Transformers Can Learn Multiclass Classification In-Context: Isotropy Governs Generalization

Daehan Yoon AXIOM0102@KAIST.AC.KR
Kim Jaechul Graduate School of AI, KAIST, Seoul, Republic of Korea

Chulhee Yun CHULHEE.YUN@KAIST.AC.KR
Kim Jaechul Graduate School of AI, KAIST, Seoul, Republic of Korea

Abstract

In-context learning plays a central role in transformer-based large language models, yet its theoretical understanding remains limited. In this work, we study multiclass in-context classification under more realistic settings, including anisotropic class centers and label imbalance, by introducing a spectral data generation framework that constructs class-center matrices with a prescribed singular spectrum. We first show that the isotropy of class-center vectors, as quantified by the stable rank, improves ICL generalization performance. From a meta-learning perspective, our theorem shows that linear transformers can learn multiclass in-context classification with near-optimal per-label sample complexity, extending prior guarantees beyond the binary setting. In the test label imbalance regime, our analysis reveals that queries from majority classes are easier to classify, while those from minority classes are more error-prone; moreover, robustness to this bias improves with the stable rank. Finally, we empirically demonstrate that our theory is consistent with observations on transformers and pretrained large language models.

1. Introduction

Transformers [25] have introduced a new paradigm across a broad range of machine learning domains. In particular, the in-context learning (**ICL**) phenomenon [5, 7], the ability to make predictions for a given query by conditioning on a small set of input–output examples provided in the prompt, has been empirically demonstrated in transformer-based large language models (LLMs). Recent studies have made substantial progress toward a theoretical understanding of ICL, with several works focusing on linear regression settings [1, 16, 27, 28]. However, since ICL can be viewed as a special case of next-token prediction, it is more naturally aligned with classification tasks. (More details in these related works is provided in Appendix B).

Our Contributions. We study generalization guarantees in multiclass in-context classification. We introduce a spectral data-generating procedure which captures anisotropic class-center vectors as well as imbalanced label distributions. Employing the linear attention model used in prior work, we analyze the induced objective, characterize the max-margin limit selected by gradient-based training, and derive generalization bounds that explicitly account for both test label imbalance and class-center geometry.

- We demonstrate that the isotropy of class-center vectors, as quantified by the stable rank, improves ICL performance for multiclass classification. We further show that transformers achieve near-optimal sample complexity in the balanced-label setting (Theorem 5).

© D. Yoon & C. Yun.

- Under test label imbalance, we show that majority-class queries are easier to classify, whereas minority-class queries are more error-prone. Our theorem also demonstrates that robustness to this majority label bias is governed by spectral structure, with higher stable rank of the class-center matrix leading to stronger robustness (Theorem 6).
- We empirically observe that the implications from our theory aligns with softmax transformers and pretrained LLMs (Appendix I).

2. Preliminaries

In this section, we formalize the model, data distribution, and the assumptions.
Notation. We use $\mathrm{diag}(d_1, \ldots, d_{m\wedge n})$ to denote the (possibly rectangular) matrix $\boldsymbol{D} \in \mathbb{R}^{m\times n}$ whose entries are $\boldsymbol{D}_{ii} = d_i$ for all $i \le m \wedge n$ and $\boldsymbol{D}_{ij} = 0$ for $i \neq j$. The remaining notation follows standard conventions and is summarized in Appendix A.

2.1. In-context Multiclass Classification

We consider in-context multiclass classification which aims to predict the label $\boldsymbol{y}^{(n+1)}$ of a *query* input $\boldsymbol{x}^{(n+1)}$ given n labeled *examples* $(\boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)})_{j=1}^{n}$. Here, $\boldsymbol{y}^{(j)} \in \mathbb{R}^K$ denotes the one-hot vector representing the class label of $\boldsymbol{x}^{(j)} \in \mathbb{R}^d$, where K is the number of classes. The model utilizes these examples through a prompt, and we encode the data in the following form.

$$\boldsymbol{E} = \begin{bmatrix} \boldsymbol{x}^{(1)} & \cdots & \boldsymbol{x}^{(n)} & \boldsymbol{x}^{(n+1)} \\ \boldsymbol{y}^{(1)} & \cdots & \boldsymbol{y}^{(n)} & \boldsymbol{0} \end{bmatrix} \in \mathbb{R}^{(d+K)\times(n+1)}.$$

Given the above data encoding, we now introduce the model. We model in-context learning using a transformer architecture, where self-attention plays a central role. The standard self-attention [25] is defined as

$$\mathrm{Attn}_\theta(\boldsymbol{E}) = \boldsymbol{W}^V \boldsymbol{E}\, \mathrm{Softmax}\left(\tfrac{1}{\sqrt{d}} \boldsymbol{E}^\top (\boldsymbol{W}^K)^\top \boldsymbol{W}^Q \boldsymbol{E}\right).$$

We consider a linearized version of self-attention, called *Linear Self-Attention*, which is defined as

$$\mathrm{LinAttn}_\theta(\boldsymbol{E}) = \boldsymbol{W}^V \boldsymbol{E} \boldsymbol{M}_\star \boldsymbol{E}^\top \boldsymbol{W}^{KQ} \boldsymbol{E}, \boldsymbol{M}_\star := \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{W}^{KQ} = (\boldsymbol{W}^K)^\top \boldsymbol{W}^Q.$$

Here, $\boldsymbol{M}_\star$ is a masking matrix that restricts attention to the context examples and prevents the query token from attending to itself. Although this masking does not appear in the original transformer formulation, it is commonly adopted in theoretical analyses [1, 27] of in-context learning. We directly train the merged matrix $\boldsymbol{W}^{KQ}$, rather than learning $\boldsymbol{W}^K$ and $\boldsymbol{W}^Q$ separately. Following common practice in theoretical studies of transformers, we analyze a linear attention model as a representative tractable model for understanding generalization of transformers.[1] With a residual connection, the prediction by the one-layer Transformer for the query token reads

$$\hat{\boldsymbol{y}} := \left(\boldsymbol{E} + \tfrac{1}{n}\mathrm{LinAttn}_\theta(\boldsymbol{E})\right)_{(d+1):(d+K),(n+1)}.$$

1. As presented in Appendix I, we empirically evaluate a transformer architecture composed of a multi-head softmax attention module and a feedforward network (FFN), and observe that the results are consistent with our theory.

We adopt a parametrization similar to prior works [1, 22]. Specifically,

$$\boldsymbol{W}^V := \begin{bmatrix} * \\ \boldsymbol{W}_\star^V \end{bmatrix}, \ \boldsymbol{W}^{KQ} := \begin{bmatrix} \boldsymbol{W} & * \\ \boldsymbol{0} & * \end{bmatrix}, \quad \boldsymbol{W}_\star^V \in \mathbb{R}^{K\times(d+K)}, \ \boldsymbol{W} \in \mathbb{R}^{d\times d}.$$

The blocks marked with $*$ do not affect the model's predictions. We fix $\boldsymbol{W}_\star^V = \begin{bmatrix} 0_{K\times d} & \boldsymbol{I}_K \end{bmatrix}$, and only optimize over $\boldsymbol{W}$. Then, the prediction reads

$$\hat{\boldsymbol{y}}(\boldsymbol{E};\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}^{(i)}\boldsymbol{x}^{(i)\top}\boldsymbol{W}\boldsymbol{x}^{(n+1)} = \frac{1}{n}\begin{bmatrix} \sum_{j\in\mathcal{C}_1} \boldsymbol{x}^{(j)\top}\boldsymbol{W}\boldsymbol{x}^{(n+1)} \\ \vdots \\ \sum_{j\in\mathcal{C}_K} \boldsymbol{x}^{(j)\top}\boldsymbol{W}\boldsymbol{x}^{(n+1)} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{\mu}}_1^\top\boldsymbol{W}\boldsymbol{x}^{(n+1)} \\ \vdots \\ \hat{\boldsymbol{\mu}}_K^\top\boldsymbol{W}\boldsymbol{x}^{(n+1)} \end{bmatrix},$$

where $\mathcal{C}_k := \{j \in [n] : \boldsymbol{y}^{(j)} = \boldsymbol{e}_k\}$ denotes the index set of examples belonging to class k, and $\hat{\boldsymbol{\mu}}_k := \frac{1}{n}\sum_{j\in\mathcal{C}_k} \boldsymbol{x}^{(j)}$ represents the signal vectors of the inputs in class k.

2.2. Data Generation and Training Process

We define the following distribution for generating multi-class ICL tasks.

Definition 1 *For a given spectrum $\boldsymbol{D} = \mathrm{diag}(d_1, \dots d_K) \in \mathbb{R}^{K\times d}$ with $K \le d$, a label allocation $\{S_k\}_{k=1}^K \in \mathbb{N}^K$ with $\sum_{k=1}^K S_k = n$, and a query class $k^\star \in [K]$, the distribution over ICL tasks $\mathcal{D}\left(\boldsymbol{D}, \{S_k\}_{k=1}^K, k^\star\right)$ is defined as follows:*

1. *There are S_k context examples belonging to class k. In other words, letting $\boldsymbol{y}^{(j)} \in \mathbb{R}^K$ be the one-hot label, there are exactly S_k indices $j \in [n]$ such that $\boldsymbol{y}^{(j)} = \boldsymbol{e}_k$. Note that the query label $\boldsymbol{y}^{(n+1)} = \boldsymbol{e}_{k^\star}$ is independent of this class allocation.*

2. *Construct a class-center matrix $\boldsymbol{M} \in \mathbb{R}^{K\times d}$ whose rows correspond to class centers:*

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top,$$

   where sample the left and right singular vector matrices $\boldsymbol{U} \in \mathbb{R}^{K\times K}$, $\boldsymbol{V} \in \mathbb{R}^{d\times d}$ uniformly at random from the set of orthogonal matrices of the corresponding dimensions. We denote by $\boldsymbol{\mu}_k^\top$ the k-th row of $\boldsymbol{M}$, corresponding to the center of class k.

3. *Given label $\boldsymbol{y}^{(j)}$, generate the corresponding input vector according to*

$$\boldsymbol{x}^{(j)} = \boldsymbol{M}^\top\boldsymbol{y}^{(j)} + \boldsymbol{z}^{(j)} = \boldsymbol{\mu}_{k(j)} + \boldsymbol{z}^{(j)},$$

   where the noise vectors $\boldsymbol{z}^{(j)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ are independently sampled across j, and $k(j)$ denotes the class corresponding to the j-th sample.

Our data generation process is motivated by the prior works [6, 15] and can be viewed as a natural extension from binary to multiclass settings. More details are provided in Appendix B.1. For fixed $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \dots, \sigma_K) \in \mathbb{R}^{K\times d}$ with $\sigma_1 \ge \cdots \ge \sigma_K \ge 0$, we consider the balanced case $S_k = S_\star$ for all $k \in [K]$. For each $\tau \in [B]$, we sample a task $(\boldsymbol{x}_\tau^{(j)}, \boldsymbol{y}_\tau^{(j)})_{j=1}^{N+1} \sim \mathcal{D}\left(\boldsymbol{\Sigma}, \{S_\star\}^K, k_\tau^\star\right)$ in an i.i.d. manner, where each $k_\tau^\star$ is independently chosen arbitrarily across tasks. We denote by N the total number of examples for training tasks, so that $N = S_\star K$. We use the subscript τ to

denote quantities associated with the task indexed by τ. Therefore, $\boldsymbol{y}_\tau^{(N+1)}$ denotes the true label of the query token in task τ, $\hat{\boldsymbol{y}}_\tau$ denotes the corresponding prediction as defined in Section 2.1. Given sampled tasks $\{(\boldsymbol{x}_\tau^{(j)}, \boldsymbol{y}_\tau^{(j)})_{j=1}^{N+1}\}_{\tau=1}^{B}$, we define the loss function as:

$$\mathcal{L}(\theta) = \frac{1}{B}\sum\nolimits_{\tau=1}^{B} \ell\big(\boldsymbol{y}_\tau^{(N+1)}, \hat{\boldsymbol{y}}_\tau\big),$$

where $\ell(\cdot)$ denotes the standard cross-entropy loss applied to the softmax of the logits in $\hat{\boldsymbol{y}}_\tau$, as in conventional multiclass classification. After training, we sample a test task from $\mathcal{D}\left(\boldsymbol{\Lambda}, \{S_k\}_{k=1}^{K}, k^\star\right)$, where the spectrum $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_K) \in \mathbb{R}^{K\times d}$ with $\lambda_1 \geq \cdots \geq \lambda_K \geq 0$, which may differ from $\boldsymbol{\Sigma}$. In the test task, the label allocation $\{S_k\}_{k=1}^{K}$ is not necessarily balanced, and we denote the total number of examples by $M := \sum_{k=1}^{K} S_k$ (which may differ from N). We make predictions for a test task with M examples, as $\hat{\boldsymbol{y}} = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{y}^{(i)}\boldsymbol{x}^{(i)\top}\boldsymbol{W}\boldsymbol{x}^{(M+1)}$, using the hard-max output.

2.3. Assumptions

We impose the following assumptions:

Assumption 2 *There exists* $0 < c_s \leq 1$ *such that* $\mathrm{sr}(\boldsymbol{\Sigma}) \geq c_s K$ *(i.e.* $\mathrm{sr}(\boldsymbol{\Sigma}) = \Theta(K)$*).*

Assumption 3 *The first singular value of training tasks is sufficiently large:* $\sigma_1^2 \geq c_\sigma d$ *for some absolute constant* $c_\sigma > 0$.

Assumption 4 *The ambient dimension* d *and the number of the classes* K *should satisfy* $d \geq K \geq 3 \vee C\log^2\left(B^2K^2/\delta\right)$ *for some absolute constant* $C > 0$.

The first two assumptions ensure non-degeneracy of the training set and are not imposed on the test spectrum $\boldsymbol{\Lambda}$. Assumption 2 and 4 guarantee that the training tasks are realizable (Assumption 9) with high probability, enabling the implicit bias induced by training. The data generation assumptions of Shen et al. [22] satisfy Assumptions 2 and 3 which are related to the data geometry.

3. Main Results

In this section, we characterize the implicit bias induced by the training tasks and leverage it to analyze the generalization behavior of transformers in in-context learning.

3.1. Implicit Bias on Pre-trained Transformers

We study the implicit bias of Linear Self-Attention trained via gradient descent. Under Assumptions 9 and 10, standard in the implicit bias literature [21, 23], and a sufficiently small step size η, the solution converges in direction to the max-margin solution

$$\boldsymbol{W}_{MM} := \arg\min \|\boldsymbol{W}\|_F^2, \quad \text{s.t. } (\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k})^\top \boldsymbol{W}\boldsymbol{x}_\tau^{(N+1)} \geq 1, \quad \forall\tau,\ k \neq k_\tau^\star,$$

where $\hat{\boldsymbol{\mu}}_{\tau,k} = \frac{1}{N}\sum_{j\in\mathcal{C}_{\tau,k}} \boldsymbol{x}_\tau^{(j)}$ denotes the signal vector for class k. More details can be found in Appendix D, which presents a rigorous analysis of the gradient descent dynamics and establishes the implicit max-margin bias of the attention parameters.

3.2. Transformers Can Learn Multiclass Classification In-Context via Stable Rank

In this section, we study multiclass in-context classification with transformers under a balanced test setting, where each class has exactly S examples. We also use balanced prompts during training. We adopt a hard-max prediction rule and denote the resulting max-margin classifier by $\boldsymbol{W}$. Defining $\hat{k} := \arg\max_k \hat{\boldsymbol{y}}(\boldsymbol{E}; \boldsymbol{W})$, the classification error satisfies

$$\mathbb{P}\left(\hat{k} \neq k^\star\right) = \mathbb{P}\left(\exists k \neq k^\star \text{ s.t. } \hat{\boldsymbol{\mu}}_k^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)} > \hat{\boldsymbol{\mu}}_{k^\star}^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)}\right)$$
$$= \mathbb{P}\Big(\bigcup_{k\neq k^\star} \{\hat{\boldsymbol{\mu}}_k^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)} > \hat{\boldsymbol{\mu}}_{k^\star}^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)}\}\Big) \leq \sum_{k\neq k^\star} \mathbb{P}\left(\hat{\boldsymbol{\mu}}_k^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)} > \hat{\boldsymbol{\mu}}_{k^\star}^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)}\right).$$

We bound the generalization error by controlling the misclassification probability for each class k.

Theorem 5 *Fix $k \neq k^\star$, and let $\delta \in (0, \frac{1}{6})$ be arbitrary. Suppose Assumptions 2, 3, 4, and 10 hold. Let $\{(\boldsymbol{x}_\tau^{(j)}, \boldsymbol{y}_\tau^{(j)})_{j=1}^{N+1}\}_{\tau=1}^B$ be B training tasks with $(\boldsymbol{x}_\tau^{(j)}, \boldsymbol{y}_\tau^{(j)})_{j=1}^{N+1} \sim \mathcal{D}\left(\boldsymbol{\Sigma}, \{S_\star\}^K, k_\tau^\star\right)$ for each $\tau \in [B]$. The max-margin solution $\boldsymbol{W}$ satisfies the following with probability at least $1 - 6\delta$: when sampling a new task with balanced prompt $(\boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)})_{j=1}^{M+1} \sim \mathcal{D}\left(\boldsymbol{\Lambda}, \{S\}^K, k^\star\right)$, we have*

$$\mathbb{P}\left(\hat{\boldsymbol{\mu}}_k^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)} > \hat{\boldsymbol{\mu}}_{k^\star}^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)}\right)$$
$$\leq C' \exp\left(-c\left(\frac{d\,\mathrm{sr}(\boldsymbol{\Lambda})}{\xi^2 K} \wedge \frac{\sqrt{d}}{\xi} \wedge \mathrm{sr}(\boldsymbol{\Lambda})\right) \cdot \frac{\mathrm{sr}(\boldsymbol{\Lambda})}{K}\right) + 6\exp\left(-c\left(\frac{\lambda_1\,\mathrm{sr}(\boldsymbol{\Lambda})}{\xi K} \wedge \sqrt{\frac{S\|\boldsymbol{\Lambda}\|_F^4}{d\xi^2 K^2}}\right)\right)$$

for some $C', c > 0$, where $\xi = O(\frac{d}{B} \vee \log^2\left(B^2K^2/\delta\right))$.

Theorem 5 provides a bound on the probability of misclassifying a sample into a fixed class k. The bound decreases as the stable rank $\mathrm{sr}(\boldsymbol{\Lambda})$ increases, indicating that more isotropic and well-separated class geometry leads to better generalization performance. Under suitable conditions, we show that transformers achieve near-optimal sample complexity for multiclass in-context learning. We provide further discussion in Appendix G.1.

3.3. Multiclass In-Context Classification under Test Label Imbalance

We study test-time label imbalance, where different numbers of examples are sampled for each class while training uses balanced prompts. Unlike prior work [6, 15, 22], our bound depends on the test label counts and captures phenomena such as majority-label bias.

Theorem 6 *Fix $k \neq k^\star$, and let $\delta \in (0, \frac{1}{6})$ be arbitrary. Suppose Assumptions 2, 3, 4, and 10 hold. Let $\{(\boldsymbol{x}_\tau^{(j)}, \boldsymbol{y}_\tau^{(j)})_{j=1}^{N+1}\}_{\tau=1}^B$ be B training tasks with $(\boldsymbol{x}_\tau^{(j)}, \boldsymbol{y}_\tau^{(j)})_{j=1}^{N+1} \sim \mathcal{D}\left(\boldsymbol{\Sigma}, \{S_\star\}^K, k_\tau^\star\right)$ for each $\tau \in [B]$. The max-margin solution $\boldsymbol{W}$ satisfies the following with probability at least $1 - 6\delta$: when sampling a new task with **imbalanced** prompt $(\boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)})_{j=1}^{M+1} \sim \mathcal{D}\left(\boldsymbol{\Lambda}, \{S_k\}_{k=1}^K, k^\star\right)$, we have*

$$\mathbb{P}\left(\hat{\boldsymbol{\mu}}_k^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)} > \hat{\boldsymbol{\mu}}_{k^\star}^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)}\right)$$
$$\leq C' \exp\left(-c\left(\frac{d\psi_{k,k^\star}\,\mathrm{sr}(\boldsymbol{\Lambda})}{\xi^2 K} \wedge \frac{\sqrt{d}}{\xi} \wedge \psi_{k,k^\star}\,\mathrm{sr}(\boldsymbol{\Lambda})\right) \cdot \frac{\psi_{k,k^\star}\,\mathrm{sr}(\boldsymbol{\Lambda})}{K}\right)$$
$$+ 6\exp\left(-c\left(\frac{\lambda_1 \psi_{k,k^\star}\,\mathrm{sr}(\boldsymbol{\Lambda})}{\xi K} \wedge \frac{S_{k^\star}}{\sqrt{S_k + S_{k^\star}}} \cdot \frac{\|\boldsymbol{\Lambda}\|_F^2}{\sqrt{d}\xi K}\right)\right)$$

for some $C', c > 0$, where $\xi = O(\frac{d}{B} \vee \log^2\left(B^2K^2/\delta\right))$, $\psi_{k,k^\star} := S_{k^\star}/\sqrt{S_k^2 + S_{k^\star}^2}$.

The imbalance factor $\psi_{k,k^\star}$ quantifies the effect of test-time label imbalance, where a smaller value—corresponding to class k being dominant—leads to a larger error bound and thus stronger majority-label bias toward class k. Importantly, this effect is mitigated by the stable rank $\mathrm{sr}(\mathbf{\Lambda})$, suggesting that higher-dimensional and more isotropic class-center geometry provides robustness against label imbalance. We provide further discussion in Appendix H.1.

4. Conclusion and Future Direction

In this work, we mainly focused on the in-context learning (ICL) capability of Transformers for multi-class classification. We showed that, under gradient descent, the implicit bias of the model enables ICL to generalize effectively. In balanced training settings, the model achieves near-optimal sample complexity, and in particular, the stable rank of the class center matrix plays a critical role in determining the generalization bound. We further analyzed both theoretically and empirically that this stable rank becomes even more important under label imbalance, and that test-time label imbalance significantly affects ICL performance.

One limitation of our analysis is that we rely on the max-margin solution, which can be viewed as the limiting direction of the training dynamics, and we do not explicitly consider the training procedure under a finite-time training budget. A natural direction for future work is to extend this analysis to more complex transformer architectures, including nonlinear layers and multi-head attention, where characterizing the role of representation geometry under label imbalance remains an open problem.

References

[1] Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=LziniAXEI9.

[2] Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=0uI5415ry7.

[3] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models, 2023. URL https://arxiv.org/abs/2211.15661.

[4] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=liMSqUuVg9.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin,

Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[6] Spencer Frei and Gal Vardi. Trained transformer classifiers generalize and exhibit benign overfitting in-context. In *The Thirteenth International Conference on Learning Representations*, 2024.

[7] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 30583–30598. Curran Associates, Inc., 2022.

[8] Angeliki Giannou, Liu Yang, Tianhao Wang, Dimitris Papailiopoulos, and Jason D. Lee. How well can transformers emulate in-context newton's method?, 2024. URL https://arxiv.org/abs/2403.03183.

[9] Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed k-means. *Mathematical Statistics and Learning*, 1(3):317–374, 2019.

[10] Karan Gupta, Sumegh Roychowdhury, Siva Rajesh Kasa, Santhosh Kumar Kasa, Anish Bhanushali, Nikhil Pattisapu, and Prasanna Srinivasa Murthy. How robust are llms to in-context majority label bias?, 2023.

[11] Juno Kim, Tai Nakamaki, and Taiji Suzuki. Transformers are minimax optimal nonparametric in-context learners. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=hF6vatntqc.

[12] Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. How do nonlinear transformers learn and generalize in in-context learning? In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=I4HTPws9P6.

[13] Zihao Li, Yuan Cao, Cheng Gao, Yihan He, Han Liu, Jason Matthew Klusowski, Jianqing Fan, and Mengdi Wang. One-layer transformer provably learns one-nearest neighbor in context. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=WDX45LNZXE.

[14] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

[15] Roey Magen and Gal Vardi. Transformers are almost optimal metalearners for linear classification. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[16] Arvind V. Mahankali, Tatsunori Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8p3fu56lKc.

