Minimax Optimal Early-Stopped Gradient Descent for Gaussian Mixture Classification

Alex Buna **Shirley Xiaoqi Liu** **Patrick Rebeschini**

Abstract

In overparameterised classification, training data can be linearly separable even when the underlying distribution is not. In this setting, gradient descent (GD) on the logistic loss diverges in norm while converging in direction to a max-margin interpolating classifier, whose implicit bias can be statistically suboptimal. In this work, we show that early stopping can overcome this suboptimality: in a Gaussian mixture model with label-flipping noise, GD stopped at an appropriate oracle time achieves minimax-optimal excess zero-one risk for covariance spectra with fast and continuous decay, including polynomial and exponential spectral decays. Our analysis combines a sharp upper bound for the early-stopped iterate with a matching statistical lower bound over arbitrary classifiers, yielding optimal rates that are validated by experiments. A central technical contribution is a new calibration result that converts excess logistic risk into excess zero-one risk; it handles the model misspecification induced by the label-flipping noise, and removes the square-root rate in standard bounds. We also establish a lower bound for linear interpolators, showing that interpolation can require exponentially more samples than early stopping to achieve the same excess risk.

1 Introduction

Machine learning models are routinely trained in overparameterised regimes, where the number of parameters d exceeds the sample size n. While interpolation [25] and early-stopping in gradient descent (GD) [40] are both known to generalise in classification settings under suitable conditions, their relative statistical efficiency remains less well-understood than in regression settings. In this work, we study classification in Gaussian mixtures with label-flipping noise, a canonical model that captures key challenges of high-dimensional classification, including misspecification under the logistic loss. In this setting, we show that early-stopped GD, with an appropriately chosen stopping time, is *minimax-optimal* in certain regimes, whereas interpolating estimators can require exponentially more samples to achieve the same excess risk.

Overparameterisation has been extensively studied in least-squares regression. Foundational studies [6, 7, 14, 35, 26] show that GD on least-squares converges to the minimum ℓ_2-norm interpolator, which, under sufficient overparameterisation and suitable spectral conditions on the data covariance, achieves vanishing excess risk as $n \to \infty$, a phenomenon termed "benign overfitting". Complementary to these results, early stopping for GD methods has been shown to act as an implicit regulariser in overparameterised regimes, closely connected to explicit ℓ_2-regularisation in least-squares regression [41, 33]. Under suitable spectral and signal regularity conditions, appropriately early-stopped iterates achieve statistically optimal rates for oracle stopping times [8, 41, 22, 27] and, despite the added difficulty, data-dependent stopping times [28, 38, 2, 19]. Most recently, it was shown in [39] that GD with a data-dependent stopping rule is minimax-optimal over a range of power-law spectral classes defined via the capacity and source conditions [10].

The authors are with the Department of Statistics, University of Oxford. Emails: alex.bunamarginean@spc.ox.ac.uk, shirley.liu@stats.ox.ac.uk, and patrick.rebeschini@stats.ox.ac.uk.

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

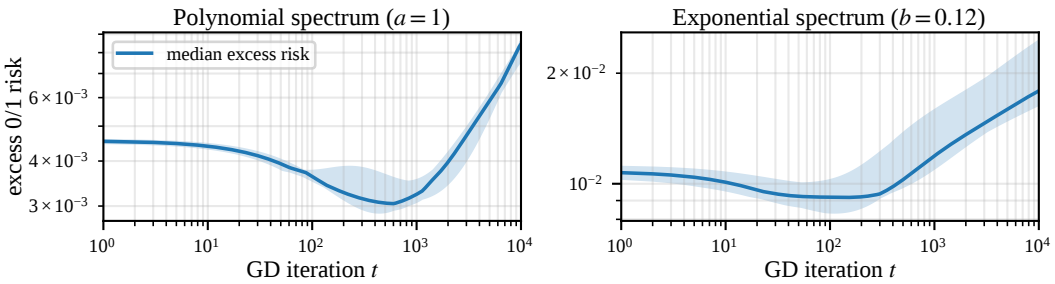


Figure 1: **Early stopping along the GD trajectory.** Excess zero-one population risk along GD iterates trained on $n = 972$ samples from the $d = 3888$-dimensional Gaussian mixture model $\boldsymbol{X} \sim \mathcal{N}(\pm\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with label-flip probability $p = 0.1$ (see (1)). The centres $\boldsymbol{\mu}$ satisfy $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\| \lesssim 1$. Left: polynomial spectrum $\lambda_i(\boldsymbol{\Sigma}) \asymp i^{-2a}$ with $a = 1$. Right: exponential spectrum $\lambda_i(\boldsymbol{\Sigma}) \asymp e^{-bi}$ with $b = 0.12$. The solid curve shows the median over independent repetitions, and the shaded band shows the interquartile range. In both cases, the excess risk first decreases and then increases as GD approaches interpolation, illustrating the statistical benefit of early stopping.

These regression results rely heavily on closed-form solutions and linear GD dynamics under least-squares loss. In contrast, even for linear classifiers, the logistic loss admits neither property, making the theory for classification far less developed despite its central role in machine learning. Specifically, in overparameterised logistic regression, GD behaves fundamentally differently from least-squares regression. Its iterates diverge in ℓ_2-norm while converging in direction to the max-margin on linearly-separable data [32, 31], and to the same direction up to a bounded offset on nonseparable data [17].

As a consequence of this implicit bias, classification analyses [11, 37, 9, 13] depart from those for least-squares regression, taking the max-margin classifier as the key object of study. A complementary line of work [25, 16, 1] shows that, under sufficient overparameterisation where all samples become support vectors, the max-margin classifier coincides with the minimum ℓ_2-norm interpolator, enabling tools from least-squares analysis to transfer to classification [25, 16, 37, 36]. Both strands of work characterise sufficient spectral conditions under which the max-margin classifier exhibits benign overfitting. Moreover, [37, 36] explicitly compare interpolation with the ℓ_2-regularised classifier. However, none of these works on classification establishes how interpolation compares to early-stopped GD in terms of sample complexity, i.e., the number of samples n required to achieve a target excess zero-one risk, or more fundamentally, whether interpolation or early-stopped GD is *statistically optimal*.

This question is motivated by the phenomenon illustrated in Figure 1. When running GD with the logistic loss for binary classification on linearly separable data, the population excess zero-one risk initially decreases, but then increases again as GD approaches the interpolating regime, illustrating that early stopping improves generalisation, whereas interpolation can be statistically suboptimal.

Characterising this statistical benefit of early stopping is the focus of this work. We study a canonical *generative* model in classification: Gaussian mixtures with label-flipping noise. In this model, the covariate vector $\boldsymbol{X} \in \mathbb{R}^d$ clusters around class-conditional means $\pm\boldsymbol{\mu} \in \mathbb{R}^d$, and the observed label Y is obtained by flipping a latent label $\widetilde{Y}$ independently with probability $p \in [0, 0.5)$. Formally, each sample $(\boldsymbol{X}, Y) \in \mathbb{R}^d \times \{\pm 1\}$ is generated by first drawing a latent label $\widetilde{Y} \in \{\pm 1\}$ uniformly, then setting

$$\boldsymbol{X} = \widetilde{Y}\boldsymbol{\mu} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \quad \text{and} \quad Y = \begin{cases} -\widetilde{Y} & \text{w.p. } p, \\ \widetilde{Y} & \text{w.p. } 1 - p, \end{cases} \tag{1}$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ is symmetric positive definite. A key challenge in this model is the presence of label-flipping noise, as studied for Gaussian mixtures in the interpolation regime in [11, 37, 13, 36]. When $p > 0$, the distribution of $Y \mid \boldsymbol{X}$ differs from the sigmoid model and is thus *misspecified* under the logistic loss. As a result, standard calibration bounds that relate excess zero-one and logistic risks [42, 5] yield suboptimal rates, necessitating new proof techniques.

In this work, we overcome the challenge due to misspecification and prove that, early-stopped GD on the logistic loss is *minimax-optimal* for classification in the model in (1) when the covariance spectrum is fast and continuously decaying (FCD), in line with the terminology used in earlier regression literature [39, Assumption 3]. To our knowledge, this is the first result establishing

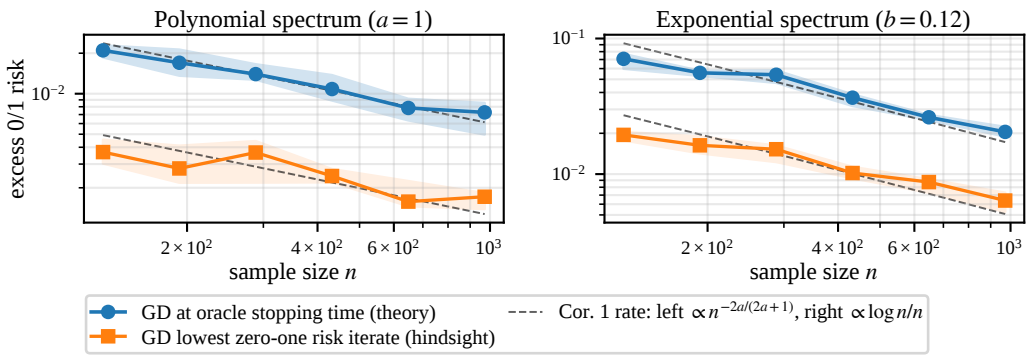


Figure 2: **Empirical rates for early-stopped GD.** Excess zero-one population risk as a function of the sample size n, under the same experimental setting as Figure 1. The blue curve shows GD at the oracle stopping time used in our theory, while the orange curve shows the GD iterate with the lowest (in hindsight) excess zero-one risk, along the same GD trajectory. Solid curves and shaded bands show medians and interquartile ranges over independent repetitions. The thin black dashed lines show the rates from Corollary 1, $n^{-2a/(2a+1)}$ and $\log n/n$, up to multiplicative constants.

the statistical optimality of early-stopped GD for Gaussian mixture classification. The FCD class covers a range of spectral decays, including both polynomial and exponential spectra, whereas [39] establishes minimax rates for GD in regression within power-law spectral classes, which correspond to polynomial decay. Figure 2 provides a finite-sample validation of our theory: it plots the empirical excess risk as a function of the sample size under both polynomial and exponential spectra, showing close agreement with the minimax rates predicted by our theory, summarised in Table 1.

1.1 Main Contributions

We study GD under the logistic loss for Gaussian mixture classification with label-flipping noise. We establish that in certain regimes early stopping is *sufficient and necessary* for statistical optimality:

- (Sufficiency) We prove that for FCD covariance spectra, there exists an oracle stopping time at which the GD iterate achieves minimax-optimal excess zero-one risk (Theorem 3). This is established via an upper bound for the early-stopped GD iterate (Theorem 1) and a matching statistical lower bound for arbitrary classifiers (Theorem 2). To the best of our knowledge, this is the first minimax optimality result for early-stopped GD for Gaussian mixture classification.
- For common FCD spectral regimes, specialising Theorem 3 yields Corollary 1, which provides explicit minimax rates for polynomially and exponentially-decaying spectra. These rates are summarised in Table 1, and are closely matched by the empirical rates in Figure 2.
- A key technical contribution is a calibration result (Lemma 4) linking excess logistic and zero-one risks in the Gaussian mixture model with label-flipping noise, which renders the model misspecified under the logistic loss. Unlike standard calibration bounds [42, 5], which incur a square-root rate and optimise over all measurable predictors, our result achieves a linear rate and is restricted to linear classifiers. This sharper calibration is essential for deriving our minimax-optimal upper bound.
- (Necessity) Finally, we prove an excess risk lower bound over all linear interpolators (Theorem 4), including the max-margin classifier that GD converges to in direction. As summarised in Table 1, in scaling regimes such as $d \asymp n \log n$, this interpolation lower bound vanishes strictly more slowly in n than the minimax rate of early-stopped GD for polynomial and exponential spectra. This result builds on and generalises Theorem 4.2 of [40] by removing its signal-sparsity assumption. This generalisation is essential for bringing this interpolation lower bound within the scope of our minimax theory, since the worst-case signal directions that determine the minimax rates need not be sparse.

1.2 Related Work

A first step towards understanding the statistical benefit of early stopping is taken by [40], which analyses early-stopped GD in *well-specified* logistic regression, a *discriminative* model. It shows that

Table 1: Summary of main results: minimax-optimal excess zero-one risk achieved by early-stopped GD at the oracle stopping time τ; excess zero-one risk lower bound at interpolation.

Excess zero-one risk	Polynomial decay $\lambda_i(\mathbf{\Sigma}) \asymp i^{-2a},\ a > \frac{1}{2}$	Exponential decay $\lambda_i(\mathbf{\Sigma}) \asymp e^{-bi},\ b > 0$
At oracle stopping time τ	$\asymp n^{-2a/(2a+1)}$	$\asymp \log n/n$
At interpolation (for $d \asymp n \log n$)	$\gtrsim 1/\log n$	$\gtrsim 1/\log n$

early-stopped GD incurs only polynomial sample complexity in certain overparameterised regimes where interpolation requires exponential sample complexity. Our work goes beyond [40] in two directions. First, we study a qualitatively distinct Gaussian mixture model, a generative model that becomes misspecified under the logistic loss in the presence of label-flipping noise $p > 0$. This misspecification requires new calibration bounds for the analysis. Second, we show that early-stopped GD is not only more sample-efficient than interpolation in common overparametrised regimes, but it is, in fact, minimax-optimal for the class of FCD covariance spectra.

Analyses of GD often characterise problem difficulty through the capacity-source conditions, e.g., [10], where capacity measures spectral decay and source measures signal alignment via a parameter r. A closely related work to ours on the regression side is [39], which establishes the minimax optimality of early-stopped GD over power-law spectral classes. These classes are defined via the capacity-source conditions, thus corresponding to polynomial spectral decay, and allowing a general alignment parameter $r \geq 0$. Their minimax rates capture explicit dependence on r. Our work is similar in spirit but focuses on classification under the logistic loss. We show minimax optimality of early-stopped GD for a broader class of spectral decays than [39], namely the FCD class, which includes both polynomial and exponential spectral decays. However, our minimax rates can be seen as corresponding to $r = \frac{1}{2}$ and therefore do not explicitly capture dependence on r.

Interpolation in Gaussian mixtures is studied as a canonical linear classification setting for benign overfitting in recent works [11, 25, 37, 9, 13, 36]; see [13, Table 1] for a summary of results across different regimes depending on the covariance spectrum, the signal-to-noise ratio, and label-flip probability. Our work studies the same model with general covariance spectrum, but focuses on the statistical performance of early-stopped GD, using interpolation as an important point of comparison.

In contrast to the finite-sample results reviewed above and developed in this paper, a long line of work [34, 29, 12, 20, 23, 18, 30, 24] studies overparameterised linear classification in the proportional *asymptotic* regime, where $d, n \to \infty$ with d/n converging to a constant. This regime enables precise asymptotic characterisation of the generalisation error. However, these results often rely on systems of equations that are hard to approach analytically and focus on isotropic covariance, with less attention given to anisotropic settings, which are important for generalisation of ℓ_2-regularised solutions [6, 25].

Notation: We use ordinary letters for scalars, boldface letters for vectors or matrices, uppercase letters for random variables, and lowercase for their realisations. For a symmetric positive definite matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$, let $\operatorname{tr}(\mathbf{\Sigma}) := \sum_{i=1}^{d} \mathbf{\Sigma}_{ii}$ denote its trace. For a vector $\boldsymbol{x} \in \mathbb{R}^d$ we denote $\|\boldsymbol{x}\| := \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$ and $\|\boldsymbol{x}\|_{\mathbf{\Sigma}} := \sqrt{\boldsymbol{x}^\top \mathbf{\Sigma} \boldsymbol{x}}$. For an arbitrary matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, let $\|\mathbf{X}\| := \sup_{\boldsymbol{v}\in\mathbb{R}^d, \|\boldsymbol{v}\|=1} \|\mathbf{X}\boldsymbol{v}\|$ denote its spectral norm. We use $\mathtt{c}$ with various subscripts to denote positive universal constants independent of n and d. Throughout, $a \lesssim b$ (resp. $a \gtrsim b$) means $a \leq \mathtt{c}_1 b$ (resp. $a \geq \mathtt{c}_2 b$) for universal constants $\mathtt{c}_1, \mathtt{c}_2 > 0$, and $a \asymp b$ means $a \lesssim b$ and $a \gtrsim b$ hold simultaneously.

2 Problem Setup

We study binary classification in the Gaussian mixtures model defined in (1). For a score function $h : \mathbb{R}^d \to \mathbb{R}$ with associated classifier $\operatorname{sign}(h) : \mathbb{R}^d \to \{\pm 1\}$, define its population zero-one risk as

$$\mathcal{E}(h) := \mathbb{P}\left(\operatorname{sign}(h(\boldsymbol{X})) \neq Y\right),$$

and for a linear score parametrised by $\boldsymbol{w} \in \mathbb{R}^d$, $h_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{w}$, we abbreviate

$$\mathcal{E}(\boldsymbol{w}) := \mathcal{E}(h_{\boldsymbol{w}}).$$

It is known (see, e.g., [15]) that the Bayes-optimal score function is linear, unique up to positive scaling, and given by

$$h_{\boldsymbol{w}^*} \in \operatorname*{arg\,min}_{h:\mathbb{R}^d \to \mathbb{R}} \mathcal{E}(h), \quad \text{where} \quad \boldsymbol{w}^* := \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}.$$

Hence, the goal of a training algorithm is to learn a linear classifier $\boldsymbol{w}$ that minimises the excess zero-one risk $\mathcal{E}(\boldsymbol{w}) - \mathcal{E}(\boldsymbol{w}^*)$.[1]

Since the zero-one loss is discontinuous, training commonly minimises the logistic loss $\ell(z) = \log(1+\exp(-z))$, a smooth convex surrogate. As before, we define the logistic risk of a general score $h : \mathbb{R}^d \to \mathbb{R}$ or a linear score $h_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{w}$, respectively, as:

$$\mathcal{L}(h) := \mathbb{E}[\ell(Yh(\boldsymbol{X}))] \quad \text{and} \quad \mathcal{L}(\boldsymbol{w}) := \mathcal{L}(h_{\boldsymbol{w}})\,.$$

The following lemma characterises the minimiser

$$\boldsymbol{w}^*_{\mathcal{L}} \in \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{w})\,,$$

which plays a central role in our early-stopped GD analysis:

Lemma 1 (Population logistic risk minimiser)**.** *The minimiser of $\mathcal{L}(\boldsymbol{w})$ over $\mathbb{R}^d$ is unique and has the form $\boldsymbol{w}^*_{\mathcal{L}} = \rho^* \boldsymbol{w}^*$ where $\rho^* \in (0, 2]$.*

Training via GD: Let $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ be n training samples drawn i.i.d. from the joint distribution of $(\boldsymbol{X}, Y)$ in (1). The most common training procedure is to minimise the empirical logistic risk

$$\widehat{\mathcal{L}}(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^n \ell(Y_i \boldsymbol{X}_i^\top \boldsymbol{w}),$$

over $\boldsymbol{w} \in \mathbb{R}^d$ via GD, initialised at the origin with a fixed step size $\gamma > 0$:

$$\boldsymbol{w}_0 = \boldsymbol{0}, \qquad \boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \gamma \nabla \widehat{\mathcal{L}}(\boldsymbol{w}_t) \quad \text{for } t \geq 0\,.$$

Let $\mathbf{X}_{1:n} \in \mathbb{R}^{n\times d}$ denote the design matrix whose rows are $\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top$, and let $\|\mathbf{X}_{1:n}\|$ denote its spectral norm. To ensure monotone descent, we set the step size to $\gamma \leq \min(1/\beta, 1)$, where $\beta := \|\mathbf{X}_{1:n}\|^2/n$ is the smoothness constant of $\widehat{\mathcal{L}}$ (see Lemma 7 in Appendix B.2).

Early stopping: We focus on the overparameterised regime with $d \geq n$, where two contrasting phenomena arise that explain the empirical observations in Figure 1. On one hand, the training samples $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ are almost surely *linearly separable*, that is, there exists some $\boldsymbol{w} \in \mathbb{R}^d$ such that

$$\min_{1\leq i\leq n} Y_i \boldsymbol{X}_i^\top \boldsymbol{w} > 0\,.$$

On the other hand, the Bayes risk $\mathcal{E}(\boldsymbol{w}^*) > 0$, implying the population distribution is *not* linearly separable, even when $p = 0$. In this regime, the implicit bias of GD under the empirical logistic risk [32, 17] drives the GD iterates to diverge in norm while converging in direction to the max-margin classifier, an interpolating solution that can incur suboptimal excess risk, as we prove in Section 4. To prevent this, we propose to stop GD early, at iteration

$$\tau := \inf\{t \geq 1 : \|\boldsymbol{w}_t\| \geq 4\,\|\boldsymbol{w}^*\|\}. \tag{2}$$

Lemma 2 (Existence of τ)**.** *Suppose $n \leq d$. Then $\|\boldsymbol{w}_t\| \to \infty$ as $t \to \infty$ almost surely, and so $\tau = \inf\{t \geq 1 : \|\boldsymbol{w}_t\| \geq 4\,\|\boldsymbol{w}^*\|\}$ is finite.*

Our stopping rule in (2) ensures that the early-stopped iterate $\boldsymbol{w}_\tau$ remains bounded in the sense of Lemma 3 (i) below. On a technical level, this allows us to localise the iterate $\boldsymbol{w}_\tau$, and thus apply a refined concentration analysis yielding the sharp upper bound in Theorem 1 below. At the same time, stopping too early would prevent the empirical logistic risk $\widehat{\mathcal{L}}(\boldsymbol{w}_\tau)$ from decreasing sufficiently, which is needed to ensure progress towards a good classifier. As is customary in risk decomposition, we show that our stopping time also guarantees $\widehat{\mathcal{L}}(\boldsymbol{w}_\tau) \leq \widehat{\mathcal{L}}(\boldsymbol{w}^*_{\mathcal{L}})$ (Lemma 3 (ii)).

[1] If, instead, one is interested in measuring the excess zero-one risk with respect to the latent labels $\widetilde{Y}$, note that the identity $\mathcal{E}(\boldsymbol{w}) = p + (1-2p)\mathbb{P}(\operatorname{sign}(\boldsymbol{X}^\top \boldsymbol{w}) \neq \widetilde{Y})$ gives an exact characterisation. Hence, the Bayes-optimal scores coincide under both risk measures and the excess risks differ only by a factor of $1-2p$. By the same identity, when $p = 0.5$, i.e., Y is random noise, $\mathcal{E}(\boldsymbol{w}) = 0.5$ for any $\boldsymbol{w} \in \mathbb{R}^d$ and the excess risk is always 0. To avoid such degenerate cases, we fix $p < 0.5$.

Lemma 3 (Properties of the early-stopped iterate)**.** *The early-stopped iterate $\boldsymbol{w}_\tau$ satisfies the following: (i) $\|\boldsymbol{w}_\tau\| < 4\|\boldsymbol{w}^*\| + 1$; and (ii) $\widehat{\mathcal{L}}(\boldsymbol{w}_\tau) \leq \widehat{\mathcal{L}}(\boldsymbol{w}^*_{\mathcal{L}})$.*

Similar oracle stopping rules, typically derived from bias-variance tradeoffs, are commonly used in the statistical analysis of early stopping [8, 41, 22, 27, 40].

We detail the proof of Lemma 1 in Appendix B.1 and the proofs of Lemmas 2–3 in Appendix B.2.

Remark 1 (Scaling regime)**.** *The condition $d \geq n$ is not inherently needed for our early-stopped GD minimax-optimality results below to hold. The upper bound in Theorem 1 below applies to any $\boldsymbol{w} \in \mathbb{R}^d$ satisfying the two properties of Lemma 3, and in particular, to the early-stopped iterate $\boldsymbol{w}_\tau$. The only role of $d \geq n$ is to guarantee that the iterates diverge in norm, hence τ is finite (Lemma 2). However, divergence can still occur, for instance, when $d < n$ and the data has a non-empty separable part [17]. The statistical lower bound in Theorem 2 below does not require $d \geq n$ either.*

3 Minimax Optimality of Early-Stopped GD

In this section, we present our main results, establishing the statistical optimality of GD at the oracle stopping time τ. We first prove a high-probability upper bound on the excess zero-one risk of the early-stopped GD iterate $\boldsymbol{w}_\tau$ (Section 3.1), and then a lower bound that applies to any arbitrary classifier, including the linear classifier induced by $\boldsymbol{w}_\tau$ (Section 3.2). Finally, we show that these bounds match as functions of the sample size n, for the FCD spectral class, including polynomially and exponentially-decaying spectra (Section 3.3). Throughout, our results depend on the spectrum of $\boldsymbol{\Sigma}$, whose eigenvalues we denote by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d > 0$.

