A Global Geometric Analysis of Maximal Coding Rate Reduction

Peng Wang [1] **Huikang Liu** [2] **Druv Pai** [3] **Yaodong Yu** [3] **Zhihui Zhu** [4] **Qing Qu** [1] **Yi Ma** [3 5]

Abstract

The maximal coding rate reduction (MCR^2) objective for learning structured and compact deep representations is drawing increasing attention, especially after its recent usage in the derivation of fully explainable and highly effective deep network architectures. However, it lacks a complete theoretical justification: only the properties of its global optima are known, and its global landscape has not been studied. In this work, we give a complete characterization of the properties of all its local and global optima, as well as other types of critical points. Specifically, we show that each (local or global) maximizer of the MCR^2 problem corresponds to a low-dimensional, discriminative, and diverse representation, and furthermore, each critical point of the objective is either a local maximizer or a strict saddle point. Such a favorable landscape makes MCR^2 a natural choice of objective for learning diverse and discriminative representations via first-order optimization methods. To validate our theoretical findings, we conduct extensive experiments on both synthetic and real data sets.

1. Introduction

In the past decade, deep learning has exhibited remarkable empirical success across a wide range of engineering and scientific applications (LeCun et al., 2015), such as computer vision (He et al., 2016; Simonyan & Zisserman, 2014), natural language processing (Sutskever et al., 2014; Vaswani et al., 2017), and health care (Esteva et al., 2019), to name a few. As argued by Bengio et al. (2013); Ma et al. (2022), one major factor contributing to the success of deep learning is the ability of deep networks to perform powerful nonlinear feature learning by converting the data distribution to a *compact* and *structured* representation. This representation greatly facilitates various downstream tasks, including classification (Dosovitskiy et al., 2020), segmentation (Kirillov et al., 2023), and generation (Saharia et al., 2022).

Based on the theory of data compression and optimal coding (Ma et al., 2007), Chan et al. (2022); Yu et al. (2020) proposed a principled and unified framework for deep learning to learn a compact and structured representation. Specifically, they proposed to maximize the difference between the *coding rate* of all features and the sum of coding rates of features in each class, which is referred to as *maximal coding rate reduction* (MCR^2). This problem is presented in Problem (4) and visualized in Figure 1(a). Here, the coding rate measures the "compactness" of the features, which is interpreted as the volume of a particular set spanned by the learned features: a lower coding rate implies a more compact feature set[1]. Consequently, the MCR^2 objective aims to maximize the volume of the set of all features while minimizing the volumes of the sets of features from each class. Motivated by the structural similarities between deep networks and unrolled optimization schemes for sparse coding (Gregor & LeCun, 2010; Monga et al., 2021), Chan et al. (2022) constructed a new deep network based on an iterative gradient descent scheme to maximize the MCR^2 objective.[2] Notably, each component of this deep network has precise optimization and geometric interpretations. Moreover, it has achieved strong empirical performance on various vision and language tasks (Chu et al., 2023; Yu et al., 2023a).

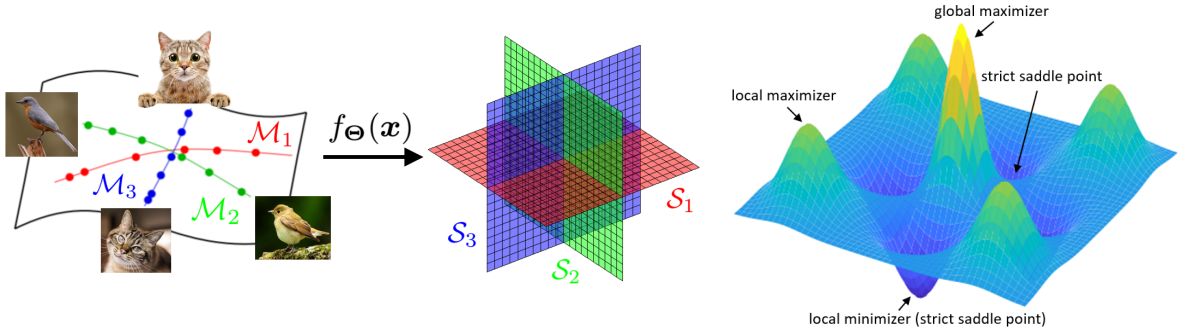
Although the MCR^2-based approach to deep learning is conceptually "white-box" and has achieved remarkable empirical performance, its theoretical foundations have been relatively under-explored. In fact, the effective feature learning mechanism and "white-box" network architecture design based on MCR^2 are direct consequences of these foundations, and understanding them will pave the way to improv-

[1]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor [2]Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai [3]Department of Electrical Engineering and Computer Science, University of California, Berkeley [4]Department of Computer Science and Engineering, The Ohio State University, Columbus [5]Institute of Data Science, University of Hong Kong. Correspondence to: Peng Wang <peng8wang@gmail.com>, Yi Ma <mayi@hku.hk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]Please refer to Chan et al. (2022, Section 2.1) for more details on measuring compactness of feature sets via coding rates.

[2]When performing maximization, we actually mean that we use gradient *ascent*. However, we write gradient descent to maintain consistency with existing optimization literature.



(a) **Visualization of feature learning via the principle of MCR^2.** (b) **Idealized landscape of an MCR^2 objective.**

Figure 1. **An illustration of the properties of MCR^2.** (a) The high-dimensional data $\{\boldsymbol{x}_i\} \subseteq \mathbb{R}^n$ lies on a union of low-dimensional submanifolds. The objective of MCR^2 is to learn a feature mapping $f_{\boldsymbol{\Theta}}(\cdot) : \mathbb{R}^n \to \mathbb{R}^d$ such that $\boldsymbol{z}_i = f_{\boldsymbol{\Theta}}(\boldsymbol{x}_i)$ for all i are low-dimensional, discriminative, and diverse. (b) According to Theorems 3.1 and 3.3, the regularized MCR^2 problem has a benign optimization landscape: each critical point is either a local maximizer or a strict saddle point. Furthermore, each local maximizer, just like the global maximizer, corresponds to a feature representation that consists of a family of orthogonal subspaces, as illustrated in the middle.

ing model interpretability and training efficiency of deep networks. Nevertheless, a comprehensive theoretical understanding of the MCR^2 problem remains lacking. In this work, we take a step towards filling this gap by studying its optimization properties. Notably, analyzing these properties, including local optimality and global landscape, of the MCR^2 objective is extremely challenging. To be precise, its objective function (see Problem (4)) is highly non-concave[3] and complicated, as it involves quadratic functions and the difference between log-determinant functions. To the best of our knowledge, characterizing the local optimality and global optimization landscape of the MCR^2 problem remains an open question.

1.1. Our Contributions

In this work, we study the optimization foundations of the MCR^2-based approach to deep learning. Towards this goal, we characterize the local and global optimality of the regularized MCR^2 problem and analyze its global optimization landscape (see Problem (5)). Our contributions can be highlighted as follows.

Characterizing the local and global optimality. For the regularized MCR^2 problem, we derive the closed-form expressions for its local and global optima for the first time. Our characterization shows that each local maximizer of the regularized MCR^2 problem is within-class compressible and between-class discriminative in the sense that features from the same class belong to a low-dimensional subspace, while features from different classes belong to different orthogonal subspaces. Besides these favorable properties, each global maximizer corresponds to a maximally diverse representation, which attains the highest possible dimension in the space.

[3]We are maximizing the MCR^2 objective. Maximizing a concave function is equivalent to minimizing a convex function.

Studying the global optimization landscape. Next, we show that the regularized MCR^2 function possesses a benign global optimization landscape, despite its complicated structures. More precisely, each critical point is either a local maximizer or strict saddle point of the regularized MCR^2 problem; see Figure 1(b). Consequently, any gradient-based optimization, such as (stochastic) gradient descent, with random initialization can escape saddle points and at least converge to a local maximizer efficiently.

Finally, we conduct extensive numerical experiments on synthetic data sets to validate our theoretical results. Moreover, we use the regularized MCR^2 objective to train deep networks on real data sets. These experimental results constitute an application of the rigorously derived MCR^2 theory to more realistic and complex deep learning problems.

Our results not only establish optimization foundations for the MCR^2 problem but also yield some important implications for the MCR^2-based approach to deep learning. Namely, our theoretical characterizations of local and global optimality offer a compelling explanation for the empirical observations that both deep networks constructed via gradient descent applied to the MCR^2 objective and over-parameterized deep networks trained by optimizing the MCR^2 objective learn low-dimensional, discriminative, and diverse representations. These results align with the motivations of Chan et al. (2022); Yu et al. (2020) for employing the MCR^2 principle for deep learning, and elucidate the outstanding performance of MCR^2-based neural networks across a wide range of vision and language tasks (Chu et al., 2023; Yu et al., 2024). Moreover, our results underscore the potential of MCR^2-based approaches to serve as a cornerstone for future advancements in deep learning, offering a principled approach to pursuing structured and compact representations in practical applications.

1.2. Related Work

Low-dimensional structures in deep representation learning. In the literature, it has long been believed that the role of deep networks is to learn certain (nonlinear) low-dimensional and informative representations of the data (Hinton & Salakhutdinov, 2006; Ma et al., 2022). For example, Papyan et al. (2020) showed that the features learned by cross-entropy (CE) loss exhibit a neural collapse phenomenon during the terminal phase of training, where the features from the same class are mapped to a vector while the features from different classes are maximally linearly separable. Ansuini et al. (2019); Recanatesi et al. (2019) demonstrated that the dimension of the intermediate features first rapidly increases and then decreases from shallow to deep layers. Masarczyk et al. (2023) concluded that the deep layers of neural networks progressively compress within-class features to learn low-dimensional features. Notably, Wang et al. (2023) proposed a theoretical framework to analyze hierarchical feature learning for learning low-dimensional representations. They showed that each layer of deep linear networks progressively compresses within-class features and discriminates between-class features in classification problems.

The MCR^2-based approach to deep learning. The MCR^2-based approach to deep learning for seeking structured and compact representations was first proposed by Yu et al. (2020). Notably, they provided a global optimality analysis of the MCR^2 problem (4) with additional rank constraints on the feature matrix of each class. Chan et al. (2022) designed a new multi-layer deep network architecture, named ReduNet, based on an iterative gradient descent scheme for maximizing the MCR^2 objective. To learn self-consistent representations, Dai et al. (2022) extended this approach to the closed-loop transcription (CTRL) framework, which is formulated as a max-min game to optimize a modified MCR^2 objective. This game was shown to have global equilibria corresponding to compact and structured representations (Pai et al., 2022). Recently, Yu et al. (2023b) showed that a transformer-like architecture named CRATE, which obtains strong empirical performance (Chu et al., 2023; Yu et al., 2023a; 2024), can be naturally derived through an iterative optimization scheme for maximizing the sparse rate reduction objective, which is an adaptation to sequence data of the MCR^2 objective studied in this work.

Notation. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we use $\|\boldsymbol{A}\|$ to denote its spectral norm, $\|\boldsymbol{A}\|_F$ to denote its Frobenius norm, and a_{ij} its (i,j)-th element. Given a vector $\boldsymbol{a} \in \mathbb{R}^d$, we use $\|\boldsymbol{a}\|$ to denote its ℓ_2-norm, a_i its i-th element, and $\mathrm{diag}(\boldsymbol{a})$ the diagonal matrix with $\boldsymbol{a}$ on its diagonal. Given a positive integer n, we denote by $[n]$ the set $\{1,\dots,n\}$. Given a set of integers $\{n_k\}_{k=1}^K$, let $n_{\max} = \max\{n_k : k \in [K]\}$. Let $\mathcal{O}^{m\times n} = \{\boldsymbol{Z} \in \mathbb{R}^{m\times n} : \boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{I}_n\}$ denote the set of all $m \times n$ orthonormal matrices.

2. Problem Setup

In this section, we first review the basic concepts of MCR^2 for deep representation learning in Section 2.1, and then introduce our studied problem in Section 2.2.

2.1. An Overview of MCR^2

In deep representation learning, given data $\{\boldsymbol{x}_i\}_{i=1}^m \subseteq \mathbb{R}^n$ from multiple (say K) classes, the goal is to learn neural-network representations of these samples that facilitate downstream tasks. Recent empirical studies have shown that good features can be learned for tasks such as classification or autoencoding by using heuristics to promote either the contraction of samples in the same class (Rifai et al., 2011) or the contrast of samples between different classes (He et al., 2020; Oord et al., 2018) during the training of neural networks. Notably, Chan et al. (2022); Yu et al. (2020) unified and formalized these practices and demonstrated that the MCR^2 objective is an effective objective to learn within-class compressible and between-class discriminative representations of the data.

The formulation of MCR^2. In this work, we mainly consider an MCR^2 objective for supervised learning problems. Specifically, let $\boldsymbol{z}_i = f_{\boldsymbol{\Theta}}(\boldsymbol{x}_i)$ for all $i \in [m]$ denote the features learned via the feature mapping $f_{\boldsymbol{\Theta}}(\cdot) : \mathbb{R}^n \to \mathbb{R}^d$ parameterized by $\boldsymbol{\Theta}$. For each $k \in [K]$, let $\boldsymbol{\pi}^k \in \{0,1\}^m$ be a label vector denoting membership of the samples in the k-th class, i.e., $\pi_i^k = 1$ if sample i belongs to class k and $\pi_i^k = 0$ otherwise for all $i \in [m]$, and $m_k := \sum_{i=1}^m \pi_i^k$ be the number of samples in the k-th class.

For each $k \in [K]$, let $\boldsymbol{Z}_k \in \mathbb{R}^{d\times m_k}$ be the matrix whose columns are the features in the k-th class. Without loss of generality, we reorder the samples in a class-by-class manner, so that we can write the matrix of all features as

$$\boldsymbol{Z} = [\boldsymbol{Z}_1,\dots,\boldsymbol{Z}_K] \in \mathbb{R}^{d\times m}. \tag{1}$$

On one hand, to make features between different classes discriminative or contrastive, one can maximize the lossy coding rate of all features in $\boldsymbol{Z}$, as argued in (Chan et al., 2022; Yu et al., 2020), as follows:

$$R(\boldsymbol{Z}) := \frac{1}{2}\log\det\left(\boldsymbol{I} + \frac{d}{m\epsilon^2}\boldsymbol{Z}\boldsymbol{Z}^T\right), \tag{2}$$

where $\epsilon > 0$ is a prescribed quantization error.[4] On the other hand, to make features from the same class compressible or contractive, one can minimize the average lossy coding rate of features in the k-th class as follows:

$$R_c\left(\boldsymbol{Z};\boldsymbol{\pi}^k\right) = \frac{m_k}{2m}\log\det\left(\boldsymbol{I} + \frac{d}{m_k\epsilon^2}\boldsymbol{Z}_k\boldsymbol{Z}_k^T\right). \tag{3}$$

[4] Here, $R(\boldsymbol{Z})$ is also known as the rate-distortion function in information theory (Cover, 1999), which represents the average number of binary bits needed to encode the data $\boldsymbol{Z}$.

Consequently, a good representation tends to maximize the difference between the coding rate for the whole and that for each class as follows:

$$\max_{\boldsymbol{Z}\in\mathbb{R}^{d\times m}} \quad R(\boldsymbol{Z}) - \sum_{k=1}^{K} R_c\left(\boldsymbol{Z};\boldsymbol{\pi}^k\right) \qquad (4)$$
$$\text{s.t.} \quad \|\boldsymbol{Z}_k\|_F^2 = m_k, \quad \forall k \in [K].$$

This is referred to as the principle of *maximal coding rate reduction* in (Chan et al., 2022; Yu et al., 2020). It is worth mentioning that this principle can be extended to self-supervised and even unsupervised learning settings, where we learn the label vectors $\{\boldsymbol{\pi}^k\}_{k=1}^K$ during training.

2.2. The Regularized MCR^2 Problem

Due to the Frobenius norm constraints, it is a tremendously difficult task to analyze Problem (4) from an optimization-theoretic perspective, as all the analysis would occur on a product of spheres instead of on Euclidean space. Therefore, we consider the Lagrangian formulation of (4). This can be viewed as a tight relaxation or even an equivalent problem of (4) whose optimal solutions agree under specific settings of the regularization parameter; see Proposition 3.2. Specifically, the formulation we study, referred to henceforth as the *regularized MCR^2 problem*, is as follows:

$$\max_{\boldsymbol{Z}} \; F(\boldsymbol{Z}) := R(\boldsymbol{Z}) - \sum_{k=1}^{K} R_c(\boldsymbol{Z};\boldsymbol{\pi}^k) - \frac{\lambda}{2}\|\boldsymbol{Z}\|_F^2, \quad (5)$$

where $\lambda > 0$ is the regularization parameter. Remark that our study on this problem applies meaningfully to at least two approaches to learning deep representations using the MCR^2 principle.

Applications of our formulation to deep representation learning via unrolled optimization. The first approach, as argued by Chan et al. (2022); Yu et al. (2023a), is to construct a new deep network architecture, i.e., ReduNet (Chan et al., 2022) or CRATE (Yu et al., 2023a), based on an iterative gradient descent scheme to optimize the MCR^2-type objective. In this approach, each layer of the constructed network approximates a gradient descent step to optimize the MCR^2-type objective given the input representation. The key takeaway is that these networks approximately implement gradient descent directly on the representations, so our analysis of the optimization properties of the MCR^2-type objective translates to explanations of the corresponding properties of the learned representations and architectures of these deep networks. In particular, our argument that the optima and optimization landscape of (5) are favorable directly translates to a justification of the correctness of learned representations of the ReduNet and a validation of its architecture design. Moreover, this study enables principled improvement of deep network architectures constructed via unrolled optimization by leveraging more advanced optimization techniques better suited for problems with benign landscapes. This can improve model interpretability and efficiency.

Applications of our formulation to deep representation learning with standard neural networks. In the second approach, one parameterizes the feature mapping $f_{\boldsymbol{\Theta}}(\cdot)$ via standard deep neural networks such as a multi-layer perceptron or ResNet (He et al., 2016), and treats the MCR^2-type objective like other loss functions applied to outputs of a neural network, such as mean-squared error or cross-entropy loss. Studying Problem (5) from this perspective would require us to optimize over $\boldsymbol{\Theta}$ instead of over $\boldsymbol{Z}$. This new optimization problem would be extraordinarily difficult to analyze, because modern neural networks have nonlinear interactions across many layers, so the parameters $\boldsymbol{\Theta}$ would affect the final representation $\boldsymbol{Z}$ in a complex way. Fortunately, since modern neural networks are often highly over-parameterized, they can interpolate or approximate any continuous function in the feature space (Lu et al., 2017), so we may omit these constraints by assuming the unconstrained feature model, where $\boldsymbol{z}_i$ for all $i \in [N]$ are treated as free optimization variables (Mixon et al., 2020; Yaras et al., 2022; Zhu et al., 2021; Wang et al., 2022a). Consequently, studying the optimization properties of Problem (5) provides valuable insights into the structures of learned representations and the efficiency of training deep networks using MCR^2-type objectives.

Difficulties of analyzing Problem (5). Although Problem (5) has no constraints, one can observe that Problem (5) is highly non-concave due to the quadratic form $\boldsymbol{Z}_k\boldsymbol{Z}_k^T$ and the difference of log-determinant functions. Notably, this problem shares similarities with low-rank matrix factorization problems. However, it employs the log-determinant function instead of the Frobenius norm, and the computation of the objective gradient involves matrix inverses. Therefore, from an optimization point of view, it is extremely challenging to analyze Problem (5).

3. Main Results

In this section, we first characterize the local and global optimal solutions of Problem (5) in Section 3.1, and then analyze the global landscape of the objective function in Section 3.2.

3.1. Characterization of Local and Global Optimality

Although Problem (5) is highly non-concave and involves matrix inverses in its gradient computation, we can still explicitly characterize its local and global optima as follows.

