_{t+1}, \boldsymbol{B}_t, \boldsymbol{J}_t, \boldsymbol{K}_t, \boldsymbol{E}_f, \boldsymbol{E}_l, \{\boldsymbol{M}_{i,t}\}_{i=1}^5\right), \tag{27}$$

namely

$$\begin{aligned} \mathbf{0} &\in \partial\left(\lambda_1\|\boldsymbol{Z}_{t+1}\|_*\right) + \boldsymbol{M}_{3,t} + \mu_t(\boldsymbol{Z}_{t+1} - \boldsymbol{J}_t) \\ &= \partial\left(\lambda_1\|\boldsymbol{Z}_{t+1}\|_*\right) + \boldsymbol{M}_{3,t} + \mu_t(\boldsymbol{Z}_{t+1} - \boldsymbol{J}_{t+1}) + \mu_t(\boldsymbol{J}_{t+1} - \boldsymbol{J}_t). \end{aligned} \tag{28}$$

As

$$\boldsymbol{M}_{3,t+1} = \boldsymbol{M}_{3,t} + \mu_t(\boldsymbol{Z}_{t+1} - \boldsymbol{J}_{t+1}), \tag{29}$$

we see that

$$\| - \boldsymbol{M}_{3,t+1} - \mu_t(\boldsymbol{J}_{t+1} - \boldsymbol{J}_t)\|_*^* \leq 1. \tag{30}$$

Combining Eq. (30) with

$$\lim_{t\to+\infty} \mu_t(\boldsymbol{J}_{t+1} - \boldsymbol{J}_t) = 0, \tag{31}$$

we conclude that the sequence $\{\boldsymbol{M}_{3,t}\}$ is bounded.
In the same way, considering

$$\begin{aligned} \mathbf{0} &\in \partial \mathcal{L}_X \left(\boldsymbol{X}_{t+1}, \boldsymbol{Z}_t, \boldsymbol{B}_t, \boldsymbol{J}_t, \boldsymbol{K}_t, \boldsymbol{E}_f, \boldsymbol{E}_l, \{\boldsymbol{M}_{i,t}\}_{i=1}^5\right) \\ &\in \partial\|\boldsymbol{X}_{t+1}\|_* - \boldsymbol{M}_{1,t} + \boldsymbol{M}_{5,t} + \mu_t\left(-\left(\tilde{\boldsymbol{X}} - \boldsymbol{X}_{t+1} - \boldsymbol{E}_{f,t}\right) + (\boldsymbol{X}_{t+1} - \boldsymbol{K}_t)\right) \\ &\in \partial\|\boldsymbol{X}_{t+1}\|_* - \boldsymbol{M}_{1,t+1} + \boldsymbol{M}_{5,t+1} - \mu_t(\boldsymbol{E}_{f,t+1} - \boldsymbol{E}_{f,t}) + \mu_t(\boldsymbol{K}_{t+1} - \boldsymbol{K}_t), \end{aligned} \tag{32}$$

we have

$$\|\boldsymbol{M}_{1,t+1} - \boldsymbol{M}_{5,t+1} + \mu_t(\boldsymbol{E}_{f,t+1} - \boldsymbol{E}_{f,t}) - \mu_t(\boldsymbol{K}_{t+1} - \boldsymbol{K}_t)\|_*^* \leq 1. \tag{33}$$

Also, we combine Eq. (33) with

$$\begin{aligned} \lim_{t\to+\infty} \mu_t(\boldsymbol{K}_{t+1} - \boldsymbol{K}_t) &= 0, \\ \lim_{t\to+\infty} \mu_t(\boldsymbol{E}_{f,t+1} - \boldsymbol{E}_{f,t}) &= 0, \end{aligned} \tag{34}$$

and the boundedness of sequence $\{\boldsymbol{M}_{1,t}\}$, the boundedness of sequence $\{\boldsymbol{M}_{5,t}\}$ can be proved.
Next, by the optimality of $\boldsymbol{B}_{t+1}$, we have

$$\mathbf{0} = -\boldsymbol{M}_{2,t} + \boldsymbol{M}_{4,t} - \mu_t\left(\tilde{\boldsymbol{Y}} - \boldsymbol{B}_{t+1} - \boldsymbol{E}_{l,t}\right) + \mu_t\left(\boldsymbol{B}_{t+1} - \boldsymbol{K}_t\boldsymbol{J}_t\right), \tag{35}$$

which leads to

$$\boldsymbol{M}_{2,t+1} = \boldsymbol{M}_{4,t+1} - \mu_t(\boldsymbol{E}_{l,t+1} - \boldsymbol{E}_{l,t}) + \mu_t(\boldsymbol{K}_{t+1}\boldsymbol{J}_{t+1} - \boldsymbol{K}_t\boldsymbol{J}_t). \tag{36}$$

Based on the assumptions of $\lim_{t\to+\infty} \mu_t(\boldsymbol{K}_{t+1} - \boldsymbol{K}_t) = 0$, and $\lim_{t\to+\infty} \mu_t(\boldsymbol{J}_{t+1} - \boldsymbol{J}_t) = 0$, we have $\lim_{t\to+\infty} \mu_t(\boldsymbol{K}_{t+1}\boldsymbol{J}_t - \boldsymbol{K}_t\boldsymbol{J}_t) = 0$. Also, we have proved the boundedness of $\boldsymbol{M}_{2,t+1}$ and assume $\lim_{t\to+\infty} \mu_t(\boldsymbol{E}_{l,t+1} - \boldsymbol{E}_{l,t}) = 0$, so it is obviously the sequence $\boldsymbol{M}_{4,t+1}$ is bounded.
Thus far, we have proved that these sequences $\{\boldsymbol{M}_{i,t}\}_{i=1}^5$ are all bounded.
In addition, we here derive the following chain of equations:

