A Local Geometric Analysis of Maximal Coding Rate Reduction via Error Bounds

Peng Wang[*]
Department of Artificial Intelligence &
Institute of Artificial Intelligence and Brain Sciences
University of Macau
pengw@um.edu.mo

Po Chen[*]
School of Data Science
Fudan University
chenp24@m.fudan.edu.cn

Huikang Liu
Antai College of Economics and Management
Shanghai Jiao Tong University
hkl1u@sjtu.edu.cn

Rujun Jiang[†]
School of Data Science
Fudan University
rjjiang@fudan.edu.cn

Abstract

While the maximal coding rate reduction (MCR^2) objective has become a useful principle for learning compact and discriminative representations, a theoretical understanding of its optimization geometry remains limited, especially on why first-order methods often exhibit fast convergence near its local maximizers. In this work, we show that the MCR^2 objective satisfies an error-bound property in a neighborhood of its local maximizers. This property guarantees that the distance to the local maximizer set can be bounded above by the gradient norm at any point in the neighborhood, directly implying regularity conditions such as the Polyak–Łojasiewicz inequality and quadratic growth. Leveraging this property, we prove that first-order methods converge linearly to a local maximizer of the MCR^2 objective under mild conditions. Numerical experiments validate the predicted local convergence behavior and provide empirical evidence that optimizing the MCR^2 objective improves representation geometry while maintaining competitive downstream accuracy.

1 Introduction

In deep representation learning, the maximal coding rate reduction (MCR^2) objective [Ma et al., 2007, Yu et al., 2020], grounded in the lossy coding rate, has attracted increasing attention in recent years. Thanks to the remarkable performance in learning compact and structured representations that facilitate downstream tasks, MCR^2-based approaches have been widely applied to multi-class classification [Wen et al., 2025, Yu et al., 2020], image segmentation [Chan et al., 2022, Meng et al., 2025, Yu et al., 2024b], and image and language generation [Yu et al., 2024a, Wang et al., 2025]. In particular, neural network architectures obtained by unrolling algorithmic updates for optimizing the MCR^2 objective, such as ReduNet [Chan et al., 2022], CRATE [Yu et al., 2024a], and their variants [Yang et al., 2024, Pai et al., 2024, Wu et al., 2025a], have been demonstrated to work effectively in practice.

As a guiding principle for learning linear discriminative representations, the MCR^2 objective has achieved remarkable empirical success. Nevertheless, due to the highly complicated and non-convex nature of the objective (see Problem (3)), the theoretical understanding of its optimization

[*]The first two authors contributed equally to this work.
[†]Corresponding author.

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

properties remains limited. A seminal work [Chan et al., 2022] proved that the global solutions of the MCR^2 problem are within-class compressible, between-class discriminative, and maximally diverse. Recently, Wang et al. [2024] explicitly characterized the set of local maximizers of the regularized MCR^2 problem. Moreover, they showed that this problem has a benign loss landscape in the sense that every critical point is either a local maximizer or a strict saddle point (see Fact 1). More interestingly, they observed that gradient ascent applied to this problem can achieve linear convergence, but a formal theoretical justification remains elusive. In addition, MCR^2-based neural network architectures have been shown to be effective at learning feature representations [Chan et al., 2022, Wang et al., 2025, Yu et al., 2024a]. These observations naturally raise the question of whether the MCR^2 objective possesses favorable optimization properties that ensure efficient convergence when optimizing it, while effectively learning meaningful feature representations when unrolled as a neural network.

1.1 Main Contributions

In this work, we study the local optimization geometry of the MCR^2 objective around its local maximizers. This local perspective is important: since prior work shows that first-order methods with random initialization almost surely avoid strict saddles [Lee et al., 2016, 2019], the key question is not only where the iterates may converge, but also how regular the landscape is once the trajectory enters a basin of a local maximizer. Our work addresses this question by establishing local regularity of the MCR^2 landscape and using it to explain the fast local convergence of first-order methods near local maximizers. Our contributions are summarized as follows:

- Despite the complexity and inherent non-convexity of the MCR^2 problem, we prove that it satisfies a local error-bound property [Luo and Tseng, 1993, Pang, 1997, Zhou and So, 2017] in a neighborhood of its local maximizer set; see Theorem 1. Specifically, we show that for any point $\boldsymbol{Z}$ in this neighborhood, the distance from $\boldsymbol{Z}$ to the local maximizer set is controlled by the gradient norm. This error-bound property captures favorable local geometric regularity of the objective and further implies standard regularity conditions, including the Polyak–Lojasiewicz (PŁ) inequality [Karimi et al., 2016, Polyak, 1963] and quadratic growth [Bonnans and Ioffe, 1995, Drusvyatskiy and Lewis, 2018].
- Leveraging the error-bound property and the benign loss landscape of the MCR^2 problem (see Fact 1), we show that first-order methods for optimizing the MCR^2 objective converge to a local maximizer at a local linear rate (see Theorem 2). This result explains the fast local convergence of gradient-based methods observed in practice and shows that, once the iterates enter a neighborhood of the local maximizer set, the MCR^2 objective has sufficient regularity to ensure geometric decay of the optimization error. It further provides theoretical support for MCR^2-based representation learning methods and unrolled architectures [Chan et al., 2022, Wang et al., 2025, Wu et al., 2025b, Pai et al., 2024], where first-order updates are used as building blocks for learning discriminative feature representations.

To support our theory, we conduct numerical experiments on both synthetic and real-world datasets. The results corroborate the predicted local convergence behavior and further illustrate the connection between optimizing the MCR^2 objective and learning discriminative feature representations.

1.2 Related Works

Deep representation learning using MCR^2. Current approaches using MCR^2 for learning representations can be broadly categorized into two types. The first uses MCR^2 as a training objective to learn discriminative and compact representations for different tasks. For example, an early work [Ma et al., 2007] demonstrated the effectiveness of the MCR^2 objective in classical clustering problems. Recently, Yu et al. [2020] optimized the MCR^2 objective to train a neural network for multi-class classification tasks. The work of Pai et al. [2023] leverages MCR^2 to learn discriminative representations for data drawn from multiple subspaces by formulating the problem as a sequential game. The work of Meng et al. [2025] applies MCR^2 in temporal rate reduction clustering to segment human motion sequences.

The second approach is to unroll neural networks based on algorithmic updates for optimizing the MCR^2 objective or its variants. For example, Yu et al. [2020] proposed a neural network named ReduNet by unrolling gradient ascent for optimizing the MCR^2 objective, enabling interpretable and

efficient representation learning. Recently, Yu et al. [2024a] designed a transformer-like network architecture named CRATE by unrolling alternating minimization for optimizing the sparse rate reduction objective. Using different objectives or different optimization algorithms, Wang et al. [2025] proposed an attention-only transformer-like architecture, while Wu et al. [2025a] developed a linear-time variant of a transformer-like architecture. Please refer to [Pai et al., 2024, Wang et al., 2025, Wu et al., 2025b, Yang et al., 2024] for additional neural network architecture designs based on the MCR^2 objective.

Error bound analysis of non-convex problems. Error bounds have been extensively studied as a fundamental tool for characterizing the local geometry of optimization problems and establishing convergence guarantees, such as *linear* convergence and *superlinear* convergence rates, for iterative algorithms [Luo and Tseng, 1993, Yue et al., 2019, Zhou and So, 2017]. While early works focused primarily on convex settings [Luo and Tseng, 1993, Zhou and So, 2017], recent studies have shown that various classes of non-convex problems also satisfy error-bound properties under structural assumptions. For example, Liu et al. [2019] showed that quadratic programs with orthogonality constraints satisfy an error-bound property, and Wang et al. [2023] showed the same result for linear programs with orthogonality constraints. Jiang and Li [2022] proved an error bound for the trust region subproblem. Recently, Chen et al. [2025] established error bounds for the regularized loss of deep matrix factorization problems. As demonstrated in these works, establishing error-bound properties generally requires dedicated analyses and advanced technical tools.

Notation. For a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\|\boldsymbol{A}\|$ denotes its spectral norm, $\|\boldsymbol{A}\|_F$ its Frobenius norm, and a_{ij} its (i,j)-th entry. For a vector $\boldsymbol{a} \in \mathbb{R}^d$, $\|\boldsymbol{a}\|$ denotes its ℓ_2 norm, a_i its i-th entry, and $\mathrm{diag}(\boldsymbol{a})$ the diagonal matrix with $\boldsymbol{a}$ on the diagonal. Given a positive integer n, we denote by $[n]$ the set $\{1,\dots,n\}$. We denote by $\mathrm{BlkD}(\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)$ the block diagonal matrix with diagonal blocks $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$. Given a set of numbers $\{a_k\}_{k=1}^K$, let $a_{\max} := \max\{a_k : k \in [K]\}$ and $a_{\min} := \min\{a_k : k \in [K]\}$. Let $\mathcal{O}^{m\times n} := \left\{\boldsymbol{X} \in \mathbb{R}^{m\times n} : \boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_n\right\}$ denote the set of all $m \times n$ orthonormal matrices. In particular, we write $\mathcal{O}^m := \mathcal{O}^{m\times m}$ to denote the set of all $m \times m$ orthogonal matrices. Given a point $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, its distance to a non-empty and closed set $\mathcal{X} \subseteq \mathbb{R}^{m\times n}$ is defined as $\mathrm{dist}(\boldsymbol{X},\mathcal{X}) := \min_{\boldsymbol{Y}\in\mathcal{X}} \|\boldsymbol{X}-\boldsymbol{Y}\|_F$.

Definition 1. *Consider the maximization of a twice continuously differentiable function F. We say*
(i) $\boldsymbol{Z}^*$ is a critical point if $\nabla F(\boldsymbol{Z}^*) = \boldsymbol{0}$;
(ii) $\boldsymbol{Z}^*$ is a local maximizer if $F(\boldsymbol{Z}^*) \geq F(\boldsymbol{Z})$ for all $\boldsymbol{Z}$ in some neighborhood of $\boldsymbol{Z}^*$;
(iii) $\boldsymbol{Z}^*$ is a strict saddle point if it is a critical point and its Hessian has at least one strictly positive eigenvalue.[3]

2 Main Results

2.1 Problem Setup

To begin, we introduce the MCR^2 objective for supervised learning problems. Consider a training dataset $\{\boldsymbol{x}_i\}_{i=1}^N \subseteq \mathbb{R}^n$ drawn from $K \geq 2$ classes. Let $\boldsymbol{z}_i \in \mathbb{R}^d$ be the feature representation of $\boldsymbol{x}_i$ obtained via a feature mapping, such as a neural network, for each $i \in [N]$. Without loss of generality, we assume that the samples are organized in a class-by-class manner. For each $k \in [K]$, let N_k be the number of samples in the k-th class and $\boldsymbol{Z}_k \in \mathbb{R}^{d\times N_k}$ be the matrix whose columns are the feature representations of the samples in the k-th class. We then define

$$\boldsymbol{Z} := [\boldsymbol{Z}_1 \quad \cdots \quad \boldsymbol{Z}_K] \in \mathbb{R}^{d\times N}. \tag{1}$$

Moreover, let

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

for each $k \in [K]$, where $\epsilon > 0$ is a coding precision. Notably, $R(\boldsymbol{Z})$ approximates the rate-distortion function of a Gaussian source when ϵ is sufficiently small [Cover and Thomas, 2006]. As argued in [Chan et al., 2022, Wang et al., 2024], $R(\boldsymbol{Z})$ measures the global coding rate of all features and therefore rewards a representation that spans a large volume, which encourages diversity and

[3] As defined in Lee et al. [2019], our definition of a *strict saddle point* allows for the inclusion of local minima.

separation across classes. In contrast, $R_c(\boldsymbol{Z}_k)$ measures the coding rate within the k-th class; subtracting $\sum_{k=1}^K R_c(\boldsymbol{Z}_k)$ penalizes within-class variability and encourages features from the same class to be compressible. Thus, maximizing the difference between the global and classwise coding rates promotes between-class discrimination and within-class compactness. This motivates the following MCR^2 problem:

$$\max_{\boldsymbol{Z}\in\mathbb{R}^{d\times N}} F(\boldsymbol{Z}) := R(\boldsymbol{Z}) - \sum_{k=1}^K R_c(\boldsymbol{Z}_k) - \frac{\lambda}{2}\|\boldsymbol{Z}\|_F^2, \tag{3}$$

where $\lambda > 0$ is a regularization parameter.

[Wang et al., 2024, Theorem 1] has shown that any local optimal solution of this problem satisfies the property that features from the same class are compressible, while features from different classes are orthogonal to each other. Moreover, [Wang et al., 2024, Theorem 2] has proven that Problem (3) admits the benign loss landscape in the sense that each critical point is either a local maximizer or a strict saddle point. For completeness and to make the subsequent theorem statements self-contained, we recall these two results below.

Fact 1. *Given the coding precision $\epsilon > 0$, suppose that the regularization parameter λ satisfies*

$$\lambda \in \left(0, \frac{d(\sqrt{N/N_{\max}} - 1)}{N(\sqrt{N/N_{\max}} + 1)\epsilon^2}\right]. \tag{4}$$

(i) Every critical point of Problem (3) *is either a local maximizer or a strict saddle point.*
(ii) It holds that $\boldsymbol{Z} = [\boldsymbol{Z}_1 \ \cdots \ \boldsymbol{Z}_K]$ is a local maximizer if and only if each block $\boldsymbol{Z}_k$ admits the following singular value decomposition

$$\boldsymbol{Z}_k = \sigma_k^* \boldsymbol{U}_k \boldsymbol{V}_k^T, \tag{5}$$

where (a) $r_k := \operatorname{rank}(\boldsymbol{Z}_k)$ satisfies $r_k \in [0, \min\{N_k, d\})$ and $\sum_{k=1}^K r_k \le \min\{N, d\}$, (b) $\boldsymbol{U}_k \in \mathcal{O}^{d\times r_k}$ satisfies $\boldsymbol{U}_k^T\boldsymbol{U}_l = \boldsymbol{0}$ for all $l \neq k$ and $\boldsymbol{V}_k \in \mathcal{O}^{N_k\times r_k}$ for all $k \in [K]$,[4] and (c) the singular value

$$\sigma_k^* := \left(\frac{\eta_k + \sqrt{\eta_k^2 - 4\lambda^2 N/N_k}}{2\lambda\alpha_k}\right)^{1/2}, \quad \eta_k := (\alpha_k - \alpha) - \lambda\left(\frac{N}{N_k} + 1\right), \tag{6}$$

where $\alpha = d/(N\epsilon^2)$ and $\alpha_k = d/(N_k\epsilon^2)$.

Based on the above theorem, we denote the set of local maximizers by

$$\begin{aligned}\mathcal{Z} := \Big\{[\boldsymbol{Z}_1 \ \cdots \ \boldsymbol{Z}_K] :\ & \boldsymbol{Z}_k = \sigma_k^*\boldsymbol{U}_k\boldsymbol{V}_k^T,\ \boldsymbol{U}_k \in \mathcal{O}^{d\times r_k},\ \boldsymbol{V}_k \in \mathcal{O}^{N_k\times r_k},\\ & \boldsymbol{U}_k^T\boldsymbol{U}_l = \boldsymbol{0},\ \forall k \neq l, \quad r_k \in [0, \min\{N_k, d\}), \quad \sum_{k=1}^K r_k \le \min\{N, d\}\Big\}.\end{aligned} \tag{7}$$

In particular, Lee et al. [2019, 2016] have proved that first-order methods with random initialization, such as gradient ascent, almost surely escape strict saddle points. This, together with Fact 1(i), implies that the iterates generated by first-order methods almost surely converge to a local maximizer. Therefore, it suffices to restrict our attention to the local maximizers of Problem (3), as defined in (7), in the subsequent analysis.

2.2 Main Theorems

Based on the above setup, we establish the error-bound property for Problem (3) in a neighborhood of its local maximizers in the following theorem, whose proof is deferred to Appendix B.1.

Theorem 1. *Given the coding precision $\epsilon > 0$, suppose that the regularization parameter λ satisfies* (4)*. There exist constants $\delta, \kappa > 0$ such that for all $\boldsymbol{Z}$ satisfying $\operatorname{dist}(\boldsymbol{Z}, \mathcal{Z}) \le \delta$, it holds that*

$$\operatorname{dist}(\boldsymbol{Z}, \mathcal{Z}) \le \kappa\|\nabla F(\boldsymbol{Z})\|_F, \tag{8}$$

where $\mathcal{Z}$, defined in (7)*, denotes the set of all local maximizers of Problem* (3)*.*

[4] When $r_k = 0$, we define $\boldsymbol{U}_k = \boldsymbol{0}$ and $\boldsymbol{V}_k = \boldsymbol{0}$.

To clarify the theorem's significance, we make the following comments:

- The error-bound property (8) shows that, for any point $\boldsymbol{Z}$ in a neighborhood of the local maximizer set $\mathcal{Z}$, the distance from $\boldsymbol{Z}$ to $\mathcal{Z}$ is bounded by the gradient norm. This establishes a form of local geometric regularity for the MCR^2 objective, ensuring that the landscape around its local maximizers is well behaved despite the non-convexity of the problem. Together with Fact 1, this error-bound property provides a refined local picture of the MCR^2 landscape, connecting global saddle avoidance with local convergence behavior.
- As demonstrated in [Rebjock and Boumal, 2025, Liao et al., 2024], the error-bound property can be leveraged to establish other regularity properties, such as Polyak–Łojasiewicz (PŁ) inequality [Karimi et al., 2016, Polyak, 1963] and quadratic growth [Bonnans and Ioffe, 1995, Drusvyatskiy and Lewis, 2018]. Notably, all the above regularity properties, together with some algorithm-dependent properties, can be used to facilitate the convergence-rate analysis of first-order methods. Specifically, as shown in [Karimi et al., 2016, Luo and Tseng, 1993, Schneider and Uschmajew, 2015, Zhou and So, 2017], a linear convergence rate for first-order iterative methods can be established when the iterates satisfy the sufficient decrease condition, a cost-to-go estimate, and a safeguard condition.
- The proof is technically involved because $\operatorname{dist}(\boldsymbol{Z}, \mathcal{Z})$ does not admit a simple closed-form expression in this problem. Unlike error-bound analyses for problems with simpler solution sets [Liu et al., 2019, Wang et al., 2023, Jiang and Li, 2019], our argument constructs an intermediate point and separately controls its distance to $\boldsymbol{Z}$ and to $\mathcal{Z}$; see Section 3.1 and Appendix A.

