Journal of Machine Learning Research 25 (2024) 1-48 Submitted 9/23; Published 5/24

Neural Collapse for Unconstrained Feature Model under Cross-entropy Loss with Imbalanced Data

Wanli Hong *† WH992@NYU.EDU
Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning
New York University Shanghai
Shanghai, 200124, China

Center for Data Science
New York University
New York, NY 10011, USA

Shuyang Ling* SL3635@NYU.EDU
Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning
New York University Shanghai
Shanghai, 200124, China

Editor: Maxim Raginsky

Abstract

Neural Collapse ($\mathcal{NC}$) is a fascinating phenomenon that arises during the terminal phase of training (TPT) of deep neural networks (DNNs). Specifically, for balanced training datasets (each class shares the same number of samples), it is observed that the feature vectors of samples from the same class converge to their corresponding in-class mean features and their pairwise angles are the same. In this paper, we study the extension of $\mathcal{NC}$ phenomenon to imbalanced datasets under cross-entropy loss function in the context of the unconstrained feature model (UFM). Our contribution is multi-fold compared with the state-of-the-art results: (a) we show that the feature vectors within the same class still collapse to the same mean vector; (b) the mean feature vectors no longer share the same pairwise angle. Instead, those angles depend on sample sizes; (c) we also characterize the sharp threshold on which the minority collapse (the feature vectors of the minority groups collapse to one single vector) will happen; (d) finally, we argue that the effect of the imbalance in datasets diminishes as the sample size grows. Our results provide a complete picture of the $\mathcal{NC}$ under the cross-entropy loss for imbalanced datasets. Numerical experiments confirm our theories.

Keywords: Neural Collapse, Minority Collapse, Unconstrained Feature Model, Singular Value Thresholding

*. Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning, New York University Shanghai, China. S.L. and W.H. are (partially) financially supported by the National Key R&D Program of China, Project Number 2021YFA1002800, National Natural Science Foundation of China (NSFC) No.12001372, Shanghai Municipal Education Commission (SMEC) via Grant 0920000112, and NYU Shanghai Boost Fund. W.H. is also supported by NYU Shanghai Ph.D. fellowship and acknowledges the NSF/NRT support.

†. Center for Data Science, New York University.
Correspondence to: Shuyang Ling (sl3635@nyu.edu)

1. Introduction

Deep neural networks (DNNs) have achieved impressive results in various classification tasks He et al. (2016); Krizhevsky et al. (2012); LeCun et al. (2015); Simonyan and Zisserman (2014); Szegedy et al. (2015). However, its highly nonconvex nature along with the massive number of parameters and distinct training paradigms pose great challenges for conducting theoretical analysis. A recent thread of works studies the ways to keep optimizing the model in the terminal phase of training (TPT) when the training loss is very close to zero to achieve a better generalization power Hoffer et al. (2017); Belkin et al. (2019b,a). Therefore, theoretical studies about such over-parametrized neural networks in this regime become helpful in demystifying DNNs so as to design better training paradigms.

Neural collapse ($\mathcal{NC}$) is a phenomenon observed in Papyan et al. (2020) that some particular structures emerge in the feature representation layer and the classification layer of DNNs in the TPT regime for classification tasks when the training dataset is balanced. It has been also observed and studied under the mean-squared loss Han et al. (2021); Poggio and Liao (2021); Zhou et al. (2022a) and in many different settings Ergen and Pilanci (2021); Ji et al. (2021); Tishby and Zaslavsky (2015). For the simplicity of future discussion, we restate the four types of collapses introduced in Papyan et al. (2020):

- $\mathcal{NC}_1$: the feature of samples from the same class converge to a unique mean feature vector;
- $\mathcal{NC}_2$: these feature vectors (after centering by their global mean) form an equiangular tight frame (ETF), i.e., they share the same pairwise angles and length;
- $\mathcal{NC}_3$: the weight of the linear classifier converges to the corresponding feature mean (up to scalar product);
- $\mathcal{NC}_4$: the trained DNN classifies the sample by finding the closest mean feature vectors to the sample feature.

After this empirical finding, many works follow to theoretically explain why $\mathcal{NC}$ occurs in DNNs. Staring from E and Wojtowytsch (2022); Fang et al. (2021); Lu and Steinerberger (2022); Mixon et al. (2020), a thread of works consider the unconstrained feature model (UFM) to simulate the process of training DNNs. The UFM simplifies a deep neural network into an optimization program by treating the features of training data as free variables to optimize over. Such simplification is based upon the rationale of universal approximation theorem Hornik et al. (1989): deep neural networks can well approximate a large variety of functions provided that the neural network is sufficiently over-parameterized. Various versions of UFMs with different loss functions and regularizations are proposed in these works E and Wojtowytsch (2022); Mixon et al. (2020); Zhu et al. (2021); Fang et al. (2021); Dang et al. (2023); Zhou et al. (2022b); Lu and Steinerberger (2022); Tirer and Bruna (2022); Tirer et al. (2023); Mixon et al. (2020); Yaras et al. (2022). They all manage to find that the global minimizers of the empirical risk function under the UFMs match the characterization of $\mathcal{NC}$ proposed in Papyan et al. (2020).

While there are many recent works focusing on balanced datasets, we take a step further to see how $\mathcal{NC}$ generalizes to imbalanced datasets. Several works have already addressed

phenomena in the imbalanced scenario. In particular, Fang et al. (2021) found a phenomenon called *minority collapse* in the TPT regime for the training on imbalanced data. They empirically observed that under cross-entropy loss, as the imbalance ratio goes to infinity, the pairwise angles among minority classes become zero, which means the predictions on the minority classes become indistinguishable. It is believed that if the imbalance ratio is above a certain threshold, this minority collapse occurs but the exact threshold is unknown. A recent work Dang et al. (2023) obtained this exact threshold under the mean-square error (MSE) loss. The work Thrampoulidis et al. (2022) considered the neural collapse for the imbalanced dataset under the unconstrained-feature SVM (UF-SVM) and proposed Simplex-Encoded-Labels Interpolation (SELI) geometry that characterized the structure of global minimizers to the UF-SVM, and later Behnia et al. (2023) extended Thrampoulidis et al. (2022) to several cross-entropy parameterizations.

However, to the best of our knowledge, the $\mathcal{NC}$ on imbalanced datasets under the cross-entropy loss is not fully understood. Our work will try to address a few questions that are not yet answered in the current literature:

(a) Does $\mathcal{NC}_1$ still occur for the UFMs under cross-entropy loss if the data are imbalanced?

(b) If $\mathcal{NC}_1$ holds, what is the structure of the mean feature or prediction matrices?

(c) Can we provide a sharp threshold for the minority collapse?

(d) How does the imbalance ratio affect the structure of the mean feature vectors if the sample size is sufficiently large?

For (a) and (b), these questions are answered under the MSE loss in Dang et al. (2023). However, it becomes much more challenging under the cross-entropy loss. While Fang et al. (2021) studied a special case when there are two giant clusters, a clear characterization of the general case is unknown under the cross-entropy loss. For (c), the threshold for minority collapse remains unknown and we aim to fill this gap. For (d), we have not observed any recent works investigating this issue.

By adopting the UFMs under the cross-entropy loss, we provide a complete picture of $\mathcal{NC}$ for the imbalanced scenario. Here are our main contributions: (i) We provide a concise proof for $\mathcal{NC}_1$ under the cross-entropy loss for imbalanced datasets. This argument is flexible and can be easily applied to other UFM settings and different loss functions. Additionally, we find $\mathcal{NC}_2$ and $\mathcal{NC}_3$ do not hold for imbalanced datasets (Theorem 2). (ii) By working with imbalanced datasets where the classes are partitioned into clusters such that classes from the same cluster share the same number of samples, we show that the mean feature, prediction and classifier weight vector of the classes from the same cluster form an ETF-like structure. Moreover, we show the bias terms corresponding to the same cluster also share the same value (Theorem 2). (iii) When there are only two clusters (majority and minority cluster), which is the same setting adopted in Fang et al. (2021), we provide an exact threshold for the minority collapse. Moreover, we also characterize the threshold for complete collapse (Theorem 3), in which case all the classes collapse to a single vector. (iv) We provide an asymptotic characterization when the number of samples in the majority and minority cluster goes to infinity, but the imbalance ratio stays constant. We find $\mathcal{NC}_2$

and $\mathcal{NC}_3$ hold asymptotically and the convergence rate w.r.t. the sample size is provided (Theorem 5).

1.1 Notation

We let boldface letter $\boldsymbol{X}$ and $\boldsymbol{x}$ be a matrix and a vector respectively; $\boldsymbol{X}^\top$ and $\boldsymbol{x}^\top$ are the transpose of $\boldsymbol{X}$ and $\boldsymbol{x}$ respectively. The matrices $\boldsymbol{I}_n$, $\boldsymbol{J}_n$, and $\boldsymbol{e}_k$ are the $n \times n$ identity matrix, a constant matrix with all entries equal to 1, and the one-hot vector with k-th entry equal to 1. For the simplicity of notation, we also let

$$\boldsymbol{C}_K := \boldsymbol{I}_K - \boldsymbol{J}_K/K \tag{1.1}$$

be the $K \times K$ centering matrix. For any vector $\boldsymbol{x}$, $\mathrm{diag}(\boldsymbol{x})$ denotes the diagonal matrix whose diagonal entries equal $\boldsymbol{x}$. For any matrix $\boldsymbol{X}$, we let $\|\boldsymbol{X}\|$, $\|\boldsymbol{X}\|_F$, and $\|\boldsymbol{X}\|_*$ be the operator norm, Frobenius form, and nuclear norm.

1.2 Organization

The following sections are organized in the following way. In Section 2, we formally introduce $\mathcal{NC}$ and our UFM along with a review of more recent works about $\mathcal{NC}$. In Section 3, we present our main theoretical results. In Section 4, we provide numerical experiments to support our theoretical findings and Section 5 justifies all our theorems.

2. Preliminaries

In this section, we briefly introduce DNNs, and then define the UFM and $\mathcal{NC}$ formally. A deep neural network (DNN) is often in the form of

$$f_\Theta(\boldsymbol{x}) = \boldsymbol{W}^\top \boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{b}$$

where $\boldsymbol{h}_{\boldsymbol{\theta}} \in \mathbb{R}^d$ represents the feature vector on the last layer, $\boldsymbol{W} \in \mathbb{R}^{K\times d}$ and $\boldsymbol{b} \in \mathbb{R}^K$ stand for the weight and bias respectively. The capital letter Θ consists of all the training parameters $(\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{b})$ in the DNN. In addition, we call $\boldsymbol{x}$ the input and $f_\Theta(\boldsymbol{x})$ the prediction vector of $\boldsymbol{x}$. Given the training data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$, we try to find a model via empirical risk minimization (ERM):

$$\min_{\Theta} \ \frac{1}{N}\sum_{i=1}^{N} \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \boldsymbol{y}_i) + \frac{\lambda}{2}\|\Theta\|^2$$

where $\ell(\cdot,\cdot)$ denotes a loss function, $\boldsymbol{y}_i$ is a one-hot vector representing the label of the i-th training data $\boldsymbol{x}_i$, and $\lambda > 0$ is the regularization parameter (i.e., weight decay parameter of SGD). For the classification tasks, we will use the cross-entropy (CE) function $\ell_{CE}(\cdot,\cdot)$, i.e.,

$$\ell_{CE}(\boldsymbol{z}, \boldsymbol{e}_k) = \log \frac{\sum_{\ell=1}^K e^{z_\ell}}{e^{z_k}} = \log \sum_{\ell=1}^K e^{z_\ell} - z_k.$$

We let $\boldsymbol{h}_{ki} := \boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}_{ki})$ be the feature of the i-th data point in the k-th class with $1 \le i \le n_k$ and $1 \le k \le K$, and $N = \sum_{k=1}^K n_k$ is the total number of samples. In other words,

there are in total K different classes with the k-th class containing n_k samples. Without loss of generality, we let $\{n_k\}_{k=1}^K$ be a non-increasing sequence, i.e., $n_1 \geq n_2 \geq \cdots \geq n_K$. To simplify the notation, we let $\boldsymbol{H} \in \mathbb{R}^{d\times N}$ be the feature matrix of all training samples with $\boldsymbol{h}_{ki}$ denoting the $\left(\sum_{i=1}^{k-1} n_k + i\right)$-th column of $\boldsymbol{H}$. Now we are ready to introduce UFMs and $\mathcal{NC}$ related results.

2.1 Unconstrained feature model

For general DNNs, the feature $\boldsymbol{h}_\theta(\cdot)$ is always highly nonlinear and thus challenging to analyze. The unconstrained feature model (UFM) simplifies the DNN model by assuming $\boldsymbol{h}_\theta(\cdot)$ as a free vector, by using the idea that if a neural network is sufficiently parameterized, it can interpolate any data. Under the UFM, we instead study the regularized empirical risk minimization (ERM):

$$\min_{\boldsymbol{W}\in\mathbb{R}^{d\times K},\boldsymbol{H}\in\mathbb{R}^{d\times N}} \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k} \ell_{CE}(\boldsymbol{W}^\top \boldsymbol{h}_{ki}+\boldsymbol{b},\boldsymbol{e}_k) + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_H}{2}\|\boldsymbol{H}\|_F^2 + \frac{\lambda_b}{2}\|\boldsymbol{b}\|^2$$

where $(\lambda_W, \lambda_H, \lambda_b)$ are positive regularization parameters. It has an equivalent matrix form:

$$\min_{\boldsymbol{W}\in\mathbb{R}^{d\times K},\boldsymbol{H}\in\mathbb{R}^{d\times N}} \mathcal{L}(\boldsymbol{W},\boldsymbol{H},\boldsymbol{b}) := \frac{1}{N}\ell_{CE}(\boldsymbol{W}^\top\boldsymbol{H}+\boldsymbol{b}\mathbf{1}_N^\top,\boldsymbol{Y}) + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_H}{2}\|\boldsymbol{H}\|_F^2 + \frac{\lambda_b}{2}\|\boldsymbol{b}\|^2 \tag{2.1}$$

where $\boldsymbol{W} \in \mathbb{R}^{d\times K}$, $\boldsymbol{H} = \{\boldsymbol{h}_{ki}\}_{1\leq i\leq n_k, 1\leq k\leq K} \in \mathbb{R}^{d\times N}$,

$$\boldsymbol{Y} = [\boldsymbol{e}_1\mathbf{1}_{n_1}^\top, \cdots, \boldsymbol{e}_K\mathbf{1}_{n_K}^\top] \in \mathbb{R}^{K\times N}, \tag{2.2}$$

and $\ell_{CE}(\boldsymbol{W}^\top\boldsymbol{H}+\boldsymbol{b}\mathbf{1}_N^\top,\boldsymbol{Y})$ computes the cross entropy column-wisely and then takes the sum.

It is a great convenience to work with model (2.1) as we can convexify the problem. Under $d \geq K$, i.e., in the regime of over-parameterization, then $\boldsymbol{W}^\top\boldsymbol{H}$ can represent any $K \times N$ matrix. Therefore, let $\boldsymbol{Z} = \boldsymbol{W}^\top\boldsymbol{H} \in \mathbb{R}^{K\times N}$ and we have

$$\min_{\boldsymbol{W}^\top\boldsymbol{H}=\boldsymbol{Z}} \lambda_W\|\boldsymbol{W}\|_F^2 + \lambda_H\|\boldsymbol{H}\|_F^2 = 2\sqrt{\lambda_W\lambda_H}\|\boldsymbol{Z}\|_* \tag{2.3}$$

which follows from (Recht et al., 2010, Lemma 5.1) and (Zhu et al., 2021, Lemma A.3). By letting $\lambda_Z := \sqrt{\lambda_W\lambda_H}$, (2.1) becomes

$$\min_{\boldsymbol{Z}\in\mathbb{R}^{K\times N},\boldsymbol{b}\in\mathbb{R}^K} \mathcal{L}(\boldsymbol{Z},\boldsymbol{b}) := \frac{1}{N}\ell_{CE}(\boldsymbol{Z}+\boldsymbol{b}\mathbf{1}_N^\top,\boldsymbol{Y}) + \lambda_Z\|\boldsymbol{Z}\|_* + \frac{\lambda_b}{2}\|\boldsymbol{b}\|^2 \qquad \textbf{(UFM)}$$

which is a convex optimization problem.

Therefore, it suffices to focus on the structure of global minimizers to (**UFM**), as it implies the global minimizer to (2.1). To see this, let $\boldsymbol{Z}^*$ be a global minimizer to (**UFM**) and $\boldsymbol{Z}^* = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be its SVD. Then $\boldsymbol{W} \propto \boldsymbol{\Sigma}^{1/2}\boldsymbol{U}^\top$ and $\boldsymbol{H} \propto \boldsymbol{\Sigma}^{1/2}\boldsymbol{V}^\top$ are actually the global minimizer to (2.1), which follows from (Recht et al., 2010, Lemma 5.1). As a result, our focus will be on analyzing $\boldsymbol{Z}$ instead of its factorized form $\boldsymbol{Z} = \boldsymbol{W}^\top\boldsymbol{H}$. In particular, we will call $\boldsymbol{Z}$ the *prediction matrix*.

2.2 Neural collapse

In this section, we will review more recent works relevant to ours. Regarding the theoretical understanding of $\mathcal{NC}$, the research on the UFMs has become popular in the past few years. Besides the UFM we have introduced above, there are several variants of the UFMs in the state-of-the-art literature. Most works focus on characterizing the global solution of the corresponding regularized empirical risk function and aim to show that it captures $\mathcal{NC}$ phenomenon on the balanced dataset. For more details, we refer readers to works such as Kothapalli et al. (2022); Zhu et al. (2021) and the references therein. With the notation introduced in Section 2.1 at hand, we can describe the $\mathcal{NC}$ more precisely.

- $\mathcal{NC}_1$ - within-class variability collapse: $\boldsymbol{h}_{ki} = \bar{\boldsymbol{h}}_k$ for $1 \leq i \leq n_k$ and $1 \leq k \leq K$;
- $\mathcal{NC}_2$ - convergence of the mean features to an ETF. Let $\bar{\boldsymbol{H}} = [\bar{\boldsymbol{h}}_1, \bar{\boldsymbol{h}}_2, \cdots, \bar{\boldsymbol{h}}_k] \in \mathbb{R}^{d\times K}$ be the mean feature matrix. Then it holds $\bar{\boldsymbol{H}}^\top \bar{\boldsymbol{H}} \propto \boldsymbol{C}_K$, i.e., the mean features form an equiangular tight frame (a regular simplex);
- $\mathcal{NC}_3$ - self-duality. The weight matrix $\boldsymbol{W}$ is proportional to $\bar{\boldsymbol{H}}^\top$.

$\mathcal{NC}$ on balanced datasets: The $\mathcal{NC}$ is first empirically observed on the balanced dataset in Papyan et al. (2020). Hence most follow-up works focus on the balanced scenario, i.e., $n_1 = \cdots = n_K$. The goal is to prove the global minimizer associated with the ERM satisfies $\mathcal{NC}_1$-$\mathcal{NC}_3$ in Papyan et al. (2020) under certain UFMs. The work Fang et al. (2021) studies the neural collapse under the bias-free unconstrained feature model (termed as the layer-peeled model in Fang et al. (2021)), and shows $\mathcal{NC}_1$-$\mathcal{NC}_3$ hold in the balanced scenario with an ℓ_2-norm constraint on $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$. Several works have provided similar results such as E and Wojtowytsch (2022); Lu and Steinerberger (2022); Zhu et al. (2021). In particular, the authors in Zhu et al. (2021) characterize the benign landscape of the regularized ERM by showing that there is only one local minimizer that is also global, modulo a global rotation.

The $\mathcal{NC}$ under the UFMs with MSE loss has also been studied in Dang et al. (2023); Han et al. (2021); Tirer and Bruna (2022); Zhou et al. (2022a). While the within-class collapse $\mathcal{NC}_1$ still holds, $\mathcal{NC}_2$ exhibits a slightly different structure: the mean feature vectors in $\bar{\boldsymbol{H}}$ become pairwise orthogonal, i.e., $\bar{\boldsymbol{H}}^\top \bar{\boldsymbol{H}} \propto \boldsymbol{I}_K$. This is due to the difference between the CE and MSE loss. Other loss functions including loss label smoothing and focal loss have been considered in Zhou et al. (2022b) to demonstrate the universality of $\mathcal{NC}$.

$\mathcal{NC}$ on imbalanced data: The work Fang et al. (2021) is likely the first to consider the $\mathcal{NC}$ for the imbalanced data under the UFM and cross-entropy loss. They work with a dataset consisting of two giant clusters A and B: each cluster A (or B) contains k_A (or k_B) classes and each class contains n_A (or n_B) samples, i.e., $n_A := n_1 = n_2 = \cdots = n_{k_A}$ and $n_B := n_{k_A+1} = n_{k_A+2} \cdots = n_{k_A+k_B}$. Without loss of generality, we assume $n_A > n_B$, and A and B are referred to as majority and minority class respectively. In Fang et al. (2021), it is empirically observed that the $\mathcal{NC}_1$ occurs. Moreover, when the imbalance ratio $r := n_A/n_B$ is greater than some threshold, all the mean feature vectors w.r.t. the minority class become the same, which means the prediction on the classes in B becomes indistinguishable. This phenomenon is termed as the *minority collapse*. Theoretically, Fang et al. (2021) shows minority collapse when the imbalance ratio r is sufficiently large but the exact threshold remains unknown.

For the minority collapse under MSE loss, Dang et al. (2023) has explicitly characterized the collapse threshold for each class in terms of the regularization parameters and number of samples. The argument essentially follows from the idea of singular value thresholding Cai et al. (2010). In particular, Behnia et al. (2023) provides the pairwise angle within the minority and majority classes under two parameterizations of the CE loss.

$\mathcal{NC}$ beyond the UFMs: There are a few other works concerning slightly more complicated models beyond the UFMs. Recently, Yaras et al. (2022) has taken one step forward from the UFMs by restricting the weight $\boldsymbol{w}_k$ and feature $\boldsymbol{h}_{ki}$ on the unit ball, also known as the normalized features, and has analyzed the $\mathcal{NC}$ under this restricted setting. One disadvantage of the UFM is that the model ignores the network depth and nonlinearity, and also the dependence of the feature vector on the input sample. A few progress in this direction include Dang et al. (2023) which considers deep linear networks and explores the $\mathcal{NC}$ under the MSE. In addition, Tirer and Bruna (2022) adds a bit of nonlinearity by applying the ReLU activation to the features $\boldsymbol{H}$ before feeding to the linear classifier. Recently, Seleznova et al. (2023) has explored the connection between the neural collapse and neural tangent kernel Jacot et al. (2018).

$\mathcal{NC}$ and training/generalization Now we briefly review a few other works that are relevant to the $\mathcal{NC}$. Regarding the stability of $\mathcal{NC}$, the work Tirer et al. (2023) considers initializing the input feature near the collapsed solution and conducts perturbation analysis in the near-collapse regime. Motivated by the ETF type mean feature vectors, Yang et al. (2022); Zhu et al. (2021) consider training with the last layer fixed as an ETF; this training scheme achieves on-par performance compared with that with the classifier not fixed. This may be used as a potential way to decrease the computational costs of training DNNs. The works Galanti et al. (2021, 2022) show that few-shot learning achieves good performance by adopting transfer learning on a trained-to-collapsed network except for the last classifier layer. A similar setting of transfer learning is also considered in Li et al. (2022).

Another important aspect is the connection between $\mathcal{NC}$ and generalization Elad et al. (2020); Hui et al. (2022). The recent work Hui et al. (2022) has examined their relation empirically. They find the collapse on the testing dataset does not take place on benchmark datasets including MNIST, FMNIST and Cifar10. They point out that $\mathcal{NC}$ is not desirable in certain transfer learning settings. Additionally, Hui et al. (2022) also observes the $\mathcal{NC}$ starts to occur on a few layers before the last layer, known as the cascading collapse.

3. Main results

The global minimizer to (**UFM**) in the balanced scenario forms exactly an equiangular tight frame E and Wojtowytsch (2022); Fang et al. (2021); Zhu et al. (2021). However, it is unclear how this phenomenon is affected by the number of samples in each class. In this section, we will present our findings on neural collapse under the imbalanced scenario. Before proceeding to our main results, we need to introduce the cluster structure.

Definition 1 (Cluster structure). *Let $\{N_j\}_{j=1}^J$ be the distinct values of $\{n_k\}_{k=1}^K$ with $J \leq K$ and*

$$\Gamma_j = \{k : n_k = N_j,\ 1 \leq k \leq K\} \tag{3.1}$$

We call Γ_j the j-th cluster, i.e., every class in Γ_j has N_j samples.

