ADVERSARIALLY TRAINED LINEAR TRANSFORMERS ARE OPTIMAL ROBUST IN-CONTEXT LEARNERS FOR GAUSSIAN MIXTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial training is one of the most reliable defenses against adversarial attacks, but its high computational cost must generally be paid anew for each task. Robust foundation models offer a promising alternative: adversarially pretrain a model once and then transfer its robustness to downstream tasks through lightweight adaptation. However, a fundamental question remains open: can robustness acquired during pretraining transfer to unseen tasks without further adversarial training? In this study, we answer this question affirmatively. A single model adversarially pretrained at scale can achieve optimal robustness on new tasks without additional task-specific training. Specifically, we show that, for a family of Gaussian-mixture classification tasks, a sufficiently deep linear transformer adversarially trained across tasks can asymptotically attain the robust Bayes error on previously unseen tasks through in-context learning from clean demonstrations. By contrast, a standardly trained model cannot. We further analyze convergence under gradient flow, an accuracy–robustness trade-off, and demonstration complexity.

1 INTRODUCTION

Adversarial examples, inputs that are slightly perturbed to cause machine learning models to make incorrect predictions, expose a fundamental vulnerability of deep learning systems (Szegedy et al., 2014). Adversarial training is one of the most reliable defenses against adversarial attacks (Goodfellow et al., 2015; Madry et al., 2018). It improves robustness by minimizing the loss under worst-case perturbations. However, the resulting min–max optimization makes adversarial training substantially more expensive than standard training. While many cheaper defenses have been proposed, some offer only apparent robustness that disappears under stronger attacks (Carlini & Wagner, 2017a;b; Athalye et al., 2018; Croce & Hein, 2020; Tramer et al., 2020). Consequently, obtaining reliable robustness generally requires paying the cost of adversarial training anew for each task.

Meanwhile, foundation models have transformed how models are developed and deployed. Rather than training a separate model for each task, practitioners pretrain a single model at scale and adapt it to downstream tasks using lightweight methods. This paradigm suggests a compelling possibility: robust foundation models. Adversarial training would be performed only once during pretraining, and the resulting robustness could then be transferred to downstream tasks at little additional cost. Although adversarial pretraining incurs an initial cost, model developers (typically large organizations) could recoup it by licensing the model. The fundamental question here is: *Can an adversarially pretrained model adapt to unseen downstream tasks while retaining robustness, without further adversarial training?*

Kumano et al. (2026) provided the first theoretical evidence that such transfer might be possible. They showed that an adversarially pretrained single-layer transformer with linear attention can adapt robustly to unseen classification tasks through in-context learning. In-context learning is a particularly lightweight adaptation method that enables transformers to learn from demonstrations in the prompt without updating their parameters (Brown et al., 2020). This result suggests that robustness acquired during pretraining can transfer to new tasks without additional training. However, their analysis is limited to a restricted data model and a single-layer architecture. Moreover, they provided only

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

bounds on the expected robust margin. Thus, it does not provide direct guarantees on robust accuracy or evaluate robust optimality. Consequently, our central question remains open in broader settings.

In this study, we provide a stronger affirmative answer to this question under substantially more general conditions. We consider multilayer linear transformers adversarially pretrained on a family of Gaussian-mixture classification tasks. We focus on the robust zero-one error of these transformers on tasks not encountered during pretraining. In particular, we investigate whether this error approaches the robust Bayes error, which is the minimum robust error attainable by any classifier, even one with full knowledge of the task distribution.

Specifically, we show that, through in-context learning from clean demonstrations, a sufficiently deep linear transformer can asymptotically attain the robust Bayes error on every task in this Gaussian family, including tasks unseen during pretraining. In other words, using only standard samples and without task-specific adversarial training or access to adversarial examples, a single linear transformer can asymptotically serve as an optimal robust predictor for each task. We should note that the transformer's parameters remain fixed and shared across all downstream tasks. Nevertheless, the transformer can adapt to each task and implement its optimal robust prediction rule.

Moreover, we show that a standardly pretrained transformer cannot achieve optimal robustness, even with infinite pretraining tasks and in-context demonstrations. This result indicates that optimal robust in-context learning cannot be attributed solely to model capacity or task diversity; rather, it stems from an inherent advantage of adversarial pretraining.

We further investigate several properties of robust in-context learning. Whereas the discussion above focuses on empirical-risk minimizers, we also prove that gradient flow in a two-layer architecture converges to a predictor that attains robust Bayes optimality. We also identify two fundamental challenges in developing robust transformers. The first is an inherent accuracy–robustness trade-off: adversarial pretraining can degrade clean accuracy even in the infinite-data limit. The second is greater demonstration complexity: even at the same Bayes excess risk, robust prediction requires more demonstrations than standard prediction.

In summary, we prove the following:

- A sufficiently deep linear transformer asymptotically achieves optimal robustness across a family of Gaussian-mixture classification tasks through in-context learning from clean demonstrations.
- Adversarial pretraining makes the transformer an optimal robust in-context learner, whereas standard pretraining does not, even with unlimited data.
- Under mild conditions, (i) gradient flow leads the transformer to become such an in-context learner, (ii) an accuracy–robustness trade-off arises, and (iii) robust in-context prediction requires more demonstrations than standard prediction.

2 Related Work

Adversarial training. Numerous defenses against adversarial examples (Szegedy et al., 2014) have been proposed, but many have subsequently been circumvented by stronger attacks (Carlini & Wagner, 2017a;b; Athalye et al., 2018; Croce & Hein, 2020; Tramer et al., 2020). Although adversarial training (Goodfellow et al., 2015; Madry et al., 2018) is one of the most reliable defenses against a wide range of strong attacks, its computational cost is high due to multiple additional backpropagation steps beyond those needed for standard training. Prior work has sought to reduce this cost by approximating multi-step backpropagation (Shafahi et al., 2019; Zhang et al., 2019; Zheng et al., 2020), reducing the number of backpropagation steps (Andriushchenko & Flammarion, 2020; Wong et al., 2020; Kim et al., 2021; Park & Lee, 2021; Jia et al., 2022), or adversarially finetuning models after standard training (Jeddi et al., 2021; Hou et al., 2022; Suzuki et al., 2023; Gowda et al., 2024). Nevertheless, all these approaches still require adversarial training to be performed separately for each task. In contrast, we adversarially pretrain a linear transformer once and adapt it to downstream tasks through in-context learning from clean demonstrations. This approach requires neither new adversarial examples nor adversarial training during downstream-task adaptation.

In-context learning. In-context learning enables a transformer to learn a task from demonstrations in the prompt, without updating its parameters (Brown et al., 2020). Existing theoretical work on

in-context learning has largely focused on the algorithms that transformers can implement in context. For example, appropriately parameterized transformers can implement one gradient descent step per layer in context, so that propagation through multiple layers corresponds to multiple gradient descent steps (Ahn et al., 2023; Bai et al., 2023; Cheng et al., 2024; Gatmiry et al., 2024; Mahankali et al., 2024; Zhang et al., 2024). In-context learning can also simulate second-order optimization (Fu et al., 2024; Giannou et al., 2024), reinforcement learning algorithms (Lee et al., 2023; Lin et al., 2024), and Bayesian inference (Xie et al., 2022; Zhang et al., 2025). Unlike these studies, we do not seek to identify the algorithms that in-context learning can simulate. Instead, we study the ability of in-context learning to transfer robustness acquired in adversarial pretraining to downstream tasks.

Adversarial robustness of in-context learning. The adversarial robustness of in-context learning remains relatively underexplored from a theoretical perspective. Anwar et al. (2025) studied context hijacking and showed that modifying in-context tokens can cause a transformer to output an arbitrary value for a clean query. Li et al. (2025) showed that deeper transformers are more robust to context hijacking. Fu et al. (2025) showed that adversarial training with short adversarial suffixes in the context and clean queries can improve robustness to longer adversarial suffixes at test time. Fu & Wang (2026) used the theory of in-context learning to analyze continuous adversarial training (Xhonneux et al., 2024). Both our work and that of Kumano et al. (2026) differ from these studies in terms of problem setting and objective. Specifically, we ask whether a transformer can learn in context from clean demonstrations to make robust predictions on adversarially perturbed queries. Following our theoretical results, we provide a detailed comparison with Kumano et al. (2026) in Section 3.5.

3 Theoretical Results

Notation. For each $n \in \mathbb{N}$, define $[n] := \{1, \dots, n\}$. We denote the i-th element of a vector $\boldsymbol{a}$ by a_i and the entry in row i and column j of a matrix $\boldsymbol{A}$ by $A_{i,j}$. For $m, n \in \mathbb{N}$, let $\mathbf{0}_n$, $\mathbf{0}_{m,n}$, and $\boldsymbol{I}_n$ denote the n-dimensional zero vector, the $m \times n$ zero matrix, and the $n \times n$ identity matrix, respectively. For $d \in \mathbb{N}$, define the real orthogonal group in dimension d by $\mathcal{O}(d) := \{\boldsymbol{R} \in \mathbb{R}^{d \times d} : \boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}_d\}$. We write $\| \cdot \|_2$ for the Euclidean norm and $\| \cdot \|_{\mathrm{F}}$ for the Frobenius norm. Let $(X_{\boldsymbol{n}})_{\boldsymbol{n} \in \mathbb{N}^m}$ be nonnegative random variables and $(a_{\boldsymbol{n}})_{\boldsymbol{n} \in \mathbb{N}^m}$ be positive deterministic values; we write $X_{\boldsymbol{n}} \lesssim_{\mathbb{P}} a_{\boldsymbol{n}}$ if, for every $\delta \in (0, 1)$, there exists $c_\delta > 0$ such that $\mathbb{P}(X_{\boldsymbol{n}} \le c_\delta a_{\boldsymbol{n}}) \ge 1 - \delta$ for all $\boldsymbol{n} \in \mathbb{N}^m$.

3.1 Problem Setting

Overview. We consider pretraining an L-layer linear transformer f on T tasks. For each $t \in [T]$, the t-th task consists of N in-context demonstrations $\mathcal{D}_{t,N} := \{(\boldsymbol{x}_{t,n}, y_{t,n})\}_{n=1}^N$ and a test pair $(\boldsymbol{x}_{t,N+1}, y_{t,N+1})$. Given the clean demonstrations $\mathcal{D}_{t,N}$, the transformer uses in-context learning to predict the binary label of the adversarially perturbed query $\boldsymbol{x}_{t,N+1} + \boldsymbol{\Delta}$. The adversary chooses $\boldsymbol{\Delta}$ subject to $\|\boldsymbol{\Delta}\|_2 \le \epsilon\sqrt{d}$. We train the transformer parameters $\boldsymbol{\Theta}_L$ by minimizing the empirical risk $\mathcal{R}_{\epsilon,T,N}(\boldsymbol{\Theta}_L)$. We then evaluate the resulting predictor on a freshly sampled test task using the taskwise zero-one error $\mathcal{E}_{\epsilon,\boldsymbol{U}}(f(\,\cdot\,; \boldsymbol{\Theta}_L, \mathcal{D}_N))$, where $\boldsymbol{U}$ denotes the task-specific variable. Our main interest is the excess error $\mathcal{E}_{\epsilon,\boldsymbol{U}}(f(\,\cdot\,; \boldsymbol{\Theta}_L, \mathcal{D}_N)) - \mathcal{E}^\star_{\epsilon,\boldsymbol{U}}$, where $\mathcal{E}^\star_{\epsilon,\boldsymbol{U}}$ denotes the corresponding taskwise Bayes error.

Task structure. Let $T, N \in \mathbb{N}$ denote the number of pretraining tasks and the number of in-context demonstrations per task, respectively. For each $t \in [T]$, the t-th task consists of N demonstrations $\mathcal{D}_{t,N} := \{(\boldsymbol{x}_{t,n}, y_{t,n})\}_{n=1}^N \subset \mathbb{R}^d \times \{\pm 1\}$ and a test pair $(\boldsymbol{x}_{t,N+1}, y_{t,N+1})$. Let $\boldsymbol{U}_t := [\boldsymbol{u}_{t,1}, \dots, \boldsymbol{u}_{t,d}] \in \mathcal{O}(d)$ denote an orthogonal matrix whose columns form a task-specific orthonormal basis. For each $(t, n) \in [T] \times [N+1]$, let $\boldsymbol{a}_{t,n} := [a_{t,n,1}, \dots, a_{t,n,d}]^\top \in \mathbb{R}^d$ denote the coefficient vector for the n-th sample of the t-th task, and define the corresponding input vector as

$$\boldsymbol{x}_{t,n} := y_{t,n} \boldsymbol{U}_t \boldsymbol{a}_{t,n} = y_{t,n}(a_{t,n,1}\boldsymbol{u}_{t,1} + \cdots + a_{t,n,d}\boldsymbol{u}_{t,d}). \tag{1}$$

Task distribution. The matrices $\{\boldsymbol{U}_t\}_{t=1}^T$ are independently drawn according to the normalized Haar measure on $\mathcal{O}(d)$. In other words, the columns of each $\boldsymbol{U}_t$ form an orthonormal basis chosen uniformly at random. The labels $\{y_{t,n}\}_{t \in [T], n \in [N+1]}$ are independently drawn from the uniform distribution on $\{\pm 1\}$. The coefficient vectors $\{\boldsymbol{a}_{t,n}\}_{t \in [T], n \in [N+1]}$ are independently drawn from the multivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$. These three collections are mutually independent. The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are common across tasks.

Dimension of the Krylov subspace. Let $r \in [d]$ denote the dimension of $\mathcal{K}_d(\boldsymbol{\Sigma}, \boldsymbol{\mu})$, the order-$d$ Krylov subspace generated by $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$. More precisely, define

$$\mathcal{K}_d(\boldsymbol{\Sigma}, \boldsymbol{\mu}) := \operatorname{span}(\boldsymbol{\mu}, \boldsymbol{\Sigma}\boldsymbol{\mu}, \boldsymbol{\Sigma}^2\boldsymbol{\mu}, \ldots, \boldsymbol{\Sigma}^{d-1}\boldsymbol{\mu}), \qquad r := \dim \mathcal{K}_d(\boldsymbol{\Sigma}, \boldsymbol{\mu}). \tag{2}$$

Equivalently, r is the largest integer $k \in [d]$ such that the vectors $\boldsymbol{\mu}, \boldsymbol{\Sigma}\boldsymbol{\mu}, \ldots, \boldsymbol{\Sigma}^{k-1}\boldsymbol{\mu}$ are linearly independent. Consequently, $\mathcal{K}_d(\boldsymbol{\Sigma}, \boldsymbol{\mu}) = \mathcal{K}_r(\boldsymbol{\Sigma}, \boldsymbol{\mu})$. Moreover, r equals the number of distinct eigenvalues of $\boldsymbol{\Sigma}$ for which $\boldsymbol{\mu}$ has a nonzero projection onto the corresponding eigenspace. If $\boldsymbol{\Sigma} = \boldsymbol{I}_d$, so that the Gaussian coordinates are mutually independent and have unit variance, then $r = 1$. If $\boldsymbol{\mu}$ is the all-ones vector and $\boldsymbol{\Sigma}$ is diagonal, then r equals the number of distinct diagonal entries of $\boldsymbol{\Sigma}$. In general, r characterizes the representational complexity of the Bayes classifier for the Gaussian data $\boldsymbol{z} := y\boldsymbol{a}$. Without adversarial perturbations, this classifier is $(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^\top \boldsymbol{z}$. Its weight vector $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ belongs to $\mathcal{K}_r(\boldsymbol{\Sigma}, \boldsymbol{\mu})$, and r is the smallest Krylov order required to represent this vector.

Input sequence. Given the clean in-context demonstrations $\mathcal{D}_N$ and a possibly perturbed query $\boldsymbol{z} \in \mathbb{R}^d$, we define the input sequence $\boldsymbol{Z}^{(0)}$ for a linear transformer as

$$\boldsymbol{Z}^{(0)} := \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_N & \boldsymbol{z} \\ y_1 & y_2 & \cdots & y_N & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(N+1)}. \tag{3}$$

The bottom-right entry serves as a placeholder for the query prediction.

Linear transformer. In this study, we focus on transformers with linear self-attention, which are commonly used in recent theoretical work (Ahn et al., 2023; Mahankali et al., 2024; Gatmiry et al., 2024; Cheng et al., 2024; Zhang et al., 2024; Kumano et al., 2026). Let $L \in \mathbb{N}$ denote the number of layers. For each $\ell \in [L]$, we recursively define

$$\boldsymbol{Z}^{(\ell)} := \boldsymbol{Z}^{(\ell-1)} + \frac{1}{N}\boldsymbol{P}_\ell \boldsymbol{Z}^{(\ell-1)} \boldsymbol{M} \boldsymbol{Z}^{(\ell-1)\top} \boldsymbol{Q}_\ell \boldsymbol{Z}^{(\ell-1)}, \qquad \boldsymbol{M} := \begin{bmatrix} \boldsymbol{I}_N & \boldsymbol{0}_N \\ \boldsymbol{0}_N^\top & 0 \end{bmatrix}, \tag{4}$$

where $\boldsymbol{P}_\ell, \boldsymbol{Q}_\ell \in \mathbb{R}^{(d+1)\times(d+1)}$ are, respectively, the trainable value and key–query matrices in layer ℓ. As in prior work on in-context learning (Ahn et al., 2023; Gatmiry et al., 2024; Cheng et al., 2024; Kumano et al., 2026), we use the masking matrix $\boldsymbol{M}$ to prevent the tokens from attending to the query token. The resulting scalar predictor and the collection of trainable parameters are defined as

$$f(\boldsymbol{z}; \boldsymbol{\Theta}_L, \mathcal{D}_N) := \left[\boldsymbol{Z}^{(L)}\right]_{d+1,N+1}, \qquad \boldsymbol{\Theta}_L := \{(\boldsymbol{P}_\ell, \boldsymbol{Q}_\ell)\}_{\ell=1}^L. \tag{5}$$

Training objective. Let $\mathcal{L} : \mathbb{R} \to [0, \infty)$ be a loss function. Based on prior work (Bai et al., 2023; Ahn et al., 2023; Mahankali et al., 2024; Zhang et al., 2024; Kumano et al., 2026), we define the empirical risk over T pretraining tasks, each with N demonstrations, as follows:

$$\mathcal{R}_{\epsilon,T,N}(\boldsymbol{\Theta}_L) := \frac{1}{T}\sum_{t=1}^{T} \max_{\|\boldsymbol{\Delta}\|_2 \le \epsilon\sqrt{d}} \mathcal{L}\Big(y_{t,N+1} f(\boldsymbol{x}_{t,N+1} + \boldsymbol{\Delta}; \boldsymbol{\Theta}_L, \mathcal{D}_{t,N})\Big). \tag{6}$$

Minimizing $\mathcal{R}_{0,T,N}$ corresponds to standard training, whereas minimizing $\mathcal{R}_{\epsilon,T,N}$ for $\epsilon > 0$ corresponds to adversarial training. Training the transformer on multiple tasks with a single shared set of parameters encourages it to adaptively infer the task-specific structure from the demonstrations and generalize across tasks, rather than specialize in any particular task.

Taskwise zero-one error. For each fixed $\boldsymbol{U} \in \mathcal{O}(d)$, let $(\boldsymbol{x}, y)$ be a fresh test pair independently drawn from the same task-conditional distribution as the pretraining samples. We define the taskwise robust zero-one error of a predictor $g : \mathbb{R}^d \to \mathbb{R}$ by

$$\mathcal{E}_{\epsilon,\boldsymbol{U}}(g) := \mathbb{P}_{(\boldsymbol{x},y)|\boldsymbol{U}}\left(\inf_{\|\boldsymbol{\Delta}\|_2 \le \epsilon\sqrt{d}} y g(\boldsymbol{x} + \boldsymbol{\Delta}) \le 0\right). \tag{7}$$

Setting $\epsilon = 0$ yields the taskwise standard zero-one error $\mathcal{E}_{0,\boldsymbol{U}}$.

Taskwise Bayes error. Let $\mathcal{G}$ denote the class of all measurable functions $g : \mathbb{R}^d \to \mathbb{R}$. For each $\boldsymbol{U} \in \mathcal{O}(d)$, we define the taskwise robust Bayes error by $\mathcal{E}^\star_{\epsilon,\boldsymbol{U}} := \inf_{g\in\mathcal{G}} \mathcal{E}_{\epsilon,\boldsymbol{U}}(g)$. Setting $\epsilon = 0$ yields the taskwise standard Bayes error $\mathcal{E}^\star_{0,\boldsymbol{U}}$.

Assumptions. The assumptions used in this study are as follows:

Assumption 3.1.

(a) The mean vector $\boldsymbol{\mu}$ is nonzero, and the covariance matrix $\boldsymbol{\Sigma}$ is symmetric and positive definite; $\boldsymbol{\mu} \neq \mathbf{0}_d$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^\top \succ \mathbf{0}_{d,d}$.

(b) The perturbation radius $\epsilon\sqrt{d}$ is smaller than the norm of the mean vector $\boldsymbol{\mu}$; $0 \leq \epsilon\sqrt{d} < \|\boldsymbol{\mu}\|_2$.

(c) The number of layers L is at least the dimension r of the Krylov subspace; $L \geq r$.

(d) The loss function $\mathcal{L} : \mathbb{R} \to [0, \infty)$ is twice differentiable and satisfies $\mathcal{L}'(z) < 0$ and $\mathcal{L}''(z) > 0$ for every $z \in \mathbb{R}$, and $\sup_{z\in\mathbb{R}}\{|\mathcal{L}'(z)| + \mathcal{L}''(z)\} < \infty$.

Condition (a) rules out degenerate classification problems. Condition (b) prevents an adversary from completely canceling the mean signal. The role of Condition (c) is explained in the next section. Condition (d) ensures that the empirical risk minimization problem is well-posed. The logistic loss is an example satisfying Condition (d).

3.2 A Single Linear Transformer Can Represent Robust Bayes Classifiers Across Tasks

In this section, we first show that, given sufficient depth, a single linear transformer can asymptotically achieve the taskwise robust Bayes error through in-context learning. At this stage, we do not address whether the required parameters can be learned by empirical risk minimization.

Theorem 3.2 (A single linear transformer can be Bayes optimal across tasks)**.** *Suppose that (a) to (c) in Assumption 3.1 hold. Then, there exists a parameter $\boldsymbol{\Theta}_L$ such that, for every $\boldsymbol{U} \in \mathcal{O}(d)$,*

$$\mathcal{E}_{\epsilon,\boldsymbol{U}}\Big(f(\,\cdot\,;\boldsymbol{\Theta}_L, \mathcal{D}_N)\Big) - \mathcal{E}^\star_{\epsilon,\boldsymbol{U}} \lesssim_{\mathbb{P}} \frac{1}{N}. \tag{8}$$

The implied constant may depend on d, ϵ, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$.

The proof is given in Section B. The theorem establishes that a single linear transformer with a shared parameter asymptotically attains the robust Bayes error for every task as the number of demonstrations grows. In other words, without any task-specific parameter updates, the same linear transformer can adapt to every task through in-context learning and match the performance of the corresponding task-specific robust Bayes classifier. This uniform guarantee demonstrates the expressive capacity of linear transformers for in-context adaptation across tasks. Moreover, the excess error decays at a rate of N^{-1}. The rate is consistent with empirical observations of strong few-shot performance in in-context learning. Recall that r reflects the complexity of the Gaussian mixture model and the resulting difficulty of representing its robust Bayes classifier. Accordingly, the condition $L \geq r$ suggests that more complex data models call for deeper linear transformers to achieve Bayes optimality.

We next provide intuition about how a linear transformer can attain the Bayes error and why the depth condition $L \geq r$ arises.