Theorem 3.1 (Local and global optimality). *Suppose that the number of training samples in the k-th class is $m_k > 0$*

for each $k \in [K]$. Given a coding precision $\epsilon > 0$, if the regularization parameter satisfies

$$\lambda \in \left(0, \frac{d(\sqrt{m/m_{\max}} - 1)}{m(\sqrt{m/m_{\max}} + 1)\epsilon^2}\right], \tag{6}$$

then the following statements hold:
*(i) (***Characterization of local maximizers***) $\boldsymbol{Z} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_K]$ is a local maximizer of Problem (5) if and only if the k-th block admits the following decomposition*

$$\boldsymbol{Z}_k = \overline{\sigma}_k \boldsymbol{U}_k \boldsymbol{V}_k^T, \tag{7}$$

where (a) $r_k = \operatorname{rank}(\boldsymbol{Z}_k)$ satisfies $r_k \in [0, \min\{m_k, d\})$ and $\sum_{k=1}^K r_k \leq \min\{m, d\}$, (b) $\boldsymbol{U}_k \in \mathcal{O}^{d \times r_k}$ satisfies $\boldsymbol{U}_k^T \boldsymbol{U}_l = \boldsymbol{0}$ for all $l \neq k$, $\boldsymbol{V}_k \in \mathcal{O}^{m_k \times r_k}$, and (c) the singular value $\overline{\sigma}_k$ is given in (16) for each $k \in [K]$.
*(ii) (***Characterization of global maximizers***) $\boldsymbol{Z} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_K]$ is a global maximizer of Problem (5) if and only if (a) it satisfies the above all conditions and $\sum_{k=1}^K r_k = \min\{m, d\}$, and (b) for all $k \neq l \in [K]$ satisfying $m_k < m_l$ and $r_l > 0$, we have $r_k = \min\{m_k, d\}$.*

We defer the proof to Section 4.1 and Appendix D.1. In this theorem, we explicitly characterize the local and global optima of Problem (5). Intuitively, this demonstrates that the features represented by each local maximizer of Problem (5) are low-dimensional and discriminative in the sense that (i) *Within-class compressible:* According to (7), at each local maximizer, the features from the same class belong to the same low-dimensional linear subspace.
(ii) *Between-class discriminative:* It follows from (7) and $\boldsymbol{U}_k^T \boldsymbol{U}_l = \boldsymbol{0}$ for all $k \neq l$ that, at each local maximizer, the features from different classes belong to different subspaces that are orthogonal to each other.
Moreover, the features represented by each global maximizer of Problem (5) are not only low-dimensional and discriminative but also diverse in the sense that
(iii) *Maximally Diverse Representation:* According to $\sum_{k=1}^K r_k = \min\{m, d\}$, at each global maximizer, the total dimension of all features is maximized to match the highest dimension that it can achieve in the feature space.

Quality of local versus global optima. Our above discussion explains the merits of achieving both local and global optima. At each maximizer, the representations are within-class compressible and between-class discriminative (Theorem 3.1 (i)). Moreover, global maximizers further satisfy that the representations are all maximally diverse (Theorem 3.1 (ii) (a)). If all classes were balanced, i.e., $m_1 = \cdots = m_K$, then Theorem 3.1 (ii) (b) would not apply, and these properties would be all that Theorem 3.1 asserts. In this case, global optima would clearly be desired over local optima. However, in the unbalanced case, the situation is more complex, because Theorem 3.1 (ii) (b) would apply. It says that for global optima, the classes with the smallest numbers of samples would fill to the largest dimension possible, and the very largest classes could collapse to $\boldsymbol{0}$, an undesirable situation. A dramatic example of this is when $m_1 > \cdots > m_K > d$, for then any global optimum would have $\operatorname{rank}(\boldsymbol{Z}_K) = d$ and $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{K-1}$ all collapse to $\boldsymbol{0}$. Overall, in the unbalanced case, global optima may not always correspond to the best representations. In particular, local optima with more equitable rank distributions (like bigger classes span more dimensions) which are still maximally diverse (i.e., ranks of each class sum to the dimension d) could be preferred in applications. As demonstrated in Section 5.1, these kinds of potentially useful local optima are realized in experiments, even with unbalanced classes.

Relation between Problems (4) and (5). Based on the characterization of global optimality in Theorem 3.1, we show the following proposition that establishes the relationship between the constrained MCR^2 problem (4) and the regularized MCR^2 problem (5) in terms of their global solutions under an appropriate choice of the regularization parameter. The proof of this result can be found in Appendix D.2.

Proposition 3.2. *Suppose that the number of training samples in each class is the same, i.e., $m_1 = \cdots = m_K$, and the coding precision $\epsilon > 0$ satisfies*

$$\epsilon \leq \frac{1}{6}\sqrt{\frac{d}{m}} \exp\left(-\frac{1}{2}\right) K^{-\frac{1}{K-1}} \left(1 + \frac{1}{\sqrt{K}}\right)^{-\frac{m}{K-1}}. \tag{8}$$

The following statements hold:
(i) If $m < d$ and the regularization parameter in Problem (5) is set as

$$\lambda = \frac{\alpha}{1+\alpha} - \frac{\alpha}{1+K\alpha}, \tag{9}$$

Problems (4) and (5) have the same global solution set.
(ii) If $m \geq d$, d/K is an integer, and the regularization parameter in Problem (5) is set as

$$\lambda = \frac{\alpha}{1+\alpha m/d} - \frac{\alpha}{1+\alpha K m/d}, \tag{10}$$

the global solution set of Problem (4) is a subset of that of Problem (5).

According to this proposition, if ϵ and λ are appropriately chosen for Problem (5), when $m < d$, Problems (4) and (5) are equivalent in terms of their global optimal solutions; when $m \leq d$, Problem (5) is a tight Lagrangian relaxation of Problem (4) such that the global solution set of the former contains that of the latter.

3.2. Analysis of Global Optimization Landscape

While we have characterized the local and global optimal solutions in Theorem 3.1, it remains unknown whether these

solutions can be computed efficiently using GD to solve Problem (5), as GD may get stuck at a saddle point. Fortunately, Sun et al. (2015); Lee et al. (2016) showed that if a function is twice continuously differentiable and satisfies *strict saddle property*, i.e., each critical point is either a local minimizer or a strict saddle point[5], GD converges to its local minimizer almost surely with random initialization. We investigate the global optimization landscape of Problem (5) by characterizing all of its critical points as follows.

Theorem 3.3 (Benign optimization landscape). *Suppose that the number of training samples in the k-th class is $m_k > 0$ for each $k \in [K]$. Given a coding precision $\epsilon > 0$, if the regularization parameter satisfies (6), it holds that any critical point $\boldsymbol{Z}$ of Problem (5) that is not a local maximizer is a strict saddle point.*

We defer the proof to Section 4.2 and Appendix D.3. Here, we make some remarks on this theorem and also on the consequences of the results derived so far.

Differences from existing results on the MCR^2 problem. Chan et al. (2022); Yu et al. (2020) have characterized the global optimality of Problem (4) with Frobenius norm constraints on each $\boldsymbol{Z}_k$ in the UFM. However, their analysis requires an additional rank constraint on each $\boldsymbol{Z}_k$ and only characterizes globally optimal representations. In contrast, our analysis eliminates the need for the rank constraint, and we characterize local and global optimality in Problem (5), as well as its optimization landscape. Interestingly, we demonstrate that the features represented by each local maximizer — not just global maximizers — are also compact and structured. Furthermore, we demonstrate that the regularized MCR^2 objective (5) is a strict saddle function. To the best of our knowledge, Theorems 3.1 and 3.3 constitute the first analysis of local optima and optimization landscapes for MCR^2 objectives. According to Daneshmand et al. (2018); Lee et al. (2016); Xu et al. (2018), Theorems 3.1 and 3.3 imply that low-dimensional and discriminative representations can be efficiently found by (stochastic) GD on Problem (5) from a random initialization.

Comparison to existing landscape analyses in non-convex optimization. In recent years, there has been a growing body of literature exploring optimization landscapes of non-convex problems in machine learning and deep learning. These include low-rank matrix factorization (Ge et al., 2017; Sun et al., 2018; Chi et al., 2019; Zhang et al., 2020), community detection (Wang et al., 2021; 2022b), dictionary learning (Sun et al., 2017; Qu et al., 2020; Zhai et al., 2020), and deep neural networks (Sun et al., 2020; Yaras et al., 2022; Zhou et al., 2022; Zhu et al., 2021; Jiang et al., 2024). The existing analyses in the literature cannot be applied to the MCR^2 problem due to its special structure, which involves the log-determinant of all features minus the sum of the log-determinant of features in each class. Our work contributes to the literature on optimization landscape analyses of non-convex problems by showing that the MCR^2 problem has a benign optimization landscape. Our approach may be of interest to analyses of the landscapes of other intricate loss functions in practical applications.

[5] We say that a critical point is a strict saddle point of Problem (5) if it has a direction with strictly positive curvature; see Definition A.2. This includes classical saddle points with strictly positive curvature as well as local minimizers.

4. Proofs of Main Results

In this section, we sketch the proofs of our main theorems in Section 3. The complete proofs can be found in Sections B and C of the appendix. For ease of exposition, let

$$\alpha := \frac{d}{m\epsilon^2}, \quad \alpha_k := \frac{d}{m_k\epsilon^2}, \quad \forall k \in [K]. \tag{11}$$

4.1. Analysis of Optimality Conditions

Our goal in this subsection is to characterize the local and global optima of Problem (5). Towards this goal, we first provide an upper bound on the objective function F in Problem (5). In particular, this upper bound is tight when the blocks $\{\boldsymbol{Z}_k\}_{k=1}^K$ are orthogonal to each other. This result is a direct consequence of (Chan et al., 2022, Lemma 10).

Lemma 4.1. *For any $\boldsymbol{Z} = [\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_K] \in \mathbb{R}^{d\times m}$ with $\boldsymbol{Z}_k \in \mathbb{R}^{d\times m_k}$, we have*

$$F(\boldsymbol{Z}) \le \sum_{k=1}^{K} \left(\frac{1}{2}\log\det\left(\boldsymbol{I}_n + \alpha \boldsymbol{Z}_k\boldsymbol{Z}_k^T\right) - \frac{m_k}{2m}\log\det\left(\boldsymbol{I}_n + \alpha_k \boldsymbol{Z}_k\boldsymbol{Z}_k^T\right) - \frac{\lambda}{2}\|\boldsymbol{Z}_k\|_F^2 \right), \tag{12}$$

where the inequality becomes equality if and only if $\boldsymbol{Z}_k^T\boldsymbol{Z}_l = \boldsymbol{0}$ for all $1 \le k \ne l \le K$.

Next, we study the following set of critical points, which are between-class discriminative (i.e., $\boldsymbol{Z}_k^T\boldsymbol{Z}_l = \boldsymbol{0}$):

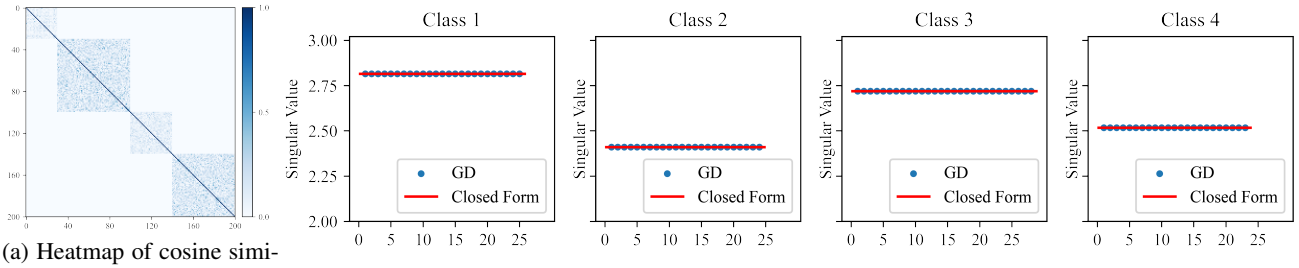
$$\mathcal{Z} := \left\{ \boldsymbol{Z} : \nabla F(\boldsymbol{Z}) = 0,\ \boldsymbol{Z}_k^T\boldsymbol{Z}_l = \boldsymbol{0}, \forall k \ne l \right\}. \tag{13}$$

Proposition 4.2. *Consider the setting of Theorem 3.1. It holds that $\boldsymbol{Z} = [\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_K] \in \mathcal{Z}$ if and only if each $\boldsymbol{Z}_k$ admits the following singular value decomposition*

$$\boldsymbol{Z}_k = \boldsymbol{U}_k\tilde{\boldsymbol{\Sigma}}_k\boldsymbol{V}_k^T,\ \tilde{\boldsymbol{\Sigma}}_k = \operatorname{diag}\left(\sigma_{k,1}, \dots, \sigma_{k,r_k}\right), \tag{14}$$

where (i) $r_k \in [0, \min\{m_k, d\})$ satisfies $\sum_{k=1}^K r_k \le d$, (ii) $\boldsymbol{U}_k \in \mathcal{O}^{n\times r_k}$ satisfies $\boldsymbol{U}_k^T\boldsymbol{U}_l = \boldsymbol{0}$ for all $1 \le k \ne l \le K$, $\boldsymbol{V}_k \in \mathcal{O}^{m_k\times r_k}$ for all $k \in [K]$, and (iii) the singular values satisfy

$$\sigma_{k,i} \in \{\overline{\sigma}_k, \underline{\sigma}_k\},\ \forall i \in [r_k], \tag{15}$$



(a) Heatmap of cosine similarity among features.

(b) The number and magnitude of singular values of each class matrix.

Figure 2. **Validation of theory for the MCR^2 problem.** (a) We visualize the heatmap of cosine similarity among learned features by GD for solving Problem (5). The lighter pixels represent lower cosine similarities between pairwise features. (b) The blue dots are plotted based on the singular values by applying SVD to the solution returned by GD, and the red line is plotted according to the closed-form solution in (7). The number of nonzero singular values in each subspace is $24, 23, 27, 26$, respectively.

where $\eta_k = (\alpha_k - \alpha) - \lambda\,(m/m_k + 1)$ *and*

$$\overline{\sigma}_k = \left(\frac{\eta_k + \sqrt{\eta_k^2 - 4\lambda^2 m/m_k}}{2\lambda\alpha_k}\right)^{1/2}, \tag{16}$$

$$\underline{\sigma}_k = \left(\frac{\eta_k - \sqrt{\eta_k^2 - 4\lambda^2 m/m_k}}{2\lambda\alpha_k}\right)^{1/2}. \tag{17}$$

This proposition shows that each critical point that is between-class discriminative (i.e., $\boldsymbol{Z}_k^T \boldsymbol{Z}_l = \boldsymbol{0}$) exhibits a specific structure: the singular values of $\boldsymbol{Z}_k$ can only take on two possible values, $\overline{\sigma}_k$ and $\underline{\sigma}_k$. We will leverage this structure and further show that $\boldsymbol{Z}$ is a strict saddle if there exists a $\boldsymbol{Z}_k$ with a singular value $\underline{\sigma}_k$.

4.2. Analysis of Optimization Landscape

Our goal in this subsection is to show that the function F in Problem (5) has a benign optimization landscape. Towards this goal, we denote the set of critical point of F by

$$\mathcal{X} = \left\{\boldsymbol{Z} \in \mathbb{R}^{d\times m} : \nabla F(\boldsymbol{Z}) = \boldsymbol{0}\right\}. \tag{18}$$

According to (13), we divide the critical point set $\mathcal{X}$ into two disjoint sets $\mathcal{Z}$ and $\mathcal{Z}^c$, i.e., $\mathcal{X} = \mathcal{Z} \cup \mathcal{Z}^c$, where

$$\mathcal{Z}^c := \left\{\boldsymbol{Z} : \nabla F(\boldsymbol{Z}) = 0,\ \boldsymbol{Z}_k^T \boldsymbol{Z}_l \neq \boldsymbol{0},\ \exists k \neq l\right\}. \tag{19}$$

Moreover, according to Proposition 4.2, we further divide $\mathcal{Z}$ into two disjoint sets $\mathcal{Z}_1$ and $\mathcal{Z}_2$, i.e., $\mathcal{Z} = \mathcal{Z}_1 \cup \mathcal{Z}_2$. Here,

$$\begin{aligned}\mathcal{Z}_1 &:= \mathcal{Z} \cap \left\{\boldsymbol{Z} : \sigma_{k,i}(\boldsymbol{Z}_k) = \overline{\sigma}_k, \forall i \in [r_k], k \in [K]\right\},\\ \mathcal{Z}_2 &:= \mathcal{Z} \setminus \mathcal{Z}_1,\end{aligned} \tag{20}$$

where $\sigma_{k,i}(\boldsymbol{Z}_k)$ denotes the i-th largest singular value of $\boldsymbol{Z}_k$. Our first step is to show that any point belonging to $\mathcal{Z}_1$ is a local maximizer, while any point belonging to $\mathcal{Z}_2$ is a strict saddle point.

Proposition 4.3. *Consider the setting of Theorem 3.3. Suppose that $\boldsymbol{Z} \in \mathcal{Z}$. Then, the following statements hold:*

(i) If $\boldsymbol{Z}_k$ takes the form of (14) with $\sigma_{k,i} = \overline{\sigma}_k$ for all $i \in [r_k]$ and all $k \in [K]$, i.e., $\boldsymbol{Z} \in \mathcal{Z}_1$, then $\boldsymbol{Z}$ is a local maximizer.
(ii) If there exists a $k \in [K]$ and $i \in [r_k]$ with $r_k \geq 1$ such that $\sigma_{k,i} = \underline{\sigma}_k$, i.e., $\boldsymbol{Z} \in \mathcal{Z}_2$, then $\boldsymbol{Z}$ is a strict saddle point.

Next, we proceed to the second step to show that any point belonging to $\mathcal{Z}^c$ is a strict saddle point. It suffices to find a direction $\boldsymbol{D} \in \mathbb{R}^{d\times m}$ such that $\nabla^2 F(\boldsymbol{Z})[\boldsymbol{D}, \boldsymbol{D}] > 0$ for each $\boldsymbol{Z} \in \mathcal{Z}^c$ according to Definition A.2.

Proposition 4.4. *Consider the setting of Theorem 3.3. If $\boldsymbol{Z} \in \mathbb{R}^{d\times m}$ is a critical point and there exists $1 \leq k \neq l \leq K$ such that $\boldsymbol{Z}_k^T \boldsymbol{Z}_l \neq \boldsymbol{0}$, i.e., $\boldsymbol{Z} \in \mathcal{Z}^c$, then $\boldsymbol{Z}$ is a strict saddle point.*

With the above preparations that characterize all the critical points, we can prove Theorem 3.1 and Theorem 3.3. We refer the reader to Appendix D for the detailed proof.

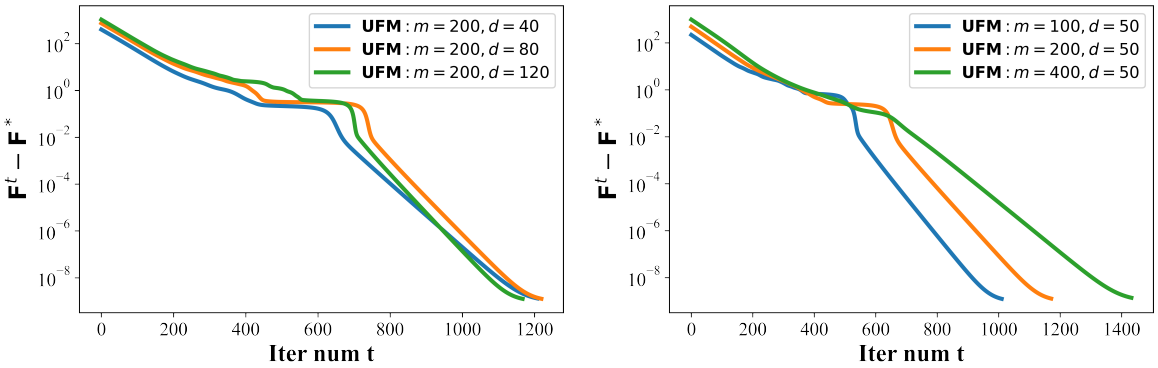
5. Experimental Results

In this section, we first conduct numerical experiments on synthetic data in Section 5.1 to validate our theoretical results, and then on real-world data sets using deep neural networks in Section 5.2 to further support our theory. All codes are implemented in Python mainly using NumPy and PyTorch. All of our experiments are executed on a computing server equipped with NVIDIA A40 GPUs. Due to space limitations, we defer some implementation details and additional experimental results to Appendix E.