$$\begin{aligned} &\mathcal{L}\left(\boldsymbol{X}_t, \boldsymbol{Z}_t, \boldsymbol{B}_t, \boldsymbol{J}_t, \boldsymbol{K}_t, \boldsymbol{E}_f, \boldsymbol{E}_l, \{\boldsymbol{M}_{i,t}\}_{i=1}^5\right) \\ &= \|\boldsymbol{X}_t\|_* + \lambda_1\|\boldsymbol{Z}_t\|_* + \lambda_2\|\boldsymbol{E}_{f,t}\|_1 + \lambda_3\|\boldsymbol{E}_{l,t}\|_{2,1} \\ &+ \Phi(\boldsymbol{M}_{1,t}, \tilde{\boldsymbol{X}} - \boldsymbol{X}_t - \boldsymbol{E}_{f,t}) + \Phi(\boldsymbol{M}_{2,t}, \tilde{\boldsymbol{Y}} - \boldsymbol{B}_t - \boldsymbol{E}_{l,t}) \\ &+ \Phi(\boldsymbol{M}_{3,t}, \boldsymbol{Z}_t - \boldsymbol{J}_t) + \Phi(\boldsymbol{M}_{4,t}, \boldsymbol{B}_t - \boldsymbol{K}_t\boldsymbol{J}_t) + \Phi(\boldsymbol{M}_{5,t}, \boldsymbol{X}_t - \boldsymbol{K}_t), \end{aligned} \tag{37}$$

where $\Phi(\boldsymbol{A},\boldsymbol{B}) = \frac{\mu}{2}\|\boldsymbol{B}\|_F^2 + \langle \boldsymbol{A},\boldsymbol{B}\rangle$;

$$\begin{aligned}
&\mathcal{L}\left(\boldsymbol{X}_t,\boldsymbol{Z}_t,\boldsymbol{B}_t,\boldsymbol{J}_t,\boldsymbol{K}_t,\boldsymbol{E}_f,\boldsymbol{E}_l,\{\boldsymbol{M}_{i,t-1}\}_{i=1}^5\right)\\
&= \|\boldsymbol{X}_t\|_* + \lambda_1\|\boldsymbol{Z}_t\|_* + \lambda_2\|\boldsymbol{E}_{f,t}\|_1 + \lambda_3\|\boldsymbol{E}_{l,t}\|_{2,1}\\
&+ \Phi(\boldsymbol{M}_{1,t-1},\tilde{\boldsymbol{X}} - \boldsymbol{X}_t - \boldsymbol{E}_{f,t}) + \Phi(\boldsymbol{M}_{2,t-1},\tilde{\boldsymbol{Y}} - \boldsymbol{B}_t - \boldsymbol{E}_{l,t})\\
&+ \Phi(\boldsymbol{M}_{3,t-1},\boldsymbol{Z}_t - \boldsymbol{J}_t) + \Phi(\boldsymbol{M}_{4,t-1},\boldsymbol{B}_t - \boldsymbol{K}_t\boldsymbol{J}_t) + \Phi(\boldsymbol{M}_{5,t-1},\boldsymbol{X}_t - \boldsymbol{K}_t),
\end{aligned} \tag{38}$$

namely

$$\begin{aligned}
&\mathcal{L}\left(\boldsymbol{X}_t,\boldsymbol{Z}_t,\boldsymbol{B}_t,\boldsymbol{J}_t,\boldsymbol{K}_t,\boldsymbol{E}_{f,t},\boldsymbol{E}_{l,t},\{\boldsymbol{M}_{i,t}\}_{i=1}^5\right)\\
&= \mathcal{L}\left(\boldsymbol{X}_t,\boldsymbol{Z}_t,\boldsymbol{B}_t,\boldsymbol{J}_t,\boldsymbol{K}_t,\boldsymbol{E}_{f,t},\boldsymbol{E}_{l,t},\{\boldsymbol{M}_{i,t-1}\}_{i=1}^5\right)\\
&+ \langle \boldsymbol{M}_{1,t} - \boldsymbol{M}_{1,t-1},\tilde{\boldsymbol{X}} - \boldsymbol{X}_t - \boldsymbol{E}_{f,t}\rangle\\
&+ \langle \boldsymbol{M}_{2,t} - \boldsymbol{M}_{2,t-1},\tilde{\boldsymbol{Y}} - \boldsymbol{B}_t - \boldsymbol{E}_{l,t}\rangle + \langle \boldsymbol{M}_{3,t} - \boldsymbol{M}_{3,t-1},\boldsymbol{Z}_t - \boldsymbol{J}_t\rangle\\
&+ \langle \boldsymbol{M}_{4,t} - \boldsymbol{M}_{4,t-1},\boldsymbol{B}_t - \boldsymbol{K}_t\boldsymbol{J}_t\rangle + \langle \boldsymbol{M}_{5,t} - \boldsymbol{M}_{5,t-1},\boldsymbol{X}_t - \boldsymbol{K}_t\rangle\\
&+ \frac{\mu_t - \mu_{t-1}}{2}(\|\tilde{\boldsymbol{X}} - \boldsymbol{X}_t - \boldsymbol{E}_{f,t}\|_F^2 + \|\tilde{\boldsymbol{Y}} - \boldsymbol{B}_t - \boldsymbol{E}_{l,t}\|_F^2\\
&+ \|\boldsymbol{Z}_t - \boldsymbol{J}_t\|_F^2 + \|\boldsymbol{B}_t - \boldsymbol{K}_t\boldsymbol{J}_t\|_F^2 + \|\boldsymbol{X}_t - \boldsymbol{K}_t\|_F^2)\\
&= \mathcal{L}\left(\boldsymbol{X}_t,\boldsymbol{Z}_t,\boldsymbol{B}_t,\boldsymbol{J}_t,\boldsymbol{K}_t,\boldsymbol{E}_{f,t},\boldsymbol{E}_{l,t},\{\boldsymbol{M}_{i,t-1}\}_{i=1}^5\right)\\
&+ \frac{\mu_t + \mu_{t-1}}{2\mu_{t-1}^2}\sum_{i=1}^5\|\boldsymbol{M}_{i,t} - \boldsymbol{M}_{i,t-1}\|_F^2.
\end{aligned} \tag{39}$$