Lemma 3.3 (Proof sketch of Theorem 3.2)**.** *Suppose that (a) and (b) in Assumption 3.1 hold. Suppose also that $N \to \infty$; write $\mathcal{D}_\infty := \mathcal{D}_N$. Let $\tau \geq 0$ be the unique solution of*

$$\tau\|(\boldsymbol{\Sigma} + \tau\boldsymbol{I}_d)^{-1}\boldsymbol{\mu}\|_2 = \epsilon\sqrt{d}. \tag{9}$$

(Bayes classifier) *For every $\boldsymbol{U} \in \mathcal{O}(d)$,*

$$g(\boldsymbol{x}) := \{(\boldsymbol{\Sigma} + \tau\boldsymbol{I}_d)^{-1}\boldsymbol{\mu}\}^\top \boldsymbol{U}^\top \boldsymbol{x}, \qquad \mathcal{E}_{\epsilon,\boldsymbol{U}}(g) = \mathcal{E}^\star_{\epsilon,\boldsymbol{U}}. \tag{10}$$

(Krylov order of the optimal direction) *Let $\mathcal{K}_0(\boldsymbol{\Sigma}, \boldsymbol{\mu}) := \{\mathbf{0}\}$. Then,*

$$(\boldsymbol{\Sigma} + \tau\boldsymbol{I}_d)^{-1}\boldsymbol{\mu} \in \mathcal{K}_r(\boldsymbol{\Sigma}, \boldsymbol{\mu}) \setminus \mathcal{K}_{r-1}(\boldsymbol{\Sigma}, \boldsymbol{\mu}). \tag{11}$$

(Effective weights of a linear transformer) *There exists a parameter family $\{\boldsymbol{\Theta}_{L,\xi} : \xi \in \mathbb{R}\}$ whose corresponding effective weights $\boldsymbol{\theta}(\boldsymbol{\Theta}_{L,\xi})$ satisfy, for every $\xi \in \mathbb{R}$, $\boldsymbol{U} \in \mathcal{O}(d)$, and $\boldsymbol{x} \in \mathbb{R}^d$,*

$$f(\boldsymbol{x}; \boldsymbol{\Theta}_{L,\xi}, \mathcal{D}_\infty) = \boldsymbol{\theta}(\boldsymbol{\Theta}_{L,\xi})^\top \boldsymbol{U}^\top \boldsymbol{x}. \tag{12}$$

(Span of the effective weights) *For the same parameter family,*

$$\operatorname{span}\{\boldsymbol{\theta}(\boldsymbol{\Theta}_{L,\xi}) : \xi \in \mathbb{R}\} = \mathcal{K}_L(\boldsymbol{\Sigma}, \boldsymbol{\mu}). \tag{13}$$

The proof is given in Section B. Although this lemma does not constitute a complete proof of Theorem 3.2, it provides a roadmap for the proof.

Bayes classifier. The first part of the lemma gives the Bayes classifier for $(\boldsymbol{x}, y)$ generated according to Eq. (1). The form of this function provides two insights into this classifier.[1]

First, the optimal weight is the solution to the following regularized mean–variance problem:

$$\operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d}\left\{\frac{1}{2}\mathbb{V}_{(\boldsymbol{x},y)|\boldsymbol{U}}[y\boldsymbol{w}^\top\boldsymbol{x}] - \mathbb{E}_{(\boldsymbol{x},y)|\boldsymbol{U}}[y\boldsymbol{w}^\top\boldsymbol{x}] + \frac{\tau}{2}\|\boldsymbol{w}\|_2^2\right\} = \boldsymbol{U}(\boldsymbol{\Sigma}+\tau\boldsymbol{I}_d)^{-1}\boldsymbol{\mu}. \tag{14}$$

The regularization parameter τ is determined by the perturbation level ϵ: it is zero when $\epsilon = 0$ and increases monotonically with ϵ. This ℓ_2 regularization controls the sensitivity to adversarial perturbations, since $|\boldsymbol{w}^\top\boldsymbol{\Delta}| \le \|\boldsymbol{w}\|_2\|\boldsymbol{\Delta}\|_2$. The taskwise Bayes classifier minimizing $\mathcal{E}_{\epsilon,\boldsymbol{U}}$ is a linear classifier obtained from a perturbation-dependent ℓ_2-regularized optimization problem.

Second, the classifier makes predictions while accounting for both variance and the effects of adversarial attacks. Substituting $\boldsymbol{x} = y\boldsymbol{U}\boldsymbol{a}$ gives

$$g(\boldsymbol{x}) := \left\{(\boldsymbol{\Sigma}+\tau\boldsymbol{I}_d)^{-1}\boldsymbol{\mu}\right\}^\top\boldsymbol{U}^\top\boldsymbol{x} = y\left\{(\boldsymbol{\Sigma}+\tau\boldsymbol{I}_d)^{-1}\boldsymbol{\mu}\right\}^\top\boldsymbol{a}. \tag{15}$$

For simplicity, suppose that $\boldsymbol{\Sigma} = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ for $\lambda_1, \ldots, \lambda_d > 0$. Then the i-th term in the inner product, for $i \in [d]$, can be expressed as $\mu_i a_i/(\lambda_i+\tau)$. When the perturbation level is small and hence τ is small, this term can be approximated by $\mu_i a_i/\lambda_i$. Thus, components with large variance have less influence on the prediction, whereas components with small variance have greater influence. In contrast, when τ is large, $g(\boldsymbol{x}) \approx y\boldsymbol{\mu}^\top\boldsymbol{a}/\tau$. The classifier relies more directly on strong features in the data without accounting for their variances.

Krylov order of the optimal direction. The second part of the lemma shows that the optimal direction belongs to the order-r Krylov subspace generated by $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$. This vector cannot belong to the order-$(r-1)$ Krylov subspace. The proof uses the Cayley–Hamilton theorem. This result suggests that, to attain the taskwise Bayes error for arbitrary $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$, a linear transformer must be expressive enough to internally represent weight vectors in the order-r Krylov subspace.

Effective weights of a linear transformer. The third part of the lemma shows that a suitable parameter family makes a linear transformer behave as a context-dependent linear classifier. We restrict $\boldsymbol{P}_{\ell,\xi}$ and $\boldsymbol{Q}_{\ell,\xi}$ for $\ell \in [L]$ as follows, and, under $N \to \infty$, the context matrix is as follows:

$$\boldsymbol{P}_{\ell,\xi} := \begin{bmatrix}\boldsymbol{0}_{d,d} & \boldsymbol{0}_d \\ \boldsymbol{0}_d^\top & 1\end{bmatrix},\ \boldsymbol{Q}_{\ell,\xi} := \begin{bmatrix}\xi\boldsymbol{I}_d & \boldsymbol{0}_d \\ \boldsymbol{0}_d^\top & 0\end{bmatrix},\ \frac{\boldsymbol{Z}^{(0)}\boldsymbol{M}\boldsymbol{Z}^{(0)\top}}{N} = \begin{bmatrix}\boldsymbol{U}(\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^\top)\boldsymbol{U}^\top & \boldsymbol{U}\boldsymbol{\mu} \\ \boldsymbol{\mu}^\top\boldsymbol{U}^\top & 1\end{bmatrix}. \tag{16}$$

The outputs after the first, second, and third layers are

$$\boldsymbol{Z}^{(1)} = \left\{\boldsymbol{I}_{d+1} + \frac{1}{N}\boldsymbol{P}_{1,\xi}\boldsymbol{Z}^{(0)}\boldsymbol{M}\boldsymbol{Z}^{(0)\top}\boldsymbol{Q}_{1,\xi}\right\}\boldsymbol{Z}^{(0)} = \begin{bmatrix}\boldsymbol{I}_d & \boldsymbol{0}_d \\ \xi\boldsymbol{\mu}^\top\boldsymbol{U}^\top & 1\end{bmatrix}\boldsymbol{Z}^{(0)}, \tag{17}$$

$$\boldsymbol{Z}^{(2)} = \begin{bmatrix}\boldsymbol{I}_d & \boldsymbol{0}_d \\ \{2\xi\boldsymbol{\mu}^\top + \xi^2\boldsymbol{\mu}^\top(\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^\top)\}\boldsymbol{U}^\top & 1\end{bmatrix}\boldsymbol{Z}^{(0)}, \tag{18}$$

$$\boldsymbol{Z}^{(3)} = \begin{bmatrix}\boldsymbol{I}_d & \boldsymbol{0}_d \\ \{3\xi\boldsymbol{\mu}^\top + 3\xi^2\boldsymbol{\mu}^\top(\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^\top) + \xi^3\boldsymbol{\mu}^\top(\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^\top)^2\}\boldsymbol{U}^\top & 1\end{bmatrix}\boldsymbol{Z}^{(0)}. \tag{19}$$

Consequently, the three-layer model's output on a query $\boldsymbol{x}$ and its effective weight are

$$f(\boldsymbol{x}; \boldsymbol{\Theta}_{3,\xi}, \mathcal{D}_\infty) = \boldsymbol{\theta}(\boldsymbol{\Theta}_{3,\xi})^\top\boldsymbol{U}^\top\boldsymbol{x},\ \boldsymbol{\theta}(\boldsymbol{\Theta}_{3,\xi}) := 3\xi\boldsymbol{\mu} + 3\xi^2(\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^\top)\boldsymbol{\mu} + \xi^3(\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^\top)^2\boldsymbol{\mu}. \tag{20}$$

The same construction shows inductively that, with suitable parameters, a linear transformer of any depth can act as a context-dependent linear classifier whose effective weight $\boldsymbol{\theta}$ is independent of $\boldsymbol{U}$.

Span of the effective weights. The fourth part of the lemma states that, as ξ varies, the effective weights associated with this parameter family span the order-L Krylov subspace. The equations above illustrate this mechanism: stacking L layers yields effective weights that are linear combinations of

[1] Closely related characterizations of the Bayes classifier were previously established by Bhagoji et al. (2019); Dan et al. (2020), and we therefore do not claim this characterization as a novel contribution. Here, it serves only as an intermediate result in the proof of Theorem 3.2.

$\boldsymbol{\mu}, \boldsymbol{\Sigma}\boldsymbol{\mu}, \ldots, \boldsymbol{\Sigma}^{L-1}\boldsymbol{\mu}$. As shown in the second part, the optimal direction lies in the order-r Krylov subspace. Hence, when $L \geq r$, the optimal direction lies in the span of the effective weights. These results explain how linear transformers achieve Bayes optimality and why the depth condition is required. Note that, even when $L \geq r$, this lemma alone does not establish that an effective weight $\boldsymbol{\theta}$ can exactly match the optimal direction $(\boldsymbol{\Sigma} + \tau \boldsymbol{I}_d)^{-1}\boldsymbol{\mu}$. See Section B for the complete proof.

3.3 Adversarial Training Enables Linear Transformers to Achieve Robust Bayes Optimality Across Tasks; Standard Training Does Not

The preceding section establishes that sufficiently deep linear transformers have the expressive capacity to asymptotically attain robust Bayes optimality across tasks. However, this existence result does not show whether such parameters can be learned by a training procedure. We address this question by analyzing empirical risk minimization under adversarial and standard training.

Theorem 3.4 (Adversarial training yields robust linear transformers)**.** *Suppose that Assumption 3.1 holds. For sufficiently large $B > 0$, choose any empirical risk minimizer*

$$\widehat{\boldsymbol{\Theta}}_{L,\epsilon,T,N} \in \operatorname*{argmin}_{\max_{\ell\in[L]}\{\|\boldsymbol{P}_\ell\|_{\mathrm{F}},\|\boldsymbol{Q}_\ell\|_{\mathrm{F}}\}\leq B} \mathcal{R}_{\epsilon,T,N}(\boldsymbol{\Theta}_L). \tag{21}$$

Then, for every $\boldsymbol{U} \in \mathcal{O}(d)$,

$$\mathcal{E}_{\epsilon,\boldsymbol{U}}\left(f\left(\cdot;\widehat{\boldsymbol{\Theta}}_{L,\epsilon,T,N},\mathcal{D}_N\right)\right) - \mathcal{E}^{\star}_{\epsilon,\boldsymbol{U}} \lesssim_{\mathbb{P}} \frac{1}{\sqrt{T}} + \frac{1}{N}. \tag{22}$$

The implied constant may depend on d, B, L, $\mathcal{L}$, ϵ, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$.

The proof is given in Section C. This theorem complements the representability result in Theorem 3.2 by showing that adversarial training can produce a transformer that asymptotically achieves robust Bayes optimality across tasks. For $d \geq 2$, a test-task basis $\boldsymbol{U}$ differs almost surely from every pretraining basis under the Haar measure on $\mathcal{O}(d)$, even when there are countably infinitely many pretraining tasks. Thus, the theorem establishes robust generalization to previously unseen tasks, which cannot be attributed to exact memorization of the pretraining bases. Moreover, the excess-error bound decomposes into a term depending on T and a term depending on N, rather than taking a joint form such as $1/(TN)$. This decomposition isolates the contributions of cross-task learning from pretraining and within-task inference from in-context demonstrations. The bound converges to zero only as both the number of pretraining tasks and the number of in-context demonstrations grow; increasing either one alone might be insufficient for the convergence to Bayes optimality.

We next contrast this result with standard training.

Theorem 3.5 (Standard training does not yield robust linear transformers)**.** *Suppose that Assumption 3.1 holds. Suppose also that $T, N \to \infty$; write $\mathcal{D}_\infty := \mathcal{D}_N$ and $\mathcal{R}_{0,\infty,\infty} := \mathcal{R}_{0,T,N}$. For sufficiently large $B > 0$, choose any population risk minimizer*

$$\widehat{\boldsymbol{\Theta}}_{L,0,\infty,\infty} \in \operatorname*{argmin}_{\max_{\ell\in[L]}\{\|\boldsymbol{P}_\ell\|_{\mathrm{F}},\|\boldsymbol{Q}_\ell\|_{\mathrm{F}}\}\leq B} \mathcal{R}_{0,\infty,\infty}(\boldsymbol{\Theta}_L). \tag{23}$$

Then, there exists a constant $C > 0$ such that, for every $\boldsymbol{U} \in \mathcal{O}(d)$,

$$\mathcal{E}_{\epsilon,\boldsymbol{U}}\left(f\left(\cdot;\widehat{\boldsymbol{\Theta}}_{L,0,\infty,\infty},\mathcal{D}_\infty\right)\right) - \mathcal{E}^{\star}_{\epsilon,\boldsymbol{U}} = \begin{cases} 0 & (\epsilon = 0 \text{ or } r = 1) \\ C & (\epsilon > 0 \text{ and } r \geq 2) \end{cases}. \tag{24}$$

The constant C may depend on d, ϵ, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$.

The proof is given in Section C. In contrast to Theorem 3.4, this theorem establishes an impossibility result for standard training: when $\epsilon > 0$ and $r \geq 2$, standard training cannot produce a linear transformer whose predictions attain robust Bayes optimality. This impossibility is not due to a finite amount of training data; it persists in the population limit, even when both the number of pretraining tasks and the number of demonstrations per task tend to infinity. These results reveal a fundamental difference between standard and adversarial training and suggest that, at least for linear transformers, adversarial training is necessary to obtain adversarially robust in-context learners.

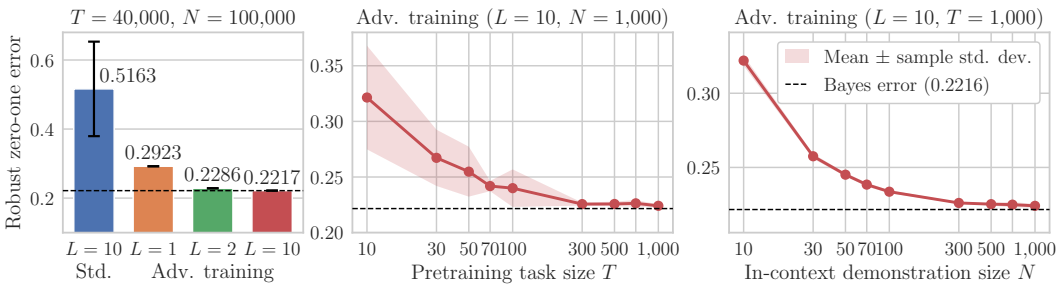

Figure 1: Robust zero-one errors for $d = 100$, $r = 10$, and $\epsilon = 0.08$. Means and sample standard deviations are shown over three runs in the left panel and five runs in the other panels.

3.4 Other Theoretical Results

Gradient flow to Bayes optimality. The preceding sections show that global minimizers of the adversarial empirical risk asymptotically achieve taskwise Bayes optimality. However, practical optimization algorithms such as gradient descent are not guaranteed to find a global minimizer. To bridge this gap, Theorem D.1 analyzes the case $L = 2$ and $r \leq 2$, in which only the final key–query matrix is trained. We show that gradient flow yields predictors whose robust errors converge to the corresponding Bayes errors.

Accuracy–robustness trade-off. Robustness often comes at the cost of clean accuracy. In particular, adversarially trained models may be less accurate on unperturbed samples than their standardly trained counterparts (Su et al., 2018; Tsipras et al., 2019; Ilyas et al., 2019). In Theorem E.1, we establish that the same trade-off arises in our setting. Moreover, the trade-off persists even with infinitely many pretraining tasks and infinitely many demonstrations per task. This result highlights a fundamental challenge in developing robust foundation models.

Demonstration complexity. Achieving a target level of robust accuracy often requires more training samples than achieving the corresponding level of standard accuracy (Schmidt et al., 2018; Dan et al., 2020). In Theorem F.1, we establish an analogous gap in the number of in-context demonstrations. Attaining the same fixed excess error relative to the respective Bayes errors requires only one demonstration for standard prediction but $\Theta(\sqrt{d})$ demonstrations for robust prediction. This result reveals another fundamental challenge in developing robust foundation models.

3.5 Comparison with Prior Work

Compared to Kumano et al. (2026), we substantially relax their assumptions on the data distribution and the model architecture while establishing stronger robustness guarantees for linear transformers.

Data distribution. Their data model restricts feature scales to two discrete levels, such as $y\boldsymbol{x} = [\alpha, \ldots, \alpha, \beta, \ldots, \beta]$, where $\alpha, \beta > 0$ represent large and small values, respectively. This is a major limitation they acknowledge. Our framework allows arbitrary feature scales. They also consider only the standard basis, which roughly corresponds to fixing $\boldsymbol{U}$ to a particular value in our framework. In contrast, our theory applies to any orthonormal basis. Although they allow a broader class of coefficient distributions than our Gaussian family, they impose strong constraints on the covariance structure, whereas we do not. Overall, our framework broadens the class of admissible data models.

Architecture. Theoretically, their analysis is limited to single-layer linear transformers, whereas ours applies to multilayer linear transformers. Their experiments are also limited to single-layer linear transformers trained on relatively small datasets, including MNIST (Deng, 2012), Fashion-MNIST (Xiao et al., 2017), and CIFAR-10 (Krizhevsky, 2009). By contrast, we conduct experiments with multilayer transformers equipped with softmax attention on larger datasets, including CIFAR-100 (Krizhevsky, 2009), Tiny ImageNet (Le & Yang, 2015), and Caltech-256 (Griffin et al., 2007), thereby evaluating the effectiveness of robust foundation models in more realistic settings.

Robustness guarantees. Although differences in assumptions make direct comparisons difficult, our results provide more direct and comprehensive robustness guarantees. First, they establish only

Table 1: Clean/robust accuracy (%) of six-layer transformers with softmax attention on real-world data, reported as mean ± sample standard deviation across five independent runs.

Training	CIFAR-100	Tiny ImageNet	Caltech-256
Standard	**79.1** ± **2.0** / 26.8 ± 3.4	**77.4** ± **3.4** / 27.6 ± 7.6	**78.0** ± **3.6** / 23.8 ± 6.5
Adversarial	77.9 ± 1.7 / **63.2** ± **1.8**	67.0 ± 6.0 / **52.8** ± **2.2**	74.6 ± 3.9 / **62.4** ± **4.6**

bounds on the expected robust margin $\mathbb{E}[yf(\boldsymbol{x} + \boldsymbol{\Delta})]$. Because the expected robust margin does not generally determine robust accuracy, these bounds alone do not provide a direct guarantee of robustness. By contrast, we prove that the robust zero-one error converges to the taskwise robust Bayes error, thereby directly guaranteeing transformer robustness. Second, their analysis does not determine whether linear transformers are robust or vulnerable in certain data regimes. As they acknowledge, their single-layer restriction limits model capacity and prevents their analysis from capturing the full robustness potential of linear transformers. By considering a multilayer model, we show that linear transformers asymptotically achieve taskwise robust Bayes optimality when they are sufficiently deep relative to the data complexity (i.e., $L \geq r$). Overall, our analysis provides more definitive guarantees and more fully characterizes the robustness of linear transformers.

4 Experimental Results

Detailed experimental settings and additional results are provided in Section A.

Synthetic datasets. We first empirically validate our theoretical results using synthetic datasets generated according to Eq. (1). We set $d = 100$, $r = 10$, and $\epsilon = 0.08$. We train the linear transformer in Eq. (5) by minimizing the logistic loss with Adam. Each model is evaluated on 40,000 test tasks with 200 queries per task. The left panel of Fig. 1 compares robust zero-one errors for $T = 40{,}000$ and $N = 100{,}000$. Under adversarial training, the robust zero-one error decreases with depth, and the depth-10 model essentially attains the robust Bayes error, supporting Theorems 3.2 and 3.4. The near-optimal performance of the depth-2 model does not contradict the theorem because $L \geq r$ is a sufficient but not necessary condition. In contrast, despite having sufficient depth and being trained on abundant data, the depth-10 model obtained by standard training remains far from optimal, consistent with Theorem 3.5. The center and right panels show that the depth-10 adversarially trained model approaches the robust Bayes error as T and N increase, respectively, supporting Theorem 3.4.

Real-world datasets. We then evaluate six-layer transformers with softmax attention on binary classification tasks from CIFAR-100, Tiny ImageNet, and Caltech-256. The pretraining, validation, and test classes are mutually disjoint. Clean and robust test accuracies are shown in Tab. 1. Adversarial training substantially improves robust accuracy, albeit at the cost of lower clean accuracy.

5 Conclusion and Limitations

We showed that adversarial pretraining can enable a sufficiently deep linear transformer to adapt to unseen Gaussian-mixture classification tasks through in-context learning from clean demonstrations and asymptotically attain the taskwise robust Bayes error. By contrast, standard pretraining does not, even in the population limit. We further established convergence under gradient flow and identified two challenges in robust in-context learning. These results provide theoretical evidence that robustness acquired once during pretraining can transfer to downstream tasks without additional adversarial training, supporting the feasibility of robust foundation models.

Our analysis is restricted to Gaussian-mixture tasks that share a common latent mean and covariance, so the unseen tasks considered in our theory still belong to a structured task family. Although such cross-task sharing is common in theoretical analyses of in-context learning (Ahn et al., 2023; Mahankali et al., 2024; Kumano et al., 2026), extending our guarantees to settings with task-dependent means or covariances and to non-Gaussian inputs is an important direction for future theoretical work. Moreover, the practical utility of robust foundation models remains unknown at the scale of modern foundation-model pretraining. We hope that our theoretical results will motivate large-scale empirical studies by large organizations with the computational resources required to conduct them.

AI USE STATEMENT

In this work, we used generative AI tools to formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, implement methods, and assist with translation. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, and interpret results. Dataset cleaning and reformatting, and support for qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create or edit software code, draft parts of a research paper, summarize or analyze existing literature, brainstorm, source or search for information, edit a research paper to improve readability, identify relevant literature, and suggest a structure for a research paper. We did not use generative AI tools to create or modify scientific figures or images, suggest experimental parameters, create artifacts, discover research topics or identify gaps, format references, and propose a title or keywords for a research paper. Formulating questions for surveys or interviews and transcribing recordings of research materials are not applicable to this work. We have reviewed all AI-assisted work. We independently checked all proofs for mathematical correctness, tested all experimental code, and read and evaluated every cited work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is a theoretical study of adversarial robustness supported by experiments on synthetic and real-world benchmark datasets. We use only publicly available datasets and do not collect new data or involve human participants. Our experiments do not interact with deployed systems.

REPRODUCIBILITY STATEMENT

We state all assumptions explicitly in the main text and provide complete proofs of all theoretical results in the appendix. Detailed experimental settings are provided in Section A. The supplementary material contains the code to reproduce the reported experimental results.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *NeurIPS*, volume 36, pp. 45614–45650, 2023.

Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. In *NeurIPS*, volume 33, pp. 16048–16059, 2020.

Usman Anwar, Johannes Von Oswald, Louis Kirsch, David Krueger, and Spencer Frei. Understanding in-context learning of linear models in transformers through an adversarial lens. *TMLR*, 2025. ISSN 2835-8856.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, pp. 274–283, 2018.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *NeurIPS*, volume 36, pp. 57125–57211, 2023.

Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport. In *NeurIPS*, volume 32, 2019.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, volume 33, pp. 1877–1901, 2020.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *SP*, pp. 39–57, 2017a.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop*, pp. 3–14, 2017b.

Xiang Cheng, Yuxin Chen, and Suvrit Sra. Transformers implement functional gradient descent to learn non-linear functions in context. In *ICML*, 2024.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, pp. 2206–2216, 2020.