Now we present the first main theorem, regarding the structure of the global minimizer to (**UFM**).

Theorem 2. *The global minimizer* $(\boldsymbol{Z}, \boldsymbol{b})$ *to* (**UFM**) *is unique and satisfies the following properties:*

(a) **(Within-class feature collapse)** *The* $\mathcal{NC}_1$ *occurs for unconstrained feature models under cross-entropy loss: the prediction vectors* $\boldsymbol{z}_{ki}$, $1 \le i \le n_k$ *within each class collapse to their sample mean* $\bar{\boldsymbol{z}}_k$*:*

$$\boldsymbol{z}_{ki} = \bar{\boldsymbol{z}}_k, \quad 1 \le i \le n_k, \quad \langle \bar{\boldsymbol{z}}_k, \mathbf{1}_K \rangle = 0, \quad 1 \le k \le K.$$

In other words, the prediction matrix $\boldsymbol{Z}$ *is in the following factorized form:*

$$\boldsymbol{Z} = \bar{\boldsymbol{Z}}\boldsymbol{Y} =\in \mathbb{R}^{K \times N}$$

where

$$\bar{\boldsymbol{Z}} = [\bar{\boldsymbol{z}}_1, \cdots, \bar{\boldsymbol{z}}_K], \quad \boldsymbol{Y} \text{ is defined in (2.2).} \tag{3.2}$$

From now on, we refer to $\bar{\boldsymbol{Z}}$ *as the mean prediction matrix.*

(b) **(Block structure of** $\bar{\boldsymbol{Z}}$**)** *The mean prediction matrix* $\bar{\boldsymbol{Z}}$ *and the bias term* $\boldsymbol{b}$ *exhibit the block structure:*

$$\bar{\boldsymbol{Z}} = \sum_{j=1}^{J} a_j \boldsymbol{I}_{\Gamma_j} + \sum_{1 \le j, j' \le J} a_{jj'} \mathbf{1}_{\Gamma_j} \mathbf{1}_{\Gamma_{j'}}^\top, \quad \boldsymbol{b} = \sum_{j=1}^{J} c_j \mathbf{1}_{\Gamma_j}, \quad a_j + \sum_{j'=1}^{J} a_{j'j} |\Gamma_{j'}| = 0,$$

where $\mathbf{1}_{\Gamma_j}$ *is an indicator vector, defined by*

$$\mathbf{1}_{\Gamma_j}(\ell) = \begin{cases} 1, & \ell \in \Gamma_j \\ 0, & \ell \in \Gamma_j^c \end{cases}, \qquad \boldsymbol{I}_{\Gamma_j} = \operatorname{diag}(\mathbf{1}_{\Gamma_j}).$$

In other words, the mean prediction vectors $\{\bar{\boldsymbol{z}}_k\}_{k \in N_j}$ *in the same cluster have the same pairwise angle, so do the mean feature matrix* $\bar{\boldsymbol{H}}$*.*

(c) **(Balanced scenario as a special case)** *If* $n_1 = n_2 = \cdots = n_K = N/K$*, then*

$$\bar{\boldsymbol{Z}} = a\left(K\boldsymbol{I}_K - \boldsymbol{J}_K\right), \quad \boldsymbol{b} = 0.$$

In particular, we have

i. if $N\lambda_Z \ge \sqrt{\frac{N}{K}}$*, then* $a = 0$*;*

ii. if $N\lambda_Z < \sqrt{\frac{N}{K}}$*, then*

$$a = \frac{1}{K} \log\left(\frac{\sqrt{K}}{\sqrt{N\lambda_Z}} - K + 1\right).$$

(d) The weight $\boldsymbol{W}$ *and feature matrix* $\boldsymbol{H}$ *also have a block structure. More precisely, let* $\bar{\boldsymbol{U}}\bar{\boldsymbol{\Sigma}}\bar{\boldsymbol{V}}^\top$ *be the SVD of* $\bar{\boldsymbol{Z}}\boldsymbol{D}^{1/2}$ *where*

$$\boldsymbol{D} := \boldsymbol{Y}\boldsymbol{Y}^\top = \mathrm{diag}(n_1, \cdots, n_K). \tag{3.3}$$

Then $\boldsymbol{H} = \bar{\boldsymbol{H}}\boldsymbol{Y}$ *and the mean prediction* $\bar{\boldsymbol{Z}}$ *equals* $\boldsymbol{W}^\top\bar{\boldsymbol{H}}$ *where*

$$\boldsymbol{W} = \bar{\boldsymbol{\Sigma}}^{1/2}\bar{\boldsymbol{U}}^\top, \quad \bar{\boldsymbol{H}} = \bar{\boldsymbol{\Sigma}}^{1/2}\bar{\boldsymbol{V}}^\top\boldsymbol{D}^{-1/2}.$$

The theorem above indicates the block structure of the global minimizer to (**UFM**). For a numerical illustration, we refer the readers to Figure 2 in our numerical section. In particular, Theorem 2(c) exactly recovers the existing results on the neural collapse for balanced datasets in Fang et al. (2021); Zhu et al. (2021). It is worth pointing out that our proof technique is much more general than those in Fang et al. (2021); Zhu et al. (2021), and a similar argument also applies to the normalized features Yaras et al. (2022).

Despite Theorem 2 characterizes the structure of global minimizers, it does not give insights into how the sample size in each class affects the global minimizers. Next, we focus on a special case where there are two giant clusters, denoted by A and B. In the cluster A (or B), there are k_A (or k_B) classes with the sample size of each individual class equal to n_A (or n_B). Without loss of generality, we let $n_A > n_B$ and refer A (B) as the majority (minority) class. Hence $N = k_A n_A + k_B n_B$ and $K = k_A + k_B$.

Note that Theorem 2(a) implies that the global minimizer of $\boldsymbol{Z}$ and $\boldsymbol{b}$ exhibit block structures. Therefore, we will frequently use the following 2×2 block matrix. We say a matrix $\boldsymbol{X} \in \mathbb{R}^{K\times K}$ equals $\mathcal{B}(a_X, b_X, c_X, d_X)$ if

$$\boldsymbol{X} = \begin{bmatrix} a_X(k_A\boldsymbol{I}_{k_A} - \boldsymbol{J}_{k_A\times k_A}) + c_X k_B \boldsymbol{I}_{k_A} & -b_X\boldsymbol{J}_{k_A\times k_B} \\ -c_X\boldsymbol{J}_{k_B\times k_A} & d_X(k_B\boldsymbol{I}_{k_B} - \boldsymbol{J}_{k_B\times k_B}) + b_X k_A \boldsymbol{I}_{k_B} \end{bmatrix} \in \mathbb{R}^{K\times K}. \tag{3.4}$$

Without loss of generality, we assume $\bar{\boldsymbol{Z}}$ (the within-class mean of $\boldsymbol{Z}$) and $\boldsymbol{b}$ are

$$\bar{\boldsymbol{Z}} = \mathcal{B}(a, b, c, d) \in \mathbb{R}^{K\times K}, \quad \boldsymbol{b} = m\begin{bmatrix} k_B\mathbf{1}_{k_A} \\ -k_A\mathbf{1}_{k_B} \end{bmatrix} \in \mathbb{R}^K \tag{3.5}$$

for some parameter a, b, c, d and m. The next theorem provides a detailed characterization of how the solution structure of $\bar{\boldsymbol{Z}}$ depends on λ_Z. This theorem will be crucial in characterizing the threshold for minority collapse.

Theorem 3 (Block structure v.s. λ_Z). *Assume* $n_A > n_B$, *and* $N = k_A n_A + k_B n_B$ *with* $k_A \geq 2$ *and* $k_B \geq 2$. *For* λ_Z *of different regimes, the optimal solution is* $\boldsymbol{Z} = \bar{\boldsymbol{Z}}\boldsymbol{Y}$ *with the mean prediction matrix* $\bar{\boldsymbol{Z}}$ *in the form of* (3.5).

(a) If $N\lambda_Z \leq \min\{\sqrt{n_A}, \sqrt{n_B}\}$, $\bar{\boldsymbol{Z}}$ *is unique in the form of* (3.5) *that satisfies* $a, b, c, d > 0$ *and*

$$a - c + d - b \leq 0.$$

(b) If $\sqrt{n_B} < N\lambda_Z < \sqrt{n_A}$ *and* $\xi(\lambda_Z, \lambda_b) < 0$ *for some nonlinear function* ξ *in* (5.24), *then*

$$\bar{\boldsymbol{Z}} = \begin{bmatrix} a(k_A\boldsymbol{I}_{k_A} - \boldsymbol{J}_{k_A\times k_A}) + c k_B \boldsymbol{I}_{k_A} & -b\boldsymbol{J}_{k_A\times k_B} \\ -c\boldsymbol{J}_{k_B\times k_A} & \frac{k_A}{k_B} b\boldsymbol{J}_{k_B\times k_B} \end{bmatrix} \in \mathbb{R}^{K\times K}.$$

Moreover, (a, b, c, d) *satisfies* $b > 0$*, and* $c > 0$*. In particular,* $\exists \varepsilon > 0$ *such that for any* $\lambda_Z \in [\sqrt{n_B}/N, \sqrt{n_B}/N + \varepsilon]$*,* $\xi(\lambda_Z, \lambda_b) < 0$ *holds.*

(c) If $\sqrt{n_B} < N\lambda_Z < \sqrt{n_A}$ *and* $\xi(\lambda_Z, \lambda_b) > 0$ *for some nonlinear function* ξ *in* (5.24)*, then*

$$\bar{\boldsymbol{Z}} = \begin{bmatrix} a(k_A \boldsymbol{I}_{k_A} - \boldsymbol{J}_{k_A \times k_A}) & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{K \times K}.$$

Moreover, (a, b, c, d) *satisfies* $b = c = d = 0$*. In particular,* $\exists \varepsilon > 0$ *such that for any* $\lambda_Z \in [\sqrt{n_A}/N - \varepsilon, \sqrt{n_A}/N]$*,* $\xi(\lambda_Z, \lambda_b) > 0$ *holds.*

(d) If $N\lambda_Z > \max\{\sqrt{n_A}, \sqrt{n_B}\}$*, then* $\bar{\boldsymbol{Z}} = 0$*.*

(e) In particular, for the bias-free scenario, i.e., $\lambda_b = \infty$*, then* $\xi(\lambda_Z, \infty) < 0$ (> 0) *is equivalent to* $\sqrt{n_B}/N < \lambda_Z < \lambda^*$ $(\lambda^* < \lambda_Z < \sqrt{n_A}/N)$ *respectively for some* $\lambda^* \in (\sqrt{n_B}/N, \sqrt{n_A}/N)$*. The threshold* λ^* *is the unique solution to a nonlinear equation.*

In Theorem 3, the sign of $\xi(\lambda_{\boldsymbol{Z}}, \lambda_{\boldsymbol{b}})$ signifies the phase transition of parameters b, c, d being nonzero or not. This nonlinear function arises from the discussion about the solution of the first order optimality system concerning a, b, c, d, and m. As a result, its expression is quite complicated. For the clarity of the presentation, we defer its formal definition to the proof. We notice that the threshold for cases (b) and (c) in the bias-free scenario is simpler. This is because it is challenging to prove the monotonicity of the nonlinear function $\xi(\lambda_Z, \lambda_b)$ in λ_Z for any fixed $\lambda_b > 0$ where $\xi(\cdot, \cdot)$ is defined in (5.24). However, for any $\lambda_b > 0$, we are able to show that when λ_Z is close to $\sqrt{n_B}/N$ (or $\sqrt{n_A}/N$), the corresponding ξ satisfies $\xi < 0$ $(\xi > 0)$. But a clear characterization of how $\bar{\boldsymbol{Z}}$ transits from case (b) to (c) is unavailable now. However, it is certain that for $\sqrt{n_B} < N\lambda_Z < \sqrt{n_A}$, the solution is either in case (b) or (c); moreover, the minority collapse occurs as long as $\lambda_Z > \sqrt{n_B}/N$ for the unconstrained feature model, i.e., the mean prediction $\bar{\boldsymbol{Z}}$ on the minority group becomes a single vector, as we can see the right block of $\bar{\boldsymbol{Z}}$ is rank-1 in both case (b) and (c).

The theorem above immediately leads to the following corollary which characterizes the sharp threshold on the minority collapse. It is empirically observed by Fang et al. (2021) that the mean prediction of minority classes collapse to one vector when fixing λ_Z and λ_b as the imbalance ratio $r = n_A/n_B$ increases and is greater than some threshold. Based on Theorem 3, we are able to give an explicit characterization of this critical threshold.

Corollary 4 (Minority collapse threshold ($r \to \infty$, n_B is fixed)). *Suppose* $r = n_A/n_B > 1$*, and* λ_Z, k_A, k_B*, and* n_B *are fixed. Assume*

$$r = \frac{n_A}{n_B} \geq \frac{1}{k_A} \left[\frac{1}{\lambda_Z \sqrt{n_B}} - k_B \right],$$

the mean prediction matrix on minority classes collapses to one vector.

We proceed to provide some asymptotic characterization of the mean prediction $\bar{\boldsymbol{Z}}$, when n_A and n_B go to infinity but their ratio stays constant.

Theorem 5 ($r = n_A/n_B > 1$ **is fixed,** $n_B \to \infty$**).** *Assume*

$$\begin{aligned}
&N\lambda_Z = \lambda < \sqrt{n_B} \ \text{ is constant},\\
&r = n_A/n_B \ \text{ is constant},\\
&N = k_A n_A + k_B n_B, \ \text{with } (k_A, k_B) \text{ fixed},\\
&\lambda_b^{-1} = o(\sqrt{N}\log N),
\end{aligned}$$

then the global minimizer $\bar{\boldsymbol{Z}}$ *of the form* $(a_N^*, b_N^*, c_N^*, d_N^*)$ *in* (3.5) *satisfies*

$$\lim_{N\to\infty} \max\left\{\left|\frac{b_N^*}{c_N^*} - 1\right|, \left|\frac{a_N^*}{c_N^*} - 1\right|, \left|\frac{d_N^*}{b_N^*} - 1\right|\right\} = O\left(\frac{1}{\log N}\right).$$

In other words, the columns of $\bar{\boldsymbol{Z}}$ *converge to the ETF as* $N \to +\infty$ *with* λ *and* $r = n_A/n_B$ *fixed, and so do the corresponding weight* $\boldsymbol{W}^\top$ *and the mean feature matrix* $\bar{\boldsymbol{H}}^\top$.

As the original model (2.1) is non-convex, a natural concern is about the landscape of the original programming. Theorem 3.2 in Zhu et al. (2021) proves a benign optimization landscape for (2.1) in the balanced scenario: all the critical points are either global minimum of (2.1) (also critical points of (**UFM**)) or saddle points. This result has been extended in Zhou et al. (2022b) to characterize the optimization landscape for other loss functions. Regarding the imbalanced scenario under the UFM and cross-entropy loss, the optimization landscape of (2.1) is also benign.

Theorem 6. *Assume the feature dimension* $d > K$*, then* $\mathcal{L}(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ *in* (2.1) *is a strict saddle function with no spurious local minimum, in the sense that*

- *Any local minimizer of* (2.1) *is a global minimizer.*
- *Any critical point* $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ *that is not a local minimizer is a strict saddle point with negative curvature, i.e. the Hessian* $\nabla^2\mathcal{L}(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$*, at this critical point, is non-degenerate and has at least one negative eigenvalue.*

Theorem 6 is a direct generalization of Theorem 3.2 in Zhu et al. (2021) from the balanced case to the imbalanced case. The proof (see Section C.1 in Zhu et al. (2021)) also directly applies without any changes, and thus we do not repeat the proof here. For the completeness of the presentation, we briefly discuss the proof idea, which is to classify the critical points of (2.1) into two categories. We denote the cross-entropy loss part of (2.1) as $\mathcal{L}_1(\boldsymbol{X}) = N^{-1}\ell_{CE}(\boldsymbol{X}, \boldsymbol{Y})$. For any critical points $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ of (2.1), if

- $\left\|\nabla\mathcal{L}_1(\boldsymbol{W}^\top\boldsymbol{H} + \boldsymbol{b}\mathbf{1}_N^\top)\right\| \le \sqrt{\lambda_W\lambda_H}$: one can show these points are also critical points of the convexified programming (**UFM**), and thus the global minimum. Intuitively, the inequality constrains the norm of the gradient, so it becomes a legal subgradient of the nuclear norm.
- $\left\|\nabla\mathcal{L}_1(\boldsymbol{W}^\top\boldsymbol{H} + \boldsymbol{b}\mathbf{1}_N^\top)\right\| > \sqrt{\lambda_W\lambda_H}$: One can construct a negative curvature direction in the null space of $\boldsymbol{W}$, which is nonempty since $d > K$, and the singular vector corresponding to the largest singular value of $\nabla^2\mathcal{L}_1(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$.

We conclude this section by discussing our results and pointing out a few possible future directions. In conclusion, we present a rigorous and in-depth study into the neural collapse phenomenon for imbalanced dataset using the UFM with cross-entropy loss. In particular, we give a full characterization of the solution when the dataset has two clusters under the UFM, thus precisely finding the minority collapse threshold. This sharp threshold in Corollary 4 is also confirmed in real experiments. As a result, one can select a suitable oversampling rate of minority classes to avoid minority collapse while saving computational resources and also not impairing test performance due to the high oversampling rate. One may also wonder whether the minority collapse and the emergence of block structure in the mean features still occur in absence of the regularization, i.e., $\lambda_{\boldsymbol{Z}} = 0$. We confirm numerically that these structures do not occur when training networks by using the SGD with zero weight decay. In addition, the feature collapse is not guaranteed either. These numerical observations imply that the regularization is essential to the neural collapse and minority collapse.

As later shown in Section 4, our theory can only partially explain the behavior of real deep neural networks. For example, we can see non-negligible difference arises in certain regime between the predicted solution by the UFM and the actual prediction by the DNNs. Despite the landscape of UFM is benign by Theorem 6, the landscape of DNN is inherently different. This calls for more complicated models to explain DNNs. Several works Tirer and Bruna (2022); Dang et al. (2023) try to add more linear layers to UFM, but adding even one layer of nonlinearity remains unexplored, which could be a future direction. It will be also interesting to find a weaker substitute of the UFM that interpolates between the DNNs and UFM. Finally, our paper does not discuss the relation between neural collapse and generalization error. To the best of our knowledge, most studies on that topic remain empirical. Addressing this issue theoretically will lead to a deeper understanding of the interplay between generalization and implicit bias broadly. We will leave these possible directions for future work.

4. Numerics

In this section, we present numerical results that confirm our theory and also give new insights. Our code is available on Github and is adapted from the code by Fang et al. (2021). In the next four subsections, we show (a) the neural collapse phenomena arising from the imbalanced dataset; (b) the block structure of the mean prediction matrix $\bar{\boldsymbol{Z}}$, and the difference between the feature mean and the exact solution obtained from solving (2.1); (c) the sharp threshold of the minority collapse; and (d) the asymptotic behavior of $\bar{\boldsymbol{Z}}$ as the sample size grows to infinity with fixed imbalance ratio.

We first briefly describe the training details. To make the experiments and settings in (2.1) consistent, we place activation regularization on the last layer before the classification layer to model the regularization on features for every network and dataset we have trained. All the networks, if not specified, are trained with a diminishing stepsize, as adopted in Fang et al. (2021): the initial learning rate is 0.1 for the first 1/6 epochs; and after the first 1/6 epochs, we divide the learning rate by 10, i.e., learning rate equals 0.01, and train for another 1/6 epochs. After that, we set the learning rate as 10^{-3} for the rest of the training process. All the networks are trained by SGD with momentum 0.9, and batch

size 128. Additionally, except for networks trained in subsection 4.3 where we turn off the weight decay of SGD, we set the weight decay of SGD to be 5e-4 during training. Since we are comparing the minority collapse threshold against regularization parameters in 4.3, we turn the weight decay off to make the regularization effect exact.

In Section 4.1 and 4.2, we train three neural network including VGG11, VGG13, and ResNet18 on Fashion MNIST (FMNIST) and Cifar10. To validate our theory under different imbalance levels, we pick the following two choices of parameters, denoted by Dataset_1 and Dataset_2. The whole dataset (either FMNIST or Cifar 10) contains three giant clusters A, B, and C. For each group (e.g. A), it contains k_A classes and each class contains n_A samples. In other words, there are in total $k_A + k_B + k_C$ classes and $k_A n_A + k_B n_B + k_C n_C$ samples. The sample size of each class in the same cluster is the same.

1. Dataset_1: $k_A = 4, k_B = k_C = 3, n_A = 5000, n_B = 4000, n_C = 3000$
2. Dataset_2: $k_A = k_C = 4, k_B = 2, n_A = 5000, n_B = 3000, n_C = 1000$

From the settings above, we can see Dataset_2 is more imbalanced than Dataset_1. For the regularization parameters, we use $\lambda_W = 10^{-3}, \lambda_H = 10^{-6}, \lambda_b = 10^{-2}$ and train each model for 1000 epochs.

In Section 4.3 and 4.4, the experiments are used to verify Theorem 3, and Theorem 5 and Corollary 4 respectively. Therefore, we only consider two giant clusters A and B, with the number of classes and within-class sample size equal to (k_A, n_A) and (k_B, n_B) respectively. The regularization parameters λ_W and λ_H are set as $\lambda_W = 10\lambda_Z$ and $\lambda_H = \lambda_Z/10$ for each given λ_Z. We will provide the specific parameter settings in each section. The network is ResNet18 and it is trained on Cifar10 by using SGD for 2000 epochs.

4.1 Collapse of feature and prediction vectors

We first show the collapse of within-class feature vectors, as predicted by Theorem 2(a). To quantify the level of within-class collapse, we compute the within-class and between-class covariance:

$$\boldsymbol{\Sigma}_W := \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\boldsymbol{h}_{ki} - \bar{\boldsymbol{h}}_k)(\boldsymbol{h}_{ki} - \bar{\boldsymbol{h}}_k)^\top, \quad \boldsymbol{\Sigma}_B := \frac{1}{K}\sum_{k=1}^{K}(\bar{\boldsymbol{h}}_k - \boldsymbol{h}_G)(\bar{\boldsymbol{h}}_k - \boldsymbol{h}_G)^\top$$

where

$$\boldsymbol{h}_G := \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\boldsymbol{h}_{ki}, \quad \bar{\boldsymbol{h}}_k := \frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{h}_{ki}, \quad 1 \le k \le K,$$

are the total and within-class means respectively. The level of within-class collapse is measured by

$$\mathcal{NC}_1 := \frac{1}{K}\operatorname{Tr}\left(\boldsymbol{\Sigma}_W \boldsymbol{\Sigma}_B^\dagger\right). \tag{4.1}$$

It is easy to see that if the within-class collapse occurs, $\mathcal{NC}_1$ should be very small, as $\boldsymbol{\Sigma}_W$ is close to 0.

Figure 1 plots the change of $\log \mathcal{NC}_1$ against the epochs. We can see after training 1000 epochs, $\log \mathcal{NC}_1$ is near -10 across all networks and datasets except VGG11 on the Cifar10 dataset which attains $\log \mathcal{NC}_1 \approx -6$. This is strong evidence of the within-class collapse, and it confirms our Theorem 2.

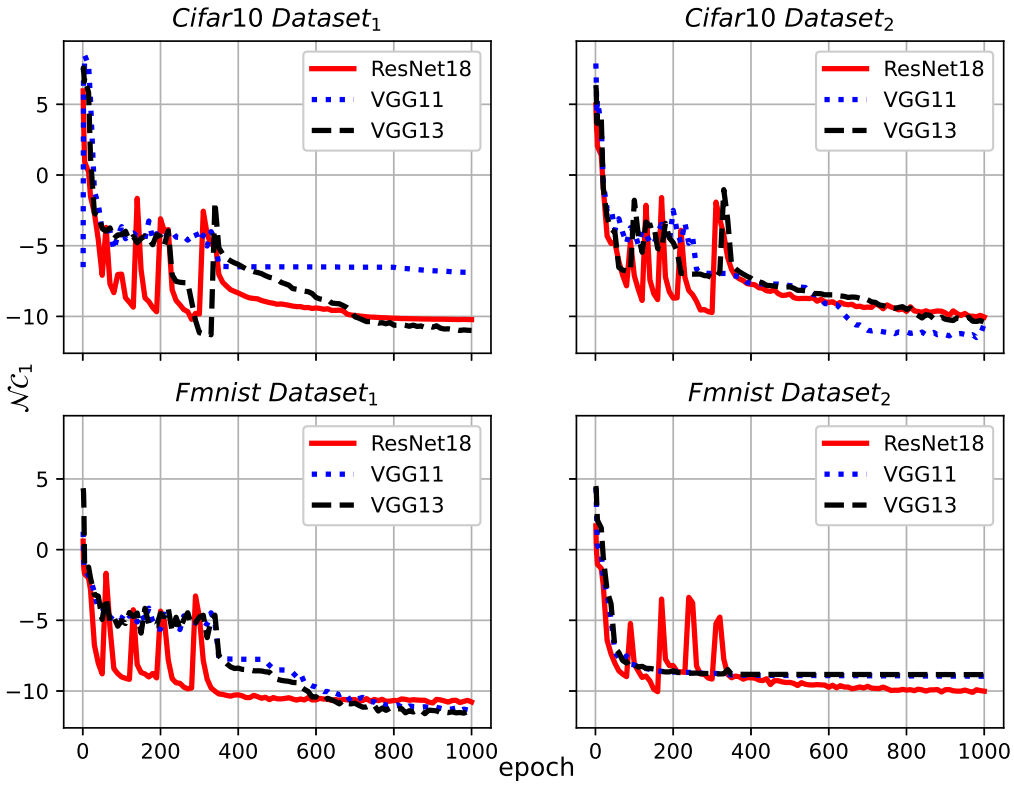

Figure 1: Plot of $\log \mathcal{NC}_1$ v.s. epochs: x-axis is the epoch number and y-axis is $\log \mathcal{NC}_1$. Datasets: Cifar10 and FMNIST with two sets of parameters Dataset$_1$ and Dataset$_2$. Network: ResNet18 (red straight line), VGG11 (blue dotted line), and VGG13 (black dashed line).