5.1. Validation of Theory for Solving Problem (5)

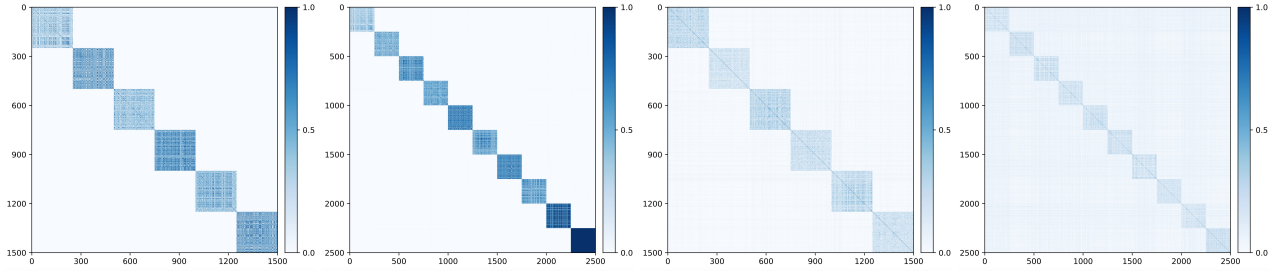
In this subsection, we employ GD for solving Problem (5) with different parameter settings. We visualize the optimization dynamics and structures of the solutions returned by GD to verify and validate Theorems 3.1 and 3.3.

Verification of Theorem 3.1. In this experiment, we set the parameters in Problem (5) as follows: the dimension of features $d = 100$, the number of classes $K = 4$, the num-



(a) Convergence for $m = 200$ and $d \in \{40, 80, 120\}$. (b) Convergence for $d = 50$ and $m \in \{100, 200, 400\}$.

Figure 3. **Convergence performance of GD for solving the regularized MCR^2 problem.** The x-axis is number of iterations (also denoted by t), and the y-axis is the function value gap $F^t - F^*$, where $F^t = F(\boldsymbol{Z}^t)$ denotes the function value at the t-th iterate $\boldsymbol{Z}^t$ generated by GD, and F^* is the optimal value of Problem (5) computed according to (7) in Theorem 3.1.



(a) MNIST: $m = 1500, K = 6$ (b) MNIST: $m = 2500, K = 10$ (c) CIFAR: $m = 1500, K = 6$ (d) CIFAR: $m = 2500, K = 10$

Figure 4. **Heatmap of cosine similarity among features produced by deep networks trained on MNIST and CIFAR-10.** The darker pixels represent higher absolute cosine similarity between features.

ber of samples in each class is $m_1 = 30, m_2 = 70, m_3 = 40, m_4 = 60$, the regularization parameter $\lambda = 0.1$, and the quantization error $\epsilon = 0.5$. Then, one can verify that λ satisfies (6). For the solution $\boldsymbol{Z}$ returned by GD, we first plot the heatmap of the cosine similarity between pairwise columns of $\boldsymbol{Z}$ in Figure 2(a). We observe that the features from different classes are orthogonal to each other, while the features from the same class are correlated. Next, we compute the singular values of $\boldsymbol{Z}_k$ via singular value decomposition (SVD) and plot the singular values using blue dots for each $k \in [K]$ in Figure 2(b). According to the closed-form solution (7) in Theorem 3.1, we also plot the theoretical bound of singular values in red in Figure 2(b). One can observe that the number of singular values of each block is respectively $24, 23, 27, 26$, summing up to 100, and the red line perfectly matches the blue dots. These results all provide strong support for Theorem 3.1.

Verification of Theorem 3.3. In this experiment, we maintain the same setting as above, except that the number of samples in each class is equal. We first fix $m = 200$ and vary $d \in \{40, 80, 120\}$, and then fix $d = 50$ and vary $d \in \{100, 200, 400\}$ to run GD. We plot the distances between function values of the iterates to the optimal value, which is computed according to (7) in Theorem 3.1, against the iteration numbers in Figure 3. We observe that GD with random initialization converges to an optimal solution at a linear rate. This indicates that the MCR^2 has a benign global landscape, which supports Theorem 3.3.

5.2. Training Deep Networks Using Regularized MCR^2

In this subsection, we conduct numerical experiments on the image datasets MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky et al., 2009) to provide evidence that our theory also applies to deep networks. More specifically, we employ a multi-layer perceptron network with ReLU activation as the feature mapping $\boldsymbol{z} = f_{\boldsymbol{\Theta}}(\boldsymbol{x})$ with output dimension 32 for MNIST and 128 for CIFAR-10. Then, we train the network parameters $\boldsymbol{\Theta}$ via Adam (Kingma & Ba, 2014) by optimizing Problem (5).

Experimental setting and results. In the experiments, we randomly sample a balanced subset with K classes and m samples from MNIST or CIFAR-10, where each class has the same number of samples. We set $\lambda = 0.001$ and $\epsilon = 0.5$. For different subsets with corresponding values of (m, K), we run experiments and report the function value $\hat{F}$ obtained by training deep networks and the optimal value F^* computed using the closed-form solution in Theorem 3.1

Table 1. Function value $\hat{F}$ obtained by training deep networks, the optimal value F^* computed by our theory on subsets of MNIST or CIFAR-10, and discrimination metric s of features.

MNIST (m, K)	$\hat{F}$	F^*	s
(1000, 4)	37.38	37.38	$5.9 \cdot 10^{-6}$
(1500, 6)	38.96	38.96	$3.8 \cdot 10^{-6}$
(2000, 8)	38.48	38.48	0.011
(2500, 10)	37.41	37.41	0.008
CIFAR-10 (m, K)	$\hat{F}$	F^*	s
(1000, 4)	215.61	215.61	0.004
(1500, 6)	229.14	229.14	0.029
(2000, 8)	230.70	230.70	0.059
(2500, 10)	228.48	228.49	0.171

in Table 1. To verify the discriminative nature of the features obtained by training deep networks across different classes, we measure the discrimination between features belonging to different classes by computing the cosine of the principal angle (Björck & Golub, 1973) between the class subspaces: $s = \max\left\{\|\boldsymbol{U}_k^T \boldsymbol{U}_l\| : k \neq l \in [K]\right\} \in [0, 1]$, where the columns of $\boldsymbol{U}_k \in \mathbb{R}^{d \times r_k}$ are the right singular vectors corresponding to the top r_k singular values of $\boldsymbol{Z}_k$ defined in (14) and r_k is its rank[6] for each $k \in [K]$. In particular, when s is smaller, the spaces spanned by each pair $\boldsymbol{Z}_k$ and $\boldsymbol{Z}_l$ for $k \neq l$ are closer to being orthogonal to each other. Then, we record the value s in Table 1 in different settings. Moreover, we visualize the pairwise cosine similarities between learned features on MNIST and CIFAR-10 when $(m, K) = (1500, 6)$ and $(2500, 10)$ in Figure 4.

We observe from Table 1 that the function value returned by training deep networks is extremely close to the global optimal value of Problem (5) and from the value s and Figure 4 that the features from different classes are nearly orthogonal to each other. These observations, together with Theorems 3.1 and 3.3, indicate that Problem (5) retains its optimization properties even when $\boldsymbol{Z}$ is parameterized by a neural network. Our theoretical analysis of Problem (5) thus illustrates a qualitative picture of training deep networks with the regularized MCR^2 objective.

6. Conclusion

In this work, we provided a complete characterization of the global landscape of the MCR^2 objective, a highly nonconcave and nonlinear function used for representation learning. We characterized all critical points, including the local and global optima, of the MCR^2 objective, and showed that — surprisingly — it has a benign global optimization landscape. These characterizations provide rigorous justifications for why such an objective can be optimized well using simple algorithms such as gradient-based methods. In particular, we show that even local optima of the objective leads to geometrically meaningful representations. Our experimental results on synthetic and real-world datasets clearly support this new theoretical characterization. With the global landscape clearly revealed, our work paves the way for exploring better optimization strategies, hence better deep neural network architectures, for optimizing the MCR^2 objective more efficiently and effectively. For future work, it is natural to extend our analysis to Problem (4) with deep network parameterizations. It is also interesting to study the sparse MCR^2 objective, which has led to high-performance transformer-like architectures (Yu et al., 2023a;b).

[6] We estimate the rank of a matrix by rounding its "stable rank" (Horn & Johnson, 2012): $r_k = \text{round}(\|\boldsymbol{Z}_k\|_F^2 / \|\boldsymbol{Z}_k\|^2)$.

Acknowledgements

The work of P.W. is supported in part by ARO YIP award W911NF1910027 and DoE award DE-SC0022186. The work of H.L. is supported in part by NSF China under Grant 12301403 and the Young Elite Scientists Sponsorship Program by CAST 2023QNRC001. The work of D.P. is supported by a UC Berkeley College of Engineering Fellowship. Y.Y. acknowledges support from the joint Simons Foundation-NSF DMS grant #2031899. Q.Q. acknowledges support from NSF CAREER CCF-2143904, NSF CCF-2212066, NSF CCF-2212326, NSF IIS 2312842, and Amazon AWS AI Award, and a gift grant from KLA. Z.Z. acknowledges support from NSF grants CCF-2240708 and IIS-2312840. Y.M. acknowledges support from the joint Simons Foundation-NSF DMS grant #2031899, the ONR grant N00014-22-1-2102, and the University of Hong Kong.

Impact Statement

This paper advances the state of knowledge in machine learning theory, including non-convex optimization for representation learning. We do not anticipate any particular societal or ethical impacts of this paper, besides those usually associated with theoretically oriented machine learning papers.

References

Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Björck, Å. and Golub, G. H. Numerical methods for computing angles between linear subspaces. *Mathematics of computation*, 27(123):579–594, 1973.

Chan, K. H. R., Yu, Y., You, C., Qi, H., Wright, J., and Ma, Y. Redunet: A white-box deep network from the principle of maximizing rate reduction, 2022.

Chi, Y., Lu, Y. M., and Chen, Y. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

Chu, T., Tong, S., Ding, T., Dai, X., Haeffele, B. D., Vidal, R., and Ma, Y. Image clustering via the principle of rate reduction in the age of pretrained models. *arXiv preprint arXiv:2306.05272*, 2023.

Cover, T. M. *Elements of information theory*. John Wiley & Sons, 1999.

Dai, X., Tong, S., Li, M., Wu, Z., Psenka, M., Chan, K. H. R., Zhai, P., Yu, Y., Yuan, X., Shum, H.-Y., et al. Ctrl: Closed-loop transcription to an ldr via minimaxing rate reduction. *Entropy*, 24(4):456, 2022.

Daneshmand, H., Kohler, J., Lucchi, A., and Hofmann, T. Escaping saddles with stochastic gradients. In *International Conference on Machine Learning*, pp. 1155–1164. PMLR, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

Ge, R., Jin, C., and Zheng, Y. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pp. 1233–1242. PMLR, 2017.

Gregor, K. and LeCun, Y. Learning fast approximations of sparse coding. In *Proceedings of the 27th international conference on international conference on machine learning*, pp. 399–406, 2010.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *science*, 313 (5786):504–507, 2006.

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge university press, 2012.

Jiang, J., Zhou, J., Wang, P., Qu, Q., Mixon, D., You, C., and Zhu, Z. Generalized neural collapse for a large number of classes. In *International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=D4B7kkB89m.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In *International conference on machine learning*, pp. 1724–1732. PMLR, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436–444, 2015.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent only converges to minimizers. In *Conference on learning theory*, pp. 1246–1257. PMLR, 2016.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-order methods almost always avoid strict saddle points. *Mathematical programming*, 176:311–337, 2019.

Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. The expressive power of neural networks: A view from the width. *Advances in neural information processing systems*, 30, 2017.

Ma, Y., Derksen, H., Hong, W., and Wright, J. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE transactions on pattern analysis and machine intelligence*, 29(9):1546–1562, 2007.

Ma, Y., Tsao, D., and Shum, H.-Y. On the principles of parsimony and self-consistency for the emergence of intelligence. *Frontiers of Information Technology & Electronic Engineering*, 23(9):1298–1323, 2022.

Masarczyk, W., Ostaszewski, M., Imani, E., Pascanu, R., Miłoś, P., and Trzciński, T. The tunnel effect: Building data representations in deep neural networks. *arXiv preprint arXiv:2305.19753*, 2023.

Mixon, D. G., Parshall, H., and Pi, J. Neural collapse with unconstrained features. *arXiv preprint arXiv:2011.11619*, 2020.

Monga, V., Li, Y., and Eldar, Y. C. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2): 18–44, 2021.

Nocedal, J. and Wright, S. J. *Numerical optimization*. Springer, 1999.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Pai, D., Psenka, M., Chiu, C.-Y., Wu, M., Dobriban, E., and Ma, Y. Pursuit of a discriminative representation for multiple subspaces via sequential games. *arXiv preprint arXiv:2206.09120*, 2022.

Papyan, V., Han, X., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Qu, Q., Zhai, Y., Li, X., Zhang, Y., and Zhu, Z. Geometric analysis of nonconvex optimization landscapes for overcomplete learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygixkHKDH.

Recanatesi, S., Farrell, M., Advani, M., Moore, T., Lajoie, G., and Shea-Brown, E. Dimensionality compression and expansion in deep neural networks. *arXiv preprint arXiv:1906.00443*, 2019.

Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th international conference on international conference on machine learning*, pp. 833–840, 2011.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Sun, J., Qu, Q., and Wright, J. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

Sun, J., Qu, Q., and Wright, J. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.

Sun, J., Qu, Q., and Wright, J. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18:1131–1198, 2018.

Sun, R., Li, D., Liang, S., Ding, T., and Srikant, R. The global landscape of neural networks: An overview. *IEEE Signal Processing Magazine*, 37(5):95–108, 2020.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, P., Liu, H., Zhou, Z., and So, A. M.-C. Optimal non-convex exact recovery in stochastic block model via projected power method. In *International Conference on Machine Learning*, pp. 10828–10838. PMLR, 2021.

Wang, P., Liu, H., Yaras, C., Balzano, L., and Qu, Q. Linear convergence analysis of neural collapse with unconstrained features. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022a.

Wang, P., Zhou, Z., and So, A. M.-C. Non-convex exact community recovery in stochastic block model. *Mathematical Programming*, 195(1):1–37, 2022b.

Wang, P., Li, X., Yaras, C., Zhu, Z., Balzano, L., Hu, W., and Qu, Q. Understanding deep representation learning via layerwise feature compression and discrimination. *arXiv preprint arXiv:2311.02960*, 2023.

Xu, Y., Jin, R., and Yang, T. First-order stochastic algorithms for escaping from saddle points in almost linear time. *Advances in neural information processing systems*, 31, 2018.

Yaras, C., Wang, P., Zhu, Z., Balzano, L., and Qu, Q. Neural collapse with normalized features: A geometric analysis over the riemannian manifold. *Advances in neural information processing systems*, 35:11547–11560, 2022.

Yu, Y., Chan, K. H. R., You, C., Song, C., and Ma, Y. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *Advances in Neural Information Processing Systems*, 33:9422–9434, 2020.

Yu, Y., Buchanan, S., Pai, D., Chu, T., Wu, Z., Tong, S., Bai, H., Zhai, Y., Haeffele, B. D., and Ma, Y. White-box transformers via sparse rate reduction: Compression is all there is? *arXiv preprint arXiv:2311.13110*, 2023a.

Yu, Y., Buchanan, S., Pai, D., Chu, T., Wu, Z., Tong, S., Haeffele, B. D., and Ma, Y. White-box transformers via sparse rate reduction. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Yu, Y., Chu, T., Tong, S., Wu, Z., Pai, D., Buchanan, S., and Ma, Y. Emergence of segmentation with minimalistic white-box transformers. In *Conference on Parsimony and Learning*, volume 234 of *Proceedings of Machine Learning Research*, pp. 72–93. PMLR, 2024.

Zhai, Y., Yang, Z., Liao, Z., Wright, J., and Ma, Y. Complete dictionary learning via l4-norm maximization over the orthogonal group. *Journal of Machine Learning Research*, 21(165):1–68, 2020.

Zhang, Y., Qu, Q., and Wright, J. From symmetry to geometry: Tractable nonconvex problems. *arXiv preprint arXiv:2007.06753*, 2020.

Zhou, J., Li, X., Ding, T., You, C., Qu, Q., and Zhu, Z. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. In *International Conference on Machine Learning*, pp. 27179–27202. PMLR, 2022.

Zhu, Z., Ding, T., Zhou, J., Li, X., You, C., Sulam, J., and Qu, Q. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34, 2021.

Supplementary Material

The organization of the supplementary material is as follows: In Appendix A, we introduce preliminary setups and auxiliary results for studying the MCR^2 problem. Then, we prove the technical results concerning the global optimality of Problem (5) in Appendix B and the optimization landscape of Problem (5) in Appendix C, respectively. In Appendix D, we prove the main theorems in Theorem 3.1 and Theorem 3.3. Finally, we provide more experimental setups and results in Appendix E.

Besides the notions introduced earlier, we shall use $\mathrm{BlkDiag}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K)$ to denote the block diagonal matrix whose diagonal blocks are $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K$.

A. Preliminaries

In this section, we first introduce the first-order optimality condition and the concept of a strict saddle point for $F(\cdot)$ in Problem (5) in Section A.1, and finally present auxiliary results about matrix computations and properties of the log-determinant function in Section A.2. Recall that $\boldsymbol{Z} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_K] \in \mathbb{R}^{d\times m}$ with $\boldsymbol{Z}_k \in \mathbb{R}^{d\times m_k}$ for each $k \in [K]$, and α, α_k are defined in (11). To simplify our development, we write $R_c(\boldsymbol{Z}, \boldsymbol{\pi}_k)$ in (3) as

$$R_c(\boldsymbol{Z}, \boldsymbol{\pi}_k) := \frac{m_k}{m} R_c(\boldsymbol{Z}_k), \quad \text{where } R_c(\boldsymbol{Z}_k) := \frac{1}{2}\log\det\left(\boldsymbol{I} + \alpha_k \boldsymbol{Z}_k \boldsymbol{Z}_k^T\right). \tag{21}$$

Therefore, we can write $F(\boldsymbol{Z})$ in Problem (5) into

$$F(\boldsymbol{Z}) = R(\boldsymbol{Z}) - \sum_{k=1}^{K} \frac{m_k}{m} R_c(\boldsymbol{Z}_k) - \frac{\lambda}{2}\|\boldsymbol{Z}\|_F^2. \tag{22}$$

A.1. Optimality Conditions and Strict Saddle Points

To begin, we compute the gradient and Hessian (in bilinear form along a direction $\boldsymbol{D} \in \mathbb{R}^{d\times m}$) of $R(\cdot)$ in (2) as follows:

$$\nabla R(\boldsymbol{Z}) = \alpha \boldsymbol{X}^{-1}\boldsymbol{Z}, \tag{23}$$

$$\nabla^2 R(\boldsymbol{Z})[\boldsymbol{D}, \boldsymbol{D}] = \alpha\langle \boldsymbol{X}^{-1}, \boldsymbol{D}\boldsymbol{D}^T\rangle - \frac{\alpha^2}{2}\mathrm{Tr}\left(\boldsymbol{X}^{-1}(\boldsymbol{Z}\boldsymbol{D}^T + \boldsymbol{D}\boldsymbol{Z}^T)\boldsymbol{X}^{-1}(\boldsymbol{Z}\boldsymbol{D}^T + \boldsymbol{D}\boldsymbol{Z}^T)\right), \tag{24}$$

where $\boldsymbol{X} := \boldsymbol{I}_n + \alpha \boldsymbol{Z}\boldsymbol{Z}^T$ and α is defined in (11). Note that we can compute the gradient and Hessian of $R_c(\cdot)$ in (21) using the same approach. Based on the above setup, we define the first-order optimality condition of Problem (5) as follows.