3.1 Upper Bound for Early-Stopped GD

We first present our excess zero-one risk upper bound for the early-stopped GD iterate $\boldsymbol{w}_\tau$. This bound involves the spectrum of $\boldsymbol{\Sigma}$ through a soft-thresholding operator, namely the *effective rank* of $\boldsymbol{\Sigma}$, defined for every $\eta > 0$ as

$$r_{\text{eff}}(\eta) := \operatorname{tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \eta \boldsymbol{I}_d)^{-1}) = \sum_{i=1}^{d} \frac{\lambda_i}{\lambda_i + \eta}\,.$$

This corresponds to the effective dimension commonly used in analyses of ridge regression, e.g., [10]. We also recall from Lemma 1 that the population logistic risk minimiser has the form $\boldsymbol{w}^*_{\mathcal{L}} = \rho^* \boldsymbol{w}^*$ where $\rho^* \in (0, 2]$. These quantities will be used in the theorem below.

Theorem 1 (Upper bound for early-stopped GD)**.** *There exist a universal constant $c_0 > 0$ and some $c_1, c_2, c_3, c_4 > 0$ depending on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ only, but not on n or p, such that the following holds:*

Fix $\delta \in (0, 1)$ and $n \geq 3$. Suppose there exists $\eta_n > 0$ satisfying $n \geq c_0(r_{\text{eff}}(\eta_n) + 1 + \log(2/\delta))$ and

$$\Psi(\eta_n) := \frac{r_{\text{eff}}(\eta_n) + \log(1/\delta) + \log\log(n) + c_2}{n} + c_3\eta_n \leq c_4(\rho^*)^2 \|\boldsymbol{w}^*\|^2_{\boldsymbol{\Sigma}}\,. \tag{3}$$

Then, with probability at least $1 - \delta$,

$$\mathcal{E}(\boldsymbol{w}_\tau) - \mathcal{E}(\boldsymbol{w}^*) \leq c_1 \frac{1 - 2p}{(\rho^*)^2 \|\boldsymbol{w}^*\|_{\boldsymbol{\Sigma}}} \Psi(\eta_n).$$

The proof of Theorem 1, including explicit constants, is provided in Appendix C, with a preview of the main technical challenges and new proof techniques provided in Section 3.1.1. We now interpret the key quantities in the bound to build intuition:

The upper bound depends on $\Psi(\eta_n)$, which must be small enough for condition (3) to hold. Since the right hand side of (3) is constant in n, this is ensured if $\Psi(\eta_n) \to 0$ as $n \to \infty$. Moreover, the bound $\mathcal{E}(\boldsymbol{w}_\tau) - \mathcal{E}(\boldsymbol{w}^*) \lesssim \Psi(\eta_n)$ shows that the same behaviour suffices for vanishing excess risk. Hence, condition (3) imposes no restrictions beyond those already sufficient for vanishing excess risk.

The tightness of the upper bound is governed by the choice of η_n through $\Psi(\eta_n)$, which consists of two terms: one involving $r_{\text{eff}}(\eta_n)$ and one linear in η_n. Since $r_{\text{eff}}(\eta_n)$ decreases with η_n, this induces a tradeoff that determines the optimal choice of η_n. For commonly studied spectral decays, such as

polynomial ($\lambda_i \asymp i^{-2a}$ for $a > \frac{1}{2}$) and exponential ($\lambda_i \asymp e^{-bi}$ for $b > 0$), this tradeoff can be solved analytically, yielding the optimal rates $n^{-2a/(2a+1)}$ and $\log n/n$, respectively. In Section 3.3, we prove that this optimality extends more generally to fast and continuously decaying spectrum (FCD).

A natural notion of signal-to-noise ratio (SNR) in the Gaussian mixtures model is $\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}}$, which is equal to $\|\boldsymbol{w}^*\|_{\boldsymbol{\Sigma}}$. Our upper bound scales inversely with the SNR, as expected: as the noise level increases, the SNR decreases, leading to a looser upper bound, consistent with the deterioration in the performance of the early-stopped GD iterate.

3.1.1 New Proof Techniques

The proof of Theorem 1 has several technical steps; we highlight the ones capturing the main novelty.

A new calibration result: Standard calibration arguments [42, 5] compare the excess zero-one risk to the excess logistic risk relative to the *unrestricted* logistic optimum:

$$\mathcal{E}(\boldsymbol{w}) - \mathcal{E}(\boldsymbol{w}^*) \lesssim \sqrt{\mathcal{L}(\boldsymbol{w}) - \min_{h:\mathbb{R}^d \to \mathbb{R}} \mathcal{L}(h)}\,. \tag{4}$$

This bound has two limitations. First, the square-root dependence leads to suboptimal rates in the excess zero-one risk guarantees. Second, and more importantly, the minimum of $\mathcal{L}(h)$ is taken over *all* measurable functions $h : \mathbb{R}^d \to \mathbb{R}$, rather than linear score functions. Consequently, the $\mathcal{L}(\boldsymbol{w}) - \min_{h:\mathbb{R}^d \to \mathbb{R}} \mathcal{L}(h)$ bound in (4) is strictly looser than one based on $\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}^*_{\mathcal{L}})$, with an irreducible difference $\mathcal{L}(\boldsymbol{w}^*_{\mathcal{L}}) - \min_{h:\mathbb{R}^d \to \mathbb{R}} \mathcal{L}(h)$.

In well-specified models such as the sigmoid, Proposition 2.1.B of [40] overcomes this second limitation, albeit with the suboptimal square-root dependence. However, this result is not directly transferable to our Gaussian mixtures model when $p > 0$, since it is misspecified under the logistic loss.

Together, the square-root rate, the misspecification of our model under logistic loss, and the fact that the unrestricted logistic risk minimiser under our model is nonlinear when $p > 0$ imply that existing calibration results yield not only suboptimal rates, but rates that do not vanish as $n \to \infty$.

We overcome these limitations via a new calibration result in Lemma 4. Its proof, provided in Appendix C.1, leverages the local α-strong convexity of $\mathcal{L}$ around $\boldsymbol{w}^*_{\mathcal{L}}$ (see Lemma 9).

Lemma 4 (Linear-rate calibration for linear classifiers)**.** *Let $\boldsymbol{w} \in \mathbb{R}^d$ satisfy $\|\boldsymbol{w}\| \le 4\|\boldsymbol{w}^*\| + 1$ and*

$$\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}^*_{\mathcal{L}}) \le \frac{\alpha(\rho^*)^2 \|\boldsymbol{w}^*\|^2_{\boldsymbol{\Sigma}}}{8}\,, \tag{5}$$

where $\alpha = \frac{1}{2}\ell''\big(3(4\|\boldsymbol{w}^*\| + 1)\sqrt{\|\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top\|}\big)$. Then,

$$\mathcal{E}(\boldsymbol{w}) - \mathcal{E}(\boldsymbol{w}^*) \le \frac{8(1-2p)}{\sqrt{2\pi}\alpha(\rho^*)^2 \|\boldsymbol{w}^*\|_{\boldsymbol{\Sigma}}} \left(\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}^*_{\mathcal{L}})\right).$$

The right hand side of (5) is constant in n, so this condition is satisfied whenever the excess logistic risk vanishes as $n \to \infty$. Thus, bounding the excess zero-one risk reduces to bounding the excess logistic risk. We now focus on this step, highlighting the technical ingredients that depart from prior work.

Bounding excess logistic risk using effective rank and localisation: Since $\|\boldsymbol{w}_\tau\| \le 4\,\|\boldsymbol{w}^*\| + 1$ (Lemma 3 (i)), it suffices to bound the excess logistic risk $\mathcal{L}(\boldsymbol{w}_\tau) - \mathcal{L}(\boldsymbol{w}^*_{\mathcal{L}})$ of the early-stopped GD iterate $\boldsymbol{w}_\tau$ over the ball $\mathcal{B} = \{\boldsymbol{w} : \|\boldsymbol{w}\| \le 4\,\|\boldsymbol{w}^*\| + 1\}$.

For this, [40, Theorem 3.2] uses a standard risk decomposition combined with Rademacher complexity, yielding a bound of the form $k/n + \sqrt{\sum_{i>k} \lambda_i / n}$. We improve this in two ways: (i) We replace hard thresholding at the k-th eigenvalue with the soft effective rank $r_{\mathrm{eff}}(\eta_n)$ to obtain finer spectral control (see Appendix C.2). In [40], the bias-variance tradeoff is governed by the discrete parameter k, whereas in our bound it is governed by the continuous parameter η_n, allowing for more effective optimisation of the bound by tuning η_n. (ii) We perform a refined risk deviation analysis by localising to low-excess-risk regions, restricting $\mathcal{B}$ to the shell $\{\boldsymbol{w} : \boldsymbol{w} \in \mathcal{B}, \mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}^*_{\mathcal{L}}) \le r\}$ for a fixed $r > 0$, and extending the resulting bounds to $\mathcal{B}$ via a peeling argument over r (see Appendix C.3). Together, these techniques allow us to remove the square-root dependence in the bound.

3.2 Statistical Lower Bound

To establish the statistical optimality of early-stopped GD, we provide a statistical lower bound for *arbitrary* scores, a strictly larger class than the linear scores obtained by logistic regression.

We work in the Gaussian mixture model (1), where the distribution of $(\boldsymbol{X}, Y)$ is parametrised by the centre $\boldsymbol{\mu}$, the noise covariance $\boldsymbol{\Sigma}$, and the label-flip probability p. For the purpose of the statistical lower bound, we fix $\boldsymbol{\Sigma}$ and p, and denote the resulting distribution of $(\boldsymbol{X}, Y)$ by $P_{\boldsymbol{\mu}}$. Then the n i.i.d. samples $S := (\boldsymbol{X}_i, Y_i)_{i=1}^n$ are distributed as $P_{\boldsymbol{\mu}}^{\otimes n}$. For any procedure $\widehat{h}$ that takes S as input, we denote its output (a learned score) by $\widehat{h}_S : \mathbb{R}^d \to \mathbb{R}$, which is measurable with respect to S.

Our lower bound involves an eigenvalue count $N(\widetilde{\eta})$, defined via the hard-thresholding operator $N(\cdot)$ as the number of eigenvalues of $\boldsymbol{\Sigma}$ larger than a fixed $\widetilde{\eta} > 0$:

$$N(\widetilde{\eta}) = \big|\{i : i \in [d],\, \lambda_i \geq \widetilde{\eta}\}\big|.$$

This hard-thresholding operator $N(\cdot)$ can be viewed as the discrete analogue of the soft-thresholding operator $r_{\mathrm{eff}}(\cdot)$ in Theorem 1; their outputs correspond to hard and soft effective dimension notions.

Theorem 2 (Statistical lower bound)**.** *There exists a universal constant $c > 0$ such that the following holds: Fix n and let $\widetilde{\eta}_n \in (0, \lambda_2/2]$ be small enough so that $N(2\widetilde{\eta}_n) > 9$. Let $\widehat{h}$ be any procedure. Then, there exists $\boldsymbol{\mu} \in \mathbb{R}^d$ with $\left\|\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right\| \leq \sqrt{2/\lambda_1}$ such that*

$$\mathop{\mathbb{E}}_{S \sim P_{\boldsymbol{\mu}}^{\otimes n}} \left[\mathcal{E}(\widehat{h}_S) - \mathcal{E}(\boldsymbol{w}^*)\right] \geq c(1-2p) \min\left\{\frac{2\widetilde{\eta}_n}{\lambda_1}, \frac{N(2\widetilde{\eta}_n) - 1}{n}\right\}.$$

This lower bound applies to *any* procedure that outputs a score function $\widehat{h}_S$ based on the samples S. In particular, this includes the linear score induced by the early-stopped iterate $\boldsymbol{w}_\tau$ of Theorem 1. As is standard in learning theory, a lower bound in expectation can be converted into a statement in probability via Markov's inequality [3], making it directly comparable to Theorem 1.

To clarify the role of the tuning parameter $\widetilde{\eta}_n$, note that $\widetilde{\eta}_n$ captures a tradeoff between the two terms in the minimum, one increasing and the other decreasing in $\widetilde{\eta}_n$. Although the effective rank $r_{\mathrm{eff}}(\eta_n)$ in the upper bound (Theorem 1) and the eigenvalue count $N(\widetilde{\eta}_n)$ in the lower bound (Theorem 2) are not directly comparable, we show in Section 3.3 that, under a common spectral regularity assumption, they are of the same order, making the two bounds comparable. In fact, with an appropriate choice of tuning parameters satisfying $\widetilde{\eta}_n = \eta_n/2$, the bounds match as functions of n (see Theorem 3).

We provide the proof of Theorem 2 in Appendix D. The proof proceeds in two steps. We first establish a triangle-inequality-type bound relating the excess zero-one risks of two centres $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ to their distance in $\boldsymbol{\Sigma}$-geometry (Lemma 18). We then construct a packing within a suitable subspace of $\{\boldsymbol{\mu} : \left\|\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right\| \leq \sqrt{2/\lambda_1}\}$ (Lemma 20), and use it to perform a reduction to hypothesis testing.

3.3 Statistical Optimality for Fast and Continuously Decaying Covariance Spectrum

In this section, we combine Theorems 1 and 2 to show that early-stopped GD is statistically optimal for the FCD spectral class. We begin by defining a FCD spectrum, following [39, Assumption 3].

Assumption 1 (Fast and Continuously Decaying spectrum)**.** *For all sufficiently small $\eta > 0$, the spectrum of $\boldsymbol{\Sigma}$ satisfies*

$$\frac{1}{\eta} \sum_{i:\lambda_i < \eta} \lambda_i \lesssim N(\eta). \tag{FCD}$$

Canonical examples of such spectral decays include polynomial ($\lambda_i \asymp i^{-2a}$ with $a > 1/2$) and exponential ($\lambda_i \asymp e^{-bi}$ with $b > 0$) decay. In linear regression, these spectra are known to prevent benign overfitting [6]. While this observation does not directly transfer to classification, it motivates early-stopped GD as an alternative to interpolation in these regimes. We next show that early-stopped GD is, in fact, statistically optimal for FCD spectra, and achieves vanishing excess zero-one risk:

Theorem 3 (Statistical optimality of early-stopped GD for FCD spectrum)**.** *Define*

$$\eta_n^* := \sup\{\eta > 0 : N(\eta) \geq n\eta\},$$

and assume that (FCD) *holds on* $(0, 2\eta_n^*]$. *Then,* $r_{\text{eff}}(\eta) \asymp N(\eta)$ *for all* $\eta \in (0, \eta_n^*]$.

Moreover, if $\eta_n^* \to 0$ *as* $n \to \infty$, *and* $\log\log(n)/n \lesssim \eta_n^*$, *then the upper bound in Theorem 1 with* $\eta_n = \eta_n^*$ *matches the statistical lower bound in Theorem 2 with* $\widetilde{\eta}_n = \eta_n^*/2$, *with common rate* $\asymp \eta_n^*$.

The implications of Theorem 3 are twofold. First, under Assumption 1, it identifies an interval $(0, \eta_n^*]$ on which the effective rank $r_{\text{eff}}(\eta)$ and the eigenvalue count $N(\eta)$, appearing in the upper and lower bounds of Theorems 1 and 2, respectively, are of the same order for all η. In this regime, these two notions of effective dimensions agree up to constants, reconciling the difference between the bounds. Second, it establishes that the optimal tuning parameters satisfy $\eta_n = \eta_n^*$ and $\widetilde{\eta}_n = \eta_n^*/2$. With this choice, the upper and lower bounds match, giving the common rate $\asymp \eta_n^*$. This rate is analytically tractable for common spectral decays, as we show next in Corollary 1. The proofs of Theorem 3 and Corollary 1 are provided in Appendix E.

Corollary 1 (Minimax rates for polynomially and exponentially-decaying spectra)**.** *For polynomial decay* $\lambda_i \asymp i^{-2a}$ *with* $a > 1/2$, *there exists* $\boldsymbol{\mu} \in \mathbb{R}^d$ *with* $\left\|\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right\| \leq \sqrt{2/\lambda_1}$ *such that*

$$\mathcal{E}(\boldsymbol{w}_\tau) - \mathcal{E}(\boldsymbol{w}^*) \asymp \eta_n^* \asymp n^{-\frac{2a}{2a+1}} .$$

For exponential decay $\lambda_i \asymp e^{-bi}$ *with* $b > 0$, *there exists* $\boldsymbol{\mu} \in \mathbb{R}^d$ *with* $\left\|\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right\| \leq \sqrt{2/\lambda_1}$ *such that*

$$\mathcal{E}(\boldsymbol{w}_\tau) - \mathcal{E}(\boldsymbol{w}^*) \asymp \eta_n^* \asymp \frac{\log n}{n} .$$

4 Interpolation Lower Bound

To show the benefit of early-stopped GD over interpolation, we prove a lower bound for *all* interpolating linear classifiers, extending [40, Theorem 4.2] by removing any *sparsity* assumption on the Bayes direction $\boldsymbol{w}^*$, so the result holds uniformly over all $\boldsymbol{w}^*$. This generalisation is essential for enabling a meaningful comparison with our minimax rates in Corollary 1: the worst case $\boldsymbol{\mu}$ does not guarantee a sparse $\boldsymbol{w}^*$, so the interpolation lower bound of [40] that assumes sparsity does not apply.

We now state this result formally. Its full proof appears in Appendix F, with a sketch in Appendix F.1.

Theorem 4 (Interpolation lower bound)**.** *Suppose* $\|\boldsymbol{w}^*\|_{\boldsymbol{\Sigma}} \asymp 1$. *Fix any* $\mathtt{c} \in (0,1)$ *and* $\delta \in (0,1)$, *and let* $k \in \{1, \ldots, n\}$ *be arbitrary. There exist constants* $\mathtt{c}_1, \mathtt{c}_2, \mathtt{c}_3 > 0$ *such that if*

$$n \geq \mathtt{c}_1 \cdot \left(k \log \tfrac{\sqrt{k}}{\mathtt{c}} + k \log\log \tfrac{2k}{\delta} + \log \tfrac{4}{\delta}\right) \quad \textit{and} \quad d - k \geq \mathtt{c}_2 n\,, \tag{6}$$

then, defining the set of interpolators $\mathcal{I} := \{\boldsymbol{w} \in \mathbb{R}^d : \min_i Y_i \boldsymbol{X}_i^\top \boldsymbol{w} > 0\}$, *with probability at least* $1\text{–}\delta$,

$$\forall \boldsymbol{w} \in \mathcal{I},\ \mathcal{E}(\boldsymbol{w}) - \mathcal{E}(\boldsymbol{w}^*) \geq \frac{\mathtt{c}_3 \cdot (1-2p)n}{\left(\sqrt{n} + \sqrt{d-k} + \sqrt{2\log(2/\delta)}\right)^2} . \tag{7}$$

The tuning parameter k trades off sample size requirement in (6) and the tightness of the lower bound in (7): larger values of k require larger n but yield tighter lower bounds. In the high-dimensional regime $n/d \to 0$, choosing k such that $k/d \to 0$ (e.g., $k = 1$) gives $\mathcal{E}(\boldsymbol{w}) - \mathcal{E}(\boldsymbol{w}^*) \gtrsim n/d$, matching the excess mean-squared error lower bound for interpolators in regression [26].

In addition, when $d \asymp n \log n$, Theorem 4 implies that the max-margin classifier $\boldsymbol{w}_\infty$, i.e., the interpolator that GD converges to in direction, satisfies the following lower bound:

$$\mathcal{E}(\boldsymbol{w}_\infty) - \mathcal{E}(\boldsymbol{w}^*) \gtrsim \frac{1}{\log n}.$$

By contrast, Corollary 1 shows that early-stopped GD achieves the minimax rates $n^{-2a/(2a+1)}$ and $\log n/n$ under polynomial and exponential spectral decays, both of which vanish strictly faster than $1/\log n$. Taken together, early stopping is sufficient for minimax optimality in these canonical FCD spectral regimes, and necessary when additionally, d is mildly overparameterised, e.g., $d \asymp n \log n$.

5 Conclusion

We show that early-stopped GD on the logistic loss for Gaussian mixture classification is minimax-optimal for covariance spectra with fast and continuous decay, including polynomial and exponential decays (Theorem 3). The corresponding minimax rates are explicit (Corollary 1) and closely matched by empirical performance of GD (Figure 2). We further show that early stopping can be substantially more sample-efficient than interpolation in common overparameterised regimes (Theorem 4), establishing a statistical separation between the two, and thus the necessity of early stopping to achieve minimax optimality.

The stopping time studied in our work is oracle-based. This is standard in statistical analyses of early stopping [8, 41, 22, 27, 40], where the stopping rule commonly identifies the statistically-optimal scale given by a bias-variance decomposition. Thus, our stopping time should be viewed as an existence result, rather than an algorithm ready to be deployed in practice. Existing data-dependent rules are best understood in regression settings such as non-parametric least squares [28], where the losses used in testing and training coincide. Extending such guarantees to noisy classification with logistic loss presents additional challenges. For instance, our theory aims to minimise the zero-one risk, but the logistic loss is used instead as a surrogate.

Acknowledgments and Disclosure of Funding

A. Buna and S. X. Liu worked closely and discussed every aspect of this work in depth. A. Buna led the development of the upper bound, statistical lower bound and statistical optimality results, and carried out the numerical experiments. S. X. Liu wrote the large majority of the main text, proved the interpolation lower bound, led comparisons with prior work, and checked and refined all proofs. P. Rebeschini helped formulate the problem, supervised the project and helped refine the paper.

A. Buna was supported by EPSRC through the StatML CDT programme [project reference EP/S023151/1]. S. X. Liu and P. Rebeschini were supported by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number EP/Y028333/1]. The authors declare no competing interests.

References

[1] Navid Ardeshir, Clayton Sanford, and Daniel J Hsu. Support vector machines and linear regression coincide with very high-dimensional features. *Advances in Neural Information Processing Systems*, 34:4907–4918, 2021.

[2] Yaroslav Averyanov and Alain Celisse. Early stopping and polynomial smoothing in regression with reproducing kernels. *arXiv preprint arXiv:2007.06827*, 2020.

[3] Francis Bach. *Learning theory from first principles*. MIT press, 2024.

[4] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4), August 2005.

[5] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[6] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[7] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

[8] Peter Bühlmann and Bin Yu. Boosting With the L2 Loss. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

[9] Yuan Cao, Quanquan Gu, and Misha Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. In *Advances in Neural Information Processing Systems*, 2021.

[10] A. Caponnetto and E. De Vito. Optimal Rates for the Regularized Least-Squares Algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, July 2007.

[11] Niladri S. Chatterji and Philip M. Long. Finite-sample Analysis of Interpolating Linear Classifiers in the Overparameterized Regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021.

[12] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2):435–495, 2022.

[13] Ichiro Hashimoto, Stanislav Volgushev, and Piotr Zwiernik. Universality of benign overfitting in binary linear classification. *arXiv preprint arXiv:2501.10538*, 2025.

[14] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

[15] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2 edition, 2009.

[16] Daniel Hsu, Vidya Muthukumar, and Ji Xu. On the proliferation of support vectors in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 91–99. PMLR, 2021.

[17] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1772–1798. PMLR, 25–28 Jun 2019.

[18] Abla Kammoun and Mohamed-Slim Alouini. On the precise error analysis of support vector machines. *IEEE Open Journal of Signal Processing*, 2:99–118, 2021.

[19] Varun Kanade, Patrick Rebeschini, and Tomas Vaškevičius. The statistical complexity of early-stopped mirror descent. *Information and Inference: A Journal of the IMA*, 12(4):3010–3041, 2023.

[20] Ganesh Ramachandra Kini and Christos Thrampoulidis. Analytic study of double descent in binary classification: The impact of loss. In *International Symposium on Information Theory*, pages 2527–2532. IEEE, 2020.

[21] Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pages 110–133, 2017.

[22] Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017.

[23] Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, and Lenka Zdeborova. The role of regularization in classification of high-dimensional noisy gaussian mixture. In *International conference on machine learning*, pages 6874–6883. PMLR, 2020.

[24] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime. *The Annals of Statistics*, 53(2):822 – 853, 2025.

[25] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: does the loss function matter? *J. Mach. Learn. Res.*, 22(1), January 2021.

[26] Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.

[27] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems*, 31, 2018.

[28] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

[29] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The Impact of Regularization on High-dimensional Logistic Regression. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[30] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The performance analysis of generalized margin maximizers on separable data. In *International conference on machine learning*, pages 8417–8426. PMLR, 2020.

[31] Ohad Shamir. Gradient methods never overfit on separable data. *Journal of Machine Learning Research*, 22(85):1–20, 2021.