[17] Elizabeth S. Meckes. *The Random Matrix Theory of the Classical Compact Groups*, volume 218 of *Cambridge Tracts in Mathematics*. Cambridge University Press, 2019. ISBN 9781108419529.

[18] George A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995. ISSN 0001-0782. doi: 10.1145/219717.219748. URL https://doi.org/10.1145/219717.219748.

[19] Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018.

[20] Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XJk19XzGq2J.

[21] Hrithik Ravi, Clayton Scott, Daniel Soudry, and Yutong Wang. The implicit bias of gradient descent on separable multiclass data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[22] Wei Shen, Ruida Zhou, Jing Yang, and Cong Shen. On the training convergence of transformers for in-context classification of gaussian mixtures. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=iSyB2yYaMx.

[23] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

[24] Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10434–10443. PMLR, 18–24 Jul 2021.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[26] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

[27] Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *ICML*, pages 35151–35174, 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

[28] Ruiqi Zhang, Spencer Frei, and Peter Bartlett. Trained transformers learn linear models in-context. In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

[29] Yedi Zhang, Aaditya K Singh, Peter E. Latham, and Andrew M Saxe. Training dynamics of in-context learning in linear attention. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=aFNq67ilos.

Contents

1 Introduction **1**

2 Preliminaries **2**
2.1 In-context Multiclass Classification . . . 2
2.2 Data Generation and Training Process . . . 3
2.3 Assumptions . . . 4

3 Main Results **4**
3.1 Implicit Bias on Pre-trained Transformers . . . 4
3.2 Transformers Can Learn Multiclass Classification In-Context via Stable Rank . . . 5
3.3 Multiclass In-Context Classification under Test Label Imbalance . . . 5

4 Conclusion and Future Direction **6**

A Notation **11**

B Related Works **11**
B.1 Motivation of Definition 1 . . . 12

C Sample Complexity Lower Bounds for GMM-Based Classification **13**

D Proof of Our Implicit Bias Results **14**
D.1 Auxilary Lemma . . . 16

E Concentrations on the Haar-Distributed Random Matrices **18**
E.1 Bilinear Hanson-Wright inequality . . . 19
E.2 Concentration on $(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W} \boldsymbol{\mu}_{k^\star}$. . . . 21
E.3 Concentration on $(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W} \boldsymbol{z}'$, $\boldsymbol{z}^\top \boldsymbol{W} \boldsymbol{\mu}_{k^\star}$ and $\boldsymbol{z}^\top \boldsymbol{W} \boldsymbol{z}'$. . . . 23

F Analysis on the Max-Margin Solution **24**
F.1 Some probabilistic concentration . . . 24
F.2 KKT condition and bound . . . 30

G Proof of Theorem 5 **33**
G.1 Dicussions on Theorem 5 . . . 34

H Proof of Theorem 6 **36**
H.1 Discussions on Theorem 6 . . . 37

I Experiment Details **38**
I.1 Nonlinear Transformer Models for Under Our Data Distribution . . . 38
I.1.1 Results and Discussions . . . 39
I.1.2 Additional discussions on Linear Attention vs (Softmax) Transformer . . . 40
I.2 Pre-trained LLMs for WordNet-based Word Classification . . . 43
I.2.1 Statistics and Discussions . . . 50

Appendix A. Notation

We denote the operator norm and the Frobenius norm of a matrix $\boldsymbol{A}$ by $\|\boldsymbol{A}\|$ and $\|\boldsymbol{A}\|_F$, respectively. The symbol $\boldsymbol{A}_{ij}$ represents the (i, j)-th entry of $\boldsymbol{A}$. We define $\boldsymbol{A}_{a:b,c}$ as a vector with dimension $b - a + 1$, whose i-th element is the $(a + i - 1,\ c)$-th entry of $\boldsymbol{A}$. We use $a \vee b = \max\{a, b\}$, $a \wedge b = \min\{a, b\}$. We use $\mathrm{diag}(d_1, \ldots, d_{m\wedge n})$ to denote the (possibly rectangular) matrix $\boldsymbol{D} \in \mathbb{R}^{m\times n}$ whose entries are $\boldsymbol{D}_{ii} = d_i$ for all $i \leq m \wedge n$ and $\boldsymbol{D}_{ij} = 0$ for $i \neq j$. Standard basis vectors are denoted by $\boldsymbol{e}_k$. We use the asymptotic notations $O(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$ for upper bounds, lower bounds, and tight bounds, respectively. We further use the notations $\tilde{O}(\cdot)$, $\tilde{\Omega}(\cdot)$, and $\tilde{\Theta}(\cdot)$ to omit poly-logarithmic factors. We use $\mathrm{sr}(\boldsymbol{A}) := \|\boldsymbol{A}\|_F^2/\|\boldsymbol{A}\|^2$ to denote the stable rank of $\boldsymbol{A}$. The stable rank quantifies the effective dimensionality, which captures the isotropy of the row vectors.

Appendix B. Related Works

Several recent works study in-context classification, mostly focusing on the *binary* case [6, 15] with two class-center vectors pointing to antipodal directions. In this setting, the problem is invariant to simultaneously flipping the signs of both the data and labels; therefore, label-imbalanced situations cannot be captured. Consequently, such settings are insufficient to capture phenomena arising under label imbalance, such as *majority label bias* [10], where in-context learning tends to favor the majority label in the prompt.

Furthermore, as a more general setting, the multiclass regime induces a more complex geometric structure than the binary setting. Shen et al. [22] study in-context *multiclass* classification, but their analysis is restricted to approximately isotropic class-center geometry, leading to an intrinsic dimension comparable to the number of classes. This contrasts with empirical observations that real-world representations, even in high-dimensional embedding spaces, often concentrate near low-dimensional subspaces [19, 20]. Moreover, the paper focuses on balanced test labels and thus fails to capture the majority label bias.

Theory of ICL. To enable tractable analysis, much of the prior work on the theory of ICL focuses on simplified architectures such as linear self-attention. Within this regime, Ahn et al. [1], von Oswald et al. [27] show that ICL can be interpreted as implicitly implementing preconditioned gradient descent for linear regression tasks, thereby framing ICL as a form of meta-learning, or learning-to-learn [24]. Akyürek et al. [3] also show that transformers can implement standard learning algorithms for linear models, including gradient descent and closed-form ridge regression. Kim et al. [11], Magen and Vardi [15] mainly focus on the optimality of ICL in terms of meta-learning. Zhang et al. [28, 29] also study linear self-attention and provide insights into the mechanism of in-context learning by analyzing the dynamics of gradient flow. Ahn et al. [2] demonstrate that in-context regression with linear attention can be useful for explaining the optimization of actual transformer models.

Recent research also investigates classification tasks [4, 6, 8, 12, 13, 15, 22]. Bai et al. [4] show that transformers can perform in-context learning by implementing gradient-based algorithms for a broad class of models, including logistic regression. Giannou et al. [8] further show that transformers can approximate higher-order optimization methods for logistic regression. Li et al. [12] analyze in-context learning capabilities in binary classification in nonlinear transformers. Similarly, Li et al. [13] study the nonlinear setting and show that a one-layer transformer can learn the binary one-nearest-neighbor problem in context. In contrast, Shen et al. [22] study gradient flow in

in-context classification in linear attention setting, and analyze how the prompt length during both training and testing affects the inference error.

Implicit bias. In over-parameterized regimes, optimization converges to specific solutions among infinitely many interpolating ones. Implicit bias is a central concept in the theoretical study of deep learning, particularly in classification settings where the direction of the learned parameters, rather than their scale, governs generalization behavior. Classical results establish convergence of gradient descent to max-margin solutions in multiclass linear classification [21, 23], with extensions to homogeneous neural networks [14].

The works most closely related to ours are Frei and Vardi [6], Magen and Vardi [15], which characterize the implicit bias of the model and leverage it to derive generalization bounds. Frei and Vardi [6] show that transformers can generalize in binary classification tasks while perfectly fitting the in-context examples, by exhibiting an implicit bias toward max-margin classifiers. Their analysis demonstrates that ICL can achieve benign overfitting, offering a principled explanation for generalization without explicit regularization. Building on a closely related perspective, Magen and Vardi [15] further show that, in the same binary setting, in-context learning implements a near optimal meta-learner. Together, these works establish implicit bias as a key mechanism underlying generalization in binary ICL.

B.1. Motivation of Definition 1

Motivation of Our Data Generation Process. In prior works [6, 15], binary classification tasks are constructed by sampling a single reference vector uniformly from the hypersphere, which can be interpreted as drawing an isotropic Gaussian vector and discarding its magnitude while retaining its direction. To extend this idea to the multiclass setting, we consider generating an i.i.d. Gaussian matrix and decomposing it into magnitude and directional components via singular value decomposition. In particular, for an i.i.d. Gaussian matrix, the left and right singular vector matrices—the directional components—are distributed uniformly on the set of orthogonal matrices and are independent of the singular values. Therefore, Definition 1 provides a natural generalization of the binary, direction-based sampling scheme to the multiclass setting. Furthermore, Definition 1 allows us to reflect the isotropy of class-center vectors via the spectrum. In particular, we study generalization on test tasks and quantify its dependence on the stable rank of $\mathbf{\Lambda}$, enabling the analysis of complex multiclass geometry beyond the scope of prior works.

Appendix C. Sample Complexity Lower Bounds for GMM-Based Classification

Solving a given test task in our setting can be viewed as a multiclass classification task under a Gaussian mixture model. In particular, Giraud and Verzelen [9] analyze clustering and classification under a sub-Gaussian mixture model and show that the misclassification error decays exponentially with the signal-to-noise ratio

$$s^2 = \frac{\Delta^2}{L^2 \max_k \|\boldsymbol{C}_k\|} \wedge \frac{m\Delta^4}{L^4 \max_k \|\boldsymbol{C}_k\|_F^2}, \qquad \Delta^2 = \min_{i \neq j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2,$$

where m denotes the minimum cluster size. Here, L denotes the sub-Gaussianity parameter, and $\boldsymbol{C}_k$ represents the noise covariance of the k-th class. In our setting, the noise is standard Gaussian with covariance $\boldsymbol{I}_d$, implying $L = 1$ and $\boldsymbol{C}_k = \boldsymbol{I}_d$ for all k. We also assume that, for a balanced test task with M examples, data points are drawn uniformly across classes, with $S = \frac{M}{K}$ examples per class. Under these assumptions, the signal-to-noise ratio simplifies to

$$s^2 = \Delta^2 \wedge \frac{S\Delta^4}{d},$$

To achieve a small constant classification error, it is therefore necessary that the number of examples per class scales as $S = \Omega\left(\frac{d}{\Delta^4}\right)$. While prior work [15] assumes an explicit separation parameter Δ^2, this quantity is not directly specified in our setting. Nevertheless, we are able to tightly bound Δ^2 using the structural properties of the class-center matrix of test task which has singular value matrix $\boldsymbol{\Lambda}$, as shown below.

Lemma 7 $\Delta^2 = \min_{i \neq j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2 \leq \frac{2}{K-1} \|\boldsymbol{\Lambda}\|_F^2$.

Proof Let $\bar{\boldsymbol{\mu}} := \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{\mu}_k$.

$$\sum_{i \neq j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2 = 2K \sum_i \|\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}\|^2 \leq 2K \sum_i \|\boldsymbol{\mu}_i\|^2 = 2K \|\boldsymbol{\Lambda}\|_F^2.$$

The left-hand side consists of $K(K-1)$ terms. By the pigeonhole principle, there exists at least one (i, j) such that

$$\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2 \leq \frac{2}{K-1} \|\boldsymbol{\Lambda}\|_F^2.$$

■

Combining this bound with the information-theoretic requirement on S we obtain the consequence.

Proposition 8 *Achieving constant classification error requires* $S = \Omega\left(\frac{dK^2}{\|\boldsymbol{\Lambda}\|_F^4}\right)$.

Appendix D. Proof of Our Implicit Bias Results

Implicit bias is a theoretically powerful concept for analyzing the behavior of parameters in over-parameterized models. To characterize the limiting dynamics induced by gradient-based training, we introduce several assumptions.

Assumption 9 *There exist $\boldsymbol{W}_\star$ such that for every τ and every $k \neq k_\tau^\star$,*

$$(\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k})^\top \boldsymbol{W}_\star \boldsymbol{x}_\tau^{(N+1)} > 0.$$

Regarding Assumption 9, since our analysis imposes a specific structural model on the data, directly assuming realizability would be potentially restrictive. Instead, under our generative assumptions, we can show that this condition holds with high probability for sufficiently large K. From the last part of Theorem 21, with probability of at least $1-\delta$,

$$\begin{aligned}&\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}, \boldsymbol{x}_\tau^{(N+1)} \rangle \\ &\geq \frac{1}{K^2}\|\boldsymbol{\Sigma}\|_F^2 - c_0 \frac{\sigma_1^2}{K\sqrt{K}} \log\left(\frac{B^2K^2}{\delta}\right) = \frac{\sigma_1^2}{K^2}\left(\mathrm{sr}(\boldsymbol{\Sigma}) - c_0\sqrt{K}\log\left(\frac{B^2K^2}{\delta}\right)\right).\end{aligned}$$

By Assumptions 2 ($\mathrm{sr}(\boldsymbol{\Sigma}) = \Theta(K)$) and 4 (the number of classes K is sufficiently large), the right-hand side is positive. Consequently, Assumption 9 holds with high probability for $\boldsymbol{W}_\star = \boldsymbol{I}_d$.

Starting from the output $\hat{\boldsymbol{y}}$:

$$\hat{\boldsymbol{y}}(\boldsymbol{E};\boldsymbol{W}) = \frac{1}{N}\begin{pmatrix}\sum_{j\in\mathcal{C}_1} \boldsymbol{x}^{(j)\top}\boldsymbol{W}\boldsymbol{x}^{(N+1)} \\ \vdots \\ \sum_{j\in\mathcal{C}_K} \boldsymbol{x}^{(j)\top}\boldsymbol{W}\boldsymbol{x}^{(N+1)}\end{pmatrix} = \begin{pmatrix}\hat{\boldsymbol{\mu}}_1^\top \boldsymbol{W}\boldsymbol{x}^{(N+1)} \\ \vdots \\ \hat{\boldsymbol{\mu}}_K^\top \boldsymbol{W}\boldsymbol{x}^{(N+1)}\end{pmatrix}$$

For each element, we can obtain

$$\hat{\boldsymbol{\mu}}_k^\top \boldsymbol{W}\boldsymbol{x}^{(N+1)} = \mathrm{Tr}(\hat{\boldsymbol{\mu}}_k^\top \boldsymbol{W}\boldsymbol{x}^{(N+1)}) = \mathrm{Tr}(\boldsymbol{W}\boldsymbol{x}^{(N+1)}\hat{\boldsymbol{\mu}}_k^\top) = \mathrm{Vec}(\boldsymbol{W}^\top)^\top \mathrm{Vec}(\boldsymbol{x}^{(N+1)}\hat{\boldsymbol{\mu}}_k^\top)$$

Using the equality $\mathrm{Vec}(\boldsymbol{u}\boldsymbol{v}^\top) = (\boldsymbol{v}\otimes\boldsymbol{I}_d)\boldsymbol{u}$, we have

$$\mathrm{Vec}(\boldsymbol{x}^{(N+1)}\hat{\boldsymbol{\mu}}_k^\top) = (\hat{\boldsymbol{\mu}}_k \otimes \boldsymbol{I}_d)\boldsymbol{x}^{(N+1)}$$

Denote $\boldsymbol{w} = \mathrm{Vec}(\boldsymbol{W}^\top)$, $\boldsymbol{A}_{\tau,k} = \hat{\boldsymbol{\mu}}_{\tau,k} \otimes \boldsymbol{I}_d$, We can re-write the loss function as follows:

$$\begin{aligned}\mathcal{L}(\boldsymbol{W}) &= -\frac{1}{B}\sum_{\tau=1}^{B}\log\frac{\exp\left(\boldsymbol{w}^\top \boldsymbol{A}_{\tau,k_\tau^\star}\boldsymbol{x}_\tau^{(N+1)}\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{w}^\top \boldsymbol{A}_{\tau,k}\boldsymbol{x}_\tau^{(N+1)}\right)} \\ &= \frac{1}{B}\sum_{\tau=1}^{B}\log\left(\sum_{k=1}^{K}\exp\left(\boldsymbol{w}^\top(\boldsymbol{A}_{\tau,k} - \boldsymbol{A}_{k_\tau^\star,\tau})\boldsymbol{x}_\tau^{(N+1)}\right)\right)\end{aligned}$$

Define $\tilde{\boldsymbol{x}}_{\tau,k} := (\boldsymbol{A}_{\tau,k_\tau^\star} - \boldsymbol{A}_{\tau,k})\boldsymbol{x}_\tau^{(N+1)}$, and we denote by $\mathcal{S}_k$ the set of tasks whose support vectors are associated with class k. Consider the max-margin problem

$$\boldsymbol{W}_{MM} := \arg\min \|\boldsymbol{W}\|_F^2, \quad \text{s.t. } (\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k})^\top \boldsymbol{W}\boldsymbol{x}_\tau^{(N+1)} \geq 1. \quad \forall\tau,\ k \neq k_\tau^\star.$$

By the KKT conditions, the solution admits the representation

$$\boldsymbol{W}_{MM} = \sum_{\tau=1}^{B}\sum_{k=1}^{K} \alpha_{\tau,k}(\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k})\boldsymbol{x}_\tau^{(N+1)\top}.$$

To characterize the implicit bias, we impose the following assumption.

Assumption 10 *There exist a vector $\tilde{\boldsymbol{w}} \in \mathbb{R}^{d^2}$ which satisfies:*

$$\forall k \in [K],\ \forall \tau \in \mathcal{S}_k :\ \eta \exp\left(-\tilde{\boldsymbol{w}}^\top \tilde{\boldsymbol{x}}_{\tau,k}\right) = \alpha_{\tau,k}$$

This assumption ensures the existence of a parameter vector that matches the effective weights induced by gradient descent dynamics. Similar assumptions play a crucial role in the analysis of implicit bias in prior works (Theorem 7 of Soudry et al. [23], Assumption 4.1. of Ravi et al. [21]).

Here, a subtle difference is that we express this using a single $\tilde{\boldsymbol{w}} \in \mathbb{R}^{d^2}$, while the existing papers represent it in the form of $\tilde{\boldsymbol{w}}_{y_n} - \tilde{\boldsymbol{w}}_k$. Since each $\tilde{\boldsymbol{w}}_k$ is a d-dimensional vector and there are K such vectors, concatenating them into a single vector gives constraints that are similar to our assumption with $\tilde{\boldsymbol{w}}$ replaced with a dk-dimensional vector. By comparison, our assumption involves more parameters, while having the same number of equality constraints as prior works. Therefore, it is milder in that the parameter vector lies in a higher-dimensional space, yielding more degrees of freedom and thus making the existence condition less restrictive.

For existence perspective, Soudry et al. [23] establish the existence of such solutions in the binary classification setting for almost all datasets. Ravi et al. [21] provides the empirical results on existence (See Appendix H of Ravi et al. [21]).

Theorem 11 *Under Assumption 9, 10, and with sufficient small step size η, the limit of gradient descent $\boldsymbol{W}(t+1) = \boldsymbol{W}(t) - \eta\nabla\mathcal{L}(\boldsymbol{W}(t))$ converges in direction to max-margin solution:*

$$\begin{aligned}
&\arg\min_{\boldsymbol{w}} \|\boldsymbol{w}\|^2 \text{ s.t. } \forall\tau, \forall k \neq k_\tau^\star : \boldsymbol{w}^\top \tilde{\boldsymbol{x}}_{\tau,k} \geq 1 \\
\Leftrightarrow &\arg\min_{\boldsymbol{W}} \|\boldsymbol{W}\|^2 \text{ s.t. } \forall\tau, \forall k \neq k_\tau^\star : \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star}^\top \boldsymbol{W}\boldsymbol{x}_\tau^{(N+1)} - \hat{\boldsymbol{\mu}}_{\tau,k}^\top \boldsymbol{W}\boldsymbol{x}_\tau^{(N+1)} \geq 1.
\end{aligned}$$

Proof

Define $\boldsymbol{r}(t) = \boldsymbol{w}(t) - \hat{\boldsymbol{w}}\log t - \tilde{\boldsymbol{w}}$, where $\hat{\boldsymbol{w}}$ is a vector representation of max-margin solution. Our goal is to show $\|\boldsymbol{r}(t)\|$ is bounded.

$$\|\boldsymbol{r}(t+1)\|^2 = \|\boldsymbol{r}(t+1) - \boldsymbol{r}(t)\|^2 + 2(\boldsymbol{r}(t+1) - \boldsymbol{r}(t))^\top \boldsymbol{r}(t) + \|\boldsymbol{r}(t)\|^2$$

Consider the first term of the right-hand side,

$$\begin{aligned}
\|\boldsymbol{r}(t+1) - \boldsymbol{r}(t)\|^2 &= \|\boldsymbol{w}(t+1) - \hat{\boldsymbol{w}}\log(t+1) - \boldsymbol{w}(t) + \hat{\boldsymbol{w}}\log(t)\|^2 \\
&= \|-\eta\nabla\mathcal{L}(\boldsymbol{W}(t)) - \hat{\boldsymbol{w}}[\log(t+1) - \log(t)]\|^2 \\
&= \eta^2\|\nabla\mathcal{L}(\boldsymbol{W}(t))\|^2 + \|\hat{\boldsymbol{w}}\|^2\log^2(1+t^{-1}) + 2\eta\hat{\boldsymbol{w}}^\top\nabla\mathcal{L}(\boldsymbol{W}(t))\log(1+t^{-1}) \\
&\leq \eta^2\|\nabla\mathcal{L}(\boldsymbol{W}(t))\|^2 + \|\hat{\boldsymbol{w}}\|^2 t^{-2}
\end{aligned}$$

The last inequality is due to Lemma 12, and $\forall x > 0 : x \geq \log(1+x) > 0$. By Lemma 13, we know that

$$\|\nabla \mathcal{L}(\boldsymbol{W}(t))\|^2 = o(1) \text{ and } \sum_{t=0}^{\infty} \|\nabla \mathcal{L}(\boldsymbol{W}(t))\|^2 < \infty$$

Considering the convergence of t^{-2} power series, we can obtain

$$\|\boldsymbol{r}(t+1) - \boldsymbol{r}(t)\|^2 = o(1) \text{ and } \sum_{t=0}^{\infty} \|\boldsymbol{r}(t+1) - \boldsymbol{r}(t)\|^2 = C_0 < \infty$$

Note that this implies that $\forall \epsilon > 0$, there exists t_0 such that for all $t > t_0$,

$$|\|\|\boldsymbol{r}(t+1)\|\| - \|\boldsymbol{r}(t)\|| < \epsilon_0.$$

By Lemma 14, we can find t_1, C_1, C_2 such that $\forall t > t_1$:

$$(\boldsymbol{r}(t+1) - \boldsymbol{r}(t))^\top \boldsymbol{r}(t) \leq C_1 t^{-\theta} + C_2 t^{-2}$$

Thus,

$$\|\boldsymbol{r}(t)\|^2 - \|\boldsymbol{r}(t_1)\|^2 = \sum_{u=t_1} [\|\boldsymbol{r}(u+1)\|^2 - \|\boldsymbol{r}(u)\|^2] \leq C_0 + 2\sum_{u=t_1}^{t-1} [C_1 u^{-\theta} + C_2 u^{-2}]$$

Note that $\theta > 1$, we can claim that $\boldsymbol{r}(t)$ is bounded. ■

D.1. Auxilary Lemma

Lemma 12 *Under Assumption 9 and for an arbitrary $\boldsymbol{W}$, $\langle \nabla \mathcal{L}(\boldsymbol{W}), \boldsymbol{W}_\star \rangle < 0$.*

Proof

$$\begin{aligned}
\mathcal{L}(\boldsymbol{W}) &= -\frac{1}{B}\sum_{\tau=1}^{B} \log\left(\frac{\exp(\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star}^\top \boldsymbol{W}\boldsymbol{x}_\tau^{(N+1)})}{\sum_{k=1}^{K}\exp(\hat{\boldsymbol{\mu}}_{\tau,k}^\top \boldsymbol{W}\boldsymbol{x}_\tau^{(N+1)})}\right) \\
&= \frac{1}{B}\sum_{\tau=1}^{B} \log\left(\sum_{k=1}^{K}\exp((\hat{\boldsymbol{\mu}}_{\tau,k} - \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star})^\top \boldsymbol{W}\boldsymbol{x}_\tau^{(N+1)})\right)
\end{aligned}$$

The gradient can be calculated as follows:

$$\nabla \mathcal{L}(\boldsymbol{W}) = \frac{1}{B}\sum_{\tau=1}^{B}\sum_{k=1}^{K} p_{\tau,k}(\boldsymbol{W})(\hat{\boldsymbol{\mu}}_{\tau,k} - \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star})\boldsymbol{x}_\tau^{(N+1)\top},$$

where

$$p_{\tau,k}(\boldsymbol{W}) = \frac{\exp((\hat{\boldsymbol{\mu}}_{\tau,k} - \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star})^\top \boldsymbol{W} \boldsymbol{x}_\tau^{(N+1)})}{\sum_{r=1}^K \exp((\hat{\boldsymbol{\mu}}_{\tau,r} - \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star})^\top \boldsymbol{W} \boldsymbol{x}_\tau^{(N+1)})}.$$

$$\begin{aligned}
\langle \nabla \mathcal{L}(\boldsymbol{W}), \boldsymbol{W}_\star \rangle &= \frac{1}{B} \sum_{\tau=1}^B \sum_{k=1}^K p_{\tau,k}(\boldsymbol{W}) \langle (\hat{\boldsymbol{\mu}}_{\tau,k} - \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star}) \boldsymbol{x}_\tau^{(N+1)\top}, \boldsymbol{W}_\star \rangle \\
&= \frac{1}{B} \sum_{\tau=1}^B \sum_{k=1}^K p_{\tau,k}(\boldsymbol{W}) \operatorname{Tr}\left(\boldsymbol{x}_\tau^{(N+1)} (\hat{\boldsymbol{\mu}}_{\tau,k} - \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star})^\top \boldsymbol{W}_\star \right) \\
&= \frac{1}{B} \sum_{\tau=1}^B \sum_{k=1}^K p_{\tau,k}(\boldsymbol{W}) (\hat{\boldsymbol{\mu}}_{\tau,k} - \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star})^\top \boldsymbol{W}_\star \boldsymbol{x}_\tau^{(N+1)} < 0
\end{aligned}$$

The last inequality is due to Assumption 9, and $p_{\tau,k}(\boldsymbol{W}) > 0$. ■

Lemma 13 (Lemma 10 in Soudry et al. [23])

Let $\mathcal{L}(\boldsymbol{w})$ be a β-smooth non-negative objective. If $\eta < 2\beta^{-1}$, then for any $\boldsymbol{w}(0)$ with the GD sequence

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \eta \nabla \mathcal{L}(\boldsymbol{w}(t))$$

we have that $\sum_{u=0}^\infty \|\nabla \mathcal{L}(\boldsymbol{w}(u))\|^2 < \infty$ and therefore $\lim_{t\to\infty} \|\nabla \mathcal{L}(\boldsymbol{w}(t))\|^2 = 0$.