Definition A.1. We say that $\boldsymbol{Z} \in \mathbb{R}^{d\times m}$ is a critical point of Problem (5) if $\nabla F(\boldsymbol{Z}) = \boldsymbol{0}$, i.e.,

$$\alpha(\boldsymbol{I} + \alpha \boldsymbol{Z}\boldsymbol{Z}^T)^{-1}\boldsymbol{Z}_k - \alpha(\boldsymbol{I} + \alpha_k \boldsymbol{Z}_k\boldsymbol{Z}_k^T)^{-1}\boldsymbol{Z}_k - \lambda \boldsymbol{Z}_k = \boldsymbol{0}, \ \forall k \in [K], \tag{25}$$

where α and α_k are defined in (11).

According to Jin et al. (2017); Lee et al. (2019), we define the strict saddle point, i.e., a critical point that has a direction with strictly positive curvature[7], of Problem (5) as follows:

Definition A.2. Suppose that $\boldsymbol{Z} \in \mathbb{R}^{d\times m}$ is a critical point of Problem (5). We say that $\boldsymbol{Z}$ is its strict saddle point if there exists a direction $\boldsymbol{D} = [\boldsymbol{D}_1, \ldots, \boldsymbol{D}_K] \in \mathbb{R}^{d\times m}$ with $\boldsymbol{D}_k \in \mathbb{R}^{d\times m_k}$ such that

$$\nabla^2 F(\boldsymbol{Z})[\boldsymbol{D}, \boldsymbol{D}] > 0,$$

where

$$\nabla^2 F(\boldsymbol{Z})[\boldsymbol{D}, \boldsymbol{D}] = \nabla^2 R(\boldsymbol{Z})[\boldsymbol{D}, \boldsymbol{D}] - \sum_{k=1}^{K} \frac{m_k}{m} \nabla^2 R_c(\boldsymbol{Z}_k)[\boldsymbol{D}_k, \boldsymbol{D}_k] - \lambda\|\boldsymbol{D}\|_F^2. \tag{26}$$

Remark that for the MCR^2 problem, strict saddle points include classical saddle points with strictly positive curvature as well as local minimizers.

[7] Note that Problem (5) is not a minimization problem but a maximization problem.

A.2. Auxiliary Results

We provide a matrix inversion lemma, which is also known as Sherman–Morrison–Woodbury formula.

Lemma A.3 (Matrix inversion lemma). *For any $\boldsymbol{Z} \in \mathbb{R}^{d\times m}$, we have*

$$(\boldsymbol{I} + \alpha \boldsymbol{Z}\boldsymbol{Z}^T)^{-1} = \boldsymbol{I} - \boldsymbol{Z}\left(\frac{1}{\alpha}\boldsymbol{I} + \boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T. \tag{27}$$

We next present a commutative property for the log-determinant function and the upper bound for the coding rate function. We refer the reader to (Chan et al., 2022, Lemma 8 & Lemma 10) for the detailed proofs. Here, let $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be a singular value decompositon (SVD) of $\boldsymbol{Z} \in \mathbb{R}^{d\times m}$, where $r = \operatorname{rank}(\boldsymbol{Z}) \le \min\{m, d\}$, $\boldsymbol{U} \in \mathcal{O}^{d\times r}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$ is a diagonal matrix, and $\boldsymbol{V} \in \mathcal{O}^{m\times r}$.

Lemma A.4 (Commutative property). *For any $\boldsymbol{Z} \in \mathbb{R}^{d\times m}$ and $\alpha > 0$, we have*

$$\frac{1}{2}\log\det\left(\boldsymbol{I} + \alpha\boldsymbol{Z}\boldsymbol{Z}^T\right) = \frac{1}{2}\log\det\left(\boldsymbol{I} + \alpha\boldsymbol{Z}^T\boldsymbol{Z}\right) = \frac{1}{2}\log\det\left(\boldsymbol{I} + \alpha\boldsymbol{\Sigma}^2\right). \tag{28}$$

Lemma A.5. *Let $\boldsymbol{Z} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_K] \in \mathbb{R}^{d\times m}$. Given $\alpha > 0$, it holds that*

$$\log\det\left(\boldsymbol{I} + \alpha\boldsymbol{Z}\boldsymbol{Z}^T\right) \le \sum_{k=1}^{K}\log\det\left(\boldsymbol{I} + \alpha\boldsymbol{Z}_k\boldsymbol{Z}_k^T\right), \tag{29}$$

where the equality holds if and only if $\boldsymbol{Z}_k^T\boldsymbol{Z}_l = \boldsymbol{0}$ for all $k \ne l \in [K]$.

Finally, we show that the objective function of Problem (5) is invariant under the block diagonal orthogonal matrices.

Lemma A.6. *For any $\boldsymbol{O} = \operatorname{BlkDiag}(\boldsymbol{O}_1, \ldots, \boldsymbol{O}_K)$, where $\boldsymbol{O}_k \in \mathcal{O}^{m_k}$ for each $k \in [K]$, we have*

$$F(\boldsymbol{Z}\boldsymbol{O}) = F(\boldsymbol{Z}), \quad \nabla F(\boldsymbol{Z}\boldsymbol{O}) = \nabla F(\boldsymbol{Z})\boldsymbol{O}, \quad \nabla^2 F(\boldsymbol{Z}\boldsymbol{O})[\boldsymbol{D}\boldsymbol{O}, \boldsymbol{D}\boldsymbol{O}] = \nabla^2 F(\boldsymbol{Z})[\boldsymbol{D}, \boldsymbol{D}]. \tag{30}$$

Proof of Lemma A.6. Let $\boldsymbol{O}_k \in \mathcal{O}^{m_k}$ be arbitrary for each $k \in [K]$ and $\boldsymbol{O} = \operatorname{BlkDiag}(\boldsymbol{O}_1, \ldots, \boldsymbol{O}_K)$. According to (2) and (21), we have $R(\boldsymbol{Z}\boldsymbol{O}) = R(\boldsymbol{Z})$ and $R_c(\boldsymbol{Z}_k\boldsymbol{O}_k) = R(\boldsymbol{Z}_k)$. This, together with (5), yields that $F(\boldsymbol{Z}\boldsymbol{O}) = F(\boldsymbol{Z})$. Moreover, it follows from (23) that $\nabla R(\boldsymbol{Z}\boldsymbol{O}) = \nabla R(\boldsymbol{Z})\boldsymbol{O}$ and $\nabla R_c(\boldsymbol{Z}_k\boldsymbol{O}_k) = \nabla R_c(\boldsymbol{Z}_k)\boldsymbol{O}_k$. This implies $\nabla F(\boldsymbol{Z}\boldsymbol{O}) = \nabla F(\boldsymbol{Z})$. Finally, using (24), we have $\nabla^2 R(\boldsymbol{Z}\boldsymbol{O})[\boldsymbol{D}\boldsymbol{O}, \boldsymbol{D}\boldsymbol{O}] = \nabla^2 R(\boldsymbol{Z})[\boldsymbol{D}, \boldsymbol{D}]$ and $\nabla^2 R_c(\boldsymbol{Z}_k\boldsymbol{O}_k)[\boldsymbol{D}_k\boldsymbol{O}_k, \boldsymbol{D}_k\boldsymbol{O}_k] = \nabla^2 R(\boldsymbol{Z}_k)[\boldsymbol{D}_k, \boldsymbol{D}_k]$. This, together with (26), implies $\nabla^2 F(\boldsymbol{Z}\boldsymbol{O})[\boldsymbol{D}\boldsymbol{O}, \boldsymbol{D}\boldsymbol{O}] = \nabla^2 F(\boldsymbol{Z})[\boldsymbol{D}, \boldsymbol{D}]$. □

B. Proofs in Section 4.1

B.1. Proof of Lemma 4.1

Proof of Lemma 4.1. It follows from Lemma A.5 that

$$\log\det\left(\boldsymbol{I}_d + \alpha\boldsymbol{Z}\boldsymbol{Z}^T\right) \le \sum_{k=1}^{K}\log\det\left(\boldsymbol{I}_d + \alpha\boldsymbol{Z}_k\boldsymbol{Z}_k^T\right), \tag{31}$$

where the equality holds if and only if $\boldsymbol{Z}_k^T\boldsymbol{Z}_l = \boldsymbol{0}$ for all $1 \le k \ne l \le K$. Substituting this into (5) directly yields (12). □

B.2. Proof of Proposition 4.2

Proof of Proposition 4.2. Let $\boldsymbol{Z} \in \mathcal{Z}$ be arbitrary, where $\mathcal{Z}$ is defined in (13). It follows from $\boldsymbol{Z} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_K] \in \mathbb{R}^{d\times m}$ that $\sum_{k=1}^{K} r_k \le d$. According to Lemma 4.1 and $\boldsymbol{Z}_k^T\boldsymbol{Z}_l = \boldsymbol{0}$ for all $k \ne l$ due to $\boldsymbol{Z} \in \mathcal{Z}$, we have $F(\boldsymbol{Z}) = \sum_{k=1}^{K} f_k(\boldsymbol{Z}_k)$, where $f_k : \mathbb{R}^{d\times m_k} \to \mathbb{R}$ takes the form of

$$f_k(\boldsymbol{Z}_k) := \frac{1}{2}\log\det\left(\boldsymbol{I}_d + \alpha\boldsymbol{Z}_k\boldsymbol{Z}_k^T\right) - \frac{m_k}{2m}\log\det\left(\boldsymbol{I}_d + \alpha_k\boldsymbol{Z}_k\boldsymbol{Z}_k^T\right) - \frac{\lambda}{2}\|\boldsymbol{Z}_k\|_F^2. \tag{32}$$

This, together with (23), yields that $\nabla F(\boldsymbol{Z}) = \boldsymbol{0}$ is equivalent to

$$\alpha \left(\boldsymbol{I}_d + \alpha \boldsymbol{Z}_k \boldsymbol{Z}_k^T\right)^{-1} \boldsymbol{Z}_k - \alpha \left(\boldsymbol{I}_d + \alpha_k \boldsymbol{Z}_k \boldsymbol{Z}_k^T\right)^{-1} \boldsymbol{Z}_k = \lambda \boldsymbol{Z}_k, \ \forall k \in [K]. \tag{33}$$

Obviously, $\boldsymbol{Z}_k = \boldsymbol{0}$ is a solution of the above equation for each $k \in [K]$, which satisfies $\boldsymbol{Z}_k^T \boldsymbol{Z}_l = \boldsymbol{0}$ for all $l \neq k$. Now, we consider $\boldsymbol{Z}_k \neq \boldsymbol{0}$, and thus $1 \leq r_k = \operatorname{rank}(\boldsymbol{Z}_k) \leq \min\{m_k, d\}$. Let

$$\boldsymbol{Z}_k = \boldsymbol{P}_k \boldsymbol{\Sigma}_k \boldsymbol{Q}_k^T = \begin{bmatrix} \boldsymbol{P}_{k,1} & \boldsymbol{P}_{k,2} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{Q}_{k,1}^T \\ \boldsymbol{Q}_{k,2}^T \end{bmatrix} \tag{34}$$

be a singular value decomposition (SVD) of $\boldsymbol{Z}_k \in \mathbb{R}^{d \times m_k}$, where $\tilde{\boldsymbol{\Sigma}}_k = \operatorname{diag}(\sigma_{k,1}, \ldots, \sigma_{k,r_k})$ with $\sigma_{k,1} \geq \cdots \geq \sigma_{k,r_k} > 0$ being positive singular values of $\boldsymbol{Z}_k$, $\boldsymbol{P}_k \in \mathcal{O}^d$ with $\boldsymbol{P}_{k,1} \in \mathbb{R}^{d \times r_k}$ and $\boldsymbol{P}_{k,2} \in \mathbb{R}^{d \times (d - r_k)}$, and $\boldsymbol{Q}_k \in \mathcal{O}^{m_k}$ with $\boldsymbol{Q}_{k,1} \in \mathbb{R}^{m_k \times r_k}$ and $\boldsymbol{Q}_{k,2} \in \mathbb{R}^{m_k \times (m_k - r_k)}$. Substituting this SVD into (33) yields for all $k \in [K]$,

$$\alpha \boldsymbol{P}_k (\boldsymbol{I}_d + \alpha \boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_k^T)^{-1} \boldsymbol{\Sigma}_k \boldsymbol{Q}_k^T - \alpha \boldsymbol{P}_k (\boldsymbol{I}_d + \alpha_k \boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_k^T)^{-1} \boldsymbol{\Sigma}_k \boldsymbol{Q}_k^T = \lambda \boldsymbol{P}_k \boldsymbol{\Sigma}_k \boldsymbol{Q}_k^T,$$

which is equivalent to

$$\alpha (\boldsymbol{I}_d + \alpha \boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_k^T)^{-1} \boldsymbol{\Sigma}_k - \alpha (\boldsymbol{I}_d + \alpha_k \boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_k^T)^{-1} \boldsymbol{\Sigma}_k = \lambda \boldsymbol{\Sigma}_k.$$

Using $\boldsymbol{\Sigma}_k = \operatorname{BlkDiag}(\tilde{\boldsymbol{\Sigma}}_k, \boldsymbol{0})$, we further obtain

$$\alpha (\boldsymbol{I}_{r_k} + \alpha \tilde{\boldsymbol{\Sigma}}_k^2)^{-1} \tilde{\boldsymbol{\Sigma}}_k - \alpha (\boldsymbol{I}_{r_k} + \alpha_k \tilde{\boldsymbol{\Sigma}}_k^2)^{-1} \tilde{\boldsymbol{\Sigma}}_k = \lambda \tilde{\boldsymbol{\Sigma}}_k.$$

Since $\tilde{\boldsymbol{\Sigma}}_k$ is a diagonal matrix with diagonal entries being positive, we have for all $k \in [K]$,

$$(\boldsymbol{I}_{r_k} + \alpha \tilde{\boldsymbol{\Sigma}}_k^2)^{-1} - (\boldsymbol{I}_{r_k} + \alpha_k \tilde{\boldsymbol{\Sigma}}_k^2)^{-1} = \frac{\lambda}{\alpha} \boldsymbol{I}_{r_k}. \tag{35}$$

This implies for each $i \in [r_k]$ and $k \in [K]$,

$$\frac{1}{1 + \alpha \sigma_{k,i}^2} - \frac{1}{1 + \alpha_k \sigma_{k,i}^2} = \frac{\lambda}{\alpha}. \tag{36}$$

Therefore, we obtain that $\sigma_{k,i}^2 > 0$ for each $i \in [r_k]$ is a positive root of the following quadratic equation with a variable $x \in \mathbb{R}$:

$$\lambda \alpha_k x^2 - \eta_k x + \frac{\lambda}{\alpha} = 0,$$

where

$$\eta_k := (\alpha_k - \alpha) - \lambda \left(1 + \frac{\alpha_k}{\alpha}\right), \ \forall k \in [K]. \tag{37}$$

According to (6), one can verify that for each $k \in [K]$,

$$\eta_k > 0, \quad \eta_k^2 - \frac{4\alpha_k}{\alpha} \lambda^2 \geq 0.$$

This yields that the above quadratic equation has positive roots as follows. For each $i \in [r_k]$ and $k \in [K]$, we have

$$\sigma_{k,i}^2 = \frac{\eta_k \pm \sqrt{\eta_k^2 - 4\lambda^2 m / m_k}}{2\lambda \alpha_k}. \tag{38}$$

Finally, using $\boldsymbol{Z}_k^T \boldsymbol{Z}_l = \boldsymbol{0}$ and (34), we obtain $\boldsymbol{Q}_{k,1}^T \boldsymbol{Q}_{l,1} = \boldsymbol{0}$ for all $1 \leq l \neq k \leq K$. These, together with (34), yields (14).

Conversely, suppose that each block $\boldsymbol{Z}_k$ of $\boldsymbol{Z}$ satisfies $\boldsymbol{Z}_k = \boldsymbol{0}$ or takes the form (14) for some $\boldsymbol{U}_k \in \mathcal{O}^{d \times r_k}$ satisfying $\boldsymbol{U}_k^T \boldsymbol{U}_l = \boldsymbol{0}$ for all $l \neq k$, $\boldsymbol{V}_k \in \mathcal{O}^{m_k \times r_k}$ for all $k \in [K]$, and $\sigma_{k,i} > 0$ satisfying (15). We are devoted to showing $\boldsymbol{Z} \in \mathcal{Z}$. It is straightforward to verify that $\boldsymbol{Z}_k^T \boldsymbol{Z}_l = \boldsymbol{0}$ for all $1 \leq k \neq l \leq K$. This, together with Lemma 4.1, implies

$F(\boldsymbol{Z}) = \sum_{k=1}^K f_k(\boldsymbol{Z}_k)$. Therefore, it suffices to verify that $\nabla f_k(\boldsymbol{Z}_k) = \boldsymbol{0}$ for each $k \in [K]$ in the rest of the proof. For each $k \in [K]$, if $\boldsymbol{Z}_k = \boldsymbol{0}$, it is obvious to verify $\nabla f(\boldsymbol{Z}_k) = \boldsymbol{0}$. Otherwise, $\boldsymbol{Z}_k$ takes the form (14) for some $\boldsymbol{U}_k \in \mathcal{O}^{d\times r_k}$, $\tilde{\boldsymbol{\Sigma}}_k \in \mathbb{R}^{r_k\times r_k}$ satisfying (15), and $\boldsymbol{V}_k \in \mathcal{O}^{m_k\times r_k}$, where $r_k \geq 1$. Now, we compute for all $i \in [r_k]$,

$$\begin{aligned}\frac{\sigma_{k,i}^2}{1/\alpha_k + \sigma_{k,i}^2} - \frac{\sigma_{k,i}^2}{1/\alpha + \sigma_{k,i}^2} &= \frac{\alpha_k\sigma_{k,i}^2}{1+\alpha_k\sigma_{k,i}^2} - \frac{\alpha\sigma_{k,i}^2}{1+\alpha\sigma_{k,i}^2} = \frac{(\alpha_k-\alpha)\sigma_{k,i}^2}{\left(1+\alpha_k\sigma_{k,i}^2\right)\left(1+\alpha\sigma_{k,i}^2\right)} \\ &= \frac{1}{1+\alpha\sigma_{k,i}^2} - \frac{1}{1+\alpha_k\sigma_{k,i}^2} = \frac{\lambda}{\alpha},\end{aligned} \tag{39}$$

where the last equality is due to (15), (16), (17), and (36). Then, we compute

$$\left(\boldsymbol{I}_d + \alpha\boldsymbol{Z}_k\boldsymbol{Z}_k^T\right)^{-1} = \left(\boldsymbol{I}_d + \alpha\boldsymbol{U}_k\tilde{\boldsymbol{\Sigma}}_k^2\boldsymbol{U}_k^T\right)^{-1} = \boldsymbol{I}_d - \boldsymbol{U}_k\tilde{\boldsymbol{\Sigma}}_k\left(\frac{1}{\alpha}\boldsymbol{I}_{r_k} + \tilde{\boldsymbol{\Sigma}}_k^2\right)^{-1}\tilde{\boldsymbol{\Sigma}}_k\boldsymbol{U}_k^T. \tag{40}$$

where the second equality follows from (27). This, together with (23), yields

$$\begin{aligned}\nabla f_k(\boldsymbol{Z}_k) &= \alpha\left(\boldsymbol{I}_d + \alpha\boldsymbol{Z}_k\boldsymbol{Z}_k^T\right)^{-1}\boldsymbol{Z}_k - \alpha\left(\boldsymbol{I}_d + \alpha_k\boldsymbol{Z}_k\boldsymbol{Z}_k^T\right)^{-1}\boldsymbol{Z}_k - \lambda\boldsymbol{Z}_k \\ &= \alpha\boldsymbol{U}_k\tilde{\boldsymbol{\Sigma}}_k\left(\left(\frac{1}{\alpha_k}\boldsymbol{I}_{r_k} + \tilde{\boldsymbol{\Sigma}}_k^2\right)^{-1} - \left(\frac{1}{\alpha}\boldsymbol{I}_{r_k} + \tilde{\boldsymbol{\Sigma}}_k^2\right)^{-1}\right)\tilde{\boldsymbol{\Sigma}}_k^2\boldsymbol{V}_k^T - \lambda\boldsymbol{Z}_k = \boldsymbol{0},\end{aligned}$$

where the last equality follows from (14) and (39). Therefore, we have $\nabla F(\boldsymbol{Z}) = \boldsymbol{0}$ as desired. This, together with $\boldsymbol{Z}_k^T\boldsymbol{Z}_l = \boldsymbol{0}$, implies $\boldsymbol{Z} \in \mathcal{Z}$. □