Chen Dan, Yuting Wei, and Pradeep Ravikumar. Sharp statistical guaratees for adversarially robust gaussian classification. In *ICML*, pp. 2345–2355, 2020.

Li Deng. The MNIST database of handwritten digit images for machine learning research. *Signal Processing Magazine*, 29(6):141–142, 2012.

Deqing Fu, Tian-qi Chen, Robin Jia, and Vatsal Sharan. Transformers learn to achieve second-order convergence rates for in-context linear regression. In *NeurIPS*, volume 37, pp. 98675–98716, 2024.

Shaopeng Fu and Di Wang. Understanding and improving continuous LLM adversarial training via in-context learning theory. In *ICLR*, 2026.

Shaopeng Fu, Liang Ding, Jingfeng Zhang, and Di Wang. Short-length adversarial training helps LLMs defend long-length jailbreak attacks: Theoretical and empirical evidence. In *NeurIPS*, volume 38, pp. 64351–64391, 2025.

Khashayar Gatmiry, Nikunj Saunshi, Sashank J Reddi, Stefanie Jegelka, and Sanjiv Kumar. Can looped transformers learn to implement multi-step gradient descent for in-context learning? In *ICML*, 2024.

Angeliki Giannou, Liu Yang, Tianhao Wang, Dimitris Papailiopoulos, and Jason D Lee. How well can transformers emulate in-context newton's method? *arXiv:2403.03183*, 2024.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

Shruthi Gowda, Bahram Zonooz, and Elahe Arani. Conserve-update-revise to cure generalization and robustness trade-off in adversarial training. In *ICLR*, pp. 15465–15485, 2024.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. Technical report, 2007.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2 edition, 2009.

Pengyue Hou, Ming Zhou, Jie Han, Petr Musilek, and Xingyu Li. Adversarial fine-tune with dynamically regulated adversary. In *IJCNN*, pp. 1–8, 2022.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *NeurIPS*, volume 32, pp. 125–136, 2019.

Ahmadreza Jeddi, Mohammad Javad Shafiee, and Alexander Wong. A simple fine-tuning is all you need: Towards robust deep learning via adversarial fine-tuning. In *CVPR Workshop*, 2021.

Xiaojun Jia, Yong Zhang, Xingxing Wei, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. Prior-guided adversarial initialization for fast adversarial training. In *ECCV*, pp. 567–584, 2022.

Hoki Kim, Woojin Lee, and Jaewook Lee. Understanding catastrophic overfitting in single-step adversarial training. In *AAAI*, volume 35, pp. 8119–8127, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Soichiro Kumano, Hiroshi Kera, and Toshihiko Yamasaki. Adversarially pretrained transformers may be universally robust in-context learners. In *ICLR*, pp. 124615–124659, 2026.

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. CS231N course report, Stanford University, 2015.

Jonathan Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. In *NeurIPS*, volume 36, pp. 43057–43083, 2023.

Tianle Li, Chenyang Zhang, Xingwu Chen, Yuan Cao, and Difan Zou. On the robustness of transformers against context hijacking for linear classification. In *NeurIPS*, volume 38, pp. 69373–69411, 2025.

Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. In *ICLR*, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Arvind Mahankali, Tatsunori B Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *ICLR*, 2024.

Geon Yeong Park and Sang Wan Lee. Reliably fast adversarial training via latent adversarial perturbation. In *ICCV*, pp. 7758–7767, 2021.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *NeurIPS*, volume 31, 2018.

Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *NeurIPS*, volume 32, 2019.

Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *ECCV*, pp. 631–648, 2018.

Satoshi Suzuki, Shin'ya Yamaguchi, Shoichiro Takeda, Sekitoshi Kanai, Naoki Makishima, Atsushi Ando, and Ryo Masumura. Adversarial finetuning with latent representation constraint to mitigate accuracy-robustness tradeoff. In *ICCV*, pp. 4367–4378, 2023.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *NeurIPS*, volume 33, pp. 1633–1645, 2020.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *ICLR*, 2019.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *ICLR*, 2020.

Sophie Xhonneux, Alessandro Sordoni, Stephan Günnemann, Gauthier Gidel, and Leo Schwinn. Efficient adversarial training in llms with continuous attacks. In *NeurIPS*, volume 37, pp. 1502–1530, 2024.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv:1708.07747*, 2017.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *ICLR*, 2022.

Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. In *NeurIPS*, volume 32, 2019.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *JMLR*, 25(49):1–55, 2024.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. In *AISTATS*, 2025.

Haizhong Zheng, Ziqi Zhang, Juncheng Gu, Honglak Lee, and Atul Prakash. Efficient adversarial training with transferable adversarial examples. In *CVPR*, pp. 1178–1187, 2020.

A Additional Experimental Results **14**
A.1 Synthetic Experiments 14
A.2 Real-World Experiments 16

B Robust Bayes Implementation by Linear Transformers: Proof of Theorem 3.2 **17**
B.1 Robust Bayes classifier 17
B.2 Query-Linear Representation 21
B.3 Krylov structure and sufficient-depth realization 21

C Adversarial and Standard Training: Proofs of Theorems 3.4 and 3.5 **25**
C.1 Directions selected by infinite-sample margin risks 25
C.2 Infinite-sample minimizers and finite-context transfer 26
C.3 Finite-task adversarial risk minimization 27

D Gradient Flow to Robust Bayes Optimality **31**

E Accuracy–Robustness Trade-Off **33**

F Demonstration Complexity **34**

A Additional Experimental Results

All experiments were conducted on a Linux server with an x86_64 CPU architecture and four NVIDIA A100-SXM4-80GB GPUs.

A.1 Synthetic Experiments

Experimental design. We choose an anisotropic Gaussian distribution for which the standard and robust Bayes directions differ, allowing the experiments to distinguish standard from robust direction recovery. Specifically, we set $d = 100$ and $\epsilon = 0.08$ and, for every $i \in [10]$, define the mean and diagonal covariance entries by

$$\mu_i := \frac{i}{10}, \qquad \Sigma_{i,i} := \mu_i 10^{-1+2(i-1)/9}. \tag{A1}$$

For every $i \in [d] \setminus [10]$, we set $\mu_i := 0$ and $\Sigma_{i,i} := 10^{-6}$. The active signal-to-variance ratios $\mu_i/\Sigma_{i,i} = 10^{1-2(i-1)/9}$ therefore range logarithmically from 10 to 10^{-1}. Because the ten active coordinates have distinct covariance eigenvalues, this construction has Krylov dimension $r = 10$. We independently draw each task basis $\boldsymbol{U}$ from the normalized Haar distribution on $\mathcal{O}(d)$ and then sample demonstrations and queries from Eq. (1). Each fixed pretraining dataset contains T task bases, N clean demonstrations per task, and one clean query per task. For evaluation, we use 40,000 independently drawn test tasks, a fresh context with the same value of N for each task, and 200 independent queries per context. We conduct three independent runs for $T = 40{,}000$ and $N = 100{,}000$ and five independent runs otherwise. Each run uses separate training and evaluation random streams, and we report means and sample standard deviations across runs.

Model and training. We use the one-head masked linear-attention architecture in Eq. (4) and restrict its parameters to a rotation-equivariant family that respects the symmetry of the task distribution. In particular, every layer has the form

$$\boldsymbol{P}_\ell := p_{\ell,x}\boldsymbol{\Pi}_x + p_{\ell,y}\boldsymbol{\Pi}_y, \qquad \boldsymbol{Q}_\ell := q_{\ell,x}\boldsymbol{\Pi}_x + q_{\ell,y}\boldsymbol{\Pi}_y, \tag{A2}$$

$$\boldsymbol{\Pi}_x := \mathrm{diag}(\boldsymbol{I}_d, 0), \qquad \boldsymbol{\Pi}_y := \mathrm{diag}(\boldsymbol{0}_d, 1). \tag{A3}$$

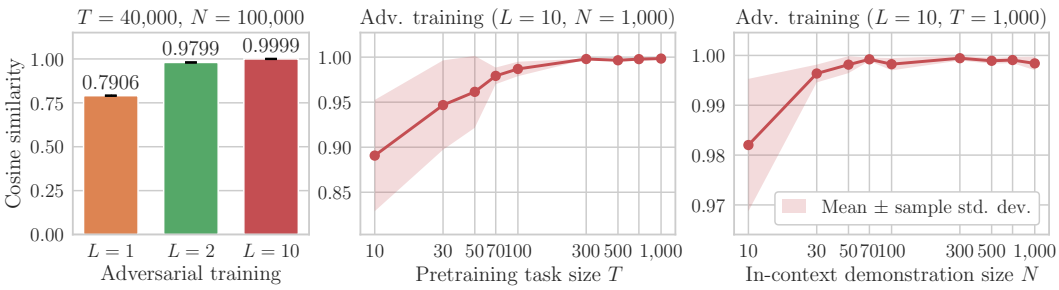

Figure A1: Cosine similarity between the population-context effective direction of an adversarially trained transformer and the robust Bayes direction. The distribution and the three experimental conditions are the same as in Fig. 1. The left panel varies the depth at $T = 40{,}000$ and $N = 100{,}000$, the center panel varies T with $L = 10$ and $N = 1{,}000$, and the right panel varies N with $L = 10$ and $T = 1{,}000$. Means and sample standard deviations are shown over three runs in the left panel and five runs in the other panels.

The four coefficients in each layer are trainable and initialized independently from $\mathcal{N}(0, 0.01^2)$. We minimize the logistic loss $\mathcal{L}(z) = \log(1+\exp(-z))$, which satisfies Assumption 3.1(d), by full-batch Adam for at most 2,500 updates on the fixed pretraining dataset. The learning rate increases linearly to 0.01 over the first 30 updates and subsequently either remains constant or follows cosine annealing to 0.001 without restarts. For each schedule, we retain the checkpoint with the smallest training loss and report the better of the two retained checkpoints under the same criterion.

Robust evaluation and Bayes baselines. Masked linear attention is linear in the query, so both adversarial training and robust evaluation can use the exact worst-case perturbation rather than an approximate attack such as the fast gradient sign method (Goodfellow et al., 2015) or projected gradient descent (PGD) (Madry et al., 2018). The adversary perturbs only the query and leaves all in-context demonstrations fixed and clean. Writing $\boldsymbol{w}(\boldsymbol{\Theta}_L, \mathcal{D}_N)$ for the context-dependent effective query direction, the exact perturbation for $\boldsymbol{w}(\boldsymbol{\Theta}_L, \mathcal{D}_N) \neq \mathbf{0}_d$ is

$$\boldsymbol{\Delta} = -y\epsilon\sqrt{d}\frac{\boldsymbol{w}(\boldsymbol{\Theta}_L, \mathcal{D}_N)}{\|\boldsymbol{w}(\boldsymbol{\Theta}_L, \mathcal{D}_N)\|_2}. \tag{A4}$$

When the effective direction is zero, we instead set $\boldsymbol{\Delta} = \mathbf{0}_d$, and during adversarial training we recompute the perturbation from the current model at every update. For comparison with the theoretical optimum, we obtain the robust Bayes direction $(\boldsymbol{\Sigma} + \tau \boldsymbol{I}_d)^{-1}\boldsymbol{\mu}$ by numerically solving

$$\tau\|(\boldsymbol{\Sigma} + \tau \boldsymbol{I}_d)^{-1}\boldsymbol{\mu}\|_2 = \epsilon\sqrt{d} \tag{A5}$$

and compute all Bayes risks analytically from the resulting Gaussian margins.

Alignment with the robust Bayes direction. To examine whether adversarial training recovers the optimal classifier rather than merely its error, Fig. A1 compares the effective direction of linear transformers with the robust Bayes direction characterized in Lemma B.4. At $T = 40{,}000$ and $N = 100{,}000$, the cosine similarity increases with depth and approaches one at depth ten. Thus, the depth-10 model recovers not only the robust Bayes error but also the unique optimal direction up to positive scaling. Increasing the number of pretraining tasks similarly improves the cosine similarity, while the values throughout the demonstration sweep remain close to one. Together with the robust errors in Fig. 1, these results support the convergence to the robust Bayes predictor established in Theorem 3.4.

Accuracy–robustness trade-off and demonstration complexity. To assess the accuracy–robustness trade-off and the qualitative dependence on context length, Tab. A1 compares clean and robust accuracy as the number of demonstrations increases. Standard training gives higher clean accuracy overall, whereas adversarial training gives substantially higher robust accuracy, supporting the accuracy–robustness trade-off in Theorems 3.5 and E.1. Moreover, standard clean accuracy saturates with relatively few demonstrations, while adversarial robust accuracy continues to improve toward the robust Bayes accuracy. This behavior is qualitatively consistent with Theorem F.1.

Table A1: Clean and robust accuracies (%), reported as mean ± sample standard deviation across five runs, for depth-10 models trained on $T = 1{,}000$ tasks. Each row uses the displayed number of demonstrations in both pretraining and evaluation. Robust accuracy uses $\epsilon = 0.08$, and every model is evaluated on 40,000 test tasks with 200 queries per task.

	Standard training		Adversarial training	
N	Clean	Robust	Clean	Robust
10	84.6 ± 4.2	63.5 ± 4.5	84.8 ± 0.7	67.8 ± 0.3
50	95.1 ± 4.3	52.8 ± 20.6	92.4 ± 0.3	75.5 ± 0.1
100	98.0 ± 0.6	45.1 ± 15.9	93.6 ± 0.3	76.6 ± 0.1
500	98.7 ± 0.2	39.6 ± 7.0	94.1 ± 0.3	77.5 ± 0.1
1,000	98.4 ± 0.6	45.1 ± 13.0	94.0 ± 0.1	77.6 ± 0.1

A.2 Real-World Experiments

Datasets and class splits. We evaluate whether the benefit of adversarial pretraining extends beyond the theoretical Gaussian model by considering CIFAR-100 (Krizhevsky, 2009), Tiny ImageNet (Le & Yang, 2015), and Caltech-256 (Griffin et al., 2007). In each of the five independent runs, the respective pretraining splits contain 80, 180, and 236 classes, while each validation and test split contains ten classes, with all three partitions mutually disjoint. We draw a different class partition in each run.

PCA features. We represent each image by an unwhitened principal component analysis (PCA) features to reduce the input dimension before constructing the in-context episodes. Let d_{PCA} denote the number of retained components and hence the dimension of each PCA feature. For every dataset and run, we scale RGB values to $[0, 1]$ and flatten each image in channel–height–width order. We fit the coordinatewise mean and PCA components using only images from that run's pretraining classes, excluding all images from validation and test classes. We compute PCA with 32 oversampling dimensions and seven power iterations. After fitting PCA, we use the same mean and components to project every image in the training and query splits. We retain $d_{\mathrm{PCA}} = 256$ components for CIFAR-100 and $d_{\mathrm{PCA}} = 512$ components for Tiny ImageNet and Caltech-256. For Caltech-256, we exclude the clutter class and select 60 training images and 20 query images from each of the remaining 256 classes. Before computing its PCA feature, we convert each Caltech-256 image to RGB, center it on a square canvas padded with the RGB value $(128, 128, 128)$, and resize it to 64×64 pixels using bicubic interpolation with antialiasing.

Binary in-context episodes. Each episode is a binary classification task whose two classes determine the meaning of the labels within that context. At the beginning of every pretraining epoch, we randomly permute the pretraining classes and pair consecutive classes, ensuring that each class appears in exactly one task during that epoch. For each class in a task, we sample 50 demonstrations and 10 queries without replacement from its training features, with the demonstration and query sets disjoint within the task. Validation and testing instead enumerate all 45 unordered pairs of the corresponding ten held-out classes. For each pair, we form one fixed context from 50 training features per class and evaluate it on 10 query-split features per class, yielding 900 queries in each final test.

Transformer and optimization. We use a six-layer pre-normalization transformer with softmax attention, residual connections, and GELU feed-forward networks, thereby testing whether the qualitative prediction of the linear-attention theory persists in a standard nonlinear architecture. Each PCA feature is augmented with one of three learned embeddings that identifies a demonstration with label -1, a demonstration with label $+1$, or an unlabeled query. CIFAR-100 uses four attention heads and a feed-forward dimension of 1,024, whereas Tiny ImageNet and Caltech-256 use eight attention heads and a feed-forward dimension of 2,048; every attention head has dimension 64. A linear head applied after the final layer normalization produces the scalar score optimized with the logistic loss. We use Xavier uniform initialization for linear weights, zero linear biases, independent $\mathcal{N}(0, 0.02^2)$ initialization for the three embeddings, and unit scales and zero biases for layer normalization. We train for 300 epochs with AdamW, exponential decay rates 0.9 and 0.95, numerical stability constant 10^{-8}, weight decay 0.05, and global gradient clipping at 1.0. The standard and adversarial runs use peak learning rates of 3×10^{-4} and 10^{-3}, respectively, with linear warmup over the first 5% of

updates followed by cosine decay to zero. We initialize the standard model from scratch and the adversarial model from the corresponding standardly trained model. Model selection minimizes the clean validation logistic loss for standard training and the 20-step PGD validation loss for adversarial training.

Adversarial training and evaluation. The real-world threat model follows the theoretical setting by perturbing only the query while keeping all demonstrations clean and fixed. Specifically, the adversary may change a query PCA feature within an L_2 ball of radius $\epsilon\sqrt{d_{\mathrm{PCA}}}$, where $\epsilon = 0.1$ in both training and evaluation. Adversarial training uses untargeted 10-step PGD with one uniformly sampled random start and step size $\epsilon\sqrt{d_{\mathrm{PCA}}}/4$. For final evaluation, we measure robust accuracy using APGD-CE and binary FAB (Croce & Hein, 2020), each run for 100 iterations. We omit the remaining AutoAttack components because the DLR loss is undefined for scalar binary scores and Square Attack relies on spatial image structure.

B Robust Bayes Implementation by Linear Transformers: Proof of Theorem 3.2

B.1 Robust Bayes classifier

In this section, we characterize the robust Bayes classifier and its error for a fixed task independently of the linear-transformer architecture. For $\boldsymbol{U} = \boldsymbol{I}_d$ and thus $\boldsymbol{x} = y\boldsymbol{a}$, we derive the robust zero-one error of every linear function in Lemma B.1, the standard Bayes error in Lemma B.2, and the robust Bayes error and all linear functions attaining it in Lemma B.3. Then, for every $\boldsymbol{U} \in \mathcal{O}(d)$ and $\boldsymbol{x} = y\boldsymbol{U}\boldsymbol{a}$, we generalize these results in Lemma B.4.

Variants or equivalent formulations of these results appear in Hastie et al. (2009); Bhagoji et al. (2019); Dan et al. (2020). These lemmas are auxiliary to our main results, and we do not claim them as novel contributions. For completeness, we provide self-contained proofs, since adapting the existing results to our setting would require essentially as much exposition as proving them directly.

Lemma B.1 (Robust error of linear functions for the identity-basis task)**.** *Suppose that y is uniform on $\{\pm 1\}$ and $\boldsymbol{x} \mid y \sim \mathcal{N}(y\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \succ \boldsymbol{0}_{d,d}$. For every $\boldsymbol{h} \in \mathbb{R}^d$, define $g_{\boldsymbol{h}} : \mathbb{R}^d \to \mathbb{R}$ by $g_{\boldsymbol{h}}(\boldsymbol{x}) := \boldsymbol{h}^\top \boldsymbol{x}$. Then, for every $\epsilon \geq 0$ and $\boldsymbol{h} \in \mathbb{R}^d \setminus \{\boldsymbol{0}_d\}$,*

$$\mathbb{P}\left(\inf_{\|\boldsymbol{\Delta}\|_2 \leq \epsilon\sqrt{d}} y g_{\boldsymbol{h}}(\boldsymbol{x} + \boldsymbol{\Delta}) \leq 0\right) = \Phi\left(-\frac{\boldsymbol{\mu}^\top \boldsymbol{h} - \epsilon\sqrt{d}\|\boldsymbol{h}\|_2}{\sqrt{\boldsymbol{h}^\top \boldsymbol{\Sigma} \boldsymbol{h}}}\right), \tag{A6}$$

where Φ denotes the cumulative distribution function of the standard normal distribution.

Proof. For every perturbation $\boldsymbol{\Delta} \in \mathbb{R}^d$, the signed margin is

$$y g_{\boldsymbol{h}}(\boldsymbol{x} + \boldsymbol{\Delta}) = y\boldsymbol{h}^\top \boldsymbol{x} + y\boldsymbol{h}^\top \boldsymbol{\Delta}. \tag{A7}$$

Cauchy–Schwarz shows that its minimum over the perturbation ball is

$$\inf_{\|\boldsymbol{\Delta}\|_2 \leq \epsilon\sqrt{d}} y g_{\boldsymbol{h}}(\boldsymbol{x} + \boldsymbol{\Delta}) = y\boldsymbol{h}^\top \boldsymbol{x} - \epsilon\sqrt{d}\|\boldsymbol{h}\|_2. \tag{A8}$$

The class-conditional model implies

$$y\boldsymbol{h}^\top \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}^\top \boldsymbol{h}, \boldsymbol{h}^\top \boldsymbol{\Sigma} \boldsymbol{h}). \tag{A9}$$

Since $\boldsymbol{h} \neq \boldsymbol{0}_d$ and $\boldsymbol{\Sigma} \succ \boldsymbol{0}_{d,d}$, the variance $\boldsymbol{h}^\top \boldsymbol{\Sigma} \boldsymbol{h}$ is strictly positive. Therefore, standardizing the Gaussian signed margin gives

$$\mathbb{P}\left(\inf_{\|\boldsymbol{\Delta}\|_2 \leq \epsilon\sqrt{d}} y g_{\boldsymbol{h}}(\boldsymbol{x} + \boldsymbol{\Delta}) \leq 0\right) \tag{A10}$$

$$= \mathbb{P}(y\boldsymbol{h}^\top \boldsymbol{x} \leq \epsilon\sqrt{d}\|\boldsymbol{h}\|_2) \tag{A11}$$

$$= \mathbb{P}_{z \sim \mathcal{N}(0,1)}\left(z \leq \frac{\epsilon\sqrt{d}\|\boldsymbol{h}\|_2 - \boldsymbol{\mu}^\top \boldsymbol{h}}{\sqrt{\boldsymbol{h}^\top \boldsymbol{\Sigma} \boldsymbol{h}}}\right) \tag{A12}$$

$$= \Phi\left(-\frac{\boldsymbol{\mu}^\top \boldsymbol{h} - \epsilon\sqrt{d}\|\boldsymbol{h}\|_2}{\sqrt{\boldsymbol{h}^\top \boldsymbol{\Sigma} \boldsymbol{h}}}\right). \tag{A13}$$

□

Lemma B.2 (Bayes error for the identity-basis task)**.** *Suppose that y is uniform on $\{\pm 1\}$ and $\boldsymbol{x} \mid y \sim \mathcal{N}(y\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} \neq \mathbf{0}_d$ and $\boldsymbol{\Sigma} \succ \mathbf{0}_{d,d}$. Then, for every measurable function $\varphi : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathbb{P}(y\varphi(\boldsymbol{x}) \leq 0) \geq \Phi\Big(-\sqrt{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}}\Big), \tag{A14}$$

where Φ denotes the cumulative distribution function of the standard normal distribution.