Since we assume the cross-entropy loss, we can easily verify that this lemma holds for our setting.

Lemma 14 (Lemma 20 in Soudry et al. [23])

Let $\theta = \min_k \left[\min_{\tau \notin \mathcal{S}_k} \tilde{\boldsymbol{x}}_{\tau,k}^\top \hat{\boldsymbol{w}} \right] > 1$. We have

$$\exists C_1, C_2, t_1, : \forall t > t_1 : (\boldsymbol{r}(t+1) - \boldsymbol{r}(t))^\top \boldsymbol{r}(t) \le C_1 t^{-\theta} + C_2 t^{-2}.$$

Although the setting of this lemma is slightly different, by substituting $\tilde{\boldsymbol{x}}_{\tau,k} := (\boldsymbol{A}_{\tau,k_\tau^\star} - \boldsymbol{A}_{\tau,k}) \boldsymbol{x}_\tau^{(N+1)}$ into the original proof and applying Assumption 10, and Lemma 13, we obtain the same result.

Appendix E. Concentrations on the Haar-Distributed Random Matrices

From Definition 1, the randomness arises from the uniform sampling from the set of orthogonal matrices. To analyze this term, we rely on concentration properties of random orthogonal matrices, which are formally characterized via the Haar measure on the orthogonal group:

Definition 15 *Let $\mathbb{O}(d)$ denote the orthogonal group of dimension d,*

$$\mathbb{O}(d) := \{\boldsymbol{Q} \in \mathbb{R}^{d\times d} \mid \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}_d\}.$$

The Haar measure on $\mathbb{O}(d)$ is the unique probability measure μ that is invariant under left and right multiplication, i.e., $\mu(\boldsymbol{Q}A) = \mu(A\boldsymbol{Q}) = \mu(A)$, for all measurable sets $A \subseteq \mathbb{O}(d)$ and $\boldsymbol{Q} \in \mathbb{O}(d)$.

We introduce two lemmas concerning Haar-distributed random matrices.

Lemma 16 (Theorem 5.17 of Meckes [17]) *Let S be one of $\mathbb{SO}(d)$ or $\mathbb{SO}^-(d)$. Let $f : S \to \mathbb{R}$ be an L-Lipschitz function with respect to Hilbert-Schmidt metrics on the S, and $\boldsymbol{U}$ be a Haar-distributed random matrices. Then for each $t > 0$, there exists a constant $c > 0$ such that*

$$\mathbb{P}\left(f(\boldsymbol{U}) \geq \mathbb{E}f(\boldsymbol{U}) + t\right) \leq \exp\left(-\frac{cdt^2}{L^2}\right).$$

Due to the disconnectedness of the orthogonal group, the lemma cannot be applied directly.

Lemma 17 *Assume that S is either $\mathbb{SO}(d)$ or $\mathbb{SO}^-(d)$ with $d \geq 3$, and let $\boldsymbol{U} \in S$ be a random matrix sampled from the uniform measure on S. For an arbitrary matrix $\boldsymbol{B} \in \mathbb{R}^{d\times d}$,*

$$\mathbb{E}[\boldsymbol{U}\boldsymbol{B}\boldsymbol{U}^\top] = \frac{\mathrm{Tr}(\boldsymbol{B})}{d}\boldsymbol{I}_d$$

Proof Let $\boldsymbol{A} \in \mathbb{SO}(d)$ be an arbitrary orthogonal matrix with determinant 1. By the translation invariance of the Haar measure, we have

$$\boldsymbol{A}\mathbb{E}_{\boldsymbol{U}}\left[\boldsymbol{U}\boldsymbol{B}\boldsymbol{U}^\top\right]\boldsymbol{A}^\top = \mathbb{E}_{\boldsymbol{U}}\left[\boldsymbol{A}\boldsymbol{U}\boldsymbol{B}(\boldsymbol{A}\boldsymbol{U})^\top\right] = \mathbb{E}_{\boldsymbol{U}}\left[\boldsymbol{U}\boldsymbol{B}\boldsymbol{U}^\top\right]$$

Hence, $\boldsymbol{D} = \mathbb{E}\left[\boldsymbol{U}\boldsymbol{B}\boldsymbol{U}^\top\right]$ commutes with any orthogonal matrix. Decompose $\boldsymbol{D}$ into symmetric and anti-symmetric parts, $\boldsymbol{D}^s, \boldsymbol{D}^a$. ($\boldsymbol{D}^s = \frac{\boldsymbol{D}+\boldsymbol{D}^\top}{2}, \boldsymbol{D}^a = \frac{\boldsymbol{D}-\boldsymbol{D}^\top}{2}$)

$$\begin{aligned}\boldsymbol{A}\boldsymbol{D}^s = \boldsymbol{A}\frac{\boldsymbol{D}+\boldsymbol{D}^\top}{2} &= \frac{1}{2}\left(\boldsymbol{A}\boldsymbol{D} + \boldsymbol{A}\boldsymbol{D}^\top\right) = \frac{1}{2}\left(\boldsymbol{D}\boldsymbol{A} + (\boldsymbol{D}\boldsymbol{A}^\top)^\top\right)\\ &= \frac{1}{2}\left(\boldsymbol{D}\boldsymbol{A} + (\boldsymbol{A}^\top\boldsymbol{D})^\top\right) = \frac{1}{2}\left(\boldsymbol{D}\boldsymbol{A} + \boldsymbol{D}^\top\boldsymbol{A}\right) = \boldsymbol{D}^s\boldsymbol{A}.\end{aligned}$$

Likewise, we can claim that $\boldsymbol{A}\boldsymbol{D}^a = \boldsymbol{D}^a\boldsymbol{A}$. Consider a matrix form of

$$\boldsymbol{P} = \begin{pmatrix} 1 & & \\ & \boldsymbol{R}_\theta & \\ & & \boldsymbol{I}_{d-3}\end{pmatrix} (d > 3), \quad \begin{pmatrix} 1 & \boldsymbol{0}_{1\times 2} \\ \boldsymbol{0}_{2\times 1} & \boldsymbol{R}_\theta\end{pmatrix} (d = 3)$$

Where $\boldsymbol{R}_\theta$ is a 2D rotation matrix associated with an angle θ. This matrix represents a rotation acting on the subspace spanned by $\boldsymbol{e}_2$ and $\boldsymbol{e}_3$. Note that $\boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{I}_d$, and $\det(\boldsymbol{P}) = 1$, so $\boldsymbol{P} \in \mathbb{SO}(d)$.

$$\boldsymbol{e}_1^\top \boldsymbol{D}^a(\cos\theta \boldsymbol{e}_2 + \sin\theta \boldsymbol{e}_3) = \boldsymbol{e}_1^\top \boldsymbol{D}^a \boldsymbol{P}\boldsymbol{e}_2 = \boldsymbol{e}_1^\top \boldsymbol{P}\boldsymbol{D}^a \boldsymbol{e}_2 = (\boldsymbol{P}^\top \boldsymbol{e}_1)^\top \boldsymbol{D}^a \boldsymbol{e}_2 = \boldsymbol{e}_1^\top \boldsymbol{D}^a \boldsymbol{e}_2$$

Since this expression does not depend on θ, we have $\boldsymbol{e}_1^\top \boldsymbol{D}^a \boldsymbol{e}_2 = 0$, and $\boldsymbol{e}_1^\top \boldsymbol{D}^a \boldsymbol{e}_3 = 0$. This can be generalized to any basis which is orthogonal to $\boldsymbol{e}_1$. Since $\boldsymbol{D}^a$ is antisymmetric, $\boldsymbol{e}_1^\top \boldsymbol{D}^a \boldsymbol{e}_1 = 0$, we can conclude that $\boldsymbol{e}_1^\top \boldsymbol{D}^a \boldsymbol{v} = 0$ for every vector $\boldsymbol{v} \in \mathbb{R}^d$. $\boldsymbol{e}_1$ can be replaced to any basis $\boldsymbol{e}_2, \cdots, \boldsymbol{e}_d$, $\boldsymbol{D}^a = 0$. We now move on to $\boldsymbol{D}^s$, the symmetric part. Let λ_i be eigenvalues of $\boldsymbol{D}^s$, and v_i be corresponding eigenvectors which are normalized. There exists $\boldsymbol{P} \in \mathbb{SO}(d)$ such that $\boldsymbol{P}\boldsymbol{v}_i = \boldsymbol{v}_j$.

$$\lambda_i \boldsymbol{v}_j = \boldsymbol{P}\lambda_i \boldsymbol{v}_i = \boldsymbol{P}\boldsymbol{D}^s \boldsymbol{v}_i = \boldsymbol{D}^s \boldsymbol{P}\boldsymbol{v}_i = \lambda_j \boldsymbol{v}_j$$

Therefore, $\lambda_i = \lambda_j$; likewise, all eigenvalues are identical. $\boldsymbol{D}^s$ is $\lambda \boldsymbol{I}$ for some $\lambda \in \mathbb{R}$, so is $\boldsymbol{D}$.

$$\mathbb{E}_{\boldsymbol{U}}\left[\boldsymbol{U}\boldsymbol{B}\boldsymbol{U}^\top\right] = \lambda \boldsymbol{I}_d, \quad \mathrm{Tr}(\mathbb{E}_{\boldsymbol{U}}\left[\boldsymbol{U}\boldsymbol{B}\boldsymbol{U}^\top\right]) = \lambda d, \quad \lambda = \frac{\mathrm{Tr}(\boldsymbol{B})}{d}.$$

Hence, $\mathbb{E}_{\boldsymbol{U}}\left[\boldsymbol{U}\boldsymbol{B}\boldsymbol{U}^\top\right] = \frac{\mathrm{Tr}(\boldsymbol{B})}{d}\boldsymbol{I}_d$. ■

E.1. Bilinear Hanson-Wright inequality

Lemma 18 (Hanson-Wright inequality for uniform on the sphere) *Let $r > 0$ and $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ be a matrix. If $\boldsymbol{v} \sim$ Unif$(r \cdot \mathbb{S}^{d-1})$, then for any $t \geq 0$,*

$$\mathbb{P}\left(\left|\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} - r^2 \frac{\mathrm{Tr}(\boldsymbol{A})}{d}\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{d^2t^2}{r^4\|\boldsymbol{A}\|_F^2}, \frac{dt}{r^2\|\boldsymbol{A}\|}\right)\right)$$

We generalize the Hanson–Wright inequality to random matrices.

Lemma 19 (Bilinear Hanson-Wright inequality for Haar-random sample) *Fix $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^d$ and let $\boldsymbol{V} \in \mathbb{O}(d)$ be a Haar-distributed random matrix. Then there exists an absolute constant $C, c > 0$ such that*

$$\mathbb{P}\left(\left|\boldsymbol{v}_1^\top \boldsymbol{V}^\top \boldsymbol{A}\boldsymbol{V}\boldsymbol{v}_2 - \frac{\mathrm{Tr}(\boldsymbol{A})}{d}\langle \boldsymbol{v}_1, \boldsymbol{v}_2\rangle\right| \geq t\right) \leq C\exp\left(-c\min\left(\frac{d^2t^2}{\|\boldsymbol{A}\|_F^2\|\boldsymbol{v}_1\|^2\|\boldsymbol{v}_2\|^2}, \frac{dt}{\|\boldsymbol{A}\|\|\boldsymbol{v}_1\|\|\boldsymbol{v}_2\|}\right)\right)$$

Proof Denote $\boldsymbol{P}_{\boldsymbol{v}_1}$ be a projection matrix onto the space spanned by $\boldsymbol{v}_1$.

$$\begin{aligned}
&\mathbb{P}\left(\left|\boldsymbol{v}_1^\top \boldsymbol{V}^\top \boldsymbol{A}\boldsymbol{V}\boldsymbol{v}_2 - \frac{\mathrm{Tr}(\boldsymbol{A})}{d}\langle \boldsymbol{v}_1, \boldsymbol{v}_2\rangle\right| \geq t\right) \\
&\leq \mathbb{P}\left(\left|\boldsymbol{v}_1^\top \boldsymbol{V}^\top \boldsymbol{A}\boldsymbol{V}\boldsymbol{P}_{\boldsymbol{v}_1}\boldsymbol{v}_2 - \frac{\mathrm{Tr}(\boldsymbol{A})}{d}\langle \boldsymbol{v}_1, \boldsymbol{v}_2\rangle\right| \geq \frac{t}{2}\right) + \mathbb{P}\left(\left|\boldsymbol{v}_1^\top \boldsymbol{V}^\top \boldsymbol{A}\boldsymbol{V}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}_1})\boldsymbol{v}_2\right| \geq \frac{t}{2}\right)
\end{aligned}$$

The first term is equivalent to a Hanson-Wright inequality for uniform on the sphere; in other word,

$$\begin{aligned}
&\mathbb{P}\left(\left|\boldsymbol{v}_1^\top \boldsymbol{V}^\top \boldsymbol{A}\boldsymbol{V}\boldsymbol{P}_{\boldsymbol{v}_1}\boldsymbol{v}_2 - \frac{\mathrm{Tr}(\boldsymbol{A})}{d}\langle \boldsymbol{v}_1, \boldsymbol{v}_2\rangle\right| \geq \frac{t}{2}\right) \\
&\leq 2\exp\left(-c\min\left(\frac{d^2t^2}{\|\boldsymbol{A}\|_F^2\|\boldsymbol{v}_1\|^2\|\boldsymbol{v}_2\|^2}, \frac{dt}{\|\boldsymbol{A}\|\|\boldsymbol{v}_1\|\|\boldsymbol{v}_2\|}\right)\right)
\end{aligned}$$

In the second term, $\boldsymbol{Y} = \boldsymbol{V}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}_1})\boldsymbol{v}_2$ is orthogonal to $\boldsymbol{X} = \boldsymbol{V}\boldsymbol{v}_1$. Without loss of generality, assume that $\|\boldsymbol{X}\| = \|\boldsymbol{Y}\| = 1$. Let $\boldsymbol{g}_1, \boldsymbol{g}_2$ be i.i.d sampled from $\mathcal{N}(0, \boldsymbol{I}_d)$, we can re-write

$$\boldsymbol{X} = \frac{\boldsymbol{g}_1}{\|\boldsymbol{g}_1\|}, \quad \boldsymbol{Y} = \frac{\boldsymbol{P}_{\boldsymbol{X}^\perp}\boldsymbol{g}_2}{\|\boldsymbol{P}_{\boldsymbol{X}^\perp}\boldsymbol{g}_2\|}.$$

And consider the events

$$\mathcal{E}_1 = \left\{\|\boldsymbol{g}_1\| \geq \sqrt{\frac{d}{2}}\right\}, \quad \mathcal{E}_2 = \left\{\|\boldsymbol{P}_{\boldsymbol{X}^\perp}\boldsymbol{g}_2\| \geq \sqrt{\frac{d-1}{2}}\right\}.$$

Then,

$$\mathbb{P}\left(\left|\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{Y}\right| \geq t'\right) \leq \mathbb{P}\left(\left|\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{Y}\right| \geq t' \mid \mathcal{E}_1, \mathcal{E}_2\right) + \mathbb{P}\left(\mathcal{E}_1^{\mathsf{c}}\right) + \mathbb{P}\left(\mathcal{E}_2^{\mathsf{c}}\right)$$

The last two terms are bounded by $\exp(-cd)$ scale, through applying the Theorem 3.1.1 of Vershynin [26].

$$\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{Y} = \frac{\boldsymbol{g}_1^\top \boldsymbol{A}\boldsymbol{g}_2 - \boldsymbol{g}_1^\top \boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{g}_2}{\|\boldsymbol{g}_1\|\|\boldsymbol{P}_{\boldsymbol{X}^\perp}\boldsymbol{g}_2\|}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\left|\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{Y}\right| \geq t' \mid \mathcal{E}_1, \mathcal{E}_2\right) \\
&\leq \mathbb{P}\left(|\boldsymbol{g}_1^\top \boldsymbol{A}\boldsymbol{g}_2| \geq \frac{\sqrt{d(d-1)}t'}{4}\right) + \mathbb{P}\left(|\frac{(\boldsymbol{g}_1^\top \boldsymbol{A}\boldsymbol{g}_1)\cdot(\boldsymbol{g}_1^\top \boldsymbol{g}_2)}{\|\boldsymbol{g}_1\|^3\|\boldsymbol{P}_{\boldsymbol{X}^\perp}\boldsymbol{g}_2\|}| \geq \frac{t'}{2} \mid \mathcal{E}_1, \mathcal{E}_2\right)
\end{aligned}$$

By following the proof of the Theorem 6.2.1 of Vershynin [26], we can easily verify that:

$$\mathbb{P}\left(|\boldsymbol{g}_1^\top \boldsymbol{A}\boldsymbol{g}_2| \geq \frac{\sqrt{d(d-1)}t'}{4}\right) \leq 2\exp\left(-c\min\left(\frac{d^2t'^2}{\|\boldsymbol{A}\|_F^2}, \frac{dt'}{\|\boldsymbol{A}\|}\right)\right)$$

for some constant $c > 0$. For the second term,

$$\begin{aligned}
&\mathbb{P}\left(|(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})\cdot\frac{(\boldsymbol{X}^\top \boldsymbol{g}_2)}{\|\boldsymbol{P}_{\boldsymbol{X}^\perp}\boldsymbol{g}_2\|}| \geq \frac{t'}{2} \mid \mathcal{E}_1, \mathcal{E}_2\right) \\
&\leq \underbrace{\mathbb{P}\left(\left|\left(\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} - \frac{\mathrm{Tr}(\boldsymbol{A})}{d}\right)\cdot\frac{(\boldsymbol{X}^\top \boldsymbol{g}_2)}{\|\boldsymbol{P}_{\boldsymbol{X}^\perp}\boldsymbol{g}_2\|}\right| \geq \frac{t'}{4} \mid \mathcal{E}_1, \mathcal{E}_2\right)}_{(1)} + \underbrace{\mathbb{P}\left(\left|\frac{\mathrm{Tr}(\boldsymbol{A})}{d}\cdot\frac{(\boldsymbol{X}^\top \boldsymbol{g}_2)}{\|\boldsymbol{P}_{\boldsymbol{X}^\perp}\boldsymbol{g}_2\|}\right| \geq \frac{t'}{4} \mid \mathcal{E}_1, \mathcal{E}_2\right)}_{(2)}
\end{aligned}$$

Consider the event $\mathcal{E}_3 = \left\{|\boldsymbol{X}^\top \boldsymbol{g}_2| \leq \sqrt{\frac{d-1}{2}}\right\}$, then the first term is bounded by

$$(1) \leq \mathbb{P}\left(\left|\left(\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} - \frac{\mathrm{Tr}(\boldsymbol{A})}{d}\right)\cdot\frac{(\boldsymbol{X}^\top \boldsymbol{g}_2)}{\|\boldsymbol{P}_{\boldsymbol{X}^\perp}\boldsymbol{g}_2\|}\right| \geq \frac{t'}{4} \mid \mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3\right) + \mathbb{P}(\mathcal{E}_3^{\mathsf{c}} \mid \mathcal{E}_1, \mathcal{E}_2)$$

Note that $\boldsymbol{X}^\top \boldsymbol{g}_2 \overset{d}{=} \mathcal{N}(0,1)$, which is independent from $\|\boldsymbol{g}_1\|$, and $\|\boldsymbol{P}_{\boldsymbol{X}^\perp}\boldsymbol{g}_2\|$; hence, $\mathbb{P}(\mathcal{E}_3^\mathsf{c} \mid \mathcal{E}_1, \mathcal{E}_2) = \mathbb{P}(\mathcal{E}_3^\mathsf{c}) \leq \exp(-cd)$. Under $\mathcal{E}_2$ and $\mathcal{E}_3$, $\frac{(\boldsymbol{X}^\top \boldsymbol{g}_2)}{\|\boldsymbol{P}_{\boldsymbol{X}^\perp}\boldsymbol{g}_2\|} \leq 1$; in contrast, it can be viewed as a Hanson-Wright inequality for uniform on the unit sphere. To sum up,

$$(1) \leq 2\exp\left(-c\min\left(\frac{d^2t'^2}{\|\boldsymbol{A}\|_F^2}, \frac{dt'}{\|\boldsymbol{A}\|}\right)\right) + \exp(-cd).$$

Note that $\frac{\mathrm{Tr}(\boldsymbol{A})}{d} \leq \frac{\|\boldsymbol{A}\|_F}{\sqrt{d}}$ holds by the Cauchy-Schwarz inequality. From $\mathcal{E}_2$, we can conclude that,

$$(2) \leq 2\exp\left(-c\frac{d^2t'^2}{\|\boldsymbol{A}\|_F^2}\right).$$

To sum up,

$$\mathbb{P}(|\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{Y}| \geq t') \leq C_1 \exp\left(-c\min\left(\frac{d^2t'^2}{\|\boldsymbol{A}\|_F^2}, \frac{dt'}{\|\boldsymbol{A}\|}\right)\right) + C_2\exp\left(-c\frac{d^2t'^2}{\|\boldsymbol{A}\|_F^2}\right) + C_3\exp(-cd).$$

For $t' \leq \|\boldsymbol{A}\|$, the first term dominates the remaining terms. For $t' > \|\boldsymbol{A}\|$, we have $\mathbb{P}(|\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{Y}| \geq t') = 0$, since $|\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{Y}| \leq \|\boldsymbol{A}\|$. Therefore, we can conclude that

$$\mathbb{P}\left(\left|\boldsymbol{v}_1^\top \boldsymbol{V}^\top \boldsymbol{A}\boldsymbol{V}\boldsymbol{v}_2 - \frac{\mathrm{Tr}(\boldsymbol{A})}{d}\langle \boldsymbol{v}_1, \boldsymbol{v}_2\rangle\right| \geq t\right) \leq C\exp\left(-c\min\left(\frac{d^2t^2}{\|\boldsymbol{A}\|_F^2\|\boldsymbol{v}_1\|^2\|\boldsymbol{v}_2\|^2}, \frac{dt}{\|\boldsymbol{A}\|\|\boldsymbol{v}_1\|\|\boldsymbol{v}_2\|}\right)\right)$$

for some constant $C, c > 0$. ■

Lemma 20 *Assume $\boldsymbol{U}, \boldsymbol{V}$ are independent Haar-distributed random matrices on $\mathbb{O}(K)$ and $\mathbb{O}(d)$ respectively. Let $\boldsymbol{\mu}_k^\top$ and $\boldsymbol{\mu}_{\boldsymbol{k}^\star}^\top$ be arbitrary row vectors from $\boldsymbol{\mu} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top$, and $z, \boldsymbol{z}'$ be i.i.d. random vectors which are sampled from $\mathcal{N}(0, \boldsymbol{I}_d)$. Then, there exists an absolute constant $C, c > 0$ such that*

$$\begin{aligned}
&\mathbb{P}\left((\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star} - \mathbb{E}\left[(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star}\right] \geq t\right), \\
&\leq C\exp\left(-c\min\left(\frac{d^2t^2}{\lambda_1^4\|\boldsymbol{W}\|_F^2}, \frac{dt}{\lambda_1^2\|\boldsymbol{W}\|}\right)\right) + \exp\left(-\frac{cd^2Kt^2}{\lambda_1^4\,\mathrm{Tr}(\boldsymbol{W})^2}\right), \\
&\mathbb{P}\left(|(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{z}'| \geq t\right) \leq 2\exp\left(-\frac{ct\sqrt{d}}{\lambda_1\|\boldsymbol{W}\|_F}\right), \\
&\mathbb{P}\left(|\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star}| \geq t\right) \leq 2\exp\left(-\frac{ct\sqrt{d}}{\lambda_1\|\boldsymbol{W}\|_F}\right) \\
&\mathbb{P}\left(|\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{z}'| \geq t\right) \leq 2\exp\left(-\frac{ct}{\|\boldsymbol{W}\|_F}\right).
\end{aligned}$$

E.2. Concentration on $(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star}$.

