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

Daehan Yoon
KAIST AI
axiom0102@kaist.ac.kr

Chulhee Yun
KAIST AI
chulheeyun@kaist.ac.kr

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 [Vaswani et al., 2017] have introduced a new paradigm across a broad range of machine learning domains, including natural language processing and computer vision. In particular, the in-context learning (**ICL**) phenomenon [Brown et al., 2020, Garg et al., 2022], 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 in developing a theoretical understanding of ICL in transformer models, primarily through an optimization-based perspective. Several studies focus on linear regression settings to analyze linear attention and ICL [von Oswald et al., 2023, Ahn et al., 2023, Zhang et al., 2023, Mahankali et al., 2024]. However, such regression settings are somewhat detached from real-world applications of LLMs, since ICL can be viewed as a special case of next-token prediction, which is more closely aligned with classification tasks. Several recent works study in-context classification, mostly focusing on the *binary* case [Frei and Vardi, 2025, Magen and Vardi, 2025] with two class-center vectors pointing in 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* [Gupta et al., 2023], 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. [2025] 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-

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

dimensional subspaces [Mu and Viswanath, 2018, Pope et al., 2021]. Moreover, the paper focuses on balanced test labels and thus fails to capture the majority label bias.

Our Contributions. In this paper, we study generalization in multiclass in-context classification. We introduce a spectral data-generating procedure that captures both anisotropic class-center geometry and label imbalance. We adopt a linear self-attention model widely used in prior theoretical work on ICL. Whereas classical analyses of multiclass GMMs typically focus on classification under a fixed mixture, we study how the model combines a matrix $\boldsymbol{W}$ learned across training tasks with class centers estimated from each new prompt to classify the query. We analyze the training objective, characterize the max-margin direction selected by gradient descent for $\boldsymbol{W}$, and use this characterization to derive generalization bounds that explicitly account for both test-time 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 3.1).
- In the test label imbalance regime, we show that queries that belong to the majority class are indeed easy to classify correctly, while queries from the minority class are prone to errors. 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 3.3).
- We empirically observe that the implications from our theory align with softmax transformers and pretrained LLMs (Section 5).

1.1 Related Work

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, von Oswald et al. [2023] and Ahn et al. [2023] 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 [Tripuraneni et al., 2021]. Akyürek et al. [2023] also show that transformers can implement standard learning algorithms for linear models, including gradient descent and closed-form ridge regression. Kim et al. [2024] and Magen and Vardi [2025] mainly focus on the optimality of ICL in terms of meta-learning. Zhang et al. [2023, 2025] 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. [2024] 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 [Bai et al., 2023, Giannou et al., 2024, Li et al., 2024a,b, Magen and Vardi, 2025, Frei and Vardi, 2025, Shen et al., 2025]. Bai et al. [2023] 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. [2024] further show that transformers can approximate higher-order optimization methods for logistic regression. Li et al. [2024a] analyze in-context learning capabilities in binary classification in nonlinear transformers. Similarly, Li et al. [2024b] 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. [2025] study gradient flow in in-context classification in a 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 [Soudry et al., 2018, Ravi et al., 2024], with extensions to homogeneous neural networks [Lyu and Li, 2020].

The works most closely related to ours are Frei and Vardi [2025] and Magen and Vardi [2025], which characterize the implicit bias of the model and leverage it to derive generalization bounds. Frei and Vardi [2025] 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 [2025] 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.

2 Preliminaries

In this section, we formalize the model, data distribution, and the assumptions.

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 $i \le m \wedge n$ 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.

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 [Vaswani et al., 2017] 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 [von Oswald et al., 2023, Ahn et al., 2023] of in-context learning—particularly in ICL regression and classification models—to explicitly enforce the causal structure of the prompt. We consider a setting in which the merged matrix $\boldsymbol{W}^{KQ}$ is directly trained, 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

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

We adopt a parametrization similar to prior works [Ahn et al., 2023, Shen et al., 2025]. Specifically,

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

[1] As presented in Section 5, 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.

Note that the blocks marked with $*$ do not influence the model's predictions. We fix $\boldsymbol{W}_\star^V = [0_{K\times d} \quad \boldsymbol{I}_K]$, and only optimize over $\boldsymbol{W}$. Under this choice, 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 vector for class k.

2.2 Data Generation and Training Process

We define the following distribution for generating multiclass ICL tasks.

Definition 2.1. *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 we sample the left and right singular vector matrices $\boldsymbol{U} \in \mathbb{R}^{K\times K}$, $\boldsymbol{V} \in \mathbb{R}^{d\times d}$ independently and 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.

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 number of context examples in each training task, 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, and $\hat{\boldsymbol{y}}_\tau$ 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:

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

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 $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1,\dots,\lambda_K) \in \mathbb{R}^{K\times d}$ with $\lambda_1 \ge \cdots \ge \lambda_K \ge 0$, 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.

Motivation of Our Data Generation Process. In prior works [Frei and Vardi, 2025, Magen and Vardi, 2025], 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 2.1 provides a natural generalization of the binary, direction-based sampling scheme to the multiclass setting. Furthermore, Definition 2.1 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.

2.3 Assumptions

We impose the following assumptions:

Assumption 2.2. *There exists* $0 < c_s \leq 1$ *such that* $\mathrm{sr}(\mathbf{\Sigma}) \geq c_s K$ *(i.e.* $\mathrm{sr}(\mathbf{\Sigma}) = \Theta(K)$*).*

Assumption 2.3. *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 2.4. *The ambient dimension* d *and the number of classes* K *should satisfy* $d \geq K \geq 3 \vee C_K \log^2\left(B^2K^2/\delta\right)$ *for some absolute constant* $C_K > 0$*.*

The first two assumptions ensure that the training set is non-degenerate, and are not required for the test task spectrum $\mathbf{\Lambda}$. Assumptions 2.2, 2.3, and 2.4 guarantee that the training tasks are realizable (Assumption B.1) with high probability, which is essential for inducing an implicit bias through training. Since the stable rank is always upper bounded by K and is maximized when the row vectors are mutually orthogonal and have identical norms, it suffices to impose only a lower bound on $\mathrm{sr}(\mathbf{\Sigma})$. We also remark that the data generation assumptions of Shen et al. [2025] satisfy Assumptions 2.2 and 2.3 which are related to the data geometry.

The largest singular value associated with the training tasks σ_1 plays a role analogous to the radius of the class centers in binary classification. Magen and Vardi [2025] relax this radius assumption and require a larger number of examples per training task N. Instead, we follow Frei and Vardi [2025], which adopts a similar assumption on the singular value norm while allowing as few as one example per class in each training task ($S_\star = 1$).

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 induced by the limiting dynamics of Linear Self-Attention trained via gradient descent with a fixed step size η. To facilitate the analysis, we adopt Assumptions B.1 and B.2, which are standard in the implicit bias literature [Soudry et al., 2018, Ravi et al., 2024] and are commonly imposed to make the analysis tractable. For a given task τ, recall the signal vector for class k, $\hat{\boldsymbol{\mu}}_{\tau,k} = \frac{1}{N}\sum_{j\in\mathcal{C}_{\tau,k}} \boldsymbol{x}_\tau^{(j)}$. Under a small step size η, the training dynamics induce an implicit bias on the attention parameters, and the solution converges in direction to the max-margin solution:

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

More details can be found in Appendix B, 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. At test time, we adopt a hard-max output rule and denote the resulting max-margin classifier $\boldsymbol{W}_{MM}$ by $\boldsymbol{W}$ for simplicity. We analyze the generalization performance by bounding the classification error. We define the error event as $\hat{k} \neq k^\star$, where $\hat{k} := \arg\max_{k\in[K]} [\hat{\boldsymbol{y}}(\boldsymbol{E}; \boldsymbol{W})]_k$. Then,

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

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

Theorem 3.1. *Fix $k \neq k^\star$, and let $\delta \in (0, \frac{1}{6})$ be arbitrary. Suppose Assumptions 2.2, 2.3, 2.4, and B.2 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 a 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 $\xi = O(\frac{d}{B} \vee \log^2\left(B^2K^2/\delta\right))$.

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.

The two terms in the bound capture different aspects of the test task. The first reflects the spectral geometry of the test class centers, while the second depends on the signal strength and the number of in-context examples per class. Controlling the overall misclassification probability therefore requires suitable conditions on both geometry and signal strength, together with sufficiently many in-context examples. The following corollary makes these sufficient conditions explicit for a prescribed target error ϵ.

Corollary 3.2 (In-context sample complexity)**.** *Under the assumptions of Theorem 3.1, fix $\epsilon \in (0, 1/2)$. There exists an absolute constant $C > 0$ such that, if $B \geq Cd$ and*

$$\mathrm{sr}(\boldsymbol{\Lambda}) \geq C \max\left\{\sqrt{K \log\frac{K}{\epsilon}}, \frac{\xi K}{\sqrt{d}} \log\frac{K}{\epsilon}\right\}, \quad \frac{\lambda_1\,\mathrm{sr}(\boldsymbol{\Lambda})}{\xi K} \geq C \log\frac{K}{\epsilon},$$

then

$$S \geq C \frac{d\xi^2 K^2}{\|\boldsymbol{\Lambda}\|_F^4} \log^2\frac{K}{\epsilon}$$

ensures that, with probability at least $1 - 6\delta$ over the training data, the learned max-margin classifier has test misclassification probability at most ϵ. Here, S denotes the number of examples per class in a balanced test prompt.

With sufficiently many training tasks, the corollary allows the test class centers to exhibit more anisotropic spectral geometry than the training assumptions permit. Specifically, the sufficient stable-rank threshold is determined by both $\sqrt{K\log(K/\epsilon)}$ and $(\xi K/\sqrt{d})\log(K/\epsilon)$. When $B = \Omega(d)$, the quantity ξ grows at most polylogarithmically; together with $d \geq K$, this yields the sufficient scaling $\mathrm{sr}(\boldsymbol{\Lambda}) = \tilde{\Omega}(\sqrt{K})$ for a fixed target error. Up to logarithmic factors, this threshold is substantially smaller than the training requirement $\mathrm{sr}(\boldsymbol{\Sigma}) = \Theta(K)$ in Assumption 2.2. Thus, provided that the signal-strength and in-context sample-size conditions also hold, the guarantee accommodates a shift toward more anisotropic class-center geometry at test time.

Once the geometric and signal-strength conditions hold, the corollary quantifies the number of in-context examples needed to achieve the target error. In particular, when $B = \Omega(d)$ and ξ is at most polylogarithmic, it suffices to take $S = \tilde{\Omega}(dK^2/\|\boldsymbol{\Lambda}\|_F^4)$ examples per class, corresponding to a total test prompt length of $M = KS$. In the regime covered by the SNR-based lower bound in Appendix A, this sufficient scaling matches the necessary number of examples per class up to logarithmic factors. The resulting guarantee therefore establishes nearly optimal in-context sample complexity in this regime, extending the binary classification result of Magen and Vardi [2025] to the multiclass setting.

3.3 Multiclass In-Context Classification under Test Label Imbalance

We now consider the test-time label imbalance setting, in which different numbers of examples are sampled for each class, while using balanced prompts during training. Our bound explicitly depends

on the class-wise label counts in the test task, unlike in prior work [Frei and Vardi, 2025, Magen and Vardi, 2025, Shen et al., 2025], thereby capturing empirical phenomena such as majority-label bias.

Theorem 3.3. *Fix $k \neq k^\star$, and let $\delta \in (0, \frac{1}{6})$ be arbitrary. Suppose Assumptions 2.2, 2.3, 2.4, and B.2 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 an **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*

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

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 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. [2023] 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.

4 Proof Sketch

In this section, we provide a sketch of the analysis underlying the classification error bound. Our goal is to establish an upper bound on the classification error under the max-margin solution $\boldsymbol{W}$:

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

We focus on the balanced label setting; the imbalanced case can be handled similarly. In this setting, $\hat{\boldsymbol{\mu}}_k$ admits the decomposition $\hat{\boldsymbol{\mu}}_k = \frac{S}{M}\boldsymbol{\mu}_k + \frac{\sqrt{S}}{M}\boldsymbol{z}_k$, where $\boldsymbol{z}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ are independent. Then,

$$\begin{aligned}
&\hat{\boldsymbol{\mu}}_k^\top \boldsymbol{W} \boldsymbol{x}^{(M+1)} > \hat{\boldsymbol{\mu}}_{k^\star}^\top \boldsymbol{W} \boldsymbol{x}^{(M+1)} \\
\Leftrightarrow\ & -\underbrace{(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W} \boldsymbol{\mu}_{k^\star}}_{\boldsymbol{X}_1} < \underbrace{(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W} \boldsymbol{z}'}_{\boldsymbol{X}_2} + \underbrace{\sqrt{2/S}\boldsymbol{z}^\top \boldsymbol{W} \boldsymbol{\mu}_{k^\star}}_{\boldsymbol{X}_3} + \underbrace{\sqrt{2/S}\boldsymbol{z}^\top \boldsymbol{W} \boldsymbol{z}'}_{\boldsymbol{X}_4},
\end{aligned}$$

for independent $\boldsymbol{z},\ \boldsymbol{z}' \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. Therefore, for any $t > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\hat{\boldsymbol{\mu}}_k^\top \boldsymbol{W} \boldsymbol{x}^{(M+1)} > \hat{\boldsymbol{\mu}}_{k^\star}^\top \boldsymbol{W} \boldsymbol{x}^{(M+1)}\right) \\
&\leq \mathbb{P}\left(\boldsymbol{X}_1 \geq -t\right) + \mathbb{P}\left(|\boldsymbol{X}_2| \geq t/4\right) + \mathbb{P}\left(|\boldsymbol{X}_3| \geq t/4\right) + \mathbb{P}\left(|\boldsymbol{X}_4| \geq t/4\right).
\end{aligned}$$

From Definition 2.1, the randomness of $\boldsymbol{X}_1$ 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 4.1. *Let $\mathbb{O}(d)$ denote the group of $d \times d$ orthogonal matrices,*

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

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 $\boldsymbol{Q} \in \mathbb{O}(d)$.

To analyze $\boldsymbol{X}_1$, we express it explicitly in terms of $\boldsymbol{U}$ and $\boldsymbol{V}$: $\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) = (\boldsymbol{e}_k - \boldsymbol{e}_{k^\star})^\top \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top \boldsymbol{W}\boldsymbol{V}\boldsymbol{\Lambda}^\top \boldsymbol{U}^\top \boldsymbol{e}_{k^\star}$. Here, $\boldsymbol{U},\boldsymbol{V}$ are independently sampled from the Haar measure on the orthogonal groups $\mathbb{O}(K)$ and $\mathbb{O}(d)$, respectively. To set the centering point for the concentration analysis, we first compute the expectation of $\boldsymbol{X}_1$. By Lemma C.2, we can derive $\mathbb{E}\boldsymbol{X}_1(\boldsymbol{U},\boldsymbol{V}) = -\frac{\mathrm{Tr}(\boldsymbol{W})}{dK}\|\boldsymbol{\Lambda}\|_F^2$. Moreover, Appendix D provides a lower bound on $\mathrm{Tr}(\boldsymbol{W})$, which ensures that this expectation is negative. For $0 < t < \frac{\mathrm{Tr}(\boldsymbol{W})}{dK}\|\boldsymbol{\Lambda}\|_F^2$ and $s = \frac{\mathrm{Tr}(\boldsymbol{W})}{dK}\|\boldsymbol{\Lambda}\|_F^2 - t > 0$, we decompose the deviation as follows:

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

The first term can be controlled using a generalized Hanson-Wright inequality, stated below.

Lemma 4.2 (Bilinear Hanson-Wright inequality for Haar-random matrix)**.** *Fix $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^d$, $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ and let $\boldsymbol{V}$ be drawn from the Haar measure on $\mathbb{O}(d)$. There exist absolute constants $C, c > 0$ such that, for every $t \geq 0$,*

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

This inequality extends classical Hanson–Wright-type results from the uniform sphere to the Haar measure on the orthogonal group; we believe this concentration result could be of independent interest. Additional details are provided in Appendix C.

The second term can be bounded using Theorem 5.17 in Meckes [2019], where the randomness arises from the orthogonal group $\mathbb{O}(K)$. Since the result applies to connected manifolds and $\mathbb{O}(K)$ is not connected, the result does not directly apply. In particular, it 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 $\mathbb{O}(K)$.

This decomposition reveals a key structural feature: the worst-case deviation of $\boldsymbol{X}_1$ is governed by the spectral norm of $\boldsymbol{\Lambda}$, while its typical behavior is controlled by the Frobenius norm. As a result, the stable rank $\mathrm{sr}(\boldsymbol{\Lambda}) = \|\boldsymbol{\Lambda}\|_F^2/\|\boldsymbol{\Lambda}\|^2$ emerges naturally as the parameter governing the concentration.