Remark 1 (Error bounds from representation space to parameter space)**.** *Although Theorem 1 characterizes the local geometry of F when the representations are treated as free optimization variables, the resulting error bound does not automatically extend to optimization over neural-network parameters. We next discuss how this error bound transfers from representation space to parameter space. To this end, let $\boldsymbol{z}_i = \phi_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \in \mathbb{R}^d$ denote the feature representation of input $\boldsymbol{x}_i$, where $\phi_{\boldsymbol{\theta}}$ is a neural-network feature map parameterized by $\boldsymbol{\theta} \in \mathbb{R}^p$. For the fixed training set $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^N$, define*

$$Z_X(\boldsymbol{\theta}) := [\phi_{\boldsymbol{\theta}}(\boldsymbol{x}_1), \ldots, \phi_{\boldsymbol{\theta}}(\boldsymbol{x}_N)] \in \mathbb{R}^{d \times N}, \quad G(\boldsymbol{\theta}) := F(Z_X(\boldsymbol{\theta})),$$

where F is defined in (3)*. The parameterized problem therefore optimizes F only over the realizable representations $\mathcal{R}_X := \{Z_X(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^p\}$, rather than the entire representation space. Suppose that Z_X is continuously differentiable in the neighborhood under consideration, and write*

$$\boldsymbol{J}_{\boldsymbol{\theta}} := D(\operatorname{vec} \circ Z_X)(\boldsymbol{\theta}) \in \mathbb{R}^{dN \times p}, \quad \boldsymbol{g}(\boldsymbol{\theta}) := \operatorname{vec}\big(\nabla F(Z_X(\boldsymbol{\theta}))\big).$$

The chain rule gives $\nabla G(\boldsymbol{\theta}) = \boldsymbol{J}_{\boldsymbol{\theta}}^T \boldsymbol{g}(\boldsymbol{\theta})$.

To establish a transfer, define $\boldsymbol{\Theta} := \{\boldsymbol{\theta} \in \mathbb{R}^p : Z_X(\boldsymbol{\theta}) \in \mathcal{Z}\}$, assume that $\boldsymbol{\Theta} \neq \emptyset$, and fix $\bar{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$. Suppose that there exist constants $c_X, \gamma > 0$ and a neighborhood $\mathcal{U}$ of $\bar{\boldsymbol{\theta}}$ such that, for all $\boldsymbol{\theta} \in \mathcal{U}$,

$$\operatorname{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta}) \leq c_X \operatorname{dist}(Z_X(\boldsymbol{\theta}), \mathcal{Z}), \quad \|\boldsymbol{J}_{\boldsymbol{\theta}}^T \boldsymbol{g}(\boldsymbol{\theta})\| \geq \gamma \|\boldsymbol{g}(\boldsymbol{\theta})\|. \tag{9}$$

Since $Z_X(\bar{\boldsymbol{\theta}}) \in \mathcal{Z}$, continuity allows us to shrink $\mathcal{U}$ so that $\operatorname{dist}(Z_X(\boldsymbol{\theta}), \mathcal{Z}) \leq \delta$ throughout $\mathcal{U}$. Then Theorem 1, together with (9)*, yields for all $\boldsymbol{\theta} \in \mathcal{U}$,*

$$\operatorname{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta}) \leq c_X \operatorname{dist}(Z_X(\boldsymbol{\theta}), \mathcal{Z}) \leq c_X \kappa \|\boldsymbol{g}(\boldsymbol{\theta})\| \leq \frac{c_X \kappa}{\gamma} \|\nabla G(\boldsymbol{\theta})\|. \tag{10}$$

Thus, G satisfies a local error bound relative to $\boldsymbol{\Theta}$, with constant $c_X \kappa / \gamma$.

For example, both conditions hold locally if the empirical neural tangent kernel is uniformly positive definite:

$$\boldsymbol{J}_{\boldsymbol{\theta}} \boldsymbol{J}_{\boldsymbol{\theta}}^T \succeq \gamma^2 \boldsymbol{I}_{dN}, \quad \forall \boldsymbol{\theta} \in \mathcal{U}. \tag{11}$$

This implies the second inequality directly and, by local path lifting, the first inequality in (9) *with $c_X = 1/\gamma$, after shrinking $\mathcal{U}$ if necessary. Thus,* (10) *holds with constant κ/γ^2. This condition requires $p \geq dN$ and is an additional assumption on the parameterization, not a consequence of Theorem 1.*

By leveraging the benign loss landscape of Problem (3) and the local Lipschitz continuity of its gradient (see Lemma 5), we establish linear convergence of the iterates under the sufficient-ascent and safeguard conditions, without requiring an explicit cost-to-go estimate.

Theorem 2. *Suppose that the sequence $\{\boldsymbol{Z}^t\}_{t\geq 0}$ generated for solving Problem* (3) *satisfies the following properties:*
(i) There exists $\mu_1 > 0$ such that

$$F(\boldsymbol{Z}^{t+1}) - F(\boldsymbol{Z}^t) \geq \mu_1 \|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_F^2.$$

(ii) There exists $\mu_2 > 0$ such that

$$\|\nabla F(\boldsymbol{Z}^t)\|_F \leq \mu_2 \|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_F.$$

Suppose in addition that there exists a sufficiently large iteration $t_0 > 0$ such that $\boldsymbol{Z}^{t_0} \in \mathcal{U}$, where $\mathcal{U}$ is a neighborhood of $\mathcal{Z}$. Then, for all $t \geq t_0$, we have

$$F(\boldsymbol{Z}^*) - F(\boldsymbol{Z}^{t+1}) \leq w^2 \left(F(\boldsymbol{Z}^*) - F(\boldsymbol{Z}^t)\right), \tag{12}$$

$$\operatorname{dist}(\boldsymbol{Z}^t, \mathcal{Z}) \leq \left(\frac{F(\boldsymbol{Z}^*) - F(\boldsymbol{Z}^{t_0})}{\mu_1}\right)^{1/2} \frac{w^{t-t_0}}{1-w}, \tag{13}$$

where $\boldsymbol{Z}^* \in \mathcal{Z}$, L_F is defined in (57)*, and $w := \left(1 + \mu_1 / \left(2\kappa L_F^2(\kappa^2\mu_2^2 + 1)\right)\right)^{-1/2}$.*

The proof is deferred to Section 3.2 and Appendix B.2. This theorem shows that the sequence $\{F(\boldsymbol{Z}^t)\}_{t\geq t_0}$ converges Q-linearly to $F(\boldsymbol{Z}^*)$ and $\{\boldsymbol{Z}^t\}_{t\geq t_0}$ converges R-linearly to the local maximizer set $\mathcal{Z}$. Note that the contraction factor in Theorem 2 depends on two algorithmic constants, μ_1 and μ_2, and two problem-geometric constants L_F and κ. A larger sufficient-ascent constant μ_1 improves the rate, whereas larger values of μ_2, L_F, or κ make the bound more conservative. Since $L_F = \lambda + 9\alpha/2 = \lambda + 9d/(2N\epsilon^2)$ by Lemma 5, the smoothness part of the rate explicitly reflects the coding precision, feature dimension, sample size, and regularization strength. The geometric constant κ is characterized in Proposition 2 through the local singular-value and subspace geometry of the maximizer set.

For gradient ascent, $\boldsymbol{Z}^{t+1} = \boldsymbol{Z}^t + \eta\nabla F(\boldsymbol{Z}^t)$ with $\eta \leq 1/L_F$. By the L_F-smoothness inequality,

$$\begin{aligned} F(\boldsymbol{Z}^{t+1}) - F(\boldsymbol{Z}^t) &\geq \eta\|\nabla F(\boldsymbol{Z}^t)\|_F^2 - \frac{L_F\eta^2}{2}\|\nabla F(\boldsymbol{Z}^t)\|_F^2 \\ &= \left(\frac{1}{\eta} - \frac{L_F}{2}\right)\|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_F^2 \geq \frac{L_F}{2}\|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_F^2. \end{aligned}$$

Thus condition (i) holds with $\mu_1 = L_F/2$. Moreover, the update gives $\|\nabla F(\boldsymbol{Z}^t)\|_F = \eta^{-1}\|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_F$, so condition (ii) holds with $\mu_2 = 1/\eta$. Combined with the strict-saddle avoidance property of randomly initialized gradient ascent [Lee et al., 2019] and the landscape in Fact 1, this yields the following corollary. We defer its formal proof to Appendix B.3.

Corollary 1. *Consider gradient ascent with random initialization for solving Problem* (3)*:*

$$\boldsymbol{Z}^{t+1} = \boldsymbol{Z}^t + \eta\nabla F(\boldsymbol{Z}^t),\ t = 0, 1, 2, \ldots,$$

where $\eta > 0$ is a constant step size with $\eta \leq 1/L_F$ and L_F is defined in (57)*. Then, it holds almost surely that gradient ascent converges to a local maximizer $\boldsymbol{Z}^*$ and there exists a $t_0 > 0$ such that* (12) *and* (13) *hold for all $t \geq t_0$, where $w = \left(1 + 1/\left(4\kappa L_F(\kappa^2/\eta^2 + 1)\right)\right)^{-1/2}$.*

Finally, Chan et al. [2022], Wang et al. [2024], Yu et al. [2024a] constructed a white-box neural network architecture by unrolling iterative optimization algorithms, such as gradient ascent, to optimize the MCR^2 and its variants. These unrolled networks have been shown to be highly effective for learning deep representations. Our characterization in Theorem 1 sheds light on their effectiveness by revealing favorable geometric properties of the MCR^2 landscape.

3 Proof Overview of Main Results

In this section, we outline the proof strategy, deferring the technical ingredients to Appendix A and the proofs of the main theorems to Appendix B. First, we prove the local error-bound property in Theorem 1 by showing that the gradient norm controls deviations from the local-maximizer geometry, including excess rank, leading singular-value errors, and non-orthogonality between class subspaces. Second, we combine this geometric error bound with sufficient ascent and a safeguard condition to prove the local linear convergence result in Theorem 2.

3.1 Proof Sketch of Error Bound

The main difficulty in proving Theorem 1 is that the local-maximizer set $\mathcal{Z}$ is non-isolated and does not admit a simple projection formula. Therefore, instead of directly computing $\mathrm{dist}(\boldsymbol{Z}, \mathcal{Z})$, we construct an explicit comparison point in $\mathcal{Z}$ from the singular structure of a nearby point $\boldsymbol{Z}$. The proof proceeds in five steps.

Step 1: Singular value decomposition. Fix $\boldsymbol{Z}^* \in \mathcal{Z}$, and let $r_k := \mathrm{rank}(\boldsymbol{Z}_k^*)$ and $r := \sum_{k=1}^K r_k$. For $\boldsymbol{Z} = [\boldsymbol{Z}_1 \ \cdots \ \boldsymbol{Z}_K]$ sufficiently close to $\boldsymbol{Z}^*$, consider the SVD

$$\boldsymbol{Z}_k = \boldsymbol{P}_k \boldsymbol{\Sigma}_k \boldsymbol{Q}_k^T = \begin{bmatrix} \boldsymbol{P}_k^{(1)} & \boldsymbol{P}_k^{(2)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_k^{(1)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_k^{(2)} \end{bmatrix} \begin{bmatrix} \boldsymbol{Q}_k^{(1)^T} \\ \boldsymbol{Q}_k^{(2)^T} \end{bmatrix}, \ \forall k \in [K]. \tag{14}$$

Here, $\boldsymbol{P}_k \in \mathcal{O}^d$, $\boldsymbol{P}_k^{(1)} \in \mathcal{O}^{d \times r_k}$, and $\boldsymbol{P}_k^{(2)} \in \mathcal{O}^{d \times (d-r_k)}$; $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times N_k}$, $\boldsymbol{\Sigma}_k^{(1)} = \mathrm{diag}(\sigma_{k,1}, \ldots, \sigma_{k,r_k})$ with $\sigma_{k,1} \geq \cdots \geq \sigma_{k,r_k}$, and $\boldsymbol{\Sigma}_k^{(2)} \in \mathbb{R}^{(d-r_k)\times(N_k-r_k)}$; $\boldsymbol{Q}_k \in \mathcal{O}^{N_k}$, $\boldsymbol{Q}_k^{(1)} \in \mathcal{O}^{N_k \times r_k}$, and $\boldsymbol{Q}_k^{(2)} \in \mathcal{O}^{N_k \times (N_k - r_k)}$. We also define the truncated and compressed representations $\boldsymbol{Z}_k^{(1)} := \boldsymbol{P}_k^{(1)} \boldsymbol{\Sigma}_k^{(1)} \boldsymbol{Q}_k^{(1)^T}$, $\boldsymbol{Z}^{(1)} := [\boldsymbol{Z}_1^{(1)} \ \cdots \ \boldsymbol{Z}_K^{(1)}] \in \mathbb{R}^{d \times N}$, $\bar{\boldsymbol{Z}}_k := \boldsymbol{P}_k^{(1)} \boldsymbol{\Sigma}_k^{(1)}$, and $\bar{\boldsymbol{Z}} := [\bar{\boldsymbol{Z}}_1 \ \cdots \ \bar{\boldsymbol{Z}}_K] \in \mathbb{R}^{d \times r}$.

Step 2: Reduction to three aggregate estimates. We reduce the desired error bound to three geometric estimates by introducing a comparison point in $\mathcal{Z}$. For any collection of matrices $\hat{\boldsymbol{U}}_k \in \mathcal{O}^{d \times r_k}$ satisfying $\hat{\boldsymbol{U}}_k^T \hat{\boldsymbol{U}}_l = \boldsymbol{0}$ for all $k \neq l$, define $\hat{\boldsymbol{Z}}_k := \sigma_k^* \hat{\boldsymbol{U}}_k \boldsymbol{Q}_k^{(1)^T}$ and $\hat{\boldsymbol{Z}} := [\hat{\boldsymbol{Z}}_1 \ \cdots \ \hat{\boldsymbol{Z}}_K]$. By Fact 1(ii), we have $\hat{\boldsymbol{Z}} \in \mathcal{Z}$. Using the SVD in (14), the triangle inequality, and the orthogonal invariance of the Frobenius norm, we obtain

$$\begin{aligned} \mathrm{dist}(\boldsymbol{Z}, \mathcal{Z}) &\leq \|\boldsymbol{Z} - \hat{\boldsymbol{Z}}\|_F \leq \sum_{k=1}^K \|\boldsymbol{Z}_k - \hat{\boldsymbol{Z}}_k\|_F \\ &\leq \sum_{k=1}^K \Big(\underbrace{\|\boldsymbol{\Sigma}_k^{(1)} - \sigma_k^* \boldsymbol{I}_{r_k}\|_F}_{A_k} + \underbrace{\sigma_k^* \|\boldsymbol{P}_k^{(1)} - \hat{\boldsymbol{U}}_k\|_F}_{B_k} + \underbrace{\|\boldsymbol{\Sigma}_k^{(2)}\|_F}_{C_k} \Big). \end{aligned} \tag{15}$$

It therefore remains to construct $\{\hat{\boldsymbol{U}}_k\}_{k=1}^K$ satisfying the above orthogonality conditions and to bound each of the aggregate quantities $\sum_{k=1}^K A_k$, $\sum_{k=1}^K B_k$, and $\sum_{k=1}^K C_k$ by a constant multiple of $\|\nabla F(\boldsymbol{Z})\|_F$.

Step 3: Controlling the tail singular values. We first control the tail singular values, which vanish at $\boldsymbol{Z}^*$. By Lemma 2, we have for $\boldsymbol{Z}$ sufficiently close to $\boldsymbol{Z}^*$,

$$\sigma_i(\boldsymbol{Z}_k) \leq \frac{2}{\lambda} \|\nabla F(\boldsymbol{Z})\|_F, \quad \forall r_k < i \leq \min\{d, N_k\}, \ k \in [K].$$

Consequently, the tail contribution of each class satisfies

$$C_k = \|\boldsymbol{\Sigma}_k^{(2)}\|_F = \left(\sum_{i=r_k+1}^{\min\{d,N_k\}} \sigma_i(\boldsymbol{Z}_k)^2 \right)^{1/2} \leq \frac{2\sqrt{\min\{d, N_k\} - r_k}}{\lambda} \|\nabla F(\boldsymbol{Z})\|_F.$$

Summing over classes yields

$$\sum_{k=1}^K C_k \leq c_C \|\nabla F(\boldsymbol{Z})\|_F, \quad c_C := \frac{2}{\lambda} \sum_{k=1}^K \sqrt{\min\{d, N_k\} - r_k} \leq \frac{2K\sqrt{d}}{\lambda}.$$

Step 4: Controlling the leading singular values and cross-subspace products. For $\boldsymbol{Z}$ sufficiently close to $\boldsymbol{Z}^*$, the tail estimate and (42) give $\|\nabla F(\boldsymbol{Z}^{(1)})\|_F \leq c_T \|\nabla F(\boldsymbol{Z})\|_F$ for some $c_T > 0$. Using Lemma 6, we express the truncated gradient in terms of $\bar{\boldsymbol{Z}}^T \bar{\boldsymbol{Z}}$ and bound the associated matrix residual by the gradient norm. The off-diagonal blocks control the cross-subspace products $\boldsymbol{P}_k^{(1)^T} \boldsymbol{P}_l^{(1)}$ via Lemma 1, while the diagonal blocks control $|\varphi_k(\sigma_{k,i})|$, where

$$\varphi_k(x) := \frac{\alpha}{1 + \alpha x^2} - \frac{\alpha}{1 + \alpha_k x^2} - \lambda, \qquad \varphi_k(\sigma_k^*) = 0.$$

The local separation estimate $|\varphi_k(x)| \geq \kappa_3 |x - \sigma_k^*|$ for $|x - \sigma_k^*| \leq \delta_1$ in Lemma 7 then bounds the leading singular-value errors. Combining these estimates as in the proof of Proposition 1 yields

$$\sum_{k=1}^{K} \|\boldsymbol{\Sigma}_k^{(1)} - \sigma_k^* \boldsymbol{I}_{r_k}\|_F + \sum_{k=1}^{K} \sum_{l \neq k} \|{\boldsymbol{P}_k^{(1)}}^T \boldsymbol{P}_l^{(1)}\|_F \leq c_2 \|\nabla F(\boldsymbol{Z})\|_F \tag{16}$$

for a local constant $c_2 > 0$. This bounds $\sum_k A_k$ and provides the cross-subspace control needed for Step 5.