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

[33] Arun Sai Suggala, Adarsh Prasad, and Pradeep Ravikumar. Connecting optimization and regularization paths. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 10631–10641, Red Hook, NY, USA, 2018. Curran Associates Inc.

[34] Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[35] Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023.

[36] Alexander Tsigler, Luiz FO Chamon, Spencer Frei, and Peter L Bartlett. Benign overfitting and the geometry of the ridge regression solution in binary classification. *arXiv preprint arXiv:2503.07966*, 2025.

[37] Ke Wang and Christos Thrampoulidis. Binary Classification of Gaussian Mixtures: Abundance of Support Vectors, Benign Overfitting, and Regularization. *SIAM Journal on Mathematics of Data Science*, 4(1):260–284, 2022.

[38] Yuting Wei, Fanny Yang, and Martin J Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. *IEEE Transactions on Information Theory*, 65(10):6685–6703, 2019.

[39] Jingfeng Wu, Peter L. Bartlett, Sham M. Kakade, Jason D. Lee, and Bin Yu. Risk comparisons in linear regression: Implicit regularization dominates explicit regularization. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 6849–6851, 2026. Extended abstract; full version at arXiv:2509.17251.

[40] Jingfeng Wu, Peter L. Bartlett, Matus Telgarsky, and Bin Yu. Benefits of early stopping in gradient descent for overparameterized logistic regression. In *Proceedings of the 42nd International Conference on Machine Learning*. JMLR.org, 2025.

[41] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On Early Stopping in Gradient Descent Learning. *Constructive Approximation*, 26(2):289–315, August 2007.

[42] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

Technical Appendices and Supplementary Material

Contents

1 Introduction **1**

1.1 Main Contributions . . . 3

1.2 Related Work . . . 3

2 Problem Setup **4**

3 Minimax Optimality of Early-Stopped GD **6**

3.1 Upper Bound for Early-Stopped GD . . . 6

3.1.1 New Proof Techniques . . . 7

3.2 Statistical Lower Bound . . . 8

3.3 Statistical Optimality for Fast and Continuously Decaying Covariance Spectrum . . . 8

4 Interpolation Lower Bound **9**

5 Conclusion **10**

A Gaussian Mixtures Model Properties **14**

B Proofs Omitted from Section 2 **15**

B.1 Population Logistic Risk Minimiser . . . 15

B.2 Optimisation Lemmas . . . 16

C Proof of the Early-Stopped GD Upper Bound **18**

C.1 Zero-One to Logistic Calibration via $\boldsymbol{\Sigma}$-geometry . . . 18

C.2 Effective Rank and Empirical Covariance . . . 21

C.2.1 Effective Rank of Population Covariances . . . 21

C.2.2 Spectral Control of the Empirical Covariance via Effective Rank . . . 22

C.3 Localisation and Concentrations . . . 23

C.3.1 Useful Properties of $\mathcal{G}(r)$ . . . 23

C.3.2 Empirical Excess Logistic Risk Concentration on $\mathcal{G}(r)$ . . . 24

C.3.3 Empirical Excess Logistic Risk Concentration on $\mathcal{B}$ . . . 27

C.4 Proof of Theorem 1 . . . 29

D Proof of the Statistical Lower Bound **30**

D.1 A Triangle Inequality for Excess Zero-One Risk . . . 30

D.2 Information Theory Tools . . . 33

D.3 Proof of Theorem 2 . . . 35

E Proofs for Fast and Continuously Decaying Spectrum **37**

F Proof of the Interpolation Lower Bound **39**
F.1 Setup via Whitening and Rotation 39
F.2 Preliminary Lemmas 40
F.3 Signal Component Induces Large Negative Margins on Constant Fraction of Samples 41
F.4 Nuisance Component Cannot Fix Large Negative Margins for Affected Samples . . 43
F.5 Proof of Theorem 4 44

Additional Notation

We will use additional notation throughout the supplementary material. For a positive integer k, we will denote $[k] = \{1, 2, \dots, k\}$. Given a symmetric positive definite matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, we use $\lambda_i(\mathbf{M})$ to denote its i-th largest eigenvalue. For vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we denote their $\mathbf{M}$-induced inner product as $\langle \boldsymbol{x}, \boldsymbol{y}\rangle_{\mathbf{\Sigma}} = \boldsymbol{x}^\top \mathbf{\Sigma} \boldsymbol{y}$. We use f_X to denote the PDF of the random variable X, and ϕ and Φ to denote the PDF and CDF, respectively, of $\mathcal{N}(0, 1)$. Throughout our proofs, we will denote the sigmoid function by

$$\sigma(z) = \frac{1}{1 + \exp(-z)},$$

and the population covariance of the covariates by

$$\mathbf{\Gamma} = \mathrm{Cov}(\boldsymbol{X}) = \mathrm{Cov}(\widetilde{Y}\boldsymbol{\mu} + \boldsymbol{\varepsilon}) = \mathbf{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top.$$

A Gaussian Mixtures Model Properties

In this section, we collect some useful properties of the Gaussian mixtures model introduced in (1).

Lemma 5 (Conditional probabilities and excess zero-one risk)**.** *For Gaussian mixtures, we have for $y \in \{\pm 1\}$ and $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\mathbb{P}(\widetilde{Y} = y \mid \boldsymbol{X} = \boldsymbol{x}) = \sigma(2y\boldsymbol{x}^\top \boldsymbol{w}^*),$$
$$\mathbb{P}(Y = y \mid \boldsymbol{X} = \boldsymbol{x}) = p + (1 - 2p)\sigma(2y\boldsymbol{x}^\top \boldsymbol{w}^*).$$

Moreover, for any $\boldsymbol{w} \in \mathbb{R}^d$, the excess zero-one risk can be expressed as

$$\mathcal{E}(\boldsymbol{w}) - \mathcal{E}(\boldsymbol{w}^*) = (1 - 2p)\mathbb{E}\left[|\tanh(\boldsymbol{X}^\top \boldsymbol{w}^*)| \mathbb{1}\{\mathrm{sign}(\boldsymbol{X}^\top \boldsymbol{w}) \neq \mathrm{sign}(\boldsymbol{X}^\top \boldsymbol{w}^*)\}\right]. \tag{8}$$

Proof. Although these results are well-known, we prove them here for completeness. Start by noting that for $\widetilde{y} \in \{\pm 1\}$,

$$\boldsymbol{X} \mid \widetilde{Y} = \widetilde{y} \sim \mathcal{N}(\widetilde{y}\boldsymbol{\mu}, \mathbf{\Sigma}).$$

Together with the fact that $\mathbb{P}(\widetilde{Y} = \pm 1) = 1/2$, the Bayes' rule gives:

$$\mathbb{P}(\widetilde{Y} = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{f_{\boldsymbol{X}|\widetilde{Y}=1}(\boldsymbol{x})}{f_{\boldsymbol{X}|\widetilde{Y}=1}(\boldsymbol{x}) + f_{\boldsymbol{X}|\widetilde{Y}=-1}(\boldsymbol{x})} = \sigma(2\boldsymbol{x}^\top \mathbf{\Sigma}^{-1}\boldsymbol{\mu}) = \sigma(2\boldsymbol{x}^\top \boldsymbol{w}^*).$$

Since $\sigma(-t) = 1 - \sigma(t)$, $\mathbb{P}(\widetilde{Y} = y \mid \boldsymbol{X} = \boldsymbol{x}) = \sigma(2y\boldsymbol{x}^\top \boldsymbol{w}^*)$. Furthermore, since $\widetilde{Y}$ is flipped with probability $p \in [0, 0.5)$ to obtain Y, we have

$$\mathbb{P}(Y = y \mid \boldsymbol{X} = \boldsymbol{x}) = p + (1 - 2p)\sigma(2y\boldsymbol{x}^\top \boldsymbol{w}^*).$$

Finally, conditioning on $\boldsymbol{X}$ and computing the excess zero-one risk gives

$$\mathcal{E}(\boldsymbol{w}) - \mathcal{E}(\boldsymbol{w}^*) = \mathbb{E}\left[|2\mathbb{P}(Y = 1 \mid \boldsymbol{X}) - 1| \mathbb{1}\{\mathrm{sign}(\boldsymbol{X}^\top \boldsymbol{w}) \neq \mathrm{sign}(\boldsymbol{X}^\top \boldsymbol{w}^*)\}\right],$$

and (8) follows from the identity $\tanh(z) = 2\sigma(2z) - 1$.

□

Lemma 6 (Subgaussian covariates)**.** *For any $\boldsymbol{w} \in \mathbb{R}^d$, $\boldsymbol{X}^\top \boldsymbol{w}$ is subgaussian with variance proxy $\|\boldsymbol{w}\|_{\boldsymbol{\Gamma}}^2$. Consequently, for any $t > 0$,*

$$\mathbb{P}(|\boldsymbol{X}^\top \boldsymbol{w}| > t) \leq 2 \exp\left(-\frac{t^2}{2\|\boldsymbol{w}\|_{\boldsymbol{\Gamma}}^2}\right). \tag{9}$$

Proof. Fix any $\boldsymbol{w} \in \mathbb{R}^d$ and any $t \in \mathbb{R}$. Then, by the independence of $\widetilde{Y}$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, we have

$$\begin{aligned}\mathbb{E}[\exp(t\boldsymbol{X}^\top \boldsymbol{w})] &= \mathbb{E}[\exp(t\widetilde{Y}\boldsymbol{\mu}^\top \boldsymbol{w})]\mathbb{E}[\exp(t\boldsymbol{\varepsilon}^\top \boldsymbol{w})] = \cosh(t\boldsymbol{\mu}^\top \boldsymbol{w}) \exp\left(\frac{t^2}{2}\boldsymbol{w}^\top \boldsymbol{\Sigma}\boldsymbol{w}\right) \\ &\leq \exp\left(\frac{t^2}{2}(\boldsymbol{w}^\top \boldsymbol{\Sigma}\boldsymbol{w} + (\boldsymbol{\mu}^\top \boldsymbol{w})^2)\right) = \exp\left(\frac{t^2}{2}\boldsymbol{w}^\top \boldsymbol{\Gamma}\boldsymbol{w}\right) = \exp\left(\frac{t^2}{2}\|\boldsymbol{w}\|_{\boldsymbol{\Gamma}}^2\right).\end{aligned}$$

Thus, $\boldsymbol{X}^\top \boldsymbol{w}$ is subgaussian with variance proxy $\|\boldsymbol{w}\|_{\boldsymbol{\Gamma}}^2$. The concentration inequality (9) follows from a Chernoff bound. □

B Proofs Omitted from Section 2

B.1 Population Logistic Risk Minimiser

We restate Lemma 1 and detail its proof.

Lemma 1 (Population logistic risk minimiser)**.** *The minimiser of $\mathcal{L}(\boldsymbol{w})$ over $\mathbb{R}^d$ is unique and has the form $\boldsymbol{w}_{\mathcal{L}}^* = \rho^* \boldsymbol{w}^*$ where $\rho^* \in (0, 2]$.*

Proof. Recall that $\boldsymbol{w}^* = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Fix an arbitrary $\boldsymbol{w} \in \mathbb{R}^d$ and decompose it as

$$\boldsymbol{w} = \rho \boldsymbol{w}^* + \boldsymbol{u},$$

where

$$\rho = \boldsymbol{w}^\top \boldsymbol{\mu}/(\boldsymbol{w}^*)^\top \boldsymbol{\mu} \in \mathbb{R} \quad \text{and} \quad \boldsymbol{u} = \boldsymbol{w} - \rho \boldsymbol{w}^*$$

By construction, $\boldsymbol{u}^\top \boldsymbol{\mu} = 0$, so $\boldsymbol{w}$ is split into a component along the signal direction $\boldsymbol{w}^*$ and a component $\boldsymbol{u}$ orthogonal to $\boldsymbol{\mu}$.

Step 1: We first prove that each minimiser of $\mathcal{L}(\cdot)$ satisfies $\boldsymbol{u} = \mathbf{0}_d$, i.e., it lies along $\boldsymbol{w}^*$. For this, let $T^* = \boldsymbol{X}^\top \boldsymbol{w}^*$ and $T_{\boldsymbol{u}} = \boldsymbol{X}^\top \boldsymbol{u}$. These random variables are independent. Indeed, since $\boldsymbol{w}^* = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and $\boldsymbol{u}^\top \boldsymbol{\mu} = 0$, we have

$$T^* = (\widetilde{Y}\boldsymbol{\mu} + \boldsymbol{\varepsilon})^\top \boldsymbol{w}^* = \widetilde{Y}\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}}^2 + \boldsymbol{\varepsilon}^\top \boldsymbol{w}^*, \tag{10}$$

$$T_{\boldsymbol{u}} = (\widetilde{Y}\boldsymbol{\mu} + \boldsymbol{\varepsilon})^\top \boldsymbol{u} = \boldsymbol{\varepsilon}^\top \boldsymbol{u}. \tag{11}$$

These are independent because $\widetilde{Y}$ is independent of $\boldsymbol{\varepsilon}$ and the Gaussians $\boldsymbol{\varepsilon}^\top \boldsymbol{w}^*$ and $\boldsymbol{\varepsilon}^\top \boldsymbol{u}$ are uncorrelated:

$$\mathrm{Cov}(\boldsymbol{\varepsilon}^\top \boldsymbol{w}^*, \boldsymbol{\varepsilon}^\top \boldsymbol{u}) = \boldsymbol{u}^\top \boldsymbol{\Sigma}\boldsymbol{w}^* = 0.$$

Now, for $t, s \in \mathbb{R}$, define

$$\zeta(t) = p + (1-2p)\sigma(2t), \qquad \vartheta_t(s) = \zeta(t)\ell(s) + (1-\zeta(t))\ell(-s)\,, \tag{12}$$

where recall $\sigma(z) = 1/(1+\exp(-z))$ denotes the sigmoid function and $\ell(z) = \log(1+\exp(-z))$ denotes the logistic loss. Then, using the decomposition of $\boldsymbol{w}$ and the tower law, we can write

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}) &= \mathbb{E}[\ell(Y\boldsymbol{X}^\top \boldsymbol{w})] = \mathbb{E}[\ell(Y\boldsymbol{X}^\top(\rho\boldsymbol{w}^* + \boldsymbol{u}))] = \mathbb{E}[\ell(Y(\rho T^* + T_{\boldsymbol{u}}))] \\ &= \mathbb{E}\left[\mathbb{E}[\ell(Y(\rho T^* + T_{\boldsymbol{u}})) \mid \boldsymbol{X}]\right] = \mathbb{E}[\vartheta_{T^*}(\rho T^* + T_{\boldsymbol{u}})]\,,\end{aligned}$$

where the last step applies $\mathbb{P}(Y = 1 \mid \boldsymbol{X}) = \zeta(T^*)$ due to Lemma 5. Since $\ell'' > 0$, the function ϑ_t is strictly convex for any $t \in \mathbb{R}$. Hence, conditioning on T^* and applying the tower law, we can write

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}) &= \mathbb{E}[\vartheta_{T^*}(\rho T^* + T_{\boldsymbol{u}})] = \mathbb{E}\left[\mathbb{E}[\vartheta_{T^*}(\rho T^* + T_{\boldsymbol{u}}) \mid T^*]\right] \\ &\overset{\text{(i)}}{\geq} \mathbb{E}\left[\vartheta_{T^*}(\rho T^* + \mathbb{E}[T_{\boldsymbol{u}} \mid T^*])\right] \overset{\text{(ii)}}{=} \mathbb{E}[\vartheta_{T^*}(\rho T^*)] = \mathcal{L}(\rho \boldsymbol{w}^*)\,,\end{aligned}$$

where (i) is Jensen's inequality applied to the strictly convex function ϑ_{T^*}, and (ii) uses the independence of T^* and $T_{\boldsymbol{u}}$, together with $\mathbb{E}[T_{\boldsymbol{u}}] = 0$. Equality in (i) holds if and only if $T_{\boldsymbol{u}} = 0$, which implies $\boldsymbol{u} = \mathbf{0}_d$.

Step 2: We have shown every minimiser of $\mathcal{L}(\cdot)$ in $\mathbb{R}^d$ takes the form $\rho \boldsymbol{w}^*$, so it remains to minimise over ρ. For this, we study the function

$$f(\rho) = \mathcal{L}(\rho \boldsymbol{w}^*).$$

To begin, by noticing $T^* = \boldsymbol{X}^\top \boldsymbol{w}^* = (\widetilde{Y}\boldsymbol{\mu} + \boldsymbol{\varepsilon})^\top \boldsymbol{w}^* = \widetilde{Y} \|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} + \boldsymbol{\varepsilon}^\top \boldsymbol{w}^*$ and $\boldsymbol{\varepsilon}^\top \boldsymbol{w}^* \perp \widetilde{Y}$, we have

$$\mathbb{E}[T^* \widetilde{Y}] = \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} = \|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} . \tag{13}$$

Thus, we can write

$$\begin{aligned}
\mathbb{E}[T^* \zeta(T^*)] &\overset{\text{(i)}}{=} \tfrac{1}{2}\mathbb{E}[T^*(2\zeta(T^*) - 1)] \\
&\overset{\text{(ii)}}{=} \tfrac{1}{2}\mathbb{E}[T^* \mathbb{E}[Y \mid \boldsymbol{X}]] = \tfrac{1}{2}\mathbb{E}[T^* Y] = \tfrac{1-2p}{2}\mathbb{E}[T^* \widetilde{Y}] \\
&\overset{\text{(iii)}}{=} \tfrac{1-2p}{2} \|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} ,
\end{aligned} \tag{14}$$

where (i) uses $\mathbb{E}[T^*] = 0$, (ii) holds because $\mathbb{P}(Y = 1 \mid \boldsymbol{X}) = \zeta(T^*)$ from Lemma 5 implies $\mathbb{E}[Y \mid \boldsymbol{X}] = 2\mathbb{P}(Y = 1 \mid \boldsymbol{X}) - 1 = 2\zeta(T^*) - 1$, and (iii) applies (13).

Let $\vartheta_t(s), \zeta(t)$ be defined as in (12), then note that the derivative $\vartheta_t'(s) = \sigma(s) - \zeta(t)$. Using this and applying (14), we obtain

$$\begin{aligned}
f'(\rho) &= \frac{\mathrm{d}}{\mathrm{d}\rho}\mathbb{E}[\vartheta_{T^*}(\rho T^*)] = \mathbb{E}[T^* \vartheta'_{T^*}(\rho T^*)] = \mathbb{E}[T^*(\sigma(\rho T^*) - \zeta(T^*))] \\
&= \mathbb{E}[T^* \sigma(\rho T^*)] - \mathbb{E}[T^* \zeta(T^*)] \\
&= \mathbb{E}[T^* \sigma(\rho T^*)] - \tfrac{1-2p}{2} \|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} .
\end{aligned} \tag{15}$$

Setting $f'(\rho^*) = 0$ and rearranging, any stationary point ρ^* of $f(\cdot)$ satisfies

$$\mathbb{E}[T^* \sigma(\rho^* T^*)] = \tfrac{1-2p}{2} \|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} .$$

Step 3: The final step is to show that $f(\cdot)$ has a *unique* stationary point $\rho^* \in (0, 2]$ and that it is a minimiser. Since $\sigma'(z) > 0$ for all $z \in \mathbb{R}$ and $T^* \neq 0$ almost surely, we have

$$f''(\rho) = \mathbb{E}[(T^*)^2 \sigma'(\rho T^*)] > 0 ,$$

meaning f is strictly convex. Hence, ρ^* is the unique minimiser of $f(\cdot)$. Finally, note that $\rho^* \in (0, 2]$ because $f'(0) < 0$ and $f'(2) \geq 0$. Indeed, Eq. (15) implies that

$$f'(0) = \tfrac{1}{2}\mathbb{E}[T^*] - \tfrac{1-2p}{2} \|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} = -\tfrac{1-2p}{2} \|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} < 0 ,$$

and

$$\begin{aligned}
f'(2) &= \mathbb{E}[T^* \sigma(2T^*)] - \tfrac{1-2p}{2} \|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} \\
&\overset{\text{(i)}}{=} \tfrac{1}{2}\mathbb{E}\left[T^*(2\sigma(2T^*) - 1)\right] - \tfrac{1-2p}{2} \|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} \\
&\overset{\text{(ii)}}{=} \tfrac{1}{2}\mathbb{E}[T^* \widetilde{Y}] - \tfrac{1-2p}{2} \|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} \\
&\overset{\text{(iii)}}{=} \tfrac{1}{2} \|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} - \tfrac{1-2p}{2} \|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} = p \|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} \geq 0 ,
\end{aligned}$$

where (i) uses $\mathbb{E}[T^*] = 0$, (ii) holds because $\mathbb{P}(\widetilde{Y} = 1 \mid \boldsymbol{X}) = \sigma(2T^*)$ from Lemma 5 implies $\mathbb{E}[\widetilde{Y} \mid \boldsymbol{X}] = 2\mathbb{P}(\widetilde{Y} = 1 \mid \boldsymbol{X}) - 1 = 2\sigma(2T^*) - 1$, and (iii) applies (13). □

B.2 Optimisation Lemmas

As detailed in Section 2, we train a linear classifier on the Gaussian mixture samples by minimising the empirical logistic risk $\widehat{\mathcal{L}}(\cdot)$ via early-stopped GD. To analyse the optimisation, we first establish the Lipschitz and smoothness properties of $\widehat{\mathcal{L}}(\cdot)$, then characterise the stopping time τ (Lemma 2) and the early-stopped GD iterate $\boldsymbol{w}_\tau$ (Lemma 3). Throughout this section, we will denote

$$\beta := \frac{\|\mathbf{X}_{1:n}\|^2}{n}.$$

Lemma 7 (Lipschitz and smoothness)**.** *The empirical logistic risk $\widehat{\mathcal{L}}(\cdot)$ is $\sqrt{\beta}$-Lipschitz and β-smooth, i.e.,*

$$\sup_{\boldsymbol{w}\in\mathbb{R}^d}\left\|\nabla\widehat{\mathcal{L}}(\boldsymbol{w})\right\| \le \sqrt{\beta} \qquad \textit{and} \qquad \sup_{\boldsymbol{w}\in\mathbb{R}^d}\left\|\nabla^2\widehat{\mathcal{L}}(\boldsymbol{w})\right\| \le \beta.$$

Proof. Recall $\|\mathbf{X}_{1:n}\|$ denotes the spectral norm of $\mathbf{X}_{1:n}$. Let $\boldsymbol{a}(\boldsymbol{w}) \in \mathbb{R}^n$ be the vector whose i-th coordinate is defined as $[\boldsymbol{a}(\boldsymbol{w})]_i = Y_i/(1+\exp(Y_i\boldsymbol{X}_i^\top\boldsymbol{w}))$. Then we have $\nabla\widehat{\mathcal{L}}(\boldsymbol{w}) = -\frac{1}{n}\mathbf{X}_{1:n}^\top\boldsymbol{a}(\boldsymbol{w}) \in \mathbb{R}^d$. Since the entry-wise absolute value $|[\boldsymbol{a}(\boldsymbol{w})]_i| \le 1$ for all $i \in [n]$, we have $\|\boldsymbol{a}(\boldsymbol{w})\| \le \sqrt{n}$ and so

$$\left\|\nabla\widehat{\mathcal{L}}(\boldsymbol{w})\right\| \le \frac{\|\mathbf{X}_{1:n}\|\,\|\boldsymbol{a}(\boldsymbol{w})\|}{n} \le \frac{\|\mathbf{X}_{1:n}\|}{\sqrt{n}} = \sqrt{\beta}.$$

For the Hessian, note that $\nabla^2\widehat{\mathcal{L}}(\boldsymbol{w}) = \frac{1}{n}\mathbf{X}_{1:n}^\top\boldsymbol{D}(\boldsymbol{w})\mathbf{X}_{1:n} \in \mathbb{R}^{d\times d}$, where $\boldsymbol{D}(\boldsymbol{w})$ is the $n\times n$ diagonal matrix whose i-th diagonal entry is $[\boldsymbol{D}(\boldsymbol{w})]_{ii} = \ell''(Y_i\boldsymbol{X}_i^\top\boldsymbol{w})$. Since $0 \le \ell''(z) \le 1/4$ for all $z \in \mathbb{R}$, we have

$$\left\|\nabla^2\widehat{\mathcal{L}}(\boldsymbol{w})\right\| \le \frac{\|\mathbf{X}_{1:n}\|\,\|\boldsymbol{D}(\boldsymbol{w})\|\,\|\mathbf{X}_{1:n}\|}{n} \le \frac{\|\mathbf{X}_{1:n}\|^2}{4n} \le \beta.$$

□

We next restate Lemmas 2 and 3 and provide their proofs, respectively.