4.2 Block structure of the mean prediction and features

In this subsection, we will illustrate the block structure of the mean prediction matrix to confirm Theorem 2(b). We also compare the difference of the mean prediction matrix $\bar{\boldsymbol{Z}} = [\boldsymbol{W}^\top \bar{\boldsymbol{h}}_k]_{1\le k\le K}$ and the solution $\bar{\boldsymbol{Z}}^*$ to the unconstrained feature model with $\lambda_Z = \sqrt{\lambda_W \lambda_H}$. The datasets and networks are exactly the same as those in Section 4.1.

In Figure 2, we plot the final mean prediction matrix $\bar{\boldsymbol{Z}}$ with the k-th column being $\bar{\boldsymbol{z}}_k = \boldsymbol{W}^\top \bar{\boldsymbol{h}}_k$ over the 12 experiments computed in Figure 1. The entries in $\bar{\boldsymbol{Z}}$ of the largest magnitude show up on the diagonal. To show a stronger contrast in the plot, we apply min-max standardization across all the mean prediction matrices. The white dashed lines separate giant clusters into 3×3 blocks which match the setting of Dataset$_1$ and Dataset$_2$. We see all the entries in each off-diagonal block, and all the off-diagonal entries in each diagonal block share a very similar magnitude in their own block. This indicates the block structure of the mean prediction matrix $\bar{\boldsymbol{Z}}$ and also that of the mean feature vectors.

Figure 3 and 4, we select two experiments to show the difference between the trained predictions $\bar{\boldsymbol{Z}}$ and $\boldsymbol{b}$ and the solution $\bar{\boldsymbol{Z}}^*$ and $\boldsymbol{b}^*$ to (2.1). For VGG13 trained on Dataset$_1$,

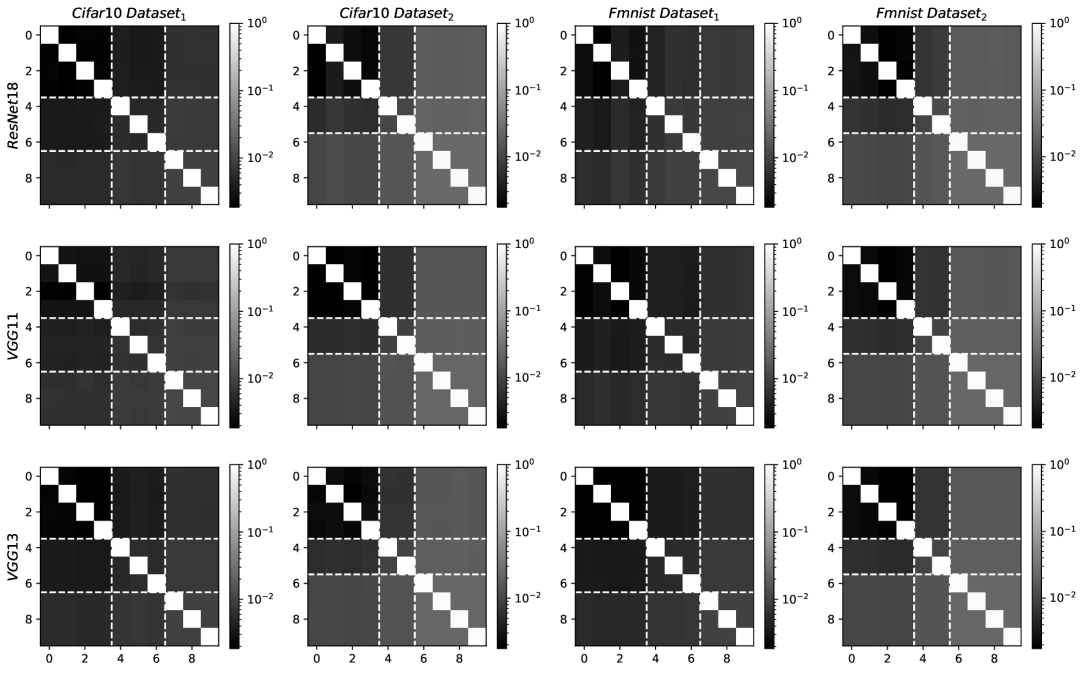

Figure 2: Standardized (over 12 matrices) mean prediction matrix $\bar{\boldsymbol{Z}}$ for all 12 experiments in Figure 1. The white dashed lines separate clusters A, B and C in Dataset_1 and Dataset_2.

Figure 3 (Left) shows a decreasing trend of the relative error between $\bar{\boldsymbol{Z}}$ and $\bar{\boldsymbol{Z}}^*$ which stabilizes around 0.04 under both Frobenius and supreme norm. The bias difference is relatively higher and stabilizes around 0.2; for VGG11 trained on Dataset_2, the relative error is higher compared to the previous one, possibly because the Dataset_2 is more imbalanced. Despite the relative error is approximately 0.1, the final mean prediction matrix shown in Figure 4 implies that $\bar{\boldsymbol{Z}}$ and $\bar{\boldsymbol{Z}}^*$ share a similar block structure.

4.3 Minority collapse

Theorem 3 and Corollary 4 show the sharp threshold on λ_Z so that the prediction made by the neural network on the minority classes collapses to a single vector. For the experiments below, we adopt a slightly different learning rate scheme. To prevent the features and weights from vanishing due to the large regularization terms, we initialize λ_H and λ_W to be ten times smaller for the first 1/6 epochs with stepsize 0.1. Then we set back the regularization parameters and keep training for the next 1/6 epochs with the stepsize 0.1.

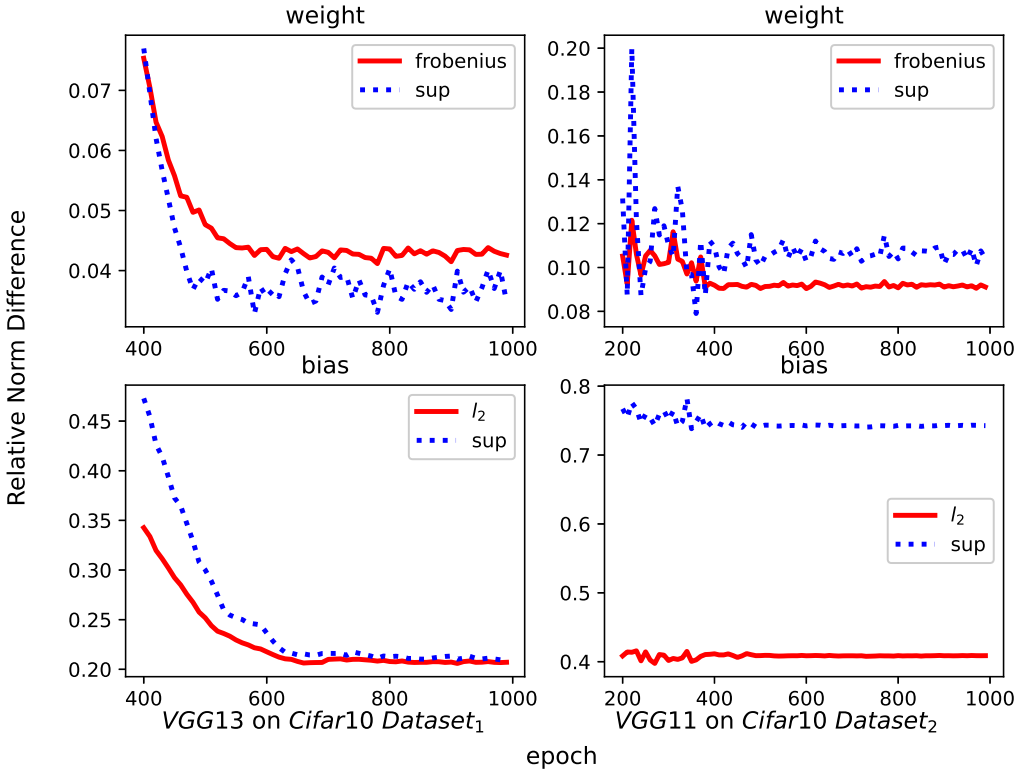

Figure 3: Relative error $\|\bar{\boldsymbol{Z}} - \bar{\boldsymbol{Z}}^*\|_F/\|\bar{\boldsymbol{Z}}^*\|_F$ ($\|\boldsymbol{b} - \boldsymbol{b}^*\|_2/\|\boldsymbol{b}^*\|_2$) and $\|\bar{\boldsymbol{Z}} - \bar{\boldsymbol{Z}}^*\|_\infty/\|\bar{\boldsymbol{Z}}^*\|_\infty$ ($\|\boldsymbol{b} - \boldsymbol{b}^*\|_\infty/\|\boldsymbol{b}^*\|_\infty$) v.s. the epoch for VGG13 on Cifar10 with Dataset$_1$ (Left) and VGG11 on Cifar10 with Dataset$_2$ (Right). The starting epoch numbers are chosen to be 400 and 200 when $\mathcal{NC}_1$ has reaches a low level.

For the next 1/3 epochs, we set the learning rate as 0.01. After that, we keep the learning rate equal to 10^{-3} for the rest.

We design two types of experiments to verify our theoretical findings. The first type fixes $n_A = 500$ and $n_B = 100$ for a given pair of (k_A, k_B), and $\lambda_b = 0.01$. Then we vary λ_Z and run ResNet18 on Cifar10 dataset for each λ_Z. For $k_A = k_B = 5$, the results are shown in Figure 5: it implies that the minority collapse occurs at $\lambda_Z = 0.0033$, which matches $\sqrt{n_B}/N = 1/300$ where $N = k_A n_A + k_B n_B = 3000$. However, our Theorem 3 fails to predict the threshold beyond which all the predictions become constant: the theoretical threshold is $\sqrt{500}/3000 \approx 0.075$ while Figure 5 shows the complete collapse for some $\lambda_Z \leq 0.069$ which is strictly smaller than 0.075. For $k_A = 3$ and $k_B = 7$, Theorem 3 predicts the minority and complete collapse occur at approximately $\lambda_Z = 0.0045$ and 0.0102 respectively. Figure 6 implies the empirical threshold for minority collapse matches our theoretical prediction while that for the complete collapse is between 0.0086 and 0.0094, strictly smaller than 0.0102.

In the second type, we fix $\lambda_Z = 0.005$, $\lambda_b = 0.01$, and $n_B = 100$. Then we let n_A increase from 100 to 1400, and compute the mean prediction matrix for each set of parameters. For

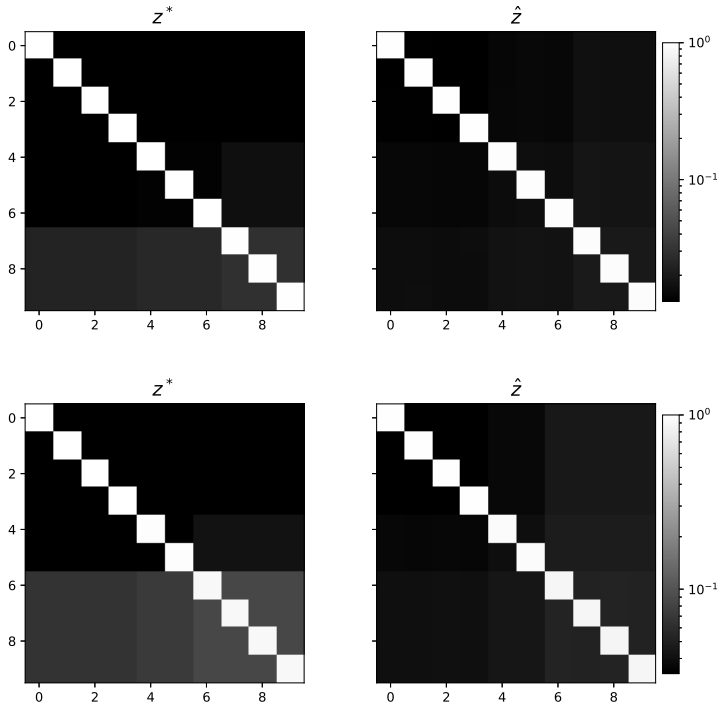

Figure 4: Comparison of the min-max standardized final mean prediction matrix $\bar{\boldsymbol{Z}}$ and $\bar{\boldsymbol{Z}}^*$. Top: VGG13 on Cifar10 with Dataset$_1$; Bottom: VGG11 on Cifar10, with Dataset$_2$.

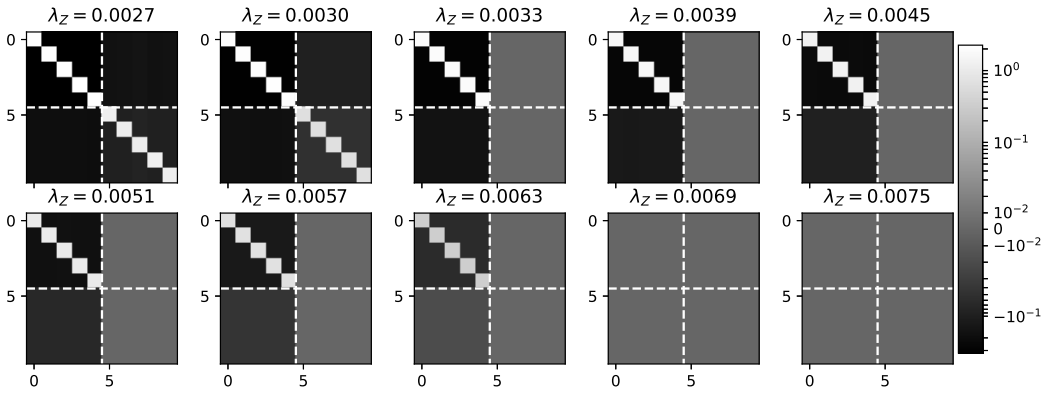

Figure 5: Plot for the mean prediction matrix $\bar{\boldsymbol{Z}}$ for 10 classes with $k_A = k_B = 5$ v.s. varying λ_Z. The white dashed lines separate clusters majority group A and minority group B.

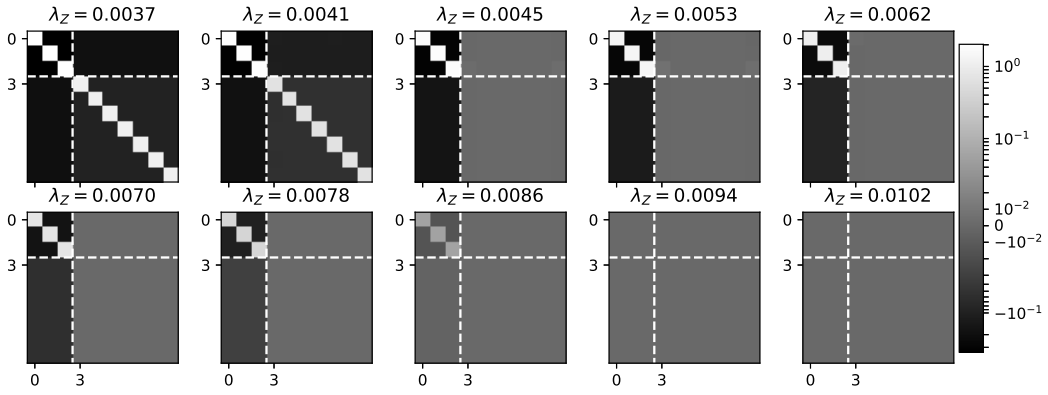

Figure 6: Plot for the mean prediction matrix $\bar{\boldsymbol{Z}}$ for 10 classes with $k_A = 3$ and $k_B = 7$ v.s. varying λ_Z. The white dashed lines separate clusters majority group A and minority group B.

$k_A = k_B = 5$, our theory predicts the threshold of n_A for minority and complete collapse are $n_A \approx 300$ and 1392 respectively. Our numerical experiments in Figure 7 confirm the threshold for minority collapse but the theory overestimates the threshold for complete collapse. Similar phenomena are also observed for $k_A = 3$ and $k_B = 7$ in which the thresholds for n_A are 433 and 3964 respectively, as shown in Figure 8.

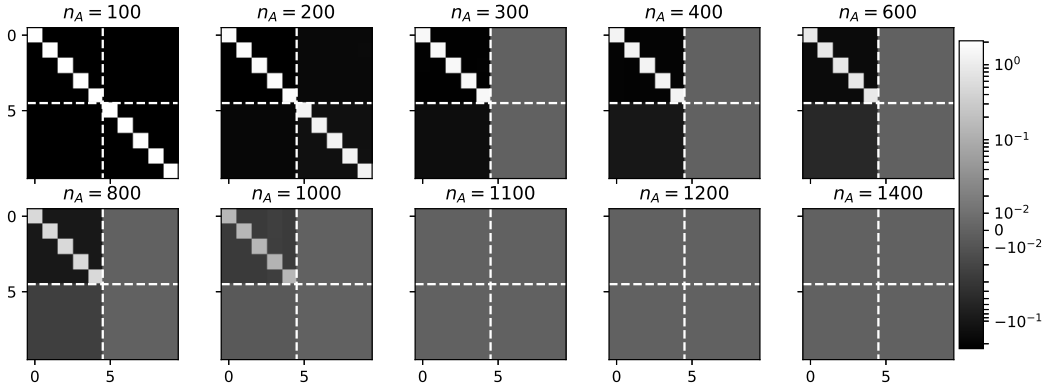

Figure 7: Plot for the mean prediction matrix $\bar{\boldsymbol{Z}}$ for 10 classes with $k_A = k_B = 5$ v.s. n_A.

Based on the Figure 5-8, we make the following main observations: (i) the threshold of minority collapse $\lambda_Z = \sqrt{n_B}/N$ matches the empirical experiments. However, the theoretical threshold for complete collapse $\lambda_Z = \sqrt{n_A}/N$ tends to overestimate; (ii) all four figures confirm the block structure of the mean prediction matrix that is characterized by cases (a), (b), and (d) in Theorem 3. For case (c), we can take a look at the 8th subfigure

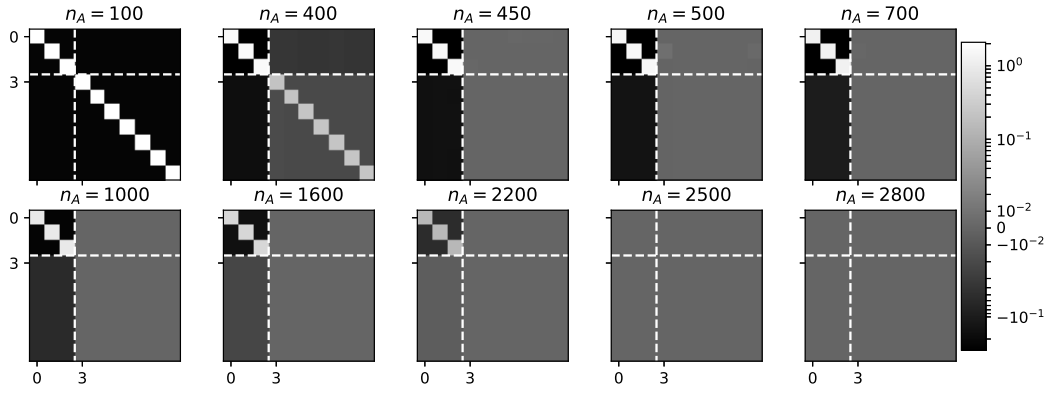

Figure 8: Plot for the mean prediction matrix $\bar{\boldsymbol{Z}}$ for 10 classes with $k_A = 3$ and $k_B = 7$ v.s. n_A.

($\lambda_Z = 0.0063$) in Figure 5. The right blocks and lower left block have a much smaller magnitude compared with the upper left blocks. However, the entries in the lower left blocks (the order is 10^{-3}) still are much larger than those on the right block (the order is 10^{-7}). Therefore, we do not see a strong signal of the case (c) for λ_Z before the complete collapse occurs.

4.4 Convergence to the ETF

In this section, we will carry out some experiments for Theorem 5. For the parameters, we fix parameters $k_A = 5, k_B = 5, r = n_A/n_B = 2, N\lambda_Z = 0.1$, and $\lambda_b = 0.01$. We train ResNet18 on the Cifar10 dataset with different n_A: n_A ranges from 500 to 5500. For each set of parameters, we run 2000 epochs and compute the pairwise correlation (i.e., cosine angle) for mean prediction vectors $\bar{\boldsymbol{Z}}$, i.e.,

$$\widehat{\boldsymbol{\Theta}} := \operatorname{diag}(\bar{\boldsymbol{Z}}^\top \bar{\boldsymbol{Z}})^{-1/2} \bar{\boldsymbol{Z}}^\top \bar{\boldsymbol{Z}} \operatorname{diag}(\bar{\boldsymbol{Z}}^\top \bar{\boldsymbol{Z}})^{-1/2}.$$

Due to the block structure of $\bar{\boldsymbol{Z}}$, the variance of angles in each block of $\widehat{\boldsymbol{\Theta}}$ is quite small, and thus here, we only plot the mean within-class correlation for A and B respectively, and the mean correlation between A and B. Here $K = k_A + k_B = 10$, and the pairwise correlation is $-1/9$ for the ETF. Figure 9 shows a clear convergence of all the three groups of mean correlation toward $-1/9$ as n_A increases, i.e., the mean prediction vectors slowly converge to an ETF. This validates our result in Theorem 5, i.e., the impact created by the imbalance in data size on the prediction of neural networks diminishes as the number of training samples increases. We also conducted experiments with higher imbalance ratio r. The trend of mean features converging towards the ETF is also observed but that will require much larger sample sizes.

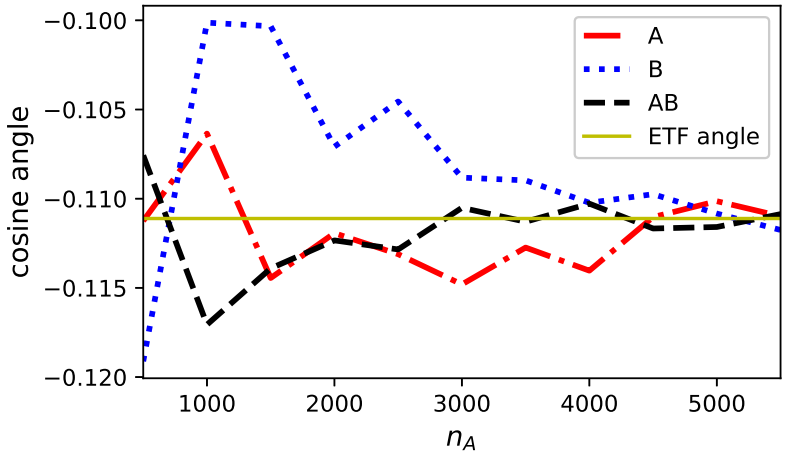

Figure 9: y-axis plots the mean pairwise correlation for mean prediction vector $\bar{\boldsymbol{Z}}$ within the cluster A (red dot-dashed) and B (blue dotted), and between clusters A and B (black dashed); x-axis is the size of n_A between 500 and 5500. The pairwise correlation of the ETF for 10 classes is denoted by the yellow straight line.

5. Proof

5.1 Basic facts and optimality condition

This subsection establishes important lemmas that will be used for proving our main theorems.