Since Algorithm 1 optimizes variables $\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{B}$, $\boldsymbol{J}$, $\boldsymbol{K}$, $\boldsymbol{E}_f$ and $\boldsymbol{E}_l$ alternatively, according to the optimality of them and Eq. (39), we have

$$\begin{aligned}
&\mathcal{L}\left(\boldsymbol{X}_{t+1},\boldsymbol{Z}_{t+1},\boldsymbol{B}_{t+1},\boldsymbol{J}_{t+1},\boldsymbol{K}_{t+1},\boldsymbol{E}_{f,t+1},\boldsymbol{E}_{l,t+1},\{\boldsymbol{M}_{i,t}\}_{i=1}^5\right)\\
&\le \mathcal{L}\left(\boldsymbol{X}_t,\boldsymbol{Z}_t,\boldsymbol{B}_t,\boldsymbol{J}_t,\boldsymbol{K}_t,\boldsymbol{E}_{f,t},\boldsymbol{E}_{l,t},\{\boldsymbol{M}_{i,t}\}_{i=1}^5\right)\\
&= \mathcal{L}\left(\boldsymbol{X}_t,\boldsymbol{Z}_t,\boldsymbol{B}_t,\boldsymbol{J}_t,\boldsymbol{K}_t,\boldsymbol{E}_{f,t},\boldsymbol{E}_{l,t},\{\boldsymbol{M}_{i,t-1}\}_{i=1}^5\right)\\
&+ \frac{\mu_t + \mu_{t-1}}{2\mu_{t-1}^2}\sum_{i=1}^5\|\boldsymbol{M}_{i,t} - \boldsymbol{M}_{i,t-1}\|_F^2,
\end{aligned} \tag{40}$$

by setting $t = n$, and taking the relationship in Eq. (40) into consideration, it holds that

$$\begin{aligned}
&\mathcal{L}\left(\boldsymbol{X}_{n+1},\boldsymbol{Z}_{n+1},\boldsymbol{B}_{n+1},\boldsymbol{J}_{n+1},\boldsymbol{K}_{n+1},\boldsymbol{E}_{f,n+1},\boldsymbol{E}_{l,n+1},\{\boldsymbol{M}_{i,n}\}_{i=1}^5\right)\\
&\le \mathcal{L}\left(\boldsymbol{X}_1,\boldsymbol{Z}_1,\boldsymbol{B}_1,\boldsymbol{J}_1,\boldsymbol{K}_1,\boldsymbol{E}_{f,1},\boldsymbol{E}_{l,1},\{\boldsymbol{M}_{i,0}\}_{i=1}^5\right)\\
&+ \sum_{t=1}^n\left(\frac{\mu_t + \mu_{t-1}}{2\mu_{t-1}^2}\sum_{i=1}^5\|\boldsymbol{M}_{i,t} - \boldsymbol{M}_{i,t-1}\|_F^2\right).
\end{aligned} \tag{41}$$

Note that we have proved the sequences $\{\boldsymbol{M}_{i,t}\}_{i=1}^5$ are bounded, so there exists one constant C, and then $\sum_{i=1}^5\|\boldsymbol{M}_{i,t} - \boldsymbol{M}_{i,t-1}\|_F^2 \le C$.

Therefore,

$$\begin{aligned}
&\sum_{t=1}^{n}\left(\frac{\mu_t+\mu_{t-1}}{2\mu_{t-1}^2}\sum_{i=1}^{5}\|\boldsymbol{M}_{i,t}-\boldsymbol{M}_{i,t-1}\|_F^2\right)\\
&\leq \sum_{t=1}^{n}\left(\frac{\mu_t+\mu_{t-1}}{2\mu_{t-1}^2}C\right)\\
&= C\sum_{t=1}^{n}\frac{\mu_0\rho^t+\mu_0\rho^{t-1}}{2(\mu_0\rho^{t-1})^2}\\
&= C\frac{\rho+1}{2\mu_0}\sum_{t=1}^{n}\frac{1}{\rho^{t-1}} \leq C\frac{(\rho+1)\rho}{2\mu_0(\rho-1)}(1-\frac{1}{\rho^n}) < +\infty,
\end{aligned} \tag{42}$$

in which ρ^t denotes the exponential power calculation and it holds that $\mu_t = \mu_0\rho^t$.
Now we can see that both sides of Eq. (41) are bounded. Based on this conclusion and the boundedness of $\{\boldsymbol{M}_{i,t}\}_{i=1}^5$, each term of the following equation is bounded, namely

$$\begin{aligned}
&\mathcal{L}\left(\boldsymbol{X}_{t+1},\boldsymbol{Z}_{t+1},\boldsymbol{B}_{t+1},\boldsymbol{J}_{t+1},\boldsymbol{K}_{t+1},\boldsymbol{E}_{f,t+1},\boldsymbol{E}_{l,t+1},\{\boldsymbol{M}_{i,t}\}_{i=1}^5\right)+\frac{1}{2\mu_t}\sum_{i=1}^{5}\|\boldsymbol{M}_{i,t}\|_F^2\\
&= \|\boldsymbol{X}_{t+1}\|_* + \lambda_1\|\boldsymbol{Z}_{t+1}\|_* + \lambda_2\|\boldsymbol{E}_{f,t+1}\|_1 + \lambda_3\|\boldsymbol{E}_{l,t+1}\|_{2,1}\\
&+ \Phi(\boldsymbol{M}_{1,t},\tilde{\boldsymbol{X}}-\boldsymbol{X}_t-\boldsymbol{E}_{f,t})+\frac{1}{2\mu_t}\|\boldsymbol{M}_{1,t}\|_F^2\\
&+ \Phi(\boldsymbol{M}_{2,t},\tilde{\boldsymbol{Y}}-\boldsymbol{B}_t-\boldsymbol{E}_{l,t})+\frac{1}{2\mu_t}\|\boldsymbol{M}_{2,t}\|_F^2\\
&+ \Phi(\boldsymbol{M}_{3,t},\boldsymbol{Z}_t-\boldsymbol{J}_t)+\frac{1}{2\mu_t}\|\boldsymbol{M}_{3,t}\|_F^2\\
&+ \Phi(\boldsymbol{M}_{4,t},\boldsymbol{B}_t-\boldsymbol{K}_t\boldsymbol{J}_t)+\frac{1}{2\mu_t}\|\boldsymbol{M}_{4,t}\|_F^2\\
&+ \Phi(\boldsymbol{M}_{5,t},\boldsymbol{X}_t-\boldsymbol{K}_t)+\frac{1}{2\mu_t}\|\boldsymbol{M}_{5,t}\|_F^2.
\end{aligned} \tag{43}$$