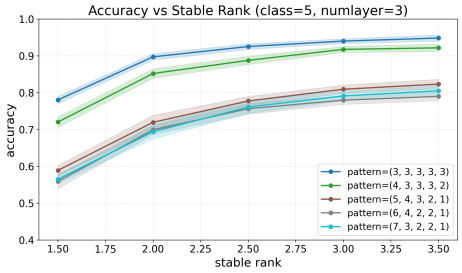
The concentration of $\boldsymbol{X}_2, \boldsymbol{X}_3$, and $\boldsymbol{X}_4$ can be established using arguments similar to those in Magen and Vardi [2025]. In particular, each of $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ is bounded with high probability by $\tilde{O}(\frac{\lambda_1\|\boldsymbol{W}\|_F}{\sqrt{d}})$, while $\boldsymbol{X}_4$ can be bounded by $\tilde{O}(\|\boldsymbol{W}\|_F)$. Since all three terms scale with $\|\boldsymbol{W}\|_F$, controlling their magnitudes requires an upper bound on $\|\boldsymbol{W}\|_F$. This motivates a careful analysis of the implicit bias of the max-margin solution, detailed in Appendix D.

5 Experiments and Discussion

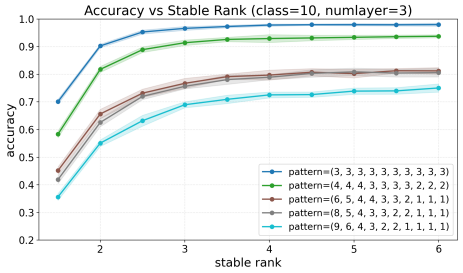
5.1 Linear and Nonlinear Transformers under Our Data Distribution

We conduct experiments on in-context multiclass classification using both the linear self-attention model studied in our theory and nonlinear Transformers with softmax self-attention layers of varying depths. The data are generated according to Definition 2.1, which allows precise control over the geometry of the class-center matrix. We evaluate classification accuracy as a function of stable rank under different label allocation patterns while fixing the Frobenius norm of the class-center matrix. For imbalanced prompts, the query label is selected from the classes with the fewest demonstrations (bottom-3 for $K = 5$ and bottom-7 for $K = 10$). The figure below presents representative results for 3-layer softmax Transformers; results for linear self-attention and other model depths, together with experimental details, are provided in Appendix G.1.

We observe two main trends in both linear and nonlinear models. First, classification accuracy generally improves as stable rank increases, consistent with the stronger generalization guarantees in Theorems 3.1 and 3.3. Second, more severe label imbalance generally reduces accuracy on minority-class queries. For such queries, a larger count advantage of a competing class reduces $\psi_{k,k^\star}$ in Theorem 3.3, weakening the corresponding error bound. The linear-model results support the qualitative predictions for the predictor analyzed in our theory, while the nonlinear results suggest that similar trends persist beyond linear self-attention.



(a) $K = 5$, 3 layers



(b) $K = 10$, 3 layers

Figure 1: In-context classification accuracy of 3-layer softmax Transformers versus stable rank under different label allocation patterns, for $K \in \{5, 10\}$.

5.2 Pre-trained LLMs for WordNet-based Word Classification

We further conduct experiments on a real-world multiclass in-context classification task constructed from WordNet [Miller, 1995], where words are mapped to semantic labels via hypernyms to study how class-center geometry influences ICL performance. We form multiclass classification problems by sampling 500 class combinations from a predefined pool of semantic categories. For each class combination, we extract input-layer token embeddings from the model, compute class center vectors by averaging them, and construct a centered matrix whose geometry is characterized by its stable rank. We conduct experiments using several transformer-based pre-trained LLMs: `Llama-3.2-1B-Instruct`, `Qwen2.5-3B-Instruct`, and `Phi-3-mini-4k-instruct`. Prompts consist of (word $\rightarrow$ label) examples followed by a query, with randomized ordering to mitigate positional biases. As in Section 5.1, we evaluate performance under different label allocation patterns, where the query label is selected from the classes with the fewest demonstrations. Pattern 1 is balanced, while Patterns 2 and 3 are imbalanced, with Pattern 3 exhibiting a higher degree of imbalance. For each combination, we construct 200 prompts and measure accuracy, where the label allocation across classes, the sampled examples, and the query label are all randomly chosen. More details on the experimental setup are provided in Appendix G.2.

Table 1: 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 2: Summary statistics characterizing the relationship between stable rank and accuracy differences across pairs of prompting patterns. The reported columns are defined as in Table 1. 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 the 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.

Class-center geometry, accuracy, and robustness to label imbalance. We also compare stable rank with mean pairwise distance, mean angular separation, and entropy-based effective dimension in a separate WordNet experiment with Llama and Qwen using an independent random seed (see Appendix G.2.1). Mean pairwise distance has the strongest Spearman correlation with accuracy for both models under all three label allocations. This finding is consistent with the role of class-center magnitude in our theoretical bound, which depends explicitly on $\|\mathbf{\Lambda}\|_F^2$. In particular, the mean squared pairwise distance is proportional to the squared Frobenius norm of the centered class-center matrix. For the accuracy drop from balanced to imbalanced prompts, stable rank and effective dimension generally show stronger negative correlations than mean pairwise distance.

Our theory offers a possible interpretation. Stable rank measures how evenly the class-center energy is distributed across directions, while the Frobenius norm measures its overall magnitude and is related to mean squared class separation. Consider a prompt with one example from the true class, $S_{k^\star} = 1$, and S examples from a competing class, $S_k = S$. Substituting these counts into the pairwise error bound gives

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

As imbalance increases (i.e., as S increases), the factor multiplying stable rank decreases as $1/S$, while that multiplying the squared Frobenius norm decreases as $1/\sqrt{S}$. The stable-rank-dependent exponent can therefore shrink faster and become a bottleneck in the error bound. A larger stable rank can offset this effect, which may help explain its stronger correlation with accuracy under greater imbalance. This may also explain why isotropy measures correlate more strongly with the accuracy drop than mean pairwise distance in most comparisons.

6 Conclusion and Future Directions

In this work, we mainly focused on the in-context learning (ICL) capability of Transformers for multiclass 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 showed, 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.

Our analysis, however, provides weaker generalization guarantees with too few training tasks relative to the input dimension, low stable rank or weak signal in class centers, or severe label imbalance favoring competing classes. In these cases, the bound may not guarantee small test error. We also rely on the max-margin solution, which describes the limiting direction of training rather than the model obtained after finitely many steps. A natural direction for future work is to extend this analysis to more complex transformer architectures, including nonlinear layers and multi-head attention, where the role of representation geometry under label imbalance remains an open question.

Acknowledgement

Daehan Yoon thanks Junsoo Oh for helpful discussions and insightful feedback. This work was supported by the InnoCORE program of the Ministry of Science and ICT (No. N10250156, 50%) and two Institute of Information & Communications Technology Planning & Evaluation (IITP) grants (No. RS-2024-00457882 National AI Research Lab Project, 30%; No. RS-2019-II190075 Artificial Intelligence Graduate School Program (KAIST), 20%) funded by the Korean government (MSIT).

References

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.

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.

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.

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.

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.

Spencer Frei and Gal Vardi. Trained transformer classifiers generalize and exhibit benign overfitting in-context. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=jwsPS8yRe4.

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.

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.

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

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.

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.

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*, 2024a. URL https://openreview.net/forum?id=I4HTPws9P6.

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*, 2024b. URL https://openreview.net/forum?id=WDX45LNZXE.

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

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz 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.

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

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.

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.

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**

1.1 Related Work 2

2 Preliminaries **3**

2.1 In-context Multiclass Classification 3

2.2 Data Generation and Training Process 4

2.3 Assumptions 5

3 Main Results **5**

3.1 Implicit Bias on Pre-trained Transformers 5

3.2 Transformers Can Learn Multiclass Classification In-Context via Stable Rank . . . 5

3.3 Multiclass In-Context Classification under Test Label Imbalance 6

4 Proof Sketch **7**

5 Experiments and Discussion **8**

5.1 Linear and Nonlinear Transformers under Our Data Distribution 8

5.2 Pre-trained LLMs for WordNet-based Word Classification 9

6 Conclusion and Future Directions **10**

A Sample Complexity Lower Bounds for GMM-Based Classification **15**

B Proof of Our Implicit Bias Results **16**

B.1 Auxiliary Lemmas 18

C Concentrations on the Haar-Distributed Random Matrices **20**

C.1 Bilinear Hanson-Wright inequality 21

C.2 Concentration on $(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top \boldsymbol{W} \boldsymbol{\mu}_{k^\star}$. 23

C.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 $\boldsymbol{z}^\top \boldsymbol{W} \boldsymbol{z}'$. 24

D Analysis of the Max-Margin Solution **25**

D.1 Concentration Bounds for Training Tasks 25

D.2 KKT Conditions and Bounds 32

E Proof of Theorem 3.1 **35**

F Proof of Theorem 3.3 **37**

G Experiment Details **39**

G.1 Linear and Nonlinear Transformers under Our Data Distribution 39

G.2 Pre-trained LLMs for WordNet-based Word Classification 47

A 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 [2019] 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 [Magen and Vardi, 2025] 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 the test task which has singular value matrix $\boldsymbol{\Lambda}$, as shown below.

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

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 following result.

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

B 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 B.1. *There exists a matrix $\boldsymbol{W}_\star$ 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 B.1, 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 D.1, 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 2.2, 2.3 and 2.4, the right-hand side is positive. Consequently, Assumption B.1 holds with high probability for $\boldsymbol{W}_\star = \boldsymbol{I}_d$.

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

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

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

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

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

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

Define $\tilde{\boldsymbol{x}}_{\tau,k} := (\boldsymbol{A}_{\tau,k_\tau^\star} - \boldsymbol{A}_{\tau,k})\boldsymbol{x}_\tau^{(N+1)}$ for $k \neq k_\tau^\star$, and consider the max-margin problem

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

Let $\hat{\boldsymbol{w}}$ denote the vectorized form of $\boldsymbol{W}_{MM}$. For each class $k \in [K]$, define the support set as the tasks for which the corresponding margin constraint is active:

$$\mathcal{S}_k := \left\{\tau \in [B] : k \neq k_\tau^\star,\ \tilde{\boldsymbol{x}}_{\tau,k}^\top \hat{\boldsymbol{w}} = 1\right\}.$$

Since every non-support constraint has margin strictly greater than one, we define

$$\theta := \min_{\substack{\tau\in[B],\, k\in[K] \\ k\neq k_\tau^\star,\, \tau\notin\mathcal{S}_k}} \tilde{\boldsymbol{x}}_{\tau,k}^\top \hat{\boldsymbol{w}} > 1.$$

By the KKT conditions, there exist multipliers $\alpha_{\tau,k} \geq 0$, with $\alpha_{\tau,k_\tau^\star} = 0$, such that

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

Complementary slackness implies that $\alpha_{\tau,k} = 0$ whenever $\tau \notin \mathcal{S}_k$. To characterize the implicit bias, we impose the following assumption on these multipliers and all constraints in the support sets.

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

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

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. [2018], Assumption 4.1 of Ravi et al. [2024]).

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.

Regarding existence, Soudry et al. [2018] establish the existence of such solutions in the binary classification setting for almost all datasets. Ravi et al. [2024] provide empirical results on existence (see Appendix H of Ravi et al. [2024]). Since our setting differs from those considered in the existing results, we also numerically examine the existence of such a solution. See Appendix B.1.1 for details.

Theorem B.3. *Under Assumption B.1 and B.2, and with a sufficiently small step size η, the gradient descent iterates $\boldsymbol{W}(t+1) = \boldsymbol{W}(t) - \eta \nabla \mathcal{L}(\boldsymbol{W}(t))$ converge in direction to the max-margin solution:*

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

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 the 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 and let $\nabla \mathcal{L}(\boldsymbol{w}(t)) = \mathrm{Vec}\left(\nabla \mathcal{L}(\boldsymbol{W}(t))^\top\right)$,

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

The last inequality is due to Lemma B.4, and $\forall x > 0 : x \geq \log(1+x) > 0$. By Lemma B.5, we know that

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

Since $\sum_{t=1}^{\infty} t^{-2} < \infty$, we can obtain

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

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

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

By Lemma B.6, 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,

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

Since $\theta > 1$, $\boldsymbol{r}(t)$ is bounded. □

B.1 Auxiliary Lemmas

Lemma B.4. *Under Assumption B.1 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

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

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

The last inequality is due to Assumption B.1, and $p_{\tau,k}(\boldsymbol{W}) > 0$. □

Lemma B.5 (Lemma 10 in Soudry et al. [2018])**.** *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 B.6 (Lemma 20 in Soudry et al. [2018])**.** *Let $\theta = \min_{\substack{\tau\in[B],\, k\in[K] \\ k\neq k_\tau^\star,\, \tau\notin\mathcal{S}_k}} \tilde{\boldsymbol{x}}_{\tau,k}^\top \hat{\boldsymbol{w}} > 1$. 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 B.2 and Lemma B.5, we obtain the same result.

B.1.1 Numerical Experiments on B.2

The existing theory characterizes the max-margin solution under Assumption B.2, which requires the existence of a vector $\widetilde{w}$ satisfying the corresponding linear system over the support constraints. Since this existence condition is not guaranteed a priori for arbitrary support sets and multipliers, we numerically examine whether it holds in the parameter regime relevant to our theoretical analysis. Let

$$\mathcal{S} := \{(\tau, k) : \alpha_{\tau,k} > 0\} = \bigcup_{k=1}^{K} (\mathcal{S}_k \times \{k\})$$

denote the set of support pairs of the max-margin solution, and let $\boldsymbol{X}_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}| \times d^2}$ be the matrix obtained by stacking the corresponding vectors $\widetilde{\boldsymbol{x}}_{\tau,k} \in \mathbb{R}^{d^2}$ row-wise, where $(\tau, k) \in \mathcal{S}$. Taking logarithms, Assumption B.2 is equivalent to

$$\boldsymbol{X}_{\mathcal{S}} \widetilde{\boldsymbol{w}} = \boldsymbol{b}_{\mathcal{S}}, \qquad \boldsymbol{b}_{\mathcal{S}} = \log(\eta/B)\mathbf{1} - \log \boldsymbol{\alpha}_{\mathcal{S}},$$

where $\boldsymbol{\alpha}_{\mathcal{S}}$ is the vector of multipliers $\alpha_{\tau,k}$ indexed by $\mathcal{S}$. We verify existence using the relative residual

$$r := \frac{\left\| \left(\boldsymbol{I} - \boldsymbol{X}_{\mathcal{S}} \boldsymbol{X}_{\mathcal{S}}^{\dagger} \right) \boldsymbol{b}_{\mathcal{S}} \right\|_2}{\|\boldsymbol{b}_{\mathcal{S}}\|_2}.$$

A value of $r \approx 0$ indicates that the system of equations in Assumption B.2 is numerically feasible. We additionally examine

$$\Delta_{\mathrm{rank}} := |\mathcal{S}| - \mathrm{rank}(\boldsymbol{X}_{\mathcal{S}})$$

and the condition number $\kappa(\boldsymbol{X}_{\mathcal{S}})$. If $\Delta_{\mathrm{rank}} = 0$, then $\boldsymbol{X}_{\mathcal{S}}$ has full row rank, and the system in Assumption B.2 is feasible for every positive support-multiplier vector. The condition number characterizes the numerical stability of the system. Our numerical experiments are conducted under

$$\sigma_1 = 2\sqrt{d}, \qquad \mathrm{sr}(\boldsymbol{\Sigma}) = 0.6K, \qquad \sigma_{2:K} = 2\sqrt{d}\sqrt{\frac{0.6K - 1}{K - 1}}, \qquad B = 4d,$$

using five different random seeds and varying d and K. These choices of stable rank and B correspond to the regime of primary interest in our theoretical framework.

Table 3: Numerical verification of Assumption B.2. Results are reported as the mean and standard deviation over five random seeds.

d	K	B	Total constraints	Support constraints	$\lvert\mathcal{S}\rvert/d^2$	r	Δ_{rank}	$\kappa(X_{\mathcal{S}})$
20	3	80	160	95.6 ± 3.8	0.239 ± 0.009	2.35×10^{-15}	0	4.99 ± 0.33
20	5	80	320	136.8 ± 5.3	0.342 ± 0.013	1.47×10^{-15}	0	6.41 ± 0.69
40	3	160	320	210.2 ± 8.2	0.131 ± 0.005	1.55×10^{-15}	0	4.20 ± 0.25
40	5	160	640	337.0 ± 12.1	0.211 ± 0.008	1.77×10^{-15}	0	5.37 ± 0.20
80	3	320	640	442.4 ± 5.8	0.069 ± 0.001	1.55×10^{-15}	0	3.98 ± 0.33
80	5	320	1280	780.8 ± 5.1	0.122 ± 0.001	1.35×10^{-15}	0	4.72 ± 0.07

The residual r is consistently on the order of 10^{-15}, indicating that $\boldsymbol{b}_{\mathcal{S}}$ lies in the column space of $\boldsymbol{X}_{\mathcal{S}}$ up to machine precision, and hence that the required solution numerically exists. Moreover, $\Delta_{\mathrm{rank}} = 0$ in every configuration shows that $\boldsymbol{X}_{\mathcal{S}}$ has full row rank. Consequently, the system remains feasible for any perturbation of $\boldsymbol{b}_{\mathcal{S}}$, including perturbations of the positive support multipliers. Finally, the moderate condition numbers, ranging approximately from 4 to 6, indicate that the resulting solutions are not excessively sensitive to small perturbations.