Denote $\boldsymbol{X}_1(\boldsymbol{U}, \boldsymbol{V}) = (\boldsymbol{e}_k - \boldsymbol{e}_{k^\star})^\top \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top \boldsymbol{W}\boldsymbol{V}\boldsymbol{\Lambda}^\top \boldsymbol{U}^\top \boldsymbol{e}_{k^\star}$, then

$$\begin{aligned}
&\mathbb{P}\left(\boldsymbol{X}_1(\boldsymbol{U}, \boldsymbol{V}) \geq \mathbb{E}\boldsymbol{X}_1(\boldsymbol{U}, \boldsymbol{V}) + t\right) \\
&\leq \mathbb{P}\left(\boldsymbol{X}_1(\boldsymbol{U}, \boldsymbol{V}) \geq \mathbb{E}_{\boldsymbol{V}}\boldsymbol{X}_1(\boldsymbol{U}, \boldsymbol{V}) + \frac{t}{2}\right) + \mathbb{P}\left(\mathbb{E}_{\boldsymbol{V}}\boldsymbol{X}_1(\boldsymbol{U}, \boldsymbol{V}) \geq \mathbb{E}\boldsymbol{X}_1(\boldsymbol{U}, \boldsymbol{V}) + \frac{t}{2}\right)
\end{aligned}$$

By Lemma 19,

$$\begin{aligned}
&\mathbb{P}\left(\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) \geq \mathbb{E}_{\boldsymbol{V}}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) + \frac{t}{2} \mid \boldsymbol{U}\right) \\
&\leq C\exp\left(-c\min\left(\frac{d^2t^2}{\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_{k^\star}\|^2\|\boldsymbol{\mu}_{k^\star}\|^2\|\boldsymbol{W}\|_F^2}, \frac{dt}{\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_{k^\star}\|\|\boldsymbol{\mu}_{k^\star}\|\|\boldsymbol{W}\|}\right)\right) \\
&\leq C\exp\left(-c\min\left(\frac{d^2t^2}{\lambda_1^4\|\boldsymbol{W}\|_F^2}, \frac{dt}{\lambda_1^2\|\boldsymbol{W}\|}\right)\right).
\end{aligned}$$

Hence,

$$\mathbb{P}\left(\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) \geq \mathbb{E}_{\boldsymbol{V}}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) + \frac{t}{2}\right) \leq C\exp\left(-c\min\left(\frac{d^2t^2}{\lambda_1^4\|\boldsymbol{W}\|_F^2}, \frac{dt}{\lambda_1^2\|\boldsymbol{W}\|}\right)\right)$$

The orthogonal group decomposes as $\mathbb{O}(K) = \mathbb{SO}(K) \sqcup \mathbb{SO}^-(K)$, where $\mathbb{SO}(K)$ contains matrices with determinant 1, and $\mathbb{SO}^-(K)$ those with determinant -1. We apply the concentration result on each component, noting that the expectation and Lipschitz constants are identical, which allows us to extend the bound to $\mathbb{O}(K)$.

Note that $g(\boldsymbol{U}) = \mathbb{E}_{\boldsymbol{V}}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) = \frac{\mathrm{Tr}(\boldsymbol{W})}{d}(\boldsymbol{e}_k-\boldsymbol{e}_{k^\star})^\top\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\boldsymbol{U}^\top\boldsymbol{e}_{k^\star}$,

$$\begin{aligned}
&|g(\boldsymbol{U}) - g(\boldsymbol{U}')| \\
&\leq \frac{\mathrm{Tr}(\boldsymbol{W})}{d}\left(|(\boldsymbol{e}_k-\boldsymbol{e}_{k^\star})^\top(\boldsymbol{U}-\boldsymbol{U}')\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\boldsymbol{U}^\top\boldsymbol{e}_{k^\star}| + |(\boldsymbol{e}_k-\boldsymbol{e}_{k^\star})^\top\boldsymbol{U}'\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top(\boldsymbol{U}-\boldsymbol{U}')^\top\boldsymbol{e}_{k^\star}|\right) \\
&\leq \frac{\mathrm{Tr}(\boldsymbol{W})}{d}\sqrt{2}\lambda_1^2\|\boldsymbol{U}-\boldsymbol{U}'\| \leq \frac{\mathrm{Tr}(\boldsymbol{W})}{d}\sqrt{2}\lambda_1^2\|\boldsymbol{U}-\boldsymbol{U}'\|_F.
\end{aligned}$$

The second term can be written as

$$\begin{aligned}
&\mathbb{P}\left(\mathbb{E}_{\boldsymbol{V}}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) \geq \mathbb{E}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) + \frac{t}{2}\right) \\
&= \sum_{S\in\{\mathbb{SO}(K),\,\mathbb{SO}^-(K)\}} \mathbb{P}\left(\mathbb{E}_{\boldsymbol{V}}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) \geq \mathbb{E}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) + \frac{t}{2} \;\mid\; \boldsymbol{U}\in S\right)\mathbb{P}\left(\boldsymbol{U}\in S\right).
\end{aligned}$$

From the Lemma 17, we have

$$\mathbb{E}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) = \mathbb{E}\left[\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) \;\mid\; \boldsymbol{U}\in S\right], \quad S\in\{\mathbb{SO}(K),\,\mathbb{SO}^-(K)\}.$$

Note that the Lipschitz constant for each subset is also bounded by $\frac{\sqrt{2}\,\mathrm{Tr}(\boldsymbol{W})\lambda_1^2}{d}$. Then, by Lemma 16, for some constant $c>0$,

$$\begin{aligned}
&\mathbb{P}\left(\mathbb{E}_{\boldsymbol{V}}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) \geq \mathbb{E}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) + \frac{t}{2} \;\mid\; \boldsymbol{U}\in S\right) \\
&= \mathbb{P}\left(\mathbb{E}_{\boldsymbol{V}}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) \geq \mathbb{E}\left[\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) \;\mid\; \boldsymbol{U}\in S\right] + \frac{t}{2} \;\mid\; \boldsymbol{U}\in S\right) \leq \exp\left(-\frac{cd^2Kt^2}{\lambda_1^4\,\mathrm{Tr}(\boldsymbol{W})^2}\right).
\end{aligned}$$

Therefore, we can obtain the concentration.

E.3. Concentration on $(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{z}'$, $\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star}$ and $\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{z}'$.

In order to derive the inequality, we first calculate the sub-Gaussian norm of $\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star}$. Recall the definition of sub-Gaussian norm of the random vector,

$$\|(\boldsymbol{e}_k - \boldsymbol{e}_{k^\star})^\top \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top\|_{\psi_2} = \sup_{\|\boldsymbol{w}\|_2=1} \|(\boldsymbol{e}_k - \boldsymbol{e}_{k^\star})^\top \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top w\|_{\psi_2}$$

Note that $\boldsymbol{V}^\top \boldsymbol{w}$ is a random vector that is uniformly distributed on the unit sphere in $\mathbb{R}^{d-1}$. By Theorem 3.4.6 from Vershynin [26], there exists a constant $C > 0$ such that $\|\boldsymbol{V}^\top \boldsymbol{w}\|_{\psi_2} \leq \frac{C}{\sqrt{d}}$. Hence,

$$\|(\boldsymbol{e}_k - \boldsymbol{e}_{k^\star})^\top \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top w\|_{\psi_2} \leq \|(\boldsymbol{e}_k - \boldsymbol{e}_{k^\star})^\top \boldsymbol{U}\boldsymbol{\Lambda}\|_2 \|\boldsymbol{V}^\top \boldsymbol{w}\|_{\psi_2} \leq \frac{\sqrt{2}C}{\sqrt{d}}\lambda_1.$$

Following the proof of Lemma C.4 of Frei and Vardi [6], we get

$$\mathbb{P}\left(|(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{z}'| \geq t\right) \leq 2\exp\left(-\frac{c_1 t\sqrt{d}}{\lambda_1 \|\boldsymbol{W}\|_F}\right)$$

where c_1 is an absolute constant. Similarly,

$$\mathbb{P}\left(|\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star}| \geq t\right) \leq 2\exp\left(-\frac{c_2 t\sqrt{d}}{\lambda_1 \|\boldsymbol{W}\|_F}\right)$$

Finally, the concentration on $\mathbb{P}\left(|\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{z}'| \geq t\right)$ is a special case of Lemma C.5 of Frei and Vardi [6], we have

$$\mathbb{P}\left(|\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{z}'| \geq t\right) \leq 2\exp\left(-\frac{c_3 t}{\|\boldsymbol{W}\|_F}\right).$$

Appendix F. Analysis on the Max-Margin Solution

In order to derive the generalization bound, we need the lower bound of $\mathrm{Tr}(\boldsymbol{W})$ and the upper bound of $\|\boldsymbol{W}\|_F$. We derive probabilistic concentration bounds for these quantities. Similar to Frei and Vardi [6], Magen and Vardi [15], start by using the KKT conditions of the max-margin solution:

$$\boldsymbol{W}_{MM} = \sum_{k=1}^{K} \sum_{\tau \in \mathcal{S}_k} \alpha_{\tau,k} (\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}) \boldsymbol{x}_\tau^{(N+1)\top}.$$

F.1. Some probabilistic concentration

In order to derive the inequality of the max-margin solution, we need to establish concentration bounds for several random variables. A notable aspect is that there are 3 different types of inner products of vectors: 1. different tasks, 2. same task but different clusters, 3. same task and same cluster. We first assume balanced regime in this section; therefore, assume that there exist $S_\star$ examples for each cluster per task.

Theorem 21 *Let $\delta \in (0, \frac{1}{6})$ be arbitrary. With probability at least $1 - 6\delta$ over the random draws of the training sets, $\{\boldsymbol{U}_\tau, \boldsymbol{V}_\tau, (\boldsymbol{x}_\tau^{(j)}, \boldsymbol{y}_\tau^{(j)})_{j=1}^N\}_{\tau=1}^B$ the following hold:*

1. *Cross-task bounds: for all $\tau \neq \tau'$.*

$$|\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_1}, \hat{\boldsymbol{\mu}}_{\tau',k_{\tau'}^\star} - \hat{\boldsymbol{\mu}}_{\tau',k_2} \rangle| \leq C \left(\frac{S_\star^2 \sigma_1^2}{N^2 \sqrt{d}} + \frac{\sigma_1 S_\star \sqrt{S_\star}}{N^2} + \frac{S_\star \sqrt{d}}{N^2} \right) \log \left(\frac{B^2 K^2}{\delta} \right).$$

2. *Same-task cross-class bounds: for all $\tau,\ k_1 \neq k_2,\ k_i \neq k_\tau^\star$.*

$$\begin{aligned} &|\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_1}, \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_2} \rangle - \frac{S_\star^2}{N^2 K} \|\boldsymbol{\Sigma}\|_F^2| \\ &\leq \frac{d S_\star}{N^2} + C \left(\frac{S_\star^2 \sigma_1^2}{N^2 \sqrt{K}} + \frac{\sigma_1 S_\star \sqrt{S_\star}}{N^2} + \frac{S_\star \sqrt{d}}{N^2} \right) \log \left(\frac{B^2 K^2}{\delta} \right). \end{aligned}$$

3. *Same-task norm bounds: for all $\tau,\ k \neq k_\tau^\star$.*

$$\begin{aligned} &|\|\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}\|^2 - \frac{2 S_\star^2}{N^2 K} \|\boldsymbol{\Sigma}\|_F^2| \\ &\leq \frac{d S_\star}{N^2} + C \left(\frac{S_\star^2 \sigma_1^2}{N^2 \sqrt{K}} + \frac{\sigma_1 S_\star \sqrt{S_\star}}{N^2} + \frac{S_\star \sqrt{d}}{N^2} \right) \log \left(\frac{B^2 K^2}{\delta} \right). \end{aligned}$$

4. *Query norm and query inner-product bounds.*

$$\begin{aligned} &|\langle \boldsymbol{x}_\tau^{(N+1)}, \boldsymbol{x}_{\tau'}^{(N+1)} \rangle| \leq C \left(\frac{\sigma_1^2}{\sqrt{d}} + \sigma_1 + \sqrt{d} \right) \log \left(\frac{B^2 K^2}{\delta} \right), \\ &|\|\boldsymbol{x}_\tau^{(N+1)}\|^2 - \frac{1}{K} \|\boldsymbol{\Sigma}\|_F^2| \leq d + C \left(\frac{\sigma_1^2}{\sqrt{K}} + \sigma_1 + \sqrt{d} \right) \log \left(\frac{B^2 K^2}{\delta} \right), \\ &|\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}, \boldsymbol{x}_\tau^{(N+1)} \rangle - \frac{S_\star}{N K} \|\boldsymbol{\Sigma}\|_F^2| \leq C \left(\frac{S_\star \sigma_1^2}{N \sqrt{K}} + \frac{S_\star \sigma_1}{N} + \frac{\sqrt{d S_\star}}{N} \right) \log \left(\frac{B^2 K^2}{\delta} \right), \end{aligned}$$

where $C > 0$ is an absolute constant.

Proof

1. Bound on $|\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_1}, \hat{\boldsymbol{\mu}}_{\tau',k_{\tau'}^\star} - \hat{\boldsymbol{\mu}}_{\tau',k_2}\rangle|$

   Recall that $\hat{\boldsymbol{\mu}}_{\tau,k} = \frac{1}{N}\sum_{j\in\mathcal{C}_k} \boldsymbol{x}_\tau^{(j)} \overset{d}{=} \frac{S_\star}{N}\boldsymbol{\mu}_{\tau,k} + \frac{\sqrt{S_\star}}{N}\boldsymbol{z}$. Note that $\boldsymbol{\mu}_{\tau,k} = \boldsymbol{V}_\tau \boldsymbol{\Sigma}^\top \boldsymbol{U}_\tau^\top \boldsymbol{e}_k$. We start from:

$$\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_1}, \hat{\boldsymbol{\mu}}_{\tau',k_{\tau'}^\star} - \hat{\boldsymbol{\mu}}_{\tau',k_2}\rangle = \frac{S_\star^2}{N^2}\underbrace{\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k_1}, \boldsymbol{\mu}_{\tau',k_{\tau'}^\star} - \boldsymbol{\mu}_{\tau',k_2}\rangle}_{(1)}$$
$$+ \frac{\sqrt{2}S_\star\sqrt{S_\star}}{N^2}\underbrace{\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k_1}, \boldsymbol{z}'\rangle}_{(2)} + \frac{\sqrt{2}S_\star\sqrt{S_\star}}{N^2}\underbrace{\langle \boldsymbol{z}, \boldsymbol{\mu}_{\tau',k_{\tau'}^\star} - \boldsymbol{\mu}_{\tau',k_2}\rangle}_{(3)} + \frac{2S_\star}{N^2}\underbrace{\langle \boldsymbol{z}, \boldsymbol{z}'\rangle}_{(4)}.$$

   Rewriting (1) in terms of $\boldsymbol{U}$ and $\boldsymbol{V}$, we obtain

$$\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k_1}, \boldsymbol{\mu}_{\tau',k_{\tau'}^\star} - \boldsymbol{\mu}_{\tau',k_2}\rangle = (\boldsymbol{e}_{k_\tau^\star} - \boldsymbol{e}_{k_1})^\top \boldsymbol{U}_\tau \boldsymbol{\Sigma}\boldsymbol{V}_\tau^\top \boldsymbol{V}_{\tau'}\boldsymbol{\Sigma}^\top \boldsymbol{U}_{\tau'}^\top (\boldsymbol{e}_{k_{\tau'}^\star} - \boldsymbol{e}_{k_2})$$

   Define the function

$$f(\boldsymbol{U}_\tau, \boldsymbol{V}_\tau, \boldsymbol{U}_{\tau'}, \boldsymbol{V}_{\tau'}) = (\boldsymbol{e}_{k_\tau^\star} - \boldsymbol{e}_{k_1})^\top \boldsymbol{U}_\tau \boldsymbol{\Sigma}\boldsymbol{V}_\tau^\top \boldsymbol{V}_{\tau'}\boldsymbol{\Sigma}^\top \boldsymbol{U}_{\tau'}^\top (\boldsymbol{e}_{k_{\tau'}^\star} - \boldsymbol{e}_{k_2}).$$

   Since each $\boldsymbol{U}$ and $\boldsymbol{V}$ is sampled from an orthogonal group, we verify that the expectation of the function does not depend on the choice of the conditional expectation respect to Haar measure. Let S be either $\mathbb{SO}(K)$ or $\mathbb{SO}^-(K)$. Consider the conditional expectation

$$\begin{aligned}
&\mathbb{E}\left[f(\boldsymbol{U}_\tau, \boldsymbol{V}_\tau, \boldsymbol{U}_{\tau'}, \boldsymbol{V}_{\tau'}) \mid \boldsymbol{U}_\tau \in S, \boldsymbol{V}_\tau, \boldsymbol{U}_{\tau'}, \boldsymbol{V}_{\tau'}\right] \\
&= (\boldsymbol{e}_{k_\tau^\star} - \boldsymbol{e}_{k_1})^\top \mathbb{E}\left[\boldsymbol{U}_\tau \boldsymbol{\Sigma}\boldsymbol{V}_\tau^\top \boldsymbol{V}_{\tau'}\boldsymbol{\Sigma}^\top \boldsymbol{U}_{\tau'}^\top \mid \boldsymbol{U}_\tau \in S, \boldsymbol{V}_\tau, \boldsymbol{U}_{\tau'}, \boldsymbol{V}_{\tau'}\right](\boldsymbol{e}_{k_{\tau'}^\star} - \boldsymbol{e}_{k_2}) \\
&= (\boldsymbol{e}_{k_\tau^\star} - \boldsymbol{e}_{k_1})^\top \mathbb{E}\left[\boldsymbol{U}_\tau \mid \boldsymbol{U}_\tau \in S\right]\boldsymbol{\Sigma}\boldsymbol{V}_\tau^\top \boldsymbol{V}_{\tau'}\boldsymbol{\Sigma}^\top \boldsymbol{U}_{\tau'}^\top (\boldsymbol{e}_{k_{\tau'}^\star} - \boldsymbol{e}_{k_2}).
\end{aligned}$$

   Choose an arbitrary rotation $\boldsymbol{R} \in \mathbb{SO}(K)$. By the property of the Haar measure,

$$\boldsymbol{R}\mathbb{E}\left[\boldsymbol{U}_\tau \mid \boldsymbol{U}_\tau \in S\right] = \mathbb{E}\left[\boldsymbol{U}_\tau \mid \boldsymbol{U}_\tau \in S\right].$$

   The only matrix satisfying the above equation is the zero matrix. As a consequence, we obtain the following remark:

Remark 22 *Let S_1, S_3 be either $\mathbb{SO}(K)$ or $\mathbb{SO}^-(K)$, and S_2, S_4 be either $\mathbb{SO}(d)$ or $\mathbb{SO}^-(d)$. Suppose that $\boldsymbol{U}_\tau \in S_1, \boldsymbol{V}_\tau \in S_2, \boldsymbol{U}_{\tau'} \in S_3$, and $\boldsymbol{V}_{\tau'} \in S_4$ are Haar distributed. Then,*

$$\mathbb{E}\left[\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k_1}, \boldsymbol{\mu}_{\tau',k_{\tau'}^\star} - \boldsymbol{\mu}_{\tau',k_2}\rangle \mid \boldsymbol{U}_\tau \in S_1, \boldsymbol{V}_\tau \in S_2, \boldsymbol{U}_{\tau'} \in S_3, \boldsymbol{V}_{\tau'} \in S_4\right] = 0$$

For fixed $\boldsymbol{U}_\tau, \boldsymbol{U}_{\tau'}$, the inner product can be viewed as an inner product of two vectors which are sampled from d-dimensional sphere. Note that each vector has sub-Gaussian norm at most $\frac{C\sigma_1}{\sqrt{d}}$ for some constant $C > 0$.

$$\mathbb{P}\left(|\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k_1}, \boldsymbol{\mu}_{\tau',k_{\tau'}^\star} - \boldsymbol{\mu}_{\tau',k_2}\rangle| \geq t \mid \boldsymbol{U}_\tau, \boldsymbol{U}_{\tau'}\right) \leq \exp\left(-\frac{cdt^2}{\sigma_1^4}\right)$$

$$\mathbb{P}\left(\cup_{\tau\neq\tau'} \cup_{\tau'\in[B]} \cup_{k_1\in[K]} \cup_{k_2\in[K]} |\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k_1}, \boldsymbol{\mu}_{\tau',k_{\tau'}^\star} - \boldsymbol{\mu}_{\tau',k_2}\rangle| \geq t\right)$$

$$\leq B^2K^2\mathbb{P}\left(|\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k_1}, \boldsymbol{\mu}_{\tau',k_{\tau'}^\star} - \boldsymbol{\mu}_{\tau',k_2}\rangle| \geq t\right) \leq \exp\left(-\frac{cdt^2}{\sigma_1^4}\right) \leq \delta,$$

for some constant $c > 0$. Therefore, with probability at least $1-\delta$, for all $\tau \neq \tau'$ and $k_1, k_2 \in [K]$, there exists an absolute constant $c_1 > 0$ such that

$$|\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k_1}, \boldsymbol{\mu}_{\tau',k_{\tau'}^\star} - \boldsymbol{\mu}_{\tau',k_2}\rangle| \leq c_1\frac{\sigma_1^2}{\sqrt{d}}\sqrt{\log\left(\frac{B^2K^2}{\delta}\right)}.$$