Then,

$$\begin{aligned}
&\mathcal{L}\left(\boldsymbol{X}_{t+1},\boldsymbol{Z}_{t+1},\boldsymbol{B}_{t+1},\boldsymbol{J}_{t+1},\boldsymbol{K}_{t+1},\boldsymbol{E}_{f,t+1},\boldsymbol{E}_{l,t+1},\{\boldsymbol{M}_{i,t}\}_{i=1}^5\right)+\frac{1}{2\mu_t}\sum_{i=1}^{5}\|\boldsymbol{M}_{i,t}\|_F^2\\
&= \|\boldsymbol{X}_{t+1}\|_* + \lambda_1\|\boldsymbol{Z}_{t+1}\|_* + \lambda_2\|\boldsymbol{E}_{f,t+1}\|_1 + \lambda_3\|\boldsymbol{E}_{l,t+1}\|_{2,1}\\
&+ \frac{\mu_t}{2}\left(\|\tilde{\boldsymbol{X}}-\boldsymbol{X}_{t+1}-\boldsymbol{E}_{f,t+1}+\frac{\boldsymbol{M}_{1,t}}{\mu_t}\|_F^2\right)+\frac{\mu_t}{2}\left(\|\tilde{\boldsymbol{Y}}-\boldsymbol{B}_{t+1}-\boldsymbol{E}_{l,t+1}+\frac{\boldsymbol{M}_{2,t}}{\mu_t}\|_F^2\right)\\
&+ \frac{\mu_t}{2}\left(\|\boldsymbol{Z}_{t+1}-\boldsymbol{J}_{t+1}+\frac{\boldsymbol{M}_{3,t}}{\mu_t}\|_F^2\right)+\frac{\mu_t}{2}\left(\|\boldsymbol{B}_{t+1}-\boldsymbol{K}_{t+1}\boldsymbol{J}_{t+1}+\frac{\boldsymbol{M}_{4,t}}{\mu_t}\|_F^2\right)\\
&+ \frac{\mu_t}{2}\left(\|\boldsymbol{X}_{t+1}-\boldsymbol{K}_{t+1}+\frac{\boldsymbol{M}_{5,t}}{\mu_t}\|_F^2\right).
\end{aligned} \tag{44}$$

Specifically, it is obvious that $\{\boldsymbol{X}_t\}$, $\{\boldsymbol{Z}_t\}$, $\{\boldsymbol{E}_{f,t}\}$, and $\{\boldsymbol{E}_{l,t}\}$ are bounded. From the ninth term of Eq. (43), we can obtain the boundedness of $\boldsymbol{K}_t$. Moreover, from the seventh term of Eq. (43), $\boldsymbol{J}_t$ is bounded.
Up to now, we have proved that the sequence $\{\Gamma_t = (\boldsymbol{X}_t,\boldsymbol{Z}_t,\boldsymbol{B}_t,\boldsymbol{J}_t,\boldsymbol{K}_t,\boldsymbol{E}_{l,t},\boldsymbol{E}_{f,t},\{\boldsymbol{M}_{i,t}\}_{i=1}^5)\}_{t=1}^{\infty}$ is bounded.

(2) According to the analysis mentioned above, and by the Bolzano-Weierstrass theorem [31], the sequence $\{\Gamma_t\}$ must have at least one accumulation point. Then $\{\Gamma^* = (\boldsymbol{X}^*, \boldsymbol{Z}^*, \boldsymbol{B}^*, \boldsymbol{J}^*, \boldsymbol{K}^*, \boldsymbol{E}_l^*, \boldsymbol{E}_f^*, \{\boldsymbol{M}_i^*\}_{i=1}^5)\}$ denotes the accumulation point. Without loss of generality, we assume that the sequence Γ_t converges to Γ^*.
As

$$\begin{aligned}
\boldsymbol{M}_{1,t+1} &= \boldsymbol{M}_{1,t} + \mu_t \left(\tilde{\boldsymbol{X}} - \boldsymbol{X}_{t+1} - \boldsymbol{E}_{f,t+1}\right), \\
\boldsymbol{M}_{2,t+1} &= \boldsymbol{M}_{2,t} + \mu_t \left(\tilde{\boldsymbol{Y}} - \boldsymbol{B}_{t+1} - \boldsymbol{E}_{l,t+1}\right), \\
\boldsymbol{M}_{3,t+1} &= \boldsymbol{M}_{3,t} + \mu_t \left(\boldsymbol{Z}_{t+1} - \boldsymbol{J}_{t+1}\right), \\
\boldsymbol{M}_{4,t+1} &= \boldsymbol{M}_{4,t} + \mu_t \left(\boldsymbol{B}_{t+1} - \boldsymbol{K}_{t+1}\boldsymbol{J}_{t+1}\right), \\
\boldsymbol{M}_{5,t+1} &= \boldsymbol{M}_{5,t} + \mu_t \left(\boldsymbol{X}_{t+1} - \boldsymbol{K}_{t+1}\right),
\end{aligned} \tag{45}$$