C. Proofs in Section 4.2

C.1. Proof of Proposition 4.3

Proof of Proposition 4.3. For each $\boldsymbol{Z} \in \mathcal{Z}$, it follows from Lemma 4.1 that

$$F(\boldsymbol{Z}) = \sum_{k=1}^K f_k(\boldsymbol{Z}_k), \tag{41}$$

where f_k is defined in (32). Suppose that there exists $k \in [K]$ such that $r_k = 0$, i.e., $\boldsymbol{Z}_k = \boldsymbol{0}$. According to (24) and (32), we compute for any $\boldsymbol{D}_k \neq \boldsymbol{0}$,

$$\nabla f_k(\boldsymbol{Z}_k)[\boldsymbol{D}_k, \boldsymbol{D}_k] = \left(\frac{\alpha}{2} - \frac{m_k}{2m}\alpha_k - \lambda\right)\|\boldsymbol{D}_k\|_F^2 = -\lambda\|\boldsymbol{D}_k\|_F^2 < 0,$$

where the second equality follows from $m_k\alpha_k/m = \alpha$ according to (11). This implies $\boldsymbol{0}$ is a local maximizer of $f_k(\boldsymbol{Z}_k)$. Suppose to the contrary that $r_k > 0$ for all $k \in [K]$. For each $\boldsymbol{Z} \in \mathcal{Z}$, using Lemma 4.1 with $\boldsymbol{Z}_k^T\boldsymbol{Z}_l = \boldsymbol{0}$ for all $k \neq l$, (14), and (32), we have

$$\begin{aligned}F(\boldsymbol{Z}) &= \sum_{k=1}^K\left(\frac{1}{2}\log\det\left(\boldsymbol{I}_n + \alpha\boldsymbol{U}_k\tilde{\boldsymbol{\Sigma}}_k^2\boldsymbol{U}_k^T\right) - \frac{m_k}{2m}\log\det\left(\boldsymbol{I}_n + \alpha_k\boldsymbol{U}_k\tilde{\boldsymbol{\Sigma}}_k^2\boldsymbol{U}_k^T\right) - \frac{\lambda}{2}\|\boldsymbol{Z}_k\|_F^2\right) \\ &= \sum_{k=1}^K\left(\frac{1}{2}\log\det\left(\boldsymbol{I} + \alpha\tilde{\boldsymbol{\Sigma}}_k^2\right) - \frac{m_k}{2m}\log\det\left(\boldsymbol{I} + \alpha_k\tilde{\boldsymbol{\Sigma}}_k^2\right) - \frac{\lambda}{2}\|\tilde{\boldsymbol{\Sigma}}_k\|_F^2\right) \\ &= \frac{1}{2}\sum_{k=1}^K\sum_{i=1}^{r_k}\left(\log\left(1+\alpha\sigma_{k,i}^2\right) - \frac{m_k}{m}\log\left(1+\alpha_k\sigma_{k,i}^2\right) - \lambda\sigma_{k,i}^2\right),\end{aligned} \tag{42}$$

where the second equality is due to (14) and Lemma A.4. For ease of exposition, let

$$h_k(x) = \log\left(1+\alpha x\right) - \frac{m_k}{m}\log\left(1+\alpha_k x\right) - \lambda x,\ \forall k \in [k]. \tag{43}$$

Using (15), (36), (37), and (38), one can verify that $h_k'(x) \leq 0$ for $x \in (0, \underline{\sigma}_k)$, $h_k'(x) \geq 0$ for $x \in [\underline{\sigma}_k, \overline{\sigma}_k)$, and $h_k'(x) \leq 0$ for $x \in [\overline{\sigma}_k, \infty)$ for all $k \in [K]$. This yields that $h_k(\underline{\sigma}_k)$ is a local minimizer and $h(\overline{\sigma}_k)$ is a local maximizer. This, together with (42) and the fact that $\boldsymbol{0}$ is a local maximizer of $f_k(\boldsymbol{Z}_k)$, implies (i) and (ii).

□

C.2. Proof of Proposition 4.4

Proof of Proposition 4.4. Note that $\boldsymbol{Z} \in \mathbb{R}^{d\times m}$ is a critical point that satisfies (25). Suppose that $\operatorname{rank}(\boldsymbol{Z}) = r$ and $\operatorname{rank}(\boldsymbol{Z}_k) = r_k$ for all $k \in [K]$. Obviously, we have $r_k \le \min\{m_k, d\}$ for all $k \in [K]$ and $\sum_{k=1}^K r_k \le r \le \min\{m, d\}$. Now, let $\boldsymbol{Z}\boldsymbol{Z}^T = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$ be an eigenvalue decomposition of $\boldsymbol{Z}\boldsymbol{Z}^T \in \mathbb{S}_+^d$, where $\boldsymbol{Q} \in \mathcal{O}^{d\times r}$ and $\boldsymbol{\Lambda} \in \mathbb{R}^{r\times r}$ is a diagonal matrix with diagonal entries being positive eigenvalues of $\boldsymbol{Z}\boldsymbol{Z}^T$. Suppose that $\boldsymbol{Z}\boldsymbol{Z}^T$ has p distinct positive eigenvalues, where $1 \le p \le r$. Let $\lambda_1 > \cdots > \lambda_p > 0$ be its distinct eigenvalue values with the corresponding multiplicities being $h_1, \ldots, h_p \in \mathbb{N}_+$, respectively. Obviously, we have $\sum_{i=1}^p h_i = r$. Therefore, we write

$$\boldsymbol{\Lambda} = \operatorname{BlkDiag}\left(\lambda_1 \boldsymbol{I}_{h_1}, \ldots, \lambda_p \boldsymbol{I}_{h_p}\right), \ \boldsymbol{Q} = \left[\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_p\right], \tag{44}$$

where $\boldsymbol{Q}_i \in \mathcal{O}^{d\times h_i}$ for all $i \in [p]$.

According to Lemma A.6, we can see that $\boldsymbol{Z}$ is a critical point with curvature if and only if $\boldsymbol{Z}\boldsymbol{O}$ is a critical point with the same curvature for each $\boldsymbol{O} = \operatorname{BlkDiag}(\boldsymbol{O}_1, \ldots, \boldsymbol{O}_K)$ with $\boldsymbol{O}_k \in \mathcal{O}^{m_k}$ for all $k \in [K]$. According to the SVD of $\boldsymbol{Z}_k$ in (34), we can take $\boldsymbol{O}_k = \boldsymbol{Q}_k$ for each $k \in [K]$. Therefore, it suffices to study $\boldsymbol{Z}_k = \boldsymbol{P}_k\boldsymbol{\Sigma}_k$ for each $k \in [K]$. Substituting this into (25) in Definition A.1 gives

$$\alpha(\boldsymbol{I} + \alpha\boldsymbol{Z}\boldsymbol{Z}^T)^{-1}\boldsymbol{P}_k\boldsymbol{\Sigma}_k - \alpha\boldsymbol{P}_k(\boldsymbol{I} + \alpha_k\boldsymbol{\Sigma}_k\boldsymbol{\Sigma}_k^T)^{-1}\boldsymbol{\Sigma}_k - \lambda\boldsymbol{P}_k\boldsymbol{\Sigma}_k = \boldsymbol{0}, \ \forall k \in [K].$$

This is equivalent to

$$\alpha(\boldsymbol{I} + \alpha\boldsymbol{Z}\boldsymbol{Z}^T)^{-1}\boldsymbol{Z}_k = \boldsymbol{Z}_k\left(\alpha(\boldsymbol{I} + \alpha_k\boldsymbol{\Sigma}_k\boldsymbol{\Sigma}_k^T)^{-1} + \lambda\boldsymbol{I}\right), \ \forall k \in [K].$$

This yields that each column of $\boldsymbol{Z}_k$ is an eigenvector of $\boldsymbol{Z}$ for each $k \in [K]$. This, together with the decomposition in (44), yields that we can permute the columns of $\boldsymbol{Z}_k$ such that the columns belonging to the space spanned by $\boldsymbol{Q}_i$ are rearranged together. Let $s_{k,i} \in \mathbb{N}$ denote the number of columns of $\boldsymbol{Z}_k$ that belong to the space spanned by $\boldsymbol{Q}_i$ for each $i \in [p]$. Obviously, we have $\sum_{i=1}^p s_{k,i} = m_k$. Consequently, for each $k \in [K]$, there exists an a column permutation matrix $\boldsymbol{\Pi}_k \in \mathbb{R}^{m_k\times m_k}$ such that

$$\boldsymbol{Z}_k\boldsymbol{\Pi}_k = \left[\boldsymbol{Z}_k^{(1)} \quad \ldots \quad \boldsymbol{Z}_k^{(p)}\right]. \tag{45}$$

where $\boldsymbol{Q}_i\boldsymbol{Q}_i^T\boldsymbol{Z}_k^{(i)} = \boldsymbol{Z}_k^{(i)} \in \mathbb{R}^{d\times s_{k,i}}$. Since $\boldsymbol{Q}_i^T\boldsymbol{Q}_j = \boldsymbol{0}$, we have ${\boldsymbol{Z}_k^{(i)}}^T\boldsymbol{Z}_k^{(j)} = \boldsymbol{0}$ for all $i \ne j$. This, together with (21) and Lemma A.5, yields

$$R_c(\boldsymbol{Z}_k) = \frac{m_k}{2m}\sum_{i=1}^p \log\det\left(\boldsymbol{I}_n + \alpha_k\boldsymbol{Z}_k^{(i)}{\boldsymbol{Z}_k^{(i)}}^T\right). \tag{46}$$

Moreover, let $s_i := \sum_{k=1}^K s_{k,i}$ and

$$\boldsymbol{Z}^{(i)} := \left[\boldsymbol{Z}_1^{(i)} \ \ldots \ \boldsymbol{Z}_K^{(i)}\right] \in \mathbb{R}^{d\times s_i}, \ \forall i \in [p]. \tag{47}$$

Using this and (45), we have

$$\boldsymbol{Z}\boldsymbol{Z}^T = \sum_{k=1}^K \boldsymbol{Z}_k\boldsymbol{Z}_k^T = \sum_{k=1}^K\sum_{i=1}^p \boldsymbol{Z}_k^{(i)}{\boldsymbol{Z}_k^{(i)}}^T = \sum_{i=1}^p \boldsymbol{Z}^{(i)}{\boldsymbol{Z}^{(i)}}^T.$$

This, together with (2), Lemma A.5, and ${\boldsymbol{Z}^{(i)}}^T\boldsymbol{Z}^{(j)} = \boldsymbol{0}$, yields that

$$R(\boldsymbol{Z}) = \frac{1}{2}\sum_{i=1}^p \log\det\left(\boldsymbol{I} + \alpha\boldsymbol{Z}^{(i)}{\boldsymbol{Z}^{(i)}}^T\right). \tag{48}$$

Characterize the structure of critical points. Now, for each $k \in [K]$ and $i \in [p]$, let $r_{k,i} = \mathrm{rank}(\boldsymbol{Z}_k^{(i)})$, where $r_{k,i} \leq \min\{d, s_{k,i}\}$. Moreover, let

$$\boldsymbol{Z}_k^{(i)} = \boldsymbol{U}_k^{(i)} \boldsymbol{\Sigma}_k^{(i)} \boldsymbol{V}_k^{(i)^T} = \begin{bmatrix} \boldsymbol{U}_{k,1}^{(i)} & \boldsymbol{U}_{k,2}^{(i)} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_k^{(i)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_{k,1}^{(i)^T} \\ \boldsymbol{V}_{k,2}^{(i)^T} \end{bmatrix} \tag{49}$$

be a singular value decomposition (SVD) of $\boldsymbol{Z}_k^{(i)}$, where $\tilde{\boldsymbol{\Sigma}}_k^{(i)} \in \mathbb{R}^{r_{k,i} \times r_{k,i}}$ is a diagonal matrix with diagonal entries being positive singular values of $\boldsymbol{Z}_k^{(i)}$; $\boldsymbol{U}_k^{(i)} \in \mathcal{O}^d$ with $\boldsymbol{U}_{k,1}^{(i)} \in \mathbb{R}^{d \times r_{k,i}}$ and $\boldsymbol{U}_{k,2}^{(i)} \in \mathbb{R}^{d \times (d - r_{k,i})}$; $\boldsymbol{V}_k^{(i)} \in \mathcal{O}^{s_{k,i}}$ with $\boldsymbol{V}_{k,1}^{(i)} \in \mathbb{R}^{s_{k,i} \times r_{k,i}}$ and $\boldsymbol{V}_{k,2}^{(i)} \in \mathbb{R}^{s_{k,i} \times (s_{k,i} - r_{k,i})}$. This, together with $\boldsymbol{Q}_i \boldsymbol{Q}_i^T \boldsymbol{Z}_k^{(i)} = \boldsymbol{Z}_k^{(i)}$, implies for all $k \in [K]$ and $i \in [p]$,

$$\boldsymbol{Q}_i \boldsymbol{Q}_i^T \boldsymbol{U}_{k,1}^{(i)} = \boldsymbol{U}_{k,1}^{(i)}. \tag{50}$$

According to (5), (23), (45), and (46), we have for all $k \in [K]$ and $i \in [p]$,

$$\alpha \boldsymbol{X}^{-1} \boldsymbol{Z}_k^{(i)} - \alpha \left(\boldsymbol{I} + \alpha_k \boldsymbol{Z}_k^{(i)} \boldsymbol{Z}_k^{(i)^T} \right)^{-1} \boldsymbol{Z}_k^{(i)} = \lambda \boldsymbol{Z}_k^{(i)}. \tag{51}$$

Substituting the block forms of $\boldsymbol{U}_k^{(i)}$ and $\boldsymbol{\Sigma}_k^{(i)}$ in (49) into the above equation and rearranging the terms, we obtain for all $k \in [K]$ and $i \in [p]$,

$$\boldsymbol{X}^{-1} \boldsymbol{U}_{k,1}^{(i)} - \boldsymbol{U}_{k,1}^{(i)} \left(\boldsymbol{I} + \alpha_k \tilde{\boldsymbol{\Sigma}}_k^{(i)^2} \right)^{-1} = \frac{\lambda}{\alpha} \boldsymbol{U}_{k,1}^{(i)}.$$

Using $\boldsymbol{X} = \boldsymbol{I} + \alpha \boldsymbol{Z} \boldsymbol{Z}^T$ and rearranging the terms, we have for all $k \in [K]$ and $i \in [p]$,

$$\boldsymbol{U}_{k,1}^{(i)} \left(\left(1 - \frac{\lambda}{\alpha} \right) \boldsymbol{I} - \left(\boldsymbol{I} + \alpha_k \tilde{\boldsymbol{\Sigma}}_k^{(i)^2} \right)^{-1} \right) = \alpha \boldsymbol{Z} \boldsymbol{Z}^T \boldsymbol{U}_{k,1}^{(i)} \left(\left(\boldsymbol{I} + \alpha_k \tilde{\boldsymbol{\Sigma}}_k^{(i)^2} \right)^{-1} + \frac{\lambda}{\alpha} \boldsymbol{I} \right). \tag{52}$$

Since $\boldsymbol{Z} \boldsymbol{Z}^T = \sum_{i=1}^p \lambda_i \boldsymbol{Q}_i \boldsymbol{Q}_i^T$, $\boldsymbol{Q}_i^T \boldsymbol{Q}_j = \boldsymbol{0}$ for all $i \neq j$, and (50), we have for all $k \in [K]$ and $i \in [p]$,

$$\boldsymbol{U}_{k,1}^{(i)} \left(\left(1 - \frac{\lambda}{\alpha} \right) \boldsymbol{I} - \left(\boldsymbol{I} + \alpha_k \tilde{\boldsymbol{\Sigma}}_k^{(i)^2} \right)^{-1} \right) = \alpha \lambda_i \boldsymbol{U}_{k,1}^{(i)} \left(\left(\boldsymbol{I} + \alpha_k \tilde{\boldsymbol{\Sigma}}_k^{(i)^2} \right)^{-1} + \frac{\lambda}{\alpha} \boldsymbol{I} \right).$$

Rearranging the terms in the above equation, we obtain for each $k \in [K]$ and $i \in [p]$,

$$\tilde{\boldsymbol{\Sigma}}_k^{(i)} = \beta_i \boldsymbol{I}, \text{ where } \beta_i := \frac{1}{\sqrt{\alpha K}} \sqrt{\frac{1 + \alpha \lambda_i}{1 - \lambda/\alpha - \lambda \lambda_i} - 1}. \tag{53}$$

Substituting this back to (52) yields for each $k \in [K]$ and $i \in [p]$,

$$\lambda_i \boldsymbol{U}_{k,1}^{(i)} = \boldsymbol{Z} \boldsymbol{Z}^T \boldsymbol{U}_{k,1}^{(i)}.$$

This, together with (49) and (53), yields $\lambda_i \boldsymbol{Z}_k^{(i)} = \boldsymbol{Z} \boldsymbol{Z}^T \boldsymbol{Z}_k^{(i)}$ for all $k \in [K]$ and $i \in [p]$. Using this and $\boldsymbol{Z}_k^{(i)^T} \boldsymbol{Z}_k^{(j)} = \boldsymbol{0}$ for all $i \neq j$, we have for all $i \in [p]$ and $k \in [K]$,

$$\lambda_i \boldsymbol{Z}_k^{(i)} = \sum_{l=1}^K \boldsymbol{Z}_l^{(i)} \boldsymbol{Z}_l^{(i)^T} \boldsymbol{Z}_k^{(i)}.$$

It follows from this and (47) that

$$\lambda_i \boldsymbol{Z}^{(i)} = \boldsymbol{Z}^{(i)} \boldsymbol{Z}^{(i)^T} \boldsymbol{Z}^{(i)}. \tag{54}$$

Since there exists $k \neq l \in [K]$ such that $\boldsymbol{Z}_k^T \boldsymbol{Z}_l \neq \boldsymbol{0}$, we can assume without loss of generality that $\boldsymbol{Z}_1^T \boldsymbol{Z}_2 \neq \boldsymbol{0}$. Then, there exist $i_1 \in [m_1]$ and $i_2 \in [m_2]$ such that $\boldsymbol{z}_{1,i_1}^T \boldsymbol{z}_{2,i_2} \neq 0$. This, together with $\boldsymbol{Z}^{(i)^T} \boldsymbol{Z}^{(j)} = \boldsymbol{0}$ for all $i \neq j$, implies that there exists $i^* \in [p]$ such that $\boldsymbol{z}_{1,i_1}, \boldsymbol{z}_{2,i_2}$ are both columns of $\boldsymbol{Z}^{(i^*)}$. Without loss of generality, suppose that $\boldsymbol{z}_{1,i_1}$ and $\boldsymbol{z}_{2,i_2}$ are

the u-th and v-th columns of $\boldsymbol{Z}^{(i^*)}$, respectively. Therefore, we have $\boldsymbol{z}_u^{(i^*)^T}\boldsymbol{z}_v^{(i^*)} \neq 0$. Using this, $\boldsymbol{z}_u^{(i^*)^T}\boldsymbol{z}_v^{(i^*)} \neq 0$, and (54), we have

$$\lambda_{i^*}\boldsymbol{z}_u^{(i^*)} = \boldsymbol{Z}^{(i^*)}\boldsymbol{Z}^{(i^*)^T}\boldsymbol{z}_u^{(i^*)}. \tag{55}$$