C Concentrations on the Haar-Distributed Random Matrices

Before proceeding with the analysis, we introduce two lemmas concerning Haar-distributed random matrices.

Lemma C.1 (Theorem 5.17 of Meckes [2019])**.** *Let S be either $\mathbb{SO}(d)$ or $\mathbb{SO}^-(d)$. Let $f : S \to \mathbb{R}$ be an L-Lipschitz function with respect to the Hilbert-Schmidt metric on the S, and let $\boldsymbol{U}$ be a Haar-distributed random matrix on S. 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 C.2. *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 every matrix in $\mathbb{SO}(d)$. 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 vector orthogonal to $\boldsymbol{e}_1$. Since $\boldsymbol{D}^a$ is antisymmetric, we have $\boldsymbol{e}_1^\top\boldsymbol{D}^a\boldsymbol{e}_1 = 0$, and hence we can conclude that $\boldsymbol{e}_1^\top\boldsymbol{D}^a\boldsymbol{v} = 0$ for every vector $\boldsymbol{v} \in \mathbb{R}^d$. Replacing $\boldsymbol{e}_1$ by any other basis $\boldsymbol{e}_2, \cdots, \boldsymbol{e}_d$ yields $\boldsymbol{D}^a = 0$. 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 $\lambda \in \mathbb{R}$, and hence $\boldsymbol{D} = \lambda\boldsymbol{I}$.

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

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

C.1 Bilinear Hanson-Wright inequality

Lemma C.3 (Hanson-Wright inequality for the uniform distribution 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 C.4 (Bilinear Hanson-Wright inequality for Haar-random matrix)**.** *Fix $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^d$, $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ and let $\boldsymbol{V} \in \mathbb{O}(d)$ be a Haar-distributed random matrix. Then there exist absolute constants $C, c > 0$ such that*

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

Proof. Let $\boldsymbol{P}_{\boldsymbol{v}_1}$ denote the orthogonal projection 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 words,

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

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 rewrite

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

Consider the events

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

Then,

$$\begin{aligned}
&\mathbb{P}\left(\left|\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{Y}\right| \geq t'\right) \\
&\leq \mathbb{P}\left(\left\{\left|\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{Y}\right| \geq t'\right\} \cap \mathcal{E}_1 \cap \mathcal{E}_2\right) + \mathbb{P}(\mathcal{E}_1^{\mathsf{c}}) + \mathbb{P}(\mathcal{E}_2^{\mathsf{c}}).
\end{aligned}$$

Each of the last two terms is bounded by $\exp(-cd)$, by Theorem 3.1.1 of Vershynin [2018].

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

Therefore,

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

By following the proof of Theorem 6.2.1 of Vershynin [2018], we can easily verify that

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

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

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

Consider the event $\mathcal{E}_3 = \left\{|\boldsymbol{X}^\top \boldsymbol{g}_2| \leq \sqrt{(d-1)/2}\right\}$. Then the first term is bounded by

$$\begin{aligned}&(1)\\ &\leq \mathbb{P}\left(\left\{\left|\left(\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} - \frac{\mathrm{Tr}(\boldsymbol{A})}{d}\right)\frac{\boldsymbol{X}^\top \boldsymbol{g}_2}{\|\boldsymbol{P}_{\boldsymbol{X}^\perp}\boldsymbol{g}_2\|}\right| \geq \frac{t'}{4}\right\} \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3\right)\\ &\quad + \mathbb{P}(\mathcal{E}_3^{\mathsf{c}} \cap \mathcal{E}_1 \cap \mathcal{E}_2).\end{aligned}$$

Note that $\boldsymbol{X}^\top \boldsymbol{g}_2 \sim \mathcal{N}(0,1)$; hence, $\mathbb{P}(\mathcal{E}_3^{\mathsf{c}} \cap \mathcal{E}_1 \cap \mathcal{E}_2) \leq \mathbb{P}(\mathcal{E}_3^{\mathsf{c}}) \leq \exp(-cd)$. Under $\mathcal{E}_2$ and $\mathcal{E}_3$, $|\boldsymbol{X}^\top \boldsymbol{g}_2|/\|\boldsymbol{P}_{\boldsymbol{X}^\perp}\boldsymbol{g}_2\| \leq 1$, so we can apply the 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 $|\mathrm{Tr}(\boldsymbol{A})|/d \leq \|\boldsymbol{A}\|_F/\sqrt{d}$ 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,

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

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

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

for some constants $C, c > 0$. □

Lemma C.5. *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 exist absolute constants $C, c > 0$ such that

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

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

Let $\boldsymbol{X}_1(\boldsymbol{U}, \boldsymbol{V}) = (\boldsymbol{e}_k - \boldsymbol{e}_{k^\star})^\top \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{V}^\top \boldsymbol{W} \boldsymbol{V} \boldsymbol{\Lambda}^\top \boldsymbol{U}^\top \boldsymbol{e}_{k^\star}$ 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 4.2,

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

Hence,

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

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

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

The second term can be written as

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

By Lemma C.2, 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 C.1, 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.

C.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 the sub-Gaussian norm of a 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 $\mathbb{S}^{d-1}$ in $\mathbb{R}^d$. By Theorem 3.4.6 from Vershynin [2018], 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 [2025], 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 [2025]. Thus

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

D Analysis of the Max-Margin Solution

We lower-bound the ratio $\mathrm{Tr}(\boldsymbol{W}_{MM})/\|\boldsymbol{W}_{MM}\|_F$ to derive the generalization guarantee. We first prove concentration bounds for the context statistics and query vectors. Following Frei and Vardi [2025], Magen and Vardi [2025], we start by using the KKT conditions of the max-margin solution:

$$\boldsymbol{W}_{MM} = \sum_{k=1}^{K} \sum_{\tau \in \mathcal{S}_k} \alpha_{\tau,k} (\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}) \boldsymbol{x}_\tau^{(N+1)\top}, \quad \alpha_{\tau,k} \geq 0.$$

D.1 Concentration Bounds for Training Tasks

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 assume the balanced regime throughout this section, with $S_\star$ examples per class in each task.

Theorem D.1. *Under the training-task data-generation model, assume that $d \geq K \geq 3$ and that each task contains $S_\star \geq 1$ context examples from each class, so that $N = KS_\star$. There exists an absolute constant $C > 0$ such that, for every $\delta \in (0, 1/6)$, with probability at least $1 - 6\delta$ over $\{\boldsymbol{U}_\tau, \boldsymbol{V}_\tau, (\boldsymbol{x}_\tau^{(j)}, \boldsymbol{y}_\tau^{(j)})_{j=1}^{N+1}\}_{\tau=1}^{B}$, the following inequalities hold simultaneously:*

1. *Cross-task bounds: for all distinct $\tau, \tau' \in [B]$ and all $k_1, k_2 \in [K]$,*

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

2. *Same-task cross-class bounds: for all $\tau \in [B]$ and distinct $k_1, k_2 \in [K] \setminus \{k_\tau^\star\}$,*

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

3. *Same-task norm bounds: for all $\tau \in [B]$ and $k \in [K] \setminus \{k_\tau^\star\}$,*

$$\left| \left\| \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k} \right\|_2^2 - \frac{2S_\star^2}{N^2K} \|\boldsymbol{\Sigma}\|_F^2 \right| \leq \frac{2dS_\star}{N^2} + C \left(\frac{S_\star^2 \sigma_1^2}{N^2\sqrt{K}} + \frac{\sigma_1 S_\star \sqrt{S_\star}}{N^2} + \frac{S_\star \sqrt{d}}{N^2} \right) \log\left(\frac{B^2K^2}{\delta} \right).$$

4. *Query norm and query inner-product bounds. For all distinct $\tau, \tau' \in [B]$,*

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

   For all $\tau \in [B]$,

$$\left| \|\boldsymbol{x}_\tau^{(N+1)}\|_2^2 - \frac{1}{K} \|\boldsymbol{\Sigma}\|_F^2 \right| \leq d + C \left(\frac{\sigma_1^2}{\sqrt{K}} + \sigma_1 + \sqrt{d} \right) \log\left(\frac{B^2K^2}{\delta} \right).$$

   For all $\tau \in [B]$ and $k \in [K] \setminus \{k_\tau^\star\}$,

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

The same constant C applies to all six inequalities and is independent of $B, K, d, S_\star, \boldsymbol{\Sigma}$, and δ.

Proof. Throughout the proof, $c, C > 0$ denote universal constants whose values may change from line to line. Each elementary concentration estimate is made uniform over its task and class indices, with failure probability at most $\delta/8$. At the end of the proof, we choose a single fixed constant large enough for all six inequalities in the theorem.

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

   We condition on all context labels and query classes, which are independent of the Haar matrices and Gaussian noises under the data-generation model. Recall that the context statistics can be jointly written as

$$\hat{\boldsymbol{\mu}}_{\tau,k} = \frac{1}{N} \sum_{j \in \mathcal{C}_{\tau,k}} \boldsymbol{x}_\tau^{(j)} = \frac{S_\star}{N} \boldsymbol{\mu}_{\tau,k} + \frac{\sqrt{S_\star}}{N} \boldsymbol{z}_{\tau,k}, \qquad \boldsymbol{\mu}_{\tau,k} = \boldsymbol{V}_\tau \boldsymbol{\Sigma}^\top \boldsymbol{U}_\tau^\top \boldsymbol{e}_k,$$

   where the vectors $\boldsymbol{z}_{\tau,k}$ are mutually independent standard Gaussian vectors, independent of the class centers. The desired bound is trivial if $k_1 = k_\tau^\star$ or $k_2 = k_{\tau'}^\star$. Otherwise, for fixed $\tau \neq \tau'$ and competitor classes k_1, k_2, define

$$\boldsymbol{z} := \frac{\boldsymbol{z}_{\tau,k_\tau^\star} - \boldsymbol{z}_{\tau,k_1}}{\sqrt{2}}, \qquad \boldsymbol{z}' := \frac{\boldsymbol{z}_{\tau',k_{\tau'}^\star} - \boldsymbol{z}_{\tau',k_2}}{\sqrt{2}}.$$

   These are independent standard Gaussian vectors, independent of the class centers. Expanding the inner product gives

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

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

   Rewriting (1) in terms of the Haar matrices, we obtain

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

   We first verify that its expectation remains zero after conditioning on the determinant components of the Haar matrices. Let S be either $\mathbb{SO}(K)$ or $\mathbb{SO}^-(K)$. By Haar invariance, for every $\boldsymbol{R} \in \mathbb{SO}(K)$,

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

   Each column of this conditional expectation is therefore invariant under every rotation and must be zero. Consequently, independence of the Haar matrices gives the following remark.

Remark D.2. *Let S_1, S_3 be either $\mathbb{SO}(K)$ or $\mathbb{SO}^-(K)$, and let S_2, S_4 be either $\mathbb{SO}(d)$ or $\mathbb{SO}^-(d)$. For $\tau \neq \tau'$, suppose that $\boldsymbol{U}_\tau, \boldsymbol{V}_\tau, \boldsymbol{U}_{\tau'}, \boldsymbol{V}_{\tau'}$ are mutually independent Haar-distributed orthogonal matrices. Then*

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

   Conditionally on $\boldsymbol{U}_\tau, \boldsymbol{U}_{\tau'}$, the two center differences are independent and uniformly distributed on spheres with fixed radii at most $\sqrt{2}\sigma_1$. If either radius is zero, their inner product vanishes. Otherwise, sphere concentration gives, for $t > 0$,

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

Averaging over the left Haar matrices and applying a union bound over the task and class indices yields

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

Choosing $t=C\sigma_1^2\sqrt{\log(16B^2K^2/\delta)/d}$ with a sufficiently large universal constant gives failure probability at most $\delta/8$. Since $\log(16B^2K^2/\delta)\leq 2\log(B^2K^2/\delta)$, with probability at least $1-\delta/8$, uniformly over the relevant indices,

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

For (2) and (3), condition on the class centers and use $\|\boldsymbol{\mu}_{\tau,k_\tau^\star}-\boldsymbol{\mu}_{\tau,k_1}\|_2\leq\sqrt{2}\sigma_1$. Conditional Gaussian tails and a union bound give each of the two elementary uniform estimates with failure probability at most $\delta/8$. Thus, with probability at least $1-\delta/4$, both bounds hold simultaneously:

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

Here and below, $\boldsymbol{z},\boldsymbol{z}'$ denote the normalized noises associated with each indexed tuple; independence across different tuples is not needed for the union bound.

For (4), the inner product of two independent standard Gaussian vectors satisfies

$$\mathbb{P}\left(|\langle\boldsymbol{z},\boldsymbol{z}'\rangle|\geq t\right)\leq 2\exp\left(-c\min\left\{\frac{t^2}{d},t\right\}\right).$$

Applying a union bound with the same probability allocation, with probability at least $1-\delta/8$, uniformly over the relevant indices,

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

Combining the four elementary estimates and using $\sqrt{\log(B^2K^2/\delta)}\leq\log(B^2K^2/\delta)$, we obtain, with probability at least $1-\delta/2$, uniformly over all $\tau\neq\tau'$ and $k_1,k_2\in[K]$,

$$\left|\left\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star}-\hat{\boldsymbol{\mu}}_{\tau,k_1},\hat{\boldsymbol{\mu}}_{\tau',k_{\tau'}^\star}-\hat{\boldsymbol{\mu}}_{\tau',k_2}\right\rangle\right|$$
$$\leq C\left(\frac{S_\star^2\sigma_1^2}{N^2\sqrt{d}}+\frac{\sigma_1S_\star\sqrt{S_\star}}{N^2}+\frac{S_\star\sqrt{d}}{N^2}\right)\log\left(\frac{B^2K^2}{\delta}\right).$$

The conditional probability bounds are uniform in the labels, so averaging over the labels gives the same unconditional guarantee. In particular, this conclusion has failure probability at most δ, as required.

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

Fix $\tau\in[B]$ and distinct $k_1,k_2\in[K]\setminus\{k_\tau^\star\}$. Using the joint Gaussian representation introduced above, recall that

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

For this calculation, set $\boldsymbol{z} := \boldsymbol{z}_{\tau,k_\tau^\star}$, $\boldsymbol{z}_1 := -\boldsymbol{z}_{\tau,k_1}$, and $\boldsymbol{z}_2 := -\boldsymbol{z}_{\tau,k_2}$. These are mutually independent standard Gaussian vectors, independent of the class centers. Expanding gives

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

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

For (1), set $\boldsymbol{v} := \boldsymbol{e}_{k_\tau^\star} - \boldsymbol{e}_{k_1}$ and $\boldsymbol{w} := \boldsymbol{e}_{k_\tau^\star} - \boldsymbol{e}_{k_2}$. Since the three class indices are distinct, $\|\boldsymbol{v}\|_2 = \|\boldsymbol{w}\|_2 = \sqrt{2}$ and $\boldsymbol{v}^\top \boldsymbol{w} = 1$. Using $\boldsymbol{V}_\tau^\top \boldsymbol{V}_\tau = \boldsymbol{I}_d$, we obtain

$$(1) = f(\boldsymbol{U}_\tau) := \boldsymbol{v}^\top \boldsymbol{U}_\tau \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \boldsymbol{U}_\tau^\top \boldsymbol{w}.$$

For either $S = \mathbb{SO}(K)$ or $S = \mathbb{SO}^-(K)$, Lemma C.2 gives

$$\mathbb{E}[\boldsymbol{U}_\tau \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \boldsymbol{U}_\tau^\top \mid \boldsymbol{U}_\tau \in S] = \frac{\|\boldsymbol{\Sigma}\|_F^2}{K} \boldsymbol{I}_K,$$
$$\mathbb{E}[f(\boldsymbol{U}_\tau) \mid \boldsymbol{U}_\tau \in S] = \frac{\|\boldsymbol{\Sigma}\|_F^2}{K}.$$

Moreover, for orthogonal matrices $\boldsymbol{U}, \boldsymbol{U}'$,

$$\begin{aligned} |f(\boldsymbol{U}) - f(\boldsymbol{U}')| &\le \left|\boldsymbol{v}^\top (\boldsymbol{U} - \boldsymbol{U}') \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \boldsymbol{U}^\top \boldsymbol{w}\right| + \left|\boldsymbol{v}^\top \boldsymbol{U}' \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top (\boldsymbol{U} - \boldsymbol{U}')^\top \boldsymbol{w}\right| \\ &\le 2\sigma_1^2 \|\boldsymbol{v}\|_2 \|\boldsymbol{w}\|_2 \|\boldsymbol{U} - \boldsymbol{U}'\|_{op} \le 4\sigma_1^2 \|\boldsymbol{U} - \boldsymbol{U}'\|_F. \end{aligned}$$

Applying Lemma C.1 to f and $-f$ on each determinant component gives a two-sided concentration bound around the same conditional mean. A union bound over at most BK^2 index tuples, with the probability allocation specified above, therefore gives, with probability at least $1 - \delta/8$,

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

uniformly over all $\tau \in [B]$ and distinct $k_1, k_2 \neq k_\tau^\star$.