Lemma 2 (Existence of τ)**.** *Suppose $n \le d$. Then $\|\boldsymbol{w}_t\| \to \infty$ as $t \to \infty$ almost surely, and so $\tau = \inf\{t \ge 1 : \|\boldsymbol{w}_t\| \ge 4\|\boldsymbol{w}^*\|\}$ is finite.*

Proof. When $n \le d$, we have $\operatorname{rank}(\mathbf{X}_{1:n}) = n$ almost surely, since the covariates $\boldsymbol{X} = \widetilde{Y}\boldsymbol{\mu} + \mathcal{N}(0, \boldsymbol{\Sigma})$ have a continuous density on $\mathbb{R}^d$ and are therefore linearly independent almost surely. Then, linear separability of the training samples follows from [40, Proposition 2.2]. On linearly-separable data, the implicit bias of GD under logistic loss [32, 31] implies $\|\boldsymbol{w}_t\| \to \infty$ as $t \to \infty$, so τ is finite. □

Lemma 3 (Properties of the early-stopped iterate)**.** *The early-stopped iterate $\boldsymbol{w}_\tau$ satisfies the following: (i) $\|\boldsymbol{w}_\tau\| < 4\|\boldsymbol{w}^*\| + 1$; and (ii) $\widehat{\mathcal{L}}(\boldsymbol{w}_\tau) \le \widehat{\mathcal{L}}(\boldsymbol{w}^*_{\mathcal{L}})$.*

Proof. **Proof of part (i):** By the definition of the stopping time $\tau := \inf\{t \ge 1 : \|\boldsymbol{w}_t\| \ge 4\|\boldsymbol{w}^*\|\}$, we have $\|\boldsymbol{w}_{\tau-1}\| < 4\|\boldsymbol{w}^*\|$. Using the GD update rule, Lemma 7, and the choice of step size $\gamma \le \min(1/\beta, 1)$ in Section 2, we can write

$$\|\boldsymbol{w}_\tau - \boldsymbol{w}_{\tau-1}\| = \gamma\left\|\nabla\widehat{\mathcal{L}}(\boldsymbol{w}_{\tau-1})\right\| \le \gamma\sqrt{\beta} \le \sqrt{\gamma}\cdot\sqrt{\gamma\beta} \le \sqrt{\gamma} \le 1\,.$$

Hence, by the triangle inequality,

$$\|\boldsymbol{w}_\tau\| \le \|\boldsymbol{w}_{\tau-1}\| + \|\boldsymbol{w}_\tau - \boldsymbol{w}_{\tau-1}\| < 4\|\boldsymbol{w}^*\| + 1.$$

Proof of part (ii): Suppose, towards a contradiction, that $\widehat{\mathcal{L}}(\boldsymbol{w}_\tau) > \widehat{\mathcal{L}}(\boldsymbol{w}^*_{\mathcal{L}})$. Since $\widehat{\mathcal{L}}(\cdot)$ is β-smooth and the step size $\gamma \le 1/\beta$, Lemma 3.3 in [40] with the comparator $\boldsymbol{w}^*_{\mathcal{L}}$ gives

$$\frac{\|\boldsymbol{w}_\tau - \boldsymbol{w}^*_{\mathcal{L}}\|^2}{2\gamma\tau} + \widehat{\mathcal{L}}(\boldsymbol{w}_\tau) \le \widehat{\mathcal{L}}(\boldsymbol{w}^*_{\mathcal{L}}) + \frac{\|\boldsymbol{w}^*_{\mathcal{L}}\|^2}{2\gamma\tau}.$$

Our assumption $\widehat{\mathcal{L}}(\boldsymbol{w}_\tau) > \widehat{\mathcal{L}}(\boldsymbol{w}^*_{\mathcal{L}})$ then forces $\|\boldsymbol{w}_\tau - \boldsymbol{w}^*_{\mathcal{L}}\| < \|\boldsymbol{w}^*_{\mathcal{L}}\|$. But by the definition of the stopping time τ, the triangle inequality, and $\|\boldsymbol{w}^*_{\mathcal{L}}\| = \|\rho^*\boldsymbol{w}^*\| \le 2\|\boldsymbol{w}^*\|$ due to Lemma 1, we arrive at a contradiction:

$$4\|\boldsymbol{w}^*\| \le \|\boldsymbol{w}_\tau\| \le \|\boldsymbol{w}_\tau - \boldsymbol{w}^*_{\mathcal{L}}\| + \|\boldsymbol{w}^*_{\mathcal{L}}\| < 2\|\boldsymbol{w}^*_{\mathcal{L}}\| \le 4\|\boldsymbol{w}^*\|\,.$$

□

C Proof of the Early-Stopped GD Upper Bound

We prove the excess zero-one risk upper bound of Theorem 1 for the early-stopped iterate $\boldsymbol{w}_\tau$ via a sequence of lemmas (Sections C.1–C.3.3), culminating in the proof of Theorem 1 at the end of this section (Section C.4).

Throughout, for symmetric matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}$, we write $\boldsymbol{A} \succ \boldsymbol{B}$ (resp. $\succeq$) to mean that $\boldsymbol{A}-\boldsymbol{B}$ is positive definite (resp. positive semidefinite), i.e., $\boldsymbol{h}^\top(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{h} > 0$ (resp. ≥ 0) for all $\boldsymbol{h} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$.

C.1 Zero-One to Logistic Calibration via $\boldsymbol{\Sigma}$-geometry

In this section, we prove the calibration result in Lemma 4, which upper bounds the excess zero-one risk by the excess logistic risk. The key technique is to introduce the $\boldsymbol{\Sigma}$-geometry distance $\|\boldsymbol{w}-\boldsymbol{w}^*_{\mathcal{L}}\|_{\boldsymbol{\Sigma}}$ as a bridge between the two risks: Lemma 8 upper bounds the excess zero-one risk by $\|\boldsymbol{w}-\boldsymbol{w}^*_{\mathcal{L}}\|_{\boldsymbol{\Sigma}}$, while Lemma 9 allows us to lower bound the excess logistic risk by $\|\boldsymbol{w}-\boldsymbol{w}^*_{\mathcal{L}}\|_{\boldsymbol{\Sigma}}$. Combining the two yields Lemma 4.

To begin, we show in Lemma 8 that if $\boldsymbol{w} \in \mathbb{R}^d$ is sufficiently close to $\boldsymbol{w}^*_{\mathcal{L}}$ in the $\boldsymbol{\Sigma}$-norm, then its excess zero-one risk $\mathcal{E}(\boldsymbol{w}) - \mathcal{E}(\boldsymbol{w}^*)$ can be upper bounded by $\|\boldsymbol{w}-\boldsymbol{w}^*_{\mathcal{L}}\|_{\boldsymbol{\Sigma}}$.

Lemma 8 ($\boldsymbol{\Sigma}$-geometry calibration of excess zero-one risk)**.** *Suppose $\boldsymbol{w} \in \mathbb{R}^d$ satisfies*

$$\|\boldsymbol{w}-\boldsymbol{w}^*_{\mathcal{L}}\|_{\boldsymbol{\Sigma}} \leq \frac{\rho^* \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}}}{2}. \tag{16}$$

Then,

$$\mathcal{E}(\boldsymbol{w}) - \mathcal{E}(\boldsymbol{w}^*) \leq \frac{4(1-2p)}{\sqrt{2\pi}(\rho^*)^2 \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}}} \|\boldsymbol{w}-\boldsymbol{w}^*_{\mathcal{L}}\|^2_{\boldsymbol{\Sigma}}\,. \tag{17}$$

Proof. Recall $\boldsymbol{w}^* = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. As in Lemma 1, decompose $\boldsymbol{w} = \rho\boldsymbol{w}^* + \boldsymbol{u}$ with $\boldsymbol{u}^\top\boldsymbol{\mu} = 0$. We first show that $\rho > 0$. Since $\boldsymbol{w}^*_{\mathcal{L}} = \rho^*\boldsymbol{w}^*$, we have

$$\|\boldsymbol{w}-\boldsymbol{w}^*_{\mathcal{L}}\|^2_{\boldsymbol{\Sigma}} = \|(\rho-\rho^*)\boldsymbol{w}^* + \boldsymbol{u}\|^2_{\boldsymbol{\Sigma}} = (\rho-\rho^*)^2 \|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} + \|\boldsymbol{u}\|^2_{\boldsymbol{\Sigma}}\,, \tag{18}$$

which implies $\|\boldsymbol{w}-\boldsymbol{w}^*_{\mathcal{L}}\|_{\boldsymbol{\Sigma}} \geq |\rho-\rho^*| \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}}$. Combining with the assumption in (16) gives $|\rho-\rho^*| \leq \rho^*/2$. Since $\rho^* > 0$ (c.f. Lemma 1), it follows that

$$\rho \geq \frac{\rho^*}{2} > 0\,. \tag{19}$$

The decomposition $\boldsymbol{w} = \rho\boldsymbol{w}^* + \boldsymbol{u}$ implies $\boldsymbol{X}^\top\boldsymbol{w} = \rho T^* + T_{\boldsymbol{u}}$, where $T^* = \boldsymbol{X}^\top\boldsymbol{w}^*$ and $T_{\boldsymbol{u}} = \boldsymbol{X}^\top\boldsymbol{u}$ are independent and each can be expanded as in (10)–(11). Applying the excess risk expression (8) of Lemma 5 and applying $\tanh(z) \leq |z|$ yields

$$\begin{aligned}
\mathcal{E}(\boldsymbol{w}) - \mathcal{E}(\boldsymbol{w}^*) &= (1-2p)\mathbb{E}\left[|\tanh(\boldsymbol{X}^\top\boldsymbol{w}^*)| \mathbb{1}\{\operatorname{sign}(\boldsymbol{X}^\top\boldsymbol{w}) \neq \operatorname{sign}(\boldsymbol{X}^\top\boldsymbol{w}^*)\}\right] \\
&\leq (1-2p)\mathbb{E}\left[|\boldsymbol{X}^\top\boldsymbol{w}^*| \mathbb{1}\{\operatorname{sign}(\boldsymbol{X}^\top\boldsymbol{w}) \neq \operatorname{sign}(\boldsymbol{X}^\top\boldsymbol{w}^*)\}\right] \\
&= (1-2p)\mathbb{E}\left[|T^*| \mathbb{1}\{\operatorname{sign}(\rho T^* + T_{\boldsymbol{u}}) \neq \operatorname{sign}(T^*)\}\right].
\end{aligned}$$

Since $\rho > 0$, we have the following event inclusion:

$$\{\operatorname{sign}(\rho T^* + T_{\boldsymbol{u}}) \neq \operatorname{sign}(T^*)\} \subseteq \left\{|T^*| \leq \frac{|T_{\boldsymbol{u}}|}{\rho}\right\}.$$

Note that conditionally on $\widetilde{Y}$, $T^* \mid \widetilde{Y} \sim \mathcal{N}(\widetilde{Y}\|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}, \|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}})$, so the conditional density satisfies $f_{T^*|\widetilde{Y}}(t) \leq 1/(\sqrt{2\pi}\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}})$. Since this bound is uniform over $\widetilde{Y} \in \{\pm 1\}$, the marginal density of

T^* satisfies the same bound $f_{T^*}(t) \le 1/(\sqrt{2\pi}\,\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}})$. Therefore, we have

$$\begin{aligned}
\mathcal{E}(\boldsymbol{w}) - \mathcal{E}(\boldsymbol{w}^*) &\le (1-2p)\mathbb{E}\left[|T^*|\mathbb{1}\{\operatorname{sign}(\rho T^* + T_{\boldsymbol{u}}) \neq \operatorname{sign}(T^*)\}\right] \\
&\le (1-2p)\mathbb{E}\left[|T^*|\mathbb{1}\left\{|T^*| \le \frac{|T_{\boldsymbol{u}}|}{\rho}\right\}\right] = (1-2p)\mathbb{E}\left[\mathbb{E}\left[|T^*|\mathbb{1}\left\{|T^*| \le \frac{|T_{\boldsymbol{u}}|}{\rho}\right\}\,\middle|\,T_{\boldsymbol{u}}\right]\right] \\
&= (1-2p)\mathbb{E}\left[\int_{-|T_{\boldsymbol{u}}|/\rho}^{|T_{\boldsymbol{u}}|/\rho} |t| f_{T^*}(t)\mathrm{d}t\right] \le \frac{1-2p}{\sqrt{2\pi}\,\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}}}\mathbb{E}\left[\int_{-|T_{\boldsymbol{u}}|/\rho}^{|T_{\boldsymbol{u}}|/\rho} |t|\mathrm{d}t\right] \\
&= \frac{1-2p}{\sqrt{2\pi}\,\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}}}\mathbb{E}\left[\frac{T_{\boldsymbol{u}}^2}{\rho^2}\right] = \frac{(1-2p)}{\sqrt{2\pi}\,\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}}} \cdot \frac{\|\boldsymbol{u}\|_{\boldsymbol{\Sigma}}^2}{\rho^2}\,,
\end{aligned}$$

where the last step holds because $T_{\boldsymbol{u}} \sim \mathcal{N}(0, \|\boldsymbol{u}\|_{\boldsymbol{\Sigma}}^2)$. Combining this with the bound $\|\boldsymbol{u}\|_{\boldsymbol{\Sigma}}^2 \le \|\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*\|_{\boldsymbol{\Sigma}}^2$ implied by (18), and with (19), we arrive at the result in (17).

□

The next step is to link the excess logistic risk to $\|\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*\|_{\boldsymbol{\Sigma}}^2$. Since $\boldsymbol{w}_{\mathcal{L}}^*$ minimises $\mathcal{L}(\cdot)$, we do so by establishing local strong convexity of $\mathcal{L}(\cdot)$ around $\boldsymbol{w}_{\mathcal{L}}^*$. By Lemma 3, we know the early-stopped iterate $\boldsymbol{w}_\tau$ satisfies $\|\boldsymbol{w}_\tau\| \le 4\|\boldsymbol{w}^*\| + 1$. We therefore work over the ball

$$\mathcal{B} = \{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\| \le R\}\,, \quad \text{where} \quad R := 4\,\|\boldsymbol{w}^*\| + 1\,. \tag{20}$$

Clearly $\boldsymbol{w}^* \in \mathcal{B}$ and, by Lemma 1, $\boldsymbol{w}_{\mathcal{L}}^* \in \mathcal{B}$ as well. As we show in Lemma 9 below, the local strong-convexity parameter of $\mathcal{L}(\cdot)$ is given by

$$\alpha = \tfrac{1}{2}\ell''(3R\sqrt{\|\boldsymbol{\Gamma}\|}), \qquad \text{where} \quad \boldsymbol{\Gamma} = \operatorname{Cov}(\boldsymbol{X}) = \boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma}. \tag{21}$$

Lemma 9 (Local strong convexity of logistic risk)**.** *For every $\boldsymbol{w} \in \mathcal{B}$ it holds that $\nabla^2\mathcal{L}(\boldsymbol{w}) \succeq \alpha\boldsymbol{\Gamma}$. Consequently,*

$$\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}_{\mathcal{L}}^*) \ge \frac{\alpha}{2}\,\|\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*\|_{\boldsymbol{\Gamma}}^2 \quad \textit{for all } \boldsymbol{w} \in \mathcal{B}.$$

Proof. **Step 1:** We first show that for every $\boldsymbol{w} \in \mathcal{B}$, $\nabla^2\mathcal{L}(\boldsymbol{w}) \succeq \alpha\boldsymbol{\Gamma}$. Note that $\ell''(z) = e^z/(1+e^z)^2$ is even, so

$$\nabla^2\mathcal{L}(\boldsymbol{w}) = \mathbb{E}[\ell''(Y\boldsymbol{X}^\top\boldsymbol{w})\boldsymbol{X}\boldsymbol{X}^\top] = \mathbb{E}[\ell''(\boldsymbol{X}^\top\boldsymbol{w})\boldsymbol{X}\boldsymbol{X}^\top]\,.$$

Hence, for any $\boldsymbol{h} \in \mathbb{R}^d$,

$$\boldsymbol{h}^\top\nabla^2\mathcal{L}(\boldsymbol{w})\boldsymbol{h} = \mathbb{E}[\ell''(\boldsymbol{X}^\top\boldsymbol{w})(\boldsymbol{X}^\top\boldsymbol{h})^2].$$

To show $\nabla^2\mathcal{L}(\boldsymbol{w}) \succeq \alpha\boldsymbol{\Gamma}$, it suffices to show that

$$\mathbb{E}[\ell''(\boldsymbol{X}^\top\boldsymbol{w})(\boldsymbol{X}^\top\boldsymbol{h})^2] \ge \alpha\boldsymbol{h}^\top\boldsymbol{\Gamma}\boldsymbol{h}\,.$$

Towards this objective, using the fact that ℓ'' is even and decreasing on $[0,\infty)$, we have

$$\ell''(\boldsymbol{X}^\top\boldsymbol{w}) \ge \ell''(3R\sqrt{\|\boldsymbol{\Gamma}\|})\mathbb{1}\{|\boldsymbol{X}^\top\boldsymbol{w}| \le 3R\sqrt{\|\boldsymbol{\Gamma}\|}\},$$

so it follows that

$$\begin{aligned}
\mathbb{E}[\ell''(\boldsymbol{X}^\top\boldsymbol{w})(\boldsymbol{X}^\top\boldsymbol{h})^2] &\ge \ell''(3R\sqrt{\|\boldsymbol{\Gamma}\|}) \cdot \mathbb{E}[(\boldsymbol{X}^\top\boldsymbol{h})^2\mathbb{1}\{|\boldsymbol{X}^\top\boldsymbol{w}| \le 3R\sqrt{\|\boldsymbol{\Gamma}\|}\}] \\
&= 2\alpha \cdot \mathbb{E}[(\boldsymbol{X}^\top\boldsymbol{h})^2\mathbb{1}\{|\boldsymbol{X}^\top\boldsymbol{w}| \le 3R\sqrt{\|\boldsymbol{\Gamma}\|}\}].
\end{aligned}$$

It remains to show that $\mathbb{E}[(\boldsymbol{X}^\top\boldsymbol{h})^2\mathbb{1}\{|\boldsymbol{X}^\top\boldsymbol{w}| \le 3R\sqrt{\|\boldsymbol{\Gamma}\|}\}] \ge \frac{1}{2}\boldsymbol{h}^\top\boldsymbol{\Gamma}\boldsymbol{h}$, or, equivalently,

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{X}^\top\boldsymbol{h})^2\mathbb{1}\{|\boldsymbol{X}^\top\boldsymbol{w}| > 3R\sqrt{\|\boldsymbol{\Gamma}\|}\}] &= \mathbb{E}[(\boldsymbol{X}^\top\boldsymbol{h})^2] - \mathbb{E}[(\boldsymbol{X}^\top\boldsymbol{h})^2\mathbb{1}\{|\boldsymbol{X}^\top\boldsymbol{w}| \le 3R\sqrt{\|\boldsymbol{\Gamma}\|}\}] \\
&= \boldsymbol{h}^\top\boldsymbol{\Gamma}\boldsymbol{h} - \mathbb{E}[(\boldsymbol{X}^\top\boldsymbol{h})^2\mathbb{1}\{|\boldsymbol{X}^\top\boldsymbol{w}| \le 3R\sqrt{\|\boldsymbol{\Gamma}\|}\}] \\
&\le \tfrac{1}{2}\boldsymbol{h}^\top\boldsymbol{\Gamma}\boldsymbol{h}.
\end{aligned} \tag{22}$$

We do this as follows: starting with Cauchy-Schwarz,

$$\mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{h})^2 \mathbb{1}\{|\boldsymbol{X}^\top \boldsymbol{w}| > 3R\sqrt{\|\boldsymbol{\Gamma}\|}\}] \le \sqrt{\mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{h})^4]}\sqrt{\mathbb{P}(|\boldsymbol{X}^\top \boldsymbol{w}| > 3R\sqrt{\|\boldsymbol{\Gamma}\|})}. \tag{23}$$

We bound each of the terms under the square roots. For the expectation, note that, since $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, $\widetilde{Y}^2 = 1$ and $\boldsymbol{\varepsilon} \perp \widetilde{Y}$, we have

$$\begin{aligned}\mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{h})^4] &= \mathbb{E}[(\widetilde{Y}\boldsymbol{\mu}^\top \boldsymbol{h} + \boldsymbol{\varepsilon}^\top \boldsymbol{h})^4] = (\boldsymbol{\mu}^\top \boldsymbol{h})^4 + 6(\boldsymbol{\mu}^\top \boldsymbol{h})^2(\boldsymbol{h}^\top \boldsymbol{\Sigma} \boldsymbol{h}) + 3(\boldsymbol{h}^\top \boldsymbol{\Sigma} \boldsymbol{h})^2 \\ &\le 3((\boldsymbol{\mu}^\top \boldsymbol{h})^2 + (\boldsymbol{h}^\top \boldsymbol{\Sigma} \boldsymbol{h}))^2 = 3(\boldsymbol{h}^\top \boldsymbol{\Gamma} \boldsymbol{h})^2. \end{aligned} \tag{24}$$

For the probability term, using Lemma 6 and the fact that $\|\boldsymbol{w}\| \le R$,

$$\mathbb{P}(|\boldsymbol{X}^\top \boldsymbol{w}| > 3R\sqrt{\|\boldsymbol{\Gamma}\|}) \le 2\exp\left(-\frac{9R^2\|\boldsymbol{\Gamma}\|}{2\boldsymbol{w}^\top \boldsymbol{\Gamma} \boldsymbol{w}}\right) \le 2\exp\left(-\tfrac{9}{2}\right). \tag{25}$$

Plugging (24) and (25) into (23), we arrive at

$$\mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{h})^2 \mathbb{1}\{|\boldsymbol{X}^\top \boldsymbol{w}| > 3R\sqrt{\|\boldsymbol{\Gamma}\|}\}] \le (\sqrt{3}\boldsymbol{h}^\top \boldsymbol{\Gamma} \boldsymbol{h}) \cdot \left(\sqrt{2}\exp\left(-\tfrac{9}{4}\right)\right) \le \tfrac{1}{2}\boldsymbol{h}^\top \boldsymbol{\Gamma} \boldsymbol{h},$$

which establishes (22) and hence for every $\boldsymbol{w} \in \mathcal{B}$, $\nabla^2 \mathcal{L}(\boldsymbol{w}) \succeq \alpha \boldsymbol{\Gamma}$.

Step 2: For the excess logistic risk lower bound, using the second-order Taylor expansion of $\boldsymbol{w} \in \mathcal{B}$ around $\boldsymbol{w}_{\mathcal{L}}^* \in \mathcal{B}$, there is some $\boldsymbol{\xi} \in \mathcal{B}$ on the line connecting $\boldsymbol{w}$ and $\boldsymbol{w}_{\mathcal{L}}^*$ such that

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}_{\mathcal{L}}^*) &\ge \nabla \mathcal{L}(\boldsymbol{w}_{\mathcal{L}}^*)^\top (\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*) + \tfrac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*)^\top \nabla^2 \mathcal{L}(\boldsymbol{\xi})(\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*) \\ &\ge \tfrac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*)^\top \nabla^2 \mathcal{L}(\boldsymbol{\xi})(\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*) \ge \frac{\alpha}{2}\|\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*\|_{\boldsymbol{\Gamma}}^2. \end{aligned}$$

where the first term vanishes by the first-order optimality condition $\nabla \mathcal{L}(\boldsymbol{w}_{\mathcal{L}}^*) = \boldsymbol{0}$ (c.f. Lemma 1). □

We now restate Lemma 4 and prove it by combining Lemmas 8–9 above.