we have

$$\begin{aligned}
\tilde{\boldsymbol{X}} - \boldsymbol{X}_{t+1} - \boldsymbol{E}_{f,t+1} &= \frac{1}{\mu_t}\left(\boldsymbol{M}_{1,t+1} - \boldsymbol{M}_{1,t}\right), \\
\tilde{\boldsymbol{Y}} - \boldsymbol{B}_{t+1} - \boldsymbol{E}_{l,t+1} &= \frac{1}{\mu_t}\left(\boldsymbol{M}_{2,t+1} - \boldsymbol{M}_{2,t}\right), \\
\boldsymbol{Z}_{t+1} - \boldsymbol{J}_{t+1} &= \frac{1}{\mu_t}\left(\boldsymbol{M}_{3,t+1} - \boldsymbol{M}_{3,t}\right), \\
\boldsymbol{B}_{t+1} - \boldsymbol{K}_{t+1}\boldsymbol{J}_{t+1} &= \frac{1}{\mu_t}\left(\boldsymbol{M}_{4,t+1} - \boldsymbol{M}_{4,t}\right), \\
\boldsymbol{X}_{t+1} - \boldsymbol{K}_{t+1} &= \frac{1}{\mu_t}\left(\boldsymbol{M}_{5,t+1} - \boldsymbol{M}_{5,t}\right).
\end{aligned} \tag{46}$$

Because of the boundedness of $\{\boldsymbol{M}_{i,t}\}_{i=1}^5$ and when $t \to +\infty$, it holds that

$$\begin{aligned}
&\tilde{\boldsymbol{X}} - \boldsymbol{X}_{t+1} - \boldsymbol{E}_{f,t+1} \to 0, \\
&\tilde{\boldsymbol{Y}} - \boldsymbol{B}_{t+1} - \boldsymbol{E}_{l,t+1} \to 0, \\
&\boldsymbol{Z}_{t+1} - \boldsymbol{J}_{t+1} \to 0, \\
&\boldsymbol{B}_{t+1} - \boldsymbol{K}_{t+1}\boldsymbol{J}_{t+1} \to 0, \\
&\boldsymbol{X}_{t+1} - \boldsymbol{K}_{t+1} \to 0,
\end{aligned} \tag{47}$$

and then we have

$$\begin{aligned}
&\tilde{\boldsymbol{X}} - \boldsymbol{X}^* - \boldsymbol{E}_f^* \to 0, \\
&\tilde{\boldsymbol{Y}} - \boldsymbol{B}^* - \boldsymbol{E}_l^* \to 0, \\
&\boldsymbol{Z}^* - \boldsymbol{J}^* \to 0, \\
&\boldsymbol{B}^* - \boldsymbol{K}^*\boldsymbol{J}^* \to 0, \\
&\boldsymbol{X}^* - \boldsymbol{K}^* \to 0.
\end{aligned} \tag{48}$$

In addition, based on the updating rule of $\boldsymbol{B}$, $\boldsymbol{B} \in [0, 1]$ always holds, and thus $\boldsymbol{B}^* \in [0, 1]$. Hence, the primal feasibility conditions of the optimization problem are satisfied by Γ^*. Moreover, we will prove that the first-order stationary condition holds when $\Gamma = \Gamma^*$. According to the updating rules for $\boldsymbol{X}_{t+1}$, $\boldsymbol{M}_{1,t+1}$ and $\boldsymbol{M}_{5,t+1}$, *i.e.*, Eq. (32), we have

$$\begin{aligned}
0 \in \partial\|\boldsymbol{X}_{t+1}\|_* &- \boldsymbol{M}_{1,t+1} + \boldsymbol{M}_{5,t+1} \\
&- \mu_t(\boldsymbol{E}_{f,t+1} - \boldsymbol{E}_{f,t}) + \mu_t(\boldsymbol{K}_{t+1} - \boldsymbol{K}_t).
\end{aligned} \tag{49}$$

Under the assumptions that

$$\begin{aligned}\lim_{t\to+\infty}\mu_t(\boldsymbol{K}_{t+1}-\boldsymbol{K}_t)=0,\\ \lim_{t\to+\infty}\mu_t(\boldsymbol{E}_{f,t+1}-\boldsymbol{E}_{f,t})=0,\end{aligned}\tag{50}$$

and let $t\to+\infty$, we have

$$\mathbf{0}\in\partial\|\boldsymbol{X}^*\|_*-\boldsymbol{M}_1^*+\boldsymbol{M}_5^*,\tag{51}$$

which leads to

$$\partial\mathcal{L}_X\left(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{B},\boldsymbol{J},\boldsymbol{K},\boldsymbol{E}_f,\boldsymbol{E}_l,\boldsymbol{M}_{i=1}^5\right)|_{(\boldsymbol{X}=\boldsymbol{X}^*)}=0.\tag{52}$$

By the rule of $\boldsymbol{Z}_{t+1}$, *i.e.*, Eq. (27) and $\lim_{t\to+\infty}\mu_t(\boldsymbol{J}_{t+1}-\boldsymbol{J}_t)=0$, we have

$$\begin{aligned}&\mathbf{0}\in\partial\left(\lambda_1\|\boldsymbol{Z}_{t+1}\|_*\right)+\boldsymbol{M}_{3,t+1}+\mu_t(\boldsymbol{J}_{t+1}-\boldsymbol{J}_t),\\ \Rightarrow&\mathbf{0}\in\partial\left(\lambda_1\|\boldsymbol{Z}^*\|_*\right)+\boldsymbol{M}_3^*.\end{aligned}\tag{53}$$

When $t\to+\infty$, we have

$$\partial\mathcal{L}_Z\left(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{B},\boldsymbol{J},\boldsymbol{K},\boldsymbol{E}_f,\boldsymbol{E}_l,\boldsymbol{M}_{i=1}^5\right)|_{(\boldsymbol{Z}=\boldsymbol{Z}^*)}=0.\tag{54}$$