Proof. **Lower bound.** Let p_+ and p_- denote the conditional densities of $\boldsymbol{x}$ given $y = +1$ and $y = -1$, respectively. Since the two labels have equal probability, conditioning on y gives

$$\mathbb{P}(y\varphi(\boldsymbol{x}) \leq 0) = \frac{1}{2}\int_{\boldsymbol{x}\in\mathbb{R}^d} [\mathbb{1}\{\varphi(\boldsymbol{x}) \leq 0\}p_+(\boldsymbol{x}) + \mathbb{1}\{\varphi(\boldsymbol{x}) \geq 0\}p_-(\boldsymbol{x})]d\boldsymbol{x}. \tag{A15}$$

For every $\boldsymbol{x} \in \mathbb{R}^d$, the integrand satisfies

$$\mathbb{1}\{\varphi(\boldsymbol{x}) \leq 0\}p_+(\boldsymbol{x}) + \mathbb{1}\{\varphi(\boldsymbol{x}) \geq 0\}p_-(\boldsymbol{x}) \geq \min\{p_+(\boldsymbol{x}), p_-(\boldsymbol{x})\}. \tag{A16}$$

Consequently,

$$\mathbb{P}(y\varphi(\boldsymbol{x}) \leq 0) \geq \frac{1}{2}\int_{\boldsymbol{x}\in\mathbb{R}^d} \min\{p_+(\boldsymbol{x}), p_-(\boldsymbol{x})\}d\boldsymbol{x}. \tag{A17}$$

Function that attains the lower bound. The Gaussian density formula gives

$$\log\left(\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \frac{1}{2}(\boldsymbol{x}+\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}+\boldsymbol{\mu}) = 2\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x}. \tag{A18}$$

Therefore, the linear function $\varphi^\star(\boldsymbol{x}) := \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x}$ is positive exactly when $p_+(\boldsymbol{x}) > p_-(\boldsymbol{x})$ and negative exactly when $p_+(\boldsymbol{x}) < p_-(\boldsymbol{x})$. Because $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \neq \mathbf{0}_d$, the hyperplane on which $\varphi^\star(\boldsymbol{x}) = 0$ has Lebesgue measure zero. Thus,

$$\frac{1}{2}\int_{\boldsymbol{x}\in\mathbb{R}^d} \min\{p_+(\boldsymbol{x}), p_-(\boldsymbol{x})\}d\boldsymbol{x} \tag{A19}$$

$$= \frac{1}{2}\int_{\boldsymbol{x}\in\mathbb{R}^d} [\mathbb{1}\{\varphi^\star(\boldsymbol{x}) \leq 0\}p_+(\boldsymbol{x}) + \mathbb{1}\{\varphi^\star(\boldsymbol{x}) \geq 0\}p_-(\boldsymbol{x})]d\boldsymbol{x} \tag{A20}$$

$$= \mathbb{P}(y\varphi^\star(\boldsymbol{x}) \leq 0). \tag{A21}$$

Value of the lower bound. The class-conditional model and the definition of $\varphi^\star$ imply

$$y\varphi^\star(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}). \tag{A22}$$

The mean and variance are both equal to $\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, which is strictly positive because $\boldsymbol{\mu} \neq \mathbf{0}_d$ and $\boldsymbol{\Sigma} \succ \mathbf{0}_{d,d}$. Standardizing therefore yields

$$\mathbb{P}(y\varphi^\star(\boldsymbol{x}) \leq 0) = \Phi\left(-\frac{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{\sqrt{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}}\right) = \Phi\Big(-\sqrt{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}\Big). \tag{A23}$$

□

Lemma B.3 (Robust Bayes characterization for the identity-basis task)**.** *Suppose that y is uniform on $\{\pm 1\}$ and $\boldsymbol{x} \mid y \sim \mathcal{N}(y\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} \neq \mathbf{0}_d$ and $\boldsymbol{\Sigma} \succ \mathbf{0}_{d,d}$. Suppose also that $0 \leq \epsilon\sqrt{d} < \|\boldsymbol{\mu}\|_2$. There exists a unique $\tau \geq 0$ satisfying*

$$\tau\|(\boldsymbol{\Sigma} + \tau\boldsymbol{I}_d)^{-1}\boldsymbol{\mu}\|_2 = \epsilon\sqrt{d}. \tag{A24}$$

For this τ, define

$$\boldsymbol{h}_\tau := (\boldsymbol{\Sigma} + \tau \boldsymbol{I}_d)^{-1} \boldsymbol{\mu}. \tag{A25}$$

Then, for every measurable function $\varphi : \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{P}\left(\inf_{\|\boldsymbol{\Delta}\|_2 \le \epsilon\sqrt{d}} y\varphi(\boldsymbol{x} + \boldsymbol{\Delta}) \le 0 \right) \ge \Phi\left(-\sqrt{\boldsymbol{h}_\tau^\top \boldsymbol{\Sigma} \boldsymbol{h}_\tau} \right). \tag{A26}$$

Among linear functions, equality is attained exactly by the functions $\boldsymbol{x} \mapsto \omega \boldsymbol{h}_\tau^\top \boldsymbol{x}$ with $\omega > 0$.

Proof. **Existence and uniqueness of τ.** Let $\lambda_1, \dots, \lambda_d > 0$ be the eigenvalues of $\boldsymbol{\Sigma}$, let $\boldsymbol{u}_1, \dots, \boldsymbol{u}_d$ be corresponding orthonormal eigenvectors, and write $\boldsymbol{\mu} = \sum_{i=1}^d m_i \boldsymbol{u}_i$. For $\tau \ge 0$, the spectral theorem gives

$$q(\tau) := \tau^2 \|(\boldsymbol{\Sigma} + \tau \boldsymbol{I}_d)^{-1} \boldsymbol{\mu}\|_2^2 = \sum_{i=1}^d m_i^2 \left(\frac{\tau}{\lambda_i + \tau} \right)^2. \tag{A27}$$

Since both sides of the equation defining τ are nonnegative, that equation is equivalent to $q(\tau) = \epsilon^2 d$. Each map $\tau \mapsto \tau/(\lambda_i + \tau)$ is continuous and strictly increasing from zero to one on $[0, \infty)$. Because $\boldsymbol{\mu} \ne \boldsymbol{0}_d$, at least one m_i is nonzero, and hence q is continuous and strictly increasing with

$$q(0) = 0, \qquad \lim_{\tau \to \infty} q(\tau) = \sum_{i=1}^d m_i^2 = \|\boldsymbol{\mu}\|_2^2. \tag{A28}$$

If $\epsilon = 0$, the equation $q(\tau) = \epsilon^2 d$ has the unique solution $\tau = 0$. If $\epsilon > 0$, the inequality $0 < \epsilon^2 d < \|\boldsymbol{\mu}\|_2^2$ and the intermediate value theorem give a unique solution $\tau > 0$.

Lower bound. For the resulting $\boldsymbol{h}_\tau$, define the label-dependent perturbation $\boldsymbol{\Delta}_y := -y\tau \boldsymbol{h}_\tau$. This perturbation is admissible because

$$\|\boldsymbol{\Delta}_y\|_2 = \tau \|\boldsymbol{h}_\tau\|_2 = \epsilon\sqrt{d}. \tag{A29}$$

Therefore, for every measurable function φ,

$$\mathbb{P}\left(\inf_{\|\boldsymbol{\Delta}\|_2 \le \epsilon\sqrt{d}} y\varphi(\boldsymbol{x} + \boldsymbol{\Delta}) \le 0 \right) \ge \mathbb{P}(y\varphi(\boldsymbol{x} + \boldsymbol{\Delta}_y) \le 0). \tag{A30}$$

The definition of $\boldsymbol{h}_\tau$ implies

$$\boldsymbol{\mu} - \tau \boldsymbol{h}_\tau = \boldsymbol{\Sigma} \boldsymbol{h}_\tau. \tag{A31}$$

Consequently, the perturbed input has the class-conditional distribution

$$\boldsymbol{x} + \boldsymbol{\Delta}_y \mid y \sim \mathcal{N}(y\boldsymbol{\Sigma}\boldsymbol{h}_\tau, \boldsymbol{\Sigma}). \tag{A32}$$

Since $\boldsymbol{\Sigma}\boldsymbol{h}_\tau \ne \boldsymbol{0}_d$, applying Lemma B.2 to this symmetric Gaussian mixture gives

$$\mathbb{P}(y\varphi(\boldsymbol{x} + \boldsymbol{\Delta}_y) \le 0) \ge \Phi\left(-\sqrt{(\boldsymbol{\Sigma}\boldsymbol{h}_\tau)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\Sigma}\boldsymbol{h}_\tau)} \right) = \Phi\left(-\sqrt{\boldsymbol{h}_\tau^\top \boldsymbol{\Sigma} \boldsymbol{h}_\tau} \right). \tag{A33}$$

Attainment of the lower bound. Applying Eq. (A6) with $\boldsymbol{h} = \boldsymbol{h}_\tau$ to the linear function $\varphi^\star(\boldsymbol{x}) = \boldsymbol{h}_\tau^\top \boldsymbol{x}$ gives

$$\mathbb{P}\left(\inf_{\|\boldsymbol{\Delta}\|_2 \le \epsilon\sqrt{d}} y\varphi^\star(\boldsymbol{x} + \boldsymbol{\Delta}) \le 0 \right) = \Phi\left(-\frac{\boldsymbol{\mu}^\top \boldsymbol{h}_\tau - \epsilon\sqrt{d}\|\boldsymbol{h}_\tau\|_2}{\sqrt{\boldsymbol{h}_\tau^\top \boldsymbol{\Sigma} \boldsymbol{h}_\tau}} \right). \tag{A34}$$

The definitions of τ and $\boldsymbol{h}_\tau$ give

$$\boldsymbol{\mu}^\top \boldsymbol{h}_\tau - \epsilon\sqrt{d}\|\boldsymbol{h}_\tau\|_2 = \boldsymbol{h}_\tau^\top (\boldsymbol{\Sigma} + \tau \boldsymbol{I}_d) \boldsymbol{h}_\tau - \tau \|\boldsymbol{h}_\tau\|_2^2 = \boldsymbol{h}_\tau^\top \boldsymbol{\Sigma} \boldsymbol{h}_\tau. \tag{A35}$$

Substitution into the preceding robust-error formula yields

$$\mathbb{P}\left(\inf_{\|\boldsymbol{\Delta}\|_2 \le \epsilon\sqrt{d}} y\varphi^\star(\boldsymbol{x} + \boldsymbol{\Delta}) \le 0 \right) = \Phi\left(-\frac{\boldsymbol{h}_\tau^\top \boldsymbol{\Sigma} \boldsymbol{h}_\tau}{\sqrt{\boldsymbol{h}_\tau^\top \boldsymbol{\Sigma} \boldsymbol{h}_\tau}} \right) = \Phi\left(-\sqrt{\boldsymbol{h}_\tau^\top \boldsymbol{\Sigma} \boldsymbol{h}_\tau} \right). \tag{A36}$$

Thus, $\varphi^\star$ attains the lower bound.

Uniqueness among linear functions. Positive scale invariance in Eq. (A6) shows that every positive multiple of $\boldsymbol{h}_\tau$ attains the same lower bound. Fix $\boldsymbol{v} \in \mathbb{R}^d \setminus \{\mathbf{0}_d\}$. The identities $\boldsymbol{\mu} = (\boldsymbol{\Sigma} + \tau \boldsymbol{I}_d)\boldsymbol{h}_\tau$ and $\epsilon\sqrt{d} = \tau\|\boldsymbol{h}_\tau\|_2$, followed by the Euclidean Cauchy–Schwarz inequality and the Cauchy–Schwarz inequality for the inner product induced by $\boldsymbol{\Sigma}$, give

$$\boldsymbol{\mu}^\top \boldsymbol{v} - \epsilon\sqrt{d}\|\boldsymbol{v}\|_2 = \boldsymbol{h}_\tau^\top \boldsymbol{\Sigma}\boldsymbol{v} + \tau(\boldsymbol{h}_\tau^\top \boldsymbol{v} - \|\boldsymbol{h}_\tau\|_2\|\boldsymbol{v}\|_2) \tag{A37}$$

$$\leq \boldsymbol{h}_\tau^\top \boldsymbol{\Sigma}\boldsymbol{v} \tag{A38}$$

$$\leq \sqrt{\boldsymbol{h}_\tau^\top \boldsymbol{\Sigma}\boldsymbol{h}_\tau}\sqrt{\boldsymbol{v}^\top \boldsymbol{\Sigma}\boldsymbol{v}}. \tag{A39}$$

The denominator $\sqrt{\boldsymbol{v}^\top \boldsymbol{\Sigma}\boldsymbol{v}}$ is strictly positive because $\boldsymbol{\Sigma} \succ \mathbf{0}_{d,d}$. Dividing by this denominator and using the strict monotonicity of Φ in Eq. (A6) show that the robust error of $g_{\boldsymbol{v}}$ is at least the lower bound. The robust error equals the lower bound only if both inequalities in the preceding display hold with equality. If $\tau > 0$, equality in the first inequality requires $\boldsymbol{h}_\tau^\top \boldsymbol{v} = \|\boldsymbol{h}_\tau\|_2\|\boldsymbol{v}\|_2$, which holds exactly when $\boldsymbol{v} = \omega \boldsymbol{h}_\tau$ for some $\omega > 0$. If $\tau = 0$, the first inequality is automatically an equality, while equality in the $\boldsymbol{\Sigma}$-inner-product Cauchy–Schwarz inequality again holds exactly when $\boldsymbol{v} = \omega \boldsymbol{h}_\tau$ for some $\omega > 0$. Conversely, every such positive multiple attains equality, as observed above. Finally, $g_{\mathbf{0}_d}$ is identically zero and therefore has robust error one, whereas the lower bound is strictly less than one half because $\boldsymbol{h}_\tau^\top \boldsymbol{\Sigma}\boldsymbol{h}_\tau > 0$. □

Lemma B.4 (Taskwise robust Bayes characterization)**.** *Suppose that (a) and (b) in Assumption 3.1 hold. For every $\boldsymbol{U} \in \mathcal{O}(d)$ and $\boldsymbol{h} \in \mathbb{R}^d$, define the linear function $g_{\boldsymbol{U},\boldsymbol{h}} : \mathbb{R}^d \to \mathbb{R}$ by*

$$g_{\boldsymbol{U},\boldsymbol{h}}(\boldsymbol{x}) := \boldsymbol{h}^\top \boldsymbol{U}^\top \boldsymbol{x}.$$

(a) For every $\boldsymbol{U} \in \mathcal{O}(d)$ and $\boldsymbol{h} \in \mathbb{R}^d \setminus \{\mathbf{0}_d\}$,

$$\mathcal{E}_{\epsilon,\boldsymbol{U}}(g_{\boldsymbol{U},\boldsymbol{h}}) = \Phi\left(-\frac{\boldsymbol{\mu}^\top \boldsymbol{h} - \epsilon\sqrt{d}\|\boldsymbol{h}\|_2}{\sqrt{\boldsymbol{h}^\top \boldsymbol{\Sigma}\boldsymbol{h}}}\right), \tag{A40}$$

(b) For every $\boldsymbol{U} \in \mathcal{O}(d)$,

$$\mathcal{E}_{\epsilon,\boldsymbol{U}}(g_{\boldsymbol{U},\boldsymbol{h}_\tau}) = \Phi\left(-\sqrt{\boldsymbol{h}_\tau^\top \boldsymbol{\Sigma}\boldsymbol{h}_\tau}\right) = \mathcal{E}^\star_{\epsilon,\boldsymbol{U}}. \tag{A41}$$

Among linear functions, the robust Bayes error is attained exactly by the functions $\boldsymbol{x} \mapsto \omega \boldsymbol{h}_\tau^\top \boldsymbol{U}^\top \boldsymbol{x}$ with $\omega > 0$.

Proof. Define $\boldsymbol{z} := \boldsymbol{U}^\top \boldsymbol{x}$ and $\widetilde{\boldsymbol{\Delta}} := \boldsymbol{U}^\top \boldsymbol{\Delta}$. Conditional on $\boldsymbol{U}$, the transformed input satisfies

$$\boldsymbol{z} \mid y \sim \mathcal{N}(y\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{A42}$$

Orthogonality of $\boldsymbol{U}$ gives

$$\|\widetilde{\boldsymbol{\Delta}}\|_2 = \|\boldsymbol{\Delta}\|_2, \qquad g_{\boldsymbol{U},\boldsymbol{h}}(\boldsymbol{x} + \boldsymbol{\Delta}) = g_{\boldsymbol{h}}(\boldsymbol{z} + \widetilde{\boldsymbol{\Delta}}). \tag{A43}$$

The map $\boldsymbol{\Delta} \mapsto \widetilde{\boldsymbol{\Delta}}$ is a bijection of the perturbation ball onto itself.

Part (a). Applying Eq. (A6) to $g_{\boldsymbol{h}}$ in the transformed coordinates proves Eq. (A40).

Part (b). For every measurable function $\varphi : \mathbb{R}^d \to \mathbb{R}$, define $\widetilde{\varphi} : \mathbb{R}^d \to \mathbb{R}$ by $\widetilde{\varphi}(\boldsymbol{z}) := \varphi(\boldsymbol{U}\boldsymbol{z})$. The map $\varphi \mapsto \widetilde{\varphi}$ is a bijection on the class of measurable functions, and the preceding change of coordinates gives

$$\mathcal{E}_{\epsilon,\boldsymbol{U}}(\varphi) = \mathbb{P}\left(\inf_{\|\widetilde{\boldsymbol{\Delta}}\|_2 \leq \epsilon\sqrt{d}} y\widetilde{\varphi}(\boldsymbol{z} + \widetilde{\boldsymbol{\Delta}}) \leq 0\right). \tag{A44}$$

Applying Lemma B.3 in the transformed coordinates proves Eq. (A41). Its uniqueness conclusion for linear functions proves the final statement of part (b). □

B.2 Query-Linear Representation

Define the population second moment by

$$\boldsymbol{S} := \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top. \tag{A45}$$

The latent finite-context matrix and its population limit are

$$\boldsymbol{G}_{a,N} := \begin{bmatrix} N^{-1}\sum_{n=1}^{N} \boldsymbol{a}_n \boldsymbol{a}_n^\top & N^{-1}\sum_{n=1}^{N} \boldsymbol{a}_n \\ N^{-1}\sum_{n=1}^{N} \boldsymbol{a}_n^\top & 1 \end{bmatrix}, \qquad \boldsymbol{G}_\star := \begin{bmatrix} \boldsymbol{S} & \boldsymbol{\mu} \\ \boldsymbol{\mu}^\top & 1 \end{bmatrix}. \tag{A46}$$

The strong law gives $\boldsymbol{G}_{a,N} \to \boldsymbol{G}_\star$ almost surely. The corresponding ambient context matrix is

$$\boldsymbol{G}_N(\boldsymbol{U}) := \frac{1}{N}\boldsymbol{Z}^{(0)}\boldsymbol{M}\boldsymbol{Z}^{(0)\top} = \mathrm{diag}(\boldsymbol{U}, 1)\boldsymbol{G}_{a,N}\,\mathrm{diag}(\boldsymbol{U}^\top, 1). \tag{A47}$$

For a context matrix $\boldsymbol{G}$, define $\boldsymbol{A}^{(0)}_{\boldsymbol{\Theta}_L}(\boldsymbol{G}) := \boldsymbol{I}_{d+1}$ and, for $\ell \in [L]$, recursively define

$$\boldsymbol{A}^{(\ell)}_{\boldsymbol{\Theta}_L}(\boldsymbol{G}) := \{\boldsymbol{I}_{d+1} + \boldsymbol{P}_\ell \boldsymbol{A}^{(\ell-1)}_{\boldsymbol{\Theta}_L}(\boldsymbol{G})\boldsymbol{G}\boldsymbol{A}^{(\ell-1)}_{\boldsymbol{\Theta}_L}(\boldsymbol{G})^\top \boldsymbol{Q}_\ell\}\boldsymbol{A}^{(\ell-1)}_{\boldsymbol{\Theta}_L}(\boldsymbol{G}). \tag{A48}$$

For any context matrix $\boldsymbol{G}$, define the depth-L effective direction by

$$\boldsymbol{\theta}_{\boldsymbol{\Theta}_L}(\boldsymbol{G}) := [\boldsymbol{I}_d \quad \boldsymbol{0}_d]\, \boldsymbol{A}^{(L)}_{\boldsymbol{\Theta}_L}(\boldsymbol{G})^\top \boldsymbol{e}_{d+1}. \tag{A49}$$

Here, $\boldsymbol{e}_{d+1}$ is the last standard basis vector of $\mathbb{R}^{d+1}$. The mask in Eq. (4) removes the query column from every context matrix. Induction over the layers therefore shows that every depth-L transformer is linear in its query:

$$f(\boldsymbol{x} + \boldsymbol{\Delta}; \boldsymbol{\Theta}_L, \mathcal{D}_N) = \boldsymbol{\theta}_{\boldsymbol{\Theta}_L}(\boldsymbol{G}_N(\boldsymbol{U}))^\top(\boldsymbol{x} + \boldsymbol{\Delta}). \tag{A50}$$

B.3 Krylov structure and sufficient-depth realization

Define the feature and label projectors by

$$\boldsymbol{\Pi}_x := \begin{bmatrix} \boldsymbol{I}_d & \boldsymbol{0}_d \\ \boldsymbol{0}_d^\top & 0 \end{bmatrix}, \qquad \boldsymbol{\Pi}_y := \begin{bmatrix} \boldsymbol{0}_{d,d} & \boldsymbol{0}_d \\ \boldsymbol{0}_d^\top & 1 \end{bmatrix}. \tag{A51}$$

We first relate the active spectra of $\boldsymbol{\Sigma}$ and $\boldsymbol{S}$.

Lemma B.5 (Active spectrum under the rank-one update)**.** *The matrix $\boldsymbol{S}$ has exactly r distinct active eigenvalues, where an eigenvalue s is active when its orthogonal spectral projector $\boldsymbol{F}_s$ satisfies $\boldsymbol{F}_s\boldsymbol{\mu} \neq \boldsymbol{0}_d$ and r is defined in Eq. (2).*

Proof. Since $\boldsymbol{S}\boldsymbol{v} = \boldsymbol{\Sigma}\boldsymbol{v} + \boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{v}$, induction in both directions gives

$$\mathrm{span}\{\boldsymbol{\mu}, \boldsymbol{S}\boldsymbol{\mu}, \boldsymbol{S}^2\boldsymbol{\mu}, \ldots\} = \mathrm{span}\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\boldsymbol{\mu}, \boldsymbol{\Sigma}^2\boldsymbol{\mu}, \ldots\}. \tag{A52}$$

The spectral decomposition of $\boldsymbol{\Sigma}$ and the Vandermonde determinant show that the right-hand side has dimension r, the number of distinct covariance eigenvalues with nonzero projection of $\boldsymbol{\mu}$. For any symmetric matrix, the dimension of the Krylov space generated by a vector equals the number of distinct eigenvalues whose spectral projectors do not annihilate that vector. Applying this fact to $\boldsymbol{S}$ proves the claim. □

We characterize the population output induced by block-selective parameters. We then choose their scalar coefficients to interpolate the desired response on the active spectrum.

Lemma B.6 (Population response of block-selective processing layers)**.** *Fix an integer $m \geq 0$. For $1 \leq \ell \leq m$, let $\kappa_\ell, \eta_\ell \in \mathbb{R}$ and define*

$$\boldsymbol{P}_\ell := \kappa_\ell \boldsymbol{\Pi}_x + \eta_\ell \boldsymbol{\Pi}_y, \qquad \boldsymbol{Q}_\ell := \boldsymbol{\Pi}_x. \tag{A53}$$

Set $\boldsymbol{\Theta}_m := \{(\boldsymbol{P}_\ell, \boldsymbol{Q}_\ell)\}_{\ell=1}^m$. *Set* $a_0(s) := 1$ *and* $b_0(s) := 0$, *and, for* $1 \le \ell \le m$, *recursively define*

$$a_\ell(s) := a_{\ell-1}(s)\{1 + \kappa_\ell s a_{\ell-1}(s)^2\}, \qquad b_\ell(s) := b_{\ell-1}(s) + \eta_\ell a_{\ell-1}(s)^2\{1 + s b_{\ell-1}(s)\}. \tag{A54}$$

Then,

$$\boldsymbol{A}_{\boldsymbol{\Theta}_m}^{(m)}(\boldsymbol{G}_\star) = \begin{bmatrix} a_m(\boldsymbol{S}) & \boldsymbol{0}_d \\ \boldsymbol{\mu}^\top b_m(\boldsymbol{S}) & 1 \end{bmatrix}. \tag{A55}$$

Consequently,

$$\boldsymbol{\theta}_{\boldsymbol{\Theta}_m}(\boldsymbol{G}_\star) = b_m(\boldsymbol{S})\boldsymbol{\mu}. \tag{A56}$$

Proof. The claim is immediate at $m = 0$. Suppose that the claimed block form holds after ℓ layers, where $0 \le \ell < m$, and denote this cumulative transformation by $\boldsymbol{T}_\ell$. Direct block multiplication gives

$$\boldsymbol{T}_\ell \boldsymbol{G}_\star \boldsymbol{T}_\ell^\top \boldsymbol{\Pi}_x = \begin{bmatrix} \boldsymbol{S} a_\ell(\boldsymbol{S})^2 & \boldsymbol{0}_d \\ \boldsymbol{\mu}^\top\{\boldsymbol{I}_d + b_\ell(\boldsymbol{S})\boldsymbol{S}\} a_\ell(\boldsymbol{S}) & 0 \end{bmatrix}. \tag{A57}$$

Substituting this expression and the $(\ell+1)$-st parameters from Eq. (A53) into Eq. (A48) gives the recursion in Eq. (A54) and proves Eq. (A55) by induction. The definition in Eq. (A49) gives $\boldsymbol{\theta}_{\boldsymbol{\Theta}_m}(\boldsymbol{G}_\star) = b_m(\boldsymbol{S})\boldsymbol{\mu}$. □

Lemma B.7 (Sufficient-depth population optimal-direction comparator)**.** *Fix* $\tau \ge 0$ *and* $\omega > 0$, *define* $\boldsymbol{h}^\star := \omega(\boldsymbol{\Sigma} + \tau \boldsymbol{I}_d)^{-1}\boldsymbol{\mu}$, *and suppose that* $L \ge r$. *There is a parameter tuple* $\boldsymbol{\Theta}_{L,\star}$ *such that, for every* $\boldsymbol{U} \in \mathcal{O}(d)$,

$$\boldsymbol{U}^\top \boldsymbol{\theta}_{\boldsymbol{\Theta}_{L,\star}}(\mathrm{diag}(\boldsymbol{U}, 1)\boldsymbol{G}_\star \,\mathrm{diag}(\boldsymbol{U}^\top, 1)) = \boldsymbol{h}^\star. \tag{A58}$$

Proof. Let $s_1, \dots, s_r$ be the active eigenvalues of $\boldsymbol{S}$, whose number is r by Lemma B.5, and set

$$C := \omega\{1 + \boldsymbol{\mu}^\top(\boldsymbol{\Sigma} + \tau \boldsymbol{I}_d)^{-1}\boldsymbol{\mu}\}. \tag{A59}$$

It is enough to construct the first r layers in latent coordinates at the population context $\boldsymbol{G}_\star$. For the first $r-1$ layers, use the notation and response recursion of Lemma B.6 with $m = r - 1$.