Lemma 4 (Linear-rate calibration for linear classifiers)**.** *Let $\boldsymbol{w} \in \mathbb{R}^d$ satisfy $\|\boldsymbol{w}\| \le 4\|\boldsymbol{w}^*\| + 1$ and*

$$\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}_{\mathcal{L}}^*) \le \frac{\alpha(\rho^*)^2\|\boldsymbol{w}^*\|_{\boldsymbol{\Sigma}}^2}{8}, \tag{5}$$

where $\alpha = \frac{1}{2}\ell''\big(3(4\|\boldsymbol{w}^*\| + 1)\sqrt{\|\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top\|}\big)$. Then,

$$\mathcal{E}(\boldsymbol{w}) - \mathcal{E}(\boldsymbol{w}^*) \le \frac{8(1-2p)}{\sqrt{2\pi}\alpha(\rho^*)^2\|\boldsymbol{w}^*\|_{\boldsymbol{\Sigma}}}\left(\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}_{\mathcal{L}}^*)\right).$$

Proof. Recall $\boldsymbol{\Gamma} = \boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma}$ so $\boldsymbol{\Gamma} \succeq \boldsymbol{\Sigma}$. Together with Lemma 9, we have

$$\|\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*\|_{\boldsymbol{\Sigma}}^2 \le \|\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*\|_{\boldsymbol{\Gamma}}^2 \le \frac{2}{\alpha}(\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}_{\mathcal{L}}^*)).$$

Using the assumption in (5), this implies

$$\|\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*\|_{\boldsymbol{\Sigma}} \le \frac{\rho^*\|\boldsymbol{w}^*\|_{\boldsymbol{\Sigma}}}{2} = \frac{\rho^*\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}}}{2}.$$

We can now apply Lemma 8 to conclude:

$$\mathcal{E}(\boldsymbol{w}) - \mathcal{E}(\boldsymbol{w}^*) \le \frac{4(1-2p)}{\sqrt{2\pi}(\rho^*)^2\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}}}\|\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*\|_{\boldsymbol{\Sigma}}^2 \le \frac{4(1-2p)}{\sqrt{2\pi}(\rho^*)^2\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}}}\frac{2}{\alpha}\left(\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}_{\mathcal{L}}^*)\right).$$

□

By Lemma 3, the early-stopped iterate $\boldsymbol{w}_\tau$ lies in the ball $\mathcal{B}$, so Lemma 4 reduces the problem of bounding the population excess zero-one risk to bounding the population excess logistic risk $\mathcal{L}(\boldsymbol{w}_\tau) - \mathcal{L}(\boldsymbol{w}_{\mathcal{L}}^*)$, which admits the decomposition below in (26). By Lemma 3 (ii), Term I of (26) is non-positive. Hence, it remains to control the uniform deviation between the empirical and population excess logistic risks, i.e., Term II of (26), and we do so using empirical process theory.

$$\mathcal{L}(\boldsymbol{w}_\tau) - \mathcal{L}(\boldsymbol{w}_\mathcal{L}^*) = \underbrace{\left(\widehat{\mathcal{L}}(\boldsymbol{w}_\tau) - \widehat{\mathcal{L}}(\boldsymbol{w}_\mathcal{L}^*)\right)}_{\text{Term I}\leq 0 \text{ by Lemma 3 (ii)}} + \underbrace{(\mathcal{L}(\boldsymbol{w}_\tau) - \mathcal{L}(\boldsymbol{w}_\mathcal{L}^*)) - \left(\widehat{\mathcal{L}}(\boldsymbol{w}_\tau) - \widehat{\mathcal{L}}(\boldsymbol{w}_\mathcal{L}^*)\right)}_{\text{Term II: controlled by empirical process theory}}. \tag{26}$$

To this end, a direct uniform convergence bound over the entire ball $\mathcal{B}$ would be too loose. To obtain a sharper upper bound in terms of the *effective rank* of $\boldsymbol{\Sigma}$, we exploit the fact that $\boldsymbol{w}_\tau$ has small excess risk, confining it to a thin shell around $\boldsymbol{w}_\mathcal{L}^*$ within $\mathcal{B}$. This localisation yields a smaller function class with better-controlled complexity, which we detail in Section C.3.

For clarity, we first recall in Section C.2 the definition of the effective rank and establish a key spectral property of the empirical covariance of the covariates $(\boldsymbol{X}_i)_{i=1}^n$ in terms of the effective rank. We then present the localisation and concentrations proofs in Section C.3.

C.2 Effective Rank and Empirical Covariance

We define the effective rank and establish two spectral properties in terms of the effective rank (Lemmas 10–11) which will be used in the sequel.

C.2.1 Effective Rank of Population Covariances

Recall that in our Gaussian mixture model, the population covariance of the noise variable $\boldsymbol{\varepsilon}$ is $\boldsymbol{\Sigma}$, while that of the covariates $\boldsymbol{X}$ is $\boldsymbol{\Gamma} = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top$. Both matrices induce geometries that play a role in our analysis, so we introduce the effective rank for each and use $\boldsymbol{\Sigma}$ or $\boldsymbol{\Gamma}$ as a superscript to avoid any confusion. For any $\eta > 0$, define the regularised matrices $\boldsymbol{\Sigma}_\eta := \boldsymbol{\Sigma} + \eta \boldsymbol{I}_d$ and $\boldsymbol{\Gamma}_\eta := \boldsymbol{\Gamma} + \eta \boldsymbol{I}_d$, and define

$$r_{\text{eff}}^{\boldsymbol{\Sigma}}(\eta) := \operatorname{tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\eta^{-1}) = \operatorname{tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \eta \boldsymbol{I}_d)^{-1}) = \sum_{i=1}^d \frac{\lambda_i(\boldsymbol{\Sigma})}{\lambda_i(\boldsymbol{\Sigma}) + \eta},$$

$$r_{\text{eff}}^{\boldsymbol{\Gamma}}(\eta) := \operatorname{tr}(\boldsymbol{\Gamma}\boldsymbol{\Gamma}_\eta^{-1}) = \operatorname{tr}(\boldsymbol{\Gamma}(\boldsymbol{\Gamma} + \eta \boldsymbol{I}_d)^{-1}) = \sum_{i=1}^d \frac{\lambda_i(\boldsymbol{\Gamma})}{\lambda_i(\boldsymbol{\Gamma}) + \eta}.$$

Note that throughout the main text we only work with the effective rank associated with $\boldsymbol{\Sigma}$, and so we write $r_{\text{eff}}(\eta)$ to refer to $r_{\text{eff}}^{\boldsymbol{\Sigma}}(\eta)$ for brevity. In the appendices, both $r_{\text{eff}}^{\boldsymbol{\Sigma}}(\eta)$ and $r_{\text{eff}}^{\boldsymbol{\Gamma}}(\eta)$ appear, so we retain the superscripts for clarity.

We show in Lemma 10 that the two effective ranks differ by at most 1.

Lemma 10 (Effective ranks of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$)**.** *For any $\eta > 0$,*

$$r_{\text{eff}}^{\boldsymbol{\Sigma}}(\eta) \leq r_{\text{eff}}^{\boldsymbol{\Gamma}}(\eta) \leq r_{\text{eff}}^{\boldsymbol{\Sigma}}(\eta) + 1.$$

Proof. It suffices to show $0 \leq r_{\text{eff}}^{\boldsymbol{\Gamma}}(\eta) - r_{\text{eff}}^{\boldsymbol{\Sigma}}(\eta) \leq 1$. Note the identities $\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\eta^{-1} = \boldsymbol{I}_d - \eta\boldsymbol{\Sigma}_\eta^{-1}$ and $\boldsymbol{\Gamma}\boldsymbol{\Gamma}_\eta^{-1} = \boldsymbol{I}_d - \eta\boldsymbol{\Gamma}_\eta^{-1}$. Then the Sherman-Morrison formula applied to $\boldsymbol{\Sigma}_\eta \succ 0$ gives

$$(\boldsymbol{\Sigma}_\eta + \boldsymbol{\mu}\boldsymbol{\mu}^\top)^{-1} = \boldsymbol{\Sigma}_\eta^{-1} - \frac{\boldsymbol{\Sigma}_\eta^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}_\eta^{-1}}{1 + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}_\eta^{-1}\boldsymbol{\mu}},$$

so we have

$$\begin{aligned}
r_{\text{eff}}^{\boldsymbol{\Gamma}}(\eta) - r_{\text{eff}}^{\boldsymbol{\Sigma}}(\eta) &= \operatorname{tr}(\boldsymbol{\Gamma}\boldsymbol{\Gamma}_\eta^{-1} - \boldsymbol{\Sigma}\boldsymbol{\Sigma}_\eta^{-1}) \\
&= \operatorname{tr}(\boldsymbol{I}_d - \eta\boldsymbol{\Gamma}_\eta^{-1} - \boldsymbol{I}_d + \eta\boldsymbol{\Sigma}_\eta^{-1}) \\
&= \eta \operatorname{tr}(\boldsymbol{\Sigma}_\eta^{-1} - (\boldsymbol{\Sigma}_\eta + \boldsymbol{\mu}\boldsymbol{\mu}^\top)^{-1}) \\
&= \frac{\eta \operatorname{tr}(\boldsymbol{\Sigma}_\eta^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}_\eta^{-1})}{1 + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}_\eta^{-1}\boldsymbol{\mu}} \\
&= \frac{\eta\boldsymbol{\mu}^\top\boldsymbol{\Sigma}_\eta^{-2}\boldsymbol{\mu}}{1 + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}_\eta^{-1}\boldsymbol{\mu}}.
\end{aligned}$$

Since $\boldsymbol{\Sigma}_\eta^{-1}, \boldsymbol{\Sigma}_\eta^{-2} \succ 0$, we clearly have $r_{\text{eff}}^{\boldsymbol{\Gamma}}(\eta) - r_{\text{eff}}^{\boldsymbol{\Sigma}}(\eta) \geq 0$. Next, to show $r_{\text{eff}}^{\boldsymbol{\Gamma}}(\eta) - r_{\text{eff}}^{\boldsymbol{\Sigma}}(\eta) \leq 1$, it suffices to show

$$\eta \boldsymbol{\mu}^\top \boldsymbol{\Sigma}_\eta^{-2} \boldsymbol{\mu} \leq \boldsymbol{\mu}^\top \boldsymbol{\Sigma}_\eta^{-1} \boldsymbol{\mu}.$$

This holds because $\boldsymbol{\Sigma}_\eta \succeq \eta \boldsymbol{I}_d$, so that $\boldsymbol{\Sigma}_\eta^{-2} \preceq \eta^{-1} \boldsymbol{\Sigma}_\eta^{-1}$, i.e., $\eta \boldsymbol{\Sigma}_\eta^{-2} \preceq \boldsymbol{\Sigma}_\eta^{-1}$. □

C.2.2 Spectral Control of the Empirical Covariance via Effective Rank

Now, define the empirical covariance of the covariates as

$$\widehat{\boldsymbol{\Gamma}} := \frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top,$$

and define the good event

$$E_{\text{cov}} := \{ ||\boldsymbol{\Gamma}_\eta^{-1/2} \widehat{\boldsymbol{\Gamma}} \boldsymbol{\Gamma}_\eta^{-1/2}|| \leq 2 \}, \tag{27}$$

on which the empirical covariance is well-behaved in that it does not introduce spurious high-variance directions, and is spectrally controlled relative to the regularised population covariance $\boldsymbol{\Gamma}_\eta$. We show in Lemma 11 that the event E_{cov} holds with high probability, whenever the sample size n is sufficiently large relative to $r_{\text{eff}}^{\boldsymbol{\Gamma}}(\eta)$.

Lemma 11 (Spectral control of the empirical covariance)**.** *There exists a universal constant $\mathsf{c} > 0$ such that the following holds: Fix $\eta > 0$. For every $\delta \in (0, 1)$, if*

$$n \geq \mathsf{c}(r_{\text{eff}}^{\boldsymbol{\Gamma}}(\eta) + \log(1/\delta)),$$

then we have $\mathbb{P}(E_{\text{cov}}) \geq 1 - \delta$.

Proof. We split this proof into two steps:

Step 1: Let $\boldsymbol{Z}_i = \boldsymbol{\Gamma}_\eta^{-1/2} \boldsymbol{X}_i$ for $1 \leq i \leq n$, so their empirical covariance takes the form

$$\frac{1}{n} \sum_{i=1}^n \boldsymbol{Z}_i \boldsymbol{Z}_i^\top = \boldsymbol{\Gamma}_\eta^{-1/2} \widehat{\boldsymbol{\Gamma}} \boldsymbol{\Gamma}_\eta^{-1/2},$$

and define its population covariance as

$$\boldsymbol{\Upsilon} = \mathbb{E}[\boldsymbol{Z}_i \boldsymbol{Z}_i^\top] = \boldsymbol{\Gamma}_\eta^{-1/2} \boldsymbol{\Gamma} \boldsymbol{\Gamma}_\eta^{-1/2} = \boldsymbol{\Gamma} \boldsymbol{\Gamma}_\eta^{-1}.$$

where note $\|\boldsymbol{\Upsilon}\| \leq 1$ and $\operatorname{tr}(\boldsymbol{\Upsilon}) = r_{\text{eff}}^{\boldsymbol{\Gamma}}(\eta)$.

In this step of the proof, we will apply [21, Theorem 9] to bound the spectral norm deviation of the sample covariance from the population covariance $\|\boldsymbol{\Gamma}_\eta^{-1/2} \widehat{\boldsymbol{\Gamma}} \boldsymbol{\Gamma}_\eta^{-1/2} - \boldsymbol{\Upsilon}\|$. To verify the conditions in the hypothesis of this theorem, note that the $\boldsymbol{Z}_i$'s are:

- weakly-square integrable, i.e., $\mathbb{E}[(\boldsymbol{Z}_i^\top \boldsymbol{u})^2] < \infty$ for all $\boldsymbol{u} \in \mathbb{R}^d$, which holds as $\|\boldsymbol{\Upsilon}\| \leq 1$;
- centred and have covariance $\boldsymbol{\Upsilon}$ and empirical covariance $\boldsymbol{\Gamma}_\eta^{-1/2} \widehat{\boldsymbol{\Gamma}} \boldsymbol{\Gamma}_\eta^{-1/2}$;
- subgaussian in the sense of [21, Definition 2], by Lemma 6 applied with $\boldsymbol{w} = \boldsymbol{\Gamma}_\eta^{-1/2} \boldsymbol{u}$;
- pregaussian, i.e., there exists a centred Gaussian random vector with the same covariance $\boldsymbol{\Upsilon}$. Indeed $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Upsilon})$ satisfies this condition.

Then, by [21, Theorem 9], there exists a universal constant $\mathsf{c}' > 0$ such that, for all $t \geq 1$, with probability at least $1 - e^{-t}$

$$\left\| \boldsymbol{\Gamma}_\eta^{-1/2} \widehat{\boldsymbol{\Gamma}} \boldsymbol{\Gamma}_\eta^{-1/2} - \boldsymbol{\Upsilon} \right\| \leq \mathsf{c}' \|\boldsymbol{\Upsilon}\| \max\left(\sqrt{\frac{\widetilde{r}(\boldsymbol{\Upsilon})}{n}}, \frac{\widetilde{r}(\boldsymbol{\Upsilon})}{n}, \sqrt{\frac{t}{n}}, \frac{t}{n} \right),$$

where, using the notation from [21, Definition 1],

$$\widetilde{r}(\boldsymbol{\Upsilon}) = \frac{\mathrm{tr}(\boldsymbol{\Upsilon})}{\|\boldsymbol{\Upsilon}\|} = \frac{r_{\mathrm{eff}}^{\boldsymbol{\Gamma}}(\eta)}{\|\boldsymbol{\Upsilon}\|}.$$

Since $\|\boldsymbol{\Upsilon}\| \le 1$, we have

$$\left\|\boldsymbol{\Gamma}_\eta^{-1/2}\widehat{\boldsymbol{\Gamma}}\boldsymbol{\Gamma}_\eta^{-1/2} - \boldsymbol{\Upsilon}\right\| \le \mathtt{c}' \max\left(\sqrt{\frac{r_{\mathrm{eff}}^{\boldsymbol{\Gamma}}(\eta)}{n}}, \frac{r_{\mathrm{eff}}^{\boldsymbol{\Gamma}}(\eta)}{n}, \sqrt{\frac{t}{n}}, \frac{t}{n}\right).$$

Step 2: This step establishes $\mathbb{P}(||\boldsymbol{\Gamma}_\eta^{-1/2}\widehat{\boldsymbol{\Gamma}}\boldsymbol{\Gamma}_\eta^{-1/2}|| \le 2) \ge 1-\delta$. Combining the last display with the triangle inequality and setting $t = \log(1/\delta)$, we obtain with probability at least $1-\delta$,

$$\begin{aligned}\left\|\boldsymbol{\Gamma}_\eta^{-1/2}\widehat{\boldsymbol{\Gamma}}\boldsymbol{\Gamma}_\eta^{-1/2}\right\| &\le \|\boldsymbol{\Upsilon}\| + \left\|\boldsymbol{\Gamma}_\eta^{-1/2}\widehat{\boldsymbol{\Gamma}}\boldsymbol{\Gamma}_\eta^{-1/2} - \boldsymbol{\Upsilon}\right\| \\ &\le 1 + \mathtt{c}' \max\left(\sqrt{\frac{r_{\mathrm{eff}}^{\boldsymbol{\Gamma}}(\eta)}{n}}, \frac{r_{\mathrm{eff}}^{\boldsymbol{\Gamma}}(\eta)}{n}, \sqrt{\frac{\log(1/\delta)}{n}}, \frac{\log(1/\delta)}{n}\right).\end{aligned}$$

Hence, if $n \ge \mathtt{c}(r_{\mathrm{eff}}^{\boldsymbol{\Gamma}}(\eta) + \log(1/\delta))$ for a sufficiently large universal constant $\mathtt{c}$, then

$$\mathtt{c}' \max\left(\sqrt{\frac{r_{\mathrm{eff}}^{\boldsymbol{\Gamma}}(\eta)}{n}}, \frac{r_{\mathrm{eff}}^{\boldsymbol{\Gamma}}(\eta)}{n}, \sqrt{\frac{\log(1/\delta)}{n}}, \frac{\log(1/\delta)}{n}\right) \le 1,$$

and this concludes the proof. □

C.3 Localisation and Concentrations

For brevity, we adopt standard notation from empirical process theory. For each $\boldsymbol{w} \in \mathbb{R}^d$, define $g_{\boldsymbol{w}} : \mathbb{R}^d \times \{\pm 1\} \to \mathbb{R}$ by

$$g_{\boldsymbol{w}}(\boldsymbol{x}, y) := \ell(y\boldsymbol{x}^\top \boldsymbol{w}) - \ell(y\boldsymbol{x}^\top \boldsymbol{w}_{\mathcal{L}}^*),$$

and let P and P_n denote the population and empirical expectation operators, so that

$$\begin{aligned}Pg_{\boldsymbol{w}} &:= \mathbb{E}[g_{\boldsymbol{w}}(\boldsymbol{X}, Y)] = \mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}_{\mathcal{L}}^*), \\ P_n g_{\boldsymbol{w}} &:= \frac{1}{n}\sum_{i=1}^n g_{\boldsymbol{w}}(\boldsymbol{X}_i, Y_i) = \widehat{\mathcal{L}}(\boldsymbol{w}) - \widehat{\mathcal{L}}(\boldsymbol{w}_{\mathcal{L}}^*).\end{aligned}$$

Then,

$$(P - P_n)g_{\boldsymbol{w}} = Pg_{\boldsymbol{w}} - P_n g_{\boldsymbol{w}} = (\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}_{\mathcal{L}}^*)) - \left(\widehat{\mathcal{L}}(\boldsymbol{w}) - \widehat{\mathcal{L}}(\boldsymbol{w}_{\mathcal{L}}^*)\right)$$

is the term that we need to bound in the risk decomposition in (26). For any $r > 0$, we work over the localised function class

$$\mathcal{G}(r) = \{g_{\boldsymbol{w}} : \boldsymbol{w} \in \mathcal{B} \text{ and } Pg_{\boldsymbol{w}} \le r\},$$

which contains $g_{\boldsymbol{w}_{\mathcal{L}}^*}$ for every $r > 0$ because $Pg_{\boldsymbol{w}_{\mathcal{L}}^*} = 0$. This class consists of functions indexed by parameters $\boldsymbol{w}$ lying in a localised shell within the ball $\mathcal{B}$.

Notation-wise, recall the ball $\mathcal{B} = \{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\| \le R\}$ with $R = 4\|\boldsymbol{w}^*\| + 1$ from (20), and the local strong-convexity parameter $\alpha = \frac{1}{2}\ell''(3R\sqrt{\|\boldsymbol{\Gamma}\|})$ of $\mathcal{L}(\cdot)$ over $\mathcal{B}$ from Eq. (21).

C.3.1 Useful Properties of $\mathcal{G}(r)$

We first present some useful properties of the function class $\mathcal{G}(r)$, which we will use in our proofs later in this section.

Lemma 12. *If $g_{\boldsymbol{w}} \in \mathcal{G}(r)$, then $\|\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*\| \le 2R$ and*

$$\|\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*\|_{\boldsymbol{\Gamma}}^2 \le \frac{2r}{\alpha}.$$

Proof. The first inequality is simply by the fact that $\boldsymbol{w}, \boldsymbol{w}^*_{\mathcal{L}} \in \mathcal{B}$ and the triangle inequality. The second inequality is obtained by rearranging the conclusion of Lemma 9 and using $Pg_{\boldsymbol{w}} \leq r$. □

Lemma 13. *For every $g_{\boldsymbol{w}} \in \mathcal{G}(r)$, $|g_{\boldsymbol{w}}(\boldsymbol{X}, Y)| \leq |\boldsymbol{X}^\top(\boldsymbol{w} - \boldsymbol{w}^*_{\mathcal{L}})|$. As a consequence,*

$$\mathbb{E}[g_{\boldsymbol{w}}(\boldsymbol{X}, Y)^2] \leq \frac{2r}{\alpha}.$$

Proof. The first inequality is a consequence of the fact that ℓ is 1-Lipschitz and $Y \in \{\pm 1\}$, whereas for the second inequality we can apply Lemma 9 to obtain

$$\mathbb{E}[g_{\boldsymbol{w}}(\boldsymbol{X}, Y)^2] \leq \mathbb{E}[(\boldsymbol{X}^\top(\boldsymbol{w} - \boldsymbol{w}^*_{\mathcal{L}}))^2] = \|\boldsymbol{w} - \boldsymbol{w}^*_{\mathcal{L}}\|^2_{\boldsymbol{\Gamma}} \leq \frac{2Pg_{\boldsymbol{w}}}{\alpha} \leq \frac{2r}{\alpha}.$$

□

C.3.2 Empirical Excess Logistic Risk Concentration on $\mathcal{G}(r)$

On the event E_{cov}, defined in (27) and characterised in Lemma 11, we provide the first uniform concentration bound: it controls how much the empirical excess logistic risk $P_n g_{\boldsymbol{w}}$ deviates from the population excess logistic risk $Pg_{\boldsymbol{w}}$ uniformly over all functions $g_{\boldsymbol{w}} \in \mathcal{G}(r)$. We introduce an auxiliary parameter $\lambda > 0$ that we will later optimise over.

Lemma 14 (Concentration on $\mathcal{G}(r)$)**.** *Assume E_{cov} holds. Fix $r, n, \lambda > 0$ and define*

$$\varrho_r = \frac{2r}{\alpha} + 4\eta R^2.$$

Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\sup_{g_{\boldsymbol{w}} \in \mathcal{G}(r)} (P - P_n) g_{\boldsymbol{w}} \leq \frac{r^{\boldsymbol{\Gamma}}_{\text{eff}}(\eta)}{\lambda n} + 4\lambda\varrho_r + 5\sqrt{\frac{\varrho_r \log(1/\delta)}{n}}$$

Proof. The proof follows a standard truncation argument. We clip each function $g_{\boldsymbol{w}} \in \mathcal{G}(r)$ at a threshold m_r into a bounded part $\overline{g_{\boldsymbol{w}}}$ and a tail residual $s_{\boldsymbol{w}}$. The bounded part is controlled via Rademacher complexity and a Bernstein-type inequality (Steps 1-2), while the tail is bounded in Step 3.

Introduce the clipping threshold

$$m_r = \sqrt{\frac{n\varrho_r}{\log(1/\delta)}},$$

and use this to clip each function $g_{\boldsymbol{w}}$ into $\overline{g_{\boldsymbol{w}}}$ with a tail residue $s_{\boldsymbol{w}}$, respectively defined as

$$\overline{g_{\boldsymbol{w}}} := \max(-m_r, \min(g_{\boldsymbol{w}}, m_r)), \quad \text{and} \quad s_{\boldsymbol{w}} := g_{\boldsymbol{w}} - \overline{g_{\boldsymbol{w}}} \quad \text{with } |s_{\boldsymbol{w}}| \leq t_{\boldsymbol{w}} := (|g_{\boldsymbol{w}}| - m_r)_+ .$$

Denote the corresponding class of clipped functions by

$$\overline{\mathcal{G}}(r) = \{\overline{g_{\boldsymbol{w}}} : g_{\boldsymbol{w}} \in \mathcal{G}(r)\}.$$

Step 1: Bounding $\mathbb{E}\mathtt{Rad}_n\overline{\mathcal{G}}(r)$: Recall that for a class $\mathcal{F}$ of functions $f : \mathbb{R}^d \times \{\pm 1\} \to \mathbb{R}$, its Rademacher complexity with respect to the training samples $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ is defined as

$$\mathtt{Rad}_n\mathcal{F} := \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(\boldsymbol{X}_i, Y_i),$$

where ϵ_i are ± 1 Rademacher random variables, and let $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$.