Lemma 7. *Define*

$$\varphi(\boldsymbol{Z}, \boldsymbol{b}) = \frac{1}{N}\ell_{CE}(\boldsymbol{Z} + \boldsymbol{b}\mathbf{1}_N^\top, \boldsymbol{Y}) = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\ell_{CE}(\boldsymbol{z}_{ki} + \boldsymbol{b}, \boldsymbol{e}_k) \tag{5.1}$$

and then $\varphi(\boldsymbol{Z}, \boldsymbol{b})$ is strongly convex in the direction $(\boldsymbol{\Delta}_Z, \boldsymbol{\Delta}_b) \in \mathbb{R}^{K\times N} \oplus \mathbb{R}^K$ that belongs to $\{(\boldsymbol{\Delta}_Z, \boldsymbol{\Delta}_b) : \mathbf{1}_K^\top(\boldsymbol{\Delta}_Z + \boldsymbol{\Delta}_b\mathbf{1}_N^\top) = 0\}$.

Proof The proof is straightforward, and it suffices to show the quadratic form

$$[\boldsymbol{\Delta}_Z, \boldsymbol{\Delta}_b] : \nabla^2_{\boldsymbol{Z},\boldsymbol{b}}\varphi(\boldsymbol{Z}, \boldsymbol{b}) : [\boldsymbol{\Delta}_Z, \boldsymbol{\Delta}_b] \geq \lambda(\boldsymbol{Z}, \boldsymbol{b}) \cdot \left\|\boldsymbol{\Delta}_Z + \boldsymbol{\Delta}_b\mathbf{1}_N^\top\right\|_F^2,$$

for every $(\boldsymbol{\Delta}_Z, \boldsymbol{\Delta}_b)$ satisfying $\mathbf{1}_K^\top(\boldsymbol{\Delta}_Z + \boldsymbol{\Delta}_b\mathbf{1}_N^\top) = 0$ where $\lambda(\boldsymbol{Z}, \boldsymbol{b})$ is a strictly positive number that only depends on $(\boldsymbol{Z}, \boldsymbol{b})$. For ease of notation, define

$$\boldsymbol{p}_{ki} := \frac{\exp(\boldsymbol{z}_{ki} + \boldsymbol{b})}{\langle \exp(\boldsymbol{z}_{ki} + \boldsymbol{b}), \mathbf{1}_K\rangle}, \quad \boldsymbol{P} = [\boldsymbol{p}_{ki}]_{1\leq i\leq n_k, 1\leq k\leq K} \in \mathbb{R}^{K\times N} \tag{5.2}$$

as the probability vector associated with $\boldsymbol{z}_{ki} + \boldsymbol{b}$. The gradient of φ is

$$\frac{\partial\varphi}{\partial\boldsymbol{z}_{ki}} = \frac{1}{N}(\boldsymbol{p}_{ki} - \boldsymbol{e}_k), \quad \frac{\partial\varphi}{\partial\boldsymbol{b}} = \frac{1}{N}\sum_{k,i}(\boldsymbol{p}_{ki} - \boldsymbol{e}_k)$$

whose matrix form is

$$\frac{\partial \varphi}{\partial \boldsymbol{Z}} = \frac{1}{N}(\boldsymbol{P} - \boldsymbol{Y}), \quad \frac{\partial \varphi}{\partial \boldsymbol{b}} = \frac{1}{N}(\boldsymbol{P} - \boldsymbol{Y})\mathbf{1}_N.$$

The corresponding Hessian is

$$\begin{aligned}
&\frac{\partial^2 \varphi}{\partial \boldsymbol{z}_{ki}^2} = \frac{1}{N}\left(\operatorname{diag}(\boldsymbol{p}_{ki}) - \boldsymbol{p}_{ki}\boldsymbol{p}_{ki}^\top\right), \quad \frac{\partial^2 \varphi}{\partial \boldsymbol{z}_{ki}\partial \boldsymbol{z}_{k'i'}} = 0, \ \ \forall (k,i) \neq (k',i'), \\
&\frac{\partial^2 \varphi}{\partial \boldsymbol{b}^2} = \frac{1}{N}\sum_{k,i}\left(\operatorname{diag}(\boldsymbol{p}_{ki}) - \boldsymbol{p}_{ki}\boldsymbol{p}_{ki}^\top\right), \quad \frac{\partial^2 \varphi}{\partial \boldsymbol{z}_{ki}\partial \boldsymbol{b}} = \frac{1}{N}(\operatorname{diag}(\boldsymbol{p}_{ki}) - \boldsymbol{p}_{ki}\boldsymbol{p}_{ki}^\top).
\end{aligned}$$

Note that $\boldsymbol{p}_{ki}\boldsymbol{p}_{ki}^\top$ is a positive matrix and thus the associated Laplacian $\operatorname{diag}(\boldsymbol{p}_{ki}) - \boldsymbol{p}_{ki}\boldsymbol{p}_{ki}^\top$ is positive semidefinite with its second smallest eigenvalue strictly positive. Let $\boldsymbol{\Delta}_{Z,ki}$ be the difference in the variable $\boldsymbol{z}_{ki}$, and then the quadratic form equals

$$\begin{aligned}
&[\boldsymbol{\Delta}_Z, \boldsymbol{\Delta}_b] : \nabla^2_{\boldsymbol{Z},\boldsymbol{b}}\varphi(\boldsymbol{Z},\boldsymbol{b}) : [\boldsymbol{\Delta}_Z, \boldsymbol{\Delta}_b] \\
&= \frac{1}{N}\sum_{k,i}(\boldsymbol{\Delta}_{Z,ki} + \boldsymbol{\Delta}_b)^\top \left(\operatorname{diag}(\boldsymbol{p}_{ki}) - \boldsymbol{p}_{ki}\boldsymbol{p}_{ki}^\top\right)(\boldsymbol{\Delta}_{Z,ki} + \boldsymbol{\Delta}_b) \\
&\geq \frac{1}{N}\sum_{k,i}\lambda_2(\operatorname{diag}(\boldsymbol{p}_{ki}) - \boldsymbol{p}_{ki}\boldsymbol{p}_{ki}^\top)\|\boldsymbol{\Delta}_{Z,ki} + \boldsymbol{\Delta}_b\|^2 \\
&\geq \frac{1}{N}\min_{k,i}\lambda_2(\operatorname{diag}(\boldsymbol{p}_{ki}) - \boldsymbol{p}_{ki}\boldsymbol{p}_{ki}^\top)\cdot\|\boldsymbol{\Delta}_Z + \boldsymbol{\Delta}_b\mathbf{1}_N^\top\|_F^2
\end{aligned}$$

where the first inequality follows from the fact that $\left(\operatorname{diag}(\boldsymbol{p}_{ki}) - \boldsymbol{p}_{ki}\boldsymbol{p}_{ki}^\top\right)\mathbf{1}_K = 0$ and $\langle \mathbf{1}_K, \boldsymbol{\Delta}_{Z,ki} + \boldsymbol{b}\rangle = 0$ for all $1 \leq i \leq n_k$ and $1 \leq k \leq K$.

■

Lemma 8 (**Optimality condition**). *The first-order optimality condition of $\mathcal{L}(\boldsymbol{Z},\boldsymbol{b})$ in* (**UFM**) *is*

$$N^{-1}(\boldsymbol{Y} - \boldsymbol{P}) \in \lambda_Z \partial\|\boldsymbol{Z}\|_*, \quad N^{-1}(\boldsymbol{Y} - \boldsymbol{P})\mathbf{1}_N = \lambda_b \boldsymbol{b} \tag{5.3}$$

where $\boldsymbol{P} = [\boldsymbol{p}_{ki}]_{1\leq i\leq n_k, 1\leq k\leq K}$, $\boldsymbol{p}_{ki}$ are defined in (5.2)*, and $\partial\|\boldsymbol{Z}\|_*$ stands for the subdifferential of the nuclear norm at $\boldsymbol{Z}$. In particular, the global minimizer $(\boldsymbol{Z},\boldsymbol{b})$ satisfies $\mathbf{1}_K^\top \boldsymbol{Z} = 0$ and $\mathbf{1}_K^\top \boldsymbol{b} = 0$.*

Proof Consider $\mathcal{L}(\boldsymbol{Z},\boldsymbol{b}) = N^{-1}\ell_{CE}(\boldsymbol{Z} + \boldsymbol{b}\mathbf{1}_N^\top, \boldsymbol{Y}) + \lambda_Z\|\boldsymbol{Z}\|_* + \lambda_b\|\boldsymbol{b}\|^2/2$ in (**UFM**). Then its gradient (subgradient) is

$$\frac{\partial L}{\partial \boldsymbol{Z}} = N^{-1}(\boldsymbol{P} - \boldsymbol{Y}) + \lambda \partial\|\boldsymbol{Z}\|_*, \quad \frac{\partial \varphi}{\partial \boldsymbol{b}} = N^{-1}(\boldsymbol{P} - \boldsymbol{Y})\mathbf{1}_N + \lambda_b \boldsymbol{b}.$$

Therefore, $(\boldsymbol{Z},\boldsymbol{b})$ is a global minimizer if

$$N^{-1}(\boldsymbol{Y} - \boldsymbol{P}) \in \lambda_Z \partial\|\boldsymbol{Z}\|_*, \quad N^{-1}(\boldsymbol{Y} - \boldsymbol{P})\mathbf{1}_N = \lambda_b \boldsymbol{b}$$

where $\partial\|\boldsymbol{Z}\|_*$ is the subdifferential of nuclear norm at $\boldsymbol{Z}$.

For any $\boldsymbol{Z}$ and $\boldsymbol{b}$, we notice that $\boldsymbol{Z}+\mathbf{1}_K\boldsymbol{v}^\top$ and $\boldsymbol{b}+\mu\mathbf{1}_K$ does not change $\varphi(\boldsymbol{Z},\boldsymbol{b})$ for any $\boldsymbol{v}\in\mathbb{R}^N$ and $\mu\in\mathbb{R}$, i.e.,

$$\varphi(\boldsymbol{Z}+\mathbf{1}_K\boldsymbol{v}^\top,\boldsymbol{b}+\mu\mathbf{1}_K)=\varphi(\boldsymbol{Z},\boldsymbol{b}).$$

Note that

$$\|\boldsymbol{b}\|^2\geq\min_{\mu}\|\boldsymbol{b}-\mu\mathbf{1}_K\|^2=\|\boldsymbol{b}-\mathbf{1}_K\langle\mathbf{1}_K,\boldsymbol{b}\rangle/K\|^2$$

and

$$\min_{\boldsymbol{v}\in\mathbb{R}^N}\|\boldsymbol{Z}-\mathbf{1}_K\boldsymbol{v}^\top\|_*=\|(\boldsymbol{I}_K-\boldsymbol{J}_K/K)\boldsymbol{Z}\|_*$$

where the subdifferential of $\|\boldsymbol{Z}-\mathbf{1}_K\boldsymbol{v}^\top\|_*$ is $(\partial\|\boldsymbol{Z}-\mathbf{1}_K\boldsymbol{v}^\top\|_*)^\top\mathbf{1}_K$ and

$$0\in(\partial\|\boldsymbol{Z}-\mathbf{1}_K\boldsymbol{v}^\top\|_*)^\top\mathbf{1}_K\Big|_{\boldsymbol{v}=\boldsymbol{Z}^\top\mathbf{1}_K/K}$$

since the column space of $(\boldsymbol{I}_K-\boldsymbol{J}_K/K)\boldsymbol{Z}$ is perpendicular to $\mathbf{1}_K$. Therefore, the global minimizer must satisfy $\mathbf{1}_K^\top\boldsymbol{Z}=0$ and $\mathbf{1}_K^\top\boldsymbol{b}=0$. ■

By considering the exact form of $\|\boldsymbol{Z}\|_*$, the optimality condition in Lemma 8 can be expressed explicitly as given by the next corollary.

Corollary 9. *Assume $\boldsymbol{z}_{ki}=\bar{\boldsymbol{z}}_k$ for $1\leq i\leq n_k, 1\leq k\leq K$ and $\langle\bar{\boldsymbol{z}}_k,\mathbf{1}_K\rangle=0$, and then it holds that*

$$\begin{aligned}N^{-1}(\boldsymbol{I}_K-\bar{\boldsymbol{P}})&=\lambda_Z\left(\left[\left(\bar{\boldsymbol{Z}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top\right)^\dagger\right]^{1/2}\bar{\boldsymbol{Z}}+\bar{\boldsymbol{R}}\right),\\ N^{-1}(\boldsymbol{I}_K-\bar{\boldsymbol{P}})\boldsymbol{n}&=\lambda_b\boldsymbol{b}\end{aligned}\tag{5.4}$$

where $\bar{\boldsymbol{Z}}=[\bar{\boldsymbol{z}}_1,\cdots,\bar{\boldsymbol{z}}_K]\in\mathbb{R}^{d\times K}$, $\bar{\boldsymbol{R}}$ satisfies $\bar{\boldsymbol{R}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top=0$, $\bar{\boldsymbol{Z}}^\top\bar{\boldsymbol{R}}=0$, $\|\bar{\boldsymbol{R}}\boldsymbol{D}^{1/2}\|\leq 1$, and $\boldsymbol{D}$ is defined in (3.3)*. In particular, if $\bar{\boldsymbol{Z}}$ is of rank $K-1$, then*

$$N^{-1}(\boldsymbol{I}_K-\bar{\boldsymbol{P}})=\lambda_Z\left[\left(\bar{\boldsymbol{Z}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top\right)^\dagger\right]^{1/2}\bar{\boldsymbol{Z}}.$$

Proof Under assumption, $\boldsymbol{z}_{ki}=\bar{\boldsymbol{z}}_k$, $1\leq i\leq n_k$ and $\bar{\boldsymbol{z}}_k^\top\mathbf{1}_K=0$, we have $\boldsymbol{Z}=\bar{\boldsymbol{Z}}\boldsymbol{Y}$ and $\boldsymbol{P}=\bar{\boldsymbol{P}}\boldsymbol{Y}$ where $\bar{\boldsymbol{Z}}$ and $\boldsymbol{Y}$ are defined in (3.2) and $\bar{\boldsymbol{P}}=[\bar{\boldsymbol{p}}_1,\cdots,\bar{\boldsymbol{p}}_K]$ with $\bar{\boldsymbol{p}}_k$ as the probability vector w.r.t. $\bar{\boldsymbol{z}}_k+\boldsymbol{b}$. Then (5.3) reduces to

$$N^{-1}(\boldsymbol{I}_K-\bar{\boldsymbol{P}})\boldsymbol{Y}\in\lambda_Z\partial\|\bar{\boldsymbol{Z}}\boldsymbol{Y}\|_*\tag{5.5}$$

Here we let $\boldsymbol{Z}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the SVD of $\boldsymbol{Z}$.

$$\partial\|\bar{\boldsymbol{Z}}\boldsymbol{Y}\|_*=\left\{\boldsymbol{U}\boldsymbol{V}^\top+\boldsymbol{R}:\|\boldsymbol{R}\|\leq 1,\ \boldsymbol{U}^\top\boldsymbol{R}=0,\ \boldsymbol{R}\boldsymbol{V}=0\right\}$$

where

$$\boldsymbol{U}\boldsymbol{V}^\top=\left[\left(\bar{\boldsymbol{Z}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top\right)^\dagger\right]^{1/2}\bar{\boldsymbol{Z}}\boldsymbol{Y}$$

and † denotes the Moore-Penrose pseudo-inverse. Then (5.5) becomes

$$N^{-1}(\boldsymbol{I}_K - \bar{\boldsymbol{P}})\boldsymbol{Y} = \lambda_Z \left(\left[\left(\bar{\boldsymbol{Z}} \boldsymbol{D} \bar{\boldsymbol{Z}}^\top \right)^\dagger \right]^{1/2} \bar{\boldsymbol{Z}} \boldsymbol{Y} + \boldsymbol{R} \right)$$

where $\boldsymbol{R}$ is in the form of

$$\boldsymbol{R} = \bar{\boldsymbol{R}} \boldsymbol{Y}, \quad \bar{\boldsymbol{R}} = [\bar{\boldsymbol{r}}_1, \cdots, \bar{\boldsymbol{r}}_K]$$

such that

$$\bar{\boldsymbol{R}} \boldsymbol{Y} \left(\bar{\boldsymbol{Z}} \boldsymbol{Y} \right)^\top = \bar{\boldsymbol{R}} \boldsymbol{D} \bar{\boldsymbol{Z}}^\top = 0, \quad \bar{\boldsymbol{Z}}^\top \bar{\boldsymbol{R}} = 0, \quad \| \bar{\boldsymbol{R}} \boldsymbol{D}^{1/2} \| \leq 1$$

where $\boldsymbol{D}$ is defined in (3.3) and $\boldsymbol{Y}\boldsymbol{Y}^\top = \boldsymbol{D}$. This leads to

$$N^{-1}(\boldsymbol{I}_K - \bar{\boldsymbol{P}}) = \lambda_Z \left(\left[\left(\bar{\boldsymbol{Z}} \boldsymbol{D} \bar{\boldsymbol{Z}}^\top \right)^\dagger \right]^{1/2} \bar{\boldsymbol{Z}} + \bar{\boldsymbol{R}} \right)$$

where $\bar{\boldsymbol{R}} \boldsymbol{D} \bar{\boldsymbol{Z}}^\top = 0$ and $\bar{\boldsymbol{Z}}^\top \bar{\boldsymbol{R}} = 0$.

In particular, if $\bar{\boldsymbol{Z}}$ is of rank $K-1$, then each column of $\bar{\boldsymbol{R}}$ is parallel to $\mathbf{1}_K$ since $\mathbf{1}_K^\top \bar{\boldsymbol{Z}} = 0$. Since $\bar{\boldsymbol{P}}$ is a positive left-stochastic matrix with $\bar{\boldsymbol{P}}^\top \mathbf{1}_K = \mathbf{1}_K$, and thus $\boldsymbol{I}_K - \bar{\boldsymbol{P}}$ is of rank $K-1$. Note that $\mathbf{1}_K$ is also in the left null space of $\boldsymbol{I}_K - \bar{\boldsymbol{P}}$, then $\bar{\boldsymbol{R}} = 0$. It means

$$N^{-1}(\boldsymbol{I}_K - \bar{\boldsymbol{P}}) = \lambda_Z \left[\left(\bar{\boldsymbol{Z}} \boldsymbol{D} \bar{\boldsymbol{Z}}^\top \right)^\dagger \right]^{1/2} \bar{\boldsymbol{Z}}.$$

For $\boldsymbol{b}$, it is straightforward to have

$$N^{-1}(\boldsymbol{Y} - \boldsymbol{P})\mathbf{1}_N = N^{-1} \sum_{k=1}^{K} n_k (\boldsymbol{e}_k - \bar{\boldsymbol{p}}_k) = N^{-1}(\boldsymbol{I}_K - \bar{\boldsymbol{P}})\boldsymbol{n} = \lambda_b \boldsymbol{b}.$$

■

Lemma 10 (Invariance under permutation).

(a) Suppose $\boldsymbol{X}$ *satisfies*

$$\boldsymbol{\Pi} \boldsymbol{X} \boldsymbol{\Pi}' = \boldsymbol{X}$$

for any permutation $\boldsymbol{\Pi}$ *and* $\boldsymbol{\Pi}'$*, then* $\boldsymbol{X} = c\boldsymbol{J}$ *for some* c*.*

(b) Suppose $\boldsymbol{X}$ *satisfies*

$$\boldsymbol{\Pi} \boldsymbol{X} \boldsymbol{\Pi}^\top = \boldsymbol{X},$$

then $\boldsymbol{X} = a\boldsymbol{I} + c\boldsymbol{J}$ *for some* a *and* c*.*

Proof For (a), we have

$$\boldsymbol{\Pi} \boldsymbol{X} = \boldsymbol{X}, \quad \boldsymbol{X} \boldsymbol{\Pi}' = \boldsymbol{X}$$

holds for any permutation matrix. Then $\boldsymbol{X}$ is invariant under either row or column permutation. Therefore, $\boldsymbol{X}$ is a constant matrix.

For (b), we first pick the permutation matrix $\mathbf{\Pi}_{\ell,\ell'}$ that only exchanges ℓ- and ℓ'-th entries. Then $\mathbf{\Pi}_{\ell,\ell'} = \mathbf{\Pi}_{\ell',\ell}$ and

$$\mathbf{\Pi}_{\ell,\ell'} \boldsymbol{X} \mathbf{\Pi}_{\ell,\ell'} = \boldsymbol{X}.$$

Now let $\boldsymbol{e}_k$ be any one-hot vector, and it holds

$$X_{\ell'\ell'} = \boldsymbol{e}_{\ell'}^\top \boldsymbol{X} \boldsymbol{e}_{\ell'} = \boldsymbol{e}_\ell^\top \mathbf{\Pi}_{\ell,\ell'} \boldsymbol{X} \mathbf{\Pi}_{\ell,\ell'} \boldsymbol{e}_\ell = \boldsymbol{e}_\ell^\top \boldsymbol{X} \boldsymbol{e}_\ell = X_{\ell\ell},$$

where $\mathbf{\Pi}_{\ell,\ell'} \boldsymbol{e}_{\ell'} = \boldsymbol{e}_\ell$. This implies that the diagonal entries of $\boldsymbol{X}$ are the same.

For any $k \neq \ell$ or ℓ', $\mathbf{\Pi}_{\ell,\ell'} \boldsymbol{e}_k = \boldsymbol{e}_k$ holds and

$$X_{k\ell} = \boldsymbol{e}_k^\top \boldsymbol{X} \boldsymbol{e}_\ell = \boldsymbol{e}_k^\top \mathbf{\Pi}_{\ell,\ell'} \boldsymbol{X} \mathbf{\Pi}_{\ell,\ell'} \boldsymbol{e}_\ell = \boldsymbol{e}_k^\top \boldsymbol{X} \boldsymbol{e}_{\ell'} = X_{k\ell'}.$$

Similarly, it also holds $X_{\ell k} = X_{\ell' k}$. Therefore, $X_{\ell\ell'}$ is the same for $\ell \neq \ell'$. ■

5.2 Proof of Theorem 2

Theorem 2(a) The proof follows from a convenient *permutation argument*. Recall (**UFM**) equals

$$\mathcal{L}(\boldsymbol{Z}, \boldsymbol{b}) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \ell_{CE}(\boldsymbol{z}_{ki} + \boldsymbol{b}, \boldsymbol{e}_k) + \lambda_Z \|\boldsymbol{Z}\|_* + \frac{\lambda_b}{2} \|\boldsymbol{b}\|^2 \tag{5.6}$$

Note that it is strongly convex in $(\boldsymbol{Z}, \boldsymbol{b})$ in the restricted direction, as shown in Lemma 7. Then the sublevel set of $\{(\boldsymbol{Z}, \boldsymbol{b}) : \mathcal{L}(\boldsymbol{Z}, \boldsymbol{b}) \leq c\}$ is a compact set when restricted on those directions, for any c if $\lambda_Z > 0$ and $\lambda_b > 0$. Therefore, the existence of a global minimizer is guaranteed. Suppose $(\boldsymbol{Z}, \boldsymbol{b})$ is a global minimizer to $\mathcal{L}(\boldsymbol{Z}, \boldsymbol{b})$ and we know from Lemma 8 that $\mathbf{1}_K^\top \boldsymbol{Z} = 0$ and $\mathbf{1}_K^\top \boldsymbol{b} = 0$. Then by using Jensen's inequality, we have

$$\frac{1}{n_k} \sum_{i=1}^{n_k} \ell_{CE}(\boldsymbol{z}_{ki} + \boldsymbol{b}, \boldsymbol{e}_k) \geq \ell_{CE}\left(\frac{1}{n_k} \sum_{i=1}^{n_k} (\boldsymbol{z}_{ki} + \boldsymbol{b}), \boldsymbol{e}_k \right) = \ell_{CE}(\bar{\boldsymbol{z}}_k + \boldsymbol{b}, \boldsymbol{e}_k)$$

where $\{\boldsymbol{z}_{ki}\}_{i=1}^{n_k}$ belong to the same class and $\bar{\boldsymbol{z}}_k = n_k^{-1} \sum_{i=1}^{n_k} \boldsymbol{z}_{ki}$.