By the updating of $\boldsymbol{B}_{t+1}$, we can see that

$$\mathbf{0}=-\boldsymbol{M}_{2,t+1}+\boldsymbol{M}_{4,t+1}-\mu_t(\boldsymbol{E}_{l,t+1}-\boldsymbol{E}_{l,t})+\mu_t(\boldsymbol{K}_{t+1}\boldsymbol{J}_{t+1}-\boldsymbol{K}_t\boldsymbol{J}_t).\tag{55}$$

With the assumption $\lim_{t\to+\infty}\mu_t(\boldsymbol{E}_{l,t+1}-\boldsymbol{E}_{l,t})=0$ and the inference $\lim_{t\to+\infty}\mu_t(\boldsymbol{K}_{t+1}\boldsymbol{J}_{t+1}-\boldsymbol{K}_t\boldsymbol{J}_t)=0$, let $t\to+\infty$, and then

$$\mathbf{0}=-\boldsymbol{M}_2^*+\boldsymbol{M}_4^*,\tag{56}$$

which leads to

$$\partial\mathcal{L}_B\left(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{B},\boldsymbol{J},\boldsymbol{K},\boldsymbol{E}_f,\boldsymbol{E}_l,\boldsymbol{M}_{i=1}^5\right)|_{(\boldsymbol{B}=\boldsymbol{B}^*)}=0.\tag{57}$$

By the updating of $\boldsymbol{J}_{t+1}$, we have

$$\begin{aligned}\mathbf{0}&=-\boldsymbol{M}_{3,t+1}-\boldsymbol{K}_t^\top\left(\boldsymbol{M}_{4,t+1}+\mu_t\left(\boldsymbol{K}_{t+1}-\boldsymbol{K}_t\right)\boldsymbol{J}_{t+1}\right)\\ &=\left(\boldsymbol{K}_{t+1}^\top-\boldsymbol{K}_t^\top\right)\boldsymbol{M}_{4,t+1}+\boldsymbol{K}_t^\top\mu_t\left(\boldsymbol{K}_{t+1}-\boldsymbol{K}_t\right)\boldsymbol{J}_{t+1}\\ &\quad+\boldsymbol{M}_{3,t+1}-\boldsymbol{K}_{t+1}^\top\boldsymbol{M}_{4,t+1},\end{aligned}\tag{58}$$

with $t\to+\infty$ and $\lim_{t\to+\infty}\mu_t(\boldsymbol{K}_{t+1}-\boldsymbol{K}_t)=0$ and its inference $\lim_{t\to+\infty}\boldsymbol{K}_{t+1}-\boldsymbol{K}_t=0$, it concludes that

$$\mathbf{0}=\boldsymbol{M}_3^*-\boldsymbol{K}^{*\top}\boldsymbol{M}_4^*,\tag{59}$$

which leads to

$$\partial\mathcal{L}_J\left(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{B},\boldsymbol{J},\boldsymbol{K},\boldsymbol{E}_f,\boldsymbol{E}_l,\boldsymbol{M}_{i=1}^5\right)|_{(\boldsymbol{J}=\boldsymbol{J}^*)}=0.\tag{60}$$

By the updating of $\boldsymbol{K}_{t+1}$,

$$\mathbf{0}=\boldsymbol{M}_{4,t+1}\boldsymbol{J}_{t+1}^\top+\boldsymbol{M}_{5,t+1},\tag{61}$$

and then let $t\to+\infty$, we have

$$\mathbf{0}=\boldsymbol{M}_4^*\boldsymbol{J}^{*\top}+\boldsymbol{M}_5^*,\tag{62}$$

and then

$$\partial\mathcal{L}_K\left(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{B},\boldsymbol{J},\boldsymbol{K},\boldsymbol{E}_f,\boldsymbol{E}_l,\boldsymbol{M}_{i=1}^5\right)|_{(\boldsymbol{K}=\boldsymbol{K}^*)}=0.\tag{63}$$

By the updating of $\boldsymbol{E}_{f,t+1}$,

$$\mathbf{0}\in\partial(\lambda_2\|\boldsymbol{E}_{f,t+1}\|_1)-\boldsymbol{M}_{1,t+1},\tag{64}$$

and let $t\to+\infty$, we have

$$\mathbf{0} \in \partial(\lambda_2 \|\boldsymbol{E}_f^*\|_1) - \boldsymbol{M}_1^*, \tag{65}$$

namely

$$\partial \mathcal{L}_{E_f} \left(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{B}, \boldsymbol{J}, \boldsymbol{K}, \boldsymbol{E}_f, \boldsymbol{E}_l, \boldsymbol{M}_{i=1}^5\right) |_{(\boldsymbol{E}_f = \boldsymbol{E}_f^*)} = 0. \tag{66}$$

Similarly, it holds that

$$\partial \mathcal{L}_{E_l} \left(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{B}, \boldsymbol{J}, \boldsymbol{K}, \boldsymbol{E}_f, \boldsymbol{E}_l, \boldsymbol{M}_{i=1}^5\right) |_{(\boldsymbol{E}_l = \boldsymbol{E}_l^*)} = 0. \tag{67}$$

Up to now, we have proved the accumulation point $\{\Gamma^* = (\boldsymbol{X}^*, \boldsymbol{Z}^*, \boldsymbol{B}^*, \boldsymbol{J}^*, \boldsymbol{K}^*, \boldsymbol{E}_l^*, \boldsymbol{E}_f^*, \{\boldsymbol{M}_i^*\}_{i=1}^5)\}$ satisfies the first-order KKT conditions of $\mathcal{L}$, and then we conclude that $\boldsymbol{X}^*$,$\boldsymbol{Z}^*$,$\boldsymbol{B}^*$,$\boldsymbol{J}^*$,$\boldsymbol{K}^*$,$\boldsymbol{E}_l^*$,$\boldsymbol{E}_f^*$ is a stationary point.