This is equivalent to

$$\sum_{j\neq u}^{s_{i^*}} \boldsymbol{z}_j^{(i^*)^T}\boldsymbol{z}_u^{(i^*)}\boldsymbol{z}_j^{(i^*)} + \left(\|\boldsymbol{z}_u^{(i^*)}\|^2 - \lambda\right)\boldsymbol{z}_u^{(i^*)} = \boldsymbol{0} \tag{56}$$

This, together with $\boldsymbol{z}_u^{(i^*)^T}\boldsymbol{z}_v^{(i^*)} \neq 0$, implies that the columns of $\boldsymbol{Z}^{(i^*)}$ are linearly dependent. By letting $t_{i^*} = \operatorname{rank}(\boldsymbol{Z}^{(i^*)})$, we have $t_{i^*} < s_{i^*}$ due to linear dependence of columns of $\boldsymbol{Z}^{(i^*)}$. Then, let $\boldsymbol{Z}^{(i^*)} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be an SVD of $\boldsymbol{Z}^{(i^*)}$, where $\boldsymbol{U} \in \mathcal{O}^{d\times t_{i^*}}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{t_{i^*}\times t_{i^*}}$, and $\boldsymbol{V} \in \mathcal{O}^{s_{i^*}\times t_{i^*}}$. Substituting this into (54) yields $\lambda_{i^*}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^3$, which implies $\boldsymbol{\Sigma} = \sqrt{\lambda_{i^*}}\boldsymbol{I}$ and

$$\boldsymbol{Z}^{(i^*)} = \sqrt{\lambda_{i^*}}\boldsymbol{U}\boldsymbol{V}^T. \tag{57}$$

Construct an ascent direction. For ease of exposition, we simply write i^* as i from now on. According to (47) and (55), we have

$$\sum_{k=1}^{K} \boldsymbol{Z}_k^{(i)}\boldsymbol{Z}_k^{(i)^T}\boldsymbol{z}_u^{(i)} = \lambda_i\boldsymbol{z}_u^{(i)}. \tag{58}$$

Recall that $\boldsymbol{z}_u^{(i)}$ and $\boldsymbol{z}_v^{(i)}$ are a column of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, respectively. Without loss of generality, suppose that $\boldsymbol{z}_u^{(i)}$ are the first column of $\boldsymbol{Z}_1^{(i)}$, i.e., $\boldsymbol{z}_u^{(i)} = \boldsymbol{Z}_1^{(i)}\boldsymbol{e}_1$. Then, let $\boldsymbol{c} = (\boldsymbol{c}_1\ \ldots\ \boldsymbol{c}_K) \in \mathbb{R}^{s_i}$ with $\boldsymbol{c}_1 = \boldsymbol{Z}_1^{(i)^T}\boldsymbol{z}_u^{(i)} - \lambda_i\boldsymbol{e}_1$ and $\boldsymbol{c}_k = \boldsymbol{Z}_k^{(i)^T}\boldsymbol{z}_u^{(i)}$ for all $k \neq 1$. This, together with $\boldsymbol{z}_u^{(i)^T}\boldsymbol{z}_v^{(i)} \neq 0$ and (58), implies $\boldsymbol{c}_2 \neq \boldsymbol{0}$ and $\boldsymbol{Z}^{(i)}\boldsymbol{c} = \boldsymbol{0}$. Now, we set $\boldsymbol{q}_k := \boldsymbol{V}_{k,1}^{(i)}\boldsymbol{V}_{k,1}^{(i)^T}\boldsymbol{c}_k$ for each $k \in [K]$ and $\boldsymbol{q} := (\boldsymbol{q}_1\ \ldots\ \boldsymbol{q}_K)$. According to $\boldsymbol{Z}_k^{(i)} = \beta_i\boldsymbol{U}_{k,1}^{(i)}\boldsymbol{V}_{k,1}^{(i)^T}$ by (49) and (53), we have for all $k \neq 1$,

$$\boldsymbol{q}_k = \boldsymbol{V}_{k,1}^{(i)}\boldsymbol{V}_{k,1}^{(i)^T}\boldsymbol{Z}_k^{(i)^T}\boldsymbol{z}_u^{(i)} = \boldsymbol{Z}_k^{(i)^T}\boldsymbol{z}_u^{(i)} = \boldsymbol{c}_k.$$

Moreover, using $\boldsymbol{Z}_k^{(i)} = \beta_i\boldsymbol{U}_{k,1}^{(i)}\boldsymbol{V}_{k,1}^{(i)^T}$ by (49) and (53), we have

$$\boldsymbol{Z}^{(i)}\boldsymbol{q} = \sum_{k=1}^{K}\boldsymbol{Z}_k^{(i)}\boldsymbol{q}_k = \beta_i\sum_{k=1}^{K}\boldsymbol{U}_{k,1}^{(i)}\boldsymbol{V}_{k,1}^{(i)^T}\boldsymbol{c}_k = \sum_{k=1}^{K}\boldsymbol{Z}_k^{(i)}\boldsymbol{c}_k = \boldsymbol{Z}^{(i)}\boldsymbol{c} = \boldsymbol{0} \tag{59}$$

and

$$\|\boldsymbol{Z}_k^{(i)}\boldsymbol{q}_k\| = \beta_i\|\boldsymbol{V}_{k,1}^{(i)^T}\boldsymbol{c}_k\| = \beta_i\|\boldsymbol{q}_k\|. \tag{60}$$

Let $\boldsymbol{u} = \boldsymbol{U}\boldsymbol{a}$, where $\boldsymbol{U}$ is given in (57) and $\boldsymbol{a} \in \mathbb{R}^{t_i}$ is chosen such that $\boldsymbol{a} \in \operatorname{span}(\boldsymbol{U}^T\boldsymbol{U}_{k,2}^{(i)})$ and $\|\boldsymbol{a}\| = 1$. We construct $\boldsymbol{D} = \left[\boldsymbol{D}^{(1)}\ \ldots\ \boldsymbol{D}^{(p)}\right]$ with $\boldsymbol{D}^{(i)} = \boldsymbol{u}\boldsymbol{q}^T$ and $\boldsymbol{D}^{(j)} = \boldsymbol{0}$ for all $j \neq i$.

Compute the bilinear form of Hessian. According to the construction of $\boldsymbol{D}$ and (59), we compute $\boldsymbol{Z}\boldsymbol{D}^T = \boldsymbol{Z}^{(i)}\boldsymbol{D}^{(i)^T} = \boldsymbol{Z}^{(i)}\boldsymbol{q}\boldsymbol{u}^T = \boldsymbol{0}$. This, together with (24) and (48), yields

$$\nabla^2 R(\boldsymbol{Z})\left[\boldsymbol{D},\boldsymbol{D}\right] = \alpha\boldsymbol{a}^T\boldsymbol{U}^T\left(\boldsymbol{I} + \alpha\boldsymbol{Z}^{(i)}\boldsymbol{Z}^{(i)^T}\right)^{-1}\boldsymbol{U}\boldsymbol{a}\|\boldsymbol{q}\|^2 = \frac{\alpha}{\alpha\lambda_i + 1}\|\boldsymbol{q}\|^2,$$

where the last equality is due to (57). With abuse of notation, let

$$R_c\left(\boldsymbol{Z}_k^{(i)}\right) = \frac{m_k}{2m}\sum_{i=1}^{p}\log\det\left(\boldsymbol{I}_n + \alpha_k\boldsymbol{Z}_k^{(i)}\boldsymbol{Z}_k^{(i)^T}\right).$$

Since $\boldsymbol{Z}_k^{(i)} = \beta_i \boldsymbol{U}_{k,1}^{(i)} \boldsymbol{V}_{k,1}^{(i)^T}$ and $\boldsymbol{D}_k^{(i)} = \boldsymbol{u}\boldsymbol{q}_k^T$, we compute for each $k \in [K]$,

$$\begin{aligned}\nabla^2 R_c\left(\boldsymbol{Z}_k^{(i)}\right)\left[\boldsymbol{D}_k^{(i)}, \boldsymbol{D}_k^{(i)}\right] &= \alpha\|\boldsymbol{q}_k\|^2 \boldsymbol{u}^T \boldsymbol{X}_k^{(i)} \boldsymbol{u} - \alpha\alpha_k \left(\boldsymbol{u}^T \boldsymbol{X}_k^{(i)} \boldsymbol{Z}_k^{(i)} \boldsymbol{q}_k\right)^2 - \\ &\quad \alpha\alpha_k \left(\boldsymbol{u}^T \boldsymbol{X}_k^{(i)} \boldsymbol{u}\boldsymbol{q}_k^T \boldsymbol{Z}_k^{(i)^T} \boldsymbol{X}_k^{(i)} \boldsymbol{Z}_k^{(i)} \boldsymbol{q}_k\right) \\ &= \alpha\|\boldsymbol{q}_k\|^2 \left(1 - \frac{\alpha_k}{\alpha_k\beta_i^2+1}\|\boldsymbol{Z}_k^{(i)^T}\boldsymbol{u}\|^2\right) - \frac{\alpha\alpha_k}{(\alpha_k\beta_i^2+1)^2}\left(\boldsymbol{u}^T\boldsymbol{Z}_k^{(i)}\boldsymbol{q}_k\right)^2 \\ &\quad - \alpha\alpha_k\left(1 - \frac{\alpha_k}{\alpha_k\beta_i^2+1}\|\boldsymbol{Z}_k^{(i)^T}\boldsymbol{u}\|^2\right)\frac{\beta_i^2\|\boldsymbol{q}_k\|^2}{\alpha_k\beta_i^2+1},\end{aligned}$$

where $\boldsymbol{X}_k^{(i)} = \left(\boldsymbol{I} + \alpha_k \boldsymbol{Z}_k^{(i)} \boldsymbol{Z}_k^{(i)^T}\right)^{-1}$, the second equality follows from $\boldsymbol{X}_k^{(i)} = \left(\boldsymbol{I} + \alpha_k\beta_i^2 \boldsymbol{U}_{k,1}^{(i)} \boldsymbol{U}_{k,1}^{(i)^T}\right)^{-1} = \boldsymbol{I} - \alpha_k\beta_i^2 \boldsymbol{U}_{k,1}^{(i)} \boldsymbol{U}_{k,1}^{(i)^T}/(\alpha_k\beta_i^2+1)$ due to (27) in Lemma A.3, $\boldsymbol{Z}_k^{(i)} = \beta_i \boldsymbol{U}_{k,1}^{(i)} \boldsymbol{V}_{k,1}^{(i)^T}$, and (60). Summing up the above equality for all $k \in [K]$ with $\alpha_k = K\alpha$ for all $k \in [K]$ yields

$$\begin{aligned}&\sum_{k=1}^K \nabla^2 R_c\left(\boldsymbol{Z}_k^{(i)}\right)\left[\boldsymbol{D}_k^{(i)}, \boldsymbol{D}_k^{(i)}\right] \\ &= \alpha\left(1 - \frac{\alpha_k\beta_i^2}{\alpha_k\beta_i^2+1}\right)\|\boldsymbol{q}\|^2 - \frac{\alpha\alpha_k}{\alpha_k\beta_i^2+1}\sum_{k=1}^K \|\boldsymbol{q}_k\|^2\|\boldsymbol{Z}_k^{(i)^T}\boldsymbol{u}\|^2 - \frac{\alpha\alpha_k}{(\alpha_k\beta_i^2+1)^2}\sum_{k=1}^K\left(\boldsymbol{u}^T\boldsymbol{Z}_k^{(i)}\boldsymbol{q}_k\right)^2 \\ &\quad + \frac{\alpha\alpha_k^2\beta_i^2}{(\alpha_k\beta_i^2+1)^2}\sum_{k=1}^K \|\boldsymbol{q}_k\|^2\|\boldsymbol{Z}_k^{(i)^T}\boldsymbol{u}\|^2 \\ &= \left(\frac{\alpha}{1+\alpha\lambda_i} - \lambda\right)\|\boldsymbol{q}\|^2 - \frac{\alpha\alpha_k}{(\alpha_k\beta_i^2+1)^2}\sum_{k=1}^K\left(\left(\boldsymbol{u}^T\boldsymbol{Z}_k^{(i)}\boldsymbol{q}_k\right)^2 + \|\boldsymbol{q}_k\|^2\|\boldsymbol{Z}_k^{(i)^T}\boldsymbol{u}\|^2\right),\end{aligned}$$

where the second equality follows from the definition of β_i in (53). Finally, we compute

$$\begin{aligned}\nabla^2 F(\boldsymbol{Z})[\boldsymbol{D}, \boldsymbol{D}] &= \nabla^2 R(\boldsymbol{Z})[\boldsymbol{D}, \boldsymbol{D}] - \sum_{j=1}^K \nabla^2 R_c(\boldsymbol{z}_j^{(i)})[\boldsymbol{d}_j^{(i)}, \boldsymbol{d}_j^{(i)}] - \lambda\|\boldsymbol{q}\|^2 \\ &= \frac{\alpha\alpha_k}{(\alpha_k\beta_i^2+1)^2}\sum_{k=1}^K\left(\left(\boldsymbol{u}^T\boldsymbol{Z}_k^{(i)}\boldsymbol{q}_k\right)^2 + \|\boldsymbol{q}_k\|^2\|\boldsymbol{Z}_k^{(i)^T}\boldsymbol{u}\|^2\right) > 0,\end{aligned}$$

where the inequality is due to $\|\boldsymbol{q}_2\| = \|\boldsymbol{c}_2\| \neq 0$ and

$$\|\boldsymbol{Z}_2^{(i)^T}\boldsymbol{u}\|^2 = \beta_2\|\boldsymbol{U}_{2,1}^{(i)^T}\boldsymbol{U}\boldsymbol{a}\| \neq 0$$

due to $\boldsymbol{a} \in \operatorname{span}(\boldsymbol{U}^T\boldsymbol{U}_{k,2}^{(i)})$. □

Given a matrix $\boldsymbol{Z} \in \mathbb{R}^{d\times m}$, let $\boldsymbol{Z}\boldsymbol{Z}^T = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$ be an eigenvalue decomposition of $\boldsymbol{Z}\boldsymbol{Z}^T \in \mathbb{S}_+^d$, where $\boldsymbol{Q} \in \mathcal{O}^{d\times r}$ and $\boldsymbol{\Lambda} \in \mathbb{R}^{r\times r}$ is a diagonal matrix with diagonal entries being positive eigenvalues of $\boldsymbol{Z}\boldsymbol{Z}^T$. Suppose that $\boldsymbol{Z}\boldsymbol{Z}^T$ has p distinct positive eigenvalues, where $1 \le p \le r$. Let $\lambda_1 > \cdots > \lambda_p > 0$ be its distinct eigenvalue values with the corresponding multiplicities being $h_1, \ldots, h_p \in \mathbb{N}_+$, respectively. Obviously, we have $\sum_{i=1}^p h_i = r$. Therefore, we write

$$\boldsymbol{\Lambda} = \operatorname{BlkDiag}\left(\lambda_1\boldsymbol{I}_{h_1}, \ldots, \lambda_p\boldsymbol{I}_{h_p}\right),\ \boldsymbol{Q} = \left[\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_p\right],$$

where $\boldsymbol{Q}_i \in \mathcal{O}^{d\times h_i}$ for all $i \in [p]$.

D. Proofs in Section 3

D.1. Proof of Theorem 3.1

Proof of Theorem 3.1. (i) Suppose that each block of $\boldsymbol{Z}$ satisfying (7), (a), (b), and (c). It directly follows from (i) in Proposition 4.3 that $\boldsymbol{Z}$ is a local maximizer. Conversely, suppose that $\boldsymbol{Z}$ is a local maximizer. According to (18), (19),

(20), (ii) in Proposition 4.3 and Proposition 4.4, if $\boldsymbol{Z} \in \mathcal{Z}^c \cup \mathcal{Z}_2$, then $\boldsymbol{Z}$ is a strict saddle point. This, together with $\mathcal{X} = \mathcal{Z}^c \cup \mathcal{Z}_1 \cup \mathcal{Z}_2$ and the fact that $\boldsymbol{Z}$ is a local maximizer, implies that $\boldsymbol{Z} \in \mathcal{Z}_1$. Using this, (20), and Proposition 4.2 yields that $\boldsymbol{Z}$ satisfying (7), (a), (b), and (c).

(ii) According to (i) in Theorem 3.1, suppose that the k-th block of a local maximizer $\boldsymbol{Z}$ admits the decomposition in (7) satisfying (a), (b), and (c) for all $k \in [K]$. This, together with (42) in the proof of Proposition 4.3, yields that

$$F(\boldsymbol{Z}) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{r_k}\left(\log\left(1+\alpha\overline{\sigma}_k^2\right) - \frac{m_k}{m}\log\left(1+\alpha_k\overline{\sigma}_k^2\right) - \lambda\overline{\sigma}_k^2\right), \tag{61}$$

where $\overline{\sigma}_k$ is defined in (16) for each $k \in [K]$. Then, we define a function $g : \mathbb{N}_+ \times \mathbb{R} \to \mathbb{R}$ as

$$g(n,x) := \log(1+\alpha x) - \frac{n}{m}\log\left(1+\frac{dx}{n\varepsilon^2}\right) - \lambda x.$$

One can verify that for all $n_1 \geq n_2$, we have $g(n_1,x) \leq g(n_2,x)$ for each x. Therefore, we have for all $m_k \leq m_l$,

$$g(m_l,\overline{\sigma}_l^2) \leq g(m_k,\overline{\sigma}_l^2) \leq g(m_k,\overline{\sigma}_k^2),$$

where the second inequality follows from $\overline{\sigma}_k^2$ is the maximizer of the function $g(m_k,x) = h_k(x)$ according to (43). This, together with (61), yields that $\boldsymbol{Z}$ is a global maximizer if and only if $\sum_{k=1}^{K} r_k = \min\{m,d\}$ and for all $k \neq l$ satisfying $m_k < m_l$ and $r_l > 0$, we have $r_k = \min\{m_k,d\}$. □

D.2. Proof of Proposition 3.2

To prove Proposition 3.2, we first need to characterize the global optimal solution set of Problem (4).

Proposition D.1. *Suppose that $m_1 = \cdots = m_K$ and (8) holds. It holds that $\boldsymbol{Z} = [\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_K] \in \mathbb{R}^{d\times m}$ with $\boldsymbol{Z}_k \in \mathbb{R}^{d\times m_k}$ for each $k \in [K]$ is a global solution of Problem (4) if and only if for each $k \in [K]$,*

$$\boldsymbol{Z}_k = \frac{m}{\min\{m,d\}}\boldsymbol{U}_k\boldsymbol{V}_k^T, \tag{62}$$

where $r_k = \min\{m,d\}/K$ for all $k \in [K]$, $\boldsymbol{U}_k \in \mathcal{O}^{d\times r_k}$ with $\boldsymbol{U}_k^T\boldsymbol{U}_l = \boldsymbol{0}$ for all $l \neq k$, and $\boldsymbol{V}_k \in \mathcal{O}^{m_k\times r_k}$ for all $k \in [K]$.