Step 5: Orthogonalization and construction of a nearby maximizer. We use the blockwise orthogonalization in Lemma 3. Set $\hat{\boldsymbol{U}}_1 := \boldsymbol{P}_1^{(1)}$ and, for $k \geq 2$, define

$$\boldsymbol{H}_k := \left(\boldsymbol{I}_d - \sum_{l=1}^{k-1} \hat{\boldsymbol{U}}_l \hat{\boldsymbol{U}}_l^T \right) \boldsymbol{P}_k^{(1)}.$$

Given a thin SVD $\boldsymbol{H}_k = \tilde{\boldsymbol{U}}_k \tilde{\boldsymbol{\Sigma}}_k \tilde{\boldsymbol{V}}_k^T$, set $\hat{\boldsymbol{U}}_k := \tilde{\boldsymbol{U}}_k \tilde{\boldsymbol{V}}_k^T$, omitting empty blocks. For $\boldsymbol{Z}$ sufficiently close to $\boldsymbol{Z}^*$, Lemma 3 ensures that the resulting matrices $\hat{\boldsymbol{U}}_k \in \mathcal{O}^{d \times r_k}$ have mutually orthogonal column spaces and satisfy

$$\|\boldsymbol{P}_k^{(1)} - \hat{\boldsymbol{U}}_k\|_F \leq c_3 \|\nabla F(\boldsymbol{Z})\|_F, \quad \forall k \in [K],$$

for a local constant $c_3 > 0$. Consequently, we have

$$\sum_{k=1}^{K} B_k \leq c_3 \sum_{k=1}^{K} \sigma_k^* \|\nabla F(\boldsymbol{Z})\|_F.$$

Combining this estimate with Steps 3–4 in (15) gives a local error bound around $\boldsymbol{Z}^*$, as formalized in Proposition 2. Finally, a finite-covering argument over the compact set $\mathcal{Z}$ yields uniform constants $\delta, \kappa > 0$, completing the proof of Theorem 1; see Appendix B.1.

3.2 Proof Sketch of Convergence Rate

The previous subsection provides the key geometric ingredient for the convergence analysis: near the local-maximizer set $\mathcal{Z}$, the distance to $\mathcal{Z}$ is controlled by the gradient norm. We use this error bound after the iterates enter a neighborhood of one connected component $\mathfrak{X}$ of $\mathcal{Z}$. This localization is necessary because $\mathcal{Z}$ is generally non-isolated, while different connected components are separated by the landscape characterization in Fact 1. Once the iterates enter a sufficiently small neighborhood of $\mathfrak{X}$, the sufficient-ascent condition ensures monotonicity of the objective, and a standard Lyapunov-stability argument keeps the trajectory inside the same basin.

For $t \geq t_0$, let $e_t := F^* - F(\boldsymbol{Z}^t) \geq 0$ and $s_t := \|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_F$. Within this neighborhood, the local PŁ estimate (66), smoothness, and the error-bound and safeguard conditions yield the cost-to-go estimate

$$\begin{aligned} e_{t+1} &\leq \kappa \|\nabla F(\boldsymbol{Z}^{t+1})\|_F^2 \leq \kappa L_F^2 \mathrm{dist}^2(\boldsymbol{Z}^{t+1}, \mathfrak{X}) \\ &\leq 2\kappa L_F^2 \big(\mathrm{dist}^2(\boldsymbol{Z}^t, \mathfrak{X}) + s_t^2 \big) \leq 2\kappa L_F^2 (\kappa^2 \mu_2^2 + 1) s_t^2. \end{aligned}$$

Combining this with sufficient ascent, $e_t - e_{t+1} \geq \mu_1 s_t^2$, gives

$$e_{t+1} \leq w^2 e_t, \quad w := \left(1 + \frac{\mu_1}{2\kappa L_F^2 (\kappa^2 \mu_2^2 + 1)} \right)^{-1/2} \in (0, 1),$$

which proves the Q-linear objective-gap decay in (12). Moreover, sufficient ascent gives $s_t \leq \sqrt{e_{t_0}/\mu_1}\, w^{t-t_0}$. These step lengths are summable, so the iterates converge. The bound $\mathrm{dist}(\boldsymbol{Z}^t, \mathfrak{X}) \leq \kappa \mu_2 s_t \to 0$ ensures that their limit belongs to $\mathfrak{X}$. Summing the geometric tail gives the R-linear bound in (13), completing the proof sketch of Theorem 2.

For fixed-step gradient ascent with $0 < \eta \leq 1/L_F$, smoothness and the update identity verify the sufficient-ascent and safeguard conditions with $\mu_1 = L_F/2$ and $\mu_2 = 1/\eta$, respectively. Substituting these constants gives the local rates in Corollary 1 after sufficiently late entry into $\mathcal{U}$.

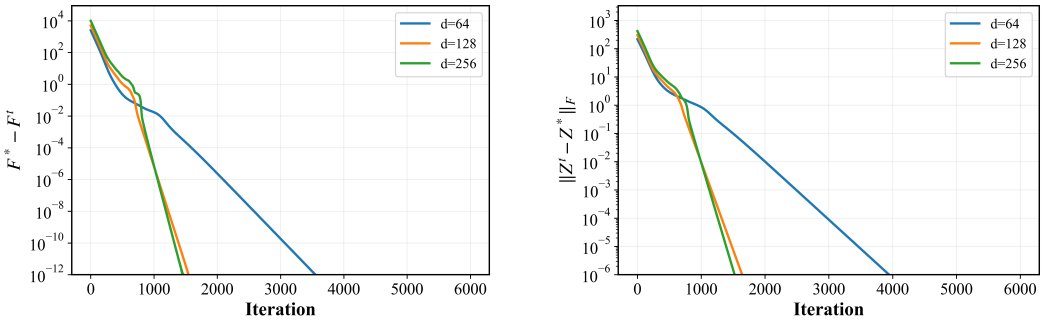


Figure 1: **Visualization of the convergence rate of gradient ascent for solving Problem** (3)**.** In both figures, the x-axis is the number of iterations, denoted by t. (Left) We plot the function value gap $F^* - F^t$ against number of iterations, where $F^t = F(\boldsymbol{Z}^t)$ and $F^* = F(\boldsymbol{Z}^*)$. (Right) We plot the iterate gap $\|\boldsymbol{Z}^t - \boldsymbol{Z}^*\|_F$ against number of iterations, where $\boldsymbol{Z}^*$ denotes the last iterate.

4 Experimental Results

In this section, we conduct numerical experiments on synthetic datasets in Section 4.1 and real-world datasets in Section 4.2 in support of our theoretical findings. Since the main contribution of this paper is a convergence-rate theory for the MCR^2 objective, the primary quantities reported here are the objective gap and the iterate gap against the number of iterations. We also provide additional experiments in Appendix D to show how optimizing the MCR^2 objective relates to feature geometry and linear-probe performance.

4.1 Convergence Rate of Gradient Ascent for MCR^2

In this subsection, we apply gradient ascent (GA) for solving Problem (3). The problem settings are as follows: We set the number of classes $K = 4$ and the number of data samples $N_k = 200$ for each $k \in [K]$, and vary the data dimension $d \in \{64, 128, 256\}$. Moreover, we set the regularization parameter $\lambda = 0.1$ and the coding precision $\epsilon = 0.5$. For the implementation of GA, the iterate is initialized using a standard Gaussian random distribution and adopts a constant step size of 0.1.

To evaluate the convergence rate of GA in terms of function values (resp., iterate gap), we plot $F(\boldsymbol{Z}^*) - F(\boldsymbol{Z}^t)$ (resp., $\|\boldsymbol{Z}^t - \boldsymbol{Z}^*\|_F$) against the number of iterations. Here, $\boldsymbol{Z}^t$ denotes the t-th iterate and $\boldsymbol{Z}^*$ denotes the last iterate generated by GA, which serves as a surrogate for a local maximizer.[5] It is observed that GA with random initialization converges to a local maximizer at a linear rate in terms of both the function value gap and the iterate distance. This is consistent with the linear convergence guarantee established in Theorem 2, which in turn provides further numerical evidence supporting the error-bound condition in Theorem 1. This controlled synthetic experiment is designed to isolate the local linear regime predicted by the theory; the resulting trajectories are shown in Figure 1.

4.2 Convergence Rate of Training Deep Networks via MCR^2

In this subsection, we extend our empirical studies to train neural networks for classification problems on real-world datasets, namely MNIST LeCun et al. [1998] and CIFAR-10 Krizhevsky and Hinton [2009]. Let $\{\boldsymbol{x}_i\}_{i=1}^N \subseteq \mathbb{R}^D$ denote the data samples. We employ a neural network $f_{\boldsymbol{\theta}}(\cdot) : \mathbb{R}^D \to \mathbb{R}^d$ to learn data representations $\boldsymbol{z}_i = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ for each $i \in [N]$. The network parameters $\boldsymbol{\theta}$ are trained by optimizing Problem (3).

In our implementations, we employ an over-parameterized multi-layer perceptron (MLP) network with residual connections to parameterize $f_{\boldsymbol{\theta}}$. We adopt a three-layer MLP with ReLU activations and optimize the MCR^2 objective using the AdamW optimizer equipped with a cosine annealing learning rate scheduler Loshchilov and Hutter [2019]. To ensure optimization stability, the schedule incorporates a linear warmup for the initial 5% of iterations, followed by a cosine decay towards a

[5] According to Lee et al. [2019], Wang et al. [2024], gradient ascent with random initialization converges to a local maximizer almost surely.

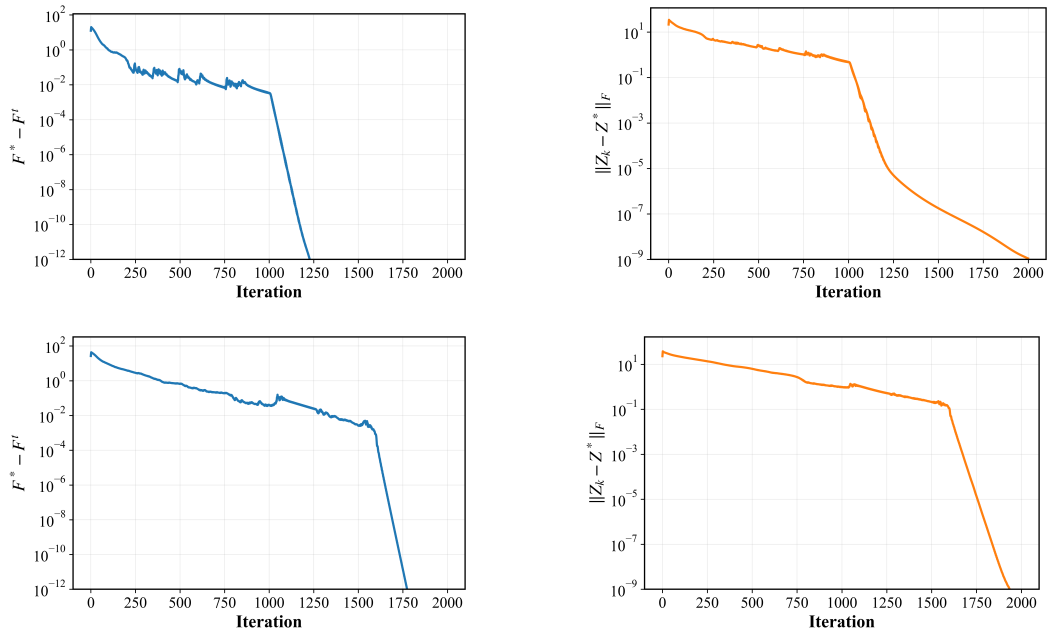


Figure 2: **Visualization of the convergence rate on MNIST and CIFAR-10 using an MLP.** We visualize the objective value gap and the iterate distance against the number of iterations. Top row: MNIST. Bottom row: CIFAR-10. Left column: objective value gap. Right column: iterate distance.

minimum learning rate of 10^{-6}. The experiments are conducted with the coding precision $\epsilon = 0.4$ and a regularization parameter $\lambda = 10^{-4}$. The initial learning rate is set to 3.5×10^{-4} for MNIST and 5×10^{-4} for CIFAR-10. All optimization runs are carried out in a deterministic full-batch setting to rigorously evaluate convergence behavior.

We plot the convergence behavior on MNIST and CIFAR-10 in Figure 2. On both datasets, the optimization exhibits a two-phase pattern: an initial exploration phase over the non-convex landscape, followed by a sharp transition into a local neighborhood of a local maximizer. After this transition, both the objective gap and the iterate distance decay, consistent with the predicted local linear convergence. The transition occurs around iteration $1{,}000$ on MNIST and around iteration $1{,}600$ on CIFAR-10, reflecting the greater complexity of CIFAR-10. Additional experiments in Appendix D further examine whether optimization progress remains aligned with representation quality under minibatch training, using linear-probe accuracy and class-mean cosine similarity.

5 Conclusions

In this work, we studied the optimization geometry of the MCR^2 problem by showing that it satisfies an error-bound property in a neighborhood of its local maximizers. This property directly implies key regularity conditions, including the PŁ inequality and quadratic growth, and allows us to establish that first-order methods converge linearly to a local maximizer. Our experimental results on synthetic and real-world datasets strongly supported our theoretical results. One future direction is to extend our analysis of Problem (3) to the constrained MCR^2 problem studied in Yu et al. [2020], which imposes a norm constraint on the features instead of using a regularizer, as well as to the sparse MCR^2 problem in Yu et al. [2024a], which considers a sparsity-inducing regularization on the features. Another promising direction is to theoretically analyze Problem (3) when the feature representations are parameterized by deep neural networks, as explored empirically in Section 4.1. Such an analysis could provide insights into the interplay between the network architecture and the optimization landscape.

Acknowledgement

R. Jiang is supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under grant JYB2025XDXM904 and the National Natural Science

Foundation of China under grants 72394360 and 72394364. P. Wang is supported in part by the National Natural Science Foundation of China under grant 12601585 and in part by the Macau Science and Technology Development Fund (FDCT) under grant 0091/2025/ITP2. H. Liu is supported by the National Natural Science Foundation of China under grants 12301403, 72192830, and 72192832.

References

J. F. Bonnans and A. Ioffe. Second-order sufficiency and quadratic growth for nonisolated minima. *Mathematics of Operations Research*, 20(4):801–817, 1995.

N. Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

K. H. R. Chan, Y. Yu, C. You, H. Qi, J. Wright, and Y. Ma. Redunet: A white-box deep network from the principle of maximizing rate reduction. *Journal of Machine Learning Research*, 23(114): 1–103, 2022.

P. Chen, R. Jiang, and P. Wang. Error bound analysis for the regularized loss of deep linear neural networks. *arXiv preprint arXiv:2502.11152*, 2025.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience (John Wiley & Sons), Hoboken, NJ, second edition, 2006.

D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, 2012.

R. Jiang and D. Li. Novel reformulations and efficient algorithms for the generalized trust region subproblem. *SIAM Journal on Optimization*, 29(2):1603–1633, 2019.

R. Jiang and X. Li. Hölderian error bounds and Kurdyka-Łojasiewicz inequality for the trust region subproblem. *Mathematics of Operations Research*, 47(4):3025–3050, 2022.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009.

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

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

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

F.-Y. Liao, L. Ding, and Y. Zheng. Error bounds, PŁ condition, and quadratic growth for weakly convex functions, and linear convergences of proximal point methods. In *Proceedings of the 6th Annual Learning for Dynamics & Control Conference*, volume 242, pages 993–1005. PMLR, 2024.

H. Liu, A. M.-C. So, and W. Wu. Quadratic optimization with orthogonality constraint: explicit Łojasiewicz exponent and linear convergence of retraction-based line-search and stochastic variance-reduced gradient methods. *Mathematical Programming*, 178(1):215–262, 2019.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Z.-Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

Y. Ma, H. Derksen, W. Hong, and J. Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9): 1546–1562, 2007.

X. Meng, Z. Tong, Z. Huang, and C.-G. Li. Temporal rate reduction clustering for human motion segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2025.

D. Pai, M. Psenka, C.-Y. Chiu, M. Wu, E. Dobriban, and Y. Ma. Pursuit of a discriminative representation for multiple subspaces via sequential games. *Journal of the Franklin Institute*, 360 (6):4135–4171, 2023.

D. Pai, S. Buchanan, Z. Wu, Y. Yu, and Y. Ma. Masked completion via structured diffusion with white-box transformers. In *International Conference on Learning Representations*, 2024.

J.-S. Pang. Error bounds in mathematical programming. *Mathematical Programming*, 79(1):299–332, 1997.

B. T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Q. Rebjock and N. Boumal. Fast convergence to non-isolated minima: four equivalent conditions for C^2 functions. *Mathematical Programming*, 213:151–199, 2025.

R. Schneider and A. Uschmajew. Convergence results for projected line-search methods on varieties of low-rank matrices via Łojasiewicz inequality. *SIAM Journal on Optimization*, 25(1):622–646, 2015.

P. Wang, H. Liu, and A. M.-C. So. Linear convergence of a proximal alternating minimization method with extrapolation for L1-norm principal component analysis. *SIAM Journal on Optimization*, 33 (2):684–712, 2023.

P. Wang, H. Liu, D. Pai, Y. Yu, Z. Zhu, Q. Qu, and Y. Ma. A global geometric analysis of maximal coding rate reduction. In *International Conference on Machine Learning*, pages 51012–51040. PMLR, 2024.

P. Wang, Y. Lu, Y. Yu, D. Pai, Q. Qu, and Y. Ma. Attention-only transformers via unrolled subspace denoising. In *International Conference on Machine Learning*, 2025.

Q. Wen, Z. Huang, and C.-G. Li. Towards interpretable and efficient attention: Compressing all by contracting a few. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Z. Wu, T. Ding, Y. Lu, D. Pai, J. Zhang, W. Wang, Y. Yu, Y. Ma, and B. D. Haeffele. Token statistics transformer: Linear-time attention via variational rate reduction. In *International Conference on Learning Representations*, 2025a.

Z. Wu, J. Zhang, D. Pai, X. Wang, C. Singh, J. Yang, J. Gao, and Y. Ma. Simplifying DINO via coding rate regularization. In *International Conference on Machine Learning*, volume 267, pages 68036–68059. PMLR, 2025b.

J. Yang, X. Li, D. Pai, Y. Zhou, Y. Ma, Y. Yu, and C. Xie. Scaling white-box transformers for vision. In *Advances in Neural Information Processing Systems*, volume 37, pages 36995–37019, 2024.

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

Y. Yu, S. Buchanan, D. Pai, T. Chu, Z. Wu, S. Tong, H. Bai, Y. Zhai, B. D. Haeffele, and Y. Ma. White-box transformers via sparse rate reduction: Compression is all there is? *Journal of Machine Learning Research*, 25(300):1–128, 2024a.

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

M.-C. Yue, Z. Zhou, and A. Man-Cho So. On the quadratic convergence of the cubic regularization method under a local error bound condition. *SIAM Journal on Optimization*, 29(1):904–932, 2019.

Z. Zhou and A. M.-C. So. A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming*, 165(2):689–728, 2017.

Appendices

In the supplementary, we provide proofs of the technical results presented in Section 2 and Section 3. Besides the notions introduced earlier, we shall use $\mathrm{BlkD}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K)$ to denote the block diagonal matrix whose diagonal blocks are $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K$.