Let $\operatorname{blkdiag}(\mathbf{\Pi}_{n_1}, \cdots, \mathbf{\Pi}_{n_k})$ be a block-diagonal permutation matrix where each $\mathbf{\Pi}_{n_k}$ is any $n_k \times n_k$ permutation. Then

$$\|\boldsymbol{Z} \operatorname{blkdiag}(\mathbf{\Pi}_{n_1}, \cdots, \mathbf{\Pi}_{n_k})\|_* = \|\boldsymbol{Z}\|_*.$$

Note that

$$\frac{1}{n_1! \cdots n_K!} \sum_{\{\mathbf{\Pi}_k, 1 \leq k \leq K\}} \operatorname{blkdiag}(\mathbf{\Pi}_{n_1}, \cdots, \mathbf{\Pi}_{n_k}) = \operatorname{blkdiag}(\boldsymbol{J}_{n_1}/n_1, \cdots, \boldsymbol{J}_{n_K}/n_K)$$

where the number of distinct permutation matrices in the form of $\operatorname{blkdiag}(\mathbf{\Pi}_{n_1}, \cdots, \mathbf{\Pi}_{n_k})$ is $n_1! \cdots n_K!$ and the summation is taken over all possible permutations in that form. Using Jensen's inequality again results in

$$\|\boldsymbol{Z}\|_* \geq \|\boldsymbol{Z} \operatorname{blkdiag}(\boldsymbol{J}_{n_1}/n_1, \cdots, \boldsymbol{J}_{n_K}/n_K)\|_* = \|[\bar{\boldsymbol{z}}_1 \mathbf{1}_{n_1}^\top, \cdots, \bar{\boldsymbol{z}}_K \mathbf{1}_{n_K}^\top]\|_* = \|\bar{\boldsymbol{Z}} \boldsymbol{Y}\|_* \tag{5.7}$$

As a result, it holds that

$$\begin{aligned}\mathcal{L}(\boldsymbol{Z},\boldsymbol{b}) &= \frac{1}{N}\sum_{k=1}^{K} n_k \left(\frac{1}{n_k}\sum_{i=1}^{n_k} \ell_{CE}(\boldsymbol{z}_{ki}+\boldsymbol{b},\boldsymbol{e}_k)\right) + \lambda_Z\|\boldsymbol{Z}\|_* + \frac{\lambda_b}{2}\|\boldsymbol{b}\|^2 \\ &\geq \sum_{k=1}^{K}\frac{n_k}{N}\ell_{CE}(\bar{\boldsymbol{z}}_k+\boldsymbol{b},\boldsymbol{e}_k) + \lambda_Z\|[\bar{\boldsymbol{z}}_1\mathbf{1}_{n_1}^\top,\cdots,\bar{\boldsymbol{z}}_K\mathbf{1}_{n_K}^\top]\|_* + \frac{\lambda_b}{2}\|\boldsymbol{b}\|^2 \\ &= \frac{1}{N}\ell_{CE}(\bar{\boldsymbol{Z}}\boldsymbol{Y}+\boldsymbol{b}\mathbf{1}_N^\top,\boldsymbol{Y}) + \lambda_Z\|\bar{\boldsymbol{Z}}\boldsymbol{Y}\|_* + \frac{\lambda_b}{2}\|\boldsymbol{b}\|^2 = R(\bar{\boldsymbol{Z}}\boldsymbol{Y},\boldsymbol{b}).\end{aligned}$$

As Lemma 7 guarantees the strong convexity of $\mathcal{L}(\boldsymbol{Z},\boldsymbol{b})$ restricted on $\{(\boldsymbol{\Delta}_Z,\boldsymbol{\Delta}_b):\mathbf{1}_K^\top(\boldsymbol{\Delta}_Z+\boldsymbol{\Delta}_b\mathbf{1}_N^\top)=0\}$, the global optimality of $(\boldsymbol{Z},\boldsymbol{b})$ implies that $\boldsymbol{Z}=\bar{\boldsymbol{Z}}\boldsymbol{Y}$, i.e., $\boldsymbol{z}_{ki}=\bar{\boldsymbol{z}}_k$ for $1\leq i\leq n_k$.

Theorem 2(b) Without loss of generality, we let $n_1\leq n_2\leq\cdots\leq n_K$ and $\Gamma_j=\{k:n_k=N_j,\ 1\leq k\leq K\}$ where $\{N_j\}_{j=1}^J$ are the distinct values of $\{n_k\}_{k=1}^K$. In other words, we re-group each class according to their individual class sizes and it holds $K=\sum_{j=1}^J|\Gamma_j|$. As Theorem 2(a) implies the global minimizer $\boldsymbol{Z}$ satisfies $\boldsymbol{z}_{ki}=\bar{\boldsymbol{z}}_k$, $1\leq i\leq n_k$. Then it holds

$$\begin{aligned}\mathcal{L}(\boldsymbol{Z},\boldsymbol{b}) &= \frac{1}{N}\sum_{k=1}^{K} n_k\ell_{CE}(\bar{\boldsymbol{z}}_k+\boldsymbol{b},\boldsymbol{e}_k) + \lambda_Z\|\bar{\boldsymbol{Z}}\boldsymbol{Y}\|_* + \frac{\lambda_b}{2}\|\boldsymbol{b}\|^2 \\ &= \frac{1}{N}\sum_{j=1}^{J} N_j\sum_{k\in\Gamma_j}\ell_{CE}(\bar{\boldsymbol{z}}_k+\boldsymbol{b},\boldsymbol{e}_k) + \lambda_Z\|\bar{\boldsymbol{Z}}\boldsymbol{D}^{1/2}\|_* + \frac{\lambda_b}{2}\|\boldsymbol{b}\|^2\end{aligned}$$

where $\boldsymbol{D}=\boldsymbol{Y}\boldsymbol{Y}^\top$ and

$$\bar{\boldsymbol{Z}}=[\boldsymbol{Z}_{\Gamma_1},\cdots,\boldsymbol{Z}_{\Gamma_J}],\quad \boldsymbol{Z}_{\Gamma_j}=[\bar{\boldsymbol{z}}_k]_{k\in\Gamma_j}.$$

Let $\boldsymbol{\Pi}_{\ell,\ell'}$ be a $K\times K$ permutation matrix that only switches the index ℓ and ℓ' but keeps the other indices unchanged. Then we claim that

$$\mathcal{L}(\boldsymbol{\Pi}_{\ell,\ell'}\bar{\boldsymbol{Z}}\boldsymbol{\Pi}_{\ell,\ell'}\boldsymbol{Y},\boldsymbol{\Pi}_{\ell,\ell'}\boldsymbol{b}) = \mathcal{L}(\bar{\boldsymbol{Z}}\boldsymbol{Y},\boldsymbol{b})$$

as long as index ℓ and ℓ' are in the same Γ_j for some j. By using the restricted strong convexity of $\mathcal{L}(\boldsymbol{Z},\boldsymbol{b})$ in Lemma 7, we have

$$\boldsymbol{\Pi}_{\ell,\ell'}\bar{\boldsymbol{Z}}\boldsymbol{\Pi}_{\ell,\ell'}=\bar{\boldsymbol{Z}},\quad \boldsymbol{\Pi}_{\ell,\ell'}\boldsymbol{b}=\boldsymbol{b},\ \ \forall\ell,\ell'\in\Gamma_j,\ \ \text{for some } j. \tag{5.8}$$

Now we assume the claim holds, and see how it leads to the desired result. For any $j\neq j'$, it holds that $\boldsymbol{\Pi}_{\Gamma_j}^\top\bar{\boldsymbol{Z}}_{j,j'}\boldsymbol{\Pi}_{\Gamma_{j'}}=\bar{\boldsymbol{Z}}_{j,j'}$ where $\boldsymbol{\Pi}_{\Gamma_j}$ is any $k_j\times k_j$ permutation matrix. Lemma 10 implies every entry in $\bar{\boldsymbol{Z}}_{j,j'}$ equals some constant $a_{jj'}$. For the diagonal blocks, we have $\boldsymbol{\Pi}_{\Gamma_j}^\top\bar{\boldsymbol{Z}}_{j,j}\boldsymbol{\Pi}_{\Gamma_j}=\bar{\boldsymbol{Z}}_{j,j}$. Therefore, $\bar{\boldsymbol{Z}}_{j,j}=a_j\boldsymbol{I}_{\Gamma_j}+a_{jj'}\mathbf{1}_{\Gamma_j}\mathbf{1}_{\Gamma_j}^\top$ follows from Lemma 10. Now we conclude that

$$\bar{\boldsymbol{Z}}=\sum_{j=1}^{J}a_j\boldsymbol{I}_{\Gamma_j}+\sum_{1\leq j,j'\leq J}a_{jj'}\mathbf{1}_{\Gamma_j}\mathbf{1}_{\Gamma_{j'}}^\top,\quad a_j+\sum_{j'=1}^{J}a_{j'j}|\Gamma_{j'}|=0 \tag{5.9}$$

where $\boldsymbol{I}_{\Gamma_j} = \mathrm{diag}(\mathbf{1}_{\Gamma_j})$, which follows from Lemma 10.

From $\boldsymbol{\Pi}_{\ell,\ell'}\boldsymbol{b} = \boldsymbol{b}$ for any $\ell, \ell' \in \Gamma_j$, then $\boldsymbol{b}$ is in the form of

$$\boldsymbol{b} = \sum_{j=1}^{J} c_j \mathbf{1}_{\Gamma_j}, \quad \sum_{j=1}^{J} c_j |\Gamma_j| = 0.$$

To complete the proof, it remains to justify the claim. For any $k \neq \ell$ or ℓ', then

$$\begin{aligned}\ell_{CE}([\boldsymbol{\Pi}_{\ell,\ell'}(\bar{\boldsymbol{Z}} + \boldsymbol{b}\mathbf{1}_K^\top)\boldsymbol{\Pi}_{\ell,\ell'}]_k, \boldsymbol{e}_k) &= \ell_{CE}(\boldsymbol{\Pi}_{\ell,\ell'}(\bar{\boldsymbol{Z}} + \boldsymbol{b}\mathbf{1}_K^\top)\boldsymbol{\Pi}_{\ell,\ell'}\boldsymbol{e}_k, \boldsymbol{e}_k) \\ &= \ell_{CE}(\boldsymbol{\Pi}_{\ell,\ell'}(\bar{\boldsymbol{z}}_k + \boldsymbol{b}), \boldsymbol{e}_k) = \ell_{CE}(\bar{\boldsymbol{z}}_k + \boldsymbol{b}, \boldsymbol{e}_k)\end{aligned}$$

where $\boldsymbol{\Pi}_{\ell,\ell'}\boldsymbol{e}_k = \boldsymbol{e}_k$ if $k \neq \ell$ or ℓ'. In addition, it holds

$$\begin{aligned}\sum_{k\in\{\ell,\ell'\}} \ell_{CE}([\boldsymbol{\Pi}_{\ell,\ell'}(\bar{\boldsymbol{Z}} + \boldsymbol{b}\mathbf{1}_K^\top)\boldsymbol{\Pi}_{\ell,\ell'}]_k, \boldsymbol{e}_k) &= \ell_{CE}(\boldsymbol{\Pi}_{\ell,\ell'}(\bar{\boldsymbol{z}}_{\ell'} + \boldsymbol{b}), \boldsymbol{e}_\ell) + \ell_{CE}(\boldsymbol{\Pi}_{\ell,\ell'}(\bar{\boldsymbol{z}}_\ell + \boldsymbol{b}), \boldsymbol{e}_{\ell'}) \\ &= \ell_{CE}(\bar{\boldsymbol{z}}_{\ell'} + \boldsymbol{b}, \boldsymbol{e}_{\ell'}) + \ell_{CE}(\bar{\boldsymbol{z}}_\ell + \boldsymbol{b}, \boldsymbol{e}_\ell)\end{aligned}$$

where $\boldsymbol{\Pi}_{\ell,\ell'}\boldsymbol{e}_\ell = \boldsymbol{e}_{\ell'}$. Also, we note that $\boldsymbol{\Pi}_{\ell,\ell'}\boldsymbol{Y} = \boldsymbol{Y}$ for ℓ and ℓ' in the same cluster, and also the nuclear norm is invariant under orthogonal transform. Hence,

$$\|\boldsymbol{\Pi}_{\ell,\ell'}\bar{\boldsymbol{Z}}\boldsymbol{\Pi}_{\ell,\ell'}\boldsymbol{Y}\|_* = \|\bar{\boldsymbol{Z}}\boldsymbol{Y}\|_*, \quad \|\boldsymbol{\Pi}_{\ell,\ell'}\boldsymbol{b}\| = \|\boldsymbol{b}\|,$$

and we have proven the claim. In other words, we have

$$\mathrm{blkdiag}(\boldsymbol{\Pi}_{\Gamma_1}, \cdots, \boldsymbol{\Pi}_{\Gamma_J})^\top \bar{\boldsymbol{Z}}\, \mathrm{blkdiag}(\boldsymbol{\Pi}_{\Gamma_1}, \cdots, \boldsymbol{\Pi}_{\Gamma_J}) = \bar{\boldsymbol{Z}}$$

where $\boldsymbol{\Pi}_{\Gamma_j}$ is any $|\Gamma_j| \times |\Gamma_j|$ permutation matrix that acts on the index set Γ_j.

Theorem 2(c) In the balanced case, $\bar{\boldsymbol{Z}} = a(K\boldsymbol{I}_K - \boldsymbol{J}_K)$ holds and $\boldsymbol{b} = 0$.

$$\bar{\boldsymbol{P}} = \frac{e^{-a}\boldsymbol{J}_K + e^{-a}(e^{aK} - 1)\boldsymbol{I}_K}{e^{-a}(K - 1 + e^{aK})} = \frac{\boldsymbol{J}_K + (e^{aK} - 1)\boldsymbol{I}_K}{K - 1 + e^{aK}}$$

and

$$\boldsymbol{I}_K - \bar{\boldsymbol{P}} = \boldsymbol{I}_K - \frac{\boldsymbol{J}_K + (e^{aK} - 1)\boldsymbol{I}_K}{K - 1 + e^{aK}} = \frac{K\boldsymbol{I}_K - \boldsymbol{J}_K}{K - 1 + e^{aK}}.$$

Then the optimality condition (5.4) indicates

$$\frac{K\boldsymbol{I}_K - \boldsymbol{J}_K}{K - 1 + e^{aK}} \in N\lambda_Z \partial \|\bar{\boldsymbol{Z}}\|_*.$$

Note that the operator norm of the left-hand side is $K/(K-1+e^{aK})$. If $N\lambda_Z \geq K/(K-1)$, then $a = 0$ and $\bar{\boldsymbol{Z}} = 0$. If $N\lambda_Z < K/(K-1)$, then $a \neq 0$ and it satisfies

$$\frac{K\boldsymbol{I}_K - \boldsymbol{J}_K}{K - 1 + e^{aK}} = N\lambda_Z(\boldsymbol{I}_K - \boldsymbol{J}_K/K) \Longrightarrow \frac{K}{K - 1 + e^{aK}} = N\lambda_Z \Longrightarrow a = \frac{1}{K}\log\left(\frac{K}{N\lambda_Z} - K + 1\right).$$

Theorem 2(d) The proof directly follows from (2.3) of (Recht et al., 2010, Lemma 5.1). More precisely, we write

$$\boldsymbol{Z} = \bar{\boldsymbol{Z}}\boldsymbol{Y} = \bar{\boldsymbol{Z}}\boldsymbol{D}^{1/2}\boldsymbol{D}^{-1/2}\boldsymbol{Y}$$

where $\boldsymbol{D}^{-1/2}\boldsymbol{Y}$ has orthogonal rows. By performing the SVD on $\bar{\boldsymbol{Z}}\boldsymbol{D}^{1/2} = \bar{\boldsymbol{U}}\bar{\boldsymbol{\Sigma}}\bar{\boldsymbol{V}}^\top$, we have

$$\boldsymbol{Z} = \bar{\boldsymbol{U}}\bar{\boldsymbol{\Sigma}}\bar{\boldsymbol{V}}^\top\boldsymbol{D}^{-1/2}\boldsymbol{Y} = \boldsymbol{W}^\top\boldsymbol{H}$$

Then (2.3) implies $\boldsymbol{W} = \bar{\boldsymbol{\Sigma}}^{1/2}\bar{\boldsymbol{U}}^\top$, $\boldsymbol{H} = \bar{\boldsymbol{\Sigma}}^{1/2}\bar{\boldsymbol{V}}^\top\boldsymbol{D}^{-1/2}\boldsymbol{Y}$, and $\bar{\boldsymbol{H}} = \bar{\boldsymbol{\Sigma}}^{1/2}\bar{\boldsymbol{V}}^\top\boldsymbol{D}^{-1/2}$.

5.3 Optimality condition and the solution structure

In this section, we focus on a special case when the dataset contains k_A classes with n_A points in each class and k_B classes with n_B points. The total number of classes is $K = k_A + k_B$ and number of points is $N = k_An_A + k_Bn_B$. In particular, we assume $k_A \geq 2$ and $k_B \geq 2$. To characterize the solution, it suffices to look into the optimality condition (5.4) in Corollary 9 which involves $\boldsymbol{I}_K - \bar{\boldsymbol{P}}$ and $(\bar{\boldsymbol{Z}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top)^{\dagger/2}\bar{\boldsymbol{Z}}$. We first provide the explicit expression for both of them and in fact, they are in the form of block-structure $\mathcal{B}$ in (3.4).

We adopt the notation (3.5) for $\bar{\boldsymbol{Z}}$ and $\boldsymbol{b}$, and further write $\bar{\boldsymbol{Z}}$ in the following form,

$$\bar{\boldsymbol{Z}} = \begin{bmatrix}(k_Aa + k_Bc)\boldsymbol{C}_{k_A} & 0\\ 0 & (k_Ab + k_Bd)\boldsymbol{C}_{k_B}\end{bmatrix} + k_Ak_B\boldsymbol{s}\boldsymbol{s}^\top\begin{bmatrix}c\boldsymbol{I}_{k_A} & 0\\ 0 & b\boldsymbol{I}_{k_B}\end{bmatrix} \in \mathbb{R}^{K\times K}$$

where $\boldsymbol{C}_{k_A}$ (and $\boldsymbol{C}_{k_B}$) is defined in (1.1) and

$$\boldsymbol{s} := \begin{bmatrix}\frac{\boldsymbol{1}_{k_A}}{k_A}\\ -\frac{\boldsymbol{1}_{k_B}}{k_B}\end{bmatrix}, \quad \|\boldsymbol{s}\| = \sqrt{\frac{1}{k_A} + \frac{1}{k_B}}, \quad \boldsymbol{D} = \begin{bmatrix}n_A\boldsymbol{I}_{k_A} & 0\\ 0 & n_B\boldsymbol{I}_{k_B}\end{bmatrix} \tag{5.10}$$

A direct computation gives

$$\bar{\boldsymbol{Z}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top = \begin{bmatrix}n_A(k_Aa + k_Bc)^2\boldsymbol{C}_{k_A} & 0\\ 0 & n_B(k_Bd + k_Ab)^2\boldsymbol{C}_{k_B}\end{bmatrix} + (k_An_Bb^2 + k_Bn_Ac^2)k_Ak_B\boldsymbol{s}\boldsymbol{s}^\top \tag{5.11}$$

and

$$\begin{aligned}\left[(\bar{\boldsymbol{Z}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top)^\dagger\right]^{\frac{1}{2}}\bar{\boldsymbol{Z}} = &\begin{bmatrix}\frac{\operatorname{sign}(k_Aa+k_Bc)}{\sqrt{n_A}}\boldsymbol{C}_{k_A} & 0\\ 0 & \frac{\operatorname{sign}(k_Ab+k_Bd)}{\sqrt{n_B}}\boldsymbol{C}_{k_B}\end{bmatrix}\\ &+ \frac{\operatorname{sign}(k_An_Bb^2 + k_Bn_Ac^2)k_Ak_B}{\sqrt{(k_An_Bb^2 + k_Bn_Ac^2)(k_A + k_B)}}\cdot\boldsymbol{s}\boldsymbol{s}^\top\begin{bmatrix}c\boldsymbol{I}_{k_A} & 0\\ 0 & b\boldsymbol{I}_{k_B}\end{bmatrix}\end{aligned}$$

is $\mathcal{B}(a_Z, b_Z, c_Z, d_Z)$ which satisfies

$$\begin{aligned}&k_Aa_Z + k_Bc_Z = n_A^{-1/2}\operatorname{sign}(k_Aa + k_Bc), \quad k_Ab_Z + k_Bd_Z = n_B^{-1/2}\operatorname{sign}(k_Ab + k_Bd),\\ &b_Z = \frac{b\operatorname{sign}(k_An_Bb^2 + k_Bn_Ac^2)}{\sqrt{(k_An_Bb^2 + k_Bn_Ac^2)(k_A + k_B)}}, \quad c_Z = \frac{c\operatorname{sign}(k_An_Bb^2 + k_Bn_Ac^2)}{\sqrt{(k_An_Bb^2 + k_Bn_Ac^2)(k_A + k_B)}}.\end{aligned} \tag{5.12}$$

In particular, if $a = b = c = d = 1/K$, then

$$\bar{\boldsymbol{Z}} = \boldsymbol{I}_K - \boldsymbol{J}_K/K$$

and

$$\bar{\boldsymbol{Z}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top = \begin{bmatrix} n_A\boldsymbol{C}_{k_A} & 0 \\ 0 & n_B\boldsymbol{C}_{k_B} \end{bmatrix} + \frac{n_Ak_B + k_An_B}{k_A+k_B}\cdot\left(\frac{k_Ak_B}{k_A+k_B}\right)\boldsymbol{s}\boldsymbol{s}^\top. \tag{5.13}$$

The eigenvalues are n_A with multiplicity k_A-1, n_B with multiplicity k_B-1, $(n_Ak_B + k_An_B)/K$ with multiplicity 1, and 0 with multiplicity 1.

For $\boldsymbol{I}_K - \bar{\boldsymbol{P}}$, direct computation implies that $\mathcal{B}(a_P, b_P, c_P, d_P)$, i.e.,

$$\boldsymbol{I}_K - \bar{\boldsymbol{P}} = \begin{bmatrix} a_Pk_A\boldsymbol{C}_{k_A} + c_Pk_B\boldsymbol{I}_{k_A} & -b_P\boldsymbol{J}_{k_A\times k_B} \\ -c_P\boldsymbol{J}_{k_B\times k_A} & d_Pk_B\boldsymbol{C}_{k_B} + b_Pk_A\boldsymbol{I}_{k_B} \end{bmatrix} \tag{5.14}$$

where

$$\begin{aligned} a_P &= \frac{e^{-a+mk_B}}{k_Be^{-c-mk_A} + e^{-a+mk_B}(k_A - 1 + e^{k_Aa+k_Bc})}, \\ b_P &= \frac{e^{-b+mk_B}}{k_Ae^{-b+mk_B} + e^{-d-mk_A}(k_B - 1 + e^{k_Ab+k_Bd})}, \\ c_P &= \frac{e^{-c-mk_A}}{k_Be^{-c-mk_A} + e^{-a+mk_B}(k_A - 1 + e^{k_Aa+k_Bc})}, \\ d_P &= \frac{e^{-d-mk_A}}{k_Ae^{-b+mk_B} + e^{-d-mk_A}(k_B - 1 + e^{k_Ab+k_Bd})}. \end{aligned} \tag{5.15}$$

To characterize the solution structure w.r.t. λ_Z and λ_b, we look into the optimality condition in Corollary 9:

$$N^{-1}(\boldsymbol{I}_K - \bar{\boldsymbol{P}}) = \lambda_Z\left(\left[(\bar{\boldsymbol{Z}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top)^\dagger\right]^{\frac{1}{2}}\bar{\boldsymbol{Z}} + \bar{\boldsymbol{R}}\right)$$

where $\bar{\boldsymbol{R}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top = 0$, $\bar{\boldsymbol{Z}}^\top\bar{\boldsymbol{R}} = 0$, and $\|\bar{\boldsymbol{R}}\boldsymbol{D}^{1/2}\| \le 1$. The key idea of the proof is straightforward: for different regimes of λ_Z, we will explicitly construct $\bar{\boldsymbol{R}}$ and show that a solution exists for the nonlinear equation system above. Then by restricted strong convexity in Lemma 7, we know that this solution must be a unique global minimizer to (**UFM**).

Now we will present the main result on how the solution structure changes w.r.t. the varying λ_Z. To characterize this, we first introduce a few functions that will be used later. Let

$$\begin{aligned} f_2(t, \lambda_Z, \lambda_b) &:= g_2(x_2(t)) - (k_A+k_B)k_Am(t, \lambda_Z, \lambda_b) \\ &= \log\left[\left(\frac{\sqrt{n_A}}{N\lambda_Z} - 1\right)(k_A + k_Bx_2(t)) + 1\right] - k_A\log x_2(t) - (k_A+k_B)k_Am(t, \lambda_Z, \lambda_b) \end{aligned} \tag{5.16}$$

where g_2 is defined in (5.23), and

$$m(t, \lambda_Z, \lambda_b) = \frac{\lambda_Z}{\lambda_b}\frac{n_A - n_Bt}{\sqrt{(k_Bn_A + k_An_Bt^2)(k_A+k_B)}}, \tag{5.17}$$

$$x_2(t) = \frac{k_A}{\sqrt{(k_B + k_An_Bt^2/n_A)(k_A+k_B)} - k_B}. \tag{5.18}$$

The main result and proof rely on a key quantity:

$$t^*(\lambda_Z, \lambda_b) = \operatorname{argmin}_{t\geq 0} |f_2(t, \lambda_Z, \lambda_b)| = \begin{cases} \text{the root of } f_2(t, \lambda_Z, \lambda_b), & \eta(\lambda_Z) < 0, \\ 0, & \eta(\lambda_Z) \geq 0, \end{cases} \tag{5.19}$$

where

$$\eta(\lambda_Z) := f_2(0, \lambda_Z, \lambda_b), \quad \lambda_Z \in (\sqrt{n_B}/N, \sqrt{n_A}/N). \tag{5.20}$$

The function $\eta(\cdot)$ is decreasing and in particular $\eta(\sqrt{n_A}/N) < 0$.