□

A.2 The Proof of Generalization Bound (Theorem 2)

Firstly, we present some basic preliminaries.

Lemma A.2. *[40] Let $\mathcal{W} = \{\boldsymbol{W} : \|\boldsymbol{W}\|_{2,1} \leq \mathcal{W}_{2,1}\}$ and $\mathcal{A} = \{\boldsymbol{A} \in \mathbb{R}^{n\times c} : \|\boldsymbol{A}\|_{2,\infty} \leq \mathcal{A}_{2,\infty}\}$, and then the empirical Rademacher complexity of the function class with $F(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{W}\|_{2,q}^2$ for $q = \frac{\ln(c)}{\ln(c)-1}$ is bounded as*

$$\mathbb{E}_\sigma[\sup_{f\in\mathcal{F}} \frac{1}{n_r} \sum_{\alpha=1}^{n_r} \sigma_\alpha \mathrm{tr}\left(\boldsymbol{W}^\top \boldsymbol{A}^{(\alpha)}\right)] \leq \mathcal{W}_{2,1}\mathcal{A}_{2,\infty}\sqrt{\frac{3\ln(c)}{n_r}}, \tag{68}$$

with the fact that the dual norm of $\ell_{2,1}$ is $\ell_{2,\infty}$.

Lemma A.3. *[40] Let $\mathcal{W} = \{\boldsymbol{W} : \|\boldsymbol{W}\|_F \leq \mathcal{W}_F\}$ and $\mathcal{A} = \{\boldsymbol{A} \in \mathbb{R}^{n\times c} : \|\boldsymbol{A}\|_F \leq \mathcal{A}_F\}$, and then the empirical Rademacher complexity of the function class with $F(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{W}\|_{2,2}^2$ is bounded as*

$$\mathbb{E}_\sigma[\sup_{f\in\mathcal{F}} \frac{1}{n_r} \sum_{\alpha=1}^{n_r} \sigma_\alpha \mathrm{tr}\left(\boldsymbol{W}^\top \boldsymbol{A}^{(\alpha)}\right)] \leq \mathcal{W}_F\mathcal{A}_F\sqrt{\frac{2}{n_r}}, \tag{69}$$

with the fact that the dual norm of Frobenius norm is Frobenius norm.

Lemma A.4. *[5] Let $\mathcal{W} = \{\boldsymbol{W} : \|\boldsymbol{W}\|_* \leq \mathcal{W}_*\}$ and $\mathcal{A} = \{\boldsymbol{A} \in \mathbb{R}^{n\times c} : \|\boldsymbol{A}\|_2 \leq \mathcal{A}_2\}$, and then the empirical Rademacher complexity of the function class with $F(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{W}\|_*^2$ is bounded as*

$$\mathbb{E}_\sigma[\sup_{f\in\mathcal{F}} \frac{1}{n_r} \sum_{\alpha=1}^{n_r} \sigma_\alpha \mathrm{tr}\left(\boldsymbol{W}^\top \boldsymbol{A}^{(\alpha)}\right)] \leq \mathcal{W}_*\mathcal{A}_2\sqrt{\frac{\ln(2n_c)}{n_r}}, \tag{70}$$

with the fact that the dual norm of the nuclear norm is spectral norm and $n_c = \max(n, c)$.

Here we give the proof of the Rademacher generalization bound theorem:

Proof. The Rademacher complexity of our model can be written as

$$\mathcal{R}(\mathcal{F}) := \mathbb{E}_\sigma\left[\sup_{\theta\in\Theta} \frac{1}{nc} \sum_{\alpha=1}^{nc} \sigma_\alpha \left(\boldsymbol{X}_{i_\alpha} \boldsymbol{Z} \boldsymbol{I}^{j_\alpha} + \boldsymbol{E}_{l,i_\alpha,j_\alpha}\right)\right]. \tag{71}$$

Since $\boldsymbol{E}_l$ is independent of $\boldsymbol{Z}$ and $\boldsymbol{X}$, the Rademacher complexity above can be rewritten as

$$
\begin{aligned}
\mathcal{R}(\mathcal{F}) &:= \mathbb{E}_\sigma\left[\sup_{X,Z\in\Theta_{X,Z}}\frac{1}{nc}\sum_{\alpha=1}^{nc}\sigma_\alpha X_{i_\alpha}ZI^{j_\alpha}\right]+\mathbb{E}_\sigma\left[\sup_{E\in\Theta_E}\frac{1}{nc}\sum_{\alpha=1}^{nc}\sigma_\alpha E_{l,i_\alpha,j_\alpha}\right]\\
&=\mathbb{E}_\sigma\left[\sup_{X,Z\in\Theta_{X,Z}}\frac{1}{nc}\sum_{\alpha=1}^{nc}\sigma_\alpha\mathrm{tr}\left(Z^\top X^\top I^{i_\alpha}I^{j_\alpha\top}\right)\right]\\
&+\mathbb{E}_\sigma\left[\sup_{E\in\Theta_E}\frac{1}{nc}\sum_{\alpha=1}^{nc}\sigma_\alpha\mathrm{tr}\left(E_l^\top I^{i_\alpha}I^{j_\alpha\top}\right)\right]\\
&=\mathbb{E}_\sigma\left[\sup_{X,Z\in\Theta_{X,Z}}\frac{1}{nc}\sum_{\alpha=1}^{nc}\sigma_\alpha\langle XZ,I^{i_\alpha}I^{j_\alpha\top}\rangle\right]\\
&+\mathbb{E}_\sigma\left[\sup_{E\in\Theta_E}\frac{1}{nc}\sum_{\alpha=1}^{nc}\sigma_\alpha\langle E_l,I^{i_\alpha}I^{j_\alpha\top}\rangle\right],
\end{aligned}
\tag{72}
$$

where $\Theta_E=\{E|\|E_l\|_{2,1}\le\mathcal{E}_{l,21}\}$, and $\Theta_{X,Z}=\{(X,Z)|\|Z\|_*\le\mathcal{Z}_*,\|X\|_*\le\mathcal{X}_*,\|\tilde{X}-X\|_1\le\mathcal{E}_{f,1},XZ\in[0,1]^{n\times c}\}$.