For (2) and (3), condition on the class centers. Each noise sum has distribution $\mathcal{N}(0, 2\boldsymbol{I}_d)$, and each center difference has norm at most $\sqrt{2}\sigma_1$. Conditional Gaussian tails and union bounds give the two elementary uniform estimates, each with failure probability at most $\delta/8$. Thus, with probability at least $1 - \delta/4$,

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

uniformly over the same indices. Independence between these two mixed terms is not required.

For (4), the shared Gaussian vector $\boldsymbol{z}$ contributes the nonzero mean $\mathbb{E}[\langle \boldsymbol{z} + \boldsymbol{z}_1, \boldsymbol{z} + \boldsymbol{z}_2 \rangle] = d$. Centering at this mean gives

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

The chi-square tail bound

$$\mathbb{P}\left(|\|\boldsymbol{z}\|_2^2 - d| \ge t\right) \le 2\exp\left(-c\min\left\{\frac{t^2}{d}, t\right\}\right)$$

and a union bound imply, with probability at least $1 - \delta/8$, the uniform estimate

$$|\|\boldsymbol{z}\|_2^2 - d| \le C\left(\sqrt{d\log\left(\frac{B^2K^2}{\delta}\right)} + \log\left(\frac{B^2K^2}{\delta}\right)\right) \le C\sqrt{d}\log\left(\frac{B^2K^2}{\delta}\right).$$

Each of the three Gaussian inner products satisfies the same final bound, with failure probability at most $\delta/8$ after taking a union bound over the indices. Consequently, with probability at least $1-\delta/2$,

$$|\langle \boldsymbol{z}+\boldsymbol{z}_1, \boldsymbol{z}+\boldsymbol{z}_2\rangle - d| \le C\sqrt{d}\log\left(\frac{B^2K^2}{\delta}\right)$$

uniformly over the relevant indices.

Combining the seven elementary estimates gives, with probability at least $1-7\delta/8$,

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

By the triangle inequality, this implies the desired bound

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

The conditional estimates are uniform in the labels, so the same guarantee holds unconditionally. Since $7\delta/8 \le \delta$, this conclusion has the required failure probability.

3. Bound on $|\|\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}\|_2^2 - \frac{2S_\star^2}{N^2K}\|\boldsymbol{\Sigma}\|_F^2|$.

Fix $\tau \in [B]$ and $k \ne k_\tau^\star$, and define $\boldsymbol{z} := (\boldsymbol{z}_{\tau,k_\tau^\star} - \boldsymbol{z}_{\tau,k})/\sqrt{2}$. This is a standard Gaussian vector independent of the class centers. Therefore,

$$\begin{aligned}\|\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}\|_2^2 = \frac{S_\star^2}{N^2}\|\boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k}\|_2^2 \\ + \frac{2\sqrt{2}S_\star\sqrt{S_\star}}{N^2}\left\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k}, \boldsymbol{z}\right\rangle + \frac{2S_\star}{N^2}\|\boldsymbol{z}\|_2^2.\end{aligned}$$

For the center term, apply the argument in part 2 with $\boldsymbol{v} = \boldsymbol{w} = \boldsymbol{e}_{k_\tau^\star} - \boldsymbol{e}_k$. The conditional mean on either determinant component is now $2\|\boldsymbol{\Sigma}\|_F^2/K$, while the Lipschitz constant remains at most $4\sigma_1^2$. Hence, with probability at least $1-\delta/8$, uniformly over all $\tau \in [B]$ and $k \ne k_\tau^\star$,

$$\left|\|\boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k}\|_2^2 - \frac{2}{K}\|\boldsymbol{\Sigma}\|_F^2\right| \le C\frac{\sigma_1^2}{\sqrt{K}}\sqrt{\log\left(\frac{B^2K^2}{\delta}\right)}.$$

Conditional Gaussian tails give, with failure probability at most $\delta/8$, the uniform estimate

$$\left|\left\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k}, \boldsymbol{z}\right\rangle\right| \le C\sigma_1\sqrt{\log\left(\frac{B^2K^2}{\delta}\right)}.$$

Likewise, the chi-square tail bound gives, with failure probability at most $\delta/8$, the uniform estimate

$$|\|\boldsymbol{z}\|_2^2 - d| \le C\sqrt{d}\log\left(\frac{B^2K^2}{\delta}\right).$$

Combining these three estimates and using $\sqrt{\log(B^2K^2/\delta)} \le \log(B^2K^2/\delta)$, with probability at least $1-3\delta/8$,

$$\begin{aligned}&\left|\|\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}\|_2^2 - \frac{2S_\star^2}{N^2K}\|\boldsymbol{\Sigma}\|_F^2 - \frac{2dS_\star}{N^2}\right| \\ &\le C\left(\frac{S_\star^2\sigma_1^2}{N^2\sqrt{K}} + \frac{\sigma_1 S_\star\sqrt{S_\star}}{N^2} + \frac{S_\star\sqrt{d}}{N^2}\right)\log\left(\frac{B^2K^2}{\delta}\right).\end{aligned}$$

Consequently,

$$\left|\|\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}\|_2^2 - \frac{2S_\star^2}{N^2K}\|\boldsymbol{\Sigma}\|_F^2\right| \leq \frac{2dS_\star}{N^2} + C\left(\frac{S_\star^2\sigma_1^2}{N^2\sqrt{K}} + \frac{\sigma_1 S_\star\sqrt{S_\star}}{N^2} + \frac{S_\star\sqrt{d}}{N^2}\right)\log\left(\frac{B^2K^2}{\delta}\right).$$

The factor 2 in the deterministic noise term arises because the difference of two independent class noises has covariance $2\boldsymbol{I}_d$. Averaging over the labels preserves the probability bound, and $3\delta/8 \leq \delta$ gives the required failure probability.

4. Bound on $|\|\boldsymbol{x}_\tau^{(N+1)}\|_2^2 - \frac{1}{K}\|\boldsymbol{\Sigma}\|_F^2|$.

Recall that $\boldsymbol{x}_\tau^{(N+1)} = \boldsymbol{\mu}_{\tau,k_\tau^\star} + \boldsymbol{z}_\tau$, where $\boldsymbol{z}_\tau$ is standard Gaussian and independent of the class centers. Thus,

$$\|\boldsymbol{x}_\tau^{(N+1)}\|_2^2 = \|\boldsymbol{\mu}_{\tau,k_\tau^\star}\|_2^2 + 2\langle\boldsymbol{\mu}_{\tau,k_\tau^\star}, \boldsymbol{z}_\tau\rangle + \|\boldsymbol{z}_\tau\|_2^2.$$

The center norm is a Haar quadratic form:

$$\|\boldsymbol{\mu}_{\tau,k_\tau^\star}\|_2^2 = \boldsymbol{e}_{k_\tau^\star}^\top \boldsymbol{U}_\tau\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\boldsymbol{U}_\tau^\top\boldsymbol{e}_{k_\tau^\star}.$$

By Lemma C.2, its conditional mean on either determinant component is $\|\boldsymbol{\Sigma}\|_F^2/K$. Its Lipschitz constant with respect to the Frobenius metric is at most $2\sigma_1^2$. Applying Lemma C.1 on each component and taking a union bound over $\tau \in [B]$ therefore gives the first estimate below. Conditional Gaussian tails, using $\|\boldsymbol{\mu}_{\tau,k_\tau^\star}\|_2 \leq \sigma_1$, give the second estimate, and chi-square tails give the third. Each estimate holds uniformly over $\tau \in [B]$ with failure probability at most $\delta/8$:

$$\left|\|\boldsymbol{\mu}_{\tau,k_\tau^\star}\|_2^2 - \frac{1}{K}\|\boldsymbol{\Sigma}\|_F^2\right| \leq C\frac{\sigma_1^2}{\sqrt{K}}\sqrt{\log\left(\frac{B^2K^2}{\delta}\right)},$$

$$|\langle\boldsymbol{\mu}_{\tau,k_\tau^\star}, \boldsymbol{z}_\tau\rangle| \leq C\sigma_1\sqrt{\log\left(\frac{B^2K^2}{\delta}\right)},$$

$$\left|\|\boldsymbol{z}_\tau\|_2^2 - d\right| \leq C\left[\sqrt{d\log\left(\frac{B^2K^2}{\delta}\right)} + \log\left(\frac{B^2K^2}{\delta}\right)\right].$$

Combining these three estimates by a union bound and using $\log(B^2K^2/\delta) > 1$, with probability at least $1 - 3\delta/8$, uniformly over $\tau \in [B]$,

$$\begin{aligned}&\left|\|\boldsymbol{x}_\tau^{(N+1)}\|_2^2 - \frac{1}{K}\|\boldsymbol{\Sigma}\|_F^2\right| \\ &\leq d + \left|\|\boldsymbol{\mu}_{\tau,k_\tau^\star}\|_2^2 - \frac{1}{K}\|\boldsymbol{\Sigma}\|_F^2\right| + 2|\langle\boldsymbol{\mu}_{\tau,k_\tau^\star}, \boldsymbol{z}_\tau\rangle| + \left|\|\boldsymbol{z}_\tau\|_2^2 - d\right| \\ &\leq d + C\left(\frac{\sigma_1^2}{\sqrt{K}} + \sigma_1 + \sqrt{d}\right)\log\left(\frac{B^2K^2}{\delta}\right).\end{aligned}$$

The conditional estimates are uniform in the labels, so averaging over the labels preserves this guarantee. Since $3\delta/8 \leq \delta$, the conclusion has the required failure probability.

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

Fix $\tau \in [B]$ and $k \neq k_\tau^\star$. Set $\boldsymbol{z} := (\boldsymbol{z}_{\tau,k_\tau^\star} - \boldsymbol{z}_{\tau,k})/\sqrt{2}$, and let $\boldsymbol{z}' := \boldsymbol{x}_\tau^{(N+1)} - \boldsymbol{\mu}_{\tau,k_\tau^\star}$ be the query noise. These are independent standard Gaussian vectors, independent of the class centers. Expanding gives

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

For the center term, apply the argument in part 2 with $\boldsymbol{v} := \boldsymbol{e}_{k_\tau^\star} - \boldsymbol{e}_k$ and $\boldsymbol{w} := \boldsymbol{e}_{k_\tau^\star}$. Since $\boldsymbol{v}^\top \boldsymbol{w} = 1$, Lemma C.2 gives conditional mean $\|\boldsymbol{\Sigma}\|_F^2/K$ on either determinant component. The Lipschitz constant is at most $2\sigma_1^2\|\boldsymbol{v}\|_2\|\boldsymbol{w}\|_2 = 2\sqrt{2}\sigma_1^2$. Lemma C.1 and a union bound therefore give the first estimate below. Conditional Gaussian tails give the two mixed-term estimates, and the Gaussian inner-product tail bound gives the last estimate. Each of these four elementary estimates holds uniformly over $\tau \in [B]$ and $k \neq k_\tau^\star$ with failure probability at most $\delta/8$:

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

$$\left|\left\langle \boldsymbol{\mu}_{\tau,k_\tau^\star} - \boldsymbol{\mu}_{\tau,k}, \boldsymbol{z}'\right\rangle\right| \le C\sigma_1\sqrt{\log\left(\frac{B^2K^2}{\delta}\right)},$$

$$\left|\left\langle \boldsymbol{z}, \boldsymbol{\mu}_{\tau,k_\tau^\star}\right\rangle\right| \le C\sigma_1\sqrt{\log\left(\frac{B^2K^2}{\delta}\right)},$$

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

Combining these estimates gives, with probability at least $1 - \delta/2$, uniformly over the relevant indices,

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

The last inequality uses $S_\star \ge 1$ and $\log(B^2K^2/\delta) > 1$. The conditional estimates are uniform in the labels, so averaging over the labels preserves this guarantee. In particular, the failure probability is at most δ.

6. Bound on $|\langle \boldsymbol{x}_\tau^{(N+1)}, \boldsymbol{x}_{\tau'}^{(N+1)}\rangle|$ for $\tau \neq \tau'$.

Write $\boldsymbol{x}_\tau^{(N+1)} = \boldsymbol{\mu}_{\tau,k_\tau^\star} + \boldsymbol{z}_\tau$, where the noise vectors $\boldsymbol{z}_\tau \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ are independent across tasks and independent of the class centers. Then,

$$\begin{aligned}
&\langle \boldsymbol{x}_\tau^{(N+1)}, \boldsymbol{x}_{\tau'}^{(N+1)}\rangle \\
&= \langle \boldsymbol{\mu}_{\tau,k_\tau^\star}, \boldsymbol{\mu}_{\tau',k_{\tau'}^\star}\rangle + \langle \boldsymbol{\mu}_{\tau,k_\tau^\star}, \boldsymbol{z}_{\tau'}\rangle \\
&\quad + \langle \boldsymbol{z}_\tau, \boldsymbol{\mu}_{\tau',k_{\tau'}^\star}\rangle + \langle \boldsymbol{z}_\tau, \boldsymbol{z}_{\tau'}\rangle.
\end{aligned}$$

Conditional on $\boldsymbol{U}_\tau$ and $\boldsymbol{U}_{\tau'}$, the two class centers are independent random rotations of fixed vectors with norms at most σ_1. Applying spherical concentration, conditional Gaussian concentration, and Gaussian inner-product concentration, followed by union bounds over the task and class indices, gives the following estimates. Each estimate holds uniformly over $\tau \neq \tau'$ with failure probability at most $\delta/8$:

$$\left|\langle \boldsymbol{\mu}_{\tau,k_\tau^\star}, \boldsymbol{\mu}_{\tau',k_{\tau'}^\star}\rangle\right| \le C\frac{\sigma_1^2}{\sqrt{d}}\sqrt{\log\left(\frac{B^2K^2}{\delta}\right)},$$

$$\left|\langle \boldsymbol{\mu}_{\tau,k_\tau^\star}, \boldsymbol{z}_{\tau'}\rangle\right| \le C\sigma_1\sqrt{\log\left(\frac{B^2K^2}{\delta}\right)},$$

$$\left|\langle \boldsymbol{z}_\tau, \boldsymbol{\mu}_{\tau',k_{\tau'}^\star}\rangle\right| \le C\sigma_1\sqrt{\log\left(\frac{B^2K^2}{\delta}\right)},$$

$$\left|\langle \boldsymbol{z}_\tau, \boldsymbol{z}_{\tau'}\rangle\right| \le C\left[\sqrt{d\log\left(\frac{B^2K^2}{\delta}\right)} + \log\left(\frac{B^2K^2}{\delta}\right)\right].$$

By a union bound, these four estimates hold simultaneously with probability at least $1 - \delta/2$. Since $d \geq 1$ and $\log(B^2K^2/\delta) \geq 1$, the triangle inequality yields, uniformly over $\tau \neq \tau'$,

$$\begin{aligned}&|\langle \boldsymbol{x}_\tau^{(N+1)}, \boldsymbol{x}_{\tau'}^{(N+1)} \rangle| \\ &\leq C\left(\frac{\sigma_1^2}{\sqrt{d}} + \sigma_1 + \sqrt{d}\right)\log\left(\frac{B^2K^2}{\delta}\right).\end{aligned}$$

Each of the six conclusions has failure probability at most δ. A union bound therefore shows that all six hold simultaneously with probability at least $1 - 6\delta$. We now fix a single absolute constant $C \geq 1$ large enough that all six inequalities in Theorem D.1 hold with this same constant.