Later, Lemma 13 will prove $f_2(t, \lambda_Z, \lambda_b)$ and $m(t, \lambda_Z, \lambda_b)$ are strictly increasing and decreasing respectively. Therefore, if $f_2(0, \lambda_Z, \lambda_b) \leq 0$, then f_2 must have a unique root and it is also equal to $t^*(\lambda_Z, \lambda_b)$; otherwise, $f_2(t, \lambda_Z, \lambda_b)$ stays positive for all $t \geq 0$ and $t = 0$ is the global minimizer of f_2 in t. Moreover, we will see that $t^*(\lambda_Z, \lambda_b)$ is increasing in λ_Z when $\lambda_b > 0$ is fixed.

For simplicity, we denote $m(t, \lambda_Z, \lambda_b)$ and $t^*(\lambda_Z, \lambda_b)$ by $m(t)$ and $t^*(\lambda_Z)$ respectively. Now we are ready to present the main result which implies Theorem 3 immediately.

Proposition 11. *Suppose $n_A > n_B$. For λ_Z of different regimes, the optimality condition satisfies the following properties:*

(a) If $N\lambda_Z \leq \sqrt{n_B}$, there is a unique solution (a, b, c, d) to

$$N^{-1}(\boldsymbol{I}_K - \bar{\boldsymbol{P}}) = \lambda_Z \left[(\bar{\boldsymbol{Z}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top)^\dagger\right]^{\frac{1}{2}} \bar{\boldsymbol{Z}}$$

that satisfies $a, b, c, d > 0$ and $a - c + d - b \leq 0$.

(b) If $\sqrt{n_B} < N\lambda_Z < \sqrt{n_A}$ and

$$k_B e^{-(k_A+k_B)m(t^*(\lambda_Z,\lambda_b))} < \frac{1}{N\lambda_Z}\sqrt{\left(\frac{k_B n_A}{t^*(\lambda_Z,\lambda_b)^2} + k_A n_B\right)(k_A + k_B)} - k_A, \tag{5.21}$$

then there is a unique solution (a, b, c, d) to

$$N^{-1}(\boldsymbol{I}_K - \bar{\boldsymbol{P}}) = \lambda_Z \left[(\bar{\boldsymbol{Z}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top)^\dagger\right]^{\frac{1}{2}} \bar{\boldsymbol{Z}} + \bar{\boldsymbol{R}}$$

where

$$\bar{\boldsymbol{R}} = \frac{1}{\sqrt{n_B}} \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{I}_{k_B} - \boldsymbol{J}_{k_B}/k_B \end{bmatrix}.$$

Moreover, (a, b, c, d) satisfies $k_A b + k_B d = 0$ and $a, b, c > 0$.

(c) If $\sqrt{n_B} < N\lambda_Z < \sqrt{n_A}$ and

$$k_B e^{-(k_A+k_B)m(t^*(\lambda_Z,\lambda_b))} > \frac{1}{N\lambda_Z}\sqrt{\left(\frac{k_B n_A}{t^*(\lambda_Z,\lambda_b)^2} + k_A n_B\right)(k_A + k_B)} - k_A, \tag{5.22}$$

then there is a unique solution (a, b, c, d, t, τ) to

$$N^{-1}(\boldsymbol{I}_K - \bar{\boldsymbol{P}}) = \lambda_Z \left[(\bar{\boldsymbol{Z}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top)^\dagger\right]^{\frac{1}{2}} \bar{\boldsymbol{Z}} + \bar{\boldsymbol{R}}$$

where

$$\bar{\boldsymbol{R}} = \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{N\lambda_Z}(\boldsymbol{I}_{k_B} - \boldsymbol{J}_{k_B}/k_B) \end{bmatrix} + \frac{k_A k_B \tau}{\sqrt{(k_B n_A + k_A n_B t^2)(k_A + k_B)}} \boldsymbol{s}\boldsymbol{s}^\top \begin{bmatrix} \boldsymbol{I}_{k_A} & 0 \\ 0 & t\boldsymbol{I}_{k_B} \end{bmatrix}$$

and $\boldsymbol{s}$ *is defined in* (5.10)*. Moreover,* (a, b, c, d) *satisfies* $b = c = d = 0$ *and* $a > 0$*.*

(d) If $N\lambda_Z > \max\{\sqrt{n_A}, \sqrt{n_B}\}$*, then* $\boldsymbol{Z} = 0$*.*

(e) If $\lambda_b = \infty$*, i.e., the bias-free scenario, then* (5.21) *and* (5.22) *are equivalent to* $\lambda_Z < \lambda^*$ *and* $\lambda_Z > \lambda^*$ *respectively for some* $\lambda^* \in (\sqrt{n_B}/N, \sqrt{n_A}/N)$*, where* λ^* *satisfies*

$$t^*(\lambda^*) = \sqrt{\frac{k_B n_A}{(N\lambda_Z)^2(k_A + k_B) - k_A n_B}}.$$

The intuition behind the change of solution structure $\bar{\boldsymbol{Z}}$ is essentially the singular value thresholding. By the 2×2 block structure of $\bar{\boldsymbol{Z}}$, $\bar{\boldsymbol{Z}}$ only has at most 3 different nonzero singular values regardless of the multiplicity. As the nuclear norm penalty parameter λ_Z grows, these singular values will gradually diminish to 0 until all the singular values become 0 after a certain finite threshold. The proof relies on the following two important lemmas.

Lemma 12. *Suppose* f_1 *and* f_2 *are continuous, and strictly decreasing and increasing respectively on* $t > 0$*. In addition, they satisfy*

$$\lim_{t\to 0} f_1(t) = +\infty, \quad \lim_{t\to\infty} f_2(t) = +\infty$$

and $I^+ = \{t : f_1 > 0, f_2 > 0\}$ *is nonempty, then*

$$f(t) = \frac{f_1(t)}{f_2(t)}$$

is decreasing on I^+ *and* $f(t) = t$ *has a unique solution.*

Proof Let $t_1 = \sup\{t : f_1(t) > 0\}$ and $t_2 = \inf\{t : f_2(t) > 0\}$. Since I^+ is nonempty, we have $t_2 < t_1$.

- if $t_1 < +\infty$, then $f_1(t_1) = 0$ and $\lim_{t\to t_1} f_2(t)$ exists and is positive, and thus $\lim_{t\to t_1^-} f(t) = 0$.
- if $t_1 = +\infty$, then $\lim_{t\to\infty} f_1(t)$ exists and is positive and $\lim_{t\to\infty} f_2(t) = \infty$, and thus $\lim_{t\to t_1^-} f(t) = 0$.
- if $t_2 = 0$, then $\lim_{t\to 0} f_2(t)$ exists and is positive and $\lim_{t\to 0} f_1(t) = \infty$, and thus $\lim_{t\to t_2^+} f(t) = \infty$.
- if $t_2 > 0$, then $f_2(t_2) = 0$ and $\lim_{t\to t_2} f_1(t)$ exists and is positive, and thus $\lim_{t\to t_2^+} f(t) = \infty$.

This implies that

$$\lim_{t\to t_1^-} f(t) = 0, \quad \lim_{t\to t_2^+} f(t) = \infty.$$

As a result, $f(t)$ is decreasing on (t_2, t_1) and the equation $f(t) = t$ has a unique solution by continuity. ■

Next, we introduce a few functions and look into their properties. These properties will be very useful in proving the existence of a solution to the optimality system under different regimes of λ_Z.

Lemma 13. *(a) Let*

$$\begin{aligned} g_1(x) &:= \log\left[\left(\frac{\sqrt{n_B}}{N\lambda_Z} - 1\right)(k_A x + k_B) + 1\right] - k_B \log x, \quad \lambda_Z \le \sqrt{n_B}/N, \\ g_2(x) &:= \log\left[\left(\frac{\sqrt{n_A}}{N\lambda_Z} - 1\right)(k_A + k_B x) + 1\right] - k_A \log x, \quad \lambda_Z \le \sqrt{n_A}/N, \end{aligned} \tag{5.23}$$

where $x > 0$. Then g_1 and g_2 are strictly decreasing in x.

(b) The function $f_2(t, \lambda_Z)$ in (5.16) and $m(t, \lambda_Z)$ in (5.17) are strictly increasing and decreasing in t respectively. Moreover, $t^*(\lambda_Z, \lambda_b)$ is increasing in $0 < \lambda_Z \le \sqrt{n_A}/N$ and $t^*(\lambda_Z, \lambda_b) \le n_A/n_B$ holds.

(c) For $N\lambda_Z \in [\sqrt{n_B}, \sqrt{n_A}]$, define

$$\xi(\lambda_Z, \lambda_b) := k_B e^{-(k_A+k_B)m(t^*(\lambda_Z,\lambda_b))} - \left(\frac{1}{N\lambda_Z}\sqrt{\left(\frac{k_B n_A}{t^*(\lambda_Z,\lambda_b)^2} + k_A n_B\right)(k_A + k_B)} - k_A\right). \tag{5.24}$$

Then for any $\lambda_b > 0$, it holds

$$\xi\left(\frac{\sqrt{n_A}}{N}, \lambda_b\right) > 0, \qquad \xi\left(\frac{\sqrt{n_B}}{N}, \lambda_b\right) < 0. \tag{5.25}$$

This implies by continuity that there exists an $\varepsilon > 0$, such that:

$$\xi\left(\frac{\sqrt{n_A}}{N} - \delta', \lambda_b\right) > 0, \qquad \xi\left(\frac{\sqrt{n_B}}{N} + \delta'', \lambda_b\right) < 0, \quad \forall \delta', \delta'' < \varepsilon.$$

(d) If $\lambda_b = \infty$, i.e., for the bias-free case, then

$$\xi(\lambda_Z, \infty) = (k_A + k_B) - \frac{1}{N\lambda_Z}\sqrt{\left(\frac{k_B n_A}{t^*(\lambda_Z)^2} + k_A n_B\right)(k_A + k_B)} \tag{5.26}$$

and it is increasing in $\lambda_Z \in [\sqrt{n_B}/N, \sqrt{n_A}/N]$. Combining with results from (5.25), we have there exists a unique $\lambda^* \in (\sqrt{n_B}/N, \sqrt{n_A}/N)$ such that

$$\xi(\lambda^*, \infty) = 0.$$

Proof of Lemma 13 (a) Consider the first-order derivative of $g_1(x)$, we have

$$g_1'(x) = \frac{k_A\left(\frac{\sqrt{n_B}}{N\lambda_Z} - 1\right)}{\left(\frac{\sqrt{n_B}}{N\lambda_Z} - 1\right)(k_A x + k_B) + 1} - \frac{k_B}{x} < \frac{1}{x} - \frac{k_B}{x} \leq 0$$

provided that $x > 0$ and $N\lambda_Z < \sqrt{n_B}$, and similarly $g_2'(x) < 0$ holds for $x > 0$ if $N\lambda_Z < \sqrt{n_A}$.

(b) We first note that $f_2(t, \lambda_Z, \lambda_b) = g_2(x_2(t)) - (k_A + k_B)k_A m(t)$. Since $x_2(t)$ in (5.18) is strictly decreasing w.r.t. t in its domain and so is g_2, as shown in Lemma 13(a), we conclude $g_2(x_2(t))$ is strictly increasing. For $m(t)$, we notice that

$$\begin{aligned} m'(t) &= -\frac{\lambda_Z}{\lambda_b}\frac{n_B(k_B n_A + n_B k_A t^2)(k_A + k_B) + k_A n_B t(n_A - n_B t)(k_A + k_B)}{[(k_B n_A + n_B k_A t^2)(k_A + k_B)]^{\frac{3}{2}}} \\ &= -\frac{\lambda_Z}{\lambda_b} n_B n_A (k_A + k_B)\frac{k_B + k_A t}{[(k_B n_A + n_B k_A t^2)(k_A + k_B)]^{\frac{3}{2}}} < 0 \end{aligned} \tag{5.27}$$

This implies m is decreasing in t, and then $f_2(t)$ is increasing in t. This proves the first half of the argument.

Now we investigate the monotonicity of t^* w.r.t. λ_Z. The idea is to apply implicit differentiation on f_2 at $t^*(\lambda_Z, \lambda_b)$. We have

$$\frac{\partial f_2}{\partial t} \cdot \frac{\mathrm{d}t^*(\lambda_Z, \lambda_b)}{\mathrm{d}\lambda_Z} + \frac{\partial f_2}{\partial \lambda_Z} = 0 \Longrightarrow \frac{\mathrm{d}t^*(\lambda_Z, \lambda_b)}{\mathrm{d}\lambda_Z} = -\frac{\partial f_2/\partial \lambda_Z}{\partial f_2/\partial t} \tag{5.28}$$

where we already know $\partial f_2/\partial t > 0$ from (b). Therefore it suffices to check the monotonicity of f_2 in λ_Z at $t^*(\lambda_Z, \lambda_b)$. Notice for any fixed $t < n_A/n_B$ ($m(t) > 0$) and $\lambda_b > 0$, we have that $f_2(t, \lambda_Z, \lambda_b)$ is decreasing in λ_Z. Finally, we claim $t^*(\lambda_Z, \lambda_b) < n_A/n_B$, so at $t^*(\lambda_Z, \lambda_b)$, $\partial f_2/\partial \lambda_Z < 0$. Therefore by (5.28), we have $\partial t^*(\lambda_Z, \lambda_b)/\partial \lambda_Z > 0$.

Now we are left with proving the claim $t^*(\lambda_Z, \lambda_b) < n_A/n_B$, which follows from the following simple argument. By plugging $t = n_A/n_B$ into f_2, it holds that

$$m(n_A/n_B) = 0, \quad x_2(n_A/n_B) < x_2(\sqrt{n_A/n_B}) = 1$$

where $n_A/n_B > 1$ and x_2 is decreasing. Therefore, by the fact g_2 is an increasing function of x, we have

$$f_2\left(\frac{n_A}{n_B}\right) = g_2\left(x_2\left(\frac{n_A}{n_B}\right)\right) > g_2\left(x_2\left(\sqrt{\frac{n_A}{n_B}}\right)\right) = g_2(1) > 0,$$

and by the monotonicity of f_2, $t^*(\lambda_Z, \lambda_b)$ must be smaller than n_A/n_B.

(c) We evaluate the value of ξ at $\lambda_Z = \sqrt{n_A}/N$ and $\lambda_Z = \sqrt{n_B}/N$ separately.

Right endpoint: at $\lambda_Z = \sqrt{n_A}/N$. Note that

$$\begin{aligned} f_2(\sqrt{n_A/n_B}, \sqrt{n_A}/N, \lambda_b) &= -(k_A + k_B)k_A m(\sqrt{n_A/n_B}, \sqrt{n_A}/N, \lambda_b) < 0, \\ f_2(n_A/n_B, \sqrt{n_A}/N, \lambda_b) &> 0, \end{aligned}$$

where $x_2(\sqrt{n_A/n_B}) = 1$. This means a root exists within the interval $(\sqrt{n_A/n_B}, n_A/n_B)$ for f_2 with $\lambda_Z = \sqrt{n_A}/N$. Denote $t^* = t^*(\lambda_Z, \lambda_b)$, $x_2^* = x_2(t^*(\lambda_Z, \lambda_b))$, and $\xi(\lambda_Z) = \xi(\lambda_Z, \lambda_b)$. By definition, the root t^* satisfies

$$\begin{aligned} x_2(t^*) = e^{-(k_A+k_B)m(t^*)} &= \frac{k_A}{\sqrt{(k_B + n_B k_A t^{*2}/n_A)(k_A+k_B)} - k_B} \\ \iff t^* \sqrt{\left(\frac{k_B}{t^{*2}} + \frac{n_B k_A}{n_A}\right)(k_A+k_B)} &= k_A e^{(k_A+k_B)m(t^*)} + k_B = \frac{k_A}{x_2^*} + k_B. \end{aligned} \tag{5.29}$$

Then applying (5.29) gives

$$\begin{aligned} \xi\left(\frac{\sqrt{n_A}}{N}\right) &= k_B e^{-(k_A+k_B)m(t^*)} - \left(\sqrt{\left(\frac{k_B}{t^{*2}} + \frac{k_A n_B}{n_A}\right)(k_A+k_B)} - k_A\right) \\ &= k_A + k_B e^{-(k_A+k_B)m(t^*)} - \frac{1}{t^*}(k_A e^{(k_A+k_B)m(t^*)} + k_B) \\ &= \left(k_A + k_B e^{-(k_A+k_B)m(t^*)}\right)\left(1 - \frac{e^{(k_A+k_B)m(t^*)}}{t^*}\right). \end{aligned}$$

Note that $1 < \sqrt{n_A/n_B} < t^* < n_A/n_B$:

$$e^{(k_A+k_B)m(t^*)} = \frac{\sqrt{(k_B + n_B t^{*2} k_A/n_A)(k_A+k_B)} - k_B}{k_A} \le \frac{1}{2}\left(\frac{n_B}{n_A}t^{*2} + 1\right) < \frac{1}{2}(t^*+1) < t^*$$

where for the first inequality, we use $\sqrt{xy} \le (x+y)/2$ for $x, y > 0$ and for next two, we use $t^* < n_A/n_B$ and $t^* > 1$ respectively. Therefore, $\xi\left(\sqrt{n_A}/N\right) > 0$ holds.

Left endpoint: at $\lambda_Z = \sqrt{n_B}/N$. There are two possible cases. If $f_2(0, \sqrt{n_B}/N, \lambda_b) \ge 0$, then $t^* = 0$ and $\xi(\sqrt{n_B}/N) = -\infty$ holds automatically. Otherwise, $t^* > 0$ and it satisfies

$$0 < \frac{1}{k_A}\log\left[\left(\sqrt{\frac{n_A}{n_B}} - 1\right)(k_A + k_B x_2^*) + 1\right] = \log x_2^* + (k_A+k_B)m(t^*) \Longrightarrow e^{-(k_A+k_B)m(t^*)} < x_2^*.$$

Then it holds that

$$\begin{aligned} \xi\left(\frac{\sqrt{n_B}}{N}\right) &= k_B e^{-(k_A+k_B)m(t^*)} - \left(\sqrt{\frac{n_A}{n_B}}\sqrt{\left(\frac{k_B}{t^{*2}} + \frac{k_A n_B}{n_A}\right)(k_A+k_B)} - k_A\right) \\ &< k_A + k_B x_2^* - \sqrt{\frac{n_A}{n_B}} \cdot \frac{1}{t^*}\left(\frac{k_A}{x_2^*} + k_B\right) \\ &\le (k_A + k_B x_2^*)\left(1 - \sqrt{\frac{n_A}{n_B}} \cdot \frac{1}{t^* x_2^*}\right) \le 0 \end{aligned}$$

where the second line follows from (5.29), and the last line uses

$$t x_2(t) = \frac{k_A t}{\sqrt{(k_B + k_A n_B t^2/n_A)(k_A+k_B)} - k_B} \le \frac{k_A t}{k_A\sqrt{n_B/n_A}\,t} \le \sqrt{\frac{n_A}{n_B}}, \quad \forall t > 0,$$

since

$$\sqrt{\left(k_B+\frac{k_An_Bt^2}{n_A}\right)(k_A+k_B)}-k_B\geq k_B\left(\sqrt{\left(1+\frac{k_An_Bt^2}{k_Bn_A}\right)\left(1+\frac{k_A}{k_B}\right)}-1\right)\geq k_A\sqrt{\frac{n_B}{n_A}}t$$

where the last inequality follows from $(1+x)(1+y)\geq(1+\sqrt{xy})^2$ for $x,y\geq 0$.

(d) For any $\lambda_Z\in[\sqrt{n_B}/N,\sqrt{n_A}/N]$ and $t\in[0,n_A/n_B]$, we know that $m(t,\lambda_Z,\lambda_b)$ is uniformly bounded. As $\lambda_b\to\infty$, we have

$$\lim_{\lambda_b\to\infty}\sup_{0\leq t\leq n_A/n_B,\sqrt{n_B}\leq N\lambda_Z\leq\sqrt{n_A}}|m(t,\lambda_Z,\lambda_b)|=0.$$

As a result, as $\lambda_b\to\infty$,

$$\lim_{\lambda_b\to\infty}\xi(\lambda_Z,\lambda_b)=(k_A+k_B)-\frac{1}{N\lambda_Z}\sqrt{\left(\frac{k_Bn_A}{t^*(\lambda_Z)^2}+k_An_B\right)(k_A+k_B)}.$$

Note that $t^*(\lambda_Z)$ is increasing in λ_Z (proven in (b)), and so is $\xi(\lambda_Z,\infty)$. From (c), we know that there exists a λ^* satisfying $\sqrt{n_B}\leq N\lambda^*\leq\sqrt{n_A}$ such that for any $\lambda_Z<\lambda^*$, $\xi(\lambda_Z,\infty)<0$ holds, and for any $\lambda_Z>\lambda^*$, $\xi(\lambda_Z,\infty)>0$ holds. This λ^* is the zero to $\xi(\lambda_Z,\infty)$, i.e.,

$$t^*(\lambda^*)=\sqrt{\frac{k_Bn_A}{(N\lambda_Z)^2(k_A+k_B)-k_An_B}}.$$

■

5.4 Proof of Proposition 11

The idea is straightforward, according to our parametrization of $\bar{\boldsymbol{Z}}$ and $\boldsymbol{b}$ in (3.5), we have $\mathbf{1}^\top(\bar{\boldsymbol{Z}},\boldsymbol{b})=0$, which is necessary for being the global minimizer by Lemma 8. By strong convexity on the set $\{(\bar{\boldsymbol{Z}},\boldsymbol{b})|\mathbf{1}^\top(\bar{\boldsymbol{Z}}+\boldsymbol{b}\mathbf{1}_N^\top)=0\}$ given by Lemma 7, it suffices to verify the first order optimality equation has a solution in each regime of λ_Z which is automatically unique and global by strong convexity. Therefore the key ideas of the proof are similar for all four cases. We first propose a candidate solution for $\bar{\boldsymbol{Z}}$ and certificate $\bar{\boldsymbol{R}}$ such that they are admissible according to Corollary 9, i.e.,

$$\bar{\boldsymbol{R}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top=0,\quad\bar{\boldsymbol{Z}}^\top\bar{\boldsymbol{R}}=0,\quad\|\bar{\boldsymbol{R}}\boldsymbol{D}^{1/2}\|\leq 1.\tag{5.30}$$

Then we derive the corresponding optimality equation system and prove the existence of a solution. In the following proof, we will let $\lambda_b>0$ be any positive number and the bias vector $\boldsymbol{b}$ is in the form of (3.5).

Proof of Proposition 11(a) We consider $N\lambda_Z<\sqrt{n_B}$.

Solution structure: In this case, we propose the structure:

$$\bar{\boldsymbol{Z}}=\mathcal{B}(a,b,c,d):=\begin{bmatrix}a(k_A\boldsymbol{I}_{k_A}-\boldsymbol{J}_{k_A\times k_A})+k_Bc\boldsymbol{I}_{k_A} & -b\boldsymbol{J}_{k_A\times k_B}\\ -c\boldsymbol{J}_{k_B\times k_A} & d(k_B\boldsymbol{I}_{k_B}-\boldsymbol{J}_{k_B\times k_B})+k_Ab\boldsymbol{I}_{k_B}\end{bmatrix},\quad\bar{\boldsymbol{R}}=0.$$

Notice conditions (5.30) are trivially satisfied as $\bar{\boldsymbol{R}}$ being zero.