As $X=\tilde{X}-E_f$, and thus

$$
\begin{aligned}
\mathcal{R}(\mathcal{F})\le\min\Bigg\{&\mathbb{E}_\sigma\left[\sup_{X_1,Z\in\Theta_{X_1,Z}}\frac{1}{nc}\sum_{\alpha=1}^{nc}\sigma_\alpha\langle XZ,I^{i_\alpha}I^{j_\alpha\top}\rangle\right],\\
&\mathbb{E}_\sigma\left[\sup_{E_f,Z\in\Theta_{E_f,Z}}\frac{1}{nc}\sum_{\alpha=1}^{nc}\sigma_\alpha\langle Z,(\tilde{X}-E_f)^\top I^{i_\alpha}I^{j_\alpha\top}\rangle\right]\Bigg\}\\
&+\mathbb{E}_\sigma\left[\sup_{E\in\Theta_E}\frac{1}{nc}\sum_{\alpha=1}^{nc}\sigma_\alpha\langle E_l,I^{i_\alpha}I^{j_\alpha\top}\rangle\right],
\end{aligned}
\tag{73}
$$

where $\Theta_{X_1,Z}=\{(X,Z)|\|XZ\|_*\le\mathcal{X}_*\mathcal{Z}_*,\|XZ\|_F\le\sqrt{n}\}$, and $\Theta_{E_f,Z}=\{(E_f,Z)|\|Z\|_*\le\mathcal{Z}_*,\|E_f\|\le\mathcal{E}_{f,1}\}$.

As $\|\tilde{X}-E_f\|_F\le\|\tilde{X}\|_F+\|E_f\|_F\le\|\tilde{X}\|_F+\sqrt{d}\|E_f\|_1$, it holds that

$$
\begin{aligned}
&\mathbb{E}_\sigma\left[\sup_{E_f,Z\in\Theta_{E_f,Z}}\frac{1}{nc}\sum_{\alpha=1}^{nc}\sigma_\alpha\langle Z,(\tilde{X}-E_f)^\top I^{i_\alpha}I^{j_\alpha\top}\rangle\right]\\
=&\frac{1}{nc}\mathbb{E}_\sigma\left[\sup_{E_f,Z\in\Theta_{E_f,Z}}\langle Z,(\tilde{X}-E_f)^\top\left(\sum_{\alpha=1}^{nc}\sigma_\alpha I^{i_\alpha}I^{j_\alpha\top}\right)\rangle\right]\\
\le&\frac{1}{nc}\mathbb{E}_\sigma\sup_{E_f,Z\in\Theta_{E_f,Z}}\|Z\|_*\|(\tilde{X}-E_f)^\top\left(\sum_{\alpha=1}^{nc}\sigma_\alpha I^{i_\alpha}I^{j_\alpha\top}\right)\|_F\\
\le&\frac{1}{nc}\mathbb{E}_\sigma\sup_{E_f,Z\in\Theta_{E_f,Z}}\|Z\|_*\|\tilde{X}-E_f\|_F\|\sum_{\alpha=1}^{nc}\sigma_\alpha I^{i_\alpha}I^{j_\alpha\top}\|_F\\
\le&\frac{1}{nc}\mathcal{Z}_*(\tilde{\mathcal{X}}_F+\sqrt{d}\mathcal{E}_{f,1})\sqrt{nc}=\frac{1}{\sqrt{nc}}\mathcal{Z}_*(\tilde{\mathcal{X}}_F+\sqrt{d}\mathcal{E}_{f,1}).
\end{aligned}
\tag{74}
$$

Taking Lemma A.2, A.3, and A.4 into consideration, Eq. (73) leads to

$$\begin{aligned}\mathcal{R}(\mathcal{F}) \leq\ & \mathcal{E}_{l,21}\|\boldsymbol{I}^i\boldsymbol{I}^{j\top}\|_{2,\infty}\sqrt{\frac{3\ln c}{nc}} \\ & + \min\left\{\mathcal{X}_*\mathcal{Z}_*\|\boldsymbol{I}^i\boldsymbol{I}^{j\top}\|_2\sqrt{\frac{\ln 2n_c}{nc}}, \frac{1}{\sqrt{nc}}\mathcal{Z}_*(\tilde{\mathcal{X}}_F + \sqrt{d}\mathcal{E}_{f,1}), \sqrt{n}\|\boldsymbol{I}^i\boldsymbol{I}^{j\top}\|_F\sqrt{\frac{2}{nc}}\right\}.\end{aligned} \tag{75}$$

Since $\max_{i,j}\|\boldsymbol{I}^i\boldsymbol{I}^{j\top}\|_{2,\infty} = 1$, $\max_{i,j}\|\boldsymbol{I}^i\boldsymbol{I}^{j\top}\|_2 = 1$, and $\max_{i,j}\|\boldsymbol{I}^i\boldsymbol{I}^{j\top}\|_F = 1$, we arrive at the upper bound of $\mathcal{R}_n(\mathcal{F}_\Theta)$, which is

$$\mathcal{R}_n(\mathcal{F}_\Theta) \leq \mathcal{E}_{l,21}\sqrt{\frac{3\ln c}{nc}} + \min\left\{\mathcal{X}_*\mathcal{Z}_*\sqrt{\frac{\ln(2n_c)}{nc}}, \frac{1}{\sqrt{nc}}\mathcal{Z}_*(\tilde{\mathcal{X}}_F + \sqrt{d}\mathcal{E}_{f,1}), \sqrt{\frac{2}{c}}\right\}. \tag{76}$$

□

Received * July 2024; revised ** ** 2024; accepted ** ** 2024