Using $N = KS_\star$, $S_\star \geq 1$, $d \geq K$, and $\sigma_1^2 \geq c_\sigma d$, define the fixed constant

$$c_0 := C\left(1 + c_\sigma^{-1/2} + c_\sigma^{-1}\right).$$

The preceding bounds then imply the following inequalities on the same event. For all distinct $\tau, \tau' \in [B]$ and all $k_1, k_2 \in [K]$,

$$\left|\left\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_1}, \hat{\boldsymbol{\mu}}_{\tau',k_{\tau'}^\star} - \hat{\boldsymbol{\mu}}_{\tau',k_2} \right\rangle\right| \leq c_0 \frac{\sigma_1^2}{K^2\sqrt{d}} \log\left(\frac{B^2K^2}{\delta}\right).$$

For all $\tau \in [B]$ and distinct $k_1, k_2 \in [K] \setminus \{k_\tau^\star\}$,

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

For all $\tau \in [B]$ and $k \in [K] \setminus \{k_\tau^\star\}$,

$$\left|\|\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}\|_2^2 - \frac{2}{K^3}\|\boldsymbol{\Sigma}\|_F^2\right| \leq \frac{2d}{S_\star K^2} + c_0 \frac{\sigma_1^2}{K^2\sqrt{K}} \log\left(\frac{B^2K^2}{\delta}\right).$$

For all distinct $\tau, \tau' \in [B]$,

$$|\langle \boldsymbol{x}_\tau^{(N+1)}, \boldsymbol{x}_{\tau'}^{(N+1)} \rangle| \leq c_0 \frac{\sigma_1^2}{\sqrt{d}} \log\left(\frac{B^2K^2}{\delta}\right).$$

For all $\tau \in [B]$,

$$\left|\|\boldsymbol{x}_\tau^{(N+1)}\|_2^2 - \frac{1}{K}\|\boldsymbol{\Sigma}\|_F^2\right| \leq d + c_0 \frac{\sigma_1^2}{\sqrt{K}} \log\left(\frac{B^2K^2}{\delta}\right).$$

Finally, for all $\tau \in [B]$ and $k \in [K] \setminus \{k_\tau^\star\}$,

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

The constant c_0 depends only on the fixed assumption constant c_σ and remains fixed throughout the subsequent analysis. □

D.2 KKT Conditions and Bounds

The quantity of interest is $\mathrm{Tr}(\boldsymbol{W}_{MM})/\|\boldsymbol{W}_{MM}\|_F$, which is scale-invariant. We focus on establishing a lower bound for it. For convenience, let $\alpha_{\tau,k_\tau^\star} = 0$. By the KKT conditions, the max-margin solution satisfies:

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

Hence,

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

Recall the last part of Theorem D.1,

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

From Assumption 2.2 and 2.4, 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 $\boldsymbol{W}_{MM}$, for every $q \in [B]$ and $s \in [K] \setminus \{k_q^\star\}$,

$$\begin{aligned}1 &\leq (\hat{\boldsymbol{\mu}}_{q,k_q^\star} - \hat{\boldsymbol{\mu}}_{q,s})^\top \sum_{\tau=1}^{B}\sum_{k=1}^{K} \alpha_{\tau,k}(\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k})\boldsymbol{x}_\tau^{(N+1)\top}\boldsymbol{x}_q^{(N+1)} \\ &= \alpha_{q,s}\|\hat{\boldsymbol{\mu}}_{q,k_q^\star} - \hat{\boldsymbol{\mu}}_{q,s}\|_2^2\|\boldsymbol{x}_q^{(N+1)}\|_2^2 + \sum_{k\neq s}\alpha_{q,k}\left\langle \hat{\boldsymbol{\mu}}_{q,k_q^\star} - \hat{\boldsymbol{\mu}}_{q,s}, \hat{\boldsymbol{\mu}}_{q,k_q^\star} - \hat{\boldsymbol{\mu}}_{q,k}\right\rangle \|\boldsymbol{x}_q^{(N+1)}\|_2^2 \\ &\quad + \sum_{\tau\neq q}\sum_{k=1}^{K}\alpha_{\tau,k}\left\langle \hat{\boldsymbol{\mu}}_{q,k_q^\star} - \hat{\boldsymbol{\mu}}_{q,s}, \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}\right\rangle \left\langle \boldsymbol{x}_\tau^{(N+1)}, \boldsymbol{x}_q^{(N+1)}\right\rangle.\end{aligned}$$

Averaging over the $B(K-1)$ valid task-class pairs (q,s) yields

$$\begin{aligned}1 &\leq \frac{1}{B(K-1)}\sum_{\substack{q\in[B]\\ s\in[K]\setminus\{k_q^\star\}}} \alpha_{q,s}\underbrace{\|\hat{\boldsymbol{\mu}}_{q,k_q^\star} - \hat{\boldsymbol{\mu}}_{q,s}\|_2^2\|\boldsymbol{x}_q^{(N+1)}\|_2^2}_{(1)} \\ &\quad + \frac{1}{B(K-1)}\sum_{\substack{q\in[B]\\ s\in[K]\setminus\{k_q^\star\}}}\sum_{k\neq s}\alpha_{q,k}\underbrace{\left\langle \hat{\boldsymbol{\mu}}_{q,k_q^\star} - \hat{\boldsymbol{\mu}}_{q,s}, \hat{\boldsymbol{\mu}}_{q,k_q^\star} - \hat{\boldsymbol{\mu}}_{q,k}\right\rangle \|\boldsymbol{x}_q^{(N+1)}\|_2^2}_{(2)} \\ &\quad + \frac{1}{B(K-1)}\sum_{\substack{q\in[B]\\ s\in[K]\setminus\{k_q^\star\}}}\sum_{\tau\neq q}\sum_{k=1}^{K}\alpha_{\tau,k}\underbrace{\left\langle \hat{\boldsymbol{\mu}}_{q,k_q^\star} - \hat{\boldsymbol{\mu}}_{q,s}, \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}\right\rangle \left\langle \boldsymbol{x}_\tau^{(N+1)}, \boldsymbol{x}_q^{(N+1)}\right\rangle}_{(3)}.\end{aligned}$$

We next derive uniform upper bounds on the three products (1), (2), and (3) using the simplified concentration bounds following Theorem D.1.

1. Bounds on (1) and (2).
   We bound both products using the norm estimates and Cauchy–Schwarz. By the simplified concentration bounds, $\mathrm{sr}(\boldsymbol{\Sigma}) \leq K$, $\sigma_1^2 \geq c_\sigma d$, and $S_\star \geq 1$, for every $\tau \in [B]$ and $k \neq k_\tau^\star$,

$$\begin{aligned}&\|\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}\|_2^2 \\ &\leq \frac{\sigma_1^2}{K^3}\left(2\,\mathrm{sr}(\boldsymbol{\Sigma}) + \frac{2dK}{S_\star\sigma_1^2} + c_0\sqrt{K}\log\left(\frac{B^2K^2}{\delta}\right)\right) \\ &\leq \frac{\sigma_1^2}{K^2}\left(2 + \frac{2}{c_\sigma} + \frac{c_0}{\sqrt{K}}\log\left(\frac{B^2K^2}{\delta}\right)\right).\end{aligned}$$

   Likewise, for every $\tau \in [B]$,

$$\begin{aligned}&\|\boldsymbol{x}_\tau^{(N+1)}\|_2^2 \\ &\leq \frac{\sigma_1^2}{K}\left(\mathrm{sr}(\boldsymbol{\Sigma}) + \frac{dK}{\sigma_1^2} + c_0\sqrt{K}\log\left(\frac{B^2K^2}{\delta}\right)\right) \\ &\leq \sigma_1^2\left(1 + \frac{1}{c_\sigma} + \frac{c_0}{\sqrt{K}}\log\left(\frac{B^2K^2}{\delta}\right)\right).\end{aligned}$$

Therefore, for all $k_1, k_2 \in [K] \setminus \{k_\tau^\star\}$, including $k_1 = k_2$, Cauchy–Schwarz gives

$$\begin{aligned}
&\left|\left\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_1}, \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_2} \right\rangle\right| \|\boldsymbol{x}_\tau^{(N+1)}\|_2^2 \\
&\le \|\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_1}\|_2 \|\hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k_2}\|_2 \|\boldsymbol{x}_\tau^{(N+1)}\|_2^2 \\
&\le \frac{\sigma_1^4}{K^2}\left(2 + \frac{2}{c_\sigma} + \frac{c_0}{\sqrt{K}}\log\left(\frac{B^2K^2}{\delta}\right)\right)\left(1 + \frac{1}{c_\sigma} + \frac{c_0}{\sqrt{K}}\log\left(\frac{B^2K^2}{\delta}\right)\right).
\end{aligned}$$

This provides a common upper bound for (1) and the absolute value of (2). Define the fixed constants $A_1 := 2(1 + c_\sigma^{-1})^2$ and $A_2 := 3(1 + c_\sigma^{-1})$. Expanding the preceding product gives the following common upper bound for (1) and the absolute value of (2):

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

2. Bound of (3). For $\tau \neq \tau'$, the cross-task concentration bounds give

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

This bounds the absolute value of (3).

Each multiplier $\alpha_{\tau,k}$ with $k \neq k_\tau^\star$ appears once in (1), $K-2$ times in (2), and $(B-1)(K-1)$ times in (3). Combining these contributions and using $(B-1)/B \le 1$ yields

$$\begin{aligned}
1 &\le \left(\sum_{\tau,k}\alpha_{\tau,k}\right)\frac{\sigma_1^4}{dK^2}\left[\frac{d}{B}\left(A_1 + \frac{A_2c_0}{\sqrt{K}}\log\left(\frac{B^2K^2}{\delta}\right) + \frac{c_0^2}{K}\log^2\left(\frac{B^2K^2}{\delta}\right)\right)\right. \\
&\qquad \left. + \frac{B-1}{B}c_0^2\log^2\left(\frac{B^2K^2}{\delta}\right)\right] \\
&\le \left(\sum_{\tau,k}\alpha_{\tau,k}\right)\frac{\sigma_1^4\xi}{dK(K-1)}.
\end{aligned}$$

where we define

$$\xi := \frac{d}{B}A_1 + \frac{d}{B\sqrt{K}}A_2c_0\log\left(\frac{B^2K^2}{\delta}\right) + c_0^2\left(1 + \frac{d}{BK}\right)\log^2\left(\frac{B^2K^2}{\delta}\right).$$

Therefore,

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

We can observe that $\xi = O\left(\frac{d}{B} \vee \log^2\left(\frac{B^2K^2}{\delta}\right)\right)$. Since $\langle \hat{\boldsymbol{\mu}}_{\tau,k_\tau^\star} - \hat{\boldsymbol{\mu}}_{\tau,k}, \boldsymbol{x}_\tau^{(N+1)}\rangle \ge \frac{\sigma_1^2}{K}\left(c_s - \frac{c_0}{\sqrt{K}}\log\left(\frac{B^2K^2}{\delta}\right)\right)$, 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,

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

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

E Proof of Theorem 3.1

For notational simplicity, denote the max-margin solution by $\boldsymbol{W}$. 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 the $\boldsymbol{z}_k$ are i.i.d. Gaussian random vectors from $\mathcal{N}(0, \boldsymbol{I}_d)$. Hence,

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

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

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 E.1. *For any $t > 0$, 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 E.1,

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

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

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

Applying Lemma C.2, 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 C.5,

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

$$\le C\exp\left(-c'\min\left(\frac{\mathrm{Tr}(\boldsymbol{W})^2\,\mathrm{sr}(\boldsymbol{\Lambda})^2}{\|\boldsymbol{W}\|_F^2 K^2}, \frac{\mathrm{Tr}(\boldsymbol{W})\,\mathrm{sr}(\boldsymbol{\Lambda})}{\|\boldsymbol{W}\|_F K}\right)\right) + \exp\left(-\frac{c'\,\mathrm{sr}(\boldsymbol{\Lambda})^2}{K}\right).$$

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

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

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,

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

F Proof of Theorem 3.3

Recall that S_k is the number of samples for a corresponding cluster k, and recall that $\hat{\boldsymbol{\mu}}_k \overset{d}{=} \frac{S_k}{M}\boldsymbol{\mu}_k + \frac{\sqrt{S_k}}{M}\boldsymbol{z}_k$. 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

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

By Lemma C.2,

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

Therefore,

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

For 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 4.2,

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

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

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

Let $t = \frac{S_{k^\star}\operatorname{Tr}(\boldsymbol{W})}{2dK}\|\boldsymbol{\Lambda}\|_F^2$. Using $\frac{\operatorname{Tr}(\boldsymbol{W})}{\|\boldsymbol{W}\|_F} = \Omega\left(\sqrt{d}\xi^{-1}\right)$, we obtain the following concentration bound for the first term:

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

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

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

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

$$\mathbb{P}\left(|\sqrt{S_k + S_{k^\star}}\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{z}'| \geq t\right) \leq 2\exp\left(-\frac{ct}{\sqrt{S_k + S_{k^\star}}\|\boldsymbol{W}\|_F}\right).$$

Substituting $t = \frac{S_{k^\star}\,\mathrm{Tr}(\boldsymbol{W})}{8dK}\|\boldsymbol{\Lambda}\|_F^2$ gives the following bound on the sum of the last three terms:

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

G Experiment Details

We perform experiments using NVIDIA GeForce RTX 3090 GPUs.

G.1 Linear and Nonlinear Transformers under Our Data Distribution

We conduct experiments on in-context multiclass classification using synthetic data generated according to Definition 2.1. We consider an input dimension of $d = 20$ and $K \in \{5, 10\}$ classes.

We evaluate both the linear self-attention model studied in our theory and nonlinear Transformers. For the linear model, we use the parameterization in Section 2.1, fixing the label-readout block $W_\star^V = [0_{K\times d}\ I_K]$ and directly training W. The nonlinear models use 2 attention heads and a model dimension of 128, without positional encoding. We vary their depth across 1, 3, and 5 layers. Each layer contains a feedforward network with hidden dimension 512, GELU activations, 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 singular values of the training class-center matrix are set to

$$(\sigma_1, \dots, \sigma_K) = \begin{cases} \sqrt{20}\,(2, \sqrt{2}, \sqrt{2}, \sqrt{2}, \sqrt{2}), & K = 5, \\ \sqrt{20}\,(2, 2, 2, 1.5, 1.5, 1.5, 1.5, 1, 1, 1), & K = 10. \end{cases}$$

These choices yield stable ranks of 3 and 6, respectively, consistent with Assumption 2.2. The noise is sampled from $\mathcal{N}(0, I_d)$, and each training prompt contains 10 demonstrations per class.

For evaluation, we vary the stable rank while fixing the squared Frobenius norm of the class-center matrix to 120 for $K = 5$ and 480 for $K = 10$. Given a target stable rank r, we set

$$\lambda_1^2 = \frac{\|\mathbf{\Lambda}\|_F^2}{r}, \qquad \lambda_j^2 = \frac{\|\mathbf{\Lambda}\|_F^2 - \lambda_1^2}{K-1}, \quad j = 2, \dots, K.$$

For $K = 5$, the stable rank ranges from 1.5 to 3.5 in increments of 0.5; for $K = 10$, it ranges from 1.5 to 6 with the same increment. Each evaluation consists of 1,000 test queries, and results are averaged over 10 independent runs with different random seeds. We evaluate different degrees of label imbalance using the allocation patterns in Table 4.

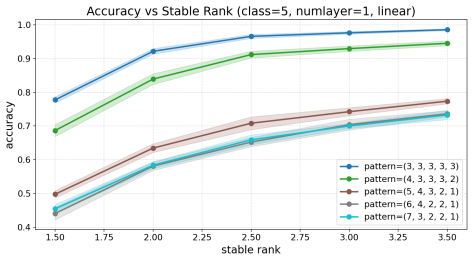
Table 4: Label allocation patterns for different numbers of classes K. Each pattern specifies the number of demonstrations per class in the test prompt.

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]

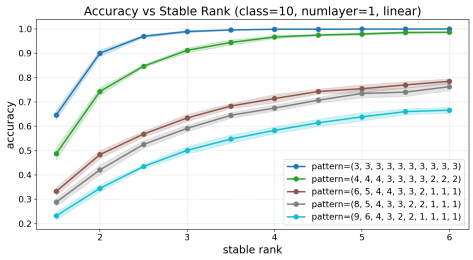
For imbalanced prompts, the query label is sampled from the bottom 3 classes for $K = 5$ and the bottom 7 classes for $K = 10$, ranked by their demonstration counts. For example, under Pattern 3 with $K = 5$, the query label is selected from the classes with 3, 2, or 1 demonstrations.

Figure 2 shows classification accuracy as a function of stable rank for the linear model and nonlinear Transformers. The linear-model results examine the predictor analyzed in our theory, while the nonlinear-model results assess whether similar trends persist in more expressive architectures. Across these models, accuracy generally improves with stable rank, while more severe imbalance reduces performance on minority-class queries. These observations qualitatively support our generalization analysis; the use of finite-step AdamW training does not directly test the asymptotic gradient-descent characterization.