These layers process $r-1$ active eigenvalues sequentially. At the beginning of stage ℓ, let $\mathcal{I}_\ell$ be the indices not yet processed and define

$$z_i^{(\ell)} := s_i a_\ell(s_i)^2, \qquad i \in \mathcal{I}_\ell. \tag{A60}$$

We maintain the invariant that these numbers are positive and pairwise distinct. This holds initially because $z_i^{(0)} = s_i$. Let k_ℓ be the unique index attaining their minimum and set

$$\kappa_{\ell+1} := -\frac{1}{z_{k_\ell}^{(\ell)}}, \qquad \eta_{\ell+1} := \frac{c/(s_{k_\ell} + \tau) - b_\ell(s_{k_\ell})}{a_\ell(s_{k_\ell})^2\{1 + s_{k_\ell} b_\ell(s_{k_\ell})\}}, \tag{A61}$$

where $c > 0$ will be chosen sufficiently small. Equation Eq. (A54) then gives

$$a_{\ell+1}(s_{k_\ell}) = 0, \qquad b_{\ell+1}(s_{k_\ell}) = \frac{c}{s_{k_\ell} + \tau}. \tag{A62}$$

Every later increment of either response at this eigenvalue contains a factor $a_j(s_{k_\ell})$, so the response remains fixed. For every remaining index $i \ne k_\ell$, the next transformed value is

$$z_i^{(\ell+1)} = z_i^{(\ell)}\left(1 - \frac{z_i^{(\ell)}}{z_{k_\ell}^{(\ell)}}\right)^2. \tag{A63}$$

The function $z \mapsto z(1 - z/z_{k_\ell}^{(\ell)})^2$ is positive and strictly increasing for $z > z_{k_\ell}^{(\ell)}$, because its derivative is $(3z/z_{k_\ell}^{(\ell)} - 1)(z/z_{k_\ell}^{(\ell)} - 1) > 0$. Thus, positivity and pairwise distinctness hold at the next stage. When $c = 0$, induction gives $b_\ell(s_i) = 0$ and $\eta_\ell = 0$ at every stage. The finitely many responses and denominators in Eq. (A61) depend continuously on c near zero, so some sufficiently small $c > 0$ makes every denominator nonzero.

Suppose first that $r \geq 2$, and let $k_\star$ denote the single index left after these $r-1$ stages. Let $\boldsymbol{T}_{r-1} := \boldsymbol{A}^{(r-1)}_{\boldsymbol{\Theta}_{r-1}}(\boldsymbol{G}_\star)$ and define

$$g_{r-1} := \boldsymbol{e}_{d+1}^\top \boldsymbol{T}_{r-1} \boldsymbol{G}_\star \boldsymbol{T}_{r-1}^\top \boldsymbol{e}_{d+1} > 0. \tag{A64}$$

The strict inequality holds because the Schur complement of the bottom-right entry of $\boldsymbol{G}_\star$ is $\boldsymbol{\Sigma} \succ \boldsymbol{0}_{d,d}$ and $\boldsymbol{T}_{r-1}^\top \boldsymbol{e}_{d+1}$ has final coordinate one. In the final layer, choose

$$\boldsymbol{P}_r := \boldsymbol{\Pi}_y, \qquad \boldsymbol{Q}_r := \alpha \boldsymbol{\Pi}_x + \beta \boldsymbol{\Pi}_y. \tag{A65}$$

Direct block multiplication gives $\boldsymbol{\theta}_{\boldsymbol{\Theta}_r}(\boldsymbol{G}_\star) = b_r(\boldsymbol{S})\boldsymbol{\mu}$, where

$$b_r(s) := (1 + \beta g_{r-1}) b_{r-1}(s) + \alpha a_{r-1}(s)^2 \{1 + s b_{r-1}(s)\}. \tag{A66}$$

Shrinking $c > 0$ if necessary ensures that $1 + s_{k_\star} b_{r-1}(s_{k_\star}) \neq 0$, because this expression is continuous in c and equals one at $c = 0$. Set

$$\beta := \frac{C/c - 1}{g_{r-1}}, \qquad \alpha := \frac{C/(s_{k_\star} + \tau) - (C/c) b_{r-1}(s_{k_\star})}{a_{r-1}(s_{k_\star})^2 \{1 + s_{k_\star} b_{r-1}(s_{k_\star})\}}. \tag{A67}$$

The denominator defining α is therefore nonzero because the induction invariant also gives $a_{r-1}(s_{k_\star}) \neq 0$. Equations Eqs. (A62), (A66) and (A67) imply

$$b_r(s_i) = \frac{C}{s_i + \tau}, \qquad i \in [r]. \tag{A68}$$

If $r = 1$, the same conclusion follows directly by choosing $\boldsymbol{P}_1 := \boldsymbol{\Pi}_y$ and $\boldsymbol{Q}_1 := C(s_1 + \tau)^{-1} \boldsymbol{\Pi}_x$. The spectral theorem and activity of $s_1, \ldots, s_r$ give

$$b_r(\boldsymbol{S})\boldsymbol{\mu} = C(\boldsymbol{S} + \tau \boldsymbol{I}_d)^{-1} \boldsymbol{\mu}, \tag{A69}$$

and the Sherman–Morrison identity gives

$$C(\boldsymbol{S} + \tau \boldsymbol{I}_d)^{-1} \boldsymbol{\mu} = \omega (\boldsymbol{\Sigma} + \tau \boldsymbol{I}_d)^{-1} \boldsymbol{\mu} = \boldsymbol{h}^\star. \tag{A70}$$

Every constructed parameter commutes with $\mathrm{diag}(\boldsymbol{U}, 1)$, so rotation equivariance gives Eq. (A58) at depth r. If $L > r$, set $\boldsymbol{P}_\ell = \boldsymbol{Q}_\ell = \boldsymbol{0}_{d+1,d+1}$ for $r < \ell \leq L$. These additional zero layers leave the effective direction unchanged, completing the proof for every $L \geq r$. □

We finally transfer the constructed comparator from the population context to finite contexts.

Lemma B.8 (Finite-context accuracy of the sufficient-depth comparator)**.** *Under the conditions of Lemma B.7, let $\boldsymbol{\Theta}_{L,\star}$ and $\boldsymbol{h}^\star$ be the comparator and target direction constructed there. There is a constant $C_{\mathrm{cmp}} > 0$ such that, for every $\boldsymbol{U} \in \mathcal{O}(d)$ and $N \in \mathbb{N}$,*

$$\mathbb{E}_{\mathcal{D}_N | \boldsymbol{U}} [\| \boldsymbol{U}^\top \boldsymbol{\theta}_{\boldsymbol{\Theta}_{L,\star}}(\boldsymbol{G}_N(\boldsymbol{U})) - \boldsymbol{h}^\star \|_2^2] \leq \frac{C_{\mathrm{cmp}}}{N}. \tag{A71}$$

Proof. Conditional on any fixed $\boldsymbol{U}$, induction in Eq. (A48) gives

$$\boldsymbol{U}^\top \boldsymbol{\theta}_{\boldsymbol{\Theta}_{L,\star}}(\boldsymbol{G}_N(\boldsymbol{U})) = \boldsymbol{\theta}_{\boldsymbol{\Theta}_{L,\star}}(\boldsymbol{G}_{a,N}). \tag{A72}$$

The right-hand side is a fixed polynomial in the entries of $N^{-1} \sum_{n=1}^N \boldsymbol{a}_n$ and $N^{-1} \sum_{n=1}^N \boldsymbol{a}_n \boldsymbol{a}_n^\top$ whose value at $(\boldsymbol{\mu}, \boldsymbol{S})$ is $\boldsymbol{h}^\star$. For every fixed integer $k \geq 1$, Gaussian sample-moment bounds give

$$\mathbb{E}_{\boldsymbol{a}_{1:N}} \left[\| \frac{1}{N} \sum_{n=1}^N \boldsymbol{a}_n - \boldsymbol{\mu} \|_2^{2k} + \| \frac{1}{N} \sum_{n=1}^N \boldsymbol{a}_n \boldsymbol{a}_n^\top - \boldsymbol{S} \|_{\mathrm{F}}^{2k} \right] \leq \frac{C_k}{N^k}, \qquad N \in \mathbb{N}, \tag{A73}$$

for a constant $C_k > 0$. Expanding the polynomial difference around $(\boldsymbol{\mu}, \boldsymbol{S})$ and applying these bounds to its finitely many monomials proves Eq. (A71). □

Lemma 3.3 (Proof sketch of Theorem 3.2)**.** *Suppose that (a) and (b) in Assumption 3.1 hold. Suppose also that $N \to \infty$; write $\mathcal{D}_\infty := \mathcal{D}_N$. Let $\tau \geq 0$ be the unique solution of*

$$\tau \|(\boldsymbol{\Sigma} + \tau \boldsymbol{I}_d)^{-1} \boldsymbol{\mu}\|_2 = \epsilon \sqrt{d}. \tag{9}$$

(Bayes classifier) *For every $\boldsymbol{U} \in \mathcal{O}(d)$,*

$$g(\boldsymbol{x}) := \{(\boldsymbol{\Sigma} + \tau \boldsymbol{I}_d)^{-1} \boldsymbol{\mu}\}^\top \boldsymbol{U}^\top \boldsymbol{x}, \qquad \mathcal{E}_{\epsilon,\boldsymbol{U}}(g) = \mathcal{E}^\star_{\epsilon,\boldsymbol{U}}. \tag{10}$$

(Krylov order of the optimal direction) *Let $\mathcal{K}_0(\boldsymbol{\Sigma}, \boldsymbol{\mu}) := \{\boldsymbol{0}\}$. Then,*

$$(\boldsymbol{\Sigma} + \tau \boldsymbol{I}_d)^{-1} \boldsymbol{\mu} \in \mathcal{K}_r(\boldsymbol{\Sigma}, \boldsymbol{\mu}) \setminus \mathcal{K}_{r-1}(\boldsymbol{\Sigma}, \boldsymbol{\mu}). \tag{11}$$

(Effective weights of a linear transformer) *There exists a parameter family $\{\boldsymbol{\Theta}_{L,\xi} : \xi \in \mathbb{R}\}$ whose corresponding effective weights $\boldsymbol{\theta}(\boldsymbol{\Theta}_{L,\xi})$ satisfy, for every $\xi \in \mathbb{R}$, $\boldsymbol{U} \in \mathcal{O}(d)$, and $\boldsymbol{x} \in \mathbb{R}^d$,*

$$f(\boldsymbol{x}; \boldsymbol{\Theta}_{L,\xi}, \mathcal{D}_\infty) = \boldsymbol{\theta}(\boldsymbol{\Theta}_{L,\xi})^\top \boldsymbol{U}^\top \boldsymbol{x}. \tag{12}$$

(Span of the effective weights) *For the same parameter family,*

$$\operatorname{span}\{\boldsymbol{\theta}(\boldsymbol{\Theta}_{L,\xi}) : \xi \in \mathbb{R}\} = \mathcal{K}_L(\boldsymbol{\Sigma}, \boldsymbol{\mu}). \tag{13}$$

Proof. Part (i) is the optimal-direction conclusion of Lemma B.4.

For part (ii), let $\lambda_1, \ldots, \lambda_r$ be the distinct active eigenvalues of $\boldsymbol{\Sigma}$. Polynomial interpolation gives a polynomial p of degree at most $r-1$ satisfying $p(\lambda_i) = (\lambda_i + \tau)^{-1}$ for every $i \in [r]$. The spectral theorem gives

$$(\boldsymbol{\Sigma} + \tau \boldsymbol{I}_d)^{-1} \boldsymbol{\mu} = p(\boldsymbol{\Sigma}) \boldsymbol{\mu} \in \mathcal{K}_r(\boldsymbol{\Sigma}, \boldsymbol{\mu}). \tag{A74}$$

The polynomial p has degree exactly $r-1$. This is immediate when $r = 1$. If $r \geq 2$ and p had smaller degree, then $(s+\tau)p(s) - 1$ would have degree at most $r-1$ and vanish at the r distinct points $\lambda_1, \ldots, \lambda_r$, so it would vanish identically. This contradicts its value -1 at $s = -\tau$. Since $\boldsymbol{\mu}, \boldsymbol{\Sigma}\boldsymbol{\mu}, \ldots, \boldsymbol{\Sigma}^{r-1}\boldsymbol{\mu}$ are linearly independent, the nonzero leading coefficient of p implies $p(\boldsymbol{\Sigma})\boldsymbol{\mu} \notin \mathcal{K}_{r-1}(\boldsymbol{\Sigma}, \boldsymbol{\mu})$, proving part (ii).

For parts (iii) and (iv), fix $t \in \mathbb{R}$ and define the shared parameter tuple $\boldsymbol{\Theta}_L(t)$ by

$$\boldsymbol{P}_\ell(t) := \boldsymbol{\Pi}_y, \qquad \boldsymbol{Q}_\ell(t) := t\boldsymbol{\Pi}_x, \qquad \ell \in [L]. \tag{A75}$$

These parameters do not depend on $\boldsymbol{U}$. Set $q_{0,t}(s) := 0$. Direct induction in Eq. (A48) at the latent population context $\boldsymbol{G}_\star$ gives

$$\boldsymbol{A}^{(\ell)}_{\boldsymbol{\Theta}_L(t)}(\boldsymbol{G}_\star) = \begin{bmatrix} \boldsymbol{I}_d & \boldsymbol{0}_d \\ \boldsymbol{\mu}^\top q_{\ell,t}(\boldsymbol{S}) & 1 \end{bmatrix}, \qquad q_{\ell,t}(s) = q_{\ell-1,t}(s) + t\{1 + s q_{\ell-1,t}(s)\}. \tag{A76}$$

Solving this scalar recursion yields

$$q_{L,t}(s) = \sum_{j=1}^{L} \binom{L}{j} t^j s^{j-1}. \tag{A77}$$

Since $\boldsymbol{\Pi}_x$ and $\boldsymbol{\Pi}_y$ commute with $\operatorname{diag}(\boldsymbol{U}, 1)$, rotation equivariance and Eq. (A49) give

$$\boldsymbol{\theta}_{\boldsymbol{\Theta}_L(t)}(\operatorname{diag}(\boldsymbol{U}, 1) \boldsymbol{G}_\star \operatorname{diag}(\boldsymbol{U}^\top, 1)) = \boldsymbol{U} q_{L,t}(\boldsymbol{S}) \boldsymbol{\mu}. \tag{A78}$$

Therefore, the population-context version of Eq. (A50) gives

$$f(\boldsymbol{x}; \boldsymbol{\Theta}_L(t), \mathcal{D}_N) = \boldsymbol{\mu}^\top q_{L,t}(\boldsymbol{S}) \boldsymbol{U}^\top \boldsymbol{x}, \qquad N \to \infty, \tag{A79}$$

which proves part (iii).

Choose distinct nonzero numbers $t_1, \ldots, t_L$. Relative to the monomial basis $1, s, \ldots, s^{L-1}$, the coefficient matrix of $q_{L,t_1}, \ldots, q_{L,t_L}$ has entries $\binom{L}{j} t_i^j$ and reduces to an invertible Vandermonde matrix after nonzero row and column scalings. Hence, $\{q_{L,t} : t \in \mathbb{R}\}$ spans all polynomials of degree

at most $L-1$. Since $\boldsymbol{S} = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top$, induction on the Krylov order gives $\mathcal{K}_L(\boldsymbol{S}, \boldsymbol{\mu}) = \mathcal{K}_L(\boldsymbol{\Sigma}, \boldsymbol{\mu})$. Combining these facts yields

$$\begin{aligned} \operatorname{span}\{q_{L,t}(\boldsymbol{S})\boldsymbol{\mu} : t \in \mathbb{R}\} &= \operatorname{span}(\boldsymbol{\mu}, \boldsymbol{S}\boldsymbol{\mu}, \ldots, \boldsymbol{S}^{L-1}\boldsymbol{\mu}) \\ &= \mathcal{K}_L(\boldsymbol{\Sigma}, \boldsymbol{\mu}), \end{aligned} \tag{A80}$$

proving part (iv). □

Theorem 3.2 (A single linear transformer can be Bayes optimal across tasks)**.** *Suppose that (a) to (c) in Assumption 3.1 hold. Then, there exists a parameter $\boldsymbol{\Theta}_L$ such that, for every $\boldsymbol{U} \in \mathcal{O}(d)$,*

$$\mathcal{E}_{\epsilon,\boldsymbol{U}}\Big(f(\cdot\,; \boldsymbol{\Theta}_L, \mathcal{D}_N)\Big) - \mathcal{E}^\star_{\epsilon,\boldsymbol{U}} \lesssim_{\mathbb{P}} \frac{1}{N}. \tag{8}$$

The implied constant may depend on d, ϵ, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$.

Proof. Fix $\boldsymbol{U} \in \mathcal{O}(d)$, $N \in \mathbb{N}$, and $\delta \in (0,1)$. Let $\boldsymbol{\Theta}_{L,\star}$ be the parameter tuple from Lemma B.7 with the solution τ of Eq. (A24) and $\omega = 1$. Define $\psi(\boldsymbol{h}) := \mathcal{E}_{\epsilon,\boldsymbol{U}}(\boldsymbol{x} \mapsto \boldsymbol{h}^\top \boldsymbol{U}^\top \boldsymbol{x})$ for $\boldsymbol{h} \in \mathbb{R}^d$. By Eq. (A40), this function does not depend on $\boldsymbol{U}$. By Lemma B.4, $\psi(\boldsymbol{h}_\tau) = \mathcal{E}^\star_{\epsilon,\boldsymbol{U}}$ and $\boldsymbol{h}_\tau$ is a global minimizer of ψ. The vector $\boldsymbol{h}_\tau$ is nonzero, so ψ is twice continuously differentiable on a neighborhood of $\boldsymbol{h}_\tau$. Its gradient vanishes at this global minimizer, and Taylor's theorem gives a quadratic upper bound for $\psi(\boldsymbol{h}) - \psi(\boldsymbol{h}_\tau)$ on a sufficiently small ball centered at $\boldsymbol{h}_\tau$. Outside that ball, the same bound holds after increasing its constant because $0 \le \psi \le 1$ and the distance from $\boldsymbol{h}_\tau$ is bounded away from zero. Consequently, there is a constant $C_\psi > 0$ such that

$$0 \le \psi(\boldsymbol{h}) - \psi(\boldsymbol{h}_\tau) \le C_\psi \min\{\|\boldsymbol{h} - \boldsymbol{h}_\tau\|_2^2, 1\}, \qquad \boldsymbol{h} \in \mathbb{R}^d. \tag{A81}$$

Let X_N denote the excess error on the left-hand side of Eq. (8) with $\boldsymbol{\Theta}_L = \boldsymbol{\Theta}_{L,\star}$. Conditional on a realized context, Lemma B.4 and Eq. (A81) give

$$0 \le X_N \le C_\psi \|\boldsymbol{U}^\top \boldsymbol{\theta}_{\boldsymbol{\Theta}_{L,\star}}(\boldsymbol{G}_N(\boldsymbol{U})) - \boldsymbol{h}_\tau\|_2^2. \tag{A82}$$

Taking the conditional expectation and applying Lemma B.8 yield

$$\mathbb{E}_{\mathcal{D}_N|\boldsymbol{U}}[X_N] \le \frac{C_\psi C_{\mathrm{cmp}}}{N}. \tag{A83}$$

Markov's inequality therefore gives

$$\mathbb{P}_{\mathcal{D}_N|\boldsymbol{U}}\left(X_N > \frac{C_\psi C_{\mathrm{cmp}}}{\delta N}\right) \le \delta. \tag{A84}$$

Since $\boldsymbol{U}$ was arbitrary and the constants are independent of $\boldsymbol{U}$, this proves Eq. (8) with $C := C_\psi C_{\mathrm{cmp}}$. □

C Adversarial and Standard Training: Proofs of Theorems 3.4 and 3.5

C.1 Directions selected by infinite-sample margin risks

For $\epsilon \ge 0$, define the population margin risk

$$\mathcal{J}_\epsilon(\boldsymbol{h}) := \mathbb{E}_{\boldsymbol{a}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\mathcal{L}(\boldsymbol{h}^\top \boldsymbol{a} - \epsilon\sqrt{d}\|\boldsymbol{h}\|_2)]. \tag{A85}$$

The following result shows that minimizing this population surrogate risk selects the corresponding Bayes-optimal direction.

Lemma C.1 (Population margin-risk minimizers align with Bayes-optimal directions)**.** *Suppose that $0 \le \epsilon\sqrt{d} < \|\boldsymbol{\mu}\|_2$. Then the function $\mathcal{J}_\epsilon$ is strictly convex and coercive and therefore has a unique*

minimizer $\boldsymbol{h}_\epsilon^\star$, which is nonzero. There is a constant $\omega_\epsilon > 0$ such that

$$\operatorname*{argmin}_{\boldsymbol{h}\in\mathbb{R}^d} \mathcal{J}_\epsilon(\boldsymbol{h}) = \{\boldsymbol{h}_\epsilon^\star\} = \{\omega_\epsilon \boldsymbol{h}_\tau\} = \{\omega_\epsilon(\boldsymbol{\Sigma} + \tau \boldsymbol{I}_d)^{-1}\boldsymbol{\mu}\}. \tag{A86}$$

Proof. For every fixed $\boldsymbol{a}$, the map $\boldsymbol{h} \mapsto \boldsymbol{h}^\top \boldsymbol{a} - \epsilon\sqrt{d}\|\boldsymbol{h}\|_2$ is concave, so its composition with the convex nonincreasing function $\mathcal{L}$ is convex. For two distinct directions, the corresponding random margins differ almost surely because $\boldsymbol{\Sigma} \succ \boldsymbol{0}_{d,d}$. Strict convexity of $\mathcal{L}$, together with concavity of the margin map, therefore makes the expected risk strictly convex. Moreover, for every unit vector $\boldsymbol{u}$ and $t > 0$,

$$\mathcal{J}_\epsilon(t\boldsymbol{u}) \geq -\mathcal{L}'(0)t\mathbb{E}_{\boldsymbol{a}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[(\epsilon\sqrt{d} - \boldsymbol{u}^\top \boldsymbol{a})_+]. \tag{A87}$$

This bound follows from the tangent inequality for $\mathcal{L}$ on the event $\boldsymbol{u}^\top \boldsymbol{a} < \epsilon\sqrt{d}$ and nonnegativity of $\mathcal{L}$ on its complement. The expectation on the right is positive and continuous in $\boldsymbol{u}$, so compactness of the unit sphere gives a positive uniform lower bound. Hence, $\mathcal{J}_\epsilon$ is coercive and has a unique minimizer. Its one-sided derivative at the origin along $\boldsymbol{\mu}/\|\boldsymbol{\mu}\|_2$ equals $\mathcal{L}'(0)(\|\boldsymbol{\mu}\|_2 - \epsilon\sqrt{d}) < 0$, so the minimizer is nonzero.