The bounds for terms (2), (3), and (4) follow identically from Magen and Vardi [15]. Using Lemma 6.2.3 of Vershynin [26], recall that $\boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k_1}$ has sub-Gaussian norm at most $\frac{C\sigma_1}{\sqrt{d}}$. Then, with probability at least $1-2\delta$, for all $\tau \neq \tau'$ and $k_1, k_2 \in [K]$, we have

$$|\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k_1}, \boldsymbol{z}'\rangle| \vee |\langle \boldsymbol{z}, \boldsymbol{\mu}_{\tau',k_{\tau'}^\star} - \boldsymbol{\mu}_{\tau',k_2}\rangle| \leq c_2\sigma_1\log\left(\frac{B^2K^2}{\delta}\right)$$

for some constant $c_2 > 0$. Likewise, with probability at least $1-\delta$, for all $\tau \neq \tau'$ and $k_1, k_2 \in [K]$,

$$|\langle \boldsymbol{z}, \boldsymbol{z}'\rangle| \leq c_3\sqrt{d}\log\left(\frac{B^2K^2}{\delta}\right).$$

To sum up,

$$|\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_1}, \hat{\boldsymbol{\mu}}_{\tau',k_{\tau'}^\star} - \hat{\boldsymbol{\mu}}_{\tau',k_2}\rangle| \leq C_0\left(\frac{S_\star^2\sigma_1^2}{N^2\sqrt{d}} + \frac{\sigma_1 S_\star\sqrt{S_\star}}{N^2} + \frac{S_\star\sqrt{d}}{N^2}\right)\log\left(\frac{B^2K^2}{\delta}\right).$$

2. Bound on $|\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_1}, \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_2}\rangle - \frac{S_\star^2}{N^2K}\|\boldsymbol{\Sigma}\|_F^2|$

   Recall that $\hat{\boldsymbol{\mu}}_{\tau,k} = \frac{1}{n}\sum_{j\in\mathcal{C}_k} \boldsymbol{x}_\tau^{(j)} \overset{d}{=} \frac{S_\star}{N}\boldsymbol{\mu}_{\tau,k} + \frac{\sqrt{S_\star}}{N}\boldsymbol{z}$.

$$\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_1}, \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_2}\rangle = \frac{S_\star^2}{N^2}\underbrace{\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k_1}, \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k_2}\rangle}_{(1)}$$

$$+ \frac{S_\star\sqrt{S_\star}}{N^2}\underbrace{\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k_1}, \boldsymbol{z} + \boldsymbol{z}_1\rangle}_{(2)} + \frac{S_\star\sqrt{S_\star}}{N^2}\underbrace{\langle \boldsymbol{z} + \boldsymbol{z}_2, \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k_2}\rangle}_{(3)} + \frac{S_\star}{N^2}\underbrace{\langle \boldsymbol{z} + \boldsymbol{z}_1, \boldsymbol{z} + \boldsymbol{z}_2\rangle}_{(4)}$$

Consider the first term:

$$(1) = (\boldsymbol{e}_{k_\tau^\star} - \boldsymbol{e}_{k_1})^\top \boldsymbol{U}_\tau \boldsymbol{\Sigma} \boldsymbol{V}_\tau^\top \boldsymbol{V}_\tau \boldsymbol{\Sigma}^\top \boldsymbol{U}_\tau^\top (\boldsymbol{e}_{k_\tau^\star} - \boldsymbol{e}_{k_2}) = (\boldsymbol{e}_{k_\tau^\star} - \boldsymbol{e}_{k_1})^\top \boldsymbol{U}_\tau \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \boldsymbol{U}_\tau^\top (\boldsymbol{e}_{k_\tau^\star} - \boldsymbol{e}_{k_2})$$

This can be regarded as a function of $\boldsymbol{U}_\tau$, which can use Lemma 16. By Lemma 17, we have

$$\mathbb{E}[\boldsymbol{U}_\tau \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \boldsymbol{U}_\tau^\top | \boldsymbol{U}_\tau \in S] = \frac{1}{K}\|\boldsymbol{\Sigma}\|_F^2 \boldsymbol{I}_K$$
$$\mathbb{E}[(\boldsymbol{e}_{k_\tau^\star} - \boldsymbol{e}_{k_1})^\top \boldsymbol{U}_\tau \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \boldsymbol{U}_\tau^\top (\boldsymbol{e}_{k_\tau^\star} - \boldsymbol{e}_{k_2}) | \boldsymbol{U}_\tau \in S] = \frac{1}{K}\|\boldsymbol{\Sigma}\|_F^2$$

where S denotes either one of $\mathbb{SO}(K)$ or $\mathbb{SO}^-(K)$. Lipschitz constant can be calculated as follows:

$$\begin{aligned}
&|\boldsymbol{v}^\top \boldsymbol{U}_\tau \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \boldsymbol{U}_\tau^\top \boldsymbol{w} - \boldsymbol{v}^\top \boldsymbol{U}_\tau' \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \boldsymbol{U}_\tau^{\top\prime} \boldsymbol{w}| \\
&\le |\boldsymbol{v}^\top (\boldsymbol{U}_\tau - \boldsymbol{U}_\tau') \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \boldsymbol{U}_\tau^\top \boldsymbol{w}| + |\boldsymbol{v}^\top \boldsymbol{U}_\tau' \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top (\boldsymbol{U}_\tau - \boldsymbol{U}_\tau')^{\top\prime} \boldsymbol{w}| \\
&\le 2\sigma_1^2 \|\boldsymbol{v}\|\|\boldsymbol{w}\|\|\boldsymbol{U}_\tau - \boldsymbol{U}_\tau'\|_{op} \le 4\sigma_1^2 d(\boldsymbol{U}_\tau, \boldsymbol{U}_\tau').
\end{aligned}$$

Hence, with probability at least $1-\delta$, for all τ and $k_1 \neq k_2 \in [K]$, there exists an absolute constant $c_4 > 0$ such that

$$|\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k_1}, \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k_2} \rangle - \frac{1}{K}\|\boldsymbol{\Sigma}\|_F^2| \le c_4 \frac{\sigma_1^2}{\sqrt{K}} \sqrt{\log\left(\frac{B^2K^2}{\delta}\right)}.$$

The term (2), (3) can be easily bounded via sub-Gaussianity.

$$|\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k_1}, \boldsymbol{z} + \boldsymbol{z}_1 \rangle| \vee |\langle \boldsymbol{z} + \boldsymbol{z}_2, \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k_2} \rangle| \le c_5 \sigma_1 \log\left(\frac{B^2K^2}{\delta}\right).$$

The term (4) can be decomposed as:

$$|\langle \boldsymbol{z} + \boldsymbol{z}_1, \boldsymbol{z} + \boldsymbol{z}_2 \rangle| \le \|\boldsymbol{z}\|^2 + |\langle \boldsymbol{z}_1, \boldsymbol{z} \rangle| + |\langle \boldsymbol{z}, \boldsymbol{z}_2 \rangle| + |\langle \boldsymbol{z}_1, \boldsymbol{z}_2 \rangle|.$$

The last three terms can be bounded by $c_6\sqrt{d}\log\left(\frac{B^2K^2}{\delta}\right)$, which is same as the previous one. The first term follows the chi-square distribution, which is well-known as sub-exponential random variable. Note that $\mathbb{E}[\|\boldsymbol{z}\|^2] = d$, from Theorem 6.2.1. of Vershynin [26],

$$\mathbb{P}\left(|\|\boldsymbol{z}\|^2 - d| \ge t\right) \le 2\exp\left(-c\min\{t^2/d, t\}\right).$$

Hence, with probability at least $1-\delta$, for all $\tau \in [B]$,

$$|\|\boldsymbol{z}\|^2 - d| \le c_7 \left[\sqrt{d\log\left(\frac{B^2K^2}{\delta}\right)} + \log\left(\frac{B^2K^2}{\delta}\right)\right].$$

To sum up,

$$\begin{aligned}
&|\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_1}, \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_2} \rangle - \frac{S_\star^2}{N^2K}\|\boldsymbol{\Sigma}\|_F^2| \\
&\le \frac{dS_\star}{N^2} + C_1\left(\frac{S_\star^2\sigma_1^2}{N^2\sqrt{K}} + \frac{\sigma_1 S_\star\sqrt{S_\star}}{N^2} + \frac{S_\star\sqrt{d}}{N^2}\right)\log\left(\frac{B^2K^2}{\delta}\right).
\end{aligned}$$

3. Bound of $|\|\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}\|^2 - \frac{2S_\star^2}{N^2K}\|\boldsymbol{\Sigma}\|_F^2|$

   Similarly,

$$\|\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}\|^2 = \frac{S_\star^2}{N^2}\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k}, \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k}\rangle + \frac{2S_\star\sqrt{S_\star}}{N^2}\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k}, \boldsymbol{z}\rangle + \frac{S_\star}{N^2}\|\boldsymbol{z}\|^2$$

   Following the same approach as 2, we can obtain the probabilistic bound as follows:

$$|\|\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}\|^2 - \frac{2S_\star^2}{N^2K}\|\boldsymbol{\Sigma}\|_F^2| \le \frac{dS_\star}{N^2} + C_2\left(\frac{S_\star^2\sigma_1^2}{N^2\sqrt{K}} + \frac{\sigma_1 S_\star\sqrt{S_\star}}{N^2} + \frac{S_\star\sqrt{d}}{N^2}\right)\log\left(\frac{B^2K^2}{\delta}\right)$$

   The different points from 2 are that the vectors in the quadratic form are the same, and the bound for Gaussian noise turns into a chi-square bound.

4. Bound of $|\langle \boldsymbol{x}_\tau^{(N+1)}, \boldsymbol{x}_{\tau'}^{(N+1)}\rangle|$

   Recall $\boldsymbol{x}_\tau^{(N+1)} \overset{d}{=} \boldsymbol{\mu}_{\tau,k_\tau^\star} + \boldsymbol{z}$, where $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Hence,

$$|\langle \boldsymbol{x}_\tau^{(N+1)}, \boldsymbol{x}_{\tau'}^{(N+1)}\rangle| \le |\langle \boldsymbol{\mu}_{\tau,k_\tau^\star}, \boldsymbol{\mu}_{\tau',k_{\tau'}^\star}\rangle| + |\langle \boldsymbol{\mu}_{\tau,k_\tau^\star}, \boldsymbol{z}'\rangle| + |\langle \boldsymbol{z}, \boldsymbol{\mu}_{\tau',k_{\tau'}^\star}\rangle| + |\langle \boldsymbol{z}, \boldsymbol{z}'\rangle|$$

   Since

$$\begin{aligned}
&|\langle \boldsymbol{\mu}_{\tau,k_\tau^\star}, \boldsymbol{\mu}_{\tau',k_{\tau'}^\star}\rangle| \le c_8\frac{\sigma_1^2}{\sqrt{d}}\sqrt{\log\left(\frac{B^2K^2}{\delta}\right)},\\
&|\langle \boldsymbol{\mu}_{\tau,k_\tau^\star}, \boldsymbol{z}'\rangle| \vee |\langle \boldsymbol{z}, \boldsymbol{\mu}_{\tau',k_{\tau'}^\star}\rangle| \le c_9\sigma_1\log\left(\frac{B^2K^2}{\delta}\right),\\
&|\langle \boldsymbol{z}, \boldsymbol{z}'\rangle| \le c_{10}\sqrt{d}\log\left(\frac{B^2K^2}{\delta}\right)
\end{aligned}$$

   for some constant $c_8, c_9, c_{10} > 0$. The first inequality is from similar approach in 1. To sum up, we can obtain:

$$|\langle \boldsymbol{x}_\tau^{(N+1)}, \boldsymbol{x}_{\tau'}^{(N+1)}\rangle| \le C_3\left(\frac{\sigma_1^2}{\sqrt{d}} + \sigma_1 + \sqrt{d}\right)\log\left(\frac{B^2K^2}{\delta}\right).$$

5. Bound of $|\|\boldsymbol{x}_\tau^{(N+1)}\|^2 - \frac{1}{K}\|\boldsymbol{\Sigma}\|_F^2|$

   Note that $\|\boldsymbol{x}_\tau^{(N+1)}\|^2 = \|\boldsymbol{\mu}_{\tau,k_\tau^\star}\|^2 + 2\langle \boldsymbol{\mu}_{\tau,k_\tau^\star}, \boldsymbol{z}\rangle + \|\boldsymbol{z}\|^2$. Similar to the previous approach, there exist some constants c_{11}, c_{12}, c_{13} such that:

$$\begin{aligned}
&|\|\boldsymbol{\mu}_{\tau,k_\tau^\star}\|^2 - \frac{1}{K}\|\boldsymbol{\Sigma}\|_F^2| \le c_{11}\frac{\sigma_1^2}{\sqrt{K}}\sqrt{\log\left(\frac{B^2K^2}{\delta}\right)},\\
&|\langle \boldsymbol{\mu}_{\tau,k_\tau^\star}, \boldsymbol{z}\rangle| \le c_{12}\sigma_1\log\left(\frac{B^2K^2}{\delta}\right),\\
&|\|\boldsymbol{z}\|^2 - d| \le c_{13}\left[\sqrt{d\log\left(\frac{B^2K^2}{\delta}\right)} + \log\left(\frac{B^2K^2}{\delta}\right)\right].
\end{aligned}$$

Therefore,

$$|\|\boldsymbol{x}_\tau^{(N+1)}\|^2 - \frac{1}{K}\|\boldsymbol{\Sigma}\|_F^2| \le d + C_4\left(\frac{\sigma_1^2}{\sqrt{K}} + \sigma_1 + \sqrt{d}\right)\log\left(\frac{B^2K^2}{\delta}\right).$$

6. Bound of $|\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}, \boldsymbol{x}_\tau^{(N+1)}\rangle - \frac{S_\star}{NK}\|\boldsymbol{\Sigma}\|_F^2|$

   Rewrite $\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}, \boldsymbol{x}_\tau^{(N+1)}\rangle$ with standard Gaussian noise $\boldsymbol{z}, \boldsymbol{z}'$,

$$\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}, \boldsymbol{x}_\tau^{(N+1)}\rangle = \frac{S_\star}{N}\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k}, \boldsymbol{\mu}_{\tau,k_\tau^\star}\rangle$$
$$+ \frac{S_\star}{N}\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k}, \boldsymbol{z}'\rangle + \frac{\sqrt{2S_\star}}{N}\langle \boldsymbol{z}, \boldsymbol{\mu}_{\tau,k_\tau^\star}\rangle + \frac{\sqrt{2S_\star}}{N}\langle \boldsymbol{z}, \boldsymbol{z}'\rangle.$$

   Following the similar approach to 2, there exist some constants c_{14}, c_{15}, c_{16} such that:

$$|\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k}, \boldsymbol{\mu}_{\tau,k_\tau^\star}\rangle - \frac{1}{K}\|\boldsymbol{\Sigma}\|_F^2| \le c_{14}\frac{\sigma_1^2}{\sqrt{K}}\sqrt{\log\left(\frac{B^2K^2}{\delta}\right)},$$
$$|\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k}, \boldsymbol{z}'\rangle| \vee |\langle \boldsymbol{z}, \sqrt{2}\boldsymbol{\mu}_{\tau,k_\tau^\star}\rangle| \le c_{15}\sigma_1\log\left(\frac{B^2K^2}{\delta}\right),$$
$$|\langle \boldsymbol{z}, \boldsymbol{z}'\rangle| \le c_{16}\sqrt{d}\log\left(\frac{B^2K^2}{\delta}\right).$$

   Hence,

$$|\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}, \boldsymbol{x}_\tau^{(N+1)}\rangle - \frac{S_\star}{NK}\|\boldsymbol{\Sigma}\|_F^2| \le C_5\left(\frac{S_\star\sigma_1^2}{N\sqrt{K}} + \frac{S_\star\sigma_1}{N} + \frac{\sqrt{dS_\star}}{N}\right)\log\left(\frac{B^2K^2}{\delta}\right)$$

Since we've assumed balanced sample(i.e. $N = S_\star K$) and by Assumptions 3 and 4, we can re-write the inequalities as follows:

$$|\langle \hat{\boldsymbol{\mu}}_{\tau,k_{\tau'}^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_1}, \hat{\boldsymbol{\mu}}_{\tau',k_{\tau'}^\star} - \hat{\boldsymbol{\mu}}_{\tau',k_2}\rangle| \le c_0\left(\frac{\sigma_1^2}{K^2\sqrt{d}}\right)\log\left(\frac{B^2K^2}{\delta}\right),$$
$$|\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_1}, \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_2}\rangle - \frac{1}{K^3}\|\boldsymbol{\Sigma}\|_F^2| \le \frac{d}{S_\star K^2} + c_0\left(\frac{\sigma_1^2}{K^2\sqrt{K}}\right)\log\left(\frac{B^2K^2}{\delta}\right),$$
$$|\|\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}\|^2 - \frac{2}{K^3}\|\boldsymbol{\Sigma}\|_F^2| \le \frac{d}{S_\star K^2} + c_0\left(\frac{\sigma_1^2}{K^2\sqrt{K}}\right)\log\left(\frac{B^2K^2}{\delta}\right),$$
$$|\langle \boldsymbol{x}_\tau^{(N+1)}, \boldsymbol{x}_{\tau'}^{(N+1)}\rangle| \le c_0\left(\frac{\sigma_1^2}{\sqrt{d}}\right)\log\left(\frac{B^2K^2}{\delta}\right),$$
$$|\|\boldsymbol{x}_\tau^{(N+1)}\|^2 - \frac{1}{K}\|\boldsymbol{\Sigma}\|_F^2| \le d + c_0\left(\frac{\sigma_1^2}{\sqrt{K}}\right)\log\left(\frac{B^2K^2}{\delta}\right),$$
$$|\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}, \boldsymbol{x}_\tau^{(N+1)}\rangle - \frac{1}{K^2}\|\boldsymbol{\Sigma}\|_F^2| \le c_0\left(\frac{\sigma_1^2}{K\sqrt{K}}\right)\log\left(\frac{B^2K^2}{\delta}\right).$$

■

F.2. KKT condition and bound

Quantity of our interest is $\mathrm{Tr}(\boldsymbol{W}_{MM})/\|\boldsymbol{W}_{MM}\|_F$, which is scale-invariant, thus we focus on establishing an lower bound for it. For convenience, let $\alpha_{\tau,k_\tau^\star} = 0$. By KKT condition, we have max-margin solution as follows:

$$\boldsymbol{W}_{MM} = \sum_{\tau=1}^{B}\sum_{k=1}^{K} \alpha_{\tau,k}(\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k})\boldsymbol{x}_\tau^{(N+1)\top}.$$

Hence,

$$\mathrm{Tr}(\boldsymbol{W}_{MM}) = \mathrm{Tr}\left(\sum_{\tau,k} \alpha_{\tau,k}(\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k})\boldsymbol{x}_\tau^{(N+1)\top}\right) = \mathrm{Tr}\left(\sum_{\tau,k} \alpha_{\tau,k}\langle\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}, \boldsymbol{x}_\tau^{(N+1)}\rangle\right).$$

Recall the last part of Theorem 21,

$$\begin{aligned}\langle\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}, \boldsymbol{x}_\tau^{(N+1)}\rangle &\geq \frac{1}{K^2}\|\boldsymbol{\Sigma}\|_F^2 - c_0\frac{\sigma_1^2}{K\sqrt{K}}\log\left(\frac{B^2K^2}{\delta}\right)\\ &= \frac{\sigma_1^2}{K^2}\left(\mathrm{sr}(\boldsymbol{\Sigma}) - c_0\sqrt{K}\log\left(\frac{B^2K^2}{\delta}\right)\right).\end{aligned}$$

Therefore, we can obtain

$$\mathrm{Tr}(\boldsymbol{W}_{MM}) = \Omega\left(\frac{\sigma_1^2}{K}\sum_{\tau,k}\alpha_{\tau,k}\right).$$

By feasibility of the solution, $\forall q \in [B], s \in [K] \setminus \{k_q^\star\}$

$$\begin{aligned}1 &\leq (\hat{\boldsymbol{\mu}}_{q,k_q^\star} - \hat{\boldsymbol{\mu}}_{q,s})^\top \sum_{\tau=1}^{B}\sum_{k=1}^{K}\alpha_{\tau,k}(\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k})\boldsymbol{x}_\tau^{(N+1)\top}\boldsymbol{x}_q^{(N+1)}\\ &= \alpha_{q,s}\|\hat{\boldsymbol{\mu}}_{q,k_q^\star} - \hat{\boldsymbol{\mu}}_{q,s}\|^2\|\boldsymbol{x}_q^{(N+1)}\|^2 + \sum_{k\neq s}\alpha_{q,k}\langle\hat{\boldsymbol{\mu}}_{q,k_q^\star} - \hat{\boldsymbol{\mu}}_{q,s}, \hat{\boldsymbol{\mu}}_{q,k_q^\star} - \hat{\boldsymbol{\mu}}_{q,k}\rangle\|\boldsymbol{x}_q^{(N+1)}\|^2\\ &\quad + \sum_{\tau\neq q}\sum_{k=1}^{K}\alpha_{\tau,k}\langle\hat{\boldsymbol{\mu}}_{q,k_q^\star} - \hat{\boldsymbol{\mu}}_{q,s}, \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}\rangle\langle\boldsymbol{x}_\tau^{(N+1)}, \boldsymbol{x}_q^{(N+1)}\rangle.\end{aligned}$$

Averaging with respect to the task and clusters, we can obtain

$$\begin{aligned}&\frac{1}{B(K-1)}\sum_{q\in[B],s\neq k_q^\star}\alpha_{q,s}\underbrace{\|\hat{\boldsymbol{\mu}}_{q,k_q^\star} - \hat{\boldsymbol{\mu}}_{q,s}\|^2\|\boldsymbol{x}_q^{(N+1)}\|^2}_{(1)}\\ &+ \frac{1}{B(K-1)}\sum_{q\in[B],s\neq k_q^\star}\sum_{k\neq s}\alpha_{q,k}\underbrace{\langle\hat{\boldsymbol{\mu}}_{q,k_q^\star} - \hat{\boldsymbol{\mu}}_{q,s}, \hat{\boldsymbol{\mu}}_{q,k_q^\star} - \hat{\boldsymbol{\mu}}_{q,k}\rangle\|\boldsymbol{x}_q^{(N+1)}\|^2}_{(2)}\\ &+ \frac{1}{B(K-1)}\sum_{q\in[B],s\neq k_q^\star}\sum_{\tau\neq q}\sum_{k=1}^{K}\alpha_{\tau,k}\underbrace{\langle\hat{\boldsymbol{\mu}}_{q,k_q^\star} - \hat{\boldsymbol{\mu}}_{q,s}, \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}\rangle\langle\boldsymbol{x}_\tau^{(N+1)}, \boldsymbol{x}_q^{(N+1)}\rangle}_{(3)}.\end{aligned}$$

We should establish the upper bound of each terms. Each bounds can be derived by Theorem 21.