A Technical Ingredients for the Error Bound

In this section, we present the proofs of the main theorem in Section 2. Recall that N denotes the number of training samples and N_k denotes the number of training samples in the k-th class for each $k \in [K]$. For ease of exposition, we define

$$\alpha := \frac{d}{N\epsilon^2},\ \alpha_k := \frac{d}{N_k\epsilon^2},\ \forall k \in [K]. \tag{17}$$

For each $k \in [K]$, we compute

$$\nabla_k F(\boldsymbol{Z}) := \frac{\partial F(\boldsymbol{Z})}{\partial \boldsymbol{Z}_k} = \alpha \left(\boldsymbol{I}_d + \alpha \boldsymbol{Z}\boldsymbol{Z}^T\right)^{-1} \boldsymbol{Z}_k - \alpha \left(\boldsymbol{I}_d + \alpha_k \boldsymbol{Z}_k\boldsymbol{Z}_k^T\right)^{-1} \boldsymbol{Z}_k - \lambda \boldsymbol{Z}_k. \tag{18}$$

A.1 Preliminary Results

We use the characterization of the local maximizer set $\mathcal{Z}$ stated in Fact 1(ii). For ease of exposition, given $\boldsymbol{Z}^* \in \mathcal{Z}$ with rank r_k and singular values $\sigma_1^*, \ldots, \sigma_{r_k}^*$, let

$$\sigma_{\max}^* := \max\left\{\sigma_i^* : i \in [r_k]\right\},\ \sigma_{\min}^* := \min\left\{\sigma_i^* : i \in [r_k]\right\}.$$

Now, we prove the following inequalities to bound the singular values for any $\boldsymbol{Z}$ in a neighborhood of a local maximizer $\boldsymbol{Z}^* \in \mathcal{Z}$.

Lemma 1. *Let $\boldsymbol{Z}^* \in \mathcal{Z}$ be arbitrary and $r_k = \mathrm{rank}(\boldsymbol{Z}_k^*)$ for each $k \in [K]$. For any $\boldsymbol{Z} = [\boldsymbol{Z}_1\ \cdots\ \boldsymbol{Z}_K]$ satisfying*

$$\|\boldsymbol{Z} - \boldsymbol{Z}^*\|_F \le \frac{\sigma_{\min}^*}{2}, \tag{19}$$

we have for each $k \in [K]$,

$$\frac{\sigma_{\min}^*}{2} \le \sigma_i(\boldsymbol{Z}_k) \le \frac{3\sigma_{\max}^*}{2},\ \forall i \in [r_k], \tag{20}$$

$$|\sigma_i(\boldsymbol{Z}_k)| \le \frac{\sigma_{\min}^*}{2},\ i = r_k + 1, \ldots, \min\{d, N_k\}. \tag{21}$$

Proof of Lemma 1. Using Weyl's inequality (see, e.g., [Horn and Johnson, 2012, Corollary 7.3.5]), we have for each $k \in [K]$ and $i \in \{1, \ldots, \min\{d, N_k\}\}$,

$$|\sigma_i(\boldsymbol{Z}_k) - \sigma_i(\boldsymbol{Z}_k^*)| \le \|\boldsymbol{Z}_k - \boldsymbol{Z}_k^*\| \le \|\boldsymbol{Z}_k - \boldsymbol{Z}_k^*\|_F \le \|\boldsymbol{Z} - \boldsymbol{Z}^*\|_F \le \frac{\sigma_{\min}^*}{2}.$$

This, together with $\sigma_{\min}^* \le \sigma_i(\boldsymbol{Z}_k^*) \le \sigma_{\max}^*$ for each $i \in [r_k]$ and $\sigma_i(\boldsymbol{Z}_k^*) = 0$ for all $i \in \{r_k + 1, \ldots, \min\{d, N_k\}\}$, directly implies (20) and (21). □

Based on the above result, we are now ready to show that the last $N_k - r_k$ singular values of $\boldsymbol{Z}$ can be bounded by its gradient norm.

Lemma 2. *Let $\boldsymbol{Z}^* \in \mathcal{Z}$ be arbitrary and $r_k = \mathrm{rank}(\boldsymbol{Z}_k^*)$ for each $k \in [K]$. For any $\boldsymbol{Z} = [\boldsymbol{Z}_1\ \cdots\ \boldsymbol{Z}_K]$ satisfying*

$$\|\boldsymbol{Z} - \boldsymbol{Z}^*\|_F \le \sqrt{\frac{\lambda}{2\alpha\alpha_{\max}}}, \tag{22}$$

it holds for all $k \in [K]$ and $i \in \{r_k + 1, \ldots, \min\{d, N_k\}\}$ that

$$\sigma_i(\boldsymbol{Z}_k) \le \frac{2\|\nabla F(\boldsymbol{Z})\|_F}{\lambda}. \tag{23}$$

Proof of Lemma 2. For ease of exposition, we denote $\sigma_{k,r_k+1} := \sigma_{r_k+1}(\boldsymbol{Z}_k)$ for each $k \in [K]$. Using Weyl's inequality and $\sigma_{r_k+1}(\boldsymbol{Z}_k^*) = 0$, we obtain

$$|\sigma_{k,r_k+1}| \leq \|\boldsymbol{Z}_k - \boldsymbol{Z}_k^*\| \leq \|\boldsymbol{Z} - \boldsymbol{Z}^*\|_F \overset{(22)}{\leq} \sqrt{\frac{\lambda}{2\alpha\alpha_{\max}}}. \tag{24}$$

Plugging the SVD of $\boldsymbol{Z}_k$ in (14) into $\nabla_k F(\boldsymbol{Z})$ in (18) yields for each $k \in [K]$,

$$\nabla_k F(\boldsymbol{Z}) = \alpha \left(\boldsymbol{I} + \alpha \boldsymbol{Z}\boldsymbol{Z}^T\right)^{-1} \boldsymbol{P}_k \boldsymbol{\Sigma}_k \boldsymbol{Q}_k^T - \alpha \boldsymbol{P}_k \left(\boldsymbol{I} + \alpha_k \boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_k^T\right)^{-1} \boldsymbol{\Sigma}_k \boldsymbol{Q}_k^T - \lambda \boldsymbol{P}_k \boldsymbol{\Sigma}_k \boldsymbol{Q}_k^T. \tag{25}$$

Multiplying the above equation on the left by $\boldsymbol{e}_{r_k+1}^T \boldsymbol{P}_k^T$ and on the right by $\boldsymbol{Q}_k \boldsymbol{e}_{r_k+1}$ yields

$$\begin{aligned}
&\left|\boldsymbol{e}_{r_k+1}^T \boldsymbol{P}_k^T \nabla_k F(\boldsymbol{Z}) \boldsymbol{Q}_k \boldsymbol{e}_{r_k+1}\right| \\
&= \left|\alpha \boldsymbol{e}_{r_k+1}^T \boldsymbol{P}_k^T \left(\boldsymbol{I} + \alpha \boldsymbol{Z}\boldsymbol{Z}^T\right)^{-1} \boldsymbol{P}_k \boldsymbol{e}_{r_k+1} - \alpha \left(1 + \alpha_k \sigma_{k,r_k+1}^2\right)^{-1} - \lambda\right| \sigma_{k,r_k+1} \\
&\geq \left(\lambda + \alpha(1 + \alpha_k \sigma_{k,r_k+1}^2)^{-1} - \alpha\right) \sigma_{k,r_k+1} \\
&= \left(\lambda - \frac{\alpha\alpha_k \sigma_{k,r_k+1}^2}{1 + \alpha_k \sigma_{k,r_k+1}^2}\right) \sigma_{k,r_k+1} \geq \left(\lambda - \alpha\alpha_k \sigma_{k,r_k+1}^2\right) \sigma_{k,r_k+1} \overset{(24)}{\geq} \frac{\lambda}{2} \sigma_{k,r_k+1},
\end{aligned}$$

where the first inequality uses the triangle inequality and $\|(\boldsymbol{I} + \alpha \boldsymbol{Z}\boldsymbol{Z}^T)^{-1}\| \leq 1$. Therefore, we have

$$\sigma_{k,r_k+1} \leq \frac{2}{\lambda} \left|\boldsymbol{e}_{r_k+1}^T \boldsymbol{P}_k^T \nabla_k F(\boldsymbol{Z}) \boldsymbol{Q}_k \boldsymbol{e}_{r_k+1}\right| \leq \frac{2}{\lambda} \|\nabla F(\boldsymbol{Z})\|_F.$$

This, together with $\sigma_{r_k+1}(\boldsymbol{Z}_k) \geq \cdots \geq \sigma_{\min\{d,N_k\}}(\boldsymbol{Z}_k)$, yields (23). □

A.2 Proof of the Error Bound

Before we proceed, let us introduce some further notions. Let $\boldsymbol{Z} = [\boldsymbol{Z}_1 \ \cdots \ \boldsymbol{Z}_K]$ be in a neighborhood of a $\boldsymbol{Z}^* \in \mathcal{Z}$ with $r_k = \operatorname{rank}(\boldsymbol{Z}_k^*)$ for each $k \in [K]$. According to Fact 1(ii), we have $r = \sum_{k=1}^K r_k$. For each $k \in [K]$, we consider the SVD in (14). Then, we construct some auxiliary matrices as follows:

$$\boldsymbol{Z}_k^{(1)} := \boldsymbol{P}_k^{(1)} \boldsymbol{\Sigma}_k^{(1)} {\boldsymbol{Q}_k^{(1)}}^T, \ \bar{\boldsymbol{Z}}_k := \boldsymbol{P}_k^{(1)} \boldsymbol{\Sigma}_k^{(1)}, \tag{26}$$

where $\boldsymbol{P}_k^{(1)}, \boldsymbol{\Sigma}_k^{(1)}, \boldsymbol{Q}_k^{(1)}$ are defined in (14) for each $k \in [K]$. Moreover, we define

$$\boldsymbol{Z}^{(1)} := \begin{bmatrix} \boldsymbol{Z}_1^{(1)} & \cdots & \boldsymbol{Z}_K^{(1)} \end{bmatrix} \in \mathbb{R}^{d \times N}, \tag{27}$$

$$\bar{\boldsymbol{Z}} := \begin{bmatrix} \bar{\boldsymbol{Z}}_1 & \cdots & \bar{\boldsymbol{Z}}_K \end{bmatrix} \in \mathbb{R}^{d \times r}. \tag{28}$$

Next, we show that the difference between the i-th largest r_k singular values of $\boldsymbol{Z}_k$ and σ_i^* can be bounded by the gradient norm. Moreover, the singular subspaces associated with the largest r_k singular values of different $\boldsymbol{Z}_k$ are nearly orthogonal to each other, and the Frobenius norm of their products can also be bounded by the gradient norm.

Proposition 1. *Let $\boldsymbol{Z}^* \in \mathcal{Z}$ be arbitrary and $r_k = \operatorname{rank}(\boldsymbol{Z}_k^*)$ for each $k \in [K]$. For any $\boldsymbol{Z} = [\boldsymbol{Z}_1 \ \cdots \ \boldsymbol{Z}_K]$ satisfying*

$$\|\boldsymbol{Z} - \boldsymbol{Z}^*\|_F \leq \min\left\{\sqrt{\frac{\lambda}{2\alpha\alpha_{\max}}}, \frac{\sigma_{\min}^*}{2}, \frac{\sigma_{\min}^*}{8\sigma_{\max}^*}, \delta_1\right\}, \tag{29}$$

$$\|\nabla F(\boldsymbol{Z})\|_F \leq \min\left\{\frac{\lambda\sigma_{\max}^*}{4\sqrt{dK}}, \frac{\lambda\sigma_{\min}^*}{16\sigma_{\max}^*\sqrt{dK}}\right\}, \tag{30}$$

where δ_1 is defined in (75) *of Lemma 7, the following statements hold:*
(i) It holds that

$$\sum_{k=1}^K \sum_{l=1, l\neq k}^K \left\|{\boldsymbol{P}_k^{(1)}}^T \boldsymbol{P}_l^{(1)}\right\| \leq \frac{4K(1+\kappa_1)}{\kappa_2 \alpha \sigma_{\min}^{*2}} \|\nabla F(\boldsymbol{Z})\|_F, \tag{31}$$

where

$$\kappa_1 := \frac{21\alpha^2\sigma^*_{\max}\sqrt{dK}}{2\lambda} + \frac{2\alpha\sqrt{dK}}{\lambda} + \frac{2(\lambda+\alpha)K\sqrt{d}}{\lambda} + \frac{14\left(\sigma^*_{\max}\right)^2\alpha\alpha_{\max}K\sqrt{d}}{\lambda}, \tag{32}$$

$$\kappa_2 := \frac{\sigma^*_{\min}}{2\left(1+4\alpha\sigma^{*2}_{\max}\right)}\left(\frac{\alpha}{1+4\alpha_{\max}\sigma^{*2}_{\max}}+\lambda\right). \tag{33}$$

(ii) It holds for all $k\in[K]$ and $i\in[r_k]$ that

$$|\sigma_i(\boldsymbol{Z}_k)-\sigma^*_i| \le \frac{(1+\kappa_1)(\alpha+\lambda)}{\kappa_2\kappa_3}\|\nabla F(\boldsymbol{Z})\|_F, \tag{34}$$

where κ_3 is defined in (75) *of Lemma 7.*

Proof of Proposition 1. Let $r := \mathrm{rank}(\boldsymbol{Z}^*)$, which satisfies $r=\sum_{k=1}^K r_k$ according to Fact 1(ii). Recall that $\boldsymbol{Z}_k$ admits the SVD introduced in (14) and $\boldsymbol{Z}_k^{(1)}$, $\bar{\boldsymbol{Z}}_k^{(1)}$ are defined in (26) for each $k\in[K]$. For notational convenience, we denote $\bar{\boldsymbol{Z}}^{(1)}$ by $\bar{\boldsymbol{Z}}$ and $\bar{\boldsymbol{Z}}_k^{(1)}$ simply by $\bar{\boldsymbol{Z}}_k$ for each $k\in[K]$. It follows from (29) and Lemma 2 that (23) holds. Using (14) and (26), we have

$$\left\|\boldsymbol{Z}_k-\boldsymbol{Z}_k^{(1)}\right\|_F = \left\|\boldsymbol{P}_k^{(2)}\boldsymbol{\Sigma}_k^{(2)}\boldsymbol{Q}_k^{(2)^T}\right\|_F = \left\|\boldsymbol{\Sigma}_k^{(2)}\right\|_F \overset{(23)}{\le} \frac{2\sqrt{d}}{\lambda}\|\nabla F(\boldsymbol{Z})\|_F, \tag{35}$$

$$\left\|\boldsymbol{Z}-\boldsymbol{Z}^{(1)}\right\|_F = \sqrt{\sum_{k=1}^K\left\|\boldsymbol{Z}_k-\boldsymbol{Z}_k^{(1)}\right\|_F^2} \overset{(35)}{\le} \frac{2\sqrt{dK}}{\lambda}\|\nabla F(\boldsymbol{Z})\|_F. \tag{36}$$

Next, we bound

$$\|\boldsymbol{Z}\| \le \|\boldsymbol{Z}-\boldsymbol{Z}^*\|+\|\boldsymbol{Z}^*\| \le \|\boldsymbol{Z}-\boldsymbol{Z}^*\|_F+\|\boldsymbol{Z}^*\| \overset{(5,29)}{\le} \frac{3}{2}\sigma^*_{\max}, \tag{37}$$

$$\|\boldsymbol{Z}^{(1)}\| \le \|\boldsymbol{Z}^{(1)}-\boldsymbol{Z}\|+\|\boldsymbol{Z}\| \overset{(36,37)}{\le} \frac{2\sqrt{dK}}{\lambda}\|\nabla F(\boldsymbol{Z})\|_F+\frac{3}{2}\sigma^*_{\max} \le 2\sigma^*_{\max}, \tag{38}$$

where the last inequality follows from $\|\nabla F(\boldsymbol{Z})\|_F \le \lambda\sigma^*_{\max}/(4\sqrt{dK})$ by (30). For each $k\in[K]$, we bound

$$\begin{aligned}
&\left\|\left(\boldsymbol{I}+\alpha_k\boldsymbol{Z}_k^T\boldsymbol{Z}_k\right)^{-1}-\left(\boldsymbol{I}+\alpha_k\boldsymbol{Z}_k^{(1)^T}\boldsymbol{Z}_k^{(1)}\right)^{-1}\right\|_F \\
&\le \alpha_k\left\|\left(\boldsymbol{I}+\alpha_k\boldsymbol{Z}_k^T\boldsymbol{Z}_k\right)^{-1}\right\|\left\|\left(\boldsymbol{I}+\alpha_k\boldsymbol{Z}_k^{(1)^T}\boldsymbol{Z}_k^{(1)}\right)^{-1}\right\|\left\|\boldsymbol{Z}_k^T\boldsymbol{Z}_k-\boldsymbol{Z}_k^{(1)^T}\boldsymbol{Z}_k^{(1)}\right\|_F \\
&\le \alpha_k\left\|\boldsymbol{Z}_k^T\boldsymbol{Z}_k-\boldsymbol{Z}_k^{(1)^T}\boldsymbol{Z}_k^{(1)}\right\|_F \le \alpha_k\left(\left\|\boldsymbol{Z}_k^T\left(\boldsymbol{Z}_k-\boldsymbol{Z}_k^{(1)}\right)\right\|_F+\left\|\left(\boldsymbol{Z}_k-\boldsymbol{Z}_k^{(1)}\right)^T\boldsymbol{Z}_k^{(1)}\right\|_F\right) \\
&\overset{(37,38)}{\le} \frac{7}{2}\sigma^*_{\max}\alpha_k\left\|\boldsymbol{Z}_k-\boldsymbol{Z}_k^{(1)}\right\|_F \overset{(35)}{\le} \frac{7\sqrt{d}\sigma^*_{\max}\alpha_k}{\lambda}\|\nabla F(\boldsymbol{Z})\|_F,
\end{aligned} \tag{39}$$

where the first inequality follows from $\boldsymbol{A}^{-1}-\boldsymbol{B}^{-1}=\boldsymbol{A}^{-1}(\boldsymbol{B}-\boldsymbol{A})\boldsymbol{B}^{-1}$ and $\|\boldsymbol{AB}\|_F\le\|\boldsymbol{A}\|\|\boldsymbol{B}\|_F$. Using a similar argument, we have

$$\begin{aligned}
&\left\|\left(\boldsymbol{I}+\alpha\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}-\left(\boldsymbol{I}+\alpha\boldsymbol{Z}^{(1)^T}\boldsymbol{Z}^{(1)}\right)^{-1}\right\|_F \\
&\le \alpha\left\|\boldsymbol{Z}^T\boldsymbol{Z}-\boldsymbol{Z}^{(1)^T}\boldsymbol{Z}^{(1)}\right\|_F \le \alpha\left(\left\|\boldsymbol{Z}^T\left(\boldsymbol{Z}-\boldsymbol{Z}^{(1)}\right)\right\|_F+\left\|\left(\boldsymbol{Z}-\boldsymbol{Z}^{(1)}\right)^T\boldsymbol{Z}^{(1)}\right\|_F\right) \\
&\overset{(37,38)}{\le} \frac{7\alpha\sigma^*_{\max}}{2}\left\|\boldsymbol{Z}-\boldsymbol{Z}^{(1)}\right\|_F \overset{(36)}{\le} \frac{7\sqrt{dK}\sigma^*_{\max}\alpha}{\lambda}\|\nabla F(\boldsymbol{Z})\|_F.
\end{aligned} \tag{40}$$