Optimality system: Note that $\boldsymbol{I}_K - \bar{\boldsymbol{P}}$ and $\left[(\bar{\boldsymbol{Z}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top)^\dagger\right]^{\frac{1}{2}}\bar{\boldsymbol{Z}}$ are in the block structure with $\mathcal{B}(a_P, b_P, c_P, d_P)$ in (5.15) and $\mathcal{B}(a_Z, b_Z, c_Z, d_Z)$ in (5.12). By comparing the coefficients, we need to ensure

$$\begin{aligned} k_A a_P + k_B c_P = N\lambda_Z(k_A a_Z + k_B c_Z), \quad k_A b_P + k_B d_P = N\lambda_Z(k_A b_Z + k_B d_Z), \\ b_P = N\lambda_Z b_Z > 0, \quad c_P = N\lambda_Z c_Z > 0. \end{aligned}$$

Then the equations above along with the second equation in (5.4), $N^{-1}(\boldsymbol{I}_K - \bar{\boldsymbol{P}})\boldsymbol{n} = \lambda_b \boldsymbol{b}$, yield to the following system,

$$\begin{aligned} \frac{k_A e^{-a+mk_B} + k_B e^{-c-mk_A}}{k_B e^{-c-mk_A} + e^{-a+mk_B}(k_A - 1 + e^{k_A a + k_B c})} &= \frac{N\lambda_Z}{\sqrt{n_A}}, \\ \frac{k_A e^{-b+mk_B} + k_B e^{-d-mk_A}}{k_A e^{-b+mk_B} + e^{-d-mk_A}(k_B - 1 + e^{k_A b + k_B d})} &= \frac{N\lambda_Z}{\sqrt{n_B}}, \\ \frac{e^{-c-mk_A}}{k_B e^{-c-mk_A} + e^{-a+mk_B}(k_A - 1 + e^{k_A a + k_B c})} &= \frac{N\lambda_Z c}{\sqrt{(k_A n_B b^2 + k_B n_A c^2)(k_A + k_B)}}, \\ \frac{e^{-b+mk_B}}{k_A e^{-b+mk_B} + e^{-d-mk_A}(k_B - 1 + e^{k_A b + k_B d})} &= \frac{N\lambda_Z b}{\sqrt{(k_A n_B b^2 + k_B n_A c^2)(k_A + k_B)}}, \\ m = \frac{\lambda_Z}{\lambda_b} &\frac{n_A c - n_B b}{\sqrt{(k_A n_B b^2 + k_B n_A c^2)(k_A + k_B)}}. \end{aligned} \tag{5.31}$$

Our goal now is to show that there exists a solution to (5.31) with $b > 0$ and $c > 0$. Let $t = b/c$, and then combining the four equations in (5.31) gives

$$\begin{aligned} \text{eq.1 in (5.31)} \quad & c = \frac{g_2(e^{a-c-m(k_A+k_B)})}{k_A + k_B} - mk_A, \\ \text{eq.2 in (5.31)} \quad & b = \frac{g_1(e^{d-b+m(k_A+k_B)})}{k_A + k_B} + mk_B, \\ \text{eq.1 and 3 in (5.31)} \quad & x_2(t) := e^{a-c-m(k_A+k_B)} = \frac{k_A}{\sqrt{(k_B + k_A n_B t^2/n_A)(k_A + k_B)} - k_B}, \\ \text{eq.2 and 4 in (5.31)} \quad & x_1(t) := e^{d-b+m(k_A+k_B)} = \frac{k_B}{\sqrt{(k_A + n_A k_B t^{-2}/n_B)(k_A + k_B)} - k_A}, \\ \text{eq.5 in(5.31)} \quad & m = \frac{\lambda_Z}{\lambda_b} \frac{n_A - n_B t}{\sqrt{(k_B n_A + k_A n_B t^2)(k_A + k_B)}} \end{aligned} \tag{5.32}$$

where g_1 and g_2 are defined in (5.23). Here

$$x_1(t) := e^{d-b+m(k_A+k_B)}, \quad x_2(t) := e^{a-c-m(k_A+k_B)} \tag{5.33}$$

respectively, and $x_1(t)$ and $x_2(t)$ are monotonic increasing and decreasing function of t on $(t > 0)$ respectively in (5.32). Using the fact $(1+x)(1+y) \geq (1+\sqrt{xy})^2$ for any $x \geq 0$ and

$y \geq 0$ leads to

$$x_1(t) \leq \frac{k_B}{k_A}\sqrt{\frac{n_B}{n_A}}t, \quad x_2(t) \leq \frac{k_A}{k_B}\sqrt{\frac{n_A}{n_B}}t^{-1}. \tag{5.34}$$

Finally, we simplify the equations (5.32) into

$$b = \frac{g_1(x_1(t))}{k_A + k_B} + mk_B, \quad c = \frac{g_2(x_2(t))}{k_A + k_B} - mk_A. \tag{5.35}$$

Now, the goal is to show the existence of solution set (a, b, c, d, m) for system (5.32).

Existence of a solution to (5.32)**:** Observe that the existence of (a, d, m) fully relies on the existence of (b, c) through eq.3-5 in (5.32). Notice, the RHS of both eq.1 and 2 only depends on $t := b/c$, so the existence of (b, c) is equivalent to whether the following single-variable equation has a solution for $t > 0$:

$$t := \frac{b}{c} = \frac{g_1(x_1(t)) + (k_A + k_B)k_B m}{g_2(x_2(t)) - (k_A + k_B)k_A m} =: \frac{f_1(t)}{f_2(t)} \tag{5.36}$$

It remains to show (5.36) has a solution. The argument will rely on the monotonicity of f_1 and f_2, and Lemma 12.

Note that g_1, g_2 and m are decreasing in t, as shown in Lemma 13. Therefore, $f_1(t)$ and $f_2(t)$ are monotonically decreasing and increasing respectively on $(0, \infty)$. Note that

$$\lim_{t\to 0^+} m(t) = \frac{\lambda_Z}{\lambda_b}\sqrt{\frac{n_A}{k_B(k_A + k_B)}}, \quad \lim_{t\to+\infty} m(t) = -\frac{\lambda_Z}{\lambda_b}\sqrt{\frac{n_B}{k_A(k_A + k_B)}}.$$

and

$$\lim_{t\to 0^+} f_1(t) = \infty, \quad \lim_{t\to\infty} f_2(t) = \infty.$$

Lemma 12 indicates that it remains to see if $\{t : f_1(t) > 0, f_2(t) > 0\}$ is empty. Due to the monotonicity of f_1 and f_2, it suffices to show that it is impossible to have $f_1(t) < 0$ and $f_2(t) < 0$ for some t. We will prove this is impossible by contradiction. Now, we assume $k_1 f_1(t) + k_B f_2(t) \leq 0$ holds. Note that

$$\begin{aligned} k_1 f_1(t) + k_B f_2(t) &= k_1 g_1(x_1(t)) + k_2 g_2(x_2(t)) \leq 0 \\ \iff k_A \log &\left[\left(\frac{\sqrt{n_B}}{N\lambda_Z} - 1\right)(k_A x_1(t) + k_B) + 1\right] \\ + k_B \log &\left[\left(\frac{\sqrt{n_A}}{N\lambda_Z} - 1\right)(k_B x_2(t) + k_A) + 1\right] \leq k_A k_B \log x_1(t)x_2(t) \end{aligned}$$

where $x_1(t)x_2(t) \leq 1$ follow from (5.34). This leads to a contradiction, as the left-hand side is strictly positive. Therefore, $k_1 f_1(t) + k_B f_2(t) > 0$ holds which implies that $\{t : f_1(t) > 0, f_2(t) > 0\}$ is nonempty. Using Lemma 12 finishes the proof, as it implies a root exists for (5.36). ■

Proof of Proposition 11(b) We consider $\sqrt{n_B} < N\lambda_Z < \sqrt{n_A}$ and λ_Z satisfies (5.21).
Solution structure: In this case, we propose the structure

$$\bar{\boldsymbol{Z}} = \begin{bmatrix} a(k_A \boldsymbol{I}_{k_A} - \boldsymbol{J}_{k_A \times k_A}) + k_B c \boldsymbol{I}_{k_A} & -b\boldsymbol{J}_{k_A \times k_B} \\ -c\boldsymbol{J}_{k_B \times k_A} & -d\boldsymbol{J}_{k_B} \end{bmatrix}, \quad \bar{\boldsymbol{R}} = \begin{bmatrix} 0 & 0 \\ 0 & \alpha(\boldsymbol{I}_{k_B} - \boldsymbol{J}_{k_B}/k_B) \end{bmatrix} \tag{5.37}$$

In other words, $\bar{\boldsymbol{Z}} = \mathcal{B}(a, b, c, d)$ with $k_A b + k_B d = 0$ and $\bar{\boldsymbol{R}}$ is $\mathcal{B}(0, 0, 0, \alpha/k_B)$. We have $\bar{\boldsymbol{R}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top = \bar{\boldsymbol{Z}}^\top \bar{\boldsymbol{R}} = 0$ and $\|\bar{\boldsymbol{R}}\boldsymbol{D}^{1/2}\| \le 1$ holds as long as $\alpha\sqrt{n_B} \le 1$.

Optimality system: Based on (5.37), $[(\bar{\boldsymbol{Z}}\boldsymbol{D}\bar{\boldsymbol{Z}})^\dagger]^{1/2}\bar{\boldsymbol{Z}}$ in (5.12) reduces to $\mathcal{B}(a_Z, b_Z, c_Z, d_Z)$ satisfying

$$\begin{aligned} &k_A a_Z + k_B c_Z = n_A^{-1/2}\,\mathrm{sign}(k_A a + k_B c), \quad k_A b_Z + k_B d_Z = 0, \\ &b_Z = \frac{b\,\mathrm{sign}(k_A n_B b^2 + k_B n_A c^2)}{\sqrt{(k_A n_B b^2 + k_B n_A c^2)(k_A + k_B)}}, \quad c_Z = \frac{c\,\mathrm{sign}(k_A n_B b^2 + k_B n_A c^2)}{\sqrt{(k_A n_B b^2 + k_B n_A c^2)(k_A + k_B)}}, \end{aligned} \tag{5.38}$$

and $\boldsymbol{I}_K - \bar{\boldsymbol{P}}$ in (5.15) becomes $\mathcal{B}(a_P, b_P, c_P, d_P)$ with

$$\begin{aligned} a_P &= \frac{e^{-a+mk_B}}{k_B e^{-c-mk_A} + e^{-a+mk_B}(k_A - 1 + e^{k_A a + k_B c})}, \\ b_P &= \frac{e^{-b+mk_B}}{k_A e^{-b+mk_B} + k_B e^{-d-mk_A}}, \\ c_P &= \frac{e^{-c-mk_A}}{k_B e^{-c-mk_A} + e^{-a+mk_B}(k_A - 1 + e^{k_A a + k_B c})}, \\ d_P &= \frac{e^{-d-mk_A}}{k_A e^{-b+mk_B} + k_B e^{-d-mk_A}}. \end{aligned} \tag{5.39}$$

Using the optimality condition $\boldsymbol{I}_K - \bar{\boldsymbol{P}} = N\lambda_Z((\bar{\boldsymbol{Z}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top)^{\dagger/2}\bar{\boldsymbol{Z}} + \bar{\boldsymbol{R}})$ leads to a normal equation similar to (5.31):

$$\begin{aligned} k_A a_P + k_B c_P &= N\lambda_Z(k_A a_Z + k_B c_Z), \\ k_A b_P + k_B d_P = 1 &= N\lambda_Z(k_A b_Z + k_B d_Z + \alpha) = N\lambda_Z \alpha, \\ b_P &= N\lambda_Z b_Z > 0, \\ c_P &= N\lambda_Z c_Z > 0, \end{aligned}$$

where $(\bar{\boldsymbol{Z}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top)^{\dagger/2}\bar{\boldsymbol{Z}} + \bar{\boldsymbol{R}}$ is $\mathcal{B}(a_Z, b_Z, c_Z, d_Z + \alpha/k_B)$ and $k_A b_Z + k_B d_Z = 0$. By comparing the coefficients, we have

$$\begin{aligned} &\frac{k_A e^{-a+mk_B} + k_B e^{-c-mk_A}}{k_B e^{-c-mk_A} + e^{mk_A - a}(k_A - 1 + e^{k_A a + k_B c})} = \frac{N\lambda_Z}{\sqrt{n_A}}, \\ &\frac{e^{-b+mk_B}}{k_A e^{-b+mk_B} + k_B e^{-d-mk_A}} = \frac{N\lambda_Z b}{\sqrt{(k_A n_B b^2 + k_B n_A c^2)(k_A + k_B)}}, \\ &\frac{e^{-c-mk_A}}{k_B e^{-c-mk_A} + e^{-a+mk_B}(k_A - 1 + e^{k_A a + k_B c})} = \frac{N\lambda_Z c}{\sqrt{(k_A n_B b^2 + k_B n_A c^2)(k_A + k_B)}}, \\ &k_A b + k_B d = 0, \quad 1 = N\lambda_Z \alpha, \\ &m = \frac{\lambda_Z}{\lambda_b} \frac{n_A c - n_B b}{\sqrt{(k_A n_B b^2 + k_B n_A c^2)(k_A + k_B)}}. \end{aligned} \tag{5.40}$$

Our goal is to show that there exists a solution $(a, b, c, -k_A b/k_B)$ and α such that the equations above have a unique solution. Here $\left\|\bar{\boldsymbol{R}}\boldsymbol{D}^{1/2}\right\| = \sqrt{n_B}\alpha = \sqrt{n_B}/N\lambda_Z \leq 1$ with $\alpha = 1/N\lambda_Z$ if $\lambda_Z \geq \sqrt{n_B}/N$, so the certificate $\bar{\boldsymbol{R}}$ is feasible. By letting $t = b/c$, we just need to show the following simplified system derived from (5.40) has a solution for (a, b, c, d, m):

$$\begin{aligned}
&\text{eq.1 in (5.40)} \quad c = \frac{g_2(e^{a-c-m(k_A+k_B)})}{k_A+k_B} - mk_A,\\
&\text{eq.4 in (5.40)} \quad b = \frac{-(d-b)k_B}{k_A+k_B},\\
&\text{eq.1 and 3 in (5.40)} \quad x_2(t) := e^{a-c-m(k_A+k_B)} = \frac{k_A}{\sqrt{(k_B + k_A n_B t^2/n_A)(k_A+k_B)} - k_B},\\
&\text{eq.2 in (5.40)} \quad x_1(t) := e^{d-b+m(k_A+k_B)} = \frac{k_B}{(N\lambda_Z)^{-1}\sqrt{(k_A n_B + k_B n_A t^{-2})(k_A+k_B)} - k_A},\\
&\text{eq.5 in (5.40)} \quad m = \frac{\lambda_Z}{\lambda_b}\frac{n_A - n_B t}{\sqrt{(k_B n_A + k_A n_B t^2)(k_A+k_B)}}
\end{aligned} \tag{5.41}$$

Existence of a solution to (5.41)**:** Similarly, we consider the following single-variable nonlinear equation for t,

$$t = \frac{-k_B(d-b)}{g_2(x_2(t)) - k_A(k_A+k_B)m(t)} =: \frac{f_1(t)}{f_2(t)} = f(t) \tag{5.42}$$

where $x_1(t)$ and $x_2(t)$ are defined in (5.41), and g_2 is in (5.23). Equivalently we can also write $f_1(t) = -k_B\left[\log x_1(t) - (k_A+k_B)m(t)\right]$.

The idea of the proof is similar to the case (a): by using the monotonicity of f_1 and f_2, and also Lemma 12. We denote the domain of $f_1(t)$ as $\mathcal{D}$, i.e.,

$$\begin{aligned}
\mathcal{D} &= \left\{t > 0 \mid (N\lambda_Z)^{-1}\sqrt{(k_B n_A t^{-2} + k_A n_B)(k_A+k_B)} > k_A\right\}\\
&= \begin{cases} t \in \mathbb{R}, & N\lambda_Z \leq \sqrt{n_B(k_A+k_B)/k_A},\\ t < \sqrt{\frac{k_B n_A(k_A+k_B)}{(N\lambda_Z)^2 k_A^2 - k_A n_B(k_A+k_B)}}, & N\lambda_Z > \sqrt{n_B(k_A+k_B)/k_A}, \end{cases}
\end{aligned}$$

and that of $f_2(t)$ is $\mathbb{R}$. In their domains, $x_1(t)$ and $x_2(t)$ are increasing and decreasing respectively, which implies $f_1(t)$ and $f_2(t)$ are strictly decreasing and increasing in t respectively. It is straightforward to verify:

$$\lim_{t\to 0^+} f_1(t) = +\infty, \quad \lim_{t\to\infty} f_2(t) = +\infty,$$

where $m(t)$ stays bounded for any t, $\lim_{t\to 0^+} f_1(t) = \infty$ due to $\lim_{t\to 0^+} x_1(t) = 0^+$, and $\lim_{t\to\infty} f_2(t) = \infty$ due to $\lim_{t\to\infty} x_2(t) \to 0^+$ and $\lim_{x\to 0^+} g_2(x) = \infty$. It suffices to show that $f_1(t)$ and $f_2(t)$ share a common positive part. The following argument divides into two subcases whether $f_2(t)$ has a root or not.

Suppose $\eta(\lambda_Z)$ in (5.20) is positive ($f_2(t)$ has no root), then f_1 and f_2 share a common positive part since as $t \to 0^+$, f_1 goes to ∞ and f_2 stays positive. For λ_Z with $\eta(\lambda_Z) < 0$,

f_2 has a unique zero $t^*(\lambda_Z)$. To ensure f_1 and f_2 share a common positive part, it suffices to have

$$f_1(t^*(\lambda_Z)) = -k_B \log x_1(t^*(\lambda_Z)) + (k_A + k_B)k_B m(t^*(\lambda_Z)) > 0,$$

which is equivalent to

$$\begin{aligned} x_1(t^*(\lambda_Z)) &= \frac{k_B}{(N\lambda_Z)^{-1}\sqrt{(k_B n_A t^*(\lambda_Z)^{-2} + k_A n_B)(k_A + k_B)} - k_A} < e^{(k_A+k_B)m(t^*(\lambda_Z))} \\ \iff k_B e^{-(k_A+k_B)m(t^*(\lambda_Z))} &< \frac{1}{N\lambda_Z}\sqrt{\left(\frac{k_B n_A}{t^*(\lambda_Z)^2} + k_A n_B\right)(k_A + k_B)} - k_A. \end{aligned}$$

This finishes the proof, as (5.21) guarantees $f_1(t^*(\lambda_Z)) > 0$ based on the argument above.

Note that Lemma 13 implies that the inequality above holds for any $\lambda_Z = \sqrt{n_B}/N + \delta$ with $\delta < \varepsilon$. Hence for any λ_Z close to $\sqrt{n_B}/N$, the case (b) in Proposition 11 holds. ■

Proof of Proposition 11(c) We consider $\sqrt{n_B} < N\lambda_Z < \sqrt{n_A}$ and λ_Z satisfies (5.22).
Solution Structure: In this case, we propose the structure:

$$\begin{aligned} \bar{\boldsymbol{Z}} = \mathcal{B}(a,0,0,0) &= \begin{bmatrix} a(k_A \boldsymbol{I}_{k_A} - \boldsymbol{J}_{k_A \times k_A}) & 0 \\ 0 & 0 \end{bmatrix}, \\ \bar{\boldsymbol{R}} = \mathcal{B}(a_R, b_R, c_R, d_R) &= \begin{bmatrix} 0 & 0 \\ 0 & \alpha(\boldsymbol{I}_{k_B} - \boldsymbol{J}_{k_B}/k_B) \end{bmatrix} + \frac{k_A k_B \tau}{\sqrt{(k_B n_A + k_A n_B t^2)(k_A + k_B)}} \boldsymbol{s}\boldsymbol{s}^\top \begin{bmatrix} \boldsymbol{I}_{k_A} & 0 \\ 0 & t\boldsymbol{I}_{k_B} \end{bmatrix} \end{aligned} \tag{5.43}$$

where

$$\begin{aligned} &k_A a_R + k_B c_R = 0, \quad k_A b_R + k_B d_R = \alpha, \\ &b_R = \frac{\tau t}{\sqrt{(k_B n_A + k_A n_B t^2)(k_A + k_B)}}, \quad c_R = \frac{\tau}{\sqrt{(k_B n_A + k_A n_B t^2)(k_A + k_B)}}, \end{aligned} \tag{5.44}$$

which satisfies $\bar{\boldsymbol{R}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top = \bar{\boldsymbol{Z}}^\top\bar{\boldsymbol{R}} = 0$ and $\|\bar{\boldsymbol{R}}\boldsymbol{D}^{1/2}\| \le 1$ holds as long as

$$\max\{\sqrt{n_B}|\alpha|, \tau\} \le 1.$$

Optimality system: Based on structure (5.43), we have (5.12) reduce to

$$k_A a_Z + k_B c_Z = n_A^{-1/2}\,\mathrm{sign}(a), \quad k_A b_Z + k_B d_Z = 0, \quad b_Z = c_Z = 0, \tag{5.45}$$

and (5.15) becomes

$$\begin{aligned} a_P &= \frac{e^{-a+mk_B}}{k_B e^{-mk_A} + e^{-a+mk_B}(k_A - 1 + e^{k_A a})}, \quad b_P = \frac{e^{mk_B}}{k_A e^{mk_B} + k_B e^{-mk_A}}, \\ c_P &= \frac{e^{-mk_A}}{k_B e^{-mk_A} + e^{-a+mk_B}(k_A - 1 + e^{k_A a})}, \quad d_P = \frac{e^{-mk_A}}{k_A e^{mk_B} + k_B e^{-mk_A}}. \end{aligned} \tag{5.46}$$

Using the optimality condition $\boldsymbol{I}_K - \bar{\boldsymbol{P}} = N\lambda_Z([(\bar{\boldsymbol{Z}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top)^\dagger]^{1/2}\bar{\boldsymbol{Z}} + \bar{\boldsymbol{R}})$ leads to a normal equation similar to (5.31):

$$k_A a_P + k_B c_P = N\lambda_Z(k_A a_Z + k_B c_Z), \quad k_A b_P + k_B d_P = 1 = N\lambda_Z \alpha,$$
$$b_P = N\lambda_Z\,(b_Z + b_R) > 0, \quad c_P = N\lambda_Z\,(c_Z + c_R) > 0.$$

where (a_Z, b_Z, c_Z, d_Z), (a_R, b_R, c_R, d_R) and (a_P, b_P, c_P, d_P) satisfy (5.45), (5.44) and (5.46) respectively. By comparing the coefficients, we have

$$\begin{aligned}
&\frac{k_A e^{-a+mk_B} + k_B e^{-mk_A}}{k_B e^{-mk_A} + e^{-a+mk_B}(k_A - 1 + e^{k_A a})} = \frac{N\lambda_Z}{\sqrt{n_A}},\\
&\frac{e^{mk_B}}{k_A e^{mk_B} + k_B e^{-mk_A}} = \frac{N\lambda_Z \tau t}{\sqrt{(k_B n_A + k_A n_B t^2)(k_A + k_B)}},\\
&\frac{e^{-mk_A}}{k_B e^{-mk_A} + e^{-a+mk_B}(k_A - 1 + e^{k_A a})} = \frac{N\lambda_Z \tau}{\sqrt{(k_B n_A + k_A n_B t^2)(k_A + k_B)}},\\
&m = \frac{\lambda_Z \tau}{\lambda_b} \frac{n_A - n_B t}{\sqrt{(k_B n_A + k_A n_B t^2)(k_A + k_B)}}, \quad \tau \le 1, \quad N\lambda_Z \alpha = 1.
\end{aligned} \tag{5.47}$$

Here $\sqrt{n_B}\alpha = \sqrt{n_B}/(N\lambda_Z) \le 1$ with $\alpha = 1/N\lambda_Z$ if $\lambda_Z \ge (\sqrt{n_B}N)$. Let $x(t,\tau) := e^{a - m(k_A + k_B)}$, and then the optimality system (5.47) becomes

$$\begin{aligned}
\text{eq.1 in (5.47)} \quad & h(t, \lambda_Z, \lambda_b, \tau) := \log\left[\left(\frac{\sqrt{n_A}}{N\lambda_Z} - 1\right)(k_A + k_B x(t,\tau)) + 1\right]\\
& \quad - k_A \log x(t,\tau) - (k_A + k_B) k_A m(t, \lambda_Z, \lambda_b, \tau) = 0,\\
\text{eq.1 and 3 in (5.47)} \quad & x(t,\tau) := e^{a - (k_A + k_B)m}\\
& \quad = \frac{k_A}{\tau^{-1}\sqrt{(k_B + t^2 k_A n_B / n_A)(k_A + k_B)} - k_B},\\
\text{eq.4 in (5.47)} \quad & m(t, \lambda_Z, \lambda_b, \tau) := \frac{\lambda_Z \tau}{\lambda_b} \frac{n_A - n_B t}{\sqrt{(k_B n_A + k_A n_B t^2)(k_A + k_B)}},\\
\text{eq.2 in (5.47), eq.2-3 in (5.48)} \quad & t = \frac{\sqrt{n_A}(k_A / x(t,\tau) + k_B)}{N\lambda_Z (k_A + k_B e^{-(k_A + k_B) m(t, \lambda_Z, \tau)})},\\
\text{eq.4 in (5.47)} \quad & \tau \le 1.
\end{aligned} \tag{5.48}$$

In particular, if $\tau = 1$, $h(t, \lambda_Z, \lambda_b, 1) = f_2(t, \lambda_Z, \lambda_b)$ holds where f_2 is given in (5.16). When no confusion arises, we denote $m(t, \lambda_Z, \lambda_b, \tau)$ and $x(t,\tau)$ by $m(t)$ and $x(t)$ respectively. Now the goal is to prove the existence of solution set (a, m, t, τ) of the above system.