Let $\boldsymbol{h} := \boldsymbol{h}_\epsilon^\star$ and set $Z := \boldsymbol{h}^\top \boldsymbol{a} - \epsilon\sqrt{d}\|\boldsymbol{h}\|_2$. The derivative bounds and Gaussian moments justify differentiation under the expectation and the Gaussian integration by parts below. The first-order condition at the nonzero minimizer is

$$\mathbb{E}_{\boldsymbol{a}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\mathcal{L}'(Z)\left(\boldsymbol{a} - \epsilon\sqrt{d}\frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}\right)\right] = \boldsymbol{0}_d. \tag{A88}$$

Gaussian integration by parts gives

$$\mathbb{E}_{\boldsymbol{a}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\mathcal{L}'(Z)\boldsymbol{a}] = \mathbb{E}_{\boldsymbol{a}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\mathcal{L}'(Z)]\boldsymbol{\mu} + \mathbb{E}_{\boldsymbol{a}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\mathcal{L}''(Z)]\boldsymbol{\Sigma}\boldsymbol{h}. \tag{A89}$$

Substitution into the first-order condition yields

$$\left(c\boldsymbol{\Sigma} + \frac{\epsilon\sqrt{d}}{\|\boldsymbol{h}\|_2}\boldsymbol{I}_d\right)\boldsymbol{h} = \boldsymbol{\mu}, \qquad c := \frac{\mathbb{E}_{\boldsymbol{a}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\mathcal{L}''(Z)]}{-\mathbb{E}_{\boldsymbol{a}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\mathcal{L}'(Z)]} > 0. \tag{A90}$$

Setting $\lambda := \epsilon\sqrt{d}/(c\|\boldsymbol{h}\|_2)$ gives

$$\boldsymbol{h} = \frac{1}{c}(\boldsymbol{\Sigma} + \lambda\boldsymbol{I}_d)^{-1}\boldsymbol{\mu}, \qquad \lambda\|(\boldsymbol{\Sigma} + \lambda\boldsymbol{I}_d)^{-1}\boldsymbol{\mu}\|_2 = \epsilon\sqrt{d}. \tag{A91}$$

Uniqueness in Lemma B.3 gives $\lambda = \tau$. Setting $\omega_\epsilon := 1/c$ proves Eq. (A86). □

C.2 Infinite-sample minimizers and finite-context transfer

The infinite-task, infinite-context risk is

$$\mathcal{R}_{\epsilon,\infty,\infty}(\boldsymbol{\Theta}_L) = \mathbb{E}_{\boldsymbol{U}\in\mathcal{O}(d)}[\mathcal{J}_\epsilon(\boldsymbol{U}^\top \boldsymbol{\theta}_{\boldsymbol{\Theta}_L}(\operatorname{diag}(\boldsymbol{U}, 1)\boldsymbol{G}_\star \operatorname{diag}(\boldsymbol{U}^\top, 1)))]. \tag{A92}$$

The preceding result also identifies the infinite-sample global optimum of the unrestricted transformer risk, independently of gradient dynamics.

Theorem C.2 (Functional identification of unrestricted fixed-depth infinite-sample minimizers)**.** *Suppose that $L \geq r$ and $0 \leq \epsilon\sqrt{d} < \|\boldsymbol{\mu}\|_2$. The unrestricted infinite-task, infinite-context risk in Eq. (A92) attains the lower bound*

$$\inf_{\boldsymbol{\Theta}_L} \mathcal{R}_{\epsilon,\infty,\infty}(\boldsymbol{\Theta}_L) = \mathcal{J}_\epsilon(\boldsymbol{h}_\epsilon^\star). \tag{A93}$$

Every global minimizer satisfies

$$\boldsymbol{U}^\top \boldsymbol{\theta}_{\boldsymbol{\Theta}_L}(\operatorname{diag}(\boldsymbol{U}, 1)\boldsymbol{G}_\star \operatorname{diag}(\boldsymbol{U}^\top, 1)) = \boldsymbol{h}_\epsilon^\star \tag{A94}$$

for every $\boldsymbol{U} \in \mathcal{O}(d)$. In particular, standard training identifies the direction $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, whereas adversarial training identifies the direction $\boldsymbol{h}_\tau$ in Eq. (A25).

Proof. Fix $\epsilon \geq 0$ satisfying $\epsilon\sqrt{d} < \|\boldsymbol{\mu}\|_2$. Conditional on $\boldsymbol{U}$, orthogonality gives

$$\mathcal{J}_\epsilon(\boldsymbol{U}^\top \boldsymbol{\theta}_{\boldsymbol{\Theta}_L}(\mathrm{diag}(\boldsymbol{U}, 1)\boldsymbol{G}_\star \,\mathrm{diag}(\boldsymbol{U}^\top, 1))) \geq \mathcal{J}_\epsilon(\boldsymbol{h}_\epsilon^\star). \tag{A95}$$

Apply Lemma B.7 with the solution τ of Eq. (A24) and $\omega = \omega_\epsilon$. Equation Eq. (A86) shows that the resulting comparator attains equality simultaneously for every $\boldsymbol{U}$. Hence, the lower bound in Eq. (A95) is attainable simultaneously for every $\boldsymbol{U}$, proving Eq. (A93). If a parameter tuple attains this value, the nonnegative conditional optimality gap in Eq. (A95) has expectation zero. The effective direction in Eq. (A49) is continuous in $\boldsymbol{U}$, so this conditional optimality gap is continuous on $\mathcal{O}(d)$. Since normalized Haar measure has full support, a nonnegative continuous function with zero Haar integral vanishes everywhere. Strict convexity from Lemma C.1 therefore gives Eq. (A94) for every $\boldsymbol{U}$. □

For $\epsilon \geq 0$ and $\boldsymbol{h} \neq \boldsymbol{0}_d$, define

$$\Gamma_\epsilon(\boldsymbol{h}) := \frac{\boldsymbol{\mu}^\top \boldsymbol{h} - \epsilon\sqrt{d}\|\boldsymbol{h}\|_2}{\sqrt{\boldsymbol{h}^\top \boldsymbol{\Sigma} \boldsymbol{h}}}, \tag{A96}$$

and set $\Gamma_\epsilon(\boldsymbol{0}_d) := -\infty$.

Lemma C.3 (Uniform finite-context transfer)**.** *Fix a depth-L parameter tuple $\boldsymbol{\Theta}_L$ and suppose that there is a nonzero vector $\boldsymbol{h} \in \mathbb{R}^d$ such that its population-context effective direction in Eq. (A49) satisfies*

$$\boldsymbol{U}^\top \boldsymbol{\theta}_{\boldsymbol{\Theta}_L}(\mathrm{diag}(\boldsymbol{U}, 1)\boldsymbol{G}_\star \,\mathrm{diag}(\boldsymbol{U}^\top, 1)) = \boldsymbol{h}, \qquad \boldsymbol{U} \in \mathcal{O}(d). \tag{A97}$$

Then, for every $\epsilon \geq 0$ satisfying $\epsilon\sqrt{d} < \|\boldsymbol{\mu}\|_2$,

$$\lim_{N\to\infty} \sup_{\boldsymbol{U}\in\mathcal{O}(d)} \left|\mathbb{E}_{\mathcal{D}_N|\boldsymbol{U}}[\mathcal{E}_{\epsilon,\boldsymbol{U}}(f(\,\cdot\,; \boldsymbol{\Theta}_L, \mathcal{D}_N))] - \Phi(-\Gamma_\epsilon(\boldsymbol{h}))\right| = 0. \tag{A98}$$

Proof. Couple the contexts for all task bases using the same latent vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N$. Equations Eqs. (A46) and (A47) and the strong law give

$$\sup_{\boldsymbol{U}\in\mathcal{O}(d)} \|\boldsymbol{G}_N(\boldsymbol{U}) - \mathrm{diag}(\boldsymbol{U}, 1)\boldsymbol{G}_\star \,\mathrm{diag}(\boldsymbol{U}^\top, 1)\|_{\mathrm{F}} = \|\boldsymbol{G}_{a,N} - \boldsymbol{G}_\star\|_{\mathrm{F}} \to 0 \tag{A99}$$

almost surely. Continuity of the fixed-depth recursion in Eq. (A48) and compactness of $\mathcal{O}(d)$ imply that the finite-context latent direction converges to $\boldsymbol{h}$ uniformly over $\boldsymbol{U}$ almost surely. Since $\boldsymbol{h} \neq \boldsymbol{0}_d$, the map $\boldsymbol{v} \mapsto \Phi(-\Gamma_\epsilon(\boldsymbol{v}))$ is continuous on a neighborhood of the limiting directions. By Lemma B.4, the conditional robust-error formula therefore converges uniformly over $\boldsymbol{U}$ almost surely. Since the robust errors and their limit belong to $[0, 1]$, bounded convergence gives

$$\mathbb{E}_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N}\left[\sup_{\boldsymbol{U}\in\mathcal{O}(d)} |\mathcal{E}_{\epsilon,\boldsymbol{U}}(f(\,\cdot\,; \boldsymbol{\Theta}_L, \mathcal{D}_N)) - \Phi(-\Gamma_\epsilon(\boldsymbol{h}))|\right] \to 0. \tag{A100}$$

Under the coupling above, the distribution of each context agrees with that of $\mathcal{D}_N$ conditional on its task basis, and hence

$$\sup_{\boldsymbol{U}\in\mathcal{O}(d)} \left|\mathbb{E}_{\mathcal{D}_N|\boldsymbol{U}}[\mathcal{E}_{\epsilon,\boldsymbol{U}}(f(\,\cdot\,; \boldsymbol{\Theta}_L, \mathcal{D}_N))] - \Phi(-\Gamma_\epsilon(\boldsymbol{h}))\right| \leq \mathbb{E}_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N}\left[\sup_{\boldsymbol{U}\in\mathcal{O}(d)} |\mathcal{E}_{\epsilon,\boldsymbol{U}}(f(\,\cdot\,; \boldsymbol{\Theta}_L, \mathcal{D}_N)) - \Phi(-\Gamma_\epsilon(\boldsymbol{h}))|\right], \tag{A101}$$

which proves Eq. (A98). □

C.3 Finite-task adversarial risk minimization

In statements and arguments involving a single generic task, we suppress the task index t from $\boldsymbol{U}_t$, $\mathcal{D}_{t,N}$, $\boldsymbol{x}_{t,n}$, $y_{t,n}$, and derived quantities, whereas we retain t when the T training tasks are considered jointly. The parameter constraint in Eq. (21) makes the empirical minimum attainable and is assumed to contain the comparator from Lemma B.7. For parameter tuples, $\|\boldsymbol{\Theta}_L\|_{\mathrm{F}}$ denotes the Euclidean product of the $2L$ matrix Frobenius norms.

Let $\mathbb{P}_{\text{task}}$ denote the distribution of a generic task $(\mathcal{D}_N, \boldsymbol{x}, y)$ induced by the generative process. Let $\mathcal{S}_{T,N} := \{(\mathcal{D}_{t,N}, \boldsymbol{x}_t, y_t)\}_{t=1}^{T}$ consist of T independent tasks drawn from $\mathbb{P}_{\text{task}}$. For the finite-context representation above, the finite-sample risk in Eq. (6) at perturbation level ϵ is

$$\mathcal{R}_{\epsilon,T,N}(\boldsymbol{\Theta}_L) = \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}(y_t \boldsymbol{\theta}_{\boldsymbol{\Theta}_L}(\boldsymbol{G}_N(\boldsymbol{U}_t))^\top \boldsymbol{x}_t - \epsilon\sqrt{d}\|\boldsymbol{\theta}_{\boldsymbol{\Theta}_L}(\boldsymbol{G}_N(\boldsymbol{U}_t))\|_2). \tag{A102}$$

Its infinite-task counterpart at context length N is

$$\mathcal{R}_{\epsilon,\infty,N}(\boldsymbol{\Theta}_L) := \mathbb{E}_{(\boldsymbol{U},\mathcal{D}_N)}[\mathcal{J}_\epsilon(\boldsymbol{U}^\top \boldsymbol{\theta}_{\boldsymbol{\Theta}_L}(\boldsymbol{G}_N(\boldsymbol{U})))]. \tag{A103}$$

Theorem 3.4 (Adversarial training yields robust linear transformers)**.** *Suppose that Assumption 3.1 holds. For sufficiently large $B > 0$, choose any empirical risk minimizer*

$$\widehat{\boldsymbol{\Theta}}_{L,\epsilon,T,N} \in \operatorname*{argmin}_{\max_{\ell\in[L]}\{\|\boldsymbol{P}_\ell\|_{\mathrm{F}},\|\boldsymbol{Q}_\ell\|_{\mathrm{F}}\}\le B} \mathcal{R}_{\epsilon,T,N}(\boldsymbol{\Theta}_L). \tag{21}$$

Then, for every $\boldsymbol{U} \in \mathcal{O}(d)$,

$$\mathcal{E}_{\epsilon,\boldsymbol{U}}\left(f\left(\cdot\,; \widehat{\boldsymbol{\Theta}}_{L,\epsilon,T,N}, \mathcal{D}_N\right)\right) - \mathcal{E}^\star_{\epsilon,\boldsymbol{U}} \lesssim_{\mathbb{P}} \frac{1}{\sqrt{T}} + \frac{1}{N}. \tag{22}$$

The implied constant may depend on d, B, L, $\mathcal{L}$, ϵ, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$.

Proof. Let $\boldsymbol{\Theta}_{L,\star}$ be the comparator from Lemma B.7 with the solution τ of Eq. (A24) and $\omega = \omega_\epsilon$ from Eq. (A86).

Uniform concentration over training tasks. Let $q_{\boldsymbol{\Theta}_L}(\mathcal{Z}_N)$ denote the summand in Eq. (A102), where $\mathcal{Z}_N$ contains one context and its independent query. Equations Eqs. (A48) and (A49) show that $\boldsymbol{\theta}_{\boldsymbol{\Theta}_L}(\boldsymbol{G}_N(\boldsymbol{U}))$ is a polynomial of fixed degree in $\boldsymbol{\Theta}_L$ and $\boldsymbol{G}_N(\boldsymbol{U})$. Compactness of the parameter constraint in Eq. (21) and boundedness of $\mathcal{L}'$ imply the existence of nonnegative envelopes F and M such that

$$|q_{\boldsymbol{\Theta}_L}(\mathcal{Z}_N)| \le F(\mathcal{Z}_N), \qquad |q_{\boldsymbol{\Theta}_L}(\mathcal{Z}_N) - q_{\widetilde{\boldsymbol{\Theta}}_L}(\mathcal{Z}_N)| \le M(\mathcal{Z}_N)\|\boldsymbol{\Theta}_L - \widetilde{\boldsymbol{\Theta}}_L\|_{\mathrm{F}}. \tag{A104}$$

The envelopes can be chosen as fixed polynomials in $1 + \|\boldsymbol{G}_N(\boldsymbol{U})\|_{\mathrm{F}}$ and $1 + \|\boldsymbol{x}\|_2$, with coefficients independent of T and N. The Gaussian task model and the averaging in $\boldsymbol{G}_N(\boldsymbol{U})$ give

$$\sup_{N\in\mathbb{N}} \mathbb{E}_{\mathcal{Z}_N}[F(\mathcal{Z}_N)^2 + M(\mathcal{Z}_N)^2] < \infty. \tag{A105}$$

The constrained parameter set is compact and finite-dimensional and hence has logarithmic Euclidean covering entropy in the covering radius. Define

$$\Delta_{T,N} := \max_{\substack{\boldsymbol{\Theta}_L=\{(\boldsymbol{P}_\ell,\boldsymbol{Q}_\ell)\}_{\ell=1}^{L} \\ \max_{\ell\in[L]}\{\|\boldsymbol{P}_\ell\|_{\mathrm{F}},\|\boldsymbol{Q}_\ell\|_{\mathrm{F}}\}\le B}} |\mathcal{R}_{\epsilon,T,N}(\boldsymbol{\Theta}_L) - \mathcal{R}_{\epsilon,\infty,N}(\boldsymbol{\Theta}_L)|. \tag{A106}$$

Symmetrization, the Euclidean covering bound, and Cauchy–Schwarz with Eq. (A105) yield a constant $C_{\text{conc}} > 0$ such that, for every $T, N \in \mathbb{N}$,

$$\mathbb{E}_{\mathcal{S}_{T,N}\sim\mathbb{P}_{\text{task}}^{\otimes T}}[\Delta_{T,N}] \le C_{\text{conc}} T^{-1/2}. \tag{A107}$$

Loss geometry and calibration. Define $\psi(\boldsymbol{h}) := \Phi(-\Gamma_\epsilon(\boldsymbol{h}))$, $\psi^\star := \psi(\boldsymbol{h}^\star_\epsilon)$, and

$$m(\boldsymbol{h}) := \min\{\|\boldsymbol{h} - \boldsymbol{h}^\star_\epsilon\|_2^2, 1\}. \tag{A108}$$

By Lemma C.1, $\boldsymbol{h}^\star_\epsilon = \omega_\epsilon \boldsymbol{h}_\tau$. Positive scale invariance of Γ_ϵ and Eqs. (A40), (A41) and (A96) therefore give

$$\psi^\star = \Phi(-\Gamma_\epsilon(\boldsymbol{h}_\tau)) = \mathcal{E}_{\epsilon,\boldsymbol{U}}(\boldsymbol{x} \mapsto \boldsymbol{h}_\tau{}^\top \boldsymbol{U}^\top \boldsymbol{x}) = \mathcal{E}^\star_{\epsilon,\boldsymbol{U}}, \qquad \boldsymbol{U} \in \mathcal{O}(d). \tag{A109}$$

The function $\mathcal{J}_\epsilon(\,\cdot\,)$ is twice continuously differentiable near the nonzero vector $\boldsymbol{h}^\star_\epsilon$, and its gradient vanishes there. Set $\boldsymbol{u}^\star := \boldsymbol{h}^\star_\epsilon/\|\boldsymbol{h}^\star_\epsilon\|_2$ and $Z^\star := \boldsymbol{h}^\star_\epsilon{}^\top \boldsymbol{a} - \epsilon\sqrt{d}\|\boldsymbol{h}^\star_\epsilon\|_2$. Direct differentiation at $\boldsymbol{h}^\star_\epsilon$ gives, for every $\boldsymbol{v} \neq \boldsymbol{0}_d$,

$$\boldsymbol{v}^\top \nabla^2 \mathcal{J}_\epsilon(\boldsymbol{h}^\star_\epsilon)\boldsymbol{v} \tag{A110}$$

$$= \mathbb{E}_{\boldsymbol{a}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\mathcal{L}''(Z^\star)(\boldsymbol{v}^\top(\boldsymbol{a}-\epsilon\sqrt{d}\boldsymbol{u}^\star))^2] \tag{A111}$$

$$- \frac{\epsilon\sqrt{d}\mathbb{E}_{\boldsymbol{a}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\mathcal{L}'(Z^\star)]}{\|\boldsymbol{h}_\epsilon^\star\|_2}(\|\boldsymbol{v}\|_2^2 - (\boldsymbol{v}^\top\boldsymbol{u}^\star)^2) \tag{A112}$$

$$> 0. \tag{A113}$$

The strict inequality holds because the first expectation is positive by $\mathcal{L}'' > 0$ and $\boldsymbol{\Sigma} \succ \mathbf{0}_{d,d}$, while the second term is nonnegative by $\mathcal{L}' < 0$. Taylor's theorem gives a quadratic upper and lower bound for the margin-risk gap on a neighborhood of $\boldsymbol{h}_\epsilon^\star$. Boundedness of $\mathcal{L}'$ makes $\mathcal{J}_\epsilon(\,\cdot\,)$ globally Lipschitz, so its quadratic upper bound extends outside this neighborhood because $\|\boldsymbol{h}-\boldsymbol{h}_\epsilon^\star\|_2$ is then bounded away from zero. Uniqueness and coercivity of $\mathcal{J}_\epsilon(\,\cdot\,)$ extend the truncated lower bound. Positive scale invariance of ψ, the identity $\boldsymbol{h}_\epsilon^\star = \omega_\epsilon \boldsymbol{h}_\tau$ from Eq. (A86), and Eq. (A81) give $\psi(\boldsymbol{h}) - \psi^\star \le C_\psi \max\{\omega_\epsilon^{-2}, 1\} m(\boldsymbol{h})$. Consequently, there are constants $C_{\mathcal{J}}, c_m, C_m > 0$ such that, for every $\boldsymbol{h} \in \mathbb{R}^d$,

$$0 \le \mathcal{J}_\epsilon(\boldsymbol{h}) - \mathcal{J}_\epsilon(\boldsymbol{h}_\epsilon^\star) \le C_{\mathcal{J}}\|\boldsymbol{h}-\boldsymbol{h}_\epsilon^\star\|_2^2, \tag{A114}$$

$$c_m m(\boldsymbol{h}) \le \mathcal{J}_\epsilon(\boldsymbol{h}) - \mathcal{J}_\epsilon(\boldsymbol{h}_\epsilon^\star), \tag{A115}$$

$$0 \le \psi(\boldsymbol{h}) - \psi^\star \le C_m m(\boldsymbol{h}). \tag{A116}$$

Finite-context comparator and finite-sample risk. Set $\boldsymbol{\delta}_N := \boldsymbol{\theta}_{\boldsymbol{\Theta}_{L,\star}}(\boldsymbol{G}_{a,N}) - \boldsymbol{h}_\epsilon^\star$. Applying Lemma B.8 with $\boldsymbol{U} = \boldsymbol{I}_d$ gives

$$\mathbb{E}_{\boldsymbol{a}_1,\dots,\boldsymbol{a}_N}[\|\boldsymbol{\delta}_N\|_2^2] \le \frac{C_{\mathrm{cmp}}}{N}. \tag{A117}$$

The infinite-context direction of $\boldsymbol{\Theta}_{L,\star}$ is $\boldsymbol{h}_\epsilon^\star$, so Eq. (A92) gives $\mathcal{R}_{\epsilon,\infty,\infty}(\boldsymbol{\Theta}_{L,\star}) = \mathcal{J}_\epsilon(\boldsymbol{h}_\epsilon^\star)$. The quadratic upper bound in Eq. (A116) therefore gives

$$\mathcal{R}_{\epsilon,\infty,N}(\boldsymbol{\Theta}_{L,\star}) - \mathcal{R}_{\epsilon,\infty,\infty}(\boldsymbol{\Theta}_{L,\star}) \le \frac{C_{\mathcal{J}}C_{\mathrm{cmp}}}{N}. \tag{A118}$$

Every conditional latent risk is at least $\mathcal{J}_\epsilon(\boldsymbol{h}_\epsilon^\star)$ by Lemma C.1. The minimizing property in Eq. (21) and the definition of $\Delta_{T,N}$ yield

$$\begin{aligned} 0 &\le \mathcal{R}_{\epsilon,\infty,N}(\widehat{\boldsymbol{\Theta}}_{L,\epsilon,T,N}) - \mathcal{J}_\epsilon(\boldsymbol{h}_\epsilon^\star) \\ &\le 2\Delta_{T,N} + \mathcal{R}_{\epsilon,\infty,N}(\boldsymbol{\Theta}_{L,\star}) - \mathcal{R}_{\epsilon,\infty,\infty}(\boldsymbol{\Theta}_{L,\star}) \\ &\le 2\Delta_{T,N} + \frac{C_{\mathcal{J}}C_{\mathrm{cmp}}}{N}. \end{aligned} \tag{A119}$$

Uniform calibration over the task basis. We next record the finite-dimensional comparison that turns Haar averages into worst-task bounds without changing the rate. Let $D_L \in \mathbb{N}$ be an upper bound on the total degree in the entries of $\boldsymbol{U}$ of the latent direction $\boldsymbol{U}^\top\boldsymbol{\theta}_{\boldsymbol{\Theta}_L}(\boldsymbol{G}_N(\boldsymbol{U}))$, and let $\mathcal{V}_{d,L}$ be the vector space of restrictions to $\mathcal{O}(d)$ of $\mathbb{R}^d$-valued polynomials of degree at most D_L. Since $\mathcal{V}_{d,L}$ is finite dimensional and normalized Haar measure has full support, its uniform norm and its L^2 norm under Haar measure are equivalent. Hence, there is $C_{d,L} > 0$ such that

$$\|\boldsymbol{q}\|_\infty^2 \le C_{d,L}\mathbb{E}_{\boldsymbol{U}\in\mathcal{O}(d)}[\|\boldsymbol{q}(\boldsymbol{U})\|_2^2], \qquad \boldsymbol{q} \in \mathcal{V}_{d,L}. \tag{A120}$$