Since the map $g \mapsto \max(-m_r, \min(g, m_r))$ is 1-Lipschitz, a standard contraction result (for example, Theorem A.6 in [4]) gives that, conditional on the samples, the expectations over the Rademacher variables $\boldsymbol{\epsilon}$ satisfy

$$\mathbb{E}_{\boldsymbol{\epsilon}}\mathtt{Rad}_n\overline{\mathcal{G}}(r) \leq \mathbb{E}_{\boldsymbol{\epsilon}}\mathtt{Rad}_n\mathcal{G}(r)\,.$$

Taking an expectation over all samples,

$$\mathbb{E}\mathtt{Rad}_n\overline{\mathcal{G}}(r) \le \mathbb{E}\mathtt{Rad}_n\mathcal{G}(r),$$

and we next upper bound the right hand side.

To do so, note that, since $\boldsymbol{w}^*_{\mathcal{L}}$ does not depend on $g_{\boldsymbol{w}}$, we have

$$\begin{aligned}\mathtt{Rad}_n\mathcal{G}(r) &= \sup_{g_{\boldsymbol{w}}\in\mathcal{G}(r)} \frac{1}{n}\sum_{i=1}^n \epsilon_i \left(\ell(Y_i\boldsymbol{X}_i^\top\boldsymbol{w}) - \ell(Y_i\boldsymbol{X}_i^\top\boldsymbol{w}^*_{\mathcal{L}})\right)\\ &= -\frac{1}{n}\sum_{i=1}^n \epsilon_i\ell(Y_i\boldsymbol{X}_i^\top\boldsymbol{w}^*_{\mathcal{L}}) + \sup_{g_{\boldsymbol{w}}\in\mathcal{G}(r)} \frac{1}{n}\sum_{i=1}^n \epsilon_i\ell(Y_i\boldsymbol{X}_i^\top\boldsymbol{w}),\end{aligned}$$

where taking expectation over $\boldsymbol{\epsilon}$ (and then over the samples) makes the first term vanish, resulting in

$$\mathbb{E}\mathtt{Rad}_n\mathcal{G}(r) = \mathbb{E}\sup_{g_{\boldsymbol{w}}\in\mathcal{G}(r)} \frac{1}{n}\sum_{i=1}^n \epsilon_i\ell(Y_i\boldsymbol{X}_i^\top\boldsymbol{w}).$$

Now, since ℓ is 1-Lipschitz, applying contraction again and using the fact that $\epsilon_i Y_i$ are still Rademacher random variables, we obtain

$$\mathbb{E}\mathtt{Rad}_n\mathcal{G}(r) \le \mathbb{E}\sup_{g_{\boldsymbol{w}}\in\mathcal{G}(r)} \frac{1}{n}\sum_{i=1}^n \epsilon_i Y_i\boldsymbol{X}_i^\top\boldsymbol{w} = \mathbb{E}\sup_{g_{\boldsymbol{w}}\in\mathcal{G}(r)} \frac{1}{n}\sum_{i=1}^n \epsilon_i\boldsymbol{X}_i^\top\boldsymbol{w}.$$

Next, we can subtract $\sum_{i=1}^n \epsilon_i\boldsymbol{X}_i^\top\boldsymbol{w}^*_{\mathcal{L}}$ because in expectation this is 0, so we arrive at

$$\mathbb{E}\mathtt{Rad}_n\mathcal{G}(r) \le \mathbb{E}\sup_{g_{\boldsymbol{w}}\in\mathcal{G}(r)} \frac{1}{n}\sum_{i=1}^n \epsilon_i\boldsymbol{X}_i^\top(\boldsymbol{w}-\boldsymbol{w}^*_{\mathcal{L}}) = \mathbb{E}\sup_{g_{\boldsymbol{w}}\in\mathcal{G}(r)} \boldsymbol{V}^\top(\boldsymbol{w}-\boldsymbol{w}^*_{\mathcal{L}}).$$

where we have introduced $\boldsymbol{V} = \frac{1}{n}\sum_{i=1}^n \epsilon_i\boldsymbol{X}_i \in \mathbb{R}^d$.

Next, recall $\boldsymbol{\Gamma}_\eta = \boldsymbol{\Gamma} + \eta\boldsymbol{I}_d$ and note that for any $\boldsymbol{h}\in\mathbb{R}^d$ and any $\lambda > 0$,

$$\boldsymbol{V}^\top\boldsymbol{h} = \boldsymbol{V}^\top\boldsymbol{h} - \lambda\left\|\boldsymbol{\Gamma}_\eta^{1/2}\boldsymbol{h}\right\|^2 + \lambda\left\|\boldsymbol{\Gamma}_\eta^{1/2}\boldsymbol{h}\right\|^2 \le \sup_{\boldsymbol{u}\in\mathbb{R}^d}\left(\boldsymbol{V}^\top\boldsymbol{u} - \lambda\left\|\boldsymbol{\Gamma}_\eta^{1/2}\boldsymbol{u}\right\|^2\right) + \lambda\left\|\boldsymbol{\Gamma}_\eta^{1/2}\boldsymbol{h}\right\|^2,$$

therefore, setting $\boldsymbol{h} = \boldsymbol{w} - \boldsymbol{w}^*_{\mathcal{L}}$ and taking supremum over $g_{\boldsymbol{w}}\in\mathcal{G}(r)$ yields

$$\sup_{g_{\boldsymbol{w}}\in\mathcal{G}(r)} \boldsymbol{V}^\top(\boldsymbol{w}-\boldsymbol{w}^*_{\mathcal{L}}) \le \sup_{\boldsymbol{u}\in\mathbb{R}^d}\left(\boldsymbol{V}^\top\boldsymbol{u} - \lambda\left\|\boldsymbol{\Gamma}_\eta^{1/2}\boldsymbol{u}\right\|^2\right) + \lambda\sup_{g_{\boldsymbol{w}}\in\mathcal{G}(r)}\left\|\boldsymbol{\Gamma}_\eta^{1/2}(\boldsymbol{w}-\boldsymbol{w}^*_{\mathcal{L}})\right\|^2.$$

Optimising over $\boldsymbol{u}$ in the first supremum we get

$$\sup_{\boldsymbol{u}\in\mathbb{R}^d}\left(\boldsymbol{V}^\top\boldsymbol{u} - \lambda\left\|\boldsymbol{\Gamma}_\eta^{1/2}\boldsymbol{u}\right\|^2\right) = \frac{1}{4\lambda}\left\|\boldsymbol{\Gamma}_\eta^{-1/2}\boldsymbol{V}\right\|^2,$$

whose expectation is

$$\begin{aligned}\frac{1}{4\lambda}\mathbb{E}\left\|\boldsymbol{\Gamma}_\eta^{-1/2}\boldsymbol{V}\right\|^2 &= \frac{1}{4\lambda}\mathbb{E}[\boldsymbol{V}^\top\boldsymbol{\Gamma}_\eta^{-1}\boldsymbol{V}] = \frac{1}{4\lambda}\mathbb{E}[\mathrm{tr}(\boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{\Gamma}_\eta^{-1})] = \frac{1}{4\lambda}\mathrm{tr}(\mathbb{E}[\boldsymbol{V}\boldsymbol{V}^\top]\boldsymbol{\Gamma}_\eta^{-1})\\ &\overset{\text{(i)}}{=} \frac{1}{4\lambda n}\mathrm{tr}(\boldsymbol{\Gamma}\boldsymbol{\Gamma}_\eta^{-1}) \overset{\text{(ii)}}{=} \frac{r_{\text{eff}}^{\boldsymbol{\Gamma}}(\eta)}{4\lambda n},\end{aligned}$$

where (i) uses $\mathbb{E}[\boldsymbol{V}\boldsymbol{V}^\top] = \boldsymbol{\Gamma}/n$ and (ii) applies the definition $r_{\text{eff}}^{\boldsymbol{\Gamma}}(\eta) := \mathrm{tr}(\boldsymbol{\Gamma}\boldsymbol{\Gamma}_\eta^{-1})$. Finally, for the second supremum, using the fact that $\boldsymbol{\Gamma}_\eta = \boldsymbol{\Gamma} + \eta\boldsymbol{I}_d$,

$$\begin{aligned}\lambda\sup_{g_{\boldsymbol{w}}\in\mathcal{G}(r)}\left\|\boldsymbol{\Gamma}_\eta^{1/2}(\boldsymbol{w}-\boldsymbol{w}^*_{\mathcal{L}})\right\|^2 &= \lambda\sup_{g_{\boldsymbol{w}}\in\mathcal{G}(r)}\left(\|\boldsymbol{w}-\boldsymbol{w}^*_{\mathcal{L}}\|_{\boldsymbol{\Gamma}}^2 + \eta\|\boldsymbol{w}-\boldsymbol{w}^*_{\mathcal{L}}\|^2\right)\\ &\le \lambda\left(\frac{2r}{\alpha} + 4\eta R^2\right) = \lambda\varrho_r,\end{aligned} \tag{28}$$

where the inequality holds because of the two inequalities from Lemma 12. Putting everything together, we arrive at

$$\mathbb{E}\mathtt{Rad}_n\overline{\mathcal{G}}(r) \le \frac{r_{\text{eff}}^{\boldsymbol{\Gamma}}(\eta)}{4\lambda n} + \lambda\varrho_r\,. \tag{29}$$

Step 2: High-probability bound for $\overline{g_{\boldsymbol{w}}}$: Each $\overline{g_{\boldsymbol{w}}}$ takes values in $[-m_r, m_r]$, an interval of length $2m_r$. Additionally, clipping cannot increase the variance, hence by Lemma 13,

$$\sup_{\overline{g_{\boldsymbol{w}}} \in \overline{\mathcal{G}}(r)} \mathrm{Var}(\overline{g_{\boldsymbol{w}}}) \le \sup_{g_{\boldsymbol{w}} \in \mathcal{G}(r)} \mathrm{Var}(g_{\boldsymbol{w}}) \le \sup_{g_{\boldsymbol{w}} \in \mathcal{G}(r)} \mathbb{E}[(g_{\boldsymbol{w}})^2] \le \frac{2r}{\alpha} \le \varrho_r.$$

We can now apply Theorem 2.1 in [4] (with $\mathcal{F} = \overline{\mathcal{G}}(r), x = \log(1/\delta), \alpha = 1, a = -m_r, b = m_r, r = \varrho_r$) to get that, with probability at least $1-\delta$,

$$\sup_{\overline{g_{\boldsymbol{w}}} \in \overline{\mathcal{G}}(r)} (P - P_n)\overline{g_{\boldsymbol{w}}} \le 4\mathbb{E}\mathtt{Rad}_n\overline{\mathcal{G}}(r) + \sqrt{\frac{2\varrho_r \log(1/\delta)}{n}} + \frac{8}{3} m_r \frac{\log(1/\delta)}{n}.$$

Using the bound (29) together with $m_r \log(1/\delta)/n = \sqrt{\varrho_r \log(1/\delta)/n}$, we have that, with probability at least $1-\delta$,

$$\sup_{\overline{g_{\boldsymbol{w}}} \in \overline{\mathcal{G}}(r)} (P - P_n)\overline{g_{\boldsymbol{w}}} \le \frac{r_{\mathrm{eff}}^{\boldsymbol{\Gamma}}(\eta)}{\lambda n} + 4\lambda\varrho_r + (\sqrt{2} + 8/3)\sqrt{\frac{\varrho_r \log(1/\delta)}{n}}. \tag{30}$$

Step 3: Bounding the tail residual: Starting from $(|g| - m_r)_+ \le g^2/4m_r$ valid for all $g \in \mathbb{R}$, Lemma 13 gives

$$\sup_{g_{\boldsymbol{w}} \in \mathcal{G}(r)} P(|g_{\boldsymbol{w}}| - m_r)_+ = \sup_{g_{\boldsymbol{w}} \in \mathcal{G}(r)} P t_{\boldsymbol{w}} \le \frac{\varrho_r}{4m_r} = \frac{1}{4}\sqrt{\frac{\varrho_r \log(1/\delta)}{n}}. \tag{31}$$

On the empirical side, the same inequality and Lemma 13 gives

$$t_{\boldsymbol{w}}(\boldsymbol{X}_i, Y_i) = (|g_{\boldsymbol{w}}(\boldsymbol{X}_i, Y_i)| - m_r)_+ \le \frac{g_{\boldsymbol{w}}(\boldsymbol{X}_i, Y_i)^2}{4m_r} \le \frac{(\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*)^\top \boldsymbol{X}_i \boldsymbol{X}_i^\top (\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*)}{4m_r},$$

so that

$$P_n t_{\boldsymbol{w}} \le \frac{1}{4m_r} (\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*)^\top \widehat{\Gamma} (\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*).$$

Now, since $(\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*)^\top \widehat{\boldsymbol{\Gamma}} (\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*) \le \left\| \boldsymbol{\Gamma}_\eta^{1/2} (\boldsymbol{w} - \boldsymbol{w}_{\mathcal{L}}^*) \right\|^2 \left\| \boldsymbol{\Gamma}_\eta^{-1/2} \widehat{\boldsymbol{\Gamma}} \boldsymbol{\Gamma}_\eta^{-1/2} \right\|$, with the assumption that the event E_{cov} holds and the fact that, by (28), $\left\| \boldsymbol{\Gamma}_\eta^{1/2} (w - \boldsymbol{w}_{\mathcal{L}}^*) \right\|^2 \le \varrho_r$,

$$\sup_{g_{\boldsymbol{w}} \in \mathcal{G}(r)} P_n t_{\boldsymbol{w}} \le \frac{2\varrho_r}{4m_r} = \frac{1}{2}\sqrt{\frac{\varrho_r \log(1/\delta)}{n}}. \tag{32}$$

Step 4: Conclusion: Using the decomposition $g_{\boldsymbol{w}} = \overline{g_{\boldsymbol{w}}} + s_{\boldsymbol{w}}$ with $|s_{\boldsymbol{w}}| \le t_{\boldsymbol{w}}$, we have

$$(P - P_n) g_{\boldsymbol{w}} = (P - P_n)\overline{g_{\boldsymbol{w}}} + P s_{\boldsymbol{w}} - P_n s_{\boldsymbol{w}} \le (P - P_n)\overline{g_{\boldsymbol{w}}} + P t_{\boldsymbol{w}} + P_n t_{\boldsymbol{w}}.$$

Taking a supremum over $g_{\boldsymbol{w}} \in \mathcal{G}(r)$ and using (30), (31), and (32), we get that, with probability $1-\delta$,

$$\sup_{g_{\boldsymbol{w}} \in \mathcal{G}(r)} (P - P_n) g_{\boldsymbol{w}} \le \frac{r_{\mathrm{eff}}^{\Gamma}(\eta)}{\lambda n} + 4\lambda\varrho_r + (\sqrt{2} + 35/12 + 1/2)\sqrt{\frac{\varrho_r \log(1/\delta)}{n}}.$$

The conclusion follows by noting that $\sqrt{2} + 35/12 + 1/2 < 5$. □

We now can pick λ conveniently in Lemma 14 and process it further to have only n^{-1} rates:

Corollary 2. *Assume E_{cov} holds. Fix any confidence $\delta > 0$ and any $r, \eta > 0$. Then, with probability $1-\delta$,*

$$\sup_{g_{\boldsymbol{w}} \in \mathcal{G}(r)} (P - P_n) g_{\boldsymbol{w}} \le \frac{128(r_{\mathrm{eff}}^{\boldsymbol{\Sigma}}(\eta) + 1)}{\alpha n} + \frac{1}{8} r + \frac{3}{16} \alpha\eta R^2 + 600 \frac{\log(1/\delta)}{\alpha n}.$$

Proof. First, take $\lambda = \alpha/128$ in Lemma 14 and then use $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ to arrive at

$$\sup_{g_{\boldsymbol{w}} \in \mathcal{G}(r)} (P - P_n) g_{\boldsymbol{w}} \le \frac{128 r_{\mathrm{eff}}^{\boldsymbol{\Gamma}}(\eta)}{\alpha n} + \frac{1}{16} r + \frac{1}{8} \alpha \eta R^2 + \sqrt{\frac{50 r \log(1/\delta)}{\alpha n}} + \sqrt{\frac{100 \eta R^2 \log(1/\delta)}{n}}, \tag{33}$$

Next, applying the inequality $\sqrt{ab} \le a/16 + 4b$ to each of the square root terms we have

$$\sqrt{\frac{50 r \log(1/\delta)}{\alpha n}} \le \frac{1}{16} r + 200 \frac{\log(1/\delta)}{\alpha n},$$
$$\sqrt{\frac{100 \eta R^2 \log(1/\delta)}{n}} = \sqrt{\frac{100 \alpha \eta R^2 \log(1/\delta)}{\alpha n}} \le \frac{1}{16} \alpha \eta R^2 + 400 \frac{\log(1/\delta)}{\alpha n}.$$

Substituting these into (33), we arrive at

$$\sup_{g_{\boldsymbol{w}} \in \mathcal{G}(r)} (P - P_n) g_{\boldsymbol{w}} \le \frac{128 r_{\mathrm{eff}}^{\boldsymbol{\Gamma}}(\eta)}{\alpha n} + \frac{1}{8} r + \frac{3}{16} \alpha \eta R^2 + 600 \frac{\log(1/\delta)}{\alpha n},$$

and the conclusion follows by using $r_{\mathrm{eff}}^{\boldsymbol{\Gamma}} \le r_{\mathrm{eff}}^{\boldsymbol{\Sigma}} + 1$ (c.f. Lemma 10).

□

Now, having proved empirical logistic risks deviation bounds on the shell $\mathcal{G}(r)$, we next extend them to the whole of $\mathcal{B} = \{\boldsymbol{w} : \|\boldsymbol{w}\| \le R\}$ using a peeling argument over r.

C.3.3 Empirical Excess Logistic Risk Concentration on $\mathcal{B}$

Lemma 15 (Concentration on $\mathcal{B}$)**.** *Fix $\delta > 0$ and $\eta > 0$ and assume that $n \ge \mathsf{c}_0 (r_{\mathrm{eff}}^{\boldsymbol{\Sigma}}(\eta) + \log(1/\delta))$ for some universal constant $\mathsf{c}_0 > 0$. With probability at least $1 - 2\delta$, for any $\boldsymbol{w} \in \mathcal{B}$:*

$$(P - P_n) g_{\boldsymbol{w}} \le \frac{1}{4} P g_{\boldsymbol{w}} + \mathsf{c}_1' \left(\frac{r_{\mathrm{eff}}^{\boldsymbol{\Sigma}}(\eta) + \log(\pi^2/(6\delta)) + \log\log(n) + \mathsf{c}_2'}{n\alpha} + \alpha \eta R^2 \right),$$

where c_1' and c_2' have the following form:

$$\mathsf{c}_1' = \frac{21251}{12},$$
$$\mathsf{c}_2' = 1 + \log\left(1 + (\log_2((\log 2 + R\sqrt{\|\boldsymbol{\Gamma}\|})\alpha))_+ + \frac{1}{\log 2} \right).$$

Proof. Given the assumption on the sample size n and Lemma 11, we will work on the event E_{cov}, which holds with probability at least $1 - \delta$. Thus, by a Union Bound, we need show that the upper-bound on $(P - P_n) g_{\boldsymbol{w}}$ from the Lemma holds with probability $1 - \delta$.

Step 1: Bounding $P g_{\boldsymbol{w}}$ for $\boldsymbol{w} \in \mathcal{B}$: First of all, note that if $\boldsymbol{w} \in \mathcal{B}$, i.e. $\|\boldsymbol{w}\| \le R$, then

$$\|\boldsymbol{w}\|_{\boldsymbol{\Gamma}} \le \sqrt{\|\boldsymbol{w}\| \, \|\boldsymbol{\Gamma}\|} \le R\sqrt{\|\boldsymbol{\Gamma}\|} := R_{\boldsymbol{\Gamma}}.$$

Let $M_{\boldsymbol{\Gamma}} = \log 2 + R_{\boldsymbol{\Gamma}}$ and note that $M_{\boldsymbol{\Gamma}} \ge \sup_{\boldsymbol{w} \in \mathcal{B}} P g_{\boldsymbol{w}} = \sup_{\boldsymbol{w} \in \mathcal{B}} (\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}_{\mathcal{L}}^*))$. Indeed, using $0 \le \ell(z) \le |z| + \log 2$ and Jensen's inequality,

$$\sup_{\boldsymbol{w} \in \mathcal{B}} (\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}_{\mathcal{L}}^*)) \le \sup_{\boldsymbol{w} \in \mathcal{B}} \mathcal{L}(\boldsymbol{w}) = \sup_{\boldsymbol{w} \in \mathcal{B}} \mathbb{E}[\ell(Y \boldsymbol{X}^\top \boldsymbol{w})] \le \log 2 + \sup_{\boldsymbol{w} \in \mathcal{B}} \mathbb{E}[|Y \boldsymbol{X}^\top \boldsymbol{w}|]$$
$$\le \log 2 + \sup_{\boldsymbol{w} \in \mathcal{B}} \sqrt{\mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{w})^2]} = \log 2 + \sup_{\boldsymbol{w} \in \mathcal{B}} \|\boldsymbol{w}\|_{\boldsymbol{\Gamma}} \le \log 2 + R_{\boldsymbol{\Gamma}} = M_{\boldsymbol{\Gamma}}.$$

Step 2: Peeling: Now, for $k \ge 0$, set $r_k = 2^{-k} M_{\boldsymbol{\Gamma}}$ and $\delta_k = 6\delta/(\pi^2 (k+1)^2)$, so that $\sum_{k \ge 0} \delta_k = \delta$. Note that

$$\log(1/\delta_k) = 2\log(k+1) + \log(\pi^2/(6\delta)).$$

Applying Corollary 2 for all r_k and δ_k (with the same η), we have, by a Union Bound, that with probability $1-\delta$, for any $k \geq 0$,

$$\sup_{g_{\boldsymbol{w}} \in \mathcal{G}(r_k)} (P - P_n) g_{\boldsymbol{w}} \leq \frac{128(r_{\text{eff}}^{\boldsymbol{\Sigma}}(\eta)+1)}{\alpha n} + \frac{1}{8} r_k + \frac{3}{16}\alpha\eta R^2 + 600\frac{2\log(k+1) + \log(\pi^2/(6\delta))}{\alpha n}. \tag{34}$$

Step 3: Bounding k: We now claim there exists c_3', depending on $R_{\boldsymbol{\Gamma}}$ (through $M_{\boldsymbol{\Gamma}}$) and α only, such that

$$\text{if } r_k = 2^{-k} M_{\boldsymbol{\Gamma}} \geq \frac{1}{n\alpha}, \quad \text{then } \log(k+1) \leq \mathsf{c}_3' + \log\log(n). \tag{35}$$

Indeed, in this case, $k \leq \log_2(n\alpha M_{\boldsymbol{\Gamma}})$, so that

$$k + 1 \leq 1 + (\log_2(M_{\boldsymbol{\Gamma}}\alpha))_+ + \log_2(n).$$

But, as $\log_2(n) = \log(n)/\log(2)$ and, assuming for simplicity $n \geq 3 > e$ (as otherwise we can replace n by $n+e$ everywhere), $\log n \geq 1$,

$$k + 1 \leq \left(1 + (\log_2(M_{\boldsymbol{\Gamma}}\alpha))_+ + \frac{1}{\log 2}\right)\log(n).$$

By taking $\log$ again, we arrive at (35) with

$$\mathsf{c}_3' = \log\left(1 + (\log_2(M_{\boldsymbol{\Gamma}}\alpha))_+ + \frac{1}{\log 2}\right)$$

For the remaining of the proof, we fix $\boldsymbol{w} \in \mathcal{B}$, define $\mathsf{c}_2' = \mathsf{c}_3' + 1$, and introduce the notation

$$\Xi(\eta) := \mathsf{c}_1'\left(\frac{r_{\text{eff}}^{\boldsymbol{\Sigma}}(\eta) + \log(\pi^2/(6\delta)) + \log\log(n) + \mathsf{c}_2'}{n\alpha} + \alpha\eta R^2\right).$$

Note that, up to constants, this is the same as $\Psi(\eta)$ in the hypothesis of Theorem 1.