Existence of a solution to (5.48)**:** The proof follows from two steps: (i) given λ_Z, $h(t, \lambda_Z, \tau)$ has a root $t(\lambda, \tau, \lambda_b)$ for any $\tau \in [\tau^*_{\lambda_Z, \lambda_b}, 1]$ where $\tau^*_{\lambda_Z, \lambda_b}$ is a number only depends on λ_Z and λ_b if

$$f_2(0, \lambda_Z, \lambda_b) = h(0, \lambda_Z, \lambda_b, 1) < 0$$

so that the first three equations in (5.48) satisfy; (ii) we show that there exists a $\tau \le 1$ such that the fourth equation in (5.48) also holds. The combination of steps (i) and (ii) is sufficient to prove the existence of the solution, we prove them respectively.

Proof for step (i): For simplicity, we use $t(\lambda_Z, \tau)$ or t to replace $t(\lambda_Z, \lambda_b, \tau)$. Observe that x and m are both determined by t and τ according to eq 2 and 3 of (5.48), so the key is to prove the existence of t as the root of h. Note that the domain of $h(t, \lambda_Z, \tau)$ is $\mathbb{R}$ for any fixed $0 < \tau \leq 1$ and $\sqrt{n_B} \leq N\lambda_Z \leq \sqrt{n_A}$, and h is strictly increasing in t with $\lim_{t\to\infty} h(t, \lambda_Z, \tau) = \infty$ and

$$h(0, \lambda_Z, \lambda_b, \tau) = \log\left[\left(\frac{\sqrt{n_A}}{N\lambda_Z} - 1\right)(k_A + k_B x(0)) + 1\right] - k_A \log x(0) - (k_A + k_B)k_A m(0)$$

where

$$m(0) = \frac{\lambda_Z \tau}{\lambda_b}\sqrt{\frac{n_A}{k_B(k_A + k_B)}}, \quad x(0) = \frac{k_A}{\tau^{-1}\sqrt{k_B(k_A + k_B)} - k_B}.$$

This implies $h(t, \lambda_Z, \lambda_b, \tau)$ has a unique solution in t for a given triple of $(\lambda_Z, \lambda_b, \tau)$ if and only if $h(0, \lambda_Z, \lambda_b, \tau) < 0$.

We can also see that $h(0, \lambda_Z, \lambda_b, \tau)$ is decreasing in λ_Z and τ. Note that

$$\lim_{\tau\to 0^+} h(0, \lambda_Z, \lambda_b, \tau) = +\infty.$$

For $h(0, \lambda_Z, \lambda_b, 1) < 0$, we define $\tau^*_{\lambda_Z, \lambda_b} \in [0, 1]$ as

$$h(0, \lambda_Z, \lambda_b, \tau^*_{\lambda_Z, \lambda_b}) = 0$$

where $\tau^*_{\lambda_Z, \lambda_b}$ is the unique zero of $h(0, \lambda_Z, \lambda_b, \tau)$ in τ as $h(0, \lambda_Z, \lambda_b, \tau)$ is continuous in τ for any λ_Z.

Now we define $t(\lambda_Z, \tau)$ be the zero to $h(t, \lambda_Z, \lambda_b, \tau)$, i.e.,

$$h(t(\lambda_Z, \tau), \lambda_Z, \lambda_b, \tau) = 0, \quad \forall \tau \in [\tau^*_{\lambda_Z, \lambda_b}, 1].$$

In particular, $t(\lambda_Z, \tau^*_{\lambda_Z, \lambda_b}) = 0$ holds, so the existence of $x(t(\lambda_Z, \tau))$ and $m(t(\lambda_Z, \tau))$ follow.

Proof for step (ii): It suffices to find a τ such that the fourth equation holds. Let

$$L(\lambda_Z, \tau) = t(\lambda_Z, \tau) - \frac{\sqrt{n_A}(k_A/x(t(\lambda_Z, \tau), \tau) + k_B)}{N\lambda_Z(k_A + k_B e^{-(k_A + k_B)m(t(\lambda_Z, \tau), \lambda_Z, \tau)})}$$

for any $\tau \in [\tau^*_{\lambda_Z, \lambda_b}, 1]$ and λ_Z with $f_2(0, \lambda_Z) \leq 0$.

To show $L(\lambda_Z, \tau)$ has a zero in τ, we check the value of $L(\lambda_Z, \tau)$ at $\tau = \tau^*_{\lambda_Z, \lambda_b}$ and $\tau = 1$. At $\tau = \tau^*_{\lambda_Z, \lambda_b}$, it holds $t(\lambda_Z, \tau^*_{\lambda_Z, \lambda_b}) = 0$ and

$$L(\lambda_Z, \tau^*_{\lambda_Z, \lambda_b}) = -\frac{\sqrt{n_A}(k_A/x(0, \tau^*_{\lambda_Z, \lambda_b}), \tau) + k_B)}{N\lambda_Z(k_A + k_B e^{-(k_A + k_B)m(0, \lambda_Z, \tau^*_{\lambda_Z, \lambda_b})})} < 0.$$

At $\tau = 1$, we have $t(\lambda_Z, 1) = t^*(\lambda_Z)$, $x(t^*(\lambda_Z), 1) = x_2(t^*(\lambda_Z))$, $m(t^*(\lambda_Z), \lambda_Z, 1) = m(t^*(\lambda_Z))$, and

$$\frac{k_A}{x(t^*(\lambda_Z), 1)} + k_B = \sqrt{(k_B + k_A n_B t^*(\lambda_Z)^2/n_A)(k_A + k_B)}$$

follows from the second equation in (5.48). Therefore,

$$\begin{aligned} L(\lambda_Z, 1) &= t^*(\lambda_Z) - \frac{\sqrt{n_A}(k_A/x(t^*(\lambda_Z), 1) + k_B)}{N\lambda_Z(k_A + k_B e^{-(k_A+k_B)m(t^*(\lambda_Z))})} \\ &= t^*(\lambda_Z) - \frac{\sqrt{(k_B n_A + k_A n_B t^*(\lambda_Z)^2)(k_A + k_B)}}{N\lambda_Z(k_A + k_B e^{-(k_A+k_B)m(t^*(\lambda_Z))})} > 0 \end{aligned}$$

which is guaranteed by the condition (5.21),

$$k_A + k_B e^{-(k_A+k_B)m(t^*(\lambda_Z))} > \frac{1}{N\lambda_Z}\sqrt{\left(\frac{k_B n_A}{t^*(\lambda_Z)^2} + k_A n_B\right)(k_A + k_B)}.$$

By continuity of $L(\lambda_Z, \tau)$ in τ, there exists a choice of τ such that the fourth equation holds. Therefore, the condition (5.21) guarantees a solution to system (5.48).

Note that Lemma 13 implies that the inequality above holds for any $\lambda_Z = \sqrt{n_A}/N - \delta$ with $\delta < \varepsilon$. Hence, for any λ_Z close to $\sqrt{n_A}/N$, the case (c) in Proposition 11 holds. ■

Proof of Proposition 11(d) We consider $N\lambda_Z \geq \sqrt{n_A}$.
Solution structure: In this case, we propose the solution structure:

$$\bar{\boldsymbol{Z}} = 0, \quad \bar{\boldsymbol{R}} = \mathcal{B}(a_R, b_R, c_R, d_R) \tag{5.49}$$

Aagin, we directly have $\bar{\boldsymbol{R}}\boldsymbol{D}\bar{\boldsymbol{Z}}^\top = \bar{\boldsymbol{Z}}^\top\bar{\boldsymbol{R}} = 0$ and we are left with verifying $\|\bar{\boldsymbol{R}}\boldsymbol{D}^{1/2}\| \leq 1$ through optimality condition.

Optimality system: The optimality condition is straightforward by (5.4) when $\boldsymbol{Z} = 0$. When $\boldsymbol{Z} = 0$, we have

$$\begin{aligned} &N^{-1}(\boldsymbol{I}_K - \bar{\boldsymbol{P}}) = \lambda_Z\bar{\boldsymbol{R}}, \\ &N^{-1}(\boldsymbol{I}_K - \bar{\boldsymbol{P}})\boldsymbol{n} = \lambda_b\boldsymbol{b}, \\ &\|\bar{\boldsymbol{R}}\boldsymbol{D}^{1/2}\| \leq 1. \end{aligned} \tag{5.50}$$

For ease of notation, we denote $u = e^{-mk_A}, v = e^{-mk_B}$ and $w = v/u$. The system (5.50) above reduces to

$$\begin{aligned} \text{eq.1 and 3 in (5.50)} \quad & \|(\boldsymbol{I}_K - \bar{\boldsymbol{P}})\boldsymbol{D}^{1/2}\| \leq N\lambda_Z, \\ \text{eq.2 (5.50)} \quad & \frac{n_A - n_B w}{k_B + k_A w} = N\lambda_b\frac{\log w}{k_A + k_B}, \\ & m = (k_A + k_B)^{-1}\log w. \end{aligned} \tag{5.51}$$

Now we proceed to find a solution to system (5.51).

Existence of a solution to (5.51)**:** We compute the SVD of $(\boldsymbol{I}_K - \bar{\boldsymbol{P}})\boldsymbol{D}^{1/2}$ directly,

$$\begin{aligned} (\boldsymbol{I}_K - \bar{\boldsymbol{P}})\boldsymbol{D}^{1/2} = &\begin{bmatrix} \sqrt{n_A}(\boldsymbol{I}_{k_A} - \boldsymbol{J}_{k_A}/k_A) & 0 \\ 0 & \sqrt{n_B}(\boldsymbol{I}_{k_B} - \boldsymbol{J}_{k_B}/k_B) \end{bmatrix} + \\ &\frac{1}{k_B u + k_A v}\begin{bmatrix} \frac{k_B\sqrt{n_A}u}{k_A}\boldsymbol{J}_{k_A} & -\sqrt{n_B}v\boldsymbol{J}_{k_A\times k_B} \\ -\sqrt{n_A}u\boldsymbol{J}_{k_B\times k_A} & \frac{k_A\sqrt{n_B}}{k_B}v\boldsymbol{J}_{k_B} \end{bmatrix} \end{aligned} \tag{5.52}$$

Hence we have,

$$(\boldsymbol{I}_K - \bar{\boldsymbol{P}})\boldsymbol{D}(\boldsymbol{I}_K - \bar{\boldsymbol{P}})^\top = \begin{bmatrix} n_A(\boldsymbol{I}_{k_A} - \boldsymbol{J}_{k_A}/k_A) & 0 \\ 0 & n_B(\boldsymbol{I}_{k_B} - \boldsymbol{J}_{k_B}/k_B) \end{bmatrix} + \frac{1}{(k_B u + k_A v)^2} \begin{bmatrix} (\frac{k_B^2 n_A}{k_A} u^2 + k_B n_B v^2)\boldsymbol{J}_{k_A} & -(k_B n_A u^2 + k_A n_B v^2)\boldsymbol{J}_{k_A \times k_B} \\ -(k_B n_A u^2 + k_A n_B v^2)\boldsymbol{J}_{k_B \times k_A} & (\frac{k_A^2 n_B}{k_B} v^2 + k_A n_A u^2)\boldsymbol{J}_{k_B} \end{bmatrix} \tag{5.53}$$

Therefore, the singular values of $(\boldsymbol{I}_K - \bar{\boldsymbol{P}})\boldsymbol{D}^{1/2}$ are $\sqrt{n_A}$ with multiplicity $k_A - 1$, $\sqrt{n_B}$ with multiplicity $k_B - 1$, $\sqrt{(k_A + k_B)(k_B n_A u^2 + k_A n_B v^2)}/(k_B u + k_A v)$ with multiplicity 1, and 0 with multiplicity 1. It suffices to show the maximum singular value is given by $\sqrt{n_A}$. Hence eq.1 in (5.51) is satisfied when $\lambda_Z \geq \sqrt{n_A}/N$. We only need to look into the squared singular value

$$\sigma(w) := \frac{(k_A + k_B)(k_B n_A u^2 + k_A n_B v^2)}{(k_B u + k_A v)^2} = \frac{(k_A + k_B)(k_B n_A + k_A n_B w^2)}{(k_B + k_A w)^2}.$$

The objective is then to show the existence of w as the solution of eq.2 in (5.51) and $\sigma(w) \leq n_A$ satisfies for that w. The idea is to constrain the range of w and bound it by the monotonicity of $\sigma(w)$.

We claim $w \in [1, n_A/n_B]$ by checking two ends of eq.2 in (5.51). On the one hand, the LHS is greater than 0 iff $w < n_A/n_B$ and the RHS is greater than 0 iff $w > 1$. On the other hand, on $w \in [1, n_A/n_B]$, the LHS (RHS) strictly decreases (increases) in w and at $w = 1$ (n_A/n_B), we have LHS>RHS (LHS<RHS). These allow us to conclude that there exists a unique $w \in [1, n_A/n_B]$ that satisfies eq.2 of (5.51).

Now we check the monotonicity of $\sigma(w)$ by computing its derivative:

$$\begin{aligned} \sigma'(w) &= 2(k_A + k_B)\frac{k_A n_B w(k_B + k_A w) - k_A(k_B n_A + k_A n_B w^2)}{(k_B + k_A w)^3} \\ &= 2k_A k_B(k_A + k_B)\frac{n_B w - n_A}{(k_B + k_A w)^3} \leq 0, \text{ for } w \in \left[1, \frac{n_A}{n_B}\right]. \end{aligned}$$

Therefore when $w = 1$, we obtain an upper bound of this singular value,

$$\sigma(1) = \frac{k_B n_A + k_A n_B}{k_A + k_B} \leq n_A,$$

which implies the largest singular value of $(\boldsymbol{I}_K - \bar{\boldsymbol{P}})\boldsymbol{D}^{1/2}$ is no larger than $\sqrt{n_A}$. This verifies eq.1 of (5.51). ■

Proof of Proposition 11(e) If $\lambda_b = \infty$, then (5.21) and (5.22) are equivalent to $\xi(\lambda_Z, \infty) < 0$ and $\xi(\lambda_Z, \infty) > 0$ in (5.26). Lemma 13(d) implies that $\xi(\lambda_Z, \infty)$ is increasing in λ_Z with $\xi(\sqrt{n_B}/N, \infty) < 0$ and $\xi(\sqrt{n_A}/N, \infty) > 0$. Therefore, there exists a λ^* which is the root to $\xi(\lambda_Z, \infty)$, such that (5.21) and (5.22) are equivalent to $\lambda_Z < \lambda^*$ and $\lambda_Z > \lambda^*$ respectively. ■

5.5 Limiting case: Proof of Corollary 4 and Theorem 5

In this section, we give some asymptotic characterization of $\bar{\boldsymbol{Z}}$, when either n_A or n_B, or both go to infinity.

Proof of Corollary 4 From Proposition 11, the minority collapse occurs when $N\lambda_Z \geq \sqrt{n_B}$. Plugging in $N = k_A n_A + k_B n_B = n_B(k_A r + k_B)$ yields

$$n_B(k_A r + k_B)\lambda_Z \geq \sqrt{n_B} \Longleftrightarrow r \geq \frac{1}{k_A}\left(\frac{1}{\sqrt{n_B}\lambda_Z} - k_B\right).$$

■

Proof of Theorem 5 Under the assumption $N\lambda_Z < \sqrt{n_B}$, the mean feature matrix $\bar{\boldsymbol{Z}}$ falls in the case (a) of Theorem 3. The key is to show that the unique solution t_N^* of (5.36) (i.e., the reduced optimality condition for case (a)) converges to 1 at the rate of $1/\log N$ as $N \to \infty$. For simplicity, we define

$$f_N(t) = \frac{g_1(x_1(t)) + (k_A + k_B)k_B m(t)}{g_2(x_2(t)) - (k_A + k_B)k_A m(t)}$$

indexed by the total sample size. Let $h_N(t) = f_N(t) - t$ and t_N be the zero of $h_N(t)$. From the previous analysis, we know t_N is unique and $h_N(t)$ is monotonically decreasing on $I_{f_N}^+ := \{t : f_N(t) > 0\}$. Moreover, it holds $f_N'(t) < 0$ on $I_{f_N}^+$ and as a result,

$$h_N'(t) = f_N'(t) - 1 < -1.$$

This implies that

$$\left|h_N(t) - h_N(t')\right| \geq \left|t - t'\right| \quad \text{for any } t, t' \in I_{f_N}^+.$$

From the following argument, it is straightforward to see that for sufficiently large N, $g_1(x_1(1))$ and $g_2(x_2(2))$ are both positive and dominate $m(1)$, which implies $1 \in I_{f_N}^+$. So we can obtain the following bound:

$$\begin{aligned}|t_N - 1| &\leq |h_N(t_N) - h_N(1)| = |h_N(1)| \\ &= \left|\frac{\log\left[\left(\frac{\sqrt{n_B}}{\lambda} - 1\right)(k_A x_1(1) + k_B) + 1\right] - k_B \log x_1(1) + (k_A + k_B)k_B m(1)}{\log\left[\left(\frac{\sqrt{n_A}}{\lambda} - 1\right)(k_A + k_B x_2(1)) + 1\right] - k_A \log x_2(1) - (k_A + k_B)k_A m(1)} - 1\right|,\end{aligned}$$

where $x_1(t)$ and $x_2(t)$ are defined in (5.33), satisfying

$$x_1(1) = \frac{k_B}{\sqrt{(rk_B + k_A)(k_A + k_B)} - k_A}, \quad x_2(1) = \frac{k_A}{\sqrt{(k_B + k_A/r)(k_A + k_B)} - k_B}. \tag{5.54}$$

It is easy to see $x_1(t)$ and $x_2(t)$ stay bounded for fixed r. As $N \to \infty$, we notice both n_A and n_B go to ∞, and also

$$m(1) \lesssim \frac{\lambda}{\sqrt{N}\lambda_b} = o\left(\log(N)\right)$$

follows from the assumption on the decay rate of λ_b. Thus sending $N \to \infty$ implies

$$|t_N - 1| \lesssim \left| \frac{\log \sqrt{n_B}/\lambda + o(\log N)}{\log \sqrt{n_A}/\lambda + o(\log N)} - 1 \right| = \frac{\log \sqrt{r}}{\log \sqrt{n_A}/\lambda} = O\left(\frac{1}{\log N}\right)$$

for sufficiently large N.

From (5.32), we know that

$$\begin{aligned} b_N &= \log\left[\left(\frac{\sqrt{n_B}}{\lambda} - 1\right)(k_A x_1(t_N) + k_B) + 1\right] - k_B \log x_1(t_N) + o(\log N) \\ &\geq \log \frac{\sqrt{n_B}}{\lambda} - k_B \log x_1(t_N) + o(\log N). \end{aligned}$$

Therefore, b_N is at least $\log(\sqrt{n_B}/\lambda)$ and similarly $c_N \geq \log(\sqrt{n_A}/\lambda)$. The uniform boundedness of x_1 and x_2 in (5.33) implies that a_N and c_N, and b_N and d_N grow at the same rate, i.e.,

$$\lim_{N\to\infty} b_N = O(\log N) = \infty, \quad \lim_{N\to\infty} c_N = O(\log N) = \infty$$

and

$$\lim_{N\to\infty} t_N = \lim_{N\to\infty} \frac{b_N}{c_N} = 1, \quad \lim_{N\to\infty} \frac{a_N}{c_N} = 1, \quad \lim_{N\to\infty} \frac{d_N}{b_N} = 1.$$

In other words, as $N \to \infty$, we have

$$\lim_{N\to\infty} \frac{1}{b_N} \bar{\boldsymbol{Z}} = (k_A + k_B)\boldsymbol{I}_{k_A+k_B} - \boldsymbol{J}_{k_A+k_B}$$

which implies the column normalized $\bar{\boldsymbol{Z}}$ in this limit should converge to the ETF, so do $\bar{\boldsymbol{H}}$ and $\boldsymbol{W}$. ■

References

Tina Behnia, Ganesh Ramachandra Kini, Vala Vakilian, and Christos Thrampoulidis. On the implicit geometry of cross-entropy parameterizations for label-imbalanced data. In *International Conference on Artificial Intelligence and Statistics*, pages 10815–10838. PMLR, 2023.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019a.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019b.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Hien Dang, Tan Nguyen, Tho Tran, Hung Tran, and Nhat Ho. Neural collapse in deep linear network: From balanced to imbalanced data. *arXiv preprint arXiv:2301.00437*, 2023.

Weinan E and Stephan Wojtowytsch. On the emergence of simplex symmetry in the final and penultimate layers of neural network classifiers. In *Mathematical and Scientific Machine Learning*, pages 270–290. PMLR, 2022.

Michael Elad, Dror Simon, and Aviad Aberdam. Another step toward demystifying deep neural networks. *Proceedings of the National Academy of Sciences*, 117(44):27070–27072, 2020.

Tolga Ergen and Mert Pilanci. Revealing the structure of deep neural networks via convex duality. In *International Conference on Machine Learning*, pages 3004–3014. PMLR, 2021.

Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021.

Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. *arXiv preprint arXiv:2112.15121*, 2021.

Tomer Galanti, András György, and Marcus Hutter. Generalization bounds for transfer learning with pretrained classifiers. *arXiv preprint arXiv:2212.12532*, 2022.

XY Han, Vardan Papyan, and David L Donoho. Neural collapse under MSE loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Like Hui, Mikhail Belkin, and Preetum Nakkiran. Limitations of neural collapse for understanding generalization in deep learning. *arXiv preprint arXiv:2202.08384*, 2022.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Wenlong Ji, Yiping Lu, Yiliang Zhang, Zhun Deng, and Weijie J Su. An unconstrained layer-peeled perspective on neural collapse. *arXiv preprint arXiv:2110.02796*, 2021.

Vignesh Kothapalli, Ebrahim Rasromani, and Vasudev Awatramani. Neural collapse: A review on modelling principles and generalization. *arXiv preprint arXiv:2206.04041*, 2022.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015.

Xiao Li, Sheng Liu, Jinxin Zhou, Xinyu Lu, Carlos Fernandez-Granda, Zhihui Zhu, and Qing Qu. Principled and efficient transfer learning of deep models via neural collapse. *arXiv preprint arXiv:2212.12206*, 2022.

Jianfeng Lu and Stefan Steinerberger. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241, 2022.

Dustin G Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *arXiv preprint arXiv:2011.11619*, 2020.

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

Tomaso Poggio and Qianli Liao. Explicit regularization and implicit bias in deep network classifiers trained with the square loss. *arXiv preprint arXiv:2101.00072*, 2021.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

Mariia Seleznova, Dana Weitzner, Raja Giryes, Gitta Kutyniok, and Hung-Hsu Chou. Neural (tangent kernel) collapse. *arXiv preprint arXiv:2305.16427*, 2023.

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

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

Christos Thrampoulidis, Ganesh Ramachandra Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. *Advances in Neural Information Processing Systems*, 35:27225–27238, 2022.

Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse. In *International Conference on Machine Learning*, pages 21478–21505. PMLR, 2022.

Tom Tirer, Haoxiang Huang, and Jonathan Niles-Weed. Perturbation analysis of neural collapse. In *International Conference on Machine Learning*, pages 34301–34329. PMLR, 2023.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (itw)*, pages 1–5. IEEE, 2015.

Yibo Yang, Liang Xie, Shixiang Chen, Xiangtai Li, Zhouchen Lin, and Dacheng Tao. Do we really need a learnable classifier at the end of deep neural network? *arXiv e-prints*, pages arXiv–2203, 2022.

Can Yaras, Peng Wang, Zhihui Zhu, Laura Balzano, and Qing Qu. Neural collapse with normalized features: A geometric analysis over the riemannian manifold. *Advances in Neural Information Processing Systems*, 35:11547–11560, 2022.

Jinxin Zhou, Xiao Li, Tianyu Ding, Chong You, Qing Qu, and Zhihui Zhu. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. In *International Conference on Machine Learning*, pages 27179–27202. PMLR, 2022a.

Jinxin Zhou, Chong You, Xiao Li, Kangning Liu, Sheng Liu, Qing Qu, and Zhihui Zhu. Are all losses created equal: A neural collapse perspective. *Advances in Neural Information Processing Systems*, 35:31697–31710, 2022b.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021.