Consequently,

$$\mathbb{E}_{\boldsymbol{U}\in\mathcal{O}(d)}[\min\{\|\boldsymbol{q}(\boldsymbol{U})\|_2^2, 1\}] \ge \frac{\mathbb{E}_{\boldsymbol{U}\in\mathcal{O}(d)}[\|\boldsymbol{q}(\boldsymbol{U})\|_2^2]}{\max\{\|\boldsymbol{q}\|_\infty^2, 1\}} \ge \frac{\min\{\|\boldsymbol{q}\|_\infty^2, 1\}}{C_{d,L}}. \tag{A121}$$

Since $\min\{\|\boldsymbol{q}\|_\infty^2, 1\} = \sup_{\boldsymbol{U}\in\mathcal{O}(d)} \min\{\|\boldsymbol{q}(\boldsymbol{U})\|_2^2, 1\}$, every $\boldsymbol{q} \in \mathcal{V}_{d,L}$ satisfies

$$\sup_{\boldsymbol{U}\in\mathcal{O}(d)} \min\{\|\boldsymbol{q}(\boldsymbol{U})\|_2^2, 1\} \le C_{d,L}\mathbb{E}_{\boldsymbol{U}\in\mathcal{O}(d)}[\min\{\|\boldsymbol{q}(\boldsymbol{U})\|_2^2, 1\}]. \tag{A122}$$

Using the coupled latent context matrix $\boldsymbol{G}_{a,N}$ from Eq. (A46), the ambient context is $\boldsymbol{G}_N(\boldsymbol{U}) = \mathrm{diag}(\boldsymbol{U}, 1)\boldsymbol{G}_{a,N}\,\mathrm{diag}(\boldsymbol{U}^\top, 1)$. The recursion in Eq. (A48) uses only matrix addition and multiplication, so for fixed latent vectors and fixed $\boldsymbol{\Theta}_L$, the map

$$\boldsymbol{U} \mapsto \boldsymbol{U}^\top\boldsymbol{\theta}_{\boldsymbol{\Theta}_L}(\boldsymbol{G}_N(\boldsymbol{U})) - \boldsymbol{h}_\epsilon^\star \tag{A123}$$

belongs to $\mathcal{V}_{d,L}$, with a degree bound independent of N, $\boldsymbol{\Theta}_L$, and the realized latent vectors. Applying Eqs. (A116) and (A122) for each fixed latent context and then taking expectations gives

$$\begin{aligned} 0 &\le \sup_{\boldsymbol{U}\in\mathcal{O}(d)} \mathbb{E}_{\mathcal{D}_N|\boldsymbol{U}}[\mathcal{E}_{\epsilon,\boldsymbol{U}}(f(\cdot;\boldsymbol{\Theta}_L,\mathcal{D}_N)) - \mathcal{E}^\star_{\epsilon,\boldsymbol{U}}] \\ &\le C_m \mathbb{E}_{\boldsymbol{a}_{1:N}}\left[\sup_{\boldsymbol{U}\in\mathcal{O}(d)} m(\boldsymbol{U}^\top \boldsymbol{\theta}_{\boldsymbol{\Theta}_L}(\boldsymbol{G}_N(\boldsymbol{U})))\right] \\ &\le C_m C_{d,L} \mathbb{E}_{(\boldsymbol{U},\boldsymbol{a}_{1:N})}[m(\boldsymbol{U}^\top \boldsymbol{\theta}_{\boldsymbol{\Theta}_L}(\boldsymbol{G}_N(\boldsymbol{U})))] \\ &\le \frac{C_m C_{d,L}}{c_m}\{\mathcal{R}_{\epsilon,\infty,N}(\boldsymbol{\Theta}_L) - \mathcal{J}_\epsilon(\boldsymbol{h}^\star_\epsilon)\}. \end{aligned} \tag{A124}$$

Conditional on the pretraining sample, apply Eq. (A119) to Eq. (A124) with $\boldsymbol{\Theta}_L = \widehat{\boldsymbol{\Theta}}_{L,\epsilon,T,N}$, and then take the expectation over the pretraining sample. Define

$$C := \frac{C_m C_{d,L}}{c_m}(2C_{\mathrm{conc}} + C_{\mathcal{J}} C_{\mathrm{cmp}}). \tag{A125}$$

Then, for every fixed $\boldsymbol{U} \in \mathcal{O}(d)$,

$$\mathbb{E}_{\mathcal{S}_{T,N},\mathcal{D}_N|\boldsymbol{U}}[\mathcal{E}_{\epsilon,\boldsymbol{U}}(f(\cdot;\widehat{\boldsymbol{\Theta}}_{L,\epsilon,T,N},\mathcal{D}_N)) - \mathcal{E}^\star_{\epsilon,\boldsymbol{U}}] \le C(T^{-1/2} + N^{-1}). \tag{A126}$$

Since the excess error is nonnegative, Markov's inequality shows that, for every $\delta \in (0,1)$,

$$\mathbb{P}_{\mathcal{S}_{T,N},\mathcal{D}_N|\boldsymbol{U}}\left(\mathcal{E}_{\epsilon,\boldsymbol{U}}(f(\cdot;\widehat{\boldsymbol{\Theta}}_{L,\epsilon,T,N},\mathcal{D}_N)) - \mathcal{E}^\star_{\epsilon,\boldsymbol{U}} > \frac{C}{\delta}(T^{-1/2} + N^{-1})\right) \le \delta. \tag{A127}$$

This proves Eq. (22). □

Theorem 3.5 (Standard training does not yield robust linear transformers)**.** *Suppose that Assumption 3.1 holds. Suppose also that $T, N \to \infty$; write $\mathcal{D}_\infty := \mathcal{D}_N$ and $\mathcal{R}_{0,\infty,\infty} := \mathcal{R}_{0,T,N}$. For sufficiently large $B > 0$, choose any population risk minimizer*

$$\widehat{\boldsymbol{\Theta}}_{L,0,\infty,\infty} \in \operatorname*{argmin}_{\max_{\ell\in[L]}\{\|\boldsymbol{P}_\ell\|_{\mathrm{F}},\|\boldsymbol{Q}_\ell\|_{\mathrm{F}}\}\le B} \mathcal{R}_{0,\infty,\infty}(\boldsymbol{\Theta}_L). \tag{23}$$

Then, there exists a constant $C > 0$ such that, for every $\boldsymbol{U} \in \mathcal{O}(d)$,

$$\mathcal{E}_{\epsilon,\boldsymbol{U}}\left(f\left(\cdot;\widehat{\boldsymbol{\Theta}}_{L,0,\infty,\infty},\mathcal{D}_\infty\right)\right) - \mathcal{E}^\star_{\epsilon,\boldsymbol{U}} = \begin{cases} 0 & (\epsilon = 0 \text{ or } r = 1) \\ C & (\epsilon > 0 \text{ and } r \ge 2) \end{cases}. \tag{24}$$

The constant C may depend on d, ϵ, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$.

Proof. Define $\boldsymbol{h}_0 := \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Applying Eq. (A86) at $\epsilon = 0$, the unique minimizer of the standard latent margin risk is $\boldsymbol{h}^\star_0 = \omega_0 \boldsymbol{h}_0$. Let $\boldsymbol{\Theta}_{L,\mathrm{std}}$ be the comparator from Lemma B.7 with $\tau = 0$ and $\omega = \omega_0$, and take B sufficiently large that this comparator is feasible in Eq. (23). The conditional lower bound in Eq. (A95) and the comparator construction give

$$\mathcal{R}_{0,\infty,\infty}(\boldsymbol{\Theta}_L) \ge \mathcal{J}_0(\boldsymbol{h}^\star_0) = \mathcal{R}_{0,\infty,\infty}(\boldsymbol{\Theta}_{L,\mathrm{std}}), \tag{A128}$$

for every parameter tuple $\boldsymbol{\Theta}_L$. Therefore, the minimizing property in Eq. (23) implies that $\widehat{\boldsymbol{\Theta}}_{L,0,\infty,\infty}$ is also an unrestricted global minimizer of the infinite-sample standard risk. Applying Theorem C.2 yields, for every $\boldsymbol{U} \in \mathcal{O}(d)$,

$$\boldsymbol{U}^\top \boldsymbol{\theta}_{\widehat{\boldsymbol{\Theta}}_{L,0,\infty,\infty}}(\mathrm{diag}(\boldsymbol{U},1)\boldsymbol{G}_\star \,\mathrm{diag}(\boldsymbol{U}^\top,1)) = \omega_0 \boldsymbol{h}_0. \tag{A129}$$

Hence, query linearity, positive scale invariance of Γ_ϵ, and Lemma B.4 give

$$\begin{aligned} &\mathcal{E}_{\epsilon,\boldsymbol{U}}(f(\cdot;\widehat{\boldsymbol{\Theta}}_{L,0,\infty,\infty},\mathcal{D}_\infty)) - \mathcal{E}^\star_{\epsilon,\boldsymbol{U}} \\ &\quad = \Phi(-\Gamma_\epsilon(\boldsymbol{h}_0)) - \Phi(-\sqrt{\boldsymbol{h}_\tau^\top \boldsymbol{\Sigma} \boldsymbol{h}_\tau}). \end{aligned} \tag{A130}$$

In particular, the right-hand side is independent of $\boldsymbol{U}$.

If $\epsilon = 0$, then $\tau = 0$ and $\boldsymbol{h}_\tau = \boldsymbol{h}_0$, so Eq. (A130) is zero. Suppose next that $r = 1$, and let $\lambda > 0$ be the unique active eigenvalue of $\boldsymbol{\Sigma}$. Then $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \lambda^{-1}\boldsymbol{\mu}$ and $\boldsymbol{h}_\tau = (\lambda + \tau)^{-1}\boldsymbol{\mu}$, so the two directions are positive multiples of each other and Eq. (A130) is again zero.

It remains to consider $\epsilon > 0$ and $r \geq 2$. Equation Eq. (A24) implies $\tau > 0$. Choose two distinct active eigenvalues $\lambda_i \neq \lambda_j$ of $\boldsymbol{\Sigma}$, with spectral projectors $\boldsymbol{E}_i$ and $\boldsymbol{E}_j$. If $\boldsymbol{h}_0 = c\boldsymbol{h}_\tau$ for some $c > 0$, then projecting this identity onto either active eigenspace gives

$$c = \frac{\lambda_k + \tau}{\lambda_k} = 1 + \frac{\tau}{\lambda_k}, \qquad k \in \{i, j\}. \tag{A131}$$

This is impossible because $\tau > 0$ and $\lambda_i \neq \lambda_j$. Thus, $\boldsymbol{h}_0$ is not a positive multiple of $\boldsymbol{h}_\tau$, so the uniqueness conclusion of Lemma B.4 shows that its robust error is strictly larger than the robust Bayes error. Consequently, the constant

$$C := \Phi(-\Gamma_\epsilon(\boldsymbol{h}_0)) - \Phi(-\sqrt{\boldsymbol{h}_\tau^\top \boldsymbol{\Sigma} \boldsymbol{h}_\tau}) \tag{A132}$$

is strictly positive, and Eq. (A130) equals C for every $\boldsymbol{U} \in \mathcal{O}(d)$. In the zero-gap cases, any positive choice of C is admissible because that constant does not appear in the corresponding branch of Eq. (24). □

D Gradient Flow to Robust Bayes Optimality

For each task t, the adversarial loss as a function of the last-layer parameter takes the form

$$\boldsymbol{Q} \mapsto \mathcal{L}(y_t \boldsymbol{\theta}_t(\boldsymbol{Q})^\top \boldsymbol{x}_t - \epsilon\sqrt{d}\|\boldsymbol{\theta}_t(\boldsymbol{Q})\|_2).$$

Because the Euclidean norm is nondifferentiable at $\mathbf{0}_d$, this objective need not be differentiable when $\boldsymbol{\theta}_t(\boldsymbol{Q}) = \mathbf{0}_d$; we therefore use subgradient flow, which agrees with ordinary gradient flow wherever all effective directions are nonzero.

Theorem D.1 (Bayes consistency of last-layer gradient flow)**.** *Assume $L = 2$ and $\boldsymbol{P}_1 = \boldsymbol{P}_2 = \boldsymbol{Q}_1 = \boldsymbol{I}_{d+1}$. Suppose that Assumption 3.1 holds. For sufficiently large $B > 0$, let $\mathcal{Q}_B := \{\boldsymbol{Q} \in \mathbb{R}^{(d+1)\times(d+1)} : \|\boldsymbol{Q}\|_{\mathrm{F}} \leq B\}$. Fix any $\boldsymbol{Q}_0 \in \mathbb{R}^{(d+1)\times(d+1)}$. Let $\boldsymbol{Q}_{T,N}(s)$ be the unique global solution to the Frobenius subgradient flow*

$$\dot{\boldsymbol{Q}}_{T,N}(s) \in -\partial_{\boldsymbol{Q}} \mathcal{R}_{\epsilon,T,N}(\boldsymbol{Q}_{T,N}(s)), \qquad \boldsymbol{Q}_{T,N}(0) = \boldsymbol{Q}_0. \tag{A133}$$

Assume that $\mathcal{Q}_B$ contains a population-risk minimizer and, almost surely, the entire unconstrained subgradient-flow trajectory $\{\boldsymbol{Q}_{T,N}(s) : s \geq 0\}$. Then, for every $\boldsymbol{U} \in \mathcal{O}(d)$ and $s > 0$,

$$\mathcal{E}_{\epsilon,\boldsymbol{U}}(f(\,\cdot\,; \boldsymbol{Q}_{T,N}(s), \mathcal{D}_N)) - \mathcal{E}^\star_{\epsilon,\boldsymbol{U}} \lesssim_{\mathbb{P}} \frac{1}{s} + \frac{1}{\sqrt{T}} + \frac{1}{N}. \tag{A134}$$

Proof. For task t, let $\boldsymbol{G}_{t,N}$ be its finite-context matrix in Eq. (A47).

$$\boldsymbol{A}_{t,1} := \boldsymbol{I}_{d+1} + \boldsymbol{G}_{t,N}, \qquad \boldsymbol{G}_{t,1} := \boldsymbol{A}_{t,1} \boldsymbol{G}_{t,N} \boldsymbol{A}_{t,1}^\top. \tag{A135}$$

The final cumulative transformation and its effective direction are

$$\boldsymbol{A}_{t,2}(\boldsymbol{Q}) = \{\boldsymbol{I}_{d+1} + \boldsymbol{G}_{t,1}\boldsymbol{Q}\}\boldsymbol{A}_{t,1}, \qquad \boldsymbol{\theta}_t(\boldsymbol{Q}) := [\boldsymbol{I}_d \quad \mathbf{0}_d]\, \boldsymbol{A}_{t,2}(\boldsymbol{Q})^\top \boldsymbol{e}_{d+1}. \tag{A136}$$

Thus, $\boldsymbol{\theta}_t(\boldsymbol{Q})$ is affine in $\boldsymbol{Q}$. Since $\mathcal{L}$ is decreasing, query linearity gives

$$\mathcal{R}_{\epsilon,T,N}(\boldsymbol{Q}) = \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}\Big(y_{t,N+1} \boldsymbol{\theta}_t(\boldsymbol{Q})^\top \boldsymbol{x}_{t,N+1} - \epsilon\sqrt{d}\|\boldsymbol{\theta}_t(\boldsymbol{Q})\|_2\Big). \tag{A137}$$

For every t, the argument of $\mathcal{L}$ in Eq. (A137) is concave in $\boldsymbol{Q}$. Because $\mathcal{L}$ is convex and decreasing, its composition with this argument is convex. Hence, $\mathcal{R}_{\epsilon,T,N}(\,\cdot\,)$ is finite, continuous, and convex.

We next show that the restricted last-layer class contains an infinite-sample optimal comparator. For $q_x, q_y \in \mathbb{R}$, define $\boldsymbol{Q}(q_x, q_y) := q_x \boldsymbol{\Pi}_x + q_y \boldsymbol{\Pi}_y$. Replacing the finite-context matrix in Eq. (A136) by the population context and multiplying the blocks at $\boldsymbol{U} = \boldsymbol{I}_d$ gives the population latent direction

$$\boldsymbol{h}(q_x, q_y) = (1 + z q_y)\boldsymbol{\mu} + q_x \boldsymbol{v}, \tag{A138}$$

$$z := \boldsymbol{\mu}^\top \boldsymbol{S}\boldsymbol{\mu} + 4\|\boldsymbol{\mu}\|_2^2 + 4 > 0, \tag{A139}$$

$$\boldsymbol{v} := (\boldsymbol{I}_d + \boldsymbol{S})\{\boldsymbol{S}^2 + 3\boldsymbol{S} + (\|\boldsymbol{\mu}\|_2^2 + 4)\boldsymbol{I}_d\}\boldsymbol{\mu}. \tag{A140}$$

The same latent direction is realized for every $\boldsymbol{U} \in \mathcal{O}(d)$ because the fixed prefix and $\boldsymbol{Q}(q_x, q_y)$ commute with $\mathrm{diag}(\boldsymbol{U}, 1)$. Let $\mathcal{K} := \mathcal{K}_r(\boldsymbol{\Sigma}, \boldsymbol{\mu})$ and recall that $r \le 2$ by assumption. The subspace $\mathcal{K}$ contains $\boldsymbol{\mu}$ and, by Eq. (A86), $\boldsymbol{h}_\epsilon^\star$, and it is invariant under $\boldsymbol{S}$. By Lemma B.5, $\boldsymbol{S}$ has exactly r distinct active eigenvalues on $\mathcal{K}$. Write

$$\boldsymbol{v} = p(\boldsymbol{S})\boldsymbol{\mu}, \qquad p(s) := (1+s)(s^2+3s+\|\boldsymbol{\mu}\|_2^2+4), \qquad p'(s) = 3s^2+8s+\|\boldsymbol{\mu}\|_2^2+7 > 0. \tag{A141}$$

If $r = 1$, then $\boldsymbol{\mu}$ spans $\mathcal{K}$. If $r = 2$, the spectral theorem and strict monotonicity of p on the positive active eigenvalues make $\boldsymbol{\mu}$ and $\boldsymbol{v}$ linearly independent. Thus, $\mathrm{span}\{\boldsymbol{\mu}, \boldsymbol{v}\} = \mathcal{K}$ in both cases. Since $z > 0$, Eq. (A140) therefore gives scalars $q_x^\star, q_y^\star$ such that $\boldsymbol{h}(q_x^\star, q_y^\star) = \boldsymbol{h}_\epsilon^\star$. Set $\boldsymbol{Q}^\star := \boldsymbol{Q}(q_x^\star, q_y^\star)$, which belongs to $\mathcal{Q}_B$ by the choice of B. Equation Eq. (A92) and the optimality of $\boldsymbol{h}_\epsilon^\star$ show that $\boldsymbol{Q}^\star$ minimizes $\mathcal{R}_{\epsilon,\infty,\infty}$ even within the unrestricted depth-two class.

Continuity of the finite-sample risk and compactness of $\mathcal{Q}_B$ give a finite-sample minimizer

$$\widehat{\boldsymbol{Q}}_{T,N} \in \operatorname*{argmin}_{\boldsymbol{Q} \in \mathcal{Q}_B} \mathcal{R}_{\epsilon,T,N}(\boldsymbol{Q}). \tag{A142}$$

Since the finite-sample risk is finite, continuous, and convex, its subdifferential is maximal monotone. Therefore, the flow in Eq. (A133) exists uniquely for all $s \ge 0$.

Because $\boldsymbol{Q}_{T,N}(s)$ and $\widehat{\boldsymbol{Q}}_{T,N}$ belong to $\mathcal{Q}_B$, the finite-sample risk gap between them is nonnegative. The subgradient inequality gives, for almost every $s \ge 0$,

$$\frac{\mathrm{d}}{\mathrm{d}s}\frac{1}{2}\|\boldsymbol{Q}_{T,N}(s) - \widehat{\boldsymbol{Q}}_{T,N}\|_{\mathrm{F}}^2 \le -\{\mathcal{R}_{\epsilon,T,N}(\boldsymbol{Q}_{T,N}(s)) - \mathcal{R}_{\epsilon,T,N}(\widehat{\boldsymbol{Q}}_{T,N})\}. \tag{A143}$$

The convex subgradient-flow chain rule shows that the finite-sample risk is nonincreasing along the flow. Integrating the preceding inequality and using this monotonicity and $\|\boldsymbol{Q}_0 - \widehat{\boldsymbol{Q}}_{T,N}\|_{\mathrm{F}} \le 2B$ gives, for every $s > 0$,

$$0 \le \mathcal{R}_{\epsilon,T,N}(\boldsymbol{Q}_{T,N}(s)) - \mathcal{R}_{\epsilon,T,N}(\widehat{\boldsymbol{Q}}_{T,N}) \le \frac{\|\boldsymbol{Q}_0 - \widehat{\boldsymbol{Q}}_{T,N}\|_{\mathrm{F}}^2}{2s} \le \frac{2B^2}{s}. \tag{A144}$$

It remains to transfer empirical optimization to robust Bayes error. Since $\boldsymbol{Q}^\star$ is fixed, Eq. (A48) makes its finite-context direction a fixed polynomial in the empirical latent moments whose population value is $\boldsymbol{h}_\epsilon^\star$. The Gaussian sample-moment bound in Eq. (A73) therefore gives, for every $\boldsymbol{U} \in \mathcal{O}(d)$,

$$\mathbb{E}_{\mathcal{D}_N|\boldsymbol{U}}[\|\boldsymbol{U}^\top \boldsymbol{\theta}_{\boldsymbol{Q}^\star}(\boldsymbol{G}_N(\boldsymbol{U})) - \boldsymbol{h}_\epsilon^\star\|_2^2] \lesssim \frac{1}{N}. \tag{A145}$$

The quadratic upper bound in Eq. (A116) therefore yields

$$\mathcal{R}_{\epsilon,\infty,N}(\boldsymbol{Q}^\star) - \mathcal{R}_{\epsilon,\infty,\infty}(\boldsymbol{Q}^\star) \lesssim \frac{1}{N}. \tag{A146}$$

Define the restricted uniform deviation

$$\Delta_{T,N} := \max_{\boldsymbol{Q} \in \mathcal{Q}_B} |\mathcal{R}_{\epsilon,T,N}(\boldsymbol{Q}) - \mathcal{R}_{\epsilon,\infty,N}(\boldsymbol{Q})|. \tag{A147}$$

The uniform concentration argument in the proof of Theorem 3.4, with parameter bound $\max\{B, \sqrt{d+1}\}$ and restricted to the compact last-layer class, gives

$$\mathbb{E}_{\mathcal{S}_{T,N} \sim \mathbb{P}_{\mathrm{task}}^{\otimes T}}[\Delta_{T,N}] \lesssim \frac{1}{\sqrt{T}}, \tag{A148}$$

uniformly over N. Fix any $\boldsymbol{U} \in \mathcal{O}(d)$. Since $\widehat{\boldsymbol{Q}}_{T,N}$ minimizes the finite-sample risk over $\mathcal{Q}_B$ and $\boldsymbol{Q}^\star \in \mathcal{Q}_B$, Eqs. (A124) and (A144) give

$$\begin{aligned} 0 &\le \mathbb{E}_{\mathcal{D}_N|\boldsymbol{U}}[\mathcal{E}_{\epsilon,\boldsymbol{U}}(f(\cdot\,;\boldsymbol{Q}_{T,N}(s),\mathcal{D}_N)) - \mathcal{E}^\star_{\epsilon,\boldsymbol{U}}] \\ &\le \frac{C_m C_{d,2}}{c_m}\{2\Delta_{T,N} + 2B^2 s^{-1} + \mathcal{R}_{\epsilon,\infty,N}(\boldsymbol{Q}^\star) - \mathcal{R}_{\epsilon,\infty,\infty}(\boldsymbol{Q}^\star)\}. \end{aligned} \tag{A149}$$

Taking the expectation over the pretraining sample and applying Eqs. (A146) and (A148) give a constant $C > 0$ such that

$$\mathbb{E}_{\mathcal{S}_{T,N},\mathcal{D}_N|\boldsymbol{U}}[\mathcal{E}_{\epsilon,\boldsymbol{U}}(f(\cdot\,;\boldsymbol{Q}_{T,N}(s),\mathcal{D}_N)) - \mathcal{E}^\star_{\epsilon,\boldsymbol{U}}] \le C\left(\frac{1}{s} + \frac{1}{\sqrt{T}} + \frac{1}{N}\right). \tag{A150}$$

Markov's inequality therefore gives, for every $\delta \in (0,1)$,

$$\mathbb{P}_{\mathcal{S}_{T,N},\mathcal{D}_N|\boldsymbol{U}}\left(\mathcal{E}_{\epsilon,\boldsymbol{U}}(f(\cdot\,;\boldsymbol{Q}_{T,N}(s),\mathcal{D}_N)) - \mathcal{E}^\star_{\epsilon,\boldsymbol{U}} > \frac{C}{\delta}\left(\frac{1}{s} + \frac{1}{\sqrt{T}} + \frac{1}{N}\right)\right) \le \delta, \tag{A151}$$

which proves Eq. (A134). □

E Accuracy–Robustness Trade-Off

Theorem E.1 (Adversarial training loses standard accuracy)**.** *Suppose that Assumption 3.1 holds. Suppose also that $T, N \to \infty$; write $\mathcal{D}_\infty := \mathcal{D}_N$ and $\mathcal{R}_{\epsilon,\infty,\infty} := \mathcal{R}_{\epsilon,T,N}$. For sufficiently large $B > 0$, choose any population risk minimizer*

$$\widehat{\boldsymbol{\Theta}}_{L,\epsilon,\infty,\infty} \in \operatorname*{argmin}_{\max_{\ell\in[L]}\{\|\boldsymbol{P}_\ell\|_{\mathrm{F}},\|\boldsymbol{Q}_\ell\|_{\mathrm{F}}\}\le B} \mathcal{R}_{\epsilon,\infty,\infty}(\boldsymbol{\Theta}_L). \tag{A152}$$

Then, there exists a constant $C > 0$ such that, for every $\boldsymbol{U} \in \mathcal{O}(d)$,

$$\mathcal{E}_{0,\boldsymbol{U}}(f(\cdot\,;\widehat{\boldsymbol{\Theta}}_{L,\epsilon,\infty,\infty},\mathcal{D}_\infty)) - \mathcal{E}^\star_{0,\boldsymbol{U}} = \begin{cases} 0 & (\epsilon = 0 \text{ or } r = 1), \\ C & (\epsilon > 0 \text{ and } r \ge 2). \end{cases} \tag{A153}$$

The constant C may depend on d, ϵ, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$.