Recall that $\nabla F(\boldsymbol{Z})=[\nabla_1 F(\boldsymbol{Z})\ \cdots\ \nabla_K F(\boldsymbol{Z})]$. To keep the display readable, define

$$\mathcal{C}(\boldsymbol{Z}) := \left[\left(\boldsymbol{I}+\alpha_1\boldsymbol{Z}_1\boldsymbol{Z}_1^T\right)^{-1}\boldsymbol{Z}_1\ \cdots\ \left(\boldsymbol{I}+\alpha_K\boldsymbol{Z}_K\boldsymbol{Z}_K^T\right)^{-1}\boldsymbol{Z}_K\right],$$

$$\mathcal{D}(\boldsymbol{Z}) := \left[\boldsymbol{Z}_1\left(\boldsymbol{I}+\alpha_1\boldsymbol{Z}_1^T\boldsymbol{Z}_1\right)^{-1}\ \cdots\ \boldsymbol{Z}_K\left(\boldsymbol{I}+\alpha_K\boldsymbol{Z}_K^T\boldsymbol{Z}_K\right)^{-1}\right].$$

Then (18) yields

$$\begin{aligned}\nabla F(\boldsymbol{Z}) &= \alpha\left(\boldsymbol{I}+\alpha\boldsymbol{Z}\boldsymbol{Z}^T\right)^{-1}\boldsymbol{Z}-\alpha\mathcal{C}(\boldsymbol{Z})-\lambda\boldsymbol{Z}\\ &= \alpha\boldsymbol{Z}\left(\boldsymbol{I}+\alpha\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}-\alpha\mathcal{D}(\boldsymbol{Z})-\lambda\boldsymbol{Z},\end{aligned}\tag{41}$$

where the second equality uses Lemma 6. Using the triangle inequality, we obtain

$$\begin{aligned}&\left\|\nabla F(\boldsymbol{Z})-\nabla F(\boldsymbol{Z}^{(1)})\right\|_F\\ &\le \alpha\left\|\boldsymbol{Z}\left((\boldsymbol{I}+\alpha\boldsymbol{Z}^T\boldsymbol{Z})^{-1}-(\boldsymbol{I}+\alpha\boldsymbol{Z}^{(1)^T}\boldsymbol{Z}^{(1)})^{-1}\right)\right\|_F+\alpha\left\|\left(\boldsymbol{Z}-\boldsymbol{Z}^{(1)}\right)\left(\boldsymbol{I}+\alpha\boldsymbol{Z}^{(1)^T}\boldsymbol{Z}^{(1)}\right)^{-1}\right\|_F\\ &\quad+\lambda\sum_{k=1}^K\left\|\boldsymbol{Z}_k-\boldsymbol{Z}_k^{(1)}\right\|_F+\alpha\sum_{k=1}^K\left\|\left(\boldsymbol{Z}_k-\boldsymbol{Z}_k^{(1)}\right)\left(\boldsymbol{I}+\alpha_k\boldsymbol{Z}_k^T\boldsymbol{Z}_k\right)^{-1}\right\|_F\\ &\quad+\alpha\sum_{k=1}^K\left\|\boldsymbol{Z}_k^{(1)}\left(\left(\boldsymbol{I}+\alpha_k\boldsymbol{Z}_k^T\boldsymbol{Z}_k\right)^{-1}-\left(\boldsymbol{I}+\alpha_k\boldsymbol{Z}_k^{(1)^T}\boldsymbol{Z}_k^{(1)}\right)^{-1}\right)\right\|_F\\ &\le \alpha\|\boldsymbol{Z}\|\left\|(\boldsymbol{I}+\alpha\boldsymbol{Z}^T\boldsymbol{Z})^{-1}-(\boldsymbol{I}+\alpha\boldsymbol{Z}^{(1)^T}\boldsymbol{Z}^{(1)})^{-1}\right\|_F+\alpha\left\|\boldsymbol{Z}-\boldsymbol{Z}^{(1)}\right\|_F+(\lambda+\alpha)\sum_{k=1}^K\left\|\boldsymbol{Z}_k-\boldsymbol{Z}_k^{(1)}\right\|_F\\ &\quad+\alpha\sum_{k=1}^K\left\|\boldsymbol{Z}_k^{(1)}\right\|\left\|\left(\boldsymbol{I}+\alpha_k\boldsymbol{Z}_k^T\boldsymbol{Z}_k\right)^{-1}-\left(\boldsymbol{I}+\alpha_k\boldsymbol{Z}_k^{(1)^T}\boldsymbol{Z}_k^{(1)}\right)^{-1}\right\|_F\\ &\le \frac{3}{2}\alpha\sigma_{\max}^*\left\|(\boldsymbol{I}+\alpha\boldsymbol{Z}^T\boldsymbol{Z})^{-1}-(\boldsymbol{I}+\alpha\boldsymbol{Z}^{(1)^T}\boldsymbol{Z}^{(1)})^{-1}\right\|_F+\frac{2\alpha\sqrt{dK}}{\lambda}\|\nabla F(\boldsymbol{Z})\|_F+\frac{2(\lambda+\alpha)K\sqrt{d}}{\lambda}\|\nabla F(\boldsymbol{Z})\|_F\\ &\quad+2\alpha\sigma_{\max}^*\sum_{k=1}^K\left\|\left(\boldsymbol{I}+\alpha_k\boldsymbol{Z}_k^T\boldsymbol{Z}_k\right)^{-1}-\left(\boldsymbol{I}+\alpha_k\boldsymbol{Z}_k^{(1)^T}\boldsymbol{Z}_k^{(1)}\right)^{-1}\right\|_F\le\kappa_1\|\nabla F(\boldsymbol{Z})\|_F,\end{aligned}$$

where the second inequality uses $\|\boldsymbol{A}\boldsymbol{B}\|_F\le\|\boldsymbol{A}\|\|\boldsymbol{B}\|_F$ for any $\boldsymbol{A},\boldsymbol{B}$, the third inequality follows from (35), (36), (37), and (38), and the last inequality is due to (39), (40), and the definition of the aggregate efficiency κ_1 in (32). This, together with the triangle inequality, implies

$$\|\nabla F(\boldsymbol{Z})\|_F\ge\frac{1}{1+\kappa_1}\left\|\nabla F(\boldsymbol{Z}^{(1)})\right\|_F.\tag{42}$$

(i) According to (18), we have

$$\begin{aligned}\left\|\nabla_k F(\boldsymbol{Z}^{(1)})\right\|_F &= \left\|\alpha\left(\boldsymbol{I}+\alpha\boldsymbol{Z}^{(1)}\boldsymbol{Z}^{(1)^T}\right)^{-1}\boldsymbol{Z}_k^{(1)}-\alpha\left(\boldsymbol{I}+\alpha_k\boldsymbol{Z}_k^{(1)}\boldsymbol{Z}_k^{(1)^T}\right)^{-1}\boldsymbol{Z}_k^{(1)}-\lambda\boldsymbol{Z}_k^{(1)}\right\|_F\\ &\overset{(26)}{=}\left\|\alpha\left(\boldsymbol{I}+\alpha\bar{\boldsymbol{Z}}\bar{\boldsymbol{Z}}^T\right)^{-1}\bar{\boldsymbol{Z}}_k-\alpha\left(\boldsymbol{I}+\alpha_k\bar{\boldsymbol{Z}}_k\bar{\boldsymbol{Z}}_k^T\right)^{-1}\bar{\boldsymbol{Z}}_k-\lambda\bar{\boldsymbol{Z}}_k\right\|_F.\end{aligned}\tag{43}$$

Using (26) and (27), we have $\bar{\boldsymbol{Z}}\bar{\boldsymbol{Z}}^T=\boldsymbol{Z}^{(1)}\boldsymbol{Z}^{(1)^T}$ and thus

$$\begin{aligned}\sigma_{\min}^2\left(\bar{\boldsymbol{Z}}\right)=\sigma_r\left(\boldsymbol{Z}^{(1)}\boldsymbol{Z}^{(1)^T}\right)&\ge\sigma_r\left(\boldsymbol{Z}^*\boldsymbol{Z}^{*^T}\right)-\left\|\boldsymbol{Z}^{(1)}\boldsymbol{Z}^{(1)^T}-\boldsymbol{Z}^*\boldsymbol{Z}^{*^T}\right\|_F\\ &\ge\sigma_{\min}^{*^2}-\|\boldsymbol{Z}^{(1)}\|\|\boldsymbol{Z}^{(1)}-\boldsymbol{Z}^*\|_F-\|\boldsymbol{Z}^*\|\|\boldsymbol{Z}^{(1)}-\boldsymbol{Z}^*\|_F\\ &\overset{(5,38)}{\ge}\sigma_{\min}^{*^2}-3\sigma_{\max}^*\left(\|\boldsymbol{Z}^{(1)}-\boldsymbol{Z}\|_F+\|\boldsymbol{Z}-\boldsymbol{Z}^*\|_F\right)\\ &\overset{(36)}{\ge}\sigma_{\min}^{*^2}-3\sigma_{\max}^*\left(\frac{2\sqrt{dK}}{\lambda}\|\nabla F(\boldsymbol{Z})\|_F+\|\boldsymbol{Z}-\boldsymbol{Z}^*\|_F\right)\overset{(29,30)}{\ge}\frac{\sigma_{\min}^{*^2}}{4}.\end{aligned}\tag{44}$$

where the first inequality uses Weyl's inequality and the second inequality follows from Fact 1(ii) and the triangle inequality. Using (18), (26), (41), and $\bar{\boldsymbol{Z}}\bar{\boldsymbol{Z}}^T=\boldsymbol{Z}^{(1)}\boldsymbol{Z}^{(1)^T}$, define

$$\Phi:=\alpha\,\mathrm{BlkD}\Big(\left(\boldsymbol{I}+\alpha_1\bar{\boldsymbol{Z}}_1^T\bar{\boldsymbol{Z}}_1\right)^{-1},\ldots,\left(\boldsymbol{I}+\alpha_K\bar{\boldsymbol{Z}}_K^T\bar{\boldsymbol{Z}}_K\right)^{-1}\Big)+\lambda\boldsymbol{I}.$$

Then we obtain the compact form

$$\begin{aligned}\left\|\nabla F(\boldsymbol{Z}^{(1)})\right\|_F^2 &= \left\|\bar{\boldsymbol{Z}}\left(\alpha\left(\boldsymbol{I}+\alpha\bar{\boldsymbol{Z}}^T\bar{\boldsymbol{Z}}\right)^{-1}-\Phi\right)\right\|_F^2 \geq \sigma_{\min}^2\left(\bar{\boldsymbol{Z}}\right)\left\|\alpha\left(\boldsymbol{I}+\alpha\bar{\boldsymbol{Z}}^T\bar{\boldsymbol{Z}}\right)^{-1}-\Phi\right\|_F^2 \\ &\geq \sigma_{\min}^2\left(\bar{\boldsymbol{Z}}\right)\sigma_{\min}^2\left(\left(\boldsymbol{I}+\alpha\bar{\boldsymbol{Z}}^T\bar{\boldsymbol{Z}}\right)^{-1}\right)\sigma_{\min}^2\left(\Phi\right)\left\|\boldsymbol{I}+\alpha\bar{\boldsymbol{Z}}^T\bar{\boldsymbol{Z}}-\Phi^{-1}\right\|_F^2 \\ &\geq \kappa_2^2\left\|\boldsymbol{I}+\alpha\bar{\boldsymbol{Z}}^T\bar{\boldsymbol{Z}}-\Phi^{-1}\right\|_F^2,\end{aligned}$$

where the equality follows from Lemma 6 and the definition of Φ. The first and second inequalities apply the lower bound $\|\boldsymbol{AB}\|_F \geq \sigma_{\min}(\boldsymbol{A})\|\boldsymbol{B}\|_F$. Furthermore, the final inequality is a consequence of (44), the definition of κ_2 in (33), and the following auxiliary results

$$\sigma_{\min}\left(\left(\boldsymbol{I}+\alpha\bar{\boldsymbol{Z}}^T\bar{\boldsymbol{Z}}\right)^{-1}\right) = \left(1+\alpha\sigma_{\max}(\bar{\boldsymbol{Z}}^T\bar{\boldsymbol{Z}})\right)^{-1} = (1+\alpha\|\bar{\boldsymbol{Z}}\|^2)^{-1} \overset{(38)}{\geq} \frac{1}{1+4\alpha\sigma_{\max}^2},$$

and

$$\begin{aligned}\sigma_{\min}\left(\Phi\right) &= \min_{k\in[K]}\left\{\sigma_{\min}\left(\alpha\left(\boldsymbol{I}+\alpha_k\bar{\boldsymbol{Z}}_k^T\bar{\boldsymbol{Z}}_k\right)^{-1}+\lambda\boldsymbol{I}\right)\right\} \\ &\geq \alpha\min_{k\in[K]}\left\{\frac{1}{1+\alpha_k\|\bar{\boldsymbol{Z}}_k\|^2}\right\}+\lambda \overset{(38)}{\geq} \frac{\alpha}{1+4\alpha_{\max}\sigma_{\max}^2}+\lambda.\end{aligned}$$

Consequently, we have

$$\begin{aligned}\left\|\nabla F(\boldsymbol{Z}^{(1)})\right\|_F \geq \kappa_2\left\|\boldsymbol{I}+\alpha\bar{\boldsymbol{Z}}^T\bar{\boldsymbol{Z}}-\Phi^{-1}\right\|_F &\geq \kappa_2\alpha\sqrt{\sum_{k=1}^K\sum_{l\neq k}^K\left\|\bar{\boldsymbol{Z}}_k^T\bar{\boldsymbol{Z}}_l\right\|_F^2} \\ &\geq \frac{\kappa_2\alpha}{K}\sum_{k=1}^K\sum_{l\neq k}^K\left\|\bar{\boldsymbol{Z}}_k^T\bar{\boldsymbol{Z}}_l\right\|_F, \qquad (45)\end{aligned}$$

where the second inequality follows by discarding the diagonal blocks in the Frobenius norms, i.e., by retaining only the blocks of $\bar{\boldsymbol{Z}}_k^T\bar{\boldsymbol{Z}}_l$ in $\bar{\boldsymbol{Z}}^T\bar{\boldsymbol{Z}}$. This, together with (42), yields

$$\begin{aligned}\frac{K(1+\kappa_1)}{\kappa_2\alpha}\left\|\nabla F(\boldsymbol{Z})\right\|_F &\geq \sum_{k=1}^K\sum_{l\neq k}^K\left\|\bar{\boldsymbol{Z}}_k^T\bar{\boldsymbol{Z}}_l\right\|_F \overset{(26)}{=} \sum_{k=1}^K\sum_{l\neq k}^K\left\|\boldsymbol{\Sigma}_k^{(1)^T}\boldsymbol{P}_k^{(1)^T}\boldsymbol{P}_l^{(1)}\boldsymbol{\Sigma}_l^{(1)}\right\|_F \\ &\geq \sum_{k=1}^K\sum_{l\neq k}^K\sigma_{\min}\left(\boldsymbol{\Sigma}_k^{(1)}\right)\sigma_{\min}\left(\boldsymbol{\Sigma}_l^{(1)}\right)\left\|\boldsymbol{P}_k^{(1)^T}\boldsymbol{P}_l^{(1)}\right\|_F \\ &\geq \frac{\sigma_{\min}^{*^2}}{4}\sum_{k=1}^K\sum_{l\neq k}^K\left\|\boldsymbol{P}_k^{(1)^T}\boldsymbol{P}_l^{(1)}\right\|_F\end{aligned}$$

where the last inequality uses (20) by Lemma 1. This directly implies (31).

(ii) Using the first inequality in (45) and considering only the diagonal blocks, we obtain for each $k \in [K]$ and $i \in [r_K]$,

$$\begin{aligned}\left\|\nabla F(\boldsymbol{Z}^{(1)})\right\|_F &\geq \kappa_2\left\|\boldsymbol{I}+\alpha\bar{\boldsymbol{Z}}_k^T\bar{\boldsymbol{Z}}_k-\alpha\left(\alpha\left(\boldsymbol{I}+\alpha_k\bar{\boldsymbol{Z}}_k^T\bar{\boldsymbol{Z}}_k\right)^{-1}+\lambda\boldsymbol{I}\right)^{-1}\right\|_F \\ &\geq \kappa_2\left|1+\alpha\sigma_{k,i}^2-\frac{\alpha}{\alpha(1+\alpha_k\sigma_{k,i}^2)^{-1}+\lambda}\right| \\ &= \frac{\kappa_2\left(1+\alpha\sigma_{k,i}^2\right)}{\alpha(1+\alpha_k\sigma_{k,i}^2)^{-1}+\lambda}\left|\frac{\alpha}{1+\alpha_k\sigma_{k,i}^2}+\lambda-\frac{\alpha}{1+\alpha\sigma_{k,i}^2}\right| \\ &\geq \frac{\kappa_2}{\alpha+\lambda}\left|\frac{\alpha}{1+\alpha_k\sigma_{k,i}^2}+\lambda-\frac{\alpha}{1+\alpha\sigma_{k,i}^2}\right|, \qquad (46)\end{aligned}$$

where the second inequality follows from (26) and $\mathbf{\Sigma}_k^{(1)} = \mathrm{diag}(\sigma_{k,1}, \ldots, \sigma_{k,r_k})$. To simplify our development, we define

$$\varphi_k(x) := \frac{\alpha}{1+\alpha x^2} - \frac{\alpha}{1+\alpha_k x^2} - \lambda. \tag{47}$$

Applying Lemma 7 to (47) yields

$$|\varphi_k(x)| \geq \kappa_3 \left|x - \sigma_k^*\right|, \ \forall \left|x - \sigma_k^*\right| \leq \delta_1. \tag{48}$$

where κ_3 and δ_1 are defined in (75) of Lemma 7. Recall the definitions in (14), (26), and (27). For all $i \in [r_k]$, we have

$$\left|\sigma_i(\bar{\boldsymbol{Z}}_k) - \sigma_k^*\right| = \left|\sigma_{k,i} - \sigma_k^*\right| \leq \|\boldsymbol{Z}_k - \boldsymbol{Z}_k^*\|_F \leq \|\boldsymbol{Z} - \boldsymbol{Z}^*\|_F \leq \delta_1,$$

where the first inequality uses Weyl's inequality and last inequality follows from (29). This, together with (47) and (48), yields

$$\left|\sigma_i(\bar{\boldsymbol{Z}}_k) - \sigma_k^*\right| \leq \frac{1}{\kappa_3}\left|\varphi_k(\sigma_i(\bar{\boldsymbol{Z}}_k))\right| \leq \frac{\alpha+\lambda}{\kappa_2\kappa_3}\left\|\nabla F(\boldsymbol{Z}^{(1)})\right\|_F.$$

where the second inequality follows from (46). Using this, (42), and $\sigma_i(\bar{\boldsymbol{Z}}_k) = \sigma_i(\boldsymbol{Z}_k)$, yields (34). □

Based on Proposition 1, we further show that there exists an orthonormal matrix such that its distance to $[\boldsymbol{P}_1^{(1)} \ \cdots \ \boldsymbol{P}_1^{(K)}]$ in (14) can be bounded by the gradient norm at $\boldsymbol{Z}$.