Step 4: The case $Pg_{\boldsymbol{w}} > \Xi(\eta)$: Since $Pg_{\boldsymbol{w}} \leq M_{\boldsymbol{\Gamma}} = r_0$, there exists some $k \geq 0$ such that $r_{k+1} < Pg_{\boldsymbol{w}} \leq r_k$. In this case $g_{\boldsymbol{w}} \in G(r_k)$, so that (34) applies. Additionally, note that

$$r_k \geq Pg_{\boldsymbol{w}} > \Xi(\eta) \geq \mathsf{c}_1'\frac{\mathsf{c}_2'}{n\alpha} \geq \frac{1}{n\alpha}, \tag{36}$$

as $\mathsf{c}_1', \mathsf{c}_2' > 1$. So, (35) also holds in this case. And, also using $r_k \leq 2Pg_{\boldsymbol{w}}$, we get:

$$\begin{aligned}
(P - P_n)g_{\boldsymbol{w}} &\leq \frac{128(r_{\text{eff}}^{\boldsymbol{\Sigma}}(\eta)+1)}{\alpha n} + \frac{1}{4}Pg_{\boldsymbol{w}} + \frac{3}{16}\alpha\eta R^2 + 600\frac{2\mathsf{c}_3' + 2\log(\log(n)) + \log(\pi^2/(6\delta))}{\alpha n} \\
&\leq \frac{1}{4}Pg_{\boldsymbol{w}} + 1200\left(\frac{r_{\text{eff}}^{\boldsymbol{\Sigma}}(\eta) + 1 + \log(\pi^2/(6\delta)) + \mathsf{c}_3' + \log\log(n)}{n\alpha} + \alpha\eta R^2\right) \\
&\leq \frac{1}{4}Pg_{\boldsymbol{w}} + \mathsf{c}_1'\left(\frac{r_{\text{eff}}^{\boldsymbol{\Sigma}}(\eta) + \log(\pi^2/(6\delta)) + \log\log(n) + \mathsf{c}_2'}{n\alpha} + \alpha\eta R^2\right),
\end{aligned}$$

where the last inequality follows from $\mathsf{c}_2' = \mathsf{c}_3' + 1$ and $\mathsf{c}_1' \geq 1200$.

Step 5: The case $Pg_{\boldsymbol{w}} \leq \Xi(\eta)$: In this case, let k be the smallest index with $r_k \leq 2\Xi(\eta)$. If $k = 0$, then $\log(k+1) \leq \mathsf{c}_3' + \log\log(n)$ clearly holds. Otherwise, by minimality, $r_{k-1} > 2\Xi(\eta)$, and then $r_k = r_{k-1}/2 > \Xi(\eta)$, so that again, by (35) and (36), $\log(k+1) \leq \mathsf{c}_3' + \log\log(n)$.

Moreover, since $Pg_{\boldsymbol{w}} \leq \Xi(\eta) < r_k$, we have $g_{\boldsymbol{w}} \in G(r_k)$ and so (34) applies. Using $r_k \leq 2\Xi(\eta)$:

$$(P - P_n)g_{\boldsymbol{w}} \leq \frac{128(r_{\text{eff}}^{\boldsymbol{\Sigma}}(\eta)+1)}{\alpha n} + \frac{1}{4}\Xi(\eta) + \frac{3}{16}\alpha\eta R^2 + 600\frac{2\mathsf{c}_3' + 2\log\log(n) + \log(\pi^2/(6\delta))}{\alpha n}.$$

By the definition of $\Xi(\eta)$ and the fact that $\mathsf{c}_2' = \mathsf{c}_3' + 1 > 1$, we have the following:

$$\begin{aligned}
\frac{128(r_{\text{eff}}^{\boldsymbol{\Sigma}}(\eta)+1)}{\alpha n} &\leq \frac{128}{\mathsf{c}_1'}\Xi(\eta), \\
\frac{3}{16}\alpha\eta R^2 &\leq \frac{3}{16\mathsf{c}_1'}\Xi(\eta), \\
600\frac{2\mathsf{c}_3' + 2\log\log(n) + \log(\pi^2/(6\delta))}{\alpha n} &\leq \frac{1200}{\mathsf{c}_1'}\Xi(\eta),
\end{aligned}$$

and thus, using the value of c_1', we have

$$\begin{aligned}
(P - P_n)g_{\boldsymbol{w}} &\leq \left(\frac{1}{4} + \frac{128}{\mathsf{c}_1'} + \frac{3}{16\mathsf{c}_1'} + \frac{1200}{\mathsf{c}_1'}\right)\Xi(\eta) = \left(\frac{1}{4} + \frac{21251}{16\mathsf{c}_1'}\right)\Xi(\eta) \\
&= \left(\frac{1}{4} + \frac{3}{4}\right)\Xi(\eta) = \Xi(\eta) \leq \frac{1}{4}Pg_{\boldsymbol{w}} + \Xi(\eta) \\
&= \frac{1}{4}Pg_{\boldsymbol{w}} + \mathsf{c}_1'\left(\frac{r_{\text{eff}}^{\boldsymbol{\Sigma}}(\eta) + \log(\pi^2/(6\delta)) + \log\log(n) + \mathsf{c}_2'}{n\alpha} + \alpha\eta R^2\right).
\end{aligned}$$

This concludes the proof. □

We are now ready to put everything together to prove the excess zero-one risk upper bound in Theorem 1.

C.4 Proof of Theorem 1

Theorem 1 (Upper bound for early-stopped GD)**.** *There exist a universal constant $\mathsf{c}_0 > 0$ and some $\mathsf{c}_1, \mathsf{c}_2, \mathsf{c}_3, \mathsf{c}_4 > 0$ depending on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ only, but not on n or p, such that the following holds:*

Fix $\delta \in (0,1)$ and $n \geq 3$. Suppose there exists $\eta_n > 0$ satisfying $n \geq \mathsf{c}_0(r_{\text{eff}}(\eta_n) + 1 + \log(2/\delta))$ and

$$\Psi(\eta_n) := \frac{r_{\text{eff}}(\eta_n) + \log(1/\delta) + \log\log(n) + \mathsf{c}_2}{n} + \mathsf{c}_3\eta_n \leq \mathsf{c}_4(\rho^*)^2\left\|\boldsymbol{w}^*\right\|_{\boldsymbol{\Sigma}}^2. \tag{3}$$

Then, with probability at least $1 - \delta$,

$$\mathcal{E}(\boldsymbol{w}_\tau) - \mathcal{E}(\boldsymbol{w}^*) \leq \mathsf{c}_1\frac{1-2p}{(\rho^*)^2\left\|\boldsymbol{w}^*\right\|_{\boldsymbol{\Sigma}}}\Psi(\eta_n).$$

Proof. Fix $\eta_n > 0$. Recall that $R = 4\left\|\boldsymbol{w}^*\right\| + 1$. Note that by Lemma 3 (i) we have $\boldsymbol{w}_\tau \in \mathcal{B}$. By Lemma 1, recall that $\boldsymbol{w}_{\mathcal{L}}^* \in \mathcal{B}$. By Lemma 3 (ii), $P_n g_{\boldsymbol{w}_\tau} \leq 0$, so we observe that

$$Pg_{\boldsymbol{w}_\tau} = (P - P_n)g_{\boldsymbol{w}_\tau} + P_n g_{\boldsymbol{w}_\tau} \leq (P - P_n)g_{\boldsymbol{w}_\tau}$$

Now, for c_1' and c_2' as in Lemma 15, consider

$$\Xi(\eta_n) = \frac{\mathsf{c}_1'}{\alpha}\left(\frac{r_{\text{eff}}^{\boldsymbol{\Sigma}}(\eta_n) + \log(\pi^2/(3\delta)) + \log\log(n) + \mathsf{c}_2'}{n} + \alpha^2R^2\eta_n\right).$$

Then, by Lemma 15 (applied with $\delta/2$ instead of δ), with probability at least $1 - \delta$,

$$Pg_{\boldsymbol{w}_\tau} \leq (P - P_n)g_{\boldsymbol{w}_\tau} \leq \frac{1}{4}Pg_{\boldsymbol{w}_\tau} + \Xi(\eta_n),$$

which can be rearranged into

$$\mathcal{L}(\boldsymbol{w}_\tau) - \mathcal{L}(\boldsymbol{w}_{\mathcal{L}}^*) = Pg_{\boldsymbol{w}_\tau} \leq \frac{4}{3}\Xi(\eta_n). \tag{37}$$

Assume that the following inequality holds, which will coincide with the assumption (3) for appropriate c_2, c_3, and c_4, stated at the end of the proof:

$$\Xi(\eta_n) \leq \frac{3}{32}\alpha(\rho^*)^2\left\|\boldsymbol{w}^*\right\|_{\boldsymbol{\Sigma}}^2.$$

Then, by (37) we have

$$\mathcal{L}(\boldsymbol{w}_\tau) - \mathcal{L}(\boldsymbol{w}^*_{\mathcal{L}}) \leq \frac{\alpha}{8}(\rho^*)^2 \left\|\boldsymbol{w}^*\right\|^2_{\boldsymbol{\Sigma}},$$

so that we can apply Lemma 4 and (37) to arrive at the conclusion that, with probability $1-\delta$,

$$\mathcal{E}(\boldsymbol{w}_\tau) - \mathcal{E}(\boldsymbol{w}^*) \leq \frac{8(1-2p)}{(\rho^*)^2\alpha\sqrt{2\pi}\left\|\boldsymbol{w}^*\right\|_{\boldsymbol{\Sigma}}}(\mathcal{L}(\boldsymbol{w}_\tau) - \mathcal{L}(\boldsymbol{w}^*_{\mathcal{L}})) \leq \frac{32(1-2p)}{3(\rho^*)^2\alpha\sqrt{2\pi}\left\|\boldsymbol{w}^*\right\|_{\boldsymbol{\Sigma}}}\Xi(\eta_n).$$

Unfolding the definition of $\Xi(\eta_n)$ and absorbing additional constants into $\mathtt{c}'_1$ and $\mathtt{c}'_2$, we have

$$\mathcal{E}(\boldsymbol{w}_\tau) - \mathcal{E}(\boldsymbol{w}^*) \leq \mathtt{c}_1\frac{1-2p}{(\rho^*)^2\left\|\boldsymbol{w}^*\right\|_{\boldsymbol{\Sigma}}}\left(\frac{r^{\boldsymbol{\Sigma}}_{\mathrm{eff}}(\eta_n) + \log(1/\delta) + \log\log(n) + \mathtt{c}_2}{n} + \mathtt{c}_3\eta_n\right).$$

The last factor corresponds exactly to $\Psi(\eta_n)$, since $r_{\mathrm{eff}}(\eta_n) = r^{\boldsymbol{\Sigma}}_{\mathrm{eff}}(\eta_n)$. The missing definitions of the constants in the Theorem statement are given by

$$\begin{aligned}
&\boldsymbol{\Gamma} = \boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma},\\
&R = 4\left\|\boldsymbol{w}^*\right\| + 1 = 4\left\|\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right\| + 1,\\
&\alpha = \frac{1}{2}\ell''(3R\sqrt{\|\boldsymbol{\Gamma}\|}),\\
&\mathtt{c}_1 = \frac{170008}{9\alpha^2\sqrt{2\pi}},\\
&\mathtt{c}_2 = 1 + \log(\pi^2/3) + \log\left(1 + (\log_2((\log 2 + R\sqrt{\|\boldsymbol{\Gamma}\|})\alpha))_+ + \frac{1}{\log 2}\right),\\
&\mathtt{c}_3 = \alpha^2R^2,\\
&\mathtt{c}_4 = \frac{9\alpha^2}{170008}.
\end{aligned}$$

□

D Proof of the Statistical Lower Bound

Recall that for the purposes of the statistical lower bound we keep the noise covariance $\boldsymbol{\Sigma}$ and the noise-flip probability p fixed, so that the class of Gaussian Mixtures introduced in (1) is effectively parametrised by $\boldsymbol{\mu}$ only. To ease the notation, we will denote the Bayes linear classifier by

$$\boldsymbol{w}^*_{\boldsymbol{\mu}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu},$$

so that, with a slight abuse of notation we write $h^*_{\boldsymbol{\mu}}(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{x}^\top\boldsymbol{w}^*_{\boldsymbol{\mu}})$ and refer to $h : \mathbb{R}^d \to \{\pm 1\}$ as a classifier, rather than a real-valued score function.

With a slight abuse of notation, we will denote by $\mathbb{P}_{\boldsymbol{\mu}}$ and $\mathbb{E}_{\boldsymbol{\mu}}$ the probability and expectation, respectively, when the underlying randomness comes from the marginal of $\boldsymbol{X}$ only.

Unlike in the proof Theorem 1, all spectral quantities will involve the spectrum of $\boldsymbol{\Sigma}$. In particular, the geometry of $\mathrm{cov}(\boldsymbol{X}) = \boldsymbol{\Gamma} = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top$ does not enter this statistical lower bound proof.

D.1 A Triangle Inequality for Excess Zero-One Risk

Lemma 16. *For any $\boldsymbol{\mu} \in \mathbb{R}^d$ with $\left\|\boldsymbol{\mu}\right\|_{\boldsymbol{\Sigma}^{-1}} = 1$ and any $h : \mathbb{R}^d \to \{\pm 1\}$,*

$$\mathcal{E}_{\boldsymbol{\mu}}(h) - \mathcal{E}_{\boldsymbol{\mu}}(\boldsymbol{w}^*_{\boldsymbol{\mu}}) \geq \frac{(1-2p)\sqrt{2\pi}}{16}\left(\mathbb{P}_{\boldsymbol{\mu}}(h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}}(\boldsymbol{X}))\right)^2.$$

Proof. Similarly to (8), we have

$$\mathcal{E}_{\boldsymbol{\mu}}(h) - \mathcal{E}_{\boldsymbol{\mu}}(\boldsymbol{w}^*_{\boldsymbol{\mu}}) = (1-2p)\mathbb{E}_{\boldsymbol{\mu}}\left[|\tanh(\boldsymbol{X}^\top\boldsymbol{w}^*_{\boldsymbol{\mu}})|\mathbb{1}\{h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}}(\boldsymbol{X})\}\right].$$

Consider the thresholding

$$t := \frac{\sqrt{2\pi}}{4}\mathbb{P}_{\boldsymbol{\mu}}(h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}}(\boldsymbol{X})) \leq 1.$$

We proceed as follows:

$$\begin{aligned}
\mathcal{E}_{\boldsymbol{\mu}}(h) - \mathcal{E}_{\boldsymbol{\mu}}(\boldsymbol{w}^*_{\boldsymbol{\mu}}) &= (1-2p)\mathbb{E}_{\boldsymbol{\mu}}\left[|\tanh(\boldsymbol{X}^\top \boldsymbol{w}^*_{\boldsymbol{\mu}})| \mathbb{1}\{h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}}(\boldsymbol{X})\}\right] \\
&\geq (1-2p)\mathbb{E}_{\boldsymbol{\mu}}\left[|\tanh(\boldsymbol{X}^\top \boldsymbol{w}^*_{\boldsymbol{\mu}})| \mathbb{1}\{h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}}(\boldsymbol{X})\} \mathbb{1}\{|\boldsymbol{X}^\top \boldsymbol{w}^*_{\boldsymbol{\mu}}| > t\}\right] \\
&\geq (1-2p)\mathbb{E}_{\boldsymbol{\mu}}\left[\tanh(t) \mathbb{1}\{h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}}(\boldsymbol{X}), |\boldsymbol{X}^\top \boldsymbol{w}^*_{\boldsymbol{\mu}}| > t\}\right] \\
&= (1-2p)\tanh(t)\mathbb{P}_{\boldsymbol{\mu}}(h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}}(\boldsymbol{X}), |\boldsymbol{X}^\top \boldsymbol{w}^*_{\boldsymbol{\mu}}| > t),
\end{aligned}$$

where the second inequality follows from the fact that $|\tanh(t)|$ is even and increasing on $[0, \infty)$. Further, by a Union Bound, we get

$$\mathcal{E}_{\boldsymbol{\mu}}(h) - \mathcal{E}_{\boldsymbol{\mu}}(\boldsymbol{w}^*_{\boldsymbol{\mu}}) \geq (1-2p)\tanh(t)\left(\mathbb{P}_{\boldsymbol{\mu}}(h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}}(\boldsymbol{X})) - \mathbb{P}_{\boldsymbol{\mu}}(|\boldsymbol{X}^\top \boldsymbol{w}^*_{\boldsymbol{\mu}}| \leq t)\right). \tag{38}$$

Now, note that $\boldsymbol{X}^\top \boldsymbol{w}^*_{\boldsymbol{\mu}} \mid \widetilde{Y} \sim \mathcal{N}(\widetilde{Y} \|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}, \|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}) \equiv \mathcal{N}(\widetilde{Y}, 1)$, so that the density of $\boldsymbol{X}^\top \boldsymbol{w}^*_{\boldsymbol{\mu}}$ is bounded above by $1/\sqrt{2\pi}$ and thus, by the definition of t,

$$\mathbb{P}_{\boldsymbol{\mu}}(|\boldsymbol{X}^\top \boldsymbol{w}^*_{\boldsymbol{\mu}}| \leq t) \leq \frac{2t}{\sqrt{2\pi}} = \frac{1}{2}\mathbb{P}_{\boldsymbol{\mu}}(h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}}(\boldsymbol{X})).$$

Plugging this back into (38) and using the fact that $\tanh(t) \geq t/2$ (because $0 \leq t \leq 1$) and the definition of t,

$$\mathcal{E}_{\boldsymbol{\mu}}(h) - \mathcal{E}_{\boldsymbol{\mu}}(\boldsymbol{w}^*_{\boldsymbol{\mu}}) \geq \frac{(1-2p)t}{4}\mathbb{P}_{\boldsymbol{\mu}}(h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}}(\boldsymbol{X})) = \frac{(1-2p)\sqrt{2\pi}}{16}\left(\mathbb{P}_{\boldsymbol{\mu}}(h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}}(\boldsymbol{X}))\right)^2.$$

□

Lemma 17. *For any $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^d$ with $\|\boldsymbol{\mu}_1\|_{\boldsymbol{\Sigma}^{-1}} = \|\boldsymbol{\mu}_2\|_{\boldsymbol{\Sigma}^{-1}} = 1$ and any classifier $h : \mathbb{R}^d \to \{\pm 1\}$,*

$$\mathbb{P}_{\boldsymbol{\mu}_1}(h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}_1}(\boldsymbol{X})) + \mathbb{P}_{\boldsymbol{\mu}_2}(h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}_2}(\boldsymbol{X})) \geq \frac{1}{\pi\sqrt{e}}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\boldsymbol{\Sigma}^{-1}}.$$

Proof. **Step 1:** First, we will prove that we can do a change of measure to the $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$-independent $Q \equiv \mathcal{N}(\mathbf{0}_d, \boldsymbol{\Sigma})$ in the sense that for $\boldsymbol{\mu} \in \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$ (with $\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}} = 1$),

$$\mathbb{P}_{\boldsymbol{\mu}}(h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}}(\boldsymbol{X})) \geq e^{-1/2}\mathbb{P}_{\boldsymbol{X} \sim Q}(h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}}(\boldsymbol{X})). \tag{39}$$

To do so, note that The Cameron-Martin Theorem, together with $\|\boldsymbol{\mu}_1\|_{\boldsymbol{\Sigma}^{-1}} = 1$, give that the Radon-Nikodym derivative of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to Q is

$$\frac{\mathrm{d}\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\mathrm{d}Q}(\boldsymbol{x}) = \exp\left(\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}\right) = \exp\left(\boldsymbol{x}^\top \boldsymbol{w}^*_{\boldsymbol{\mu}} - \frac{1}{2}\right).$$

Thus, since $\boldsymbol{X} \mid \widetilde{Y} \sim \mathcal{N}(\widetilde{Y}\boldsymbol{\mu}, \boldsymbol{\Sigma})$, taking expectation over $\widetilde{Y} = \pm 1$, we have

$$\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{\mu}}}{\mathrm{d}Q}(\boldsymbol{x}) = \frac{1}{2}\exp(-1/2)\left(\exp(\boldsymbol{x}^\top \boldsymbol{w}^*_{\boldsymbol{\mu}}) + \exp(-\boldsymbol{x}^\top \boldsymbol{w}^*_{\boldsymbol{\mu}})\right) \geq \exp(-1/2)$$

which proves (39).

Step 2: Now, using (39) for $\boldsymbol{\mu} = \boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and a Union Bound, we can lower bound the left-hand of the inequality in the conclusion of the Lemma as

$$\begin{aligned}
\text{LHS} &\geq \frac{1}{\sqrt{e}}\Big(\mathbb{P}_{\boldsymbol{X} \sim Q}\big(h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}_1}(\boldsymbol{X})\big) + \mathbb{P}_{\boldsymbol{X} \sim Q}\big(h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}_2}(\boldsymbol{X})\big)\Big) \\
&\geq \frac{1}{\sqrt{e}}\mathbb{P}_{\boldsymbol{X} \sim Q}\big(h^*_{\boldsymbol{\mu}_1}(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}_2}(\boldsymbol{X})\big) \\
&= \frac{1}{\sqrt{e}}\mathbb{P}_{\boldsymbol{X} \sim Q}\big(\operatorname{sign}(\boldsymbol{X}^\top \boldsymbol{w}^*_{\boldsymbol{\mu}_1}) \neq \operatorname{sign}(\boldsymbol{X}^\top \boldsymbol{w}^*_{\boldsymbol{\mu}_2})\big).
\end{aligned}$$

Step 3: Thus, it remains to show that

$$\mathop{\mathbb{P}}_{\boldsymbol{X}\sim Q}(\operatorname{sign}(\boldsymbol{X}^\top \boldsymbol{w}^*_{\boldsymbol{\mu}_1}) \neq \operatorname{sign}(\boldsymbol{X}^\top \boldsymbol{w}^*_{\boldsymbol{\mu}_2})) \geq \frac{1}{\pi} \left\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right\|_{\boldsymbol{\Sigma}^{-1}}. \tag{40}$$

To do so, we will work with the whitened $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{\Sigma})$. Recall the notation $\boldsymbol{X} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}$ so that $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{I}_d)$, and additionally consider $\boldsymbol{v}^*_1 := \boldsymbol{\Sigma}^{1/2}\boldsymbol{w}^*_{\boldsymbol{\mu}_1} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_1$, and $\boldsymbol{v}^*_2 := \boldsymbol{\Sigma}^{1/2}\boldsymbol{w}^*_{\boldsymbol{\mu}_2} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_2$ with $\|\boldsymbol{v}^*_1\| = \|\boldsymbol{v}^*_2\| = 1$. So, (40) reduces to

$$\mathbb{P}(\operatorname{sign}(\boldsymbol{Z}^\top \boldsymbol{v}^*_1) \neq \operatorname{sign}(\boldsymbol{Z}^\top \boldsymbol{v}^*_2)) \geq \frac{\|\boldsymbol{v}^*_1 - \boldsymbol{v}^*_2\|}{\pi}, \tag{41}$$

where the probability is taken over $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{I}_d)$ and we suppress this for convenience. To prove (41) , note that

$$\mathbb{P}(\operatorname{sign}(\boldsymbol{Z}^\top \boldsymbol{v}^*_1) \neq \operatorname{sign}(\boldsymbol{Z}^\top \boldsymbol{v}^*_2)) = 1 - 2\mathbb{P}(\boldsymbol{Z}^\top \boldsymbol{v}^*_1 > 0, \boldsymbol{Z}^\top \boldsymbol{v}^*_2 > 0).$$

Let $\theta \in [0, \pi]$ be the angle between $\boldsymbol{v}^*_1$ and $\boldsymbol{v}^*_2$, i.e. $(v^*_1)^\top \boldsymbol{v}^*_2 = \cos\theta$. Then, the scalar random variables $\boldsymbol{Z}^\top \boldsymbol{v}^*_1$ and $\boldsymbol{Z}^\top \boldsymbol{v}^*_2$ are both mean 0, variance 1, and have correlation $\cos\theta$, so we can write

$$\boldsymbol{Z}^\top \boldsymbol{v}^*_2 = \boldsymbol{Z}^\top \boldsymbol{v}^*_1 \cos\theta + |\sin\theta| N,$$

where $N \sim \mathcal{N}(0,1)$ is independent of $\boldsymbol{Z}^\top \boldsymbol{v}^*_1$. Then, recalling that ϕ and Φ are the p.d.f. and c.d.f, respectively, of $\mathcal{N}(0,1)$, we have

$$\mathbb{P}(\boldsymbol{Z}^\top \boldsymbol{v}^*_1 > 0, \boldsymbol{Z}^\top \boldsymbol{v}^*_2 > 0) = \int_0^\infty \phi(t)\, \mathbb{P}\left(N > -\frac{t\cos\theta}{|\sin\theta|}\right) \mathrm{d}t = \int_0^\infty \phi(t)\Phi\left(\frac{t\cos\theta}{|\sin\theta|}\right) \mathrm{d}t.$$

Note that this also is valid at $\theta \in \{0, \pi\}$: if $\theta = 0$, then $\boldsymbol{Z}^\top \boldsymbol{v}^*_1 = \boldsymbol{Z}^\top \boldsymbol{v}^*_2$ and the previous equality reads

$$\mathbb{P}(\boldsymbol{Z}^\top \boldsymbol{v}^*_1 > 0) = \int_0^\infty \phi(t)\mathrm{d}t,$$

whereas for $\theta = \pi$, $\boldsymbol{Z}^\top \boldsymbol{v}^*_1 = -\boldsymbol{Z}^\top \boldsymbol{v}^*_2$ and

$$\mathbb{P}(\boldsymbol{Z}^\top \boldsymbol{v}^*_1 > 0, \boldsymbol{Z}^\top \boldsymbol{v}^*_2 > 0) = 0 = \int_0^\infty \phi(t)\Phi(-\infty)\mathrm{d}t.$$