1. Bound of (1), (2).

   The bounds of (1) and (2) are identical, except the coefficient of the term related to a stable rank.

$$(1) = \|\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}\|^2 \|\boldsymbol{x}_\tau^{(N+1)}\|^2$$
$$\leq \frac{\sigma_1^4}{K^4}\left(2\,\mathrm{sr}(\boldsymbol{\Sigma}) + \frac{dK}{S_\star \sigma_1^2} + c_0\sqrt{K}\log\left(\frac{B^2K^2}{\delta}\right)\right)\left(\mathrm{sr}(\boldsymbol{\Sigma}) + \frac{dK}{\sigma_1^2} + c_0\sqrt{K}\log\left(\frac{B^2K^2}{\delta}\right)\right).$$

$$(2) \leq |\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_1}, \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_2}\rangle| \|\boldsymbol{x}_\tau^{(N+1)}\|^2$$
$$\leq \frac{\sigma_1^4}{K^4}\left(\mathrm{sr}(\boldsymbol{\Sigma}) + \frac{dK}{S_\star \sigma_1^2} + c_0\sqrt{K}\log\left(\frac{B^2K^2}{\delta}\right)\right)\left(\mathrm{sr}(\boldsymbol{\Sigma}) + \frac{dK}{\sigma_1^2} + c_0\sqrt{K}\log\left(\frac{B^2K^2}{\delta}\right)\right).$$

   Denote $a \in \{1, 2\}$.

$$\frac{\sigma_1^4}{K^4}\left(a\,\mathrm{sr}(\boldsymbol{\Sigma}) + \frac{dK}{S_\star \sigma_1^2} + c_0\sqrt{K}\log\left(\frac{B^2K^2}{\delta}\right)\right)\left(\mathrm{sr}(\boldsymbol{\Sigma}) + \frac{dK}{\sigma_1^2} + c_0\sqrt{K}\log\left(\frac{B^2K^2}{\delta}\right)\right)$$
$$\leq \frac{\sigma_1^4}{K^2}\left(a + \frac{1}{S_\star c_\sigma} + c_0\frac{1}{\sqrt{K}}\log\left(\frac{B^2K^2}{\delta}\right)\right)\left(1 + \frac{1}{c_\sigma} + c_0\frac{1}{\sqrt{K}}\log\left(\frac{B^2K^2}{\delta}\right)\right)$$

   Denote A_1, A_2 as some constants such that $A_1 \geq \left(a + \frac{1}{S_\star c_\delta}\right)\left(1 + \frac{1}{c_\delta}\right)$ and $A_2 \geq a + 1 + \frac{S_\star + 1}{S_\star c_\delta}$, then the last term can be summarized as:

$$\frac{\sigma_1^4}{K^2}\left[A_1 + A_2 c_0 \frac{1}{\sqrt{K}}\log\left(\frac{B^2K^2}{\delta}\right) + c_0^2\frac{1}{K}\log^2\left(\frac{B^2K^2}{\delta}\right)\right].$$

   Note that stable rank cannot exceed the rank: $\mathrm{sr}(\boldsymbol{\Sigma}) \leq K$. Indeed, this bound is tight up to a constant because of Assumption 2.

2. Bound of (3).

$$(3) \leq |\langle \hat{\boldsymbol{\mu}}_{\tau,k_{\tau'}^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_1}, \hat{\boldsymbol{\mu}}_{\tau',k_{\tau'}^\star} - \hat{\boldsymbol{\mu}}_{\tau',k_2}\rangle| |\langle \boldsymbol{x}_\tau^{(N+1)}, \boldsymbol{x}_{\tau'}^{(N+1)}\rangle| \leq c_0^2 \frac{\sigma_1^4}{dK^2}\log^2\left(\frac{B^2K^2}{\delta}\right).$$

The contribution of $\alpha_{\tau,k}$ from (1) is counted once, whereas (2) and (3) contribute $K-1$ and $(B-1)K$ terms, respectively. Combining these terms yields

$$1 \leq \frac{1}{B(K-1)}\sum_{\tau,k}\alpha_{\tau,k}\left[\frac{\sigma_1^4}{K}\left[A_1 + A_2c_0\frac{1}{\sqrt{K}}\log\left(\frac{B^2K^2}{\delta}\right) + c_0^2\frac{1}{K}\log^2\left(\frac{B^2K^2}{\delta}\right) + c_0^2\frac{B-1}{d}\log^2\left(\frac{B^2K^2}{\delta}\right)\right]\right]$$
$$\leq \left(\sum_{\tau,k}\alpha_{\tau,k}\right)\left(\frac{\sigma_1^4}{dK(K-1)}\left[\frac{d}{B}A_1 + \frac{d}{B\sqrt{K}}A_2c_0\log\left(\frac{B^2K^2}{\delta}\right) + c_0^2\left(1 + \frac{d}{BK}\right)\log^2\left(\frac{B^2K^2}{\delta}\right)\right]\right).$$

Denote $\xi = \frac{d}{B}A_1 + \frac{d}{B\sqrt{K}}A_2c_0\log\left(\frac{B^2K^2}{\delta}\right) + c_0^2\left(1 + \frac{d}{BK}\right)\log^2\left(\frac{B^2K^2}{\delta}\right)$, we can observe that $\xi = O\left(\frac{d}{B} \vee \log^2\left(\frac{B^2K^2}{\delta}\right)\right)$.

$$\sum_{\tau,k}\alpha_{\tau,k} \geq \frac{dK(K-1)}{\sigma_1^4\xi}, \quad \mathrm{Tr}(\boldsymbol{W}_{MM}) = \Omega\left(\frac{dK}{\sigma_1^2\xi}\right).$$

Since $\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}, \boldsymbol{x}_\tau^{(N+1)} \rangle \geq \frac{\sigma_1^2}{K}\left(c_s - \frac{c_0}{\sqrt{K}}\log\left(\frac{B^2K^2}{\delta}\right)\right)$, We can easily verify that $\boldsymbol{W}_0 = \frac{K}{\sigma_1^2\left(c_s - \frac{c_0}{\sqrt{K}}\log\left(\frac{B^2K^2}{\delta}\right)\right)}\boldsymbol{I}_d$ satisfies the condition of the max-margin problem. Hence,

$$\|\boldsymbol{W}_{MM}\|_F \leq \|\boldsymbol{W}_0\|_F = O\left(\frac{K\sqrt{d}}{\sigma_1^2}\right).$$

Therefore, $\mathrm{Tr}(\boldsymbol{W}_{MM})/\|\boldsymbol{W}_{MM}\|_F = \Omega(\sqrt{d}\xi^{-1})$.

Appendix G. Proof of Theorem 5

For notational simplicity, denote $\boldsymbol{W}$ as max-margin solution. Our goal is to upper bound $\mathbb{P}\left(\hat{\boldsymbol{\mu}}_k^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)} > \hat{\boldsymbol{\mu}}_{k^\star}^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)}\right)$. Note that $\hat{\boldsymbol{\mu}}_k \overset{d}{=} \frac{S}{M}\boldsymbol{\mu}_k + \frac{\sqrt{S}}{M}\boldsymbol{z}_k$ where $\boldsymbol{z}_k$s are i.i.d. gaussian random vectors from $\mathcal{N}(0, \boldsymbol{I}_d)$. Hence,

$$\frac{M}{S}(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_{k^\star})\boldsymbol{W}\boldsymbol{x}^{(M+1)} = \left(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star} + \sqrt{\frac{2}{S}}\boldsymbol{z}\right)^\top \boldsymbol{W}\left(\boldsymbol{\mu}_{k^\star} + \boldsymbol{z}'\right)$$
$$= (\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star} + (\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{z}' + \sqrt{\frac{2}{S}}\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star} + \sqrt{\frac{2}{S}}\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{z}'.$$

Therefore,

$$\hat{\boldsymbol{\mu}}_k^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)} > \hat{\boldsymbol{\mu}}_{k^\star}^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)}$$
$$\Leftrightarrow -(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star} < (\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{z}' + \sqrt{\frac{2}{S}}\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star} + \sqrt{\frac{2}{S}}\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{z}'.$$

Remark 23 *For some $t > 0$, and random variables $\boldsymbol{X}_1,\ \boldsymbol{X}_2,\ \boldsymbol{X}_3$, and $\boldsymbol{X}_4$,*

$$\mathbb{P}\left(-\boldsymbol{X}_1 < \boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_4\right) \le \mathbb{P}\left(-\boldsymbol{X}_1 < t \cup |\boldsymbol{X}_2| \ge \frac{t}{4} \cup |\boldsymbol{X}_3| \ge \frac{t}{4} \cup |\boldsymbol{X}_4| \ge \frac{t}{4}\right)$$
$$\le \mathbb{P}\left(\boldsymbol{X}_1 > -t\right) + \mathbb{P}(|\boldsymbol{X}_2| \ge \frac{t}{4}) + \mathbb{P}(|\boldsymbol{X}_3| \ge \frac{t}{4}) + \mathbb{P}(|\boldsymbol{X}_4| \ge \frac{t}{4})$$

By Remark 23,

$$\mathbb{P}\left(\hat{\boldsymbol{\mu}}_k^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)} > \hat{\boldsymbol{\mu}}_{k^\star}^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)}\right)$$
$$\le \mathbb{P}\left((\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star} \ge -\frac{\mathrm{Tr}(\boldsymbol{W})}{2dK}\|\boldsymbol{\Lambda}\|_F^2\right) + \mathbb{P}\left(|(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{z}'| \ge \frac{\mathrm{Tr}(\boldsymbol{W})}{8dK}\|\boldsymbol{\Lambda}\|_F^2\right)$$
$$+ \mathbb{P}\left(\sqrt{\frac{2}{S}}|\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star}| \ge \frac{\mathrm{Tr}(\boldsymbol{W})}{8dK}\|\boldsymbol{\Lambda}\|_F^2\right) + \mathbb{P}\left(\sqrt{\frac{2}{S}}|\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{z}'| \ge \frac{\mathrm{Tr}(\boldsymbol{W})}{8dK}\|\boldsymbol{\Lambda}\|_F^2\right).$$

Applying Lemma 17, we can obtain:

$$\mathbb{E}\left[(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star}\right] = -\frac{\mathrm{Tr}(\boldsymbol{W})}{dK}\|\boldsymbol{\Lambda}\|_F^2.$$

We can re-write

$$\mathbb{P}\left((\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star} \ge -\frac{\mathrm{Tr}(\boldsymbol{W})}{2dK}\|\boldsymbol{\Lambda}\|_F^2\right)$$
$$= \mathbb{P}\left((\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star} \ge \mathbb{E}\left[(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star}\right] + \frac{\mathrm{Tr}(\boldsymbol{W})}{2dK}\|\boldsymbol{\Lambda}\|_F^2\right).$$

Therefore, by applying Lemma 20,

$$\mathbb{P}\left((\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star} \ge -\frac{\mathrm{Tr}(\boldsymbol{W})}{2dK}\|\boldsymbol{\Lambda}\|_F^2\right)$$
$$\le C\exp\left(-c'\min\left(\frac{\mathrm{Tr}(\boldsymbol{W})^2\,\mathrm{sr}(\boldsymbol{\Lambda})^2}{\|\boldsymbol{W}\|_F^2 K^2}, \frac{\mathrm{Tr}(\boldsymbol{W})\,\mathrm{sr}(\boldsymbol{\Lambda})}{\|\boldsymbol{W}\|_F K}\right)\right) + \exp\left(-\frac{c'\,\mathrm{sr}(\boldsymbol{\Lambda})^2}{K}\right).$$

Using the fact $\frac{\mathrm{Tr}(\boldsymbol{W})}{\|\boldsymbol{W}\|_F} = \Omega\left(\sqrt{d}\xi^{-1}\right)$, we can obtain upper bound as:

$$C\exp\left(-c\left(\frac{d\,\mathrm{sr}(\boldsymbol{\Lambda})}{\xi^2 K} \wedge \frac{\sqrt{d}}{\xi} \wedge \mathrm{sr}(\boldsymbol{\Lambda})\right)\cdot\frac{\mathrm{sr}(\boldsymbol{\Lambda})}{K}\right).$$

The last three terms can be bounded by

$$\mathbb{P}\left(|(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{z}'| \geq \frac{\mathrm{Tr}(\boldsymbol{W})}{8dK}\|\boldsymbol{\Lambda}\|_F^2\right) \leq 2\exp\left(-\frac{c\lambda_1\,\mathrm{sr}(\boldsymbol{\Lambda})}{\xi K}\right),$$

$$\mathbb{P}\left(\sqrt{\frac{2}{S}}|\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star}| \geq \frac{\mathrm{Tr}(\boldsymbol{W})}{8dK}\|\boldsymbol{\Lambda}\|_F^2\right) \leq 2\exp\left(-\frac{c\sqrt{S}\lambda_1\,\mathrm{sr}(\boldsymbol{\Lambda})}{\xi K}\right),$$

$$\mathbb{P}\left(\sqrt{\frac{2}{S}}|\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{z}'| \geq \frac{\mathrm{Tr}(\boldsymbol{W})}{8dK}\|\boldsymbol{\Lambda}\|_F^2\right) \leq 2\exp\left(-\frac{c\sqrt{S}\|\boldsymbol{\Lambda}\|_F^2}{\xi K\sqrt{d}}\right).$$

To sum up,

$$\begin{aligned}&\mathbb{P}\left(\hat{\boldsymbol{\mu}}_k^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)} > \hat{\boldsymbol{\mu}}_{k^\star}^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)}\right)\\ &\leq C\exp\left(-c\left(\frac{d\,\mathrm{sr}(\boldsymbol{\Lambda})}{\xi^2 K} \wedge \frac{\sqrt{d}}{\xi} \wedge \mathrm{sr}(\boldsymbol{\Lambda})\right)\cdot\frac{\mathrm{sr}(\boldsymbol{\Lambda})}{K}\right) + 6\exp\left(-\left(\frac{\lambda_1\,\mathrm{sr}(\boldsymbol{\Lambda})}{\xi K} \wedge \sqrt{\frac{S\|\boldsymbol{\Lambda}\|_F^4}{d\xi^2 K^2}}\right)\right).\end{aligned}$$

G.1. Dicussions on Theorem 5

Specifically, the theorem provides an upper bound on the probability of misclassifying the label as a fixed class k. Extending this result to all classes via a union bound introduces a multiplicative factor of K in the error probability; however, this corresponds to only an additive $\log K$ term in the exponent. Hence, it does not significantly affect the scaling of the sample complexity. Substituting the definition of ξ into the bound, d appears in the denominator of the exponent; therefore, higher dimension d leads to poorer generalization performance when other parameters are fixed.

To achieve a small constant error, both terms must be controlled. The first term arises from the test data generation process and depends on the geometry of the class centers, while the second term arises from the Gaussian noise and depends on the scale of the class vectors as well as the number of examples per class S. Controlling these terms imposes requirements on the problem parameters. In particular, both terms require aggregating sufficient information across training tasks, which necessitates a sufficiently large number of tasks B. The sufficient condition is $B = \Omega(d)$, under which ξ scales only logarithmically.

The second term depends on the signal strength and the number of in-context examples in the test task. A sufficient regime that controls this term is given by $B = \Omega(d)$ and $\lambda_1\,\mathrm{sr}(\boldsymbol{\Lambda})/(\xi K) = \Omega(1)$. Under this regime, the remaining requirement comes from controlling the expression involving S. In particular, ensuring that the second term contributes only a small constant to the error requires $S = \tilde{\Omega}\left(dK^2/\|\boldsymbol{\Lambda}\|_F^4\right)$. This characterizes the required scaling of S, which—up to logarithmic factors—matches the SNR-based lower bound in Appendix C. Overall, under these assumptions, transformers achieve nearly optimal sample complexity as meta-learners for multiclass classification, extending the main result of Magen and Vardi [15] from binary to multiclass settings.

When the stable rank is low, class centers become poorly separated, making classification more difficult. Under sufficiently large number of tasks B (i.e. $B = \Omega(d)$), the sufficient condition $\mathrm{sr}(\mathbf{\Lambda}) = \tilde{\Omega}(\sqrt{K})$ arises from the first term. This is strictly weaker than the requirement imposed on the singular values of training tasks $\mathrm{sr}(\mathbf{\Sigma}) = \Theta(K)$ (Assumption 2). This suggests that transformers can robustly perform in-context learning even when the class-center matrix of a new test prompt exhibits anisotropic class geometry due to distribution shifts.

Overall, Theorem 5 characterizes the interplay between error bounds, sample size, and geometry. In particular, a larger S and a higher stable rank of the test task reduce the classification error.

Appendix H. Proof of Theorem 6

Recall S_k is number of samples for a corresponding cluster k, and recall that $\hat{\boldsymbol{\mu}}_i \overset{d}{=} \frac{S_k}{M}\boldsymbol{\mu}_k + \frac{\sqrt{S_k}}{M}\boldsymbol{z}_i$. We want to derive the upper bound of the failure probability, that is

$$\mathbb{P}\left(\hat{\boldsymbol{\mu}}_k^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)} > \hat{\boldsymbol{\mu}}_{k^\star}^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)}\right)$$

under the max-margin solution $\boldsymbol{W}$. Therefore, the condition $\hat{\boldsymbol{\mu}}_k^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)} > \hat{\boldsymbol{\mu}}_{k^\star}^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)}$ is equivalent to

$$-(S_k\boldsymbol{\mu}_k - S_{k^\star}\boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star} < (S_k\boldsymbol{\mu}_k - S_{k^\star}\boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{z}' + \sqrt{S_k + S_{k^\star}}\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star} + \sqrt{S_k + S_{k^\star}}\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{z}'.$$

By Lemma 17,

$$\mathbb{E}\left[(S_k\boldsymbol{\mu}_k - S_{k^\star}\boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star}\right] = -\frac{S_{k^\star}\operatorname{Tr}(\boldsymbol{W})}{dK}\|\boldsymbol{\Lambda}\|_F^2.$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\hat{\boldsymbol{\mu}}_k^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)} > \hat{\boldsymbol{\mu}}_{k^\star}^\top \boldsymbol{W}\boldsymbol{x}^{(M+1)}\right)\\
&\le \mathbb{P}\left((S_k\boldsymbol{\mu}_k - S_{k^\star}\boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star} \ge -\frac{S_{k^\star}\operatorname{Tr}(\boldsymbol{W})}{2dK}\|\boldsymbol{\Lambda}\|_F^2\right)\\
&\quad + \mathbb{P}\left(|(S_k\boldsymbol{\mu}_k - S_{k^\star}\boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{z}'| \ge \frac{S_{k^\star}\operatorname{Tr}(\boldsymbol{W})}{8dK}\|\boldsymbol{\Lambda}\|_F^2\right)\\
&\quad + \mathbb{P}\left(|\sqrt{S_k + S_{k^\star}}\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star}| \ge \frac{S_{k^\star}\operatorname{Tr}(\boldsymbol{W})}{8dK}\|\boldsymbol{\Lambda}\|_F^2\right)\\
&\quad + \mathbb{P}\left(|\sqrt{S_k + S_{k^\star}}\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{z}'| \ge \frac{S_{k^\star}\operatorname{Tr}(\boldsymbol{W})}{8dK}\|\boldsymbol{\Lambda}\|_F^2\right).
\end{aligned}$$

Consider the first term, define

$$\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) := (S_k\boldsymbol{\mu}_k - S_{k^\star}\boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star} = (S_k\boldsymbol{e}_k - S_{k^\star}\boldsymbol{e}_{k^\star})^\top \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top \boldsymbol{W}\boldsymbol{V}\boldsymbol{\Lambda}^\top \boldsymbol{U}^\top \boldsymbol{e}_{k^\star}.$$

$$\begin{aligned}
&\mathbb{P}\left(\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) \ge \mathbb{E}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) + t\right)\\
&\le \mathbb{P}\left(\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) \ge \mathbb{E}_{\boldsymbol{V}}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) + \frac{t}{2}\right) + \mathbb{P}\left(\mathbb{E}_{\boldsymbol{V}}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) \ge \mathbb{E}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) + \frac{t}{2}\right).
\end{aligned}$$

By Lemma 19,

$$\begin{aligned}
&\mathbb{P}\left(\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) \ge \mathbb{E}_{\boldsymbol{V}}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) + \frac{t}{2} \mid \boldsymbol{U}\right)\\
&\le C\exp\left(-c\min\left(\frac{d^2t^2}{\|S_k\boldsymbol{\mu}_k - S_{k^\star}\boldsymbol{\mu}_{k^\star}\|^2\|\boldsymbol{\mu}_{k^\star}\|^2\|\boldsymbol{W}\|_F^2}, \frac{dt}{\|S_k\boldsymbol{\mu}_k - S_{k^\star}\boldsymbol{\mu}_{k^\star}\|\|\boldsymbol{\mu}_{k^\star}\|\|\boldsymbol{W}\|}\right)\right)\\
&\le C\exp\left(-c\min\left(\frac{d^2t^2}{(S_k^2 + S_{k^\star}^2)\lambda_1^4\|\boldsymbol{W}\|_F^2}, \frac{dt}{\sqrt{S_k^2 + S_{k^\star}^2}\lambda_1^2\|\boldsymbol{W}\|}\right)\right).
\end{aligned}$$

Note that $\mathbb{E}_{\boldsymbol{V}}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) = \frac{\mathrm{Tr}(\boldsymbol{W})}{d}(S_k\boldsymbol{e}_k - S_{k^\star}\boldsymbol{e}_{k^\star})^\top\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\boldsymbol{U}^\top\boldsymbol{e}_{k^\star}$ has a Lipshitz constant which is upper bounded by $2\sqrt{S_k^2+S_{k^\star}^2}\frac{\mathrm{Tr}(\boldsymbol{W})}{d}\lambda_1^2$; hence,

$$\mathbb{P}\left(\mathbb{E}_{\boldsymbol{V}}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) \geq \mathbb{E}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) + \frac{t}{2}\right) \leq \exp\left(-c\frac{d^2Kt^2}{4(S_k^2+S_{k^\star}^2)\lambda_1^4\,\mathrm{Tr}(\boldsymbol{W})^2}\right).$$

Let $t = \frac{S_{k^\star}\,\mathrm{Tr}(\boldsymbol{W})}{2dK}\|\boldsymbol{\Lambda}\|_F^2$, and recall the fact that $\frac{\mathrm{Tr}(\boldsymbol{W})}{\|\boldsymbol{W}\|_F} = \Omega\left(\sqrt{d}\xi^{-1}\right)$, we can obtain the concentration on the first term as follows:

$$\begin{aligned}&\mathbb{P}\left(\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) \geq \mathbb{E}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) + t\right)\\ &\leq C\exp\left(-c\left(\frac{S_{k^\star}d\,\mathrm{sr}(\boldsymbol{\Lambda})}{\sqrt{S_k^2+S_{k^\star}^2}\xi^2K} \wedge \frac{\sqrt{d}}{\xi} \wedge \frac{S_{k^\star}\,\mathrm{sr}(\boldsymbol{\Lambda})}{\sqrt{S_k^2+S_{k^\star}^2}}\right)\cdot\frac{S_{k^\star}\,\mathrm{sr}(\boldsymbol{\Lambda})}{\sqrt{S_k^2+S_{k^\star}^2}K}\right)\end{aligned}$$

for some constant $C, c > 0$. For other terms,

$$\begin{aligned}\mathbb{P}\left(|(S_k\boldsymbol{\mu}_k - S_{k^\star}\boldsymbol{\mu}_{k^\star})^\top\boldsymbol{W}\boldsymbol{z}'| \geq t\right) &\leq 2\exp\left(-\frac{ct\sqrt{d}}{\sqrt{S_k^2+S_{k^\star}^2}\lambda_1\|\boldsymbol{W}\|_F}\right),\\ \mathbb{P}\left(|\sqrt{S_k+S_{k^\star}}\boldsymbol{z}^\top\boldsymbol{W}\boldsymbol{\mu}_y| \geq t\right) &\leq 2\exp\left(-\frac{ct\sqrt{d}}{\sqrt{S_k+S_{k^\star}}\lambda_1\|\boldsymbol{W}\|_F}\right),\\ \mathbb{P}\left(|\sqrt{S_k+S_{k^\star}}\boldsymbol{z}^\top\boldsymbol{W}\boldsymbol{z}'| \geq t\right) &\leq 2\exp\left(-\frac{ct}{\sqrt{S_k+S_{k^\star}}\|\boldsymbol{W}\|_F}\right).\end{aligned}$$

Plugging $t = \frac{S_{k^\star}\,\mathrm{Tr}(\boldsymbol{W})}{8dK}\|\boldsymbol{\Lambda}\|_F^2$, the sum of last three terms can be bounded as

$$6\exp\left(-c\left(\frac{S_{k^\star}\lambda_1\,\mathrm{sr}(\boldsymbol{\Lambda})}{\sqrt{S_k^2+S_{k^\star}^2}\xi K} \wedge \frac{S_{k^\star}}{\sqrt{S_k+S_{k^\star}}}\cdot\frac{\|\boldsymbol{\Lambda}\|_F^2}{\sqrt{d}\xi K}\right)\right).$$

H.1. Discussions on Theorem 6

The factor $\psi_{k,k^\star}$ captures the effect of test-time label imbalance on the error bound toward class k. When class k is the majority class, $\psi_{k,k^\star}$ becomes small, which makes the upper bound larger. As a result, the corresponding bound on the probability of misclassifying the label as class k becomes larger, suggesting an increased tendency toward majority-label bias in favor of class k.