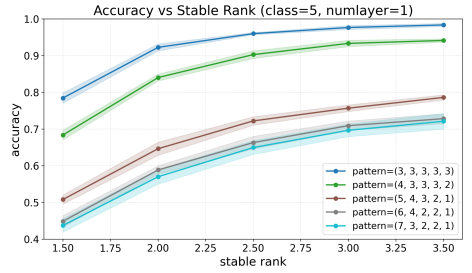
Results



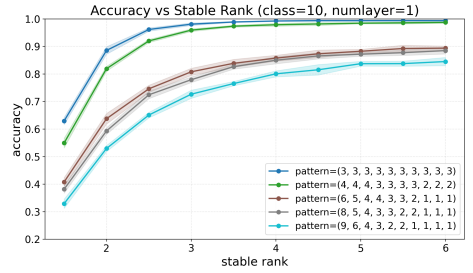
(a) $K = 5$, linear self-attention



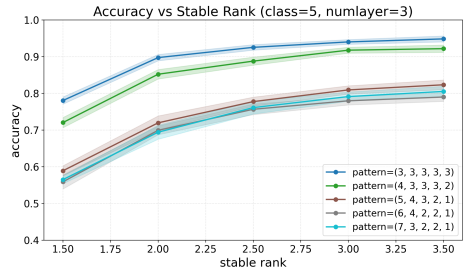
(b) $K = 10$, linear self-attention



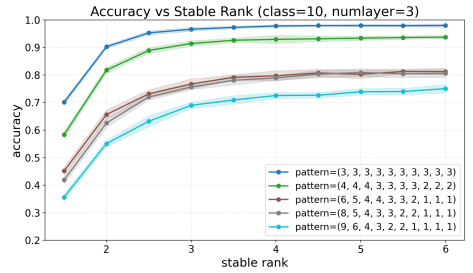
(c) $K = 5$, 1 layer



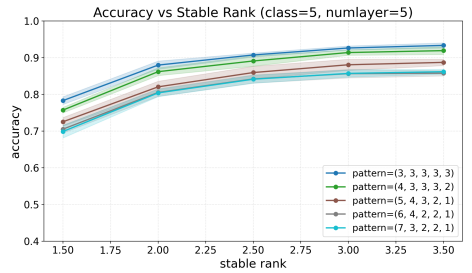
(d) $K = 10$, 1 layer



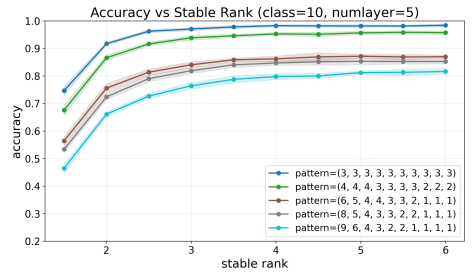
(e) $K = 5$, 3 layers



(f) $K = 10$, 3 layers



(g) $K = 5$, 5 layers



(h) $K = 10$, 5 layers

Figure 2: Classification accuracy versus stable rank under different label allocation patterns. The first row shows linear self-attention; the remaining rows show nonlinear Transformers with 1, 3, and 5 layers. Shaded bands indicate one standard deviation across runs.

G.1.1 Additional discussions on Linear Attention vs (Softmax) Transformer

We examine the learned attention matrices in the $K = 10$ experiments described above. For the one-layer linear attention model, we visualize the learned $\mathbf{W}^{KQ}$ in Figure 3.

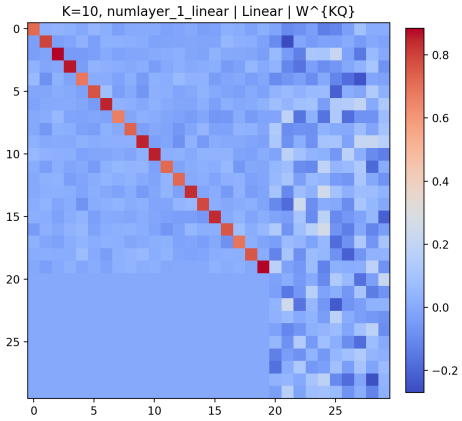


Figure 3: Heatmap of $\mathbf{W}^{KQ}$ for linear attention after training

In linear attention, we empirically observe a concentration of positive entries along the diagonal. Our theoretical analysis establishes a lower bound on the trace-to-Frobenius ratio of the max-margin solution (Section D.2). We observe a qualitatively similar pattern in nonlinear Transformers, where at least one attention head in each layer exhibits predominantly positive diagonal entries (Figure 4). These observations motivate the following conditional analysis, although the heatmaps alone do not establish the corresponding trace-to-Frobenius bound for nonlinear models.

Consider a simplified single-head softmax attention model with the same fixed label readout and context-only attention mask as in our linear model. We absorb the attention temperature into $\boldsymbol{W}$ and take $\boldsymbol{W}$ to be fixed independently of the new test task, with $\mathrm{Tr}(\boldsymbol{W}) > 0$. To isolate the interaction between class geometry and label allocation, we consider the noiseless setting $\boldsymbol{x}^{(j)} = \boldsymbol{\mu}_{k(j)}$ and $\boldsymbol{x}^{(M+1)} = \boldsymbol{\mu}_{k^\star}$. The class-$k$ output is then

$$\begin{aligned}\hat{y}_k &= \frac{1}{M}\frac{\sum_{j\in\mathcal{C}_k}\exp\left(\boldsymbol{x}^{(j)\top}\boldsymbol{W}\boldsymbol{x}^{(M+1)}\right)}{\sum_{j=1}^{M}\exp\left(\boldsymbol{x}^{(j)\top}\boldsymbol{W}\boldsymbol{x}^{(M+1)}\right)} \\ &= \frac{1}{M}\frac{S_k\exp\left(\boldsymbol{\mu}_k^\top\boldsymbol{W}\boldsymbol{\mu}_{k^\star}\right)}{\sum_{\ell=1}^{K}S_\ell\exp\left(\boldsymbol{\mu}_\ell^\top\boldsymbol{W}\boldsymbol{\mu}_{k^\star}\right)}.\end{aligned}$$

Since the denominator is shared across classes, prediction is equivalent to maximizing

$$\boldsymbol{\mu}_k^\top\boldsymbol{W}\boldsymbol{\mu}_{k^\star} + \log S_k.$$

Thus, label imbalance contributes an additive term to the log-score. In particular, the correct class outranks class k if and only if

$$\boldsymbol{\mu}_{k^\star}^\top\boldsymbol{W}\boldsymbol{\mu}_{k^\star} - \boldsymbol{\mu}_k^\top\boldsymbol{W}\boldsymbol{\mu}_{k^\star} > \log\left(\frac{S_k}{S_{k^\star}}\right).$$

This illustrates how a geometric margin can counteract the count advantage of a majority class. For a competing class $k \neq k^\star$, our class-center generation process gives

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

Consequently,

$$\begin{aligned}&\mathbb{P}(\hat{y}_k \geq \hat{y}_{k^\star}) \\ &= \mathbb{P}\left((\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top\boldsymbol{W}\boldsymbol{\mu}_{k^\star} \geq \log\left(\frac{S_{k^\star}}{S_k}\right)\right) \\ &= \mathbb{P}\left((\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top\boldsymbol{W}\boldsymbol{\mu}_{k^\star} - \mathbb{E}\left[(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^\star})^\top\boldsymbol{W}\boldsymbol{\mu}_{k^\star}\right] \geq (1+\phi_{k,k^\star})\frac{\mathrm{Tr}(\boldsymbol{W})}{dK}\|\boldsymbol{\Lambda}\|_F^2\right),\end{aligned}$$

where

$$\phi_{k,k^\star} := \frac{dK}{\mathrm{Tr}(\boldsymbol{W})\|\boldsymbol{\Lambda}\|_F^2} \log\left(\frac{S_{k^\star}}{S_k}\right).$$

For fixed $\boldsymbol{W}$ and $\boldsymbol{\Lambda}$, a smaller $\phi_{k,k^\star}$ corresponds to a larger count advantage for the competing class. We assume $\phi_{k,k^\star} > -1$, ensuring that the threshold lies above the mean and that an upper-tail concentration bound is informative. Applying the same class-center concentration argument gives

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

Motivated by the property established for the linear model, suppose additionally that

$$\frac{\mathrm{Tr}(\boldsymbol{W})}{\|\boldsymbol{W}\|_F} \geq c_0\frac{\sqrt{d}}{\xi}$$

for an absolute constant $c_0 > 0$. Then, after adjusting absolute constants, the preceding bound implies

$$\begin{aligned}&\mathbb{P}(\hat{y}_k \geq \hat{y}_{k^\star}) \\ &\quad\leq C\exp\left(-c\left(\frac{d(1+\phi_{k,k^\star})\,\mathrm{sr}(\boldsymbol{\Lambda})}{\xi^2 K} \wedge \frac{\sqrt{d}}{\xi} \wedge (1+\phi_{k,k^\star})\,\mathrm{sr}(\boldsymbol{\Lambda})\right)\frac{(1+\phi_{k,k^\star})\,\mathrm{sr}(\boldsymbol{\Lambda})}{K}\right).\end{aligned}$$

This conditional bound shows how the spectrum of the class-center matrix and the number of examples from each class together affect classification in the noiseless softmax model. With the other quantities fixed, a higher stable rank gives a tighter error bound and can offset a smaller imbalance factor. This may help explain the trends we observe in our nonlinear experiments. Two questions remain open: whether softmax training yields the assumed trace-to-Frobenius bound and whether the analysis extends to Transformers with noisy inputs, multiple heads, and multiple layers.

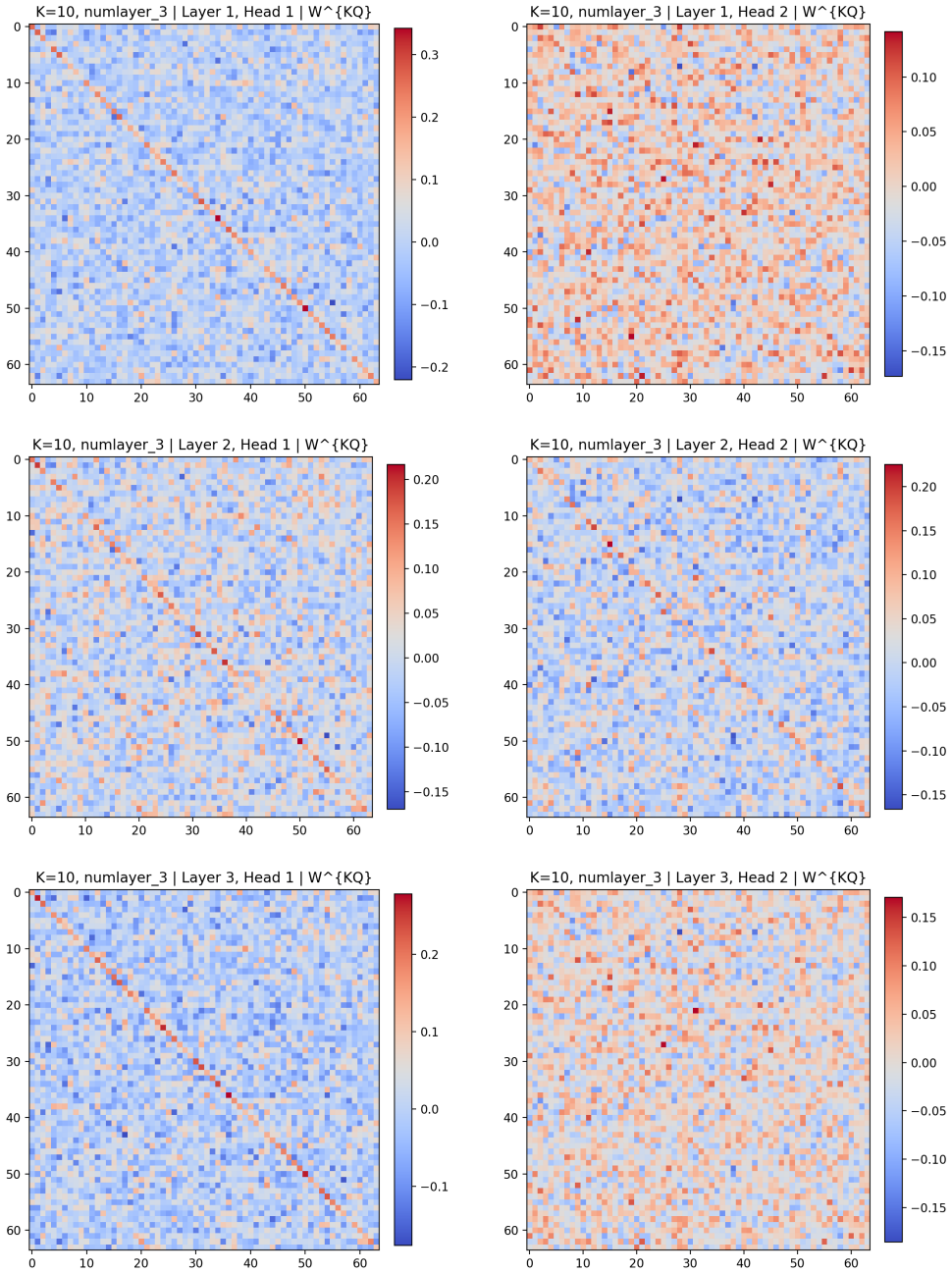


Figure 4: Heatmaps of $\mathbf{W}^{KQ}$ for each layer and attention head. Each row corresponds to a layer, and each column corresponds to a head. These heatmaps suggest that a similar pattern also emerges in softmax attention, beyond the linear attention setting considered in our analysis.

G.1.2 Controlling Minimum Class Separation

Prior work on multiclass GMMs identifies the minimum pairwise class-center distance as a key quantity governing recovery performance [Giraud and Verzelen, 2019]. In our original experiments in Appendix G.1, we vary the stable rank of the class-center matrix while fixing its Frobenius norm. However, this does not necessarily control the minimum pairwise distance, raising the question of whether changes in minimum separation alone can explain the observed accuracy trend. To address this question, we conduct an additional experiment in which every task is normalized to have the same minimum pairwise class-center distance while retaining its target stable rank. This allows us to examine whether the relationship between stable rank and accuracy persists after controlling for minimum separation.

Data generation with fixed minimum distance. Following Appendix G.1, we construct a singular-value matrix $\boldsymbol{\Lambda}_r$ for each target stable rank $r \in (1, K]$. For a fixed initial Frobenius norm $F_0 > 0$, its singular values satisfy

$$\left(\lambda_1^{(r)}\right)^2 = \frac{F_0^2}{r}, \qquad \left(\lambda_k^{(r)}\right)^2 = \frac{F_0^2(r-1)}{r(K-1)}, \qquad k = 2, \dots, K.$$

For each task τ, we independently sample Haar-distributed orthogonal matrices $\boldsymbol{U}_\tau$ and $\boldsymbol{V}_\tau$ and form

$$\boldsymbol{M}_{\tau,r}^{(0)} = \boldsymbol{U}_\tau \boldsymbol{\Lambda}_r \boldsymbol{V}_\tau^\top.$$

Let $(\boldsymbol{\mu}_{\tau,r,k}^{(0)})^\top$ denote the k-th row of $\boldsymbol{M}_{\tau,r}^{(0)}$. We define

$$\Delta_{\tau,r}^{(0)} := \min_{k \neq \ell} \left\| \boldsymbol{\mu}_{\tau,r,k}^{(0)} - \boldsymbol{\mu}_{\tau,r,\ell}^{(0)} \right\|_2.$$

We then rescale each task separately to a common target distance $\Delta_0 > 0$:

$$c_{\tau,r} := \frac{\Delta_0}{\Delta_{\tau,r}^{(0)}}, \qquad \widetilde{\boldsymbol{M}}_{\tau,r} := c_{\tau,r} \boldsymbol{M}_{\tau,r}^{(0)}.$$

This ensures that every task has the same minimum pairwise class-center distance while preserving its target stable rank:

$$\min_{k \neq \ell} \|\widetilde{\boldsymbol{\mu}}_{\tau,r,k} - \widetilde{\boldsymbol{\mu}}_{\tau,r,\ell}\|_2 = \Delta_0, \quad \mathrm{sr}(\widetilde{\boldsymbol{M}}_{\tau,r}) = \mathrm{sr}(\boldsymbol{M}_{\tau,r}^{(0)}) = r.$$

Finally, inputs are generated as in Appendix G.1:

$$\boldsymbol{x}_\tau^{(j)} = \widetilde{\boldsymbol{M}}_{\tau,r}^\top \boldsymbol{y}_\tau^{(j)} + \boldsymbol{z}_\tau^{(j)}, \qquad \boldsymbol{z}_\tau^{(j)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d).$$

Thus, the minimum class separation and noise distribution are held fixed across stable-rank settings. The Frobenius norm after normalization is allowed to vary.

Results. We evaluate a one-layer linear self-attention model and softmax Transformers with $L \in \{1, 3, 5\}$ layers for $K \in \{5, 10\}$. For each class count, we consider five label-allocation patterns ranging from balanced to strongly imbalanced, keeping the total number of demonstrations fixed at 15 for $K = 5$ and 30 for $K = 10$. For each class count K, we use a common target minimum distance Δ_0 across all tasks and stable-rank settings. Specifically, $\Delta_0^2 = 48$ for $K = 5$ and $\Delta_0^2 = 96$ for $K = 10$. These values satisfy $\Delta_0^2 = 2F_0^2/K$, where the squared Frobenius norm before normalization is $F_0^2 = 120$ and $F_0^2 = 480$, respectively. Each configuration is evaluated on 1,000 examples per seed, and the results are averaged over 10 random seeds. Figures 5 and 6 present the results, with each legend specifying the number of demonstrations from each class. Lines show mean accuracy over 10 random seeds, and shaded regions indicate one standard deviation.

The results exhibit the same qualitative trend as our original experiments in Appendix G.1: accuracy generally improves with stable rank under both balanced and imbalanced label allocations. This trend is observed in both linear self-attention and softmax Transformers, even when the minimum pairwise class-center distance is held fixed.