Proof. According to Lemma A.5, we have

$$R(\boldsymbol{Z}) - \sum_{k=1}^{K} R_c\left(\boldsymbol{Z};\boldsymbol{\pi}^k\right) \leq \frac{1}{2}\sum_{k=1}^{K}\log\det\left(\boldsymbol{I}+\alpha\boldsymbol{Z}_k\boldsymbol{Z}_k^T\right) - \sum_{k=1}^{K}\frac{m_k}{2m}\log\det\left(\boldsymbol{I}+\alpha_k\boldsymbol{Z}_k\boldsymbol{Z}_k^T\right) =: H(\boldsymbol{Z}), \tag{63}$$

where the inequality becomes equality if and only if $\boldsymbol{Z}_k^T\boldsymbol{Z}_l = \boldsymbol{0}$ for all $k \neq l$. To simplify our development, let $r_k := \operatorname{rank}(\boldsymbol{Z}_k)$ denote the rank of $\boldsymbol{Z}_k \in \mathbb{R}^{d\times m_k}$, where $r_k \leq \min\{d,m_k\}$ for each $k \in [K]$ and $\sum_{k=1}^{K} r_k \leq \min\{d,m\}$, and

$$h_k(\boldsymbol{Z}_k) := \frac{1}{2}\log\det\left(\boldsymbol{I}+\alpha\boldsymbol{Z}_k\boldsymbol{Z}_k^T\right) - \frac{m_k}{2m}\log\det\left(\boldsymbol{I}+\alpha_k\boldsymbol{Z}_k\boldsymbol{Z}_k^T\right),\ \forall k \in [K]. \tag{64}$$

Moreover, let

$$\boldsymbol{Z}_k = \boldsymbol{P}_k\boldsymbol{\Sigma}_k\boldsymbol{Q}_k^T = \begin{bmatrix}\boldsymbol{P}_{k,1} & \boldsymbol{P}_{k,2}\end{bmatrix}\begin{bmatrix}\tilde{\boldsymbol{\Sigma}}_k & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{Q}_{k,1}^T\\ \boldsymbol{Q}_{k,2}^T\end{bmatrix}$$

be a singular value decomposition (SVD) of $\boldsymbol{Z}_k$, where $\tilde{\boldsymbol{\Sigma}}_k = \operatorname{diag}(\sigma_{k,1},\ldots,\sigma_{k,r_k})$ with $\sigma_{k,1} \geq \cdots \geq \sigma_{k,r_k} > 0$ being positive singular values of $\boldsymbol{Z}_k$, $\boldsymbol{P}_k \in \mathcal{O}^d$ with $\boldsymbol{P}_{k,1} \in \mathbb{R}^{d\times r_k}$ and $\boldsymbol{P}_{k,2} \in \mathbb{R}^{d\times(d-r_k)}$, and $\boldsymbol{Q}_k \in \mathcal{O}^{m_k}$ with $\boldsymbol{Q}_{k,1} \in \mathbb{R}^{m_k\times r_k}$ and $\boldsymbol{Q}_{k,2} \in \mathbb{R}^{m_k\times(m_k-r_k)}$. Substituting this SVD into (64), together with $\|\boldsymbol{Z}_k\|_F^2 = m_k$, yields that to maximize $H(\boldsymbol{Z})$, it suffices to study for each $k \in [K]$,

$$\max_{\sigma_{k,1},\ldots,\sigma_{k,r_k}} \sum_{i=1}^{r_k}\log\left(1+\alpha\sigma_{k,i}^2\right) - \sum_{i=1}^{r_k}\frac{m_k}{m}\log\left(1+\alpha_k\sigma_{k,i}^2\right) \qquad \text{s.t. } \sum_{i=1}^{r_k}\sigma_{k,i}^2 = m_k.$$

To simplify our development, let $x_i := \sigma_{k,i}^2 \geq 0$ for each $i \in [r_k]$. This, together with $m_k = m/K$, implies that it suffices to study

$$\max_{x_1,\ldots,x_{r_k}} g(\boldsymbol{x}) := \sum_{i=1}^{r_k} \log\left(1+\alpha x_i\right) - \sum_{i=1}^{r_k} \frac{1}{K}\log\left(1+\alpha_k x_i\right) \qquad \text{s.t. } \sum_{i=1}^{r_k} x_i = \frac{m}{K},\ x_i \geq 0,\ \forall i \in [r_k]. \tag{65}$$

This, together with Lemma D.2 and (11), yields that the optimal solution for each $k \in [K]$ is

$$x_i^* = \frac{m}{r_k K},\ \forall i \in [r_k]. \tag{66}$$

(i) Suppose that $m \leq d$. Then, we have $r_k \leq m/K$ for each $k \in [K]$ and $\sum_{k=1}^K r_k \leq m$. This, together with (66) and Lemma D.3, implies that $r_k = m/K$, and thus $x_i^* = 1$ for all $i \in [r_k]$ and $k \in [k]$.

(ii) Suppose that $m > d$. Then, we have $r_k \leq \min\{d, m/K\}$ for each $k \in [K]$ and $\sum_{k=1}^K r_k \leq d$. To compute the optimal function value, we consider the following problem:

$$\max_{r_1,\ldots,r_K \in \mathbb{Z}} \sum_{k=1}^K r_k \left(\log\left(1+\frac{m\alpha}{r_k K}\right) - \frac{1}{K}\log\left(1+\frac{m\alpha}{r_k}\right)\right) \quad \text{s.t. } \sum_{k=1}^K r_k = d,\ r_k \leq \min\left\{d, \frac{m}{K}\right\},\ \forall k \in [K].$$

Now, we study the following function:

$$\phi(x) := x\left(\log\left(1+\frac{m\alpha}{Kx}\right) - \frac{1}{K}\log\left(1+\frac{m\alpha}{x}\right)\right),\ \text{where } x \in \left[1, \frac{m}{K}\right]$$

We compute

$$\phi'(x) = \log\left(1+\frac{m\alpha}{Kx}\right) - \frac{1}{K}\log\left(1+\frac{m\alpha}{x}\right) - \frac{m\alpha}{Kx+m\alpha} + \frac{m\alpha}{K(x+m\alpha)},$$
$$\phi''(x) = -\frac{m\alpha/x}{Kx+m\alpha} + \frac{m\alpha/x}{K(x+m\alpha)} + \frac{Km\alpha}{(Kx+m\alpha)^2} - \frac{m\alpha}{K(x+m\alpha)^2} = -\frac{m^2\alpha^2/x}{(Kx+m\alpha)^2} + \frac{m^2\alpha^2/x}{K(x+m\alpha)^2}.$$

Since $x \in [1, m/K]$, we have $Kx^2 \leq m^2/K \leq m^2\alpha^2$ when $\alpha \geq 1/\sqrt{K}$, and thus $\phi''(x) \leq 0$. Therefore, $\phi(x)$ is a concave function for all $x \in [1, m/K]$. Then, applying the Jensen inequality yields

$$\sum_{k=1}^K \frac{1}{K} f(r_k) \leq f\left(\sum_{k=1}^K r_k\right),$$

where the inequality becomes equality if and only if $r_1 = \cdots = r_k = d/K$. This, together with (66), yields $x_i^* = m/d$.

According to (i) and (ii), we have $x_i^* = m/\min\{m,d\}$ and $r_k = \min\{m,d\}/K$. Therefore, we have $\boldsymbol{Z}_k = m/\min\{m,d\}\boldsymbol{P}_{k,1}\boldsymbol{Q}_{k,1}^T$, where $\boldsymbol{P}_{k,1} \in \mathcal{O}^{d\times r_k}$ and $\boldsymbol{V}_k \in \mathcal{O}^{m_k \times r_k}$ for each $k \in [K]$. Then, we complete the proof.

□

Based on the above proposition, we are ready to prove Proposition 3.2.

Proof of Proposition 3.2. Let $\hat{\boldsymbol{Z}} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_k]$ denote the optimal solution of Problem (4). According to Proposition D.1 and $\alpha = d/(m\epsilon^2)$ with ϵ satisfying (8), it suffices to study the following two cases.

(i) Suppose that $m < d$. Using this and (62), we have $\hat{\boldsymbol{Z}}_k = \boldsymbol{U}_k\boldsymbol{V}_k^T$ and $\hat{r}_k = m/K$ for each $k \in [K]$. Moreover, according to Theorem 3.1, if $m_k = m/K$ for each $k \in [K]$ and λ satisfies (10), one can verify that the global solutions of Problem (5) satisfy (7) with $\overline{\sigma}_1 = \cdots = \overline{\sigma}_K = 1$ and $\sum_{k=1}^K r_k = m$. Since $r_k \leq m_k$ for each $k \in [K]$, we have $r_k = m/K$ for each $k \in [K]$. Therefore, Problem (4) and Problem (5) have the same global solution set.

(ii) Suppose that $m \geq d$. Using this and (62), we have $\hat{\boldsymbol{Z}}_k = m\boldsymbol{U}_k\boldsymbol{V}_k^T/d$ and $\hat{r}_k = d/K$ for each $k \in [K]$. Moreover, according to Theorem 3.1, if $m_k = m/K$ for each $k \in [K]$ and λ satisfies (9), one can verify that the global solutions of Problem (5) satisfy (7) with $\overline{\sigma}_1 = \cdots = \overline{\sigma}_K = m/d$ and $\sum_{k=1}^K r_k = d$. Therefore, the global solution set of Problem (4) is a subset of that of Problem (5).

□

Lemma D.2. *Suppose that m, K are integers such that m/K is a positive integer, $r \leq m/K$ is an integer, and $\alpha > 0$ is a constant. Consider the following optimization problem*

$$\min_{\boldsymbol{x}\in\mathbb{R}^r} \sum_{i=1}^{r} -\log\left(1+\alpha x_i\right) + \sum_{i=1}^{r} \frac{1}{K}\log(1+K\alpha x_i) \quad \text{s.t. } \sum_{i=1}^{r} x_i = \frac{m}{K},\ x_i \geq 0,\ \forall i \in [r]. \tag{67}$$

If

$$\alpha \geq 6K^{\frac{1}{K-1}}\exp(1)\left(1+\frac{1}{\sqrt{K}}\right)^{\frac{m}{K-1}}, \tag{68}$$

the optimal solution is

$$x_i^* = \frac{m}{rK},\ \forall i \in [r]. \tag{69}$$

Proof. If $r = 1$, it is trivial to see that (69) is the optimal solution. Therefore, it suffices to study $r \geq 2$. To simplify our development, let $f(x) := -\log(1+\alpha x) + \log(1+K\alpha x)/K$ and $F(\boldsymbol{x}) := \sum_{i=1}^{r} f(x_i)$. Then, one can verify that for all $x \geq 0$,

$$f'(x) = -\frac{\alpha}{1+\alpha x} + \frac{\alpha}{1+K\alpha x} < 0, \quad f''(x) = \frac{\alpha^2}{(1+\alpha x)^2} - \frac{K\alpha^2}{(1+K\alpha x)^2}. \tag{70}$$

Introducing dual variables λ associated with the constraint $\sum_{i=1}^{r} x_i = m/K$ and μ_i associated with the constraint $x_i \geq 0$ for each $i \in [r]$, we write the Lagrangian as follows

$$\mathcal{L}(\boldsymbol{x};\lambda,\boldsymbol{\mu}) = \sum_{i=1}^{r} f(x_i) + \lambda\left(\sum_{i=1}^{r} x_i - \frac{m}{K}\right) - \sum_{i=1}^{r}\mu_i x_i. \tag{71}$$

Then, we write the KKT system as follows:

$$-\frac{\alpha}{1+\alpha x_i} + \frac{\alpha}{1+K\alpha x_i} + \lambda - \mu_i = 0,\ x_i\mu_i = 0,\ x_i \geq 0,\ \mu_i \geq 0,\ \forall i \in [r],\ \sum_{i=1}^{r} x_i = \frac{m}{K}. \tag{72}$$

Now, let $\mathcal{S} := \{i \in [r] : x_i > 0\}$ denote the support set of a KKT point $\boldsymbol{x} \in \mathbb{R}^r$ and $s := |\mathcal{S}|$ denote the cardinality of the support set, where $1 \leq s \leq r$. This, together with (72), implies that for each $i \in \mathcal{S}$,

$$-\frac{\alpha}{1+\alpha x_i} + \frac{\alpha}{1+K\alpha x_i} + \lambda = 0,\ \sum_{i\in\mathcal{S}} x_i = \frac{m}{K}. \tag{73}$$

This is equivalent to the following quadratic equation:

$$K\lambda\alpha x_i^2 - \left((K-1)\alpha - (K+1)\lambda\right)x_i + \frac{\lambda}{\alpha} = 0. \tag{74}$$

We compute

$$\Delta = \eta^2 - 4K\lambda^2, \text{where } \eta := (K-1)\alpha - (K+1)\lambda. \tag{75}$$

Note that for all $i \in \mathcal{S}$, we have $x_i > 0$, and thus $\mu_i = 0$. This, together with $K \geq 2$ and the first equation in (73), implies $\lambda > 0$. Consequently, the quadratic equation (74) has a positive root if and only if $\eta \geq 0$ and $\Delta \geq 0$. This implies

$$0 < \lambda \leq \frac{\sqrt{K}-1}{\sqrt{K}+1}\alpha. \tag{76}$$

Then, the solution of Problem (74) is $x_i \in \{\overline{x}, \underline{x}\}$ for each $i \in \mathcal{S}$, where

$$\overline{x} = \frac{\eta+\sqrt{\Delta}}{2K\lambda\alpha}, \quad \underline{x} = \frac{\eta-\sqrt{\Delta}}{2K\lambda\alpha}. \tag{77}$$

Now, we discuss the KKT points that could potentially be optimal solutions. Let $\boldsymbol{x} \in \mathbb{R}^r$ be a KKT point satisfying $x_i \in \{\overline{x}, \underline{x}\}$ for each $i \in \mathcal{S}$, where $s \in \{1, 2, \ldots, r\}$. In particular, when $s = 1$, we have $x_i = m/K$ for all $i \in \mathcal{S}$. In the following, we consider $s \in \{2, \ldots, r\}$.

Case 1. Suppose that $x_i = x_j$ for all $i, j \in \mathcal{S}$. This, together with $\sum_{i\in\mathcal{S}} x_i = m/K$ and $x_i \in \{\overline{x}, \underline{x}\}$ for each $i \in \mathcal{S}$, yields

$$x_i = \frac{m}{sK}, \ \forall i \in \mathcal{S}. \tag{78}$$

Case 2. Suppose that there exists $i \neq j \in \mathcal{S}$ such that $x_i \neq x_j$. This, together with $x_i \in \{\overline{x}, \underline{x}\}$, implies $\overline{x} > \underline{x}$. According to (70), we have $f''(x) = 0$ at $\hat{x} = 1/(\alpha\sqrt{K})$. Then, we obtain that $f'(x)$ is strictly decreasing in $[0, \hat{x}]$ and strictly increasing in $[\hat{x}, \infty]$. Then, one can further verify that $\underline{x} < \hat{x} < \overline{x}$. This, together with (70), implies

$$f''(\underline{x}) < 0, \ f''(\overline{x}) > 0. \tag{79}$$

For ease of exposition, let $l(\boldsymbol{x}) = |\{i \in \mathcal{S} : x_i = \underline{x}\}|$ be the number of entries of $\boldsymbol{x}$ that equal to $\underline{x}$. Then, we claim that any optimal solution $\boldsymbol{x}^\star$ satisfies $l(\boldsymbol{x}^\star) \leq 1$. Now, we prove this claim by contradiction. Without loss of generality, we assume that $x_i^\star = \overline{x}$ for all $i = 1, \ldots, r - l$ and $x_i^\star = \underline{x}$ for all $i = r - l + 1, \ldots, r$ with $l \geq 2$. This, together with (79) and $l \geq 2$, yields

$$f''(x_{r-1}^\star) < 0, \quad f''(x_r^\star) < 0. \tag{80}$$

Using the second-order necessary condition for constraint optimization problems (see, e.g., (Nocedal & Wright, 1999, Theorem 12.5)) and $x_i^\star \geq 0$ for all $i \in [r]$, we obtain

$$\sum_{i=1}^{r} f''(x_i^\star) v_i^2 \geq 0, \ \forall \boldsymbol{v} \in \mathbb{R}^r \text{ s.t. } \sum_{i=1}^{r} v_i = 0. \tag{81}$$

Then, we take $\boldsymbol{v} \in \mathbb{R}^r$ such that $v_1 = \cdots = v_{r-2} = 0$ and $v_{r-1} = -v_r \neq 0$. Substituting this into (81) yields

$$f''(x_{r-1}^\star) + f''(x_r^\star) \geq 0,$$

which contradicts (80). Therefore, $\boldsymbol{x}^\star$ cannot be an optimal solution. Then, we prove the claim. In this case, we can write the KKT point that could be an optimal solution as follows: There exists $i \in \mathcal{S}$ such that

$$x_i = \underline{x}, \ x_j = \overline{x}, \ \forall j \neq i. \tag{82}$$

Now, we compute the function values for the above two cases, i.e., (78) and (82), and compare them to determine which one is the optimal solution. To simplify our analysis, let $h(x) := f(x)/x$. For (78) in **Case 1**, we compute

$$F_1 = sf\left(\frac{m}{sK}\right) = \frac{m}{K} h\left(\frac{m}{sK}\right). \tag{83}$$

For (82) in **Case 2**, we compute

$$F_2 = (s-1)f(\overline{x}) + f(\underline{x}) \geq (s-1)f\left(\frac{m}{(s-1)K}\right) + f(\underline{x}) = \frac{m}{K} h\left(\frac{m}{(s-1)K}\right) + f(\underline{x}), \tag{84}$$

where the inequality is because $\overline{x} \leq m/((s-1)K)$ and $f(x)$ is strictly dereasing in $[0, \infty)$. For all $x \in [m/(sK), m/((s-1)K)]$, we compute

$$\begin{aligned} h'(x) &= \frac{f'(x)x - f(x)}{x^2} = -\frac{1}{x}\left(\frac{\alpha}{1+\alpha x} - \frac{\alpha}{1+K\alpha x}\right) + \frac{1}{x^2}\left(\log(1+\alpha x) - \frac{1}{K}\log(1+K\alpha x)\right) \\ &\geq -\frac{(K-1)\alpha^2}{(1+\alpha x)(1+K\alpha x)} + \frac{1}{x^2}\left(\log(1+\alpha) - \frac{1}{K}\log(1+\alpha K)\right) \\ &\geq -\frac{K-1}{Kx^2}\left(1 - \log(1+\alpha) + \frac{\log K}{K-1}\right) \geq -\frac{(K-1)Ks^2}{m^2}\left(1 + \frac{\log K}{K-1} - \log(1+\alpha)\right), \end{aligned} \tag{85}$$

where the first inequality follows from $\log(1+\alpha x)-\log(1+K\alpha x)/K \geq \log(1+\alpha)-\log(1+\alpha K)/K$ due to $x \geq m/(sK) \geq 1$, the second inequality uses $\log(1+\alpha)-\log(1+\alpha K)/K = (K-1)\log(1+\alpha)/K + \log\left((1+\alpha)/(1+\alpha K)\right)/K \geq \left((K-1)\log(1+\alpha)-\log K\right)/K$, and the last inequality is because of $x \geq m/(sK)$. According to (68), we have

$$1+\frac{\log K}{K-1}-\log(1+\alpha)=\log\left(\frac{K^{\frac{1}{K-1}}\exp(1)}{1+\alpha}\right)<0. \tag{86}$$

Using the mean-value theorem, there exists $x \in (m/(sK), m/\left((s-1)K\right))$ such that

$$h\left(\frac{m}{(s-1)K}\right)-h\left(\frac{m}{sK}\right)=h'(x)\left(\frac{m}{(s-1)K}-\frac{m}{sK}\right)\geq\frac{(K-1)s}{m(s-1)}\left(\log(1+\alpha)-1-\frac{\log K}{K-1}\right), \tag{87}$$

where the inequality follows from (85). Now, we are devoted to bounding $f(\underline{x})$. According to (75) and (76), we have

$$\underline{x}=\frac{\eta-\sqrt{\Delta}}{2K\lambda\alpha}=\frac{4K\lambda^2}{2K\lambda\alpha(\eta+\sqrt{\Delta})}\leq\frac{2\lambda}{\alpha\eta}. \tag{88}$$

This, together with the fact that $f(x)$ is decreasing in $(0,\infty)$, yields

$$f(\underline{x})\geq f\left(\frac{2\lambda}{\alpha\eta}\right)=-\log\left(1+\frac{2\lambda}{\eta}\right)+\frac{1}{K}\log\left(1+\frac{2K\lambda}{\eta}\right)\geq-\log\left(1+\frac{2\lambda}{\eta}\right)\geq-\log\left(1+\frac{1}{\sqrt{K}}\right),$$

where the last inequality uses $\eta=(K-1)\alpha-(K+1)\lambda\geq 2\sqrt{K}\lambda$ due to $(K-1)\alpha\geq(\sqrt{K}+1)^2\lambda$ by (76). This, together with (83), (84), and (87), yields

$$\begin{aligned}F_2-F_1&=\frac{m}{K}\left(h\left(\frac{m}{(s-1)K}\right)-h\left(\frac{m}{sK}\right)\right)+f(\underline{x})\\&\geq\frac{(K-1)s}{m(s-1)}\left(\log(1+\alpha)-1-\frac{\log K}{K-1}\right)-\log\left(1+\frac{1}{\sqrt{K}}\right)\\&\geq\frac{K-1}{m}\log\left(\frac{1+\alpha}{K^{\frac{1}{K-1}}\exp(1)}\right)-\log\left(1+\frac{1}{\sqrt{K}}\right)>0,\end{aligned}$$

where the last inequality follows from (68). This implies that the optimal solution takes the form of (78) for some $s\in[r]$. Consequently, the function value of (78) for each $s\in[r]$ is

$$s\left(-\log\left(1+\frac{\alpha m}{sK}\right)+\frac{1}{K}\log\left(1+\frac{\alpha m}{s}\right)\right).$$

This, together with Lemma D.3, implies that when the optimal solution takes the form of (78) with $s=r$, Problem (67) achieves its global minimum. Then, we complete the proof. □

Lemma D.3. *Consider the setting in Lemma D.2 and the following function*

$$h(s):=s\left(\frac{1}{K}\log\left(1+\frac{m\alpha}{s}\right)-\log\left(1+\frac{m\alpha}{sK}\right)\right), \tag{89}$$

where $s\in[1,r]$ and α satisfies (68). Then, $h(s)$ is decreasing in $s\in[1,r]$.