Proof. Define $\boldsymbol{h}_0 := \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. By Eq. (A86), the unique minimizer of the adversarial latent margin risk is $\boldsymbol{h}^\star_\epsilon = \omega_\epsilon \boldsymbol{h}_\tau$ for some $\omega_\epsilon > 0$. Let $\boldsymbol{\Theta}_{L,\mathrm{adv}}$ be the comparator from Lemma B.7 with the solution τ of Eq. (A24) and $\omega = \omega_\epsilon$, and take B sufficiently large that this comparator is feasible in Eq. (A152). The conditional lower bound in Eq. (A95) and the comparator construction give

$$\mathcal{R}_{\epsilon,\infty,\infty}(\boldsymbol{\Theta}_L) \ge \mathcal{J}_\epsilon(\boldsymbol{h}^\star_\epsilon) = \mathcal{R}_{\epsilon,\infty,\infty}(\boldsymbol{\Theta}_{L,\mathrm{adv}}) \tag{A154}$$

for every parameter tuple $\boldsymbol{\Theta}_L$. Therefore, the minimizing property in Eq. (A152) implies that $\widehat{\boldsymbol{\Theta}}_{L,\epsilon,\infty,\infty}$ is also an unrestricted global minimizer of the infinite-sample adversarial risk. Applying Theorem C.2 yields, for every $\boldsymbol{U} \in \mathcal{O}(d)$,

$$\boldsymbol{U}^\top \boldsymbol{\theta}_{\widehat{\boldsymbol{\Theta}}_{L,\epsilon,\infty,\infty}}(\mathrm{diag}(\boldsymbol{U},1)\boldsymbol{G}_\star\,\mathrm{diag}(\boldsymbol{U}^\top,1)) = \omega_\epsilon \boldsymbol{h}_\tau. \tag{A155}$$

Hence, query linearity, positive scale invariance of Γ_0, and Lemma B.4 give

$$\begin{aligned} &\mathcal{E}_{0,\boldsymbol{U}}(f(\cdot\,;\widehat{\boldsymbol{\Theta}}_{L,\epsilon,\infty,\infty},\mathcal{D}_\infty)) - \mathcal{E}^\star_{0,\boldsymbol{U}} \\ &\quad = \Phi\left(-\frac{\boldsymbol{\mu}^\top \boldsymbol{h}_\tau}{\sqrt{\boldsymbol{h}_\tau^\top \boldsymbol{\Sigma} \boldsymbol{h}_\tau}}\right) - \Phi(-\sqrt{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}). \end{aligned} \tag{A156}$$

In particular, the right-hand side is independent of $\boldsymbol{U}$.

If $\epsilon = 0$, then $\tau = 0$ and $\boldsymbol{h}_\tau = \boldsymbol{h}_0$, so Eq. (A156) is zero. Suppose next that $r = 1$, and let $\lambda > 0$ be the unique active eigenvalue of $\boldsymbol{\Sigma}$. Then $\boldsymbol{h}_0 = \lambda^{-1}\boldsymbol{\mu}$ and $\boldsymbol{h}_\tau = (\lambda + \tau)^{-1}\boldsymbol{\mu}$, so the two directions are positive multiples of each other and Eq. (A156) is again zero.

It remains to consider $\epsilon > 0$ and $r \geq 2$. Equation Eq. (A24) implies $\tau > 0$. Choose two distinct active eigenvalues $\lambda_i \neq \lambda_j$ of $\boldsymbol{\Sigma}$, with spectral projectors $\boldsymbol{E}_i$ and $\boldsymbol{E}_j$. If $\boldsymbol{h}_\tau = c\boldsymbol{h}_0$ for some $c > 0$, then projecting this identity onto either active eigenspace gives

$$c = \frac{\lambda_k}{\lambda_k + \tau}, \qquad k \in \{i, j\}. \tag{A157}$$

This is impossible because $\tau > 0$ and $\lambda_i \neq \lambda_j$. Thus, $\boldsymbol{h}_\tau$ is not a positive multiple of $\boldsymbol{h}_0$, so the uniqueness conclusion of Lemma B.4, applied at perturbation radius zero, shows that its standard error is strictly larger than the standard Bayes error. Consequently, the constant

$$C := \Phi\left(-\frac{\boldsymbol{\mu}^\top \boldsymbol{h}_\tau}{\sqrt{\boldsymbol{h}_\tau^\top \boldsymbol{\Sigma} \boldsymbol{h}_\tau}}\right) - \Phi(-\sqrt{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}}) \tag{A158}$$

is strictly positive, and Eq. (A156) equals C for every $\boldsymbol{U} \in \mathcal{O}(d)$. □

F Demonstration Complexity

Theorem F.1 (Robust prediction requires more demonstrations)**.** *Suppose that, for constants* $0 < c < C < \infty$,

$$c\sqrt{d} \leq \|\boldsymbol{\mu}\|_2 \leq C\sqrt{d}, \qquad c\boldsymbol{I}_d \preceq \boldsymbol{\Sigma} \preceq C\boldsymbol{I}_d, \qquad c \leq \|\boldsymbol{\mu}\|_2 - \epsilon\sqrt{d} \leq C. \tag{A159}$$

Fix any depth $L \in \mathbb{N}$. *For every target error* $\eta > \mathcal{E}^\star_{\epsilon,\boldsymbol{U}}$, *define*

$$N_\epsilon(\eta) := \min\{N \in \mathbb{N} : \inf_{\boldsymbol{\Theta}_L} \sup_{\boldsymbol{U} \in \mathcal{O}(d)} \mathbb{E}_{\mathcal{D}_N | \boldsymbol{U}}[\mathcal{E}_{\epsilon,\boldsymbol{U}}(f(\,\cdot\,; \boldsymbol{\Theta}_L, \mathcal{D}_N))] \leq \eta\}. \tag{A160}$$

Then, for every fixed $\delta \in (0, 1/2)$ *and all sufficiently large* d,

$$N_0(\mathcal{E}^\star_{0,\boldsymbol{U}} + \delta) = 1, \qquad N_\epsilon(\mathcal{E}^\star_{\epsilon,\boldsymbol{U}} + \delta) = \Theta(\sqrt{d}). \tag{A161}$$

The constants hidden in Θ *depend only on* c, C, *and* δ.

Proof. Write $\overline{\boldsymbol{\mu}} := \boldsymbol{\mu}/\|\boldsymbol{\mu}\|_2$.

The robust Bayes error is asymptotically attained by the mean direction. For every unit vector $\boldsymbol{q} \in \mathbb{R}^d$, define its robust standardized margin by

$$M(\boldsymbol{q}) := \frac{\boldsymbol{\mu}^\top \boldsymbol{q} - \epsilon\sqrt{d}}{\sqrt{\boldsymbol{q}^\top \boldsymbol{\Sigma} \boldsymbol{q}}} = \frac{\|\boldsymbol{\mu}\|_2 - \epsilon\sqrt{d} - \|\boldsymbol{\mu}\|_2(1 - \overline{\boldsymbol{\mu}}^\top \boldsymbol{q})}{\sqrt{\boldsymbol{q}^\top \boldsymbol{\Sigma} \boldsymbol{q}}}. \tag{A162}$$

By Lemma B.4, the robust Bayes error is $\Phi(-M^\star)$, where $M^\star := \max_{\|\boldsymbol{q}\|_2 = 1} M(\boldsymbol{q})$. Substituting $\boldsymbol{q} = \overline{\boldsymbol{\mu}}$ gives

$$M^\star \geq \frac{\|\boldsymbol{\mu}\|_2 - \epsilon\sqrt{d}}{\sqrt{\overline{\boldsymbol{\mu}}^\top \boldsymbol{\Sigma} \overline{\boldsymbol{\mu}}}} > 0. \tag{A163}$$

Every maximizer $\boldsymbol{q}^\star$ therefore has a positive numerator in Eq. (A162), and hence

$$\|\boldsymbol{q}^\star - \overline{\boldsymbol{\mu}}\|_2^2 = 2(1 - \overline{\boldsymbol{\mu}}^\top \boldsymbol{q}^\star) < \frac{2(\|\boldsymbol{\mu}\|_2 - \epsilon\sqrt{d})}{\|\boldsymbol{\mu}\|_2} = O(d^{-1/2}). \tag{A164}$$

The spectral bounds and Eq. (A164) imply

$$|\boldsymbol{q}^{\star\top} \boldsymbol{\Sigma} \boldsymbol{q}^\star - \overline{\boldsymbol{\mu}}^\top \boldsymbol{\Sigma} \overline{\boldsymbol{\mu}}| \leq 2C\|\boldsymbol{q}^\star - \overline{\boldsymbol{\mu}}\|_2 = O(d^{-1/4}). \tag{A165}$$

The numerator of $M(\boldsymbol{q}^\star)$ is at most $\|\boldsymbol{\mu}\|_2 - \epsilon\sqrt{d}$, while every denominator in Eq. (A162) is at least $\sqrt{c}$. It follows that

$$0 \leq M^\star - \frac{\|\boldsymbol{\mu}\|_2 - \epsilon\sqrt{d}}{\sqrt{\overline{\boldsymbol{\mu}}^\top \boldsymbol{\Sigma} \overline{\boldsymbol{\mu}}}} = O(d^{-1/4}). \tag{A166}$$

Since the standard normal density is bounded, Eq. (A166) yields

$$\mathcal{E}^{\star}_{\epsilon,U} = \Phi\left(-\frac{\|\boldsymbol{\mu}\|_2 - \epsilon\sqrt{d}}{\sqrt{\overline{\boldsymbol{\mu}}^\top \boldsymbol{\Sigma}\overline{\boldsymbol{\mu}}}}\right) + O(d^{-1/4}) \tag{A167}$$

uniformly over $\boldsymbol{U} \in \mathcal{O}(d)$.

The empirical-mean predictor. Given N demonstrations, define

$$\overline{\boldsymbol{v}}_N := \frac{1}{N}\sum_{n=1}^{N} y_n \boldsymbol{x}_n, \qquad f_{\mathrm{mean}}(\boldsymbol{x}; \mathcal{D}_N) := \overline{\boldsymbol{v}}_N^\top \boldsymbol{x}. \tag{A168}$$

For every depth $L \in \mathbb{N}$, this predictor is realized by setting $\boldsymbol{P}_\ell = \boldsymbol{Q}_\ell = \boldsymbol{0}_{d+1,d+1}$ for $1 \le \ell < L$, $\boldsymbol{P}_L = \boldsymbol{\Pi}_y$, and $\boldsymbol{Q}_L = \boldsymbol{I}_{d+1}$. Conditional on a task basis $\boldsymbol{U}$, its latent direction is

$$\overline{\boldsymbol{a}}_N := \boldsymbol{\mu} + \boldsymbol{\xi}_N, \qquad \boldsymbol{\xi}_N \sim \mathcal{N}\left(\boldsymbol{0}_d, \frac{\boldsymbol{\Sigma}}{N}\right), \qquad \overline{\boldsymbol{v}}_N = \boldsymbol{U}\overline{\boldsymbol{a}}_N. \tag{A169}$$

One demonstration suffices for standard prediction. For $N = 1$, Eq. (A169) and the spectral bounds give

$$\boldsymbol{\mu}^\top \overline{\boldsymbol{a}}_1 = \|\boldsymbol{\mu}\|_2^2 + O_{\mathbb{P}}(\|\boldsymbol{\mu}\|_2), \qquad \overline{\boldsymbol{a}}_1^\top \boldsymbol{\Sigma}\overline{\boldsymbol{a}}_1 = O_{\mathbb{P}}(d). \tag{A170}$$

Hence, the standardized clean margin diverges in probability:

$$\frac{\boldsymbol{\mu}^\top \overline{\boldsymbol{a}}_1}{\sqrt{\overline{\boldsymbol{a}}_1^\top \boldsymbol{\Sigma}\overline{\boldsymbol{a}}_1}} \xrightarrow{\mathbb{P}} \infty. \tag{A171}$$

Equations Eqs. (A40) and (A171), boundedness of the error, and convergence in probability imply

$$\sup_{\boldsymbol{U}\in\mathcal{O}(d)} \mathbb{E}_{\mathcal{D}_1|\boldsymbol{U}}[\mathcal{E}_{0,\boldsymbol{U}}(f_{\mathrm{mean}}(\,\cdot\,;\mathcal{D}_1))] \to 0. \tag{A172}$$

The standard Bayes error is nonnegative, so Eq. (A172) shows that, for every fixed $\delta > 0$ and all sufficiently large d,

$$N_0(\mathcal{E}^{\star}_{0,\boldsymbol{U}} + \delta) = 1. \tag{A173}$$

Upper bound for robust prediction. Fix $a > 0$, set $N := \lceil a\sqrt{d}\rceil$, and define $\widehat{\boldsymbol{q}} := \overline{\boldsymbol{a}}_N/\|\overline{\boldsymbol{a}}_N\|_2$. Let $\mathcal{A} := \{\|\boldsymbol{\xi}_N\|_2 \le \|\boldsymbol{\mu}\|_2/2\}$. Markov's inequality and Eq. (A159) give

$$\mathbb{P}(\mathcal{A}^{\mathsf{c}}) \le \frac{4\mathbb{E}[\|\boldsymbol{\xi}_N\|_2^2]}{\|\boldsymbol{\mu}\|_2^2} = \frac{4\,\mathrm{tr}(\boldsymbol{\Sigma})}{N\|\boldsymbol{\mu}\|_2^2} = O(d^{-1/2}). \tag{A174}$$

On $\mathcal{A}$, normalization is Lipschitz and therefore

$$\|\boldsymbol{\mu}\|_2(1 - \overline{\boldsymbol{\mu}}^\top \widehat{\boldsymbol{q}}) = \frac{\|\boldsymbol{\mu}\|_2}{2}\|\widehat{\boldsymbol{q}} - \overline{\boldsymbol{\mu}}\|_2^2 \le \frac{2\|\boldsymbol{\xi}_N\|_2^2}{\|\boldsymbol{\mu}\|_2}. \tag{A175}$$

Taking expectations in Eq. (A175) yields

$$\mathbb{E}[\|\boldsymbol{\mu}\|_2(1 - \overline{\boldsymbol{\mu}}^\top \widehat{\boldsymbol{q}})\mathbb{1}\{\mathcal{A}\}] \le \frac{2\,\mathrm{tr}(\boldsymbol{\Sigma})}{N\|\boldsymbol{\mu}\|_2} \le \frac{K}{a} \tag{A176}$$

for a constant $K > 0$ depending only on c and C. Moreover, Cauchy–Schwarz, Eq. (A174), and the same normalization bound give

$$\mathbb{E}[\|\widehat{\boldsymbol{q}} - \overline{\boldsymbol{\mu}}\|_2\mathbb{1}\{\mathcal{A}\}] = O(d^{-1/4}), \qquad \mathbb{P}(\mathcal{A}^{\mathsf{c}}) = O(d^{-1/2}). \tag{A177}$$

The map $\boldsymbol{q} \mapsto (\boldsymbol{q}^\top \boldsymbol{\Sigma}\boldsymbol{q})^{-1/2}$ is uniformly Lipschitz on the unit sphere under Eq. (A159). Combining Eqs. (A40), (A167), (A174), (A176) and (A177) with the Lipschitz continuity of Φ gives

$$\sup_{\boldsymbol{U}\in\mathcal{O}(d)} \mathbb{E}_{\mathcal{D}_N|\boldsymbol{U}}[\mathcal{E}_{\epsilon,\boldsymbol{U}}(f_{\mathrm{mean}}(\,\cdot\,;\mathcal{D}_N))] - \mathcal{E}^{\star}_{\epsilon,\boldsymbol{U}} \le \frac{K'}{a} + o(1) \tag{A178}$$

for a constant $K' > 0$ depending only on c and C. Choosing $a > 2K'/\delta$ proves

$$N_\epsilon(\mathcal{E}^\star_{\epsilon,\boldsymbol{U}} + \delta) = O(\sqrt{d}) \tag{A179}$$

uniformly over L.

Lower bound for robust prediction. A standard volumetric packing argument gives unit vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{J_d} \in \overline{\boldsymbol{\mu}}^\perp$ such that

$$J_d \geq \exp(kd), \qquad \|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2 \geq 1 \quad (i \neq j) \tag{A180}$$

for a numerical constant $k > 0$. Fix a constant $B \geq 32C/c$, to be chosen sufficiently large below, set $\beta_d := \sqrt{B}d^{-1/4}$, and, for each $j \in [J_d]$, let $\boldsymbol{U}_j$ be the rotation through angle β_d in the plane spanned by $\overline{\boldsymbol{\mu}}$ and $\boldsymbol{v}_j$ that maps $\overline{\boldsymbol{\mu}}$ to

$$\boldsymbol{u}_j := \cos(\beta_d)\overline{\boldsymbol{\mu}} + \sin(\beta_d)\boldsymbol{v}_j \tag{A181}$$

and acts as the identity on the orthogonal complement of this plane. For all sufficiently large d, Eq. (A180) implies

$$\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2 \geq \frac{\beta_d}{2} \quad (i \neq j). \tag{A182}$$

Let $\mathbb{P}_j$ be the distribution of one labeled demonstration under task $\boldsymbol{U}_j$. The label is independent of the task, while the signed feature has distribution

$$y\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{U}_j\boldsymbol{\mu}, \boldsymbol{U}_j\boldsymbol{\Sigma}\boldsymbol{U}_j^\top). \tag{A183}$$

Each difference $\boldsymbol{U}_i - \boldsymbol{U}_j$ has rank at most four and Frobenius norm $O(\beta_d)$. In particular,

$$\|\boldsymbol{U}_i\boldsymbol{\mu} - \boldsymbol{U}_j\boldsymbol{\mu}\|_2 = O(\|\boldsymbol{\mu}\|_2\beta_d), \qquad \|\boldsymbol{U}_i\boldsymbol{\Sigma}\boldsymbol{U}_i^\top - \boldsymbol{U}_j\boldsymbol{\Sigma}\boldsymbol{U}_j^\top\|_{\mathrm{F}} = O(\beta_d). \tag{A184}$$

To control the covariance contribution to the Gaussian relative entropy, apply $x - 1 - \log x \leq K_0(x-1)^2$ on the compact interval $[c/C, C/c]$ to the eigenvalues of

$$(\boldsymbol{U}_j\boldsymbol{\Sigma}\boldsymbol{U}_j^\top)^{-1/2}\boldsymbol{U}_i\boldsymbol{\Sigma}\boldsymbol{U}_i^\top(\boldsymbol{U}_j\boldsymbol{\Sigma}\boldsymbol{U}_j^\top)^{-1/2}. \tag{A185}$$

The Gaussian relative-entropy formula, Eqs. (A159) and (A184), and this scalar bound give

$$\max_{i,j\in[J_d]} \mathrm{KL}(\mathbb{P}_i\|\mathbb{P}_j) \leq K''\|\boldsymbol{\mu}\|_2^2\beta_d^2 \leq K''C^2B\sqrt{d} \tag{A186}$$

for a constant $K'' > 0$ depending only on c and C. Fix $\alpha > 0$ and suppose that $N \leq \alpha\sqrt{d}$. Tensorization, Eqs. (A180) and (A186), and Fano's inequality show that every estimator $\widehat{J}$ based on N demonstrations satisfies

$$\frac{1}{J_d}\sum_{j=1}^{J_d}\mathbb{P}_j^{\otimes N}(\widehat{J} \neq j) \geq 1 - \frac{K''C^2\alpha Bd + \log 2}{kd}. \tag{A187}$$

By Eq. (A50), every linear transformer has an effective direction determined by the demonstrations. Assign an arbitrary unit direction when this effective direction is zero, and otherwise normalize it to obtain $\widehat{\boldsymbol{q}}$. Decode the task by choosing the element of $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_{J_d}\}$ nearest to $\widehat{\boldsymbol{q}}$. Whenever this decoder is incorrect, Eq. (A182) gives

$$1 - \boldsymbol{u}_j^\top\widehat{\boldsymbol{q}} = \frac{1}{2}\|\widehat{\boldsymbol{q}} - \boldsymbol{u}_j\|_2^2 \geq \frac{\beta_d^2}{32}. \tag{A188}$$

On this event, the numerator in the conditional robust-error formula is bounded below by

$$\epsilon\sqrt{d} - \|\boldsymbol{\mu}\|_2\boldsymbol{u}_j^\top\widehat{\boldsymbol{q}} = \|\boldsymbol{\mu}\|_2(1 - \boldsymbol{u}_j^\top\widehat{\boldsymbol{q}}) - (\|\boldsymbol{\mu}\|_2 - \epsilon\sqrt{d}) \geq \frac{cB}{32} - C. \tag{A189}$$

The choice $B \geq 32C/c$ makes this lower bound nonnegative. Its standard deviation is at most $\sqrt{C}$ by Eq. (A159). Thus, Eqs. (A40), (A187) and (A189) lower-bound the robust error averaged over the packed tasks by

$$\left(1 - \frac{K''C^2\alpha B}{k} - o(1)\right)\Phi\left(\frac{cB/32 - C}{\sqrt{C}}\right). \tag{A190}$$

The zero effective direction itself has robust error one, so the arbitrary assignment made above does not weaken the lower bound. The mean direction is a feasible robust classifier, so

$$\mathcal{E}^{\star}_{\epsilon,U} \le \Phi\left(-\frac{\|\boldsymbol{\mu}\|_2 - \epsilon\sqrt{d}}{\sqrt{\overline{\boldsymbol{\mu}}^{\top}\boldsymbol{\Sigma}\overline{\boldsymbol{\mu}}}}\right) \le \Phi\left(-\frac{c}{\sqrt{C}}\right) < \frac{1}{2}. \tag{A191}$$

Since $\Phi(-c/\sqrt{C}) + \delta < 1$, first choose $B \ge 32C/c$ sufficiently large and then α sufficiently small that the lower bound in Eq. (A190) is strictly larger than $\Phi(-c/\sqrt{C}) + \delta$ for all sufficiently large d. The worst-task error is at least the average over the packed tasks, so Eq. (A191) implies

$$N_{\epsilon}(\mathcal{E}^{\star}_{\epsilon,\boldsymbol{U}} + \delta) = \Omega(\sqrt{d}) \tag{A192}$$

uniformly over L. Combining Eqs. (A173), (A179) and (A192) proves Eq. (A161). □