Importantly, this imbalance factor $\psi_{k,k^\star}$ interacts multiplicatively with the stable rank $\mathrm{sr}(\boldsymbol{\Lambda})$. While a small $\psi_{k,k^\star}$ reflects an adversarial label distribution, a large stable rank counteracts this effect. In this sense, the stable rank serves as a geometric buffer against label imbalance. Consequently, even under severe label imbalance, a sufficiently large stable rank can preserve a meaningful generalization guarantee. This suggests that robustness to majority-label bias improves as the stable rank increases.

This interpretation is further supported by empirical evidence from prior work. In particular, Gupta et al. [10] observe that LLMs exhibit increased robustness to majority-label bias as model size increases. Our analysis provides a possible geometric explanation for this phenomenon: larger models may induce class-center embeddings with higher stable rank $\mathrm{sr}(\boldsymbol{\Lambda})$, which in turn mitigates the adverse effects of label imbalance, leading to improved robustness against majority-label bias.

Appendix I. Experiment Details

We perform experiments using NVIDIA GeForce RTX 3090 GPUs.

I.1. Nonlinear Transformer Models for Under Our Data Distribution

We conduct experiments on in-context multiclass classification using a Transformer architecture, where the dataset is synthetically generated from our data-generating process (Definition 1). We consider input dimension $d = 20$ and number of classes $K \in \{5, 10\}$. The model is a Transformer with 2 attention heads and model dimension 128, without positional encoding. We vary the depth across 1, 3, and 5 layers. Each layer uses GELU activations, a feedforward network (FFN) with hidden dimension 512, and dropout with rate 0.1.

For training, we use a batch size of 128 and train for 10,000 steps using the AdamW optimizer with a learning rate of 10^{-4}. The spectrum of the class center matrix is chosen to control the stable rank. Specifically, for $K = 10$, we set $\mathbf{\Sigma} = [2, 2, 2, 1.5, 1.5, 1.5, 1.5, 1, 1, 1] \cdot \sqrt{20}$ and for $K = 5$, $\mathbf{\Sigma} = [2, \sqrt{2}, \sqrt{2}, \sqrt{2}, \sqrt{2}] \cdot \sqrt{20}$. In both cases, the resulting stable rank is approximately $0.6K$ to reflect the assumption 2. The noise is sampled as $\mathcal{N}(0, I_d)$, and we use 10 labels per class.

For evaluation, we fix the input dimension while varying the stable rank under a controlled total variation of class center vectors. Specifically, we fix the squared Frobenius norm of the class center matrix to either 120 ($K = 5$) or 480 ($K = 10$), and adjust the singular values accordingly. Given a target stable rank r, we set singular values spectrum as

$$\lambda_1^2 = \frac{\|\mathbf{\Lambda}\|_F^2}{r}, \quad \lambda_{2:K}^2 = \frac{\|\mathbf{\Lambda}\|_F^2 - \lambda_1^2}{K-1}.$$

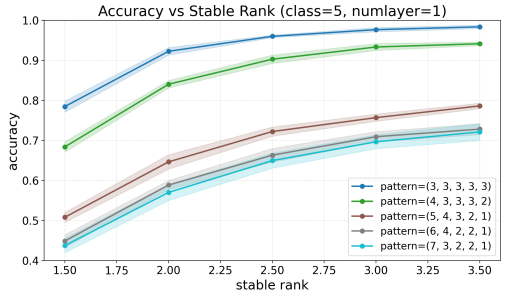
For $K = 5$, the stable rank ranges from 1.5 to 3.5 with a step size of 0.5, and for $K = 10$, from 1.5 to 6 with the same increment. Each evaluation consists of 1,000 examples, and all results are averaged over 10 independent runs with different random seeds. For the imbalance setting, we evaluate the model under various label allocation patterns across classes. The specific configurations are summarized in Table 1.

Table 1: Label allocation patterns for different numbers of classes K. Each pattern specifies the number of examples per class in the context.

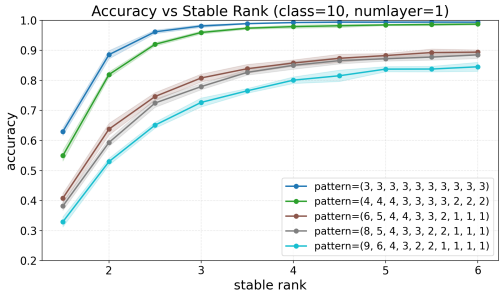
Pattern	$K = 5$	$K = 10$
Pattern 1	[3,3,3,3,3]	[3,3,3,3,3,3,3,3,3,3]
Pattern 2	[4,3,3,3,2]	[4,4,4,3,3,3,3,2,2,2]
Pattern 3	[5,4,3,2,1]	[6,5,4,4,3,3,2,1,1,1]
Pattern 4	[6,4,2,2,1]	[8,5,4,3,3,2,2,1,1,1]
Pattern 5	[7,3,2,2,1]	[9,6,4,3,2,2,1,1,1,1]

The query label is sampled from minority classes, defined as the bottom 3 classes for $K = 5$ and the bottom 7 classes for $K = 10$ in terms of label count. For example, under Pattern 3 with $K = 5$, the query label is selected from classes with 3, 2, or 1 labels. The plots below show accuracy versus stable rank for each pattern, obtained under different numbers of classes and transformer layers.

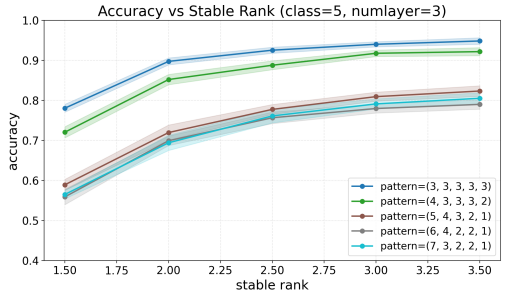
I.1.1. RESULTS AND DISCUSSIONS



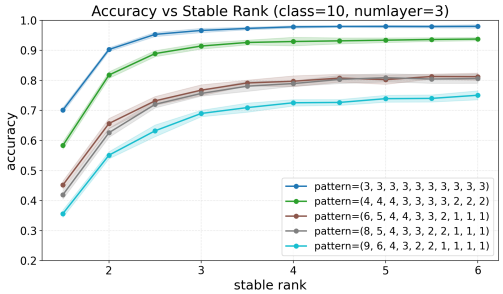
(a) $K = 5$, 1 layer



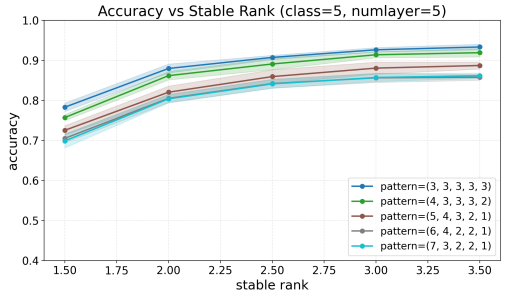
(b) $K = 10$, 1 layer



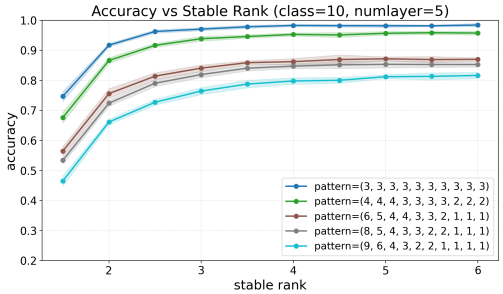
(c) $K = 5$, 3 layers



(d) $K = 10$, 3 layers



(e) $K = 5$, 5 layers



(f) $K = 10$, 5 layers

Figure 1: Performance across different numbers of classes and model depths. Each result is shown with a 1-sigma band.

We observe two main findings from the experiments. First, as the stable rank increases, the ICL accuracy improves. This observation is consistent with Theorems 5 and 6, which show that the error bound decreases as the stable rank increases. Second, as the degree of label imbalance from the label allocation pattern increases, the accuracy degrades, suggesting that class imbalance adversely affects in-context learning performance. Since the query labels are selected from the less frequent classes, $\psi_{k,k^\star}$ from Theorem 6 tends to become smaller as the degree of label imbalance increases. Therefore, this observation is consistent with our theoretical results.

I.1.2. ADDITIONAL DISCUSSIONS ON LINEAR ATTENTION VS (SOFTMAX) TRANSFORMER

We conduct the same experiment for $K = 10$ using a one-layer linear attention model in our parametrization setting. All other experimental settings (e.g., batch size, optimizer, and learning rate) are kept identical. To analyze the exact solution, we visualize the heatmap of $\mathbf{W}^{KQ}$.

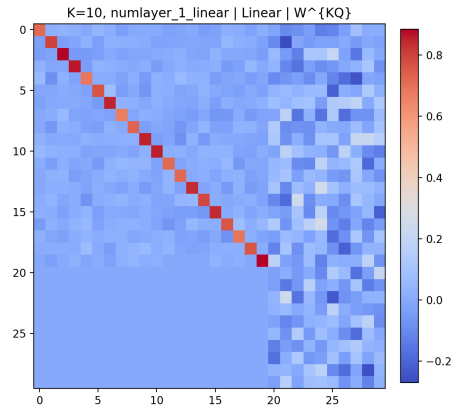


Figure 2: Heatmap of $\mathbf{W}^{KQ}$ for linear attention after training

In linear attention, we empirically observe that the learned solutions tend to concentrate along the diagonal with positive entries. Such solutions are characterized by a large trace-to-Frobenius ratio, satisfying $\mathrm{Tr}(\boldsymbol{W})/\|\boldsymbol{W}\|_F = \Omega(\sqrt{d}\xi^{-1})$, which in turn enables our theoretical analysis (Section F.2). We further observe that a similar pattern emerges in nonlinear Transformers, where in each layer, at least one attention head exhibits a diagonally concentrated structure with mainly positive entries. This suggests that a similar implicit bias characterized by $\mathrm{Tr}(\boldsymbol{W})/\|\boldsymbol{W}\|_F$ may persist in nonlinear settings. Under such solution, we can consider the output of softmax attention for class k as follows:

$$
\begin{aligned}
\hat{y} &= \left(\boldsymbol{E} + \frac{1}{M}\mathrm{Attn}_\theta(\boldsymbol{E})\right)_{(d+1):(d+K),(M+1)} \\
&= \frac{1}{M}\frac{\sum_{j\in\mathcal{C}_k}\exp\left(\boldsymbol{x}^{(j)\top}\boldsymbol{W}\boldsymbol{x}^{(M+1)}\right)}{\sum_{j\in[M]}\exp\left(\boldsymbol{x}^{(j)\top}\boldsymbol{W}\boldsymbol{x}^{(M+1)}\right)} \\
&\propto \sum_{j\in\mathcal{C}_k}\exp\left(\boldsymbol{x}^{(j)\top}\boldsymbol{W}\boldsymbol{x}^{(M+1)}\right) \text{ (Assume Gaussian noise is negligible)} \\
&\approx S_k\exp\left(\boldsymbol{\mu}_k^\top\boldsymbol{W}\boldsymbol{\mu}_{k^\star}\right) = \exp\left(\boldsymbol{\mu}_k^\top\boldsymbol{W}\boldsymbol{\mu}_{k^\star} + \log S_k\right).
\end{aligned}
$$

The quantity of interest, the classification error bound can be expressed as follows:

$$
\begin{aligned}
&\mathbb{P}\left((\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top\boldsymbol{W}\boldsymbol{\mu}_{k^\star} + \log\left(\frac{S_k}{S_{k^\star}}\right) \geq 0\right) \\
&= \mathbb{P}\left((\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top\boldsymbol{W}\boldsymbol{\mu}_{k^\star} \geq \frac{\mathrm{Tr}(\boldsymbol{W})}{dk}\|\boldsymbol{\Lambda}\|_F^2 + \mathbb{E}\left[(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top\boldsymbol{W}\boldsymbol{\mu}_{k^\star}\right] + \log\left(\frac{S_{k^\star}}{S_k}\right)\right).
\end{aligned}
$$

Let $\phi_{k,k^\star} := \log\left(\frac{S_{k^\star}}{S_k}\right)\frac{dk}{\mathrm{Tr}(\boldsymbol{W})\|\boldsymbol{\Lambda}\|_F^2}$ be an imbalance factor, which is similar to $\psi_{k,k^\star}$; that is, a lower $\phi_{k,k^\star}$ implies adversarial label allocation. (We assume $\phi_{k,k^\star} > -1$, avoiding degenerate label allocation regimes. For fixed $\boldsymbol{W}$ and $\boldsymbol{\Lambda}$, the value only depends on label allocation.)

$$\begin{aligned}
&\mathbb{P}\left((\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star} \geq \mathbb{E}\left[(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W}\boldsymbol{\mu}_{k^\star}\right] + (1+\phi_{k,k^\star})\frac{\mathrm{Tr}(\boldsymbol{W})}{dk}\|\boldsymbol{\Lambda}\|_F^2\right) \\
&\leq C\exp\left(-c\min\left(\frac{\mathrm{Tr}(\boldsymbol{W})^2(1+\phi_{k,k^\star})^2\,\mathrm{sr}(\boldsymbol{\Lambda})^2}{\|\boldsymbol{W}\|_F^2 K^2}, \frac{\mathrm{Tr}(\boldsymbol{W})(1+\phi_{k,k^\star})\,\mathrm{sr}(\boldsymbol{\Lambda})}{\|\boldsymbol{W}\|_F K}\right)\right) \\
&\quad + \exp\left(-\frac{c'(1+\phi_{k,k^\star})^2\,\mathrm{sr}(\boldsymbol{\Lambda})^2}{K}\right).
\end{aligned}$$

Through the property of implicit bias $\frac{\mathrm{Tr}(\boldsymbol{W})}{\|\boldsymbol{W}\|_F} = \Omega\left(\sqrt{d}\xi^{-1}\right)$, we can obtain upper bound as:

$$C\exp\left(-c\left(\frac{d(1+\phi_{k,k^\star})\,\mathrm{sr}(\boldsymbol{\Lambda})}{\xi^2 K} \wedge \frac{\sqrt{d}}{\xi} \wedge (1+\phi_{k,k^\star})\,\mathrm{sr}(\boldsymbol{\Lambda})\right)\cdot\frac{(1+\phi_{k,k^\star})\,\mathrm{sr}(\boldsymbol{\Lambda})}{K}\right).$$

Under solutions characterized by the lower bound of trace-to-Frobenius ratio, we establish that our theory extends consistently to softmax attention. In particular, (i) higher stable rank improves ICL performance, and (ii) robustness to majority-label bias increases with stable rank.

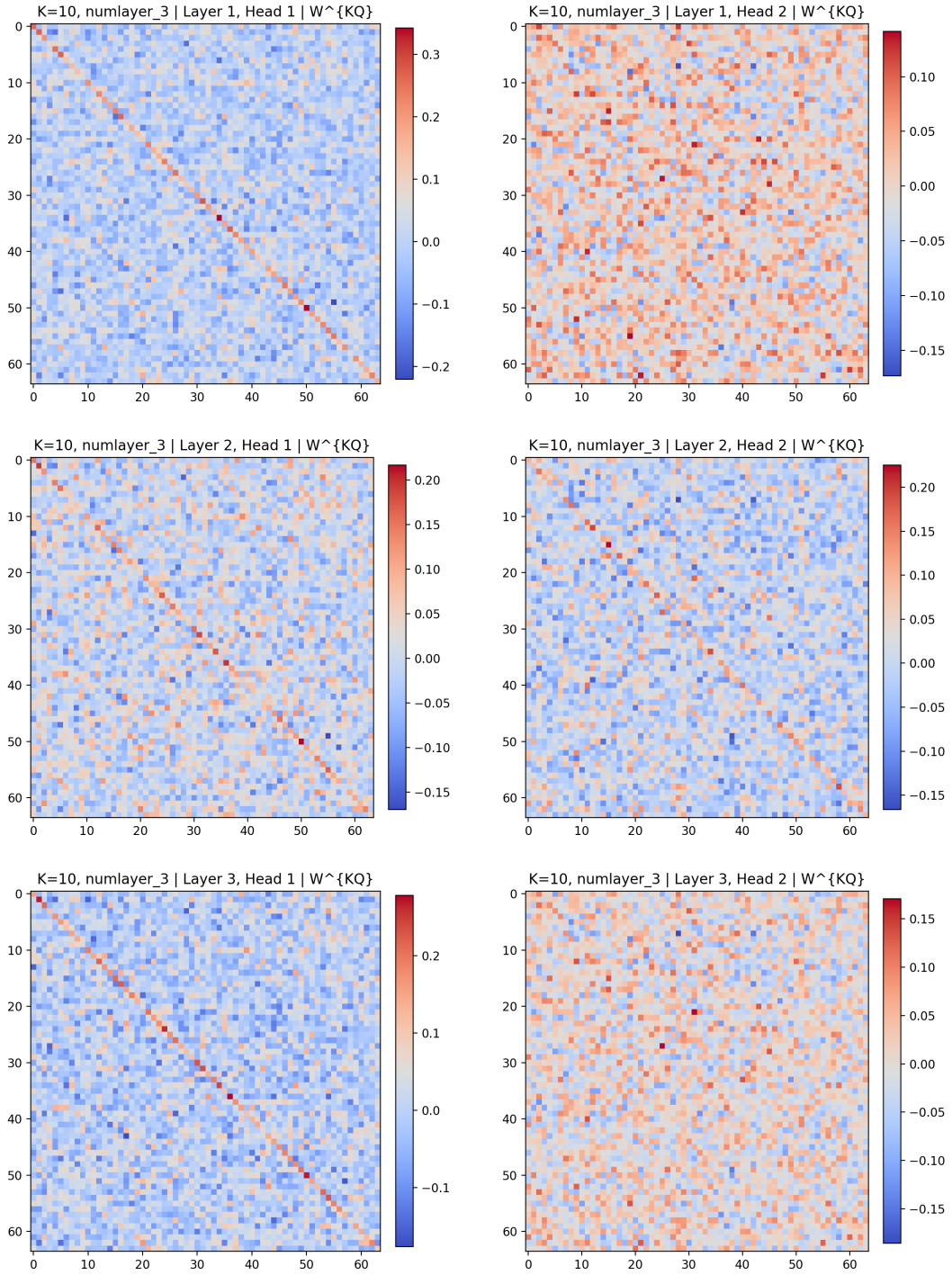


Figure 3: Heatmaps of $\mathbf{W}^{KQ}$ for each layer and attention head. Each row corresponds to a layer, and each column corresponds to a head.

I.2. Pre-trained LLMs for WordNet-based Word Classification

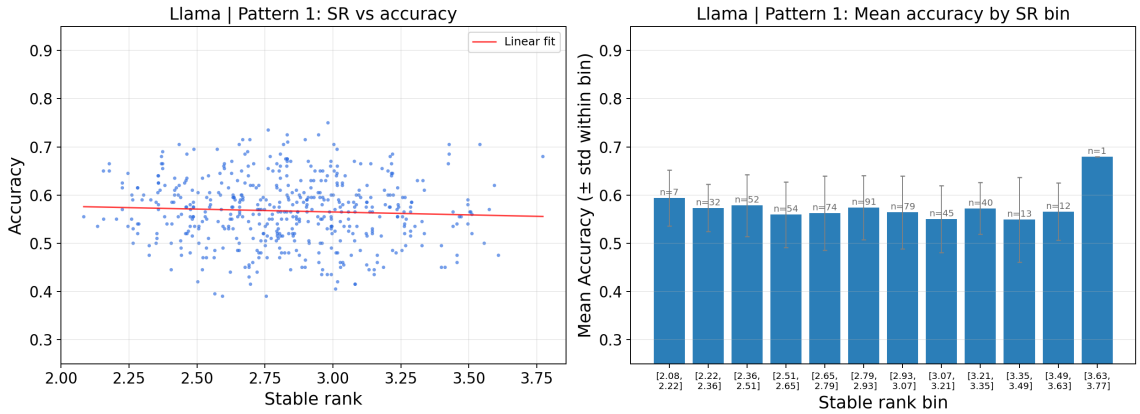
We further conduct experiments on real-world natural language datasets.We construct a multiclass classification task based on WordNet [18], defining a word-to-label mapping via hypernyms (e.g., lion → animal). Since WordNet is organized as a tree, we collect words by restricting the hyponym depth to at most 2 from each hypernym. We sample multiple 7-class subsets from a pool of 13 semantic classes (action, animal, food, group, location, object, person, plant, process, property, relation, state, substance), and evaluate performance across diverse class combinations. For each subset, we extract token embeddings from the model's input layer, group them by label, and average them to obtain class cluster center vectors. Stacking these vectors and centering the rows forms a matrix, whose geometry is quantified via the stable rank. We then examine how this geometric property relates to downstream ICL performance. To control for the correlation between the Frobenius norm and stable rank, we restrict sampled class subsets to those within one standard deviation of the empirical mean, thereby isolating the effect of stable rank from magnitude variations.

We conduct experiments using the following three pre-trained LLMs: `Llama-3.2-1B-Instruct`, `Qwen2.5-3B-Instruct`, and `Phi-3-mini-4k-instruct`. Note that, since embeddings differ across models, the class-center matrix may vary even for the same combination. Prompts consist of (word → label) demonstrations followed by a query (word →), with disjoint query sets and shuffled ordering in order to mitigate positional biases such as recency effects.

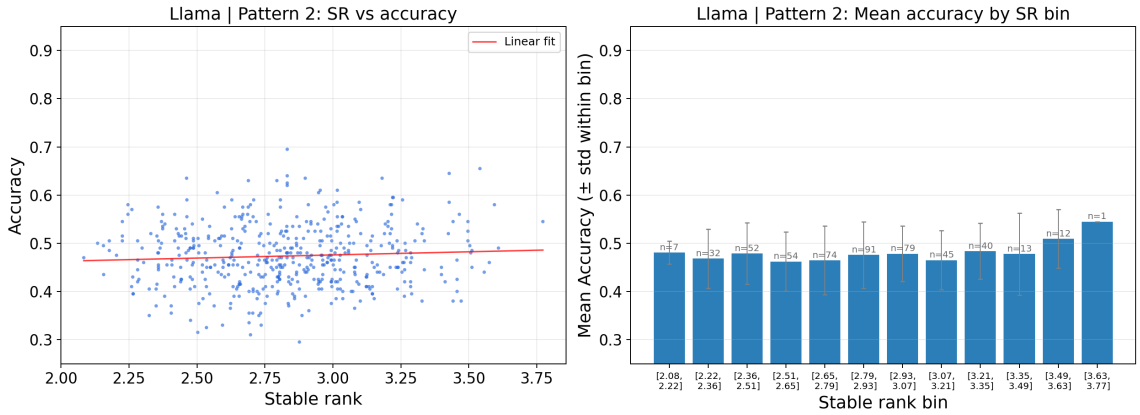
For each prompt, we sample a fixed number of labels per class according to predefined label allocation patterns: Pattern 1 (balanced) [5, 5, 5, 5, 5, 5, 5], Pattern 2 (imbalanced) [9, 8, 6, 5, 4, 2, 1], and Pattern 3 (imbalanced) [12, 9, 6, 4, 2, 1, 1]. In addition, even within the same combination of subclasses, we reshuffle the label-to-class assignments for each query, resulting in diverse prompt configurations. To evaluate imbalanced label scenario, the query label is preferentially selected from the bottom-5 classes (i.e., those with the fewest labels) in the current prompt. For example, under Pattern 3, we allocate labels as follows: food (12), location (9), object (6), plant (4), property (2), relation (1), and state (1). The query word is then sampled from the bottom-5 classes: object, plant, property, relation, and state. We perform classification using constrained decoding, restricting the output space to valid label token sequences to prevent predictions outside the label set. We sample 500 class combinations, and for each combination and label-allocation pattern, we measure accuracy over 200 prompts, analyzing the relationship between stable rank and classification accuracy.

The plots below show scatter plots of stable rank versus accuracy over 500 combinations, together with linear fits and binned averages. We also present analogous plots for each pattern, illustrating the differences across patterns using scatter plots with linear fits and binning. The bin plot below shows the mean and the 1-sigma interval.