Lemma 3. *Let $\boldsymbol{Z}^* \in \mathcal{Z}$ be arbitrary and $r_k = \mathrm{rank}(\boldsymbol{Z}_k^*)$ for each $k \in [K]$. For any $\boldsymbol{Z} = [\boldsymbol{Z}_1 \ \cdots \ \boldsymbol{Z}_K]$ satisfying* (29) *and*

$$\|\nabla F(\boldsymbol{Z})\|_F \leq \min\left\{\frac{\lambda\sigma_{\min}^*}{4\sqrt{dK}}, \frac{\lambda\sigma_{\min}^*}{16\sigma_{\max}^*\sqrt{dK}}, \frac{\kappa_2\alpha\sigma_{\min}^*}{3^{K-2}\cdot 8K(1+\kappa_1)}\right\}, \tag{49}$$

where κ_1 and κ_2 are defined in Proposition 1. Then, there exist matrices $\hat{\boldsymbol{U}}_k \in \mathcal{O}^{d\times r_k}$, $k \in [K]$, with pairwise orthogonal column spaces (i.e., $\hat{\boldsymbol{U}}_k^T\hat{\boldsymbol{U}}_l = \boldsymbol{0}$ for all $k \neq l$), such that for all $k \in [K]$

$$\left\|\boldsymbol{P}_k^{(1)} - \hat{\boldsymbol{U}}_k\right\|_F \leq \frac{3^{K-2}\cdot 8K(1+\kappa_1)}{\kappa_2\alpha\sigma_{\min}^*}\|\nabla F(\boldsymbol{Z})\|_F. \tag{50}$$

Proof of Lemma 3. Since (29) and (49) hold, we obtain that Proposition 1 holds. We define

$$\boldsymbol{H}_1 := \boldsymbol{P}_1^{(1)}, \ \hat{\boldsymbol{U}}_1 := \boldsymbol{P}_1^{(1)}, \ \boldsymbol{H}_k := \boldsymbol{P}_k^{(1)} - \sum_{l=1}^{k-1}\hat{\boldsymbol{U}}_l\hat{\boldsymbol{U}}_l^T\boldsymbol{P}_k^{(1)}, \ \forall k \geq 2. \tag{51}$$

Then, let $\boldsymbol{H}_k = \tilde{\boldsymbol{U}}_k\tilde{\boldsymbol{\Sigma}}_k\tilde{\boldsymbol{V}}_k^T$ be a thin SVD of $\boldsymbol{H}_k \in \mathbb{R}^{d\times r_k}$, where $\tilde{\boldsymbol{U}}_k \in \mathcal{O}^{d\times r_k}$, $\tilde{\boldsymbol{\Sigma}}_k \in \mathbb{R}^{r_k\times r_k}$, and $\tilde{\boldsymbol{V}}_k \in \mathcal{O}^{r_k}$ for all $k \geq 2$. We further define $\hat{\boldsymbol{U}}_k := \tilde{\boldsymbol{U}}_k\tilde{\boldsymbol{V}}_k^T$ for all $k \geq 2$.

Now, we use an inductive argument to show that the following properties hold for any $k \geq 1$:

(i) $\hat{\boldsymbol{U}}_l^T\boldsymbol{H}_k = \boldsymbol{0}, \ \forall l \in [k-1]$,

(ii) $\|\boldsymbol{P}_k^{(1)} - \boldsymbol{H}_k\|_F \leq \dfrac{3^{k-2}\cdot 4K(1+\kappa_1)}{\kappa_2\alpha\sigma_{\min}^*}\|\nabla F(\boldsymbol{Z})\|_F,$

$$\|\hat{\boldsymbol{U}}_k - \boldsymbol{P}_k^{(1)}\|_F \leq \frac{3^{k-2}\cdot 8K(1+\kappa_1)}{\kappa_2\alpha\sigma_{\min}^*}\|\nabla F(\boldsymbol{Z})\|_F,$$

(iii) $\mathrm{rank}(\boldsymbol{H}_k) = r_k$,

(iv) $\hat{\boldsymbol{U}}_k^T\hat{\boldsymbol{U}}_k = \boldsymbol{I}, \ \hat{\boldsymbol{U}}_l^T\hat{\boldsymbol{U}}_k = \boldsymbol{0}, \ \forall l \in [k-1]$.

First, consider the case where $k = 1$. Then, it is easy to verify that (i)-(iv) hold. Next, suppose that (i)-(iv) hold for all $k = 1, \ldots, s-1$. Then we aim to show (i)-(iv) hold for $k = s$. Note that $\boldsymbol{H}_s = \boldsymbol{P}_s^{(1)} - \sum_{i=1}^{s-1}\hat{\boldsymbol{U}}_i\hat{\boldsymbol{U}}_i^T\boldsymbol{P}_s^{(1)}$. For each $l \in [s-1]$, we compute

$$\hat{\boldsymbol{U}}_l^T\boldsymbol{H}_s = \hat{\boldsymbol{U}}_l^T\left(\boldsymbol{P}_s^{(1)} - \sum_{i=1}^{s-1}\hat{\boldsymbol{U}}_i\hat{\boldsymbol{U}}_i^T\boldsymbol{P}_s^{(1)}\right) = \hat{\boldsymbol{U}}_l^T\boldsymbol{P}_s^{(1)} - \hat{\boldsymbol{U}}_l^T\boldsymbol{P}_s^{(1)} = \boldsymbol{0},$$

where the second equality uses $\hat{\boldsymbol{U}}_i^T \hat{\boldsymbol{U}}_k = \boldsymbol{0}$ for all $i \in [k-1]$ and $k \in [s-1]$. This implies that (i) holds for $k = s$. Then, we compute

$$\begin{aligned}
\|\boldsymbol{H}_s - \boldsymbol{P}_s^{(1)}\|_F &= \left\| \sum_{i=1}^{s-1} \hat{\boldsymbol{U}}_i \hat{\boldsymbol{U}}_i^T \boldsymbol{P}_s^{(1)} \right\|_F \leq \sum_{i=1}^{s-1} \left\| \hat{\boldsymbol{U}}_i^T \boldsymbol{P}_s^{(1)} \right\|_F \\
&\leq \sum_{i=1}^{s-1} \left\| \hat{\boldsymbol{U}}_i - \boldsymbol{P}_i^{(1)} \right\|_F + \sum_{i=1}^{s-1} \left\| \boldsymbol{P}_i^{(1)^T} \boldsymbol{P}_s^{(1)} \right\|_F \\
&\leq \left(2 \sum_{i=2}^{s-1} 3^{i-2} + 1 \right) \frac{4K(1+\kappa_1)}{\kappa_2 \alpha \sigma_{\min}^*} \|\nabla F(\boldsymbol{Z})\|_F = \frac{3^{s-2} \cdot 4K(1+\kappa_1)}{\kappa_2 \alpha \sigma_{\min}^*} \|\nabla F(\boldsymbol{Z})\|_F,
\end{aligned} \tag{52}$$

where the second inequality follows from the triangle inequality and $\|\boldsymbol{AB}\|_F \leq \|\boldsymbol{A}\| \|\boldsymbol{B}\|_F$ and the last inequality uses (31) and the fact that (ii) holds for all $i \in [s-1]$. Moreover, we have

$$\begin{aligned}
\|\hat{\boldsymbol{U}}_s - \boldsymbol{P}_s^{(1)}\|_F &= \|\tilde{\boldsymbol{U}}_s \tilde{\boldsymbol{V}}_s^T - \boldsymbol{P}_s^{(1)}\|_F \leq \|\tilde{\boldsymbol{U}}_s \tilde{\boldsymbol{V}}_s^T - \boldsymbol{H}_s\|_F + \|\boldsymbol{H}_s - \boldsymbol{P}_s^{(1)}\|_F \\
&\leq \|\boldsymbol{I} - \tilde{\boldsymbol{\Sigma}}_s\|_F + \|\boldsymbol{H}_s - \boldsymbol{P}_s^{(1)}\|_F = \|\boldsymbol{\sigma}(\boldsymbol{P}_s^{(1)}) - \boldsymbol{\sigma}(\boldsymbol{H}_s)\|_F + \|\boldsymbol{H}_s - \boldsymbol{P}_s^{(1)}\|_F \\
&\leq 2\|\boldsymbol{H}_s - \boldsymbol{P}_s^{(1)}\|_F \leq \frac{3^{s-2} \cdot 8K(1+\kappa_1)}{\kappa_2 \alpha \sigma_{\min}^*} \|\nabla F(\boldsymbol{Z})\|_F,
\end{aligned} \tag{53}$$

where the second equality follows because $\boldsymbol{\sigma}(\boldsymbol{A})$ denotes all positive singular values of $\boldsymbol{A}$ and $\boldsymbol{\sigma}(\boldsymbol{P}_s^{(1)}) = \boldsymbol{1}$, the third inequality uses Mirsky's inequality, and the last inequality is due to (52). Therefore, (ii) holds for $k = s$. Using Mirsky's inequality again yields

$$\begin{aligned}
\left| \sigma_{r_s}(\boldsymbol{H}_s) - \sigma_{r_s}(\boldsymbol{P}_s^{(1)}) \right| &\leq \left\| \boldsymbol{H}_s - \boldsymbol{P}_s^{(1)} \right\| \leq \left\| \boldsymbol{H}_s - \boldsymbol{P}_s^{(1)} \right\|_F \\
&\leq \frac{3^{s-2} \cdot 4K(1+\kappa_1)}{\kappa_2 \alpha \sigma_{\min}^*} \|\nabla F(\boldsymbol{Z})\|_F \overset{(49)}{\leq} \frac{1}{2}.
\end{aligned}$$

Then, we have $\sigma_{r_s}(\boldsymbol{H}_s) \geq \sigma_{r_s}(\boldsymbol{P}_s^{(1)}) - 1/2 = 1/2$. This, together with $\boldsymbol{H}_s \in \mathbb{R}^{d \times r_s}$, yields $\operatorname{rank}(\boldsymbol{H}_s) = r_s$. Therefore, (iii) holds for $k = s$. Finally, we compute $\hat{\boldsymbol{U}}_s^T \hat{\boldsymbol{U}}_s = (\tilde{\boldsymbol{U}}_s \tilde{\boldsymbol{V}}_s^T)^T \tilde{\boldsymbol{U}}_s \tilde{\boldsymbol{V}}_s^T = \boldsymbol{I}$ and $\hat{\boldsymbol{U}}_l^T \hat{\boldsymbol{U}}_s = \hat{\boldsymbol{U}}_l^T \tilde{\boldsymbol{U}}_s \tilde{\boldsymbol{V}}_s^T = \boldsymbol{0}$ due to $\hat{\boldsymbol{U}}_l^T \boldsymbol{H}_s = \boldsymbol{0}$ and the fact that $\boldsymbol{H}_s$ has full rank. Then, (iv) holds for $k = s$. Consequently, we prove that (i)-(iv) hold for $k = s$. □

To maintain consistency, we convert the gradient-norm condition (49) into a bound on the distance to a local maximizer as in (29). To this end, we show that the gradient is Lipschitz continuous in a neighborhood of local maximizers. According to (18), we have

$$\nabla_k F(\boldsymbol{Z}) = \alpha(\boldsymbol{I}_d + \alpha \boldsymbol{Z}\boldsymbol{Z}^T)^{-1} \boldsymbol{Z}_k - \alpha(\boldsymbol{I}_d + \alpha_k \boldsymbol{Z}_k \boldsymbol{Z}_k^T)^{-1} \boldsymbol{Z}_k - \lambda \boldsymbol{Z}_k, \quad \forall k \in [K]. \tag{54}$$

By concatenating the blocks for all $k \in [K]$, the full gradient $\nabla F(\boldsymbol{Z}) \in \mathbb{R}^{d \times n}$ takes the form of

$$\nabla F(\boldsymbol{Z}) = T_1(\boldsymbol{Z}) - T_2(\boldsymbol{Z}) - T_3(\boldsymbol{Z}),$$

where

$$\begin{aligned}
&T_1(\boldsymbol{Z}) = \alpha(\boldsymbol{I}_d + \alpha \boldsymbol{Z}\boldsymbol{Z}^T)^{-1} \boldsymbol{Z}, \\
&T_2(\boldsymbol{Z}) = \left[\alpha(\boldsymbol{I}_d + \alpha_1 \boldsymbol{Z}_1 \boldsymbol{Z}_1^T)^{-1} \boldsymbol{Z}_1, \cdots, \alpha(\boldsymbol{I}_d + \alpha_K \boldsymbol{Z}_K \boldsymbol{Z}_K^T)^{-1} \boldsymbol{Z}_K \right], \quad T_3(\boldsymbol{Z}) = \lambda \boldsymbol{Z}.
\end{aligned}$$

Lemma 4. *For any matrix $\boldsymbol{M}$ and scalar $\gamma > 0$, let $\boldsymbol{A} := (\boldsymbol{I}_d + \gamma \boldsymbol{M}\boldsymbol{M}^T)^{-1}$. It holds that*

$$\|\boldsymbol{A}\| \leq 1, \quad \|\boldsymbol{AM}\| \leq \frac{1}{2\sqrt{\gamma}}, \quad \|\boldsymbol{M}^T \boldsymbol{AM}\| \leq \frac{1}{\gamma}. \tag{55}$$

Proof. Let $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be the singular value decomposition with singular values $\sigma_i \geq 0$. Note that $\boldsymbol{A}$ has eigenvalues $(1 + \gamma\sigma_i^2)^{-1}$. Thus, we immediately have $\|\boldsymbol{A}\| = \max_i (1 + \gamma\sigma_i^2)^{-1} \leq 1$. By the AM-GM inequality, $1 + \gamma\sigma_i^2 \geq 2\sqrt{\gamma}\sigma_i$, implying $\|\boldsymbol{AM}\| = \max_i \sigma_i (1 + \gamma\sigma_i^2)^{-1} \leq (2\sqrt{\gamma})^{-1}$. Finally, $\|\boldsymbol{M}^T \boldsymbol{AM}\| = \max_i \sigma_i^2 (1 + \gamma\sigma_i^2)^{-1} \leq \gamma^{-1}$ follows directly from the fact that $\gamma\sigma_i^2 < 1 + \gamma\sigma_i^2$. □

Lemma 5. *It holds for any $\boldsymbol{Z}, \boldsymbol{Z}' \in \mathbb{R}^{d\times N}$ that*

$$\|\nabla F(\boldsymbol{Z}) - \nabla F(\boldsymbol{Z}')\|_F \le L_F \|\boldsymbol{Z} - \boldsymbol{Z}'\|_F, \tag{56}$$

where

$$L_F := \lambda + \frac{9\alpha}{2}. \tag{57}$$

Using the above lemma, (49) holds provided that

$$\|\boldsymbol{Z} - \boldsymbol{Z}^*\|_F \le \delta_2, \tag{58}$$

where

$$\delta_2 := \frac{1}{L_F} \min\left\{ \frac{\lambda\sigma^*_{\min}}{4\sqrt{dK}}, \frac{\lambda\sigma^*_{\min}}{16\sigma^*_{\max}\sqrt{dK}}, \frac{\kappa_2\alpha\sigma^*_{\min}}{3^{K-2}8K(1+\kappa_1)} \right\}. \tag{59}$$

Proof of Lemma 5. To compute the Lipschitz constant of $\nabla F(\boldsymbol{Z})$, it suffices to separately compute the Lipschitz constant of each term T_1, T_2, T_3 by the triangle inequality. Clearly, $T_3(\boldsymbol{Z}) = \lambda\boldsymbol{Z}$ is Lipschitz continuous with constant $L_3 = \lambda$.

Let $P(\boldsymbol{Z}) := (\boldsymbol{I}_d + \alpha\boldsymbol{Z}\boldsymbol{Z}^T)^{-1}$. We compute

$$\begin{aligned} T_1(\boldsymbol{Z}) - T_1(\boldsymbol{Z}') &= \alpha\left(P(\boldsymbol{Z})\boldsymbol{Z} - P(\boldsymbol{Z}')\boldsymbol{Z}'\right) \\ &= \alpha \underbrace{P(\boldsymbol{Z})(\boldsymbol{Z} - \boldsymbol{Z}')}_{\boldsymbol{E}_1} + \alpha \underbrace{(P(\boldsymbol{Z}) - P(\boldsymbol{Z}'))\boldsymbol{Z}'}_{\boldsymbol{E}_2}. \end{aligned} \tag{60}$$

Note that $\|\boldsymbol{E}_1\|_F \le \|\boldsymbol{Z} - \boldsymbol{Z}'\|_F$ due to $\|P(\boldsymbol{Z})\| \le 1$. To bound $\boldsymbol{E}_2$, first note the identity that $P(\boldsymbol{Z}) - P(\boldsymbol{Z}') = \alpha P(\boldsymbol{Z})\left(\boldsymbol{Z}'\boldsymbol{Z}'^T - \boldsymbol{Z}\boldsymbol{Z}^T\right)P(\boldsymbol{Z}')$. This, together with $\boldsymbol{Z}'\boldsymbol{Z}'^T - \boldsymbol{Z}\boldsymbol{Z}^T = -\left(\boldsymbol{Z}(\boldsymbol{Z} - \boldsymbol{Z}')^T + (\boldsymbol{Z} - \boldsymbol{Z}')\boldsymbol{Z}'^T\right)$, yields

$$\boldsymbol{E}_2 = -\alpha P(\boldsymbol{Z})\left(\boldsymbol{Z}(\boldsymbol{Z} - \boldsymbol{Z}')^T + (\boldsymbol{Z} - \boldsymbol{Z}')\boldsymbol{Z}'^T\right)P(\boldsymbol{Z}')\boldsymbol{Z}'.$$

Now, we bound $\|\boldsymbol{E}_2\|_F$ by grouping terms:

$$\begin{aligned} \|\boldsymbol{E}_2\|_F &\le \alpha\left\|P(\boldsymbol{Z})\boldsymbol{Z}(\boldsymbol{Z} - \boldsymbol{Z}')^T P(\boldsymbol{Z}')\boldsymbol{Z}'\right\|_F + \alpha\left\|P(\boldsymbol{Z})(\boldsymbol{Z} - \boldsymbol{Z}')\boldsymbol{Z}'^T P(\boldsymbol{Z}')\boldsymbol{Z}'\right\|_F \\ &\le \alpha\|P(\boldsymbol{Z})\boldsymbol{Z}\|\|\boldsymbol{Z} - \boldsymbol{Z}'\|_F\|P(\boldsymbol{Z}')\boldsymbol{Z}'\| + \alpha\|P(\boldsymbol{Z})\|\|\boldsymbol{Z} - \boldsymbol{Z}'\|_F\|\boldsymbol{Z}'^T P(\boldsymbol{Z}')\boldsymbol{Z}'\| \\ &\le \frac{1}{4}\|\boldsymbol{Z} - \boldsymbol{Z}'\|_F + \|\boldsymbol{Z} - \boldsymbol{Z}'\|_F = \frac{5}{4}\|\boldsymbol{Z} - \boldsymbol{Z}'\|_F, \end{aligned} \tag{61}$$

where the last inequality follows from Lemma 4. This, together with (60) and $\|\boldsymbol{E}_1\|_F \le \|\boldsymbol{Z} - \boldsymbol{Z}'\|_F$, yields that the Lipschitz constant of $T_1(\boldsymbol{Z})$ is $L_1 = 9\alpha/4$.