Minimum pairwise class separation plays a central role in classical error bounds for multiclass GMMs. When a Transformer solves this classification problem in context, it uses a matrix $\boldsymbol{W}$ learned across

training tasks to compare the query with demonstrations from a new task. In the linear attention model we analyze, aggregating the bilinear terms $\boldsymbol{x}_i^\top \boldsymbol{W} \boldsymbol{x}_q$ by class yields the score $\hat{\boldsymbol{\mu}}_k^\top \boldsymbol{W} \boldsymbol{x}_q$, where $\hat{\boldsymbol{\mu}}_k$ is computed from the **current prompt**. Our concentration bound for the difference between two such scores depends on the spectrum of the class-center matrix through quantities such as stable rank. In our additional experiments, accuracy generally improves with stable rank for both linear attention and softmax Transformers even when the minimum pairwise class-center distance is held fixed. This suggests that spectral geometry helps explain in-context classification performance beyond what minimum separation alone can capture.

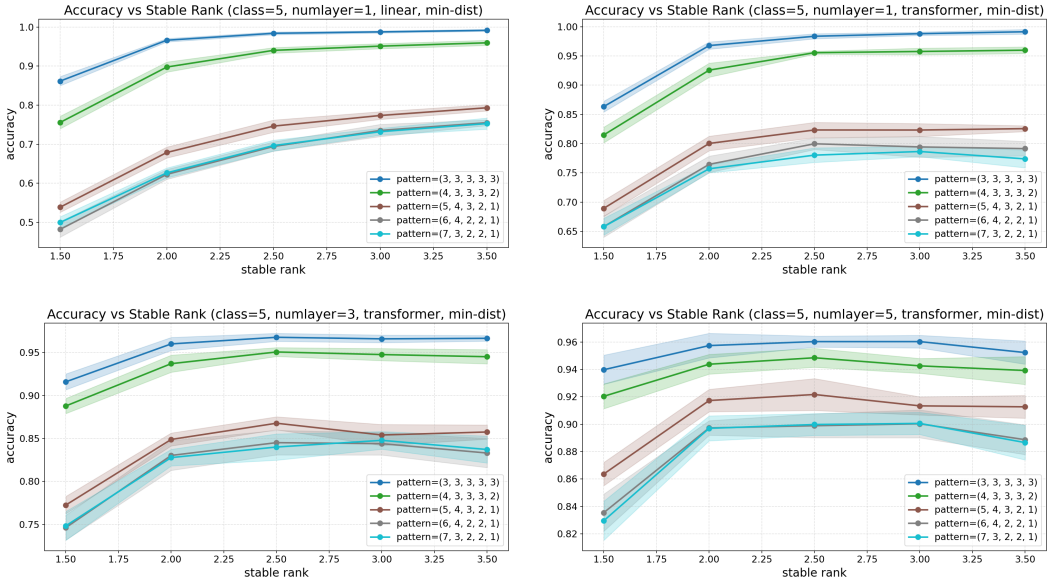


Figure 5: Accuracy versus stable rank with fixed minimum pairwise class-center distance for $K = 5$. The panels show one-layer linear self-attention (top left) and softmax Transformers with one (top right), three (bottom left), and five (bottom right) layers.

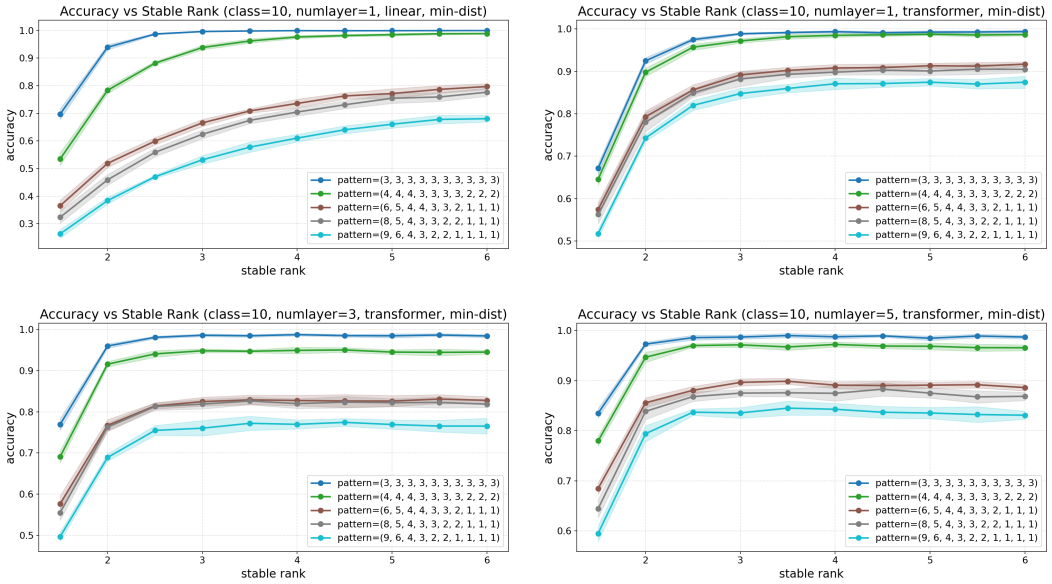


Figure 6: Accuracy versus stable rank with fixed minimum pairwise class-center distance for $K = 10$. The panels show one-layer linear self-attention (top left) and softmax Transformers with one (top right), three (bottom left), and five (bottom right) layers.

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

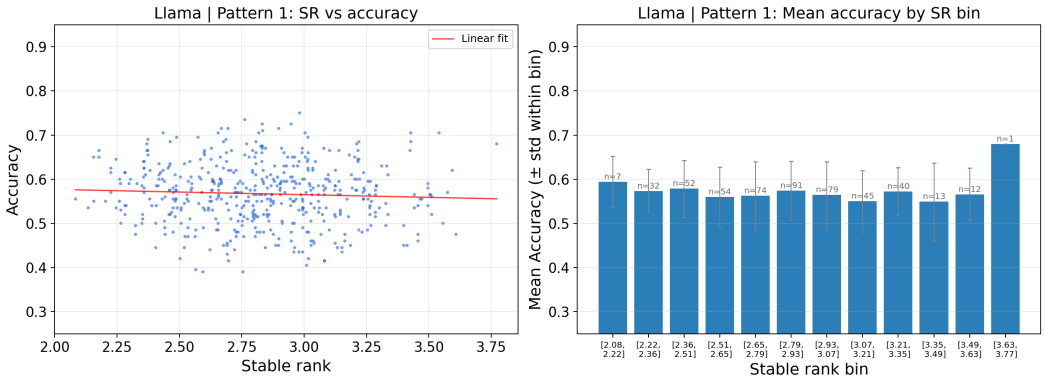
We further conduct experiments on real-world natural language datasets. We construct a multiclass classification task based on WordNet [Miller, 1995], defining a word-to-label mapping via hypernyms (e.g., lion $\rightarrow$ animal). Since WordNet is organized as a tree, we collect words by restricting 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 $\rightarrow$ label) demonstrations followed by a query (word $\rightarrow$), with disjoint query sets and shuffled ordering in order to mitigate positional biases such as recency effects.

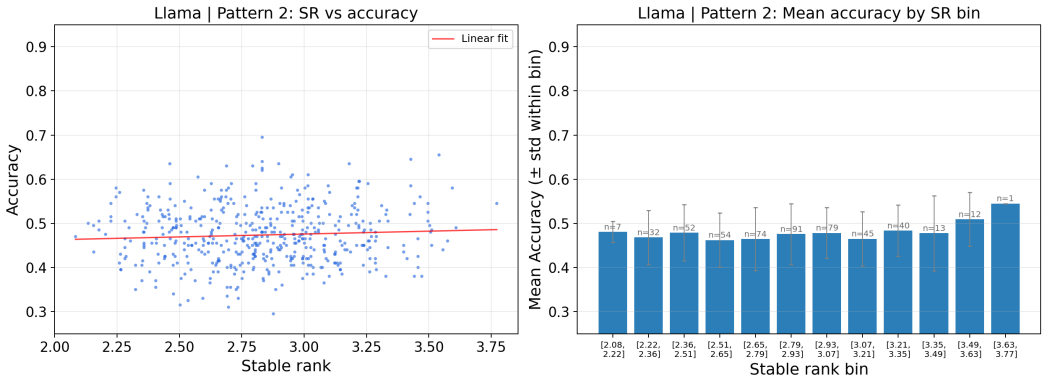
For each prompt, we sample a fixed number of demonstrations per class according to predefined label allocation patterns: Pattern 1 (balanced) [5, 5, 5, 5, 5, 5, 5], Pattern 2 (imbalanced) [9, 8, 6, 5, 4, 2, 1], and Pattern 3 (imbalanced) [12, 9, 6, 4, 2, 1, 1]. In addition, even within the same class subset, we reshuffle the label-to-class assignments for each query, resulting in diverse prompt configurations. To evaluate the imbalanced-label scenario, the query label is preferentially selected from the bottom-5 classes (i.e., those with the fewest demonstrations) in the current prompt. For example, under Pattern 3, we allocate demonstrations 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.