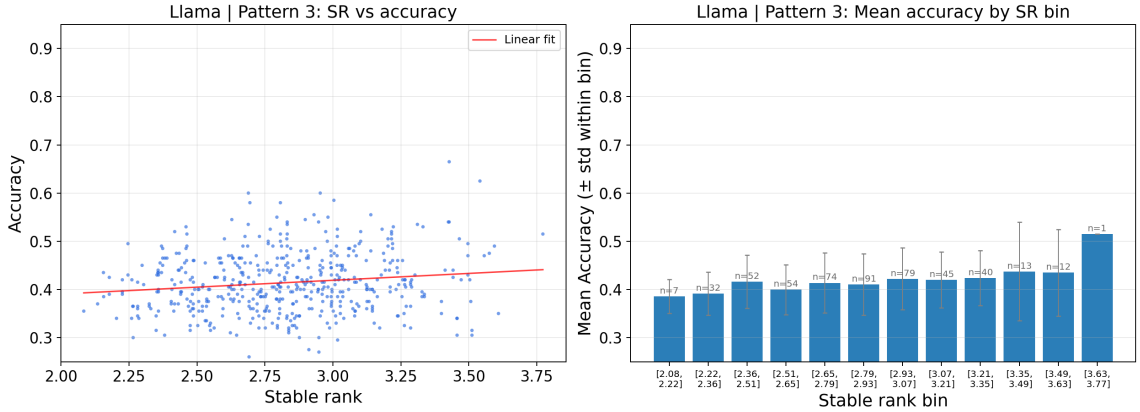
RESULTS ON LLAMA



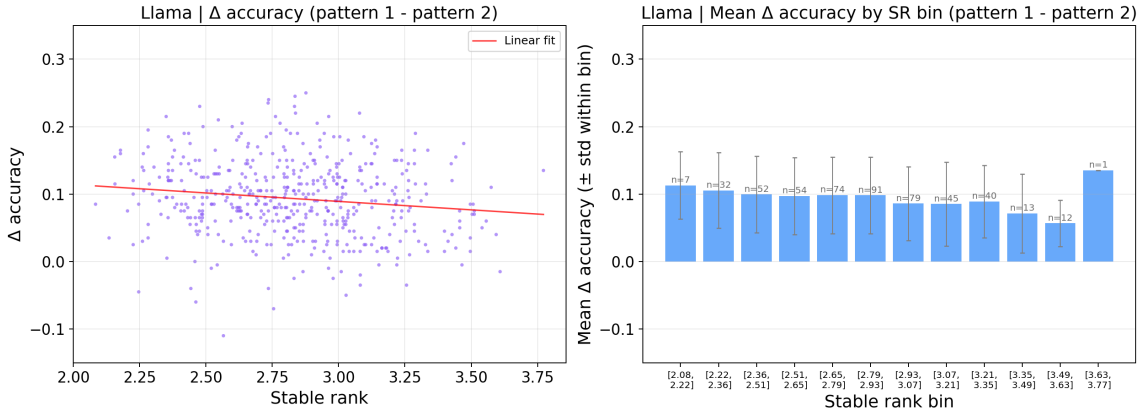
(a) Llama Pattern 1



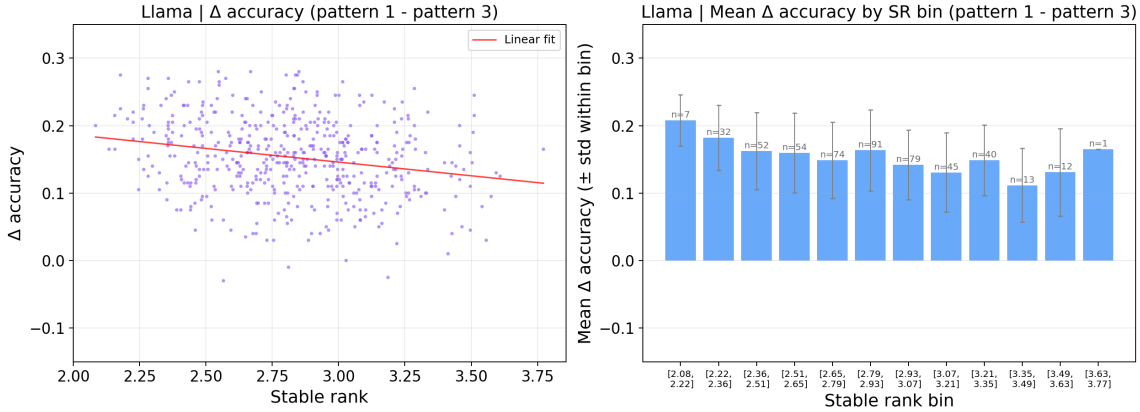
(b) Llama Pattern 2



(c) Llama Pattern 3

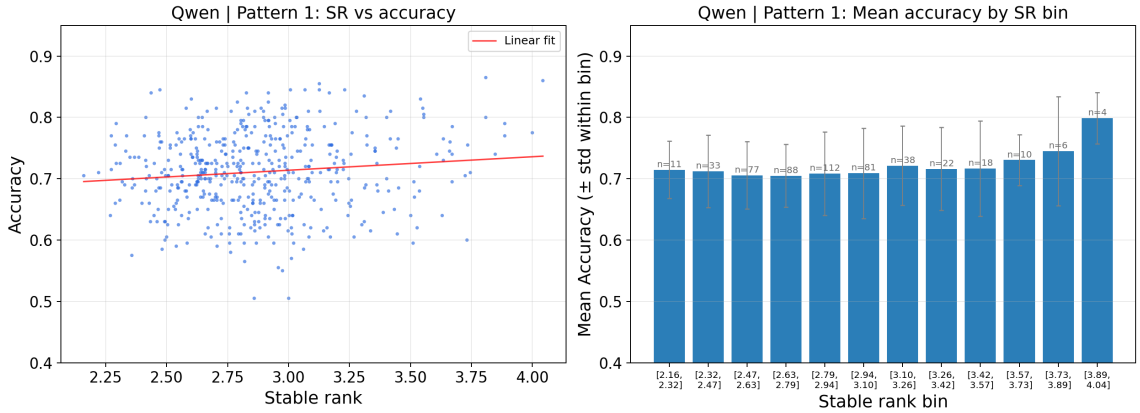


(a) Llama Difference between pattern 1 and 2

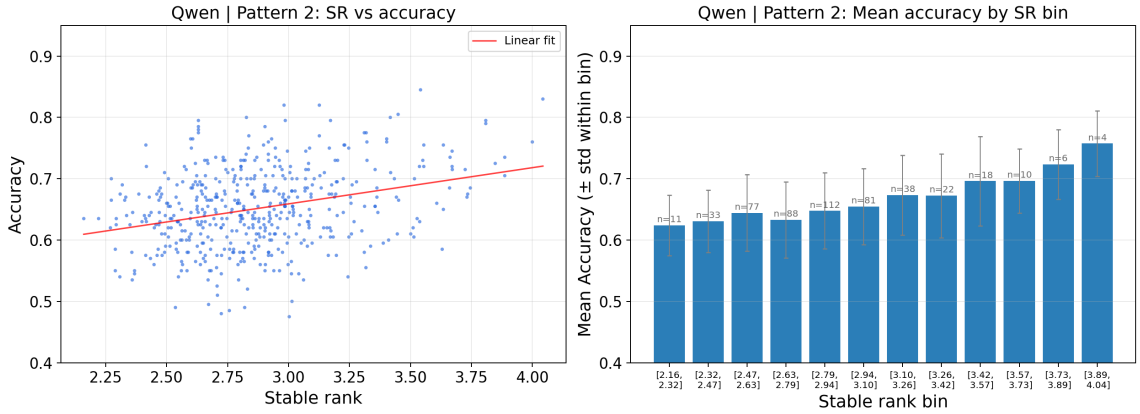


(b) Llama Difference between pattern 1 and 3

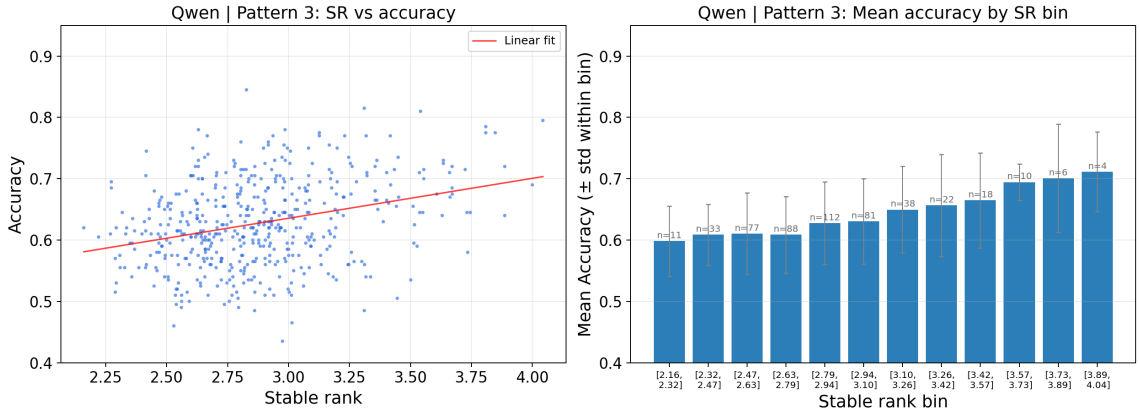
RESULTS ON QWEN



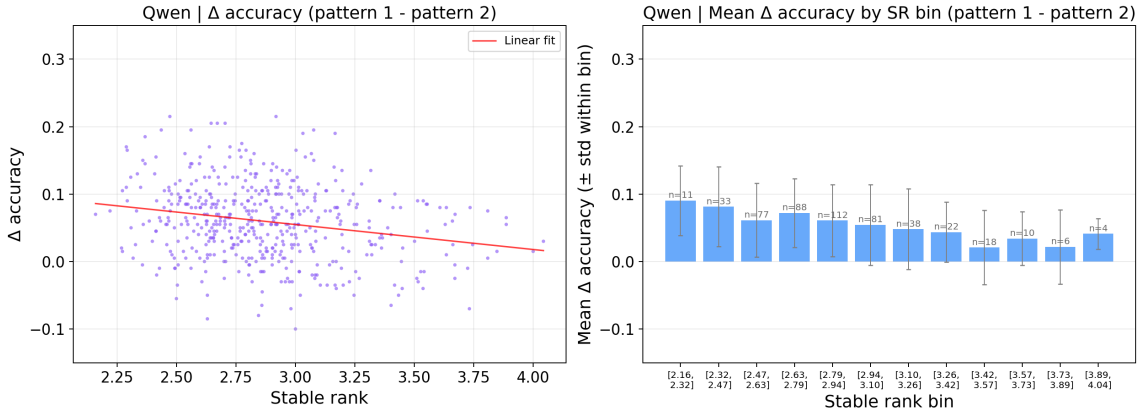
(a) Qwen Pattern 1



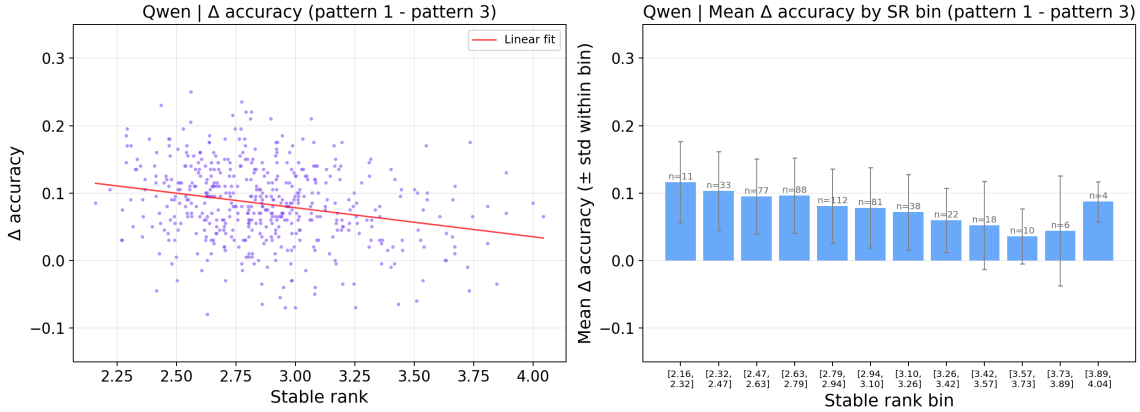
(b) Qwen Pattern 2



(c) Qwen Pattern 3

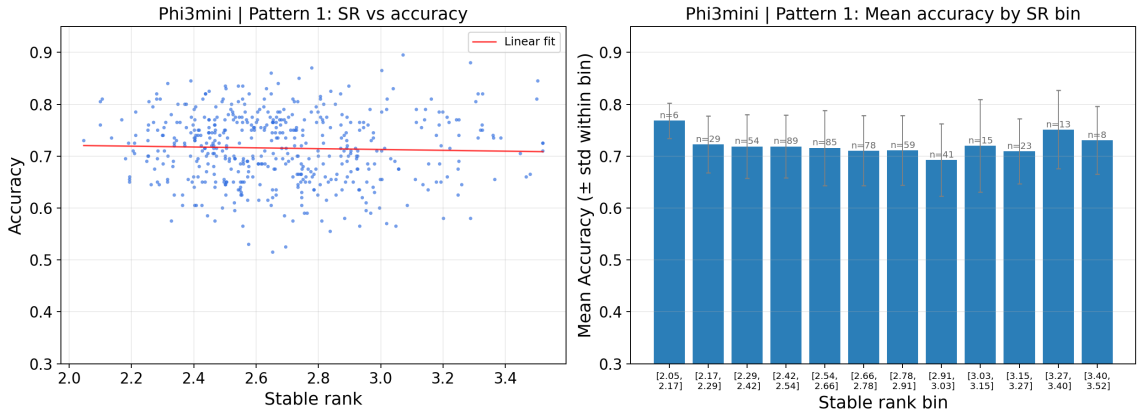


(a) Qwen Difference between pattern 1 and 2

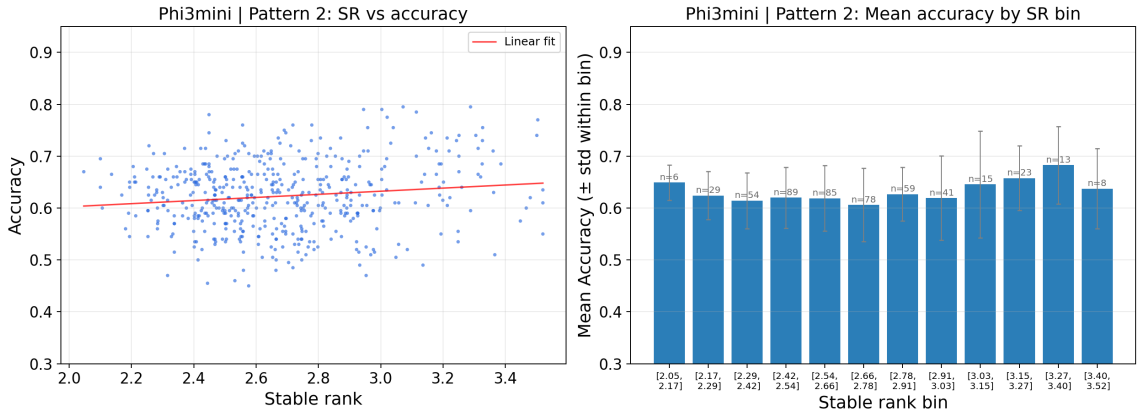


(b) Qwen Difference between pattern 1 and 3

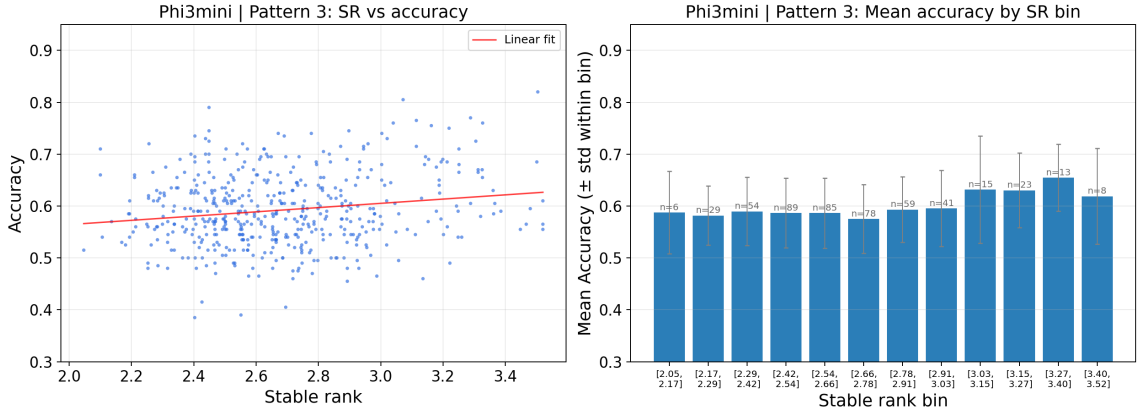
RESULTS ON PHI-3



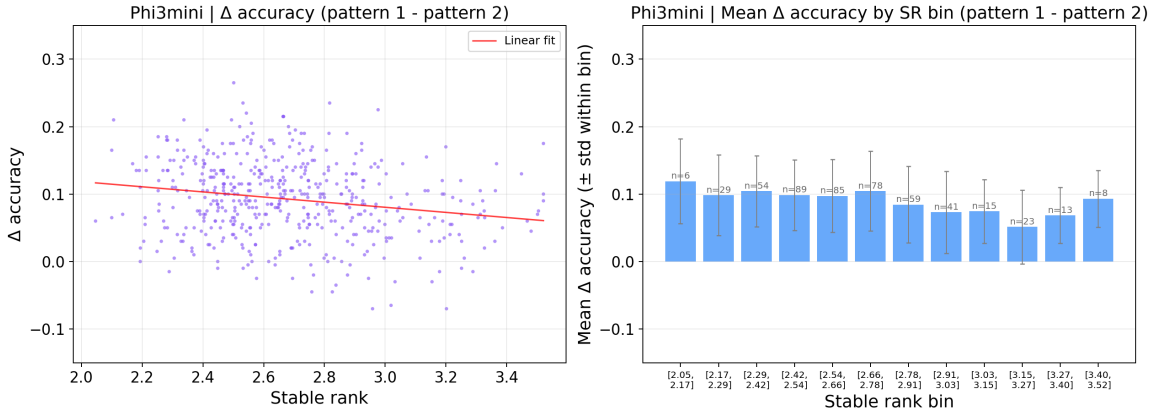
(a) Phi-3 Pattern 1



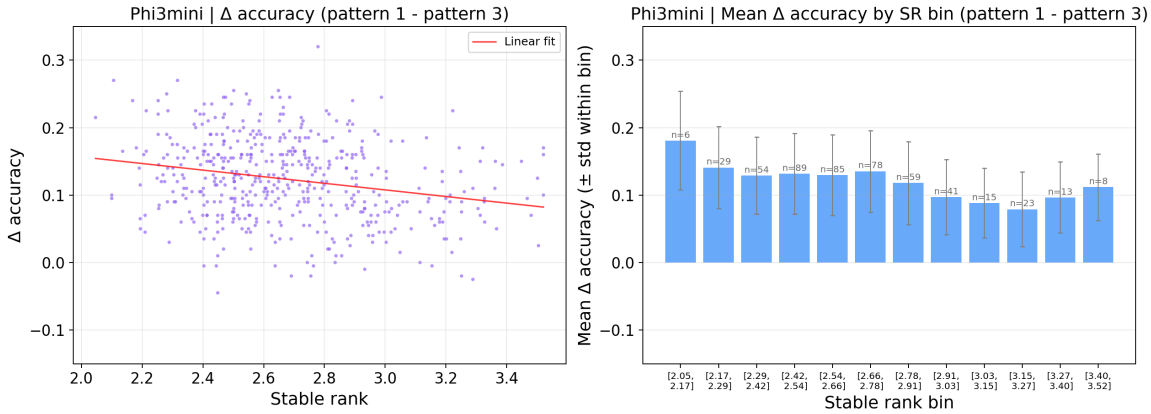
(b) Phi-3 Pattern 2



(c) Phi-3 Pattern 3



(a) Phi-3 Difference between pattern 1 and 2



(b) Phi-3 Difference between pattern 1 and 3

I.2.1. STATISTICS AND DISCUSSIONS

Table 2: Summary statistics characterizing the relationship between stable rank (independent variable) and accuracy (dependent variable) across models and prompting patterns. For each setting, we report the linear regression slope, Pearson correlation coefficient (r_{Pearson}) with its p-value for testing zero slope, and Spearman rank correlation coefficient (ρ_{Spearman}) with its corresponding p-value. Statistically significant results are highlighted in bold.

Model	Pattern	Slope	r_{Pearson}	p_{reg} (slope $= 0$)	ρ_{Spearman}	p_{Spearman} ($\rho = 0$)
Llama	1	-0.0120	-0.0556	2.14×10^{-1}	-0.0628	1.61×10^{-1}
Llama	2	0.0130	0.0644	1.51×10^{-1}	0.0351	4.33×10^{-1}
Llama	3	0.0285	0.1468	$\mathbf{9.98 \times 10^{-4}}$	0.1213	$\mathbf{6.61 \times 10^{-3}}$
Qwen	1	0.0220	0.1150	$\mathbf{1.01 \times 10^{-2}}$	0.0898	$\mathbf{4.48 \times 10^{-2}}$
Qwen	2	0.0590	0.3033	$\mathbf{4.22 \times 10^{-12}}$	0.2566	$\mathbf{5.84 \times 10^{-9}}$
Qwen	3	0.0651	0.3114	$\mathbf{1.06 \times 10^{-12}}$	0.2790	$\mathbf{2.15 \times 10^{-10}}$
Phi-3	1	-0.0080	-0.0350	4.34×10^{-1}	-0.0624	1.64×10^{-1}
Phi-3	2	0.0300	0.1340	$\mathbf{2.68 \times 10^{-3}}$	0.0800	7.39×10^{-2}
Phi-3	3	0.0409	0.1708	$\mathbf{1.24 \times 10^{-4}}$	0.1123	$\mathbf{1.19 \times 10^{-2}}$

We observe a weak but statistically significant positive correlation between the stable rank of the class center matrix and ICL performance, particularly under imbalanced regimes. The positive slope and Spearman correlation indicate that as the stable rank of the class center matrix increases, the ICL accuracy of the model increases, which is consistent with our theoretical results. We further analyze how the variation in accuracy due to label imbalance depends on the stable rank under fixed class combinations by taking the difference in accuracy between two patterns.

Table 3: Summary statistics characterizing the relationship between stable rank and accuracy differences across pairs of prompting patterns. The reported columns are defined as in Table 2. Statistically significant results are highlighted in bold.

Model	Pattern Pair	Slope	r_{Pearson}	p_{reg} (slope $= 0$)	ρ_{Spearman}	p_{Spearman} ($\rho = 0$)
Llama	1-2	-0.0250	-0.1409	$\mathbf{1.59 \times 10^{-3}}$	-0.1515	$\mathbf{6.77 \times 10^{-4}}$
Llama	1-3	-0.0405	-0.2224	$\mathbf{5.04 \times 10^{-7}}$	-0.2143	$\mathbf{1.32 \times 10^{-6}}$
Qwen	1-2	-0.0370	-0.2205	$\mathbf{6.34 \times 10^{-7}}$	-0.1967	$\mathbf{9.43 \times 10^{-6}}$
Qwen	1-3	-0.0431	-0.2440	$\mathbf{1.08 \times 10^{-8}}$	-0.2331	$\mathbf{2.36 \times 10^{-7}}$
Phi-3	1-2	-0.0380	-0.1965	$\mathbf{9.63 \times 10^{-6}}$	-0.1848	$\mathbf{3.21 \times 10^{-5}}$
Phi-3	1-3	-0.0489	-0.2339	$\mathbf{1.22 \times 10^{-7}}$	-0.2188	$\mathbf{7.79 \times 10^{-7}}$

We also observe a weak but statistically significant negative correlation between the stable rank of the class center matrix and accuracy difference between balanced and imbalanced patterns. The negative slope and Spearman correlation indicate that, as the stable rank of the class center matrix increases, the performance gap between balanced and imbalanced label allocation decreases, leading to improved robustness under majority label bias. Such an observation is consistent with our theoretical results.