Let $f_k(\boldsymbol{Z}_k) := \alpha(\boldsymbol{I}_d + \alpha_k\boldsymbol{Z}_k\boldsymbol{Z}_k^T)^{-1}\boldsymbol{Z}_k$ for each $k \in [K]$. The structure of f_k is identical to T_1 but defined on the block $\boldsymbol{Z}_k$ with scalar α_k. Following the same derivation as above, each f_k is Lipschitz continuous with constant $L_{2,k} = 9\alpha/4$. Since $T_2(\boldsymbol{Z})$ is a block-diagonal operator, we have

$$\|T_2(\boldsymbol{Z}) - T_2(\boldsymbol{Z}')\|_F^2 = \sum_{k=1}^K \|f_k(\boldsymbol{Z}_k) - f_k(\boldsymbol{Z}'_k)\|_F^2 \le \sum_{k=1}^K L_{2,k}^2 \|\boldsymbol{Z}_k - \boldsymbol{Z}'_k\|_F^2 \le \frac{81\alpha^2}{16}\|\boldsymbol{Z} - \boldsymbol{Z}'\|_F^2.$$

Therefore, we have $L_2 = 9\alpha/4$.

Combining the above computations together yields that the Lipschitz constant of $\nabla F(\boldsymbol{Z})$ is

$$L_F = L_1 + L_2 + L_3 = \lambda + \frac{9\alpha}{2}.$$

□

Using (59), Lemma 2, Proposition 1, and Lemma 3, we are ready to establish the error bound as follows.

Proposition 2. *Let $\boldsymbol{Z}^* \in \mathcal{Z}$ be arbitrary and $r_k = \operatorname{rank}(\boldsymbol{Z}_k^*)$ for each $k \in [K]$. For any $\boldsymbol{Z} = [\boldsymbol{Z}_1 \ \cdots \ \boldsymbol{Z}_K]$ satisfying*

$$\|\boldsymbol{Z} - \boldsymbol{Z}^*\|_F \leq \delta_{\boldsymbol{Z}^*}, \tag{62}$$

we have

$$\operatorname{dist}(\boldsymbol{Z}, \mathcal{Z}) \leq \kappa_{\boldsymbol{Z}^*} \|\nabla F(\boldsymbol{Z})\|_F, \tag{63}$$

where

$$\delta_{\boldsymbol{Z}^*} := \min\left\{\sqrt{\frac{\lambda}{2\alpha\alpha_{\max}}}, \frac{\sigma_{\min}^*}{2}, \frac{\sigma_{\min}^*}{8\sigma_{\max}^*}, \delta_1, \delta_2\right\}, \tag{64}$$

$$\kappa_{\boldsymbol{Z}^*} := \frac{(1+\kappa_1)(\alpha+\lambda)}{\kappa_2\kappa_3}\sum_{k=1}^{K}\sqrt{r_k} + 3^{K-2}\cdot\frac{8K^2(1+\kappa_1)\sigma_{\max}^*}{\kappa_2\alpha\sigma_{\min}^*} + \frac{2\sqrt{d}K}{\lambda}, \tag{65}$$

κ_1, κ_2 are respectively defined in (32) *and* (33) *of Proposition 1, and κ_3 is defined in* (75) *of Lemma 7.*

Proof of Proposition 2. Substituting $\boldsymbol{Z}' = \boldsymbol{Z}^*$ into (56) and using (58) and (59) yield (49). Since (29) and (49) hold, we obtain that Lemma 2, Proposition 1, and Lemma 3 hold. Recall the SVD of $\boldsymbol{Z}_k$ in (14) and the notions in (26). Moreover, let $\hat{\boldsymbol{U}}_k$ be as defined in the proof of Lemma 3 for each $k \in [K]$. Now, we construct

$$\hat{\boldsymbol{Z}} = \begin{bmatrix}\hat{\boldsymbol{Z}}_1 & \dots & \hat{\boldsymbol{Z}}_K\end{bmatrix}, \text{ where } \hat{\boldsymbol{Z}}_k = \sigma_k^*\hat{\boldsymbol{U}}_k\boldsymbol{Q}_k^{(1)^T},$$

where $\boldsymbol{Q}_k^{(1)}$ is defined in (26). By Lemma 3 and Fact 1(ii), $\hat{\boldsymbol{Z}} \in \mathcal{Z}$. Moreover, for all $k \neq l \in [K]$,

$$\hat{\boldsymbol{Z}}_k^T\hat{\boldsymbol{Z}}_l = \sigma_k^*\sigma_l^*\boldsymbol{Q}_k^{(1)^T}\hat{\boldsymbol{U}}_k^T\hat{\boldsymbol{U}}_l\boldsymbol{Q}_l^{(1)} = \boldsymbol{0},$$

where the last equality is due to Lemma 3. For each $k \in [K]$, we have

$$\begin{aligned}
\left\|\boldsymbol{Z}_k - \hat{\boldsymbol{Z}}_k\right\|_F &\overset{(14)}{=} \left\|\boldsymbol{P}_k^{(1)}\boldsymbol{\Sigma}_k^{(1)}\boldsymbol{Q}_k^{(1)^T} + \boldsymbol{P}_k^{(2)}\boldsymbol{\Sigma}_k^{(2)}\boldsymbol{Q}_k^{(2)^T} - \sigma_k^*\hat{\boldsymbol{U}}_k\boldsymbol{Q}_k^{(1)^T}\right\|_F \\
&\leq \left\|\boldsymbol{P}_k^{(1)}\boldsymbol{\Sigma}_k^{(1)}\boldsymbol{Q}_k^{(1)^T} - \sigma_k^*\hat{\boldsymbol{U}}_k\boldsymbol{Q}_k^{(1)^T}\right\|_F + \left\|\boldsymbol{P}_k^{(2)}\boldsymbol{\Sigma}_k^{(2)}\boldsymbol{Q}_k^{(2)^T}\right\|_F \\
&= \left\|\boldsymbol{P}_k^{(1)}\boldsymbol{\Sigma}_k^{(1)} - \sigma_k^*\hat{\boldsymbol{U}}_k\right\|_F + \left\|\boldsymbol{\Sigma}_k^{(2)}\right\|_F \\
&\leq \left\|\boldsymbol{P}_k^{(1)}\left(\boldsymbol{\Sigma}_k^{(1)} - \sigma_k^*\boldsymbol{I}\right)\right\|_F + \sigma_k^*\left\|\boldsymbol{P}_k^{(1)} - \hat{\boldsymbol{U}}_k\right\|_F + \left\|\boldsymbol{\Sigma}_k^{(2)}\right\|_F \\
&\leq \left(\frac{(1+\kappa_1)(\alpha+\lambda)}{\kappa_2\kappa_3}\sqrt{r_k} + \frac{3^{k-2}\cdot 8K(1+\kappa_1)\sigma_k^*}{\kappa_2\alpha\sigma_{\min}^*} + \frac{2\sqrt{d}}{\lambda}\right)\|\nabla F(\boldsymbol{Z})\|_F,
\end{aligned}$$

where the last inequality uses (23) in Lemma 2, (34) in Proposition 1, and (50) in Lemma 3. Therefore, we have

$$\begin{aligned}
\operatorname{dist}(\boldsymbol{Z}, \mathcal{Z}) &\leq \left\|\boldsymbol{Z} - \hat{\boldsymbol{Z}}\right\|_F = \sqrt{\sum_{k=1}^{K}\|\boldsymbol{Z}_k - \hat{\boldsymbol{Z}}_k\|_F^2} \\
&\leq \left(\frac{(1+\kappa_1)(\alpha+\lambda)}{\kappa_2\kappa_3}\sum_{k=1}^{K}\sqrt{r_k} + \frac{3^{K-2}\cdot 8K^2(1+\kappa_1)\sigma_{\max}^*}{\kappa_2\alpha\sigma_{\min}^*} + \frac{2\sqrt{d}K}{\lambda}\right)\|\nabla F(\boldsymbol{Z})\|_F.
\end{aligned}$$

□

B Proofs in Section 2

B.1 Proof of Theorem 1

Armed with the above setup, we now proceed to prove Theorem 1 by applying the above argument to all points in the neighborhood $\mathcal{Z}$.

Proof of Theorem 1. According to (7), it is obvious that $\mathcal{Z}$ is a compact set. Let $\boldsymbol{Z}^* \in \mathcal{Z}$ be arbitrary. Using Proposition 2, there exists a $\delta_{\boldsymbol{Z}^*} > 0$, defined in (64), such that for all $\|\boldsymbol{Z} - \boldsymbol{Z}^*\|_F \le \delta_{\boldsymbol{Z}^*}$, it holds that

$$\operatorname{dist}(\boldsymbol{Z}, \mathcal{Z}) \le \kappa_{\boldsymbol{Z}^*} \|\nabla F(\boldsymbol{Z})\|_F,$$

where $\kappa_{\boldsymbol{Z}^*}$ is defined as in (65). Then, it always holds that

$$\bigcup_{\boldsymbol{Z}^* \in \mathcal{Z}} \{\boldsymbol{Z} : \|\boldsymbol{Z} - \boldsymbol{Z}^*\|_F \le \delta_{\boldsymbol{Z}^*}\} \supseteq \mathcal{Z}.$$

As $\mathcal{Z}$ is compact, using the Heine–Borel theorem, we can select finite elements $\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(n)} \in \mathcal{Z}$ and the associated neighborhoods $\mathcal{U}_1, \ldots, \mathcal{U}_n$ such that $\bigcup_{i\in[n]} \mathcal{U}_i \supseteq \mathcal{Z}$. For each $i \in [n]$, it holds that

$$\operatorname{dist}(\boldsymbol{Z}, \mathcal{Z}) \le \kappa_{\boldsymbol{Z}^{(i)}} \|\nabla F(\boldsymbol{Z})\|_F, \ \forall \boldsymbol{Z} \in \mathcal{U}_i.$$

Therefore, we have

$$\operatorname{dist}(\boldsymbol{Z}, \mathcal{Z}) \le \kappa \|\nabla F(\boldsymbol{Z})\|_F, \ \forall \boldsymbol{Z} \in \mathcal{U},$$

where $\kappa := \max_{i\in[n]} \kappa_{\boldsymbol{Z}^{(i)}}$ and $\mathcal{U} := \bigcup_{i\in[n]} \mathcal{U}_i$. □

Remark 2. *From Theorem 1, we observe that if* $\operatorname{dist}(\boldsymbol{Z}, \mathcal{Z}) \le \delta$ *and* $\nabla F(\boldsymbol{Z}) = \boldsymbol{0}$*, then* $\boldsymbol{Z} \in \mathcal{Z}$*. This, together with* (8)*, implies that the set of local maximizers* $\mathcal{Z}$ *is well separated from strict saddle points, even though no explicit characterization of the latter is available.*

B.2 Proof of Theorem 2

Proof of Theorem 2. First, we claim that Lyapunov stability holds, i.e., there exists a neighborhood $\mathcal{U}$ such that if $\boldsymbol{Z}^{t_0} \in \mathcal{U}$, then $\boldsymbol{Z}^t \in \mathcal{U}$ for all $t \ge t_0$. We prove this claim by adapting [Boumal, 2023, Proposition 3.3] to the specific structure of our problem. By Fact 1(ii), $\mathcal{Z}$ has finitely many connected components, determined by the choice of $\boldsymbol{r}$. This, together with Remark 2, implies that any connected component (say, $\mathfrak{X}$) is isolated from other critical points; that is, there exists a neighborhood $\mathcal{H}$ of $\mathfrak{X}$ such that $\boldsymbol{Z} \in \mathcal{H}$ and $\nabla F(\boldsymbol{Z}) = \boldsymbol{0}$ imply $\boldsymbol{Z} \in \mathfrak{X}$ and $F(\cdot)$ is constant on $\mathfrak{X}$. We use $F(\boldsymbol{Z}^*)$ to denote the function value in this set for arbitrary $\boldsymbol{Z}^* \in \mathfrak{X}$. Since $\mathcal{Z}$ has only finitely many connected components, we may, without loss of generality, restrict our attention to the component $\mathfrak{X}$. Using Theorem 1 and [Boumal, 2023, Proposition 2.4], there exists a neighborhood of $\mathfrak{X}$ such that

$$F(\boldsymbol{Z}^*) - F(\boldsymbol{Z}) \le \kappa \|\nabla F(\boldsymbol{Z})\|_F^2. \tag{66}$$

Without loss of generality, we assume that this neighborhood is one on which the error bound inequality holds (see Theorem 1), since we can always take the intersection of the two neighborhoods. Because $\mathfrak{X}$ consists of local maximizers and is compact, there exists a compact neighborhood $\mathcal{V}$ of $\mathfrak{X}$ such that $F(\boldsymbol{Z}) < F(\boldsymbol{Z}^*)$ for each $\boldsymbol{Z} \in \mathcal{V} \setminus \mathfrak{X}$. By property (i) of the generated sequence, we conclude $\lim_{t\to\infty} \|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_F = 0$. Thus, there exists an open neighborhood $\mathcal{W}$ of $\mathfrak{X}$ such that $\mathcal{W} \subseteq \mathcal{V} \cap \mathcal{B}(\mathfrak{X}, \delta)$, where δ is as defined in Theorem 1, and there exists t_0 such that $\boldsymbol{Z}^t \in \mathcal{W}$ implies $\boldsymbol{Z}^{t+1} \in \mathcal{V}$ for sufficiently large $t \ge t_0$. Let $\alpha := \max_{\boldsymbol{Z}\in\mathcal{V}\setminus\mathcal{W}} F(\boldsymbol{Z})$, which is well defined because $\mathcal{V} \setminus \mathcal{W}$ is compact and F is continuous. Define $\mathcal{U} := \{\boldsymbol{Z} \in \mathcal{W} : F(\boldsymbol{Z}) > \alpha\}$. By construction, we have $\mathcal{U} \subseteq \mathcal{W} \subseteq \mathcal{B}(\mathfrak{X}, \delta)$. Note that $\boldsymbol{Z}^{t_0} \in \mathcal{U}$. Then we have $\boldsymbol{Z}^{t_0+1} \in \mathcal{V}$ by the above implication, and property (i) gives $F(\boldsymbol{Z}^{t_0+1}) \ge F(\boldsymbol{Z}^{t_0}) > \alpha$. This implies $\boldsymbol{Z}^{t_0+1} \in \mathcal{U}$. By induction, it follows that $\boldsymbol{Z}^t \in \mathcal{U}$ for all $t \ge t_0$.

According to Theorem 1 and $\boldsymbol{Z}^{t_0} \in \mathcal{U}$, we have for all $t \ge t_0$ that

$$\boldsymbol{Z}^t \in \mathcal{U} \quad \text{and} \quad \operatorname{dist}(\boldsymbol{Z}^t, \mathfrak{X}) \le \kappa \|\nabla F(\boldsymbol{Z}^t)\|_F. \tag{67}$$

Thus, we have

$$\begin{aligned} F(\boldsymbol{Z}^*) - F(\boldsymbol{Z}^{t+1}) &\le \kappa \|\nabla F(\boldsymbol{Z}^{t+1})\|_F^2 \le \kappa L_F^2 \operatorname{dist}^2(\boldsymbol{Z}^{t+1}, \mathfrak{X}) \\ &\le 2\kappa L_F^2 \left(\operatorname{dist}^2(\boldsymbol{Z}^t, \mathfrak{X}) + \|\boldsymbol{Z}^t - \boldsymbol{Z}^{t+1}\|_F^2\right) \\ &\le 2\kappa L_F^2 (\kappa^2 \mu_2^2 + 1) \|\boldsymbol{Z}^t - \boldsymbol{Z}^{t+1}\|_F^2, \end{aligned}$$

where the first inequality follows from (66), the second from Lemma 5, and the last from (67). Using (i), we have

$$\frac{F(\boldsymbol{Z}^*) - F(\boldsymbol{Z}^t)}{F(\boldsymbol{Z}^*) - F(\boldsymbol{Z}^{t+1})} \ge 1 + \frac{\mu_1 \|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_F^2}{F(\boldsymbol{Z}^*) - F(\boldsymbol{Z}^{t+1})}.$$

Substituting this into the inequality above, we have

$$\frac{F(\boldsymbol{Z}^*) - F(\boldsymbol{Z}^t)}{F(\boldsymbol{Z}^*) - F(\boldsymbol{Z}^{t+1})} \geq w^{-2}, \quad \text{where } w := \left(1 + \frac{\mu_1}{2\kappa L_F^2(\kappa^2\mu_2^2 + 1)}\right)^{-1/2}. \tag{68}$$

Using (i) again, we have

$$\mu_1\|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_F^2 \leq F(\boldsymbol{Z}^{t+1}) - F(\boldsymbol{Z}^t) \leq F(\boldsymbol{Z}^*) - F(\boldsymbol{Z}^t) \leq w^{2(t-t_0)}\left(F(\boldsymbol{Z}^*) - F(\boldsymbol{Z}^{t_0})\right).$$

Thus, we obtain that

$$\|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_F \leq \left(\frac{F(\boldsymbol{Z}^*) - F(\boldsymbol{Z}^{t_0})}{\mu_1}\right)^{1/2} w^{t-t_0}.$$

For each $m \geq t$, we have the following bound:

$$\|\boldsymbol{Z}^t - \boldsymbol{Z}^m\|_F \leq \left(\frac{F(\boldsymbol{Z}^*) - F(\boldsymbol{Z}^t)}{\mu_1}\right)^{1/2} \sum_{i=t}^{m-1} w^{i-t_0}.$$

This shows that the sequence $\{\boldsymbol{Z}^t\}$ is a Cauchy sequence. Thus, we conclude that $\{\boldsymbol{Z}^t\}$ converges to a local maximizer $\bar{\boldsymbol{Z}} \in \mathcal{Z}$. Note that $\bar{\boldsymbol{Z}}$ may differ from $\boldsymbol{Z}^*$, although $F(\bar{\boldsymbol{Z}}) = F(\boldsymbol{Z}^*)$. Finally, letting $m \to \infty$, we obtain that

$$\|\boldsymbol{Z}^t - \bar{\boldsymbol{Z}}\|_F \leq \left(\frac{F(\boldsymbol{Z}^*) - F(\boldsymbol{Z}^t)}{\mu_1}\right)^{1/2} \frac{w^{t-t_0}}{1-w}.$$

This, together with (68), implies the desired result. Since the finitely many connected components of $\mathcal{Z}$ are mutually separated, the conclusion extends from a neighborhood of a single component $\mathfrak{X}$ to a neighborhood of the entire set $\mathcal{Z}$. With a slight abuse of notation, we continue to denote this Lyapunov-stable neighborhood by $\mathcal{U}$. □

B.3 Proof of Corollary 1

Proof of Corollary 1. Combined with the fact that gradient ascent with random initialization almost surely escapes strict saddles Lee et al. [2019] and the loss landscape of Problem (3) (see Fact 1), the iterates almost always converge to a local maximizer. Moreover, since $\|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_F = \eta\|\nabla F(\boldsymbol{Z}^t)\|_F$ and the update norm vanishes asymptotically along the convergent trajectory, it follows from the argument in [Boumal, 2023, Proposition 3.3] that once $\boldsymbol{Z}^{t_0} \in \mathcal{U}$, which is defined in the proof of Theorem 2, we have $\boldsymbol{Z}^t \in \mathcal{U}$ for all $t \geq t_0$.