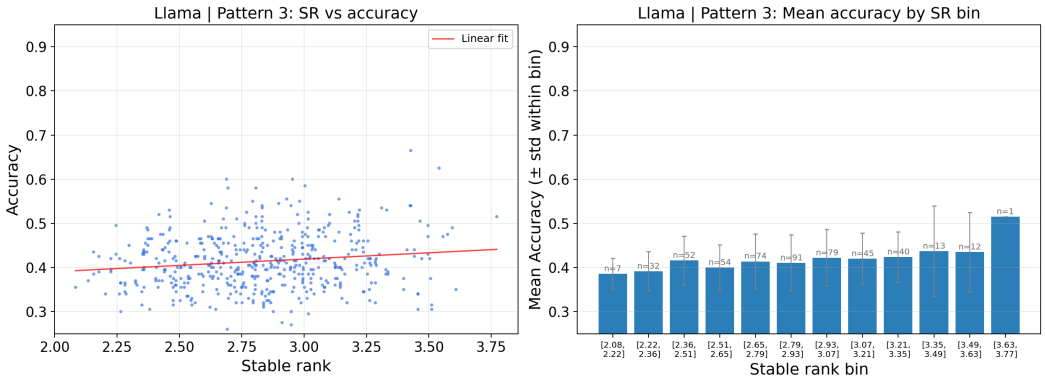
Results on Llama



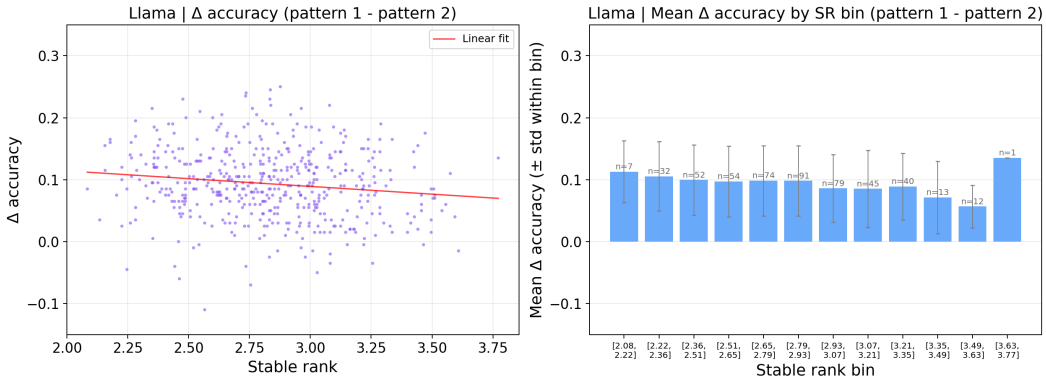
(a) Llama Pattern 1



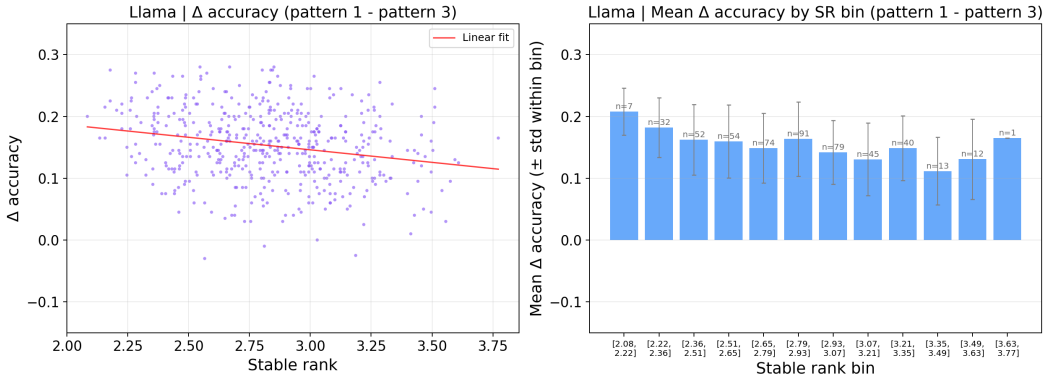
(b) Llama Pattern 2



(c) Llama Pattern 3

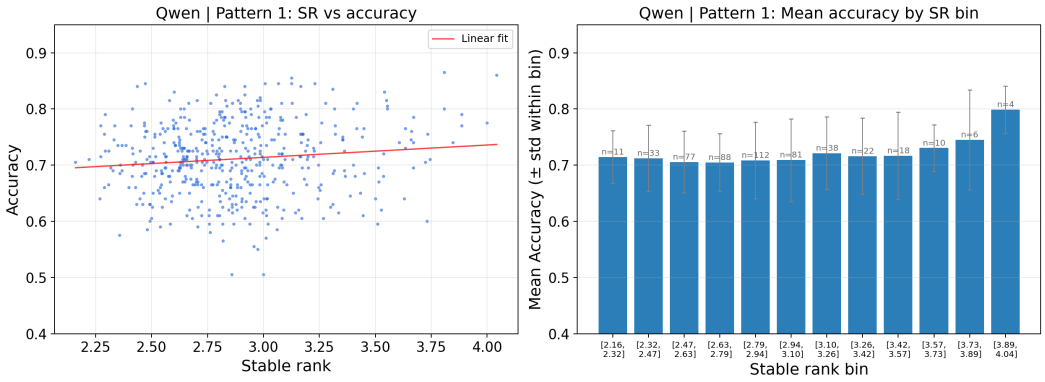


(a) Llama Difference between Patterns 1 and 2

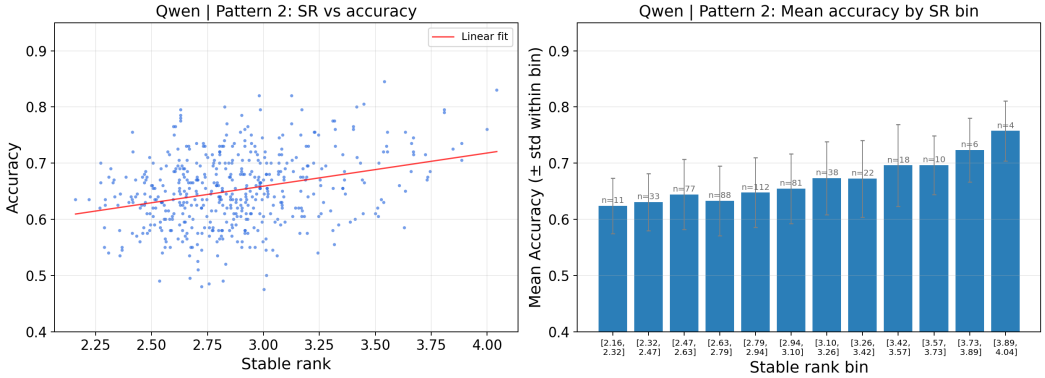


(b) Llama Difference between Patterns 1 and 3

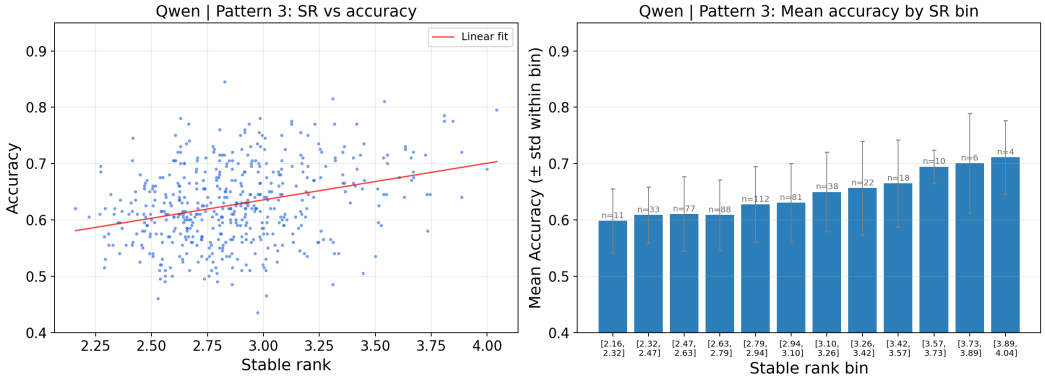
Results on Qwen



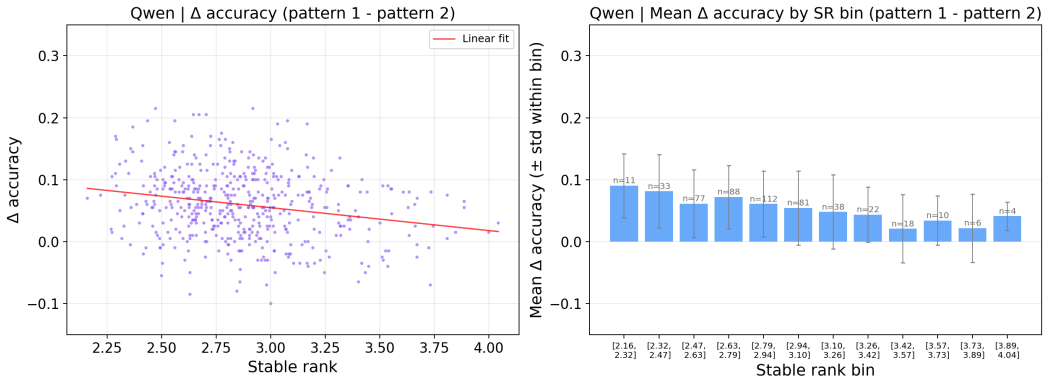
(a) Qwen Pattern 1



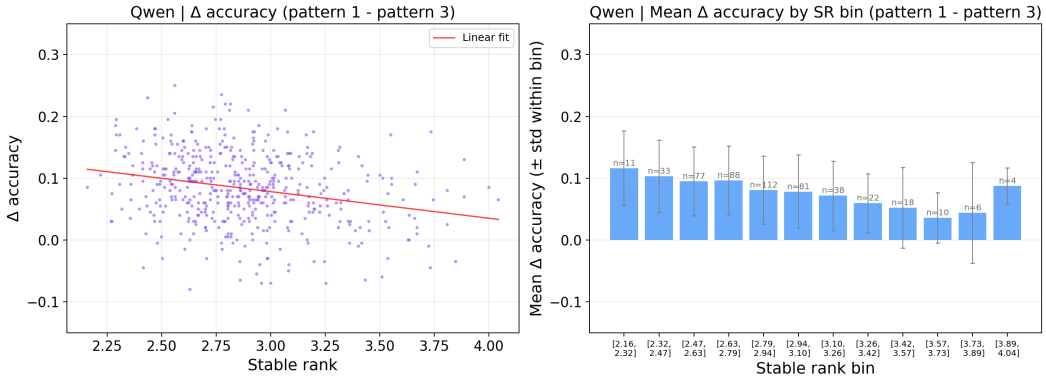
(b) Qwen Pattern 2



(c) Qwen Pattern 3

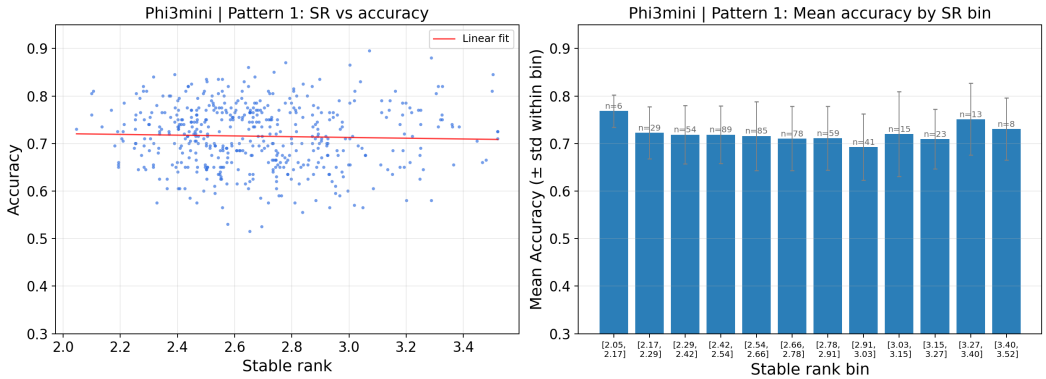


(a) Qwen Difference between Patterns 1 and 2

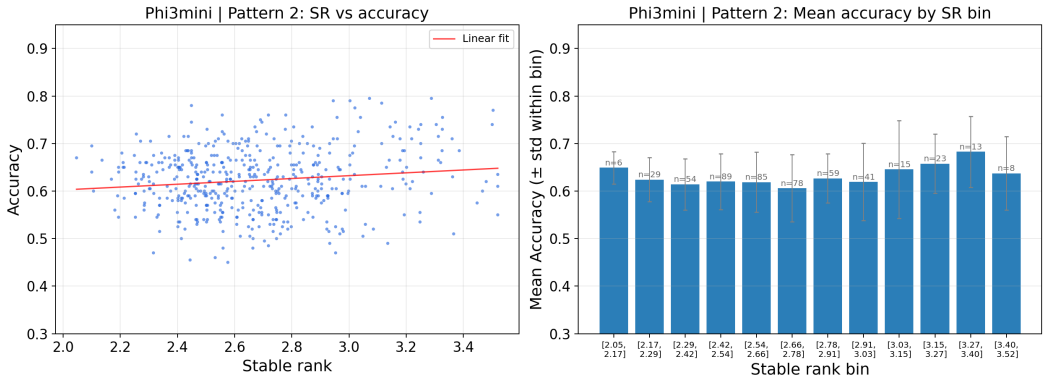


(b) Qwen Difference between Patterns 1 and 3

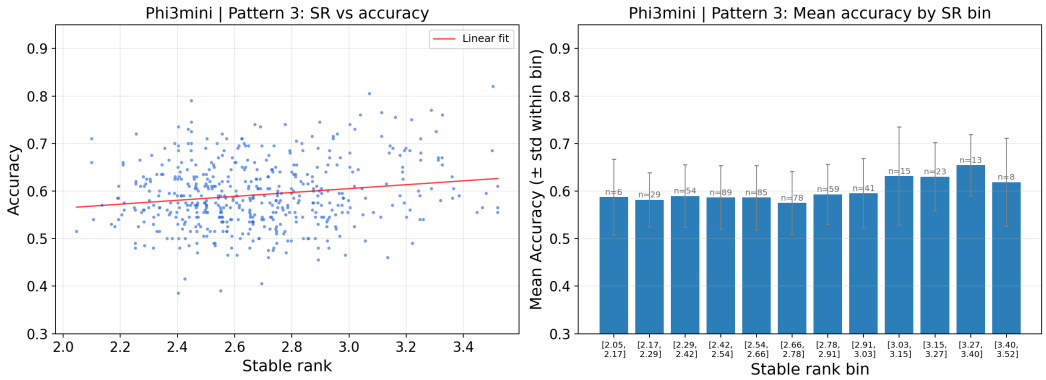
Results on Phi-3



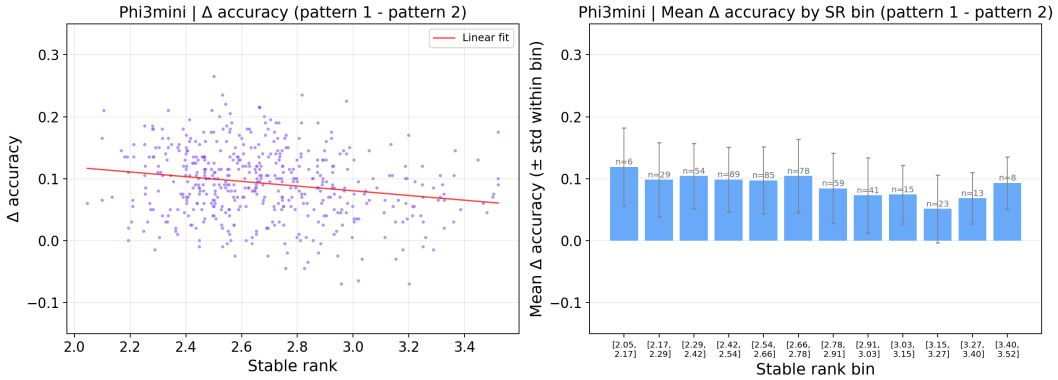
(a) Phi-3 Pattern 1



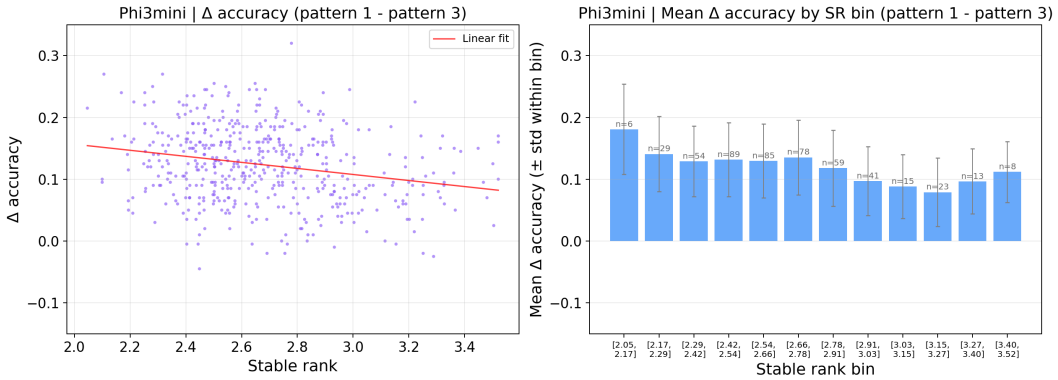
(b) Phi-3 Pattern 2



(c) Phi-3 Pattern 3



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



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

G.2.1 Comparison with Other Geometric Quantities

We repeat the WordNet experiment with an independent random seed to compare stable rank with three other measures of class-center geometry: mean pairwise distance, mean angular separation, and entropy-based effective dimension. We use `Llama-3.2-1B-Instruct` and `Qwen2.5-3B-Instruct` and follow the evaluation protocol described above. All four measures are evaluated within the same run for each model.

Geometric quantities. For each subset of $K = 7$ classes, let $\boldsymbol{m}_k \in \mathbb{R}^d$ denote the average input embedding of the WordNet words belonging to class k. We stack the class centers as rows of $\boldsymbol{M} \in \mathbb{R}^{K\times d}$ and center them across classes:

$$\bar{\boldsymbol{m}} := \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{m}_k, \qquad \boldsymbol{X} := \boldsymbol{M} - \mathbf{1}\bar{\boldsymbol{m}}^\top.$$

Let $\boldsymbol{x}_i^\top$ denote the i-th row of $\boldsymbol{X}$, and let $\sigma_1 \geq \cdots \geq \sigma_r > 0$ denote its nonzero singular values. We compute the stable rank, mean pairwise distance, and mean angular separation as

$$\begin{aligned}
\mathrm{sr}(\boldsymbol{X}) &:= \frac{\|\boldsymbol{X}\|_F^2}{\sigma_1^2}, \\
\bar{d} &:= \frac{2}{K(K-1)}\sum_{i<j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2, \\
\bar{\theta} &:= \frac{2}{K(K-1)}\sum_{i<j} \arccos\left(\frac{\boldsymbol{x}_i^\top \boldsymbol{x}_j}{\|\boldsymbol{x}_i\|_2\|\boldsymbol{x}_j\|_2}\right).
\end{aligned}$$

To quantify how evenly the squared singular values are distributed, we also compute the spectral entropy $H(\boldsymbol{X})$ and the corresponding effective dimension $d_{\mathrm{eff}}(\boldsymbol{X})$:

$$p_i := \frac{\sigma_i^2}{\sum_{j=1}^r \sigma_j^2}, \qquad H(\boldsymbol{X}) := -\sum_{i=1}^{r} p_i \log p_i, \qquad d_{\mathrm{eff}}(\boldsymbol{X}) := \exp\big(H(\boldsymbol{X})\big).$$

Since exponentiation preserves the ordering of class subsets, spectral entropy and effective dimension give identical Spearman rank correlations.

Accuracy and imbalance gaps. For each class subset, let Acc_j denote classification accuracy under Pattern j. Pattern 1 uses balanced prompts, while Patterns 2 and 3 use imbalanced prompts. We define the accuracy gaps as

$$\Delta_{1,2} := \mathrm{Acc}_1 - \mathrm{Acc}_2, \qquad \Delta_{1,3} := \mathrm{Acc}_1 - \mathrm{Acc}_3\,.$$

For each model, we compute the Spearman rank correlation across class subsets between each geometric quantity and each of the five outcomes. A negative correlation with $\Delta_{1,j}$ indicates that subsets with larger values of the geometric quantity tend to have a smaller accuracy drop from Pattern 1 to Pattern j.

Classification accuracy and class-center magnitude. Mean pairwise distance has the largest observed correlation with accuracy for both models under all three label allocations, with Spearman coefficients ranging from 0.457 to 0.556 (Table 5). This finding is consistent with the role of class-center magnitude in our theoretical bound, which depends explicitly on $\|\boldsymbol{\Lambda}\|_F^2$. The connection is exact for the mean *squared* pairwise distance of the empirical class centers. Since $\sum_{i=1}^K \boldsymbol{x}_i = \mathbf{0}$,

$$\frac{2}{K(K-1)}\sum_{i<j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 = \frac{2}{K-1}\|\boldsymbol{X}\|_F^2.$$

The squared Frobenius norm is proportional to the mean squared pairwise distance, while Table 5 reports the mean pairwise distance. Although squaring gives more weight to larger distances, both measures tend to increase as the class centers become more spread out.

Table 5: Spearman rank correlations across WordNet class subsets in the additional run with an independent random seed. The first three outcomes are accuracies under individual label allocations, and the last two are balanced-to-imbalanced accuracy gaps. Parentheses give unadjusted p-values for tests of zero Spearman correlation. Bold entries indicate $p < 0.05$.

Outcome	Stable rank	Mean distance	Mean angle	Effective dimension
`Llama-3.2-1B-Instruct`				
Acc_1	$-0.030\,(0.496)$	$\mathbf{0.457\,(< 0.0001)}$	$\mathbf{0.274\,(< 0.0001)}$	$\mathbf{0.114\,(0.011)}$
Acc_2	$0.022\,(0.619)$	$\mathbf{0.533\,(< 0.0001)}$	$\mathbf{0.281\,(< 0.0001)}$	$\mathbf{0.225\,(< 0.0001)}$
Acc_3	$0.072\,(0.108)$	$\mathbf{0.509\,(< 0.0001)}$	$\mathbf{0.289\,(< 0.0001)}$	$\mathbf{0.219\,(< 0.0001)}$
$\Delta_{1,2}$	$-0.079\,(0.077)$	$\mathbf{-0.117\,(0.009)}$	$-0.032\,(0.479)$	$\mathbf{-0.161\,(0.0003)}$
$\Delta_{1,3}$	$\mathbf{-0.098\,(0.028)}$	$-0.074\,(0.098)$	$-0.012\,(0.783)$	$\mathbf{-0.131\,(0.003)}$
`Qwen2.5-3B-Instruct`				
Acc_1	$0.061\,(0.173)$	$\mathbf{0.508\,(< 0.0001)}$	$\mathbf{0.125\,(0.005)}$	$\mathbf{0.156\,(0.0004)}$
Acc_2	$\mathbf{0.278\,(< 0.0001)}$	$\mathbf{0.556\,(< 0.0001)}$	$\mathbf{0.111\,(0.013)}$	$\mathbf{0.400\,(< 0.0001)}$
Acc_3	$\mathbf{0.250\,(< 0.0001)}$	$\mathbf{0.555\,(< 0.0001)}$	$\mathbf{0.103\,(0.021)}$	$\mathbf{0.380\,(< 0.0001)}$
$\Delta_{1,2}$	$\mathbf{-0.257\,(< 0.0001)}$	$-0.065\,(0.146)$	$0.020\,(0.659)$	$\mathbf{-0.276\,(< 0.0001)}$
$\Delta_{1,3}$	$\mathbf{-0.226\,(< 0.0001)}$	$-0.072\,(0.105)$	$0.029\,(0.519)$	$\mathbf{-0.268\,(< 0.0001)}$

Accuracy gaps and spectral geometry. Effective dimension has the largest absolute correlation with the accuracy gap in all four comparisons across the two models and two imbalanced patterns. Its coefficients range from -0.276 to -0.131, indicating that class subsets with more evenly distributed spectral energy tend to have smaller accuracy drops. Stable rank also correlates negatively with both gaps for both models. Its correlations are stronger for Qwen (-0.257 and -0.226) than for Llama (-0.079 and -0.098), and exceed those of mean pairwise distance in absolute value in three of the four comparisons.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contributions of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitation of our results in Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide our assumptions in Section 2, Appendix B. In addition we provide complete proof of our theoretical findings in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide details of our experiments in Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our experiments use synthetic data, the publicly available WordNet dataset, and open-source models. We provide full experimental details, ensuring full reproducibility.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide experimental setting and details in Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We present 1-sigma error bars for the results in Appendices G.1 and G.2, and report the results of statistical significance tests in Section 5.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide details on the computational resources in Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper does not violate the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is theoretical.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper is theoretical.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: Our experiments use publicly available datasets and open-source models.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper is theoretical and does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper is theoretical and does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: We used LLMs only for writing and editing.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.