Proof. For ease of exposition, let $\beta:=m\alpha$ and $x:=1/s\in[1/r,1]$. According to (68), we have

$$\alpha\geq 6K^{\frac{1}{K-1}}\exp(1)\left(1+\frac{1}{\sqrt{K}}\right)^{\frac{2m}{K-1}}>1\geq\frac{r\sqrt{K}}{m}.$$

This implies $\beta\geq r\sqrt{K}$. Then, we study

$$h(s)=g(x)=\frac{1}{x}\left(\frac{1}{K}\log\left(1+\beta x\right)-\log\left(1+\frac{\beta x}{K}\right)\right). \tag{90}$$

Note that showing $h(s)$ is decreasing in $s \in [1, r]$ is equivalent to proving $g(x)$ is increasing in $x \in [1/r, 1]$. Now, we compute

$$g'(x) = \frac{1}{x}\left(\frac{\beta}{K(1+\beta x)} - \frac{\beta}{K+\beta x}\right) - \frac{1}{x^2}\left(\frac{1}{K}\log(1+\beta x) - \log\left(1+\frac{\beta x}{K}\right)\right) = -\frac{1}{x^2}\phi(x), \tag{91}$$

where

$$\phi(x) := \frac{1}{K}\log(1+\beta x) - \log\left(1+\frac{\beta x}{K}\right) + \beta x\left(\frac{1}{K+\beta x} - \frac{1}{K(1+\beta x)}\right).$$

Then, it suffices to show $\phi(x) \le 0$ for all $x \in [K/m, 1]$ due to $1/r \ge K/m$. Towards this goal, we compute

$$\begin{aligned}\phi'(x) &= \frac{\beta}{K(1+\beta x)} - \frac{\beta}{K+\beta x} + \beta\left(\frac{1}{K+\beta x} - \frac{1}{K(1+\beta x)}\right) + \beta x\left(\frac{-\beta}{(K+\beta x)^2} + \frac{\beta}{K(1+\beta x)^2}\right)\\ &= -x\beta^2\frac{(K-1)(\beta^2x^2-K)}{(K+\beta x)^2K^2(1+\beta x)^2} \le 0,\end{aligned}$$

where the inequality follows from $\beta^2x^2 \ge \beta^2K^2/m^2 \ge K$ due to $x \in [K/m, 1]$ and $\beta = \alpha m > 2m > m/\sqrt{K}$. Therefore, $\phi(x)$ is decreasing in $[K/m, 1]$. Next, we compute

$$\begin{aligned}\phi\left(\frac{K}{m}\right) &= \frac{1}{K}\log\left(1+\frac{\beta K}{m}\right) - \log\left(1+\frac{\beta}{m}\right) + \frac{\beta}{m}\left(\frac{1}{1+\beta/m} - \frac{1}{1+\beta K/m}\right)\\ &= \frac{1}{K}\log(1+\alpha K) - \log(1+\alpha) + \alpha\left(\frac{1}{1+\alpha} - \frac{1}{1+\alpha K}\right)\\ &\le \frac{1}{K}\log(1+\alpha K) - \log(1+\alpha) + 1 \le \frac{1}{2}\log(1+2\alpha) - \log(1+\alpha) + 1 \le 0,\end{aligned}$$

where the second equality is due to $\beta = \alpha m$, the second inequality holds because $\log(1+\alpha K)/K$ is decreasing as $K \ge 2$ increases when $\alpha \ge 2$, and the last inequality follows from $\alpha \ge 15$ by (68). This, together with the fact that ϕ is decreasing in $[K/m, 1]$, yields $\phi(x) \le \phi(K/m) \le 0$. Using this and (91), we obtain $g'(x) \ge 0$ in $[K/m, 1]$. Therefore, $g(x)$ is increasing in $[K/m, 1]$. Then, we complete the proof. □

D.3. Proof of and Theorem 3.3

Proof of Theorem 3.3. According to (i) of Proposition 4.3, if $\boldsymbol{Z}$ is a critical point but not a local maximizer, we have $\boldsymbol{Z} \in \mathcal{Z}_2 \cup \mathcal{Z}^c$. This, together with (ii) of Proposition 4.3 and Proposition 4.4, yields that $\boldsymbol{Z}$ is a strict saddle point. □

E. Additional Experimental Setups and Results

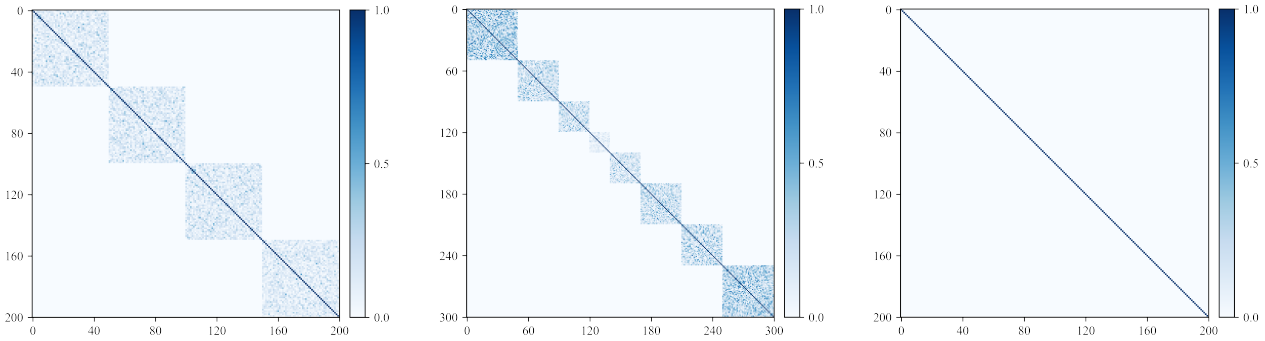
In this section, we provide additional implementation details and experimental results under different parameter settings for Sections 5.1 and 5.2 in Appendices E.1 and E.2, respectively.

E.1. Implementation Details and Additional Results in Section 5.1

Training setups. In this subsection, we employ full-batch gradient descent (GD) for solving Problem (5). Here, we use the Gaussian distribution to randomly initialize GD. More precisely, we randomly generate an initial point $\boldsymbol{Z}^{(0)}$ whose entries are i.i.d. sampled from the standard normal distribution, i.e. $z_{ij}^{(0)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. We fix the learning rate of GD as 10^{-1} in the training. We terminate the algorithm when the gradient norm at some iterate is less 10^{-5}.

In addition to the results presented in Section 5.1, we perform additional experiments under different settings as follows. To support our theorems, we visualize the heatmap of learned features and plot the number and magnitude of singular values in each class. Unless specified otherwise, we use the training setups introduced above in the following experiments.

- **Experiment 1 on balanced data.** In this experiment, we consider that the number of samples in each class is same and set the parameters in Problem (5) as follows: the dimension of features $d = 100$, the number of classes $K = 4$,



(a) Experiment 1 (b) Experiment 2 (c) Experiment 3

Figure 5. **Heatmap of cosine similarity between pairwise features under different settings.**

the number of samples in each class is $m_1 = m_2 = m_3 = m_4 = 50$, the regularization parameter $\lambda = 0.1$, and the quantization error $\epsilon = 0.5$. We visualize the heatmap between pairwise features of $\boldsymbol{Z}$ obtained by GD in Figure 5(a) and the number and magnitude of singular values in each class in Figure 6.

- **Experiment 2 on data with more classes.** In this experiment, we set the parameters in Problem (5) as follows: the dimension of features $d = 100$, the number of classes $K = 8$, the number of samples in each class is $m_1 = 50, m_2 = 40, m_3 = 30, m_4 = 20, m_5 = 30, m_6 = 40, m_7 = 40, m_8 = 50$, the regularization parameter $\lambda = 0.1$, and the quantization error $\epsilon = 0.5$. We visualize the heatmap between pairwise features of $\boldsymbol{Z}$ obtained by GD in Figure 5(b) and the number and magnitude of singular values in each class in Figure 7.
- **Experiment 3 on data where the dimension d is larger than the number of samples m.** In this experiment, we set the parameters in Problem (5) as follows: the dimension of features $d = 300$, the number of classes $K = 4$, the number of samples in each class is $m_1 = 50, m_2 = 50, m_3 = 40, m_4 = 60$, the regularization parameter $\lambda = 0.01$, and the quantization error $\epsilon = 5$. Note that in this experiment, we set the learning rate as 1. We visualize the heatmap between pairwise features of $\boldsymbol{Z}$ obtained by GD in Figure 5(c) and the number and magnitude of singular values in each class in Figure 8.

According to the results in Figures 5(a), 5(b), 6, and 7 of Experiments 1 and 2, we observe that when the number of samples is larger than its dimension, i.e., $m \geq d$, the learned features via the MCR^2 principle are within-class compressible and between-class discriminative in both balanced and unbalanced data sets. Moreover, the dimension of the space spanned by these features is maximized such that $\sum_{k=1}^{K} r_k = \min\{m, d\}$. This directly supports Theorem 3.1. By comparing the function value returned by GD and that computed by the closed-form in Theorem 3.1, we found that GD with random initialization will always converge to a global maximizer of Problem (5) when the data is balanced, while it will always converge to a local maximizer of Problem (5) when data is unbalanced. This directly supports Theorem 3.3.

According to the results in Figures 5(c) and 8 of Experiment 3, we observe that when the number of samples is smaller than its dimension, i.e., $m \leq d$, the learned features via the MCR^2 principle are orthogonal to each other and the dimension of each subspace is equal to the number of samples, i.e., $r_k = m_k$ for each $k \in [K]$. This exactly supports Theorem 3.1. Indeed, when $d \geq m$ and $r_k = m_k$ for each $k \in [K]$, it follows from Theorem 3.1 that $\boldsymbol{V}_k = \boldsymbol{I}$ and thus $\boldsymbol{Z}_k = \overline{\sigma}_k \boldsymbol{U}_k$ for each $k \in [K]$ for each local maximizer. Therefore, this also supports Theorem 3.3 as GD with random initialization converges to a local maximizer.

E.2. Implementation Details and Additional Results in Section 5.2

Network architecture and training setups for MNIST. In the experiments on MNIST, we employ a 4-layer multilayer perception (MLP) network with ReLU activation as the feature mapping with the intermediate dimension 2048 and output dimension 32. In particular, each layer of MLP networks consists of a linear layer and layer norm layer followed by ReLU activation in the implementation. We train the network parameters via Adam by optimizing the MCR^2 function. For the Adam settings, we use a momentum of 0.9, a full-batch size, and a dynamically adaptive learning rate initialized with 5×10^{-3}, modulated by a CosineAnnealing learning rate scheduler (Loshchilov & Hutter, 2016). We terminate the algorithm when it reaches 3000 epochs.

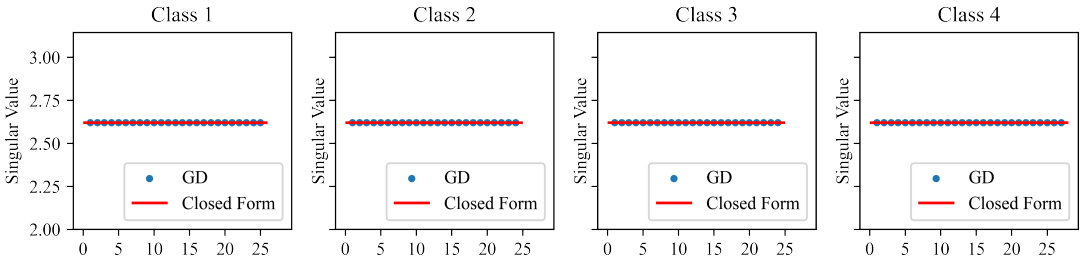


Figure 6. **The number and magnitude of singular values in each subspace in Experiment 1.** The blue dots are plotted based on the singular values by applying SVD to the solution returned by GD, and the red line is plotted according to the closed-form solution in (7). The number of singular values in each subspace is $25, 24, 24, 27$, respectively.

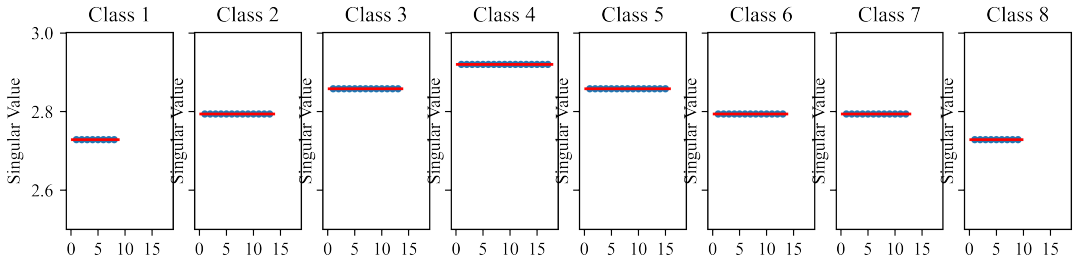


Figure 7. **The number and magnitude of singular values in each subspace in Experiment 2.** The blue dots are plotted based on the singular values by applying SVD to the solution returned by GD, and the red line is plotted according to the closed-form solution in (7). The number of singular values in each subspace is $8, 13, 13, 17, 15, 13, 12, 9$, respectively.

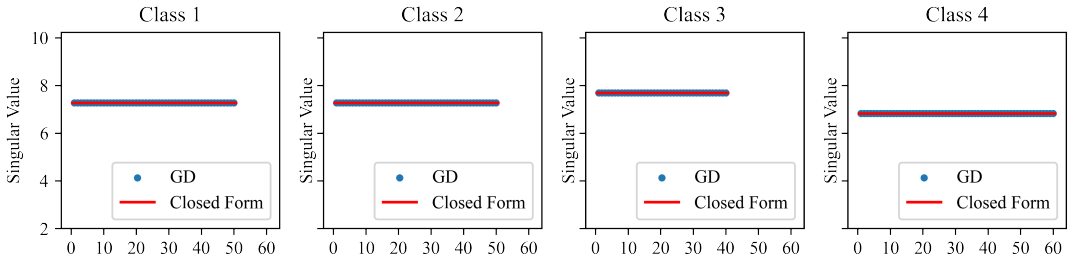
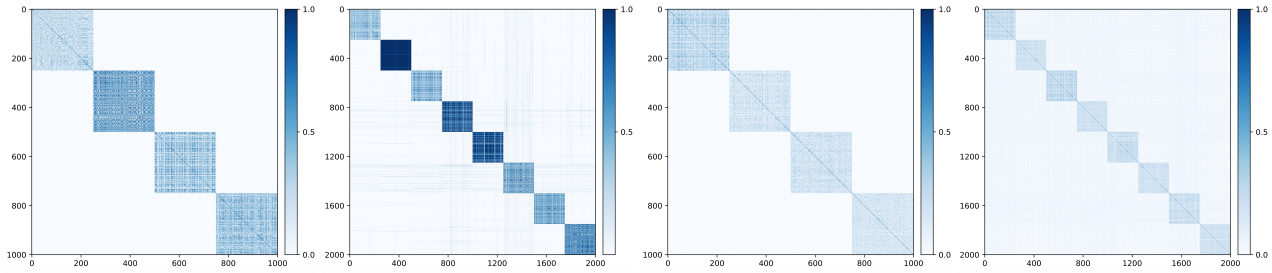


Figure 8. **The number and magnitude of singular values in each subspace in Experiment 3.** The blue dots are plotted based on the singular values by applying SVD to the solution returned by GD, and the red line is plotted according to the closed-form solution in (7). The number of singular values in each subspace is $50, 50, 40, 60$, respectively.

More experimental results on MNIST. Besides the numerical results in Table 1, we also plot the heatmap of the cosine similarity between pairwise columns of the features in $\boldsymbol{Z}$ obtained by training deep networks in Figure 9. We observe that the features from different classes are nearly orthogonal to each other, while those from the same classes are highly correlated. This supports our results in Theorem 3.1.

Network architecture and training setups for CIFAR-10. In the experiments on CIFAR10, we employ a 2-layer multilayer perceptron (MLP) network with ReLU activation as the feature mapping with the intermediate dimension 4096 and output dimension 128. In particular, each layer of MLP networks consists of a linear layer and layer norm layer followed by ReLU activation in the implementation. We train the network parameters via Adam by optimizing the MCR^2 function. For the Adam settings, we use a momentum of 0.9, a full-batch size, and a dynamically adaptive learning rate initialized with 5×10^{-3}, modulated by a CosineAnnealing learning rate scheduler (Loshchilov & Hutter, 2016). We terminate the algorithm when it reaches 4000 epochs.

More experimental results on CIFAR-10. Besides the numerical results in Table 1, we also plot the heatmap of the cosine similarity between pairwise columns of the features in $\boldsymbol{Z}$ obtained by training deep networks in Figure 9. We observe that the features from different classes are nearly orthogonal to each other, while those from the same classes are highly correlated. This supports our results in Theorem 3.1.



(a) MNIST: $m = 1000, K = 4$ (b) MNIST: $m = 2000, K = 8$ (c) CIFAR: $m = 1000, K = 4$ (d) CIFAR: $m = 2000, K = 8$

Figure 9. **Heatmap of cosine similarity among learned features by training deep networks on MNIST and CIFAR-10.** We train network parameters by optimizing the regularized MCR^2 objective (5) on m samples split equally among K classes of MNIST and CIFAR-10. In the figure, the darker pixels represent higher cosine similarity between features. In particular, when the (i, j)-th pixel is close to 0 (very light blue), the features i and j are approximately orthogonal.