By Lemma 5,

$$F(\boldsymbol{Z}^{t+1}) - F(\boldsymbol{Z}^t) \geq \langle \nabla F(\boldsymbol{Z}^t), \boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\rangle - \frac{L_F}{2}\|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_F^2 = \left(\frac{1}{\eta} - \frac{L_F}{2}\right)\|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_F^2.$$

Assuming $\eta \leq 1/L_F$, this implies

$$F(\boldsymbol{Z}^{t+1}) - F(\boldsymbol{Z}^t) \geq \frac{1}{2\eta}\|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_F^2 \geq \frac{L_F}{2}\|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_F^2.$$

Combining this with Theorem 2 yields the desired result with

$$w = \left(1 + \frac{1}{4\kappa L_F(\kappa^2/\eta^2 + 1)}\right)^{-1/2}.$$

Here we use the explicit choices $\mu_1 = L_F/2$ and $\mu_2 = 1/\eta$ from the gradient-ascent update. □

Remark 3. *Theorem 2 gives a local rate conditional on sufficiently late entry into its neighborhood. For gradient ascent under Corollary 1, almost-sure convergence implies eventual entry; no uniform bound on the entry time or global linear rate is asserted. The local constants are characterized in Proposition 2, specifically* (64)–(65)*. To clarify why the trajectory stays nearby, let $\mathfrak{X}$ be one connected component and F^* its objective value. Sufficient ascent and boundedness of F above imply $\Delta_t := \boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t \to \mathbf{0}$. As in the proof, choose a compact neighborhood $\mathcal{V}$ within the error-bound region with $F < F^*$ on $\mathcal{V} \setminus \mathfrak{X}$, and an open neighborhood $\mathcal{W}$ of $\mathfrak{X}$ with $\overline{\mathcal{W}} \subset \mathrm{int}(\mathcal{V})$. Set*

$$\rho := \mathrm{dist}(\overline{\mathcal{W}}, \mathcal{V}^c) > 0, \qquad a := \max_{\mathcal{V}\setminus\mathcal{W}} F < F^*, \qquad \mathcal{U} := \{\boldsymbol{Z} \in \mathcal{W} : F(\boldsymbol{Z}) > a\}.$$

For sufficiently large t, $\|\Delta_t\|_F < \rho$. If $\boldsymbol{Z}^t \in \mathcal{U}$, the next iterate lies in $\mathcal{V}$ and sufficient ascent gives $F(\boldsymbol{Z}^{t+1}) > a$, excluding $\mathcal{V} \setminus \mathcal{W}$. Induction therefore keeps every subsequent iterate in $\mathcal{U}$. Entry into $\mathcal{U}$ remains a separate assumption for the general algorithmic class in Theorem 2.

C Auxiliary Results

Lemma 6. *For a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, it holds that*

$$(\boldsymbol{I}_m + \boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{I}_n + \boldsymbol{A}^T\boldsymbol{A})^{-1}. \tag{69}$$

Proof. Let $r = \operatorname{rank}(\boldsymbol{A})$ and $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be an SVD of $\boldsymbol{A}$, where $\boldsymbol{U} \in \mathcal{O}^m$, $\boldsymbol{V} \in \mathcal{O}^n$, and $\boldsymbol{\Sigma} = \operatorname{BlkD}(\tilde{\boldsymbol{\Sigma}}, \boldsymbol{0}) \in \mathbb{R}^{m\times n}$ with $\tilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{r\times r}$ being a diagonal matrix. Then, we compute

$$\left(\boldsymbol{I}_m + \boldsymbol{A}\boldsymbol{A}^T\right)^{-1}\boldsymbol{A} = \boldsymbol{U}\left(\boldsymbol{I}_m + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\right)^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T, \tag{70}$$

$$\boldsymbol{A}(\boldsymbol{I}_n + \boldsymbol{A}^T\boldsymbol{A})^{-1} = \boldsymbol{U}\boldsymbol{\Sigma}\left(\boldsymbol{I}_n + \boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{V}^T. \tag{71}$$

Note that

$$\left(\boldsymbol{I}_m + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\right)^{-1}\boldsymbol{\Sigma} = \begin{bmatrix}(\boldsymbol{I}_r + \tilde{\boldsymbol{\Sigma}}^2)^{-1}\tilde{\boldsymbol{\Sigma}} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}, \quad \boldsymbol{\Sigma}\left(\boldsymbol{I}_n + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\right)^{-1} = \begin{bmatrix}\tilde{\boldsymbol{\Sigma}}(\boldsymbol{I}_r + \tilde{\boldsymbol{\Sigma}}^2)^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}.$$

This, together with $(\boldsymbol{I}_r + \tilde{\boldsymbol{\Sigma}}^2)^{-1}\tilde{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{\Sigma}}(\boldsymbol{I}_r + \tilde{\boldsymbol{\Sigma}}^2)^{-1}$, (70), and (71), yields (69). □

Lemma 7. *Given the coding precision $\epsilon > 0$, suppose that λ satisfies*

$$\lambda \in \left(0, \frac{d(\sqrt{N/N_{\max}} - 1)}{N(\sqrt{N/N_{\max}} + 1)\epsilon^2}\right]. \tag{72}$$

For each $k \in [K]$, define

$$\varphi_k(x) := \frac{\alpha}{1 + \alpha x^2} - \frac{\alpha}{1 + \alpha_k x^2} - \lambda,$$

where α and α_k are defined in (17)*. Then, the following statements hold:*
(i) For each $k \in [K]$, the equation $\varphi_k(x) = 0$ has two distinct positive roots

$$\sigma_k^* = \left(\frac{\eta_k + \sqrt{\eta_k^2 - 4\lambda^2 N/N_k}}{2\lambda\alpha_k}\right)^{1/2}, \ \sigma_k^\dagger = \left(\frac{\eta_k - \sqrt{\eta_k^2 - 4\lambda^2 N/N_k}}{2\lambda\alpha_k}\right)^{1/2} \tag{73}$$

where η_k is defined in (6) *for each $k \in [K]$.*

(ii) For all x satisfying $|x - \sigma_k^*| \le \delta_1$, it holds that

$$|\varphi_k(x)| \ge \kappa_3 |x - \sigma_k^*|, \tag{74}$$

where

$$\delta_1 := \frac{1}{2}\min\left\{\sigma_k^* - \sigma_k^\dagger : k \in [K]\right\}, \ \kappa_3 := \frac{2\alpha^2\alpha_k(\sigma_k^* - \delta_1)(\alpha_k - \alpha)\sigma_k^*\sigma_k^\dagger\delta_1^2}{\left(1 + \alpha(\sigma_k^* + \delta_1)^2\right)^2\left(1 + \alpha_k(\sigma_k^* + \delta_1)^2\right)^2}. \tag{75}$$

Proof. (i) Computing the roots of $\varphi_k(x) = 0$ is equivalent to solving

$$\lambda\alpha_k x^4 - \eta_k x^2 + \frac{\lambda}{\alpha} = 0,$$

where η_k is defined in (6). Using (72), one can verify that for each $k \in [K]$,

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

This, together with (17), yields that the above quadratic equation has positive roots as follows:

$$x^2 = \frac{\eta_k \pm \sqrt{\eta_k^2 - 4\lambda^2 N/N_k}}{2\lambda\alpha_k},$$

which directly implies (73). According to Vieta's formula, we have

$$\left(\sigma_k^*\sigma_k^\dagger\right)^2 = \frac{1}{\alpha\alpha_k}. \tag{76}$$

(ii) Let x be arbitrary such that $|x - \sigma_k^*| \le \delta_1$. This, together with $\delta_1 = \frac{1}{2}\min\left\{\sigma_k^* - \sigma_k^\dagger : k \in [K]\right\}$, implies

$$\frac{\sigma_k^* + \sigma_k^\dagger}{2} \le \sigma_k^* - \delta_1 \le x \le \sigma_k^* + \delta_1. \tag{77}$$

We compute

$$\begin{aligned}
|\varphi_k'(x)| &= 2\alpha \left| \frac{\alpha x}{(1+\alpha x^2)^2} - \frac{\alpha_k x}{(1+\alpha_k x^2)^2} \right| \\
&= \frac{2\alpha x}{(1+\alpha x^2)^2 (1+\alpha_k x^2)^2} \left| (\alpha - \alpha_k)(1 - \alpha\alpha_k x^4) \right| \\
&= \frac{2\alpha^2 \alpha_k x}{(1+\alpha x^2)^2 (1+\alpha_k x^2)^2} \left| (\alpha - \alpha_k) \left(\left(\sigma_k^* \sigma_k^\dagger\right)^2 - x^4 \right) \right| \\
&\ge \frac{2\alpha^2 \alpha_k (\sigma_k^* - \delta_1)(\alpha_k - \alpha)\sigma_k^* \sigma_k^\dagger}{(1+\alpha(\sigma_k^* + \delta_1)^2)^2 (1+\alpha_k(\sigma_k^* + \delta_1)^2)^2} \left(x^2 - \sigma_k^* \sigma_k^\dagger \right) \\
&\ge \frac{2\alpha^2 \alpha_k (\sigma_k^* - \delta_1)(\alpha_k - \alpha)\sigma_k^* \sigma_k^\dagger}{(1+\alpha(\sigma_k^* + \delta_1)^2)^2 (1+\alpha_k(\sigma_k^* + \delta_1)^2)^2} \left(\left(\frac{\sigma_k^* + \sigma_k^\dagger}{2} \right)^2 - \sigma_k^* \sigma_k^\dagger \right) \\
&\ge \frac{2\alpha^2 \alpha_k (\sigma_k^* - \delta_1)(\alpha_k - \alpha)\sigma_k^* \sigma_k^\dagger \delta_1^2}{(1+\alpha(\sigma_k^* + \delta_1)^2)^2 (1+\alpha_k(\sigma_k^* + \delta_1)^2)^2},
\end{aligned}$$

where the third equality uses (76), the first and second inequalities follow from (77), and the last inequality uses the definition of δ_1. This, together with the mean value theorem and $\varphi_k(\sigma_k^*) = 0$, yields (74). □

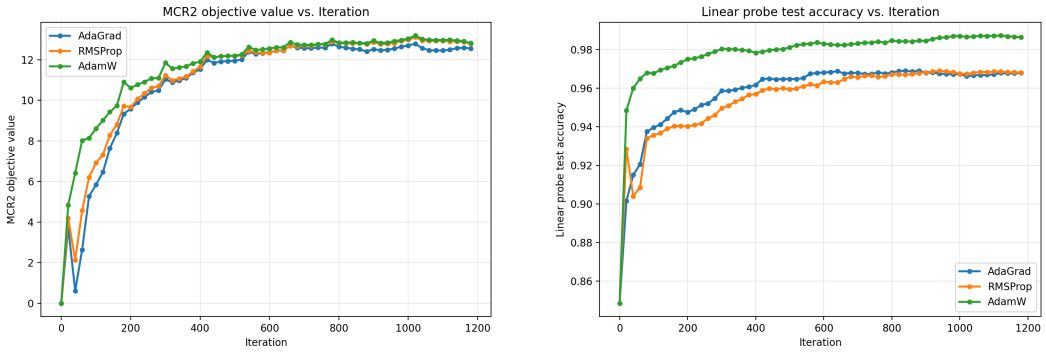
D Extension Experiments

Section 4.2 studies the qualitative two-stage convergence behavior in a deterministic full-batch setting. Here we use a more practical minibatch training setup to examine whether progress on the MCR^2 objective remains aligned with representation quality. This experiment is complementary to the convergence plots in the main text: it tracks the objective together with linear-probe accuracy and class-mean geometry along training.

Setup. We conduct supervised MCR^2 training on MNIST using a CNN encoder with a 64-dimensional normalized feature output. The network is trained for 10 epochs with batch size 512, and we compare three optimizers: *AdaGrad*, *RMSProp*, and *AdamW*. The learning rates are set to 0.01 for AdaGrad and 0.001 for both RMSProp and AdamW, with weight decay 10^{-4}; a cosine annealing scheduler is applied throughout training. To ensure a fair comparison, all optimizers use the same model initialization and the same minibatch order.

Objective value and linear-probe accuracy. During training, we record the MCR^2 objective value and linear-probe test accuracy. The trajectory in Figure 3 shows that optimizers that improve the MCR^2 objective more quickly also tend to obtain stronger downstream linear-probe accuracy. Table 1 reports the same trend at selected iterations. All methods start from the same initialization with identical accuracy, but AdamW consistently attains both a larger objective value and higher linear-probe accuracy during training.

Feature separability. To examine feature separability, we compute cosine-similarity heatmaps of class-mean feature vectors using up to 5000 randomly selected training samples, where each class mean is obtained by averaging normalized features within that class. We plot these heatmaps at selected iterations to visualize how the class structure evolves during optimization in Figure 4. As optimization proceeds, the off-diagonal entries become smaller and the heatmap becomes closer to diagonal, indicating that class-mean features become more separated. This is consistent with the local maximizer structure in Fact 1(ii), where different class subspaces are orthogonal.



(a) Objective value vs. iteration.

(b) Linear-probe accuracy vs. iteration.

Figure 3: Comparison of optimization behavior and representation quality across different optimizers.

Table 1: Objective value and linear-probe test accuracy at selected iterations on MNIST.

Iteration	Optimizer	Objective	Test Acc.
0	AdaGrad	0.0009	84.85%
0	AdamW	0.0009	84.85%
0	RMSProp	0.0009	84.85%
200	AdaGrad	9.8483	95.50%
200	AdamW	11.0919	97.50%
200	RMSProp	10.0209	93.70%
400	AdaGrad	11.7208	96.35%
400	AdamW	12.2218	97.95%
400	RMSProp	11.8589	95.80%
600	AdaGrad	11.5014	97.20%
600	AdamW	11.6748	98.30%
600	RMSProp	11.8196	96.10%
800	AdaGrad	12.7542	97.00%
800	AdamW	13.0624	98.60%
800	RMSProp	12.8779	96.55%
1000	AdaGrad	12.9536	96.70%
1000	AdamW	13.2556	98.65%
1000	RMSProp	13.1013	96.65%

Overall, these extension experiments show a limited optimization–geometry alignment: in this controlled minibatch setting, faster progress on the MCR^2 objective is accompanied by improved linear-probe accuracy and clearer class-mean separation. The theorem itself concerns the MCR^2 objective, and these empirical results should be viewed as a representation-facing diagnostic rather than a broad benchmark comparison.

Reproducibility, variability, and compute. All experiments use standard public datasets. MNIST and CIFAR-10 are accessed through their canonical public releases, and the synthetic experiment generates Gaussian random initializations with the dimensions and hyperparameters specified in Section 4.1. For the convergence plots, the main reported quantities are deterministic optimization trajectories under the stated initialization, objective, step size or learning-rate schedule, and stopping horizon. The deep-network convergence experiments in Section 4.2 are run in a deterministic full-batch setting to isolate the local linear regime. The optimizer comparison controls stochastic variability by using the same model initialization and the same minibatch order for all optimizers, so the reported trajectories are paired.

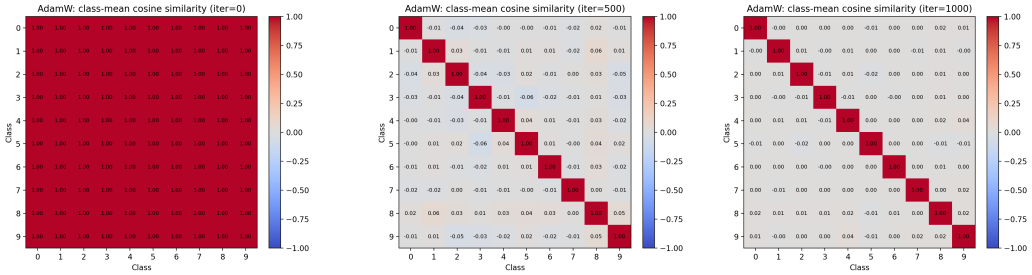


Figure 4: Evolution of the class-mean cosine similarity matrix on MNIST for AdamW.

The reported experiments are lightweight relative to large-scale deep-learning benchmarks. The synthetic and full-batch convergence experiments can be reproduced on a single workstation, and the MNIST minibatch extensions can be run on a single commodity GPU without distributed training. An anonymized supplementary code archive accompanying the submission contains scripts for generating the synthetic data, running the MCR^2 optimization experiments, training the MNIST extension models, and reproducing the reported figures and tables.