So, using the identity

$$\int_0^\infty \phi(t)\Phi(at)\mathrm{d}t = \frac{1}{4} + \frac{\arctan a}{2\pi},$$

valid for any $a \in \mathbb{R}$, we get

$$\mathbb{P}(\operatorname{sign}(\boldsymbol{Z}^\top \boldsymbol{v}^*_1) \neq \operatorname{sign}(\boldsymbol{Z}^\top \boldsymbol{v}^*_2)) = 1 - 2\left(\frac{1}{4} + \frac{\arctan(\cos\theta/|\sin\theta|)}{2\pi}\right) = \frac{\theta}{\pi}.$$

We arrive at (41) by using $\theta \geq 2\sin(\theta/2)$ (recall that $\theta \geq 0$) and $\|\boldsymbol{v}^*_1\| = \|\boldsymbol{v}^*_2\| = 1$:

$$\frac{\theta}{\pi} \geq \frac{2\sin(\theta/2)}{\pi} = \frac{\sqrt{2 - 2\cos\theta}}{\pi} = \frac{\sqrt{2 - 2(\boldsymbol{v}^*_1)^\top \boldsymbol{v}^*_2}}{\pi} = \frac{\|\boldsymbol{v}^*_1 - \boldsymbol{v}^*_2\|}{\pi}.$$

□

Putting together Lemmas 16 and 17, we arrive at:

Lemma 18 ("Triangle inequality" for zero-one excess risk)**.** *For any $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^d$ with $\|\boldsymbol{\mu}_1\|_{\boldsymbol{\Sigma}^{-1}} = \|\boldsymbol{\mu}_2\|_{\boldsymbol{\Sigma}^{-1}} = 1$ and any classifier $h : \mathbb{R}^d \to \{-1, 1\}$, we have*

$$\left(\mathcal{E}_{\boldsymbol{\mu}_1}(h) - \mathcal{E}_{\boldsymbol{\mu}_1}(\boldsymbol{w}^*_{\boldsymbol{\mu}_1})\right) + \left(\mathcal{E}_{\boldsymbol{\mu}_2}(h) - \mathcal{E}_{\boldsymbol{\mu}_2}(\boldsymbol{w}^*_{\boldsymbol{\mu}_2})\right) \geq \mathsf{c}_1(1-2p)\left\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right\|^2_{\boldsymbol{\Sigma}^{-1}},$$

where

$$\mathsf{c}_1 = \frac{\sqrt{2}}{32\pi^{3/2}e}.$$

Proof. Using Lemma 16, the inequality $a^2 + b^2 \geq (a+b)^2/2$, and Lemma 17, we have that the left hand side $\left(\mathcal{E}_{\boldsymbol{\mu}_1}(h) - \mathcal{E}_{\boldsymbol{\mu}_1}(\boldsymbol{w}^*_{\boldsymbol{\mu}_1})\right) + \left(\mathcal{E}_{\boldsymbol{\mu}_2}(h) - \mathcal{E}_{\boldsymbol{\mu}_2}(\boldsymbol{w}^*_{\boldsymbol{\mu}_2})\right)$ is at least

$$\begin{aligned}
&\geq \frac{(1-2p)\sqrt{2\pi}}{16}\left(\left(\mathbb{P}_{\boldsymbol{\mu}_1}(h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}_1}(\boldsymbol{X}))\right)^2 + \left(\mathbb{P}_{\boldsymbol{\mu}_2}(h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}_2}(\boldsymbol{X}))\right)^2\right) \\
&\geq \frac{(1-2p)\sqrt{2\pi}}{32}\left(\mathbb{P}_{\boldsymbol{\mu}_1}(h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}_1}(\boldsymbol{X})) + \mathbb{P}_{\boldsymbol{\mu}_2}(h(\boldsymbol{X}) \neq h^*_{\boldsymbol{\mu}_2}(\boldsymbol{X}))\right)^2 \\
&\geq \frac{(1-2p)\sqrt{2}}{32\pi^{3/2}e} \cdot \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2_{\boldsymbol{\Sigma}^{-1}} \ .
\end{aligned}$$

□

D.2 Information Theory Tools

Lemma 19. *For any $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^d$,*

$$\mathrm{KL}(P_{\boldsymbol{\mu}_1} \parallel P_{\boldsymbol{\mu}_2}) \leq \frac{1}{2}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2_{\boldsymbol{\Sigma}^{-1}} \ .$$

Proof. By standard KL-facts, for any $\widetilde{y} \in \{\pm 1\}$,

$$\mathrm{KL}(\mathcal{N}(\widetilde{y}\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \parallel \mathcal{N}(\widetilde{y}\boldsymbol{\mu}_2, \boldsymbol{\Sigma})) = \frac{1}{2}\|\widetilde{y}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\|^2_{\boldsymbol{\Sigma}^{-1}} = \frac{1}{2}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2_{\boldsymbol{\Sigma}^{-1}} .$$

Hence, denoting by $\widetilde{P}_{\boldsymbol{\mu}_1}$ and $\widetilde{P}_{\boldsymbol{\mu}_2}$ the laws of $(\boldsymbol{X}, \widetilde{Y})$ under $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively, since $\widetilde{Y}$ has the same law irrespective of $\boldsymbol{\mu}_1$ or $\boldsymbol{\mu}_2$:

$$\mathrm{KL}(\widetilde{P}_{\boldsymbol{\mu}_1} \parallel \widetilde{P}_{\boldsymbol{\mu}_2}) = \mathbb{E}_{\widetilde{Y}}\left[\mathrm{KL}(\mathcal{N}(\widetilde{Y}\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \parallel \mathcal{N}(\widetilde{Y}\boldsymbol{\mu}_2, \boldsymbol{\Sigma}))\right] = \frac{1}{2}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2_{\boldsymbol{\Sigma}^{-1}} .$$

Finally, when including the random classification noise $\boldsymbol{\varepsilon}$, the data processing inequality gives

$$\mathrm{KL}(P_{\boldsymbol{\mu}_1} \parallel P_{\boldsymbol{\mu}_2}) \leq \mathrm{KL}(\widetilde{P}_{\boldsymbol{\mu}_1} \parallel \widetilde{P}_{\boldsymbol{\mu}_2}) \leq \frac{1}{2}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2_{\Sigma^{-1}} \ .$$

□

Now, recall that $\boldsymbol{\Sigma}$ is a symmetric positive semidefinite matrix and let $(\boldsymbol{u}_i, \lambda_i)_{i=1}^d$ be its eigendecomposition, with $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_d > 0$ and $\|\boldsymbol{u}_i\| = 1$. We will show that the lower bound holds with respect to $\boldsymbol{\mu} \in \mathcal{U} \subseteq \mathbb{R}^d$ defined as follows:

$$\mathcal{U} = \left\{\boldsymbol{\mu} : \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}} = 1, \left\|\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right\| \leq \sqrt{\frac{2}{\lambda_1}}\right\}.$$

Additionally, recall that for $\widetilde{\eta} \in (0, \lambda_2]$ we consider the eigenvalue counting function

$$N(\widetilde{\eta}) = |\{i : 1 \leq i \leq d, \lambda_i \geq \widetilde{\eta}\}|,$$

and note that $N(\widetilde{\eta}) \geq 2$.

We are now ready to present the packing that will be used throughout our statistical lower bound.

Lemma 20 (Dense packing of $\mathcal{U}$)**.** *There exists a universal constant $\frac{1}{2} < \mathtt{c}_2 < 1$ such that the following holds:*

Fix $\widetilde{\eta} \in (0, \lambda_2/2]$ and assume $N(2\widetilde{\eta}) \geq 9$. Consider any $\kappa > 0$ with $\kappa^2 \leq 2\widetilde{\eta}/\lambda_1$. Then, there exist a subset $\mathcal{U}_\kappa \subset \mathcal{U}$ with $\log|\mathcal{U}_\kappa| \geq \mathtt{c}_2(N(2\widetilde{\eta}) - 1)$ and for any $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathcal{U}_\kappa$ with $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$,

$$\frac{\kappa}{4} \leq \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\boldsymbol{\Sigma}^{-1}} \leq 2\kappa.$$

Proof. We start with the Gilbert-Varshamov bound, which states that there exists a constant $1 > \mathtt{c}_2 > 1/2$ (in fact, it can be taken as $\mathtt{c}_2 = (1 - H_2(1/64))\log 2 \approx 0.61$) such that, if $N(2\widetilde{\eta}) \geq 9$, then there is a packing subset $\mathcal{P} \subset \{\pm 1\}^{N(2\widetilde{\eta})-1}$ with $\log|\mathcal{P}| \geq \mathtt{c}_2(N(2\widetilde{\eta}) - 1)$ and $d_H(\boldsymbol{a}, \boldsymbol{a}') \geq$

$(N(2\widetilde{\eta})-1)/64$ for any $\boldsymbol{a},\boldsymbol{a}'\in\mathcal{P}$, where d_H denotes the Hamming distance. Define the set of $N(2\widetilde{\eta})-1$ dimensional set of unit vectors

$$\mathcal{V}=\left\{\frac{1}{\sqrt{N(2\widetilde{\eta})-1}}\boldsymbol{a}:\boldsymbol{a}\in\mathcal{P}\right\}.$$

Then $\log|\mathcal{V}|\geq \mathtt{c}_2(N(2\widetilde{\eta})-1)$ and for any $\boldsymbol{v}=\boldsymbol{a}/\sqrt{(N(2\widetilde{\eta})-1)},\boldsymbol{v}'=\boldsymbol{a}'/\sqrt{(N(2\widetilde{\eta})-1)}\in\mathcal{V}$ (with $\boldsymbol{a},\boldsymbol{a}'\in\mathcal{P}$),

$$\|\boldsymbol{v}-\boldsymbol{v}'\|^2=\frac{4}{N(2\widetilde{\eta})-1}d_H(\boldsymbol{a},\boldsymbol{a}')\geq\frac{4}{N(2\widetilde{\eta})-1}\cdot\frac{N(2\widetilde{\eta})-1}{64}=\frac{1}{16},\quad\text{i.e. }\|\boldsymbol{v}-\boldsymbol{v}'\|\geq\frac{1}{4}. \tag{42}$$

We next convert elements of $\mathcal{V}\subset\mathbb{R}^{N(2\widetilde{\eta})-1}$ to elements of $\mathcal{U}$. Each $\boldsymbol{v}\in\mathcal{V}$ has $N(2\widetilde{\eta})-1$ coordinates, and we denote its i'th coordinate by v_i. Since $\widetilde{\eta}\leq\lambda_2/2$, we have

$$N(2\widetilde{\eta})-1=|\{i:1\leq i\leq d,\lambda_i\geq 2\widetilde{\eta}\}|-1=|\{i:2\leq i\leq d,\lambda_i\geq 2\widetilde{\eta}\}|\in\{2,\dots,d\}.$$

and, recalling the d eigenvectors $\boldsymbol{u}_i$ of $\boldsymbol{\Sigma}$, we can define

$$\boldsymbol{\mu}_{\boldsymbol{v}}=\sqrt{\lambda_1(1-\kappa^2)}\boldsymbol{u}_1+\kappa\sum_{i=2}^{N(2\widetilde{\eta})}v_{i-1}\sqrt{\lambda_i}\boldsymbol{u}_i\in\mathbb{R}^d.$$

We consider all these vectors to define

$$\mathcal{U}_\kappa=\{\boldsymbol{\mu}_{\boldsymbol{v}}:\boldsymbol{v}\in\mathcal{V}\}.$$

Firstly, this satisfies

$$\log|\mathcal{U}_\kappa|=\log|\mathcal{V}|\geq\mathtt{c}_2(N(2\widetilde{\eta})-1),$$

and we next show that $\mathcal{U}_\kappa\subseteq\mathcal{U}$.

Indeed, for any such $\boldsymbol{\mu}_{\boldsymbol{v}}\in\mathcal{U}_\kappa$ we have

$$\|\boldsymbol{\mu}_{\boldsymbol{v}}\|^2_{\boldsymbol{\Sigma}^{-1}}=(1-\kappa^2)+\kappa^2\|\boldsymbol{v}\|^2=1,$$

and, using the fact that $\lambda_i\geq 2\widetilde{\eta}$ for $i\in\{2,\dots,N(2\widetilde{\eta})\}$ and the fact that $\kappa^2\leq 2\widetilde{\eta}/\lambda_1$,

$$\left\|\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\boldsymbol{v}}\right\|^2=\frac{1-\kappa^2}{\lambda_1}+\kappa^2\sum_{i=2}^{N(2\widetilde{\eta})}\frac{v_{i-1}^2}{\lambda_i}\leq\frac{1}{\lambda_1}+\frac{\kappa^2}{2\widetilde{\eta}}\sum_{i=2}^{N(2\widetilde{\eta})}v_{i-1}^2=\frac{1}{\lambda_1}+\frac{\kappa^2}{2\widetilde{\eta}}\leq\frac{2}{\lambda_1}.$$

Finally, let $\boldsymbol{\mu}_{\boldsymbol{v}_1},\boldsymbol{\mu}_{\boldsymbol{v}_2}\in\mathcal{U}_\kappa$. Note that, using (42),

$$\|\boldsymbol{\mu}_{\boldsymbol{v}_1}-\boldsymbol{\mu}_{\boldsymbol{v}_2}\|_{\boldsymbol{\Sigma}^{-1}}=\left\|\kappa\sum_{i=2}^{N(2\widetilde{\eta})}(v_{1,i-1}-v_{2,i-1})\sqrt{\lambda_i}\boldsymbol{u}_i\right\|_{\boldsymbol{\Sigma}^{-1}}=\kappa\|\boldsymbol{v}_1-\boldsymbol{v}_2\|\geq\frac{\kappa}{4},$$

and, since $\boldsymbol{v}_1,\boldsymbol{v}_2$ are unit vectors, using the triangle inequality,

$$\|\boldsymbol{\mu}_{\boldsymbol{v}_1}-\boldsymbol{\mu}_{\boldsymbol{v}_2}\|_{\boldsymbol{\Sigma}^{-1}}=\kappa\|\boldsymbol{v}_1-\boldsymbol{v}_2\|\leq 2\kappa.$$

This concludes the proof. □

The final ingredient that we will need in the lower bound proof is:

Lemma 21 (Fano's inequality)**.** *Let U, S, and $\widehat{U}$ be three random variables where U and $\widehat{U}$ are conditionally independent, given S, and U and $\widehat{U}$ take values in a finite set of size M, and U is uniform. Then:*

$$\mathbb{P}(U\neq\widehat{U})\geq 1-\frac{I(U;S)+\ln 2}{\ln(M)},$$

where $I(U;S)$ denotes the mutual information between U and S.

D.3 Proof of Theorem 2

Theorem 2 (Statistical lower bound)**.** *There exists a universal constant $\mathsf{c} > 0$ such that the following holds: Fix n and let $\widetilde{\eta}_n \in (0, \lambda_2/2]$ be small enough so that $N(2\widetilde{\eta}_n) > 9$. Let $\widehat{h}$ be any procedure. Then, there exists $\boldsymbol{\mu} \in \mathbb{R}^d$ with $\left\|\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right\| \le \sqrt{2/\lambda_1}$ such that*

$$\mathop{\mathbb{E}}_{S\sim P_{\boldsymbol{\mu}}^{\otimes n}}\left[\mathcal{E}(\widehat{h}_S) - \mathcal{E}(\boldsymbol{w}^*)\right] \ge \mathsf{c}(1-2p)\min\left\{\frac{2\widetilde{\eta}_n}{\lambda_1}, \frac{N(2\widetilde{\eta}_n)-1}{n}\right\}.$$

Proof. Fix $\widetilde{\eta} \in (0, \lambda_2/2]$ and recall the constant $1/2 < \mathsf{c}_2 < 1$ from Lemma 20. We will apply this Lemma with

$$\kappa = \frac{\sqrt{\mathsf{c}_2}}{4}\min\left\{\sqrt{\frac{2\widetilde{\eta}}{\lambda_1}}, \sqrt{\frac{N(2\widetilde{\eta})-1}{n}}\right\}. \tag{43}$$

Note that, as $\mathsf{c}_2 < 1$, we have $\kappa^2 \le 2\widetilde{\eta}/\lambda_1$, so that Lemma 20 can be indeed applied. Thus, there exists a subset $\mathcal{U}_\kappa$ of $\mathcal{U}$ with $\log|\mathcal{U}_\kappa| \ge \mathsf{c}_2(N(2\widetilde{\eta})-1)$ and $\kappa/4 \le \|\boldsymbol{\mu}-\boldsymbol{\mu}'\|_{\boldsymbol{\Sigma}^{-1}} \le 2\kappa$ for any distinct $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{U}_\kappa$.

Using Lemma 18, for any $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathcal{U}_\kappa$,

$$\begin{aligned}\left(\mathcal{E}_{\boldsymbol{\mu}_1}(h) - \mathcal{E}_{\boldsymbol{\mu}_1}(\boldsymbol{w}^*_{\boldsymbol{\mu}_1})\right) + \left(\mathcal{E}_{\boldsymbol{\mu}_2}(h) - \mathcal{E}_{\boldsymbol{\mu}_2}(\boldsymbol{w}^*_{\boldsymbol{\mu}_2})\right) &\ge \mathsf{c}_1(1-2p)\left\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right\|^2_{\boldsymbol{\Sigma}^{-1}} \\ &\ge \frac{\mathsf{c}_1(1-2p)\kappa^2}{16}\mathbb{1}\{\boldsymbol{\mu}_1 \ne \boldsymbol{\mu}_2\}. \end{aligned}\tag{44}$$

Now, fix some $\boldsymbol{\mu} \in \mathcal{U}_\kappa$ and draw $S \sim P_{\boldsymbol{\mu}}^{\otimes n}$. Let $\widehat{h}$ be any procedure and $\widehat{h}_S : \boldsymbol{X} \in \mathbb{R}^d \mapsto Y \in \{\pm 1\}$ be the induced sample-based estimator. Define the decoder

$$\widehat{\boldsymbol{\mu}}_S = \mathop{\arg\min}_{\boldsymbol{\mu}' \in \mathcal{U}_\kappa}\left(\mathcal{E}_{\boldsymbol{\mu}'}(\widehat{h}_S) - \mathcal{E}_{\boldsymbol{\mu}'}(\boldsymbol{w}^*_{\boldsymbol{\mu}'})\right), \tag{45}$$

which is a function of S and an element of $\mathcal{U}_\kappa$. Note that, using (44) with $\boldsymbol{\mu}_1 := \boldsymbol{\mu}$ and $\boldsymbol{\mu}_2 = \widehat{\boldsymbol{\mu}}_S$, this further implies that

$$\begin{aligned}\mathcal{E}_{\boldsymbol{\mu}}(\widehat{h}_S) - \mathcal{E}_{\boldsymbol{\mu}}(\boldsymbol{w}^*_{\boldsymbol{\mu}}) &\ge \frac{1}{2}\left(\left(\mathcal{E}_{\boldsymbol{\mu}}(\widehat{h}_S) - \mathcal{E}_{\boldsymbol{\mu}}(\boldsymbol{w}^*_{\boldsymbol{\mu}})\right) + \left(\mathcal{E}_{\widehat{\boldsymbol{\mu}}_S}(\widehat{h}_S) - \mathcal{E}_{\widehat{\boldsymbol{\mu}}_S}(\boldsymbol{w}^*_{\widehat{\boldsymbol{\mu}}_S})\right)\right) \\ &\ge \frac{\mathsf{c}_1(1-2p)\kappa^2}{32}\mathbb{1}\{\boldsymbol{\mu} \ne \widehat{\boldsymbol{\mu}}_S\}.\end{aligned}$$

Thus, taking expectations over $S \sim P_{\boldsymbol{\mu}}^{\otimes n}$ and then supremum over $\boldsymbol{\mu} \in \mathcal{U}$, we arrive at

$$\sup_{\boldsymbol{\mu}\in\mathcal{U}_\kappa}\mathop{\mathbb{E}}_{S\sim P_{\boldsymbol{\mu}}^{\otimes n}}\left[\mathcal{E}_{\boldsymbol{\mu}}(\widehat{h}_S) - \mathcal{E}_{\boldsymbol{\mu}}(\boldsymbol{w}^*_{\boldsymbol{\mu}})\right] \ge \frac{c_1(1-2p)\kappa^2}{32}\sup_{\boldsymbol{\mu}\in\mathcal{U}_\kappa}\mathop{\mathbb{P}}_{S\sim P_{\boldsymbol{\mu}}^{\otimes n}}(\boldsymbol{\mu} \ne \widehat{\boldsymbol{\mu}}_S). \tag{46}$$

We are now ready to apply Fano's inequality (c.f. Lemma 21). Consider a uniform random variable $\boldsymbol{U}$ over $\mathcal{U}_\kappa$ and, conditional on $\boldsymbol{U} = \boldsymbol{\mu}$, draw $S \sim P_{\boldsymbol{\mu}}^{\otimes n}$, and let $\widehat{\boldsymbol{\mu}}_S$ the estimator based on the sample S defined in (45). Note that $\widehat{\boldsymbol{\mu}}_S$ and $\boldsymbol{U}$ are conditionally independent given S. Thus,

$$\mathop{\mathbb{P}}_{\boldsymbol{U},S}(\boldsymbol{U} \ne \widehat{\boldsymbol{\mu}}_S) \ge 1 - \frac{I(\boldsymbol{U};S)+\log 2}{\log|\mathcal{U}_\kappa|}.$$

By the law of total probability we have

$$\mathop{\mathbb{P}}_{\boldsymbol{U},S}(\boldsymbol{U} \ne \widehat{\boldsymbol{\mu}}_S) = \frac{1}{|\mathcal{U}_\kappa|}\sum_{\boldsymbol{\mu}\in\mathcal{U}_\kappa}\mathop{\mathbb{P}}_{S\sim P_{\boldsymbol{\mu}}^{\otimes n}}(\boldsymbol{\mu} \ne \widehat{\boldsymbol{\mu}}_S),$$

so that

$$\sup_{\boldsymbol{\mu}\in\mathcal{U}_\kappa}\mathop{\mathbb{P}}_{S\sim P_{\boldsymbol{\mu}}^{\otimes n}}(\boldsymbol{\mu} \ne \widehat{\boldsymbol{\mu}}_S) \ge \mathop{\mathbb{P}}_{\boldsymbol{U},S}(\boldsymbol{U} \ne \widehat{\boldsymbol{\mu}}_S) \ge 1 - \frac{I(\boldsymbol{U};S)+\log 2}{\log|\mathcal{U}_\kappa|}. \tag{47}$$

We now work on the right-hand side term of (47). For that, define the average law

$$\bar{P} = \frac{1}{|\mathcal{U}_\kappa|} \sum_{\boldsymbol{\mu} \in \mathcal{U}_\kappa} P_{\boldsymbol{\mu}}^{\otimes n}.$$

Then,

$$\begin{aligned}
I(\boldsymbol{U}; S) &\overset{\text{(i)}}{=} \frac{1}{|\mathcal{U}_\kappa|} \sum_{\boldsymbol{\mu} \in \mathcal{U}_\kappa} \mathrm{KL}(P_{\boldsymbol{\mu}}^{\otimes n} \parallel \bar{P}) \\
&\overset{\text{(ii)}}{\leq} \frac{1}{|\mathcal{U}_\kappa|} \sum_{\boldsymbol{\mu} \in \mathcal{U}_\kappa} \frac{1}{|\mathcal{U}_\kappa|} \sum_{\boldsymbol{\mu}' \in \mathcal{U}_\kappa} \mathrm{KL}(P_{\boldsymbol{\mu}}^{\otimes n} \parallel P_{\boldsymbol{\mu}'}^{\otimes n}) \\
&\leq \max_{\substack{\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{U}_\kappa \\ \boldsymbol{\mu} \neq \boldsymbol{\mu}'}} \mathrm{KL}(P_{\boldsymbol{\mu}}^{\otimes n} \parallel P_{\boldsymbol{\mu}'}^{\otimes n}) \\
&\overset{\text{(iii)}}{\leq} \frac{n}{2} \max_{\substack{\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{U}_\kappa \\ \boldsymbol{\mu} \neq \boldsymbol{\mu}'}} \|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_{\boldsymbol{\Sigma}^{-1}}^2 \\
&\overset{\text{(iv)}}{\leq} 2n\kappa^2 \\
&\overset{\text{(v)}}{\leq} \frac{1}{8} \mathtt{c}_2 (N(2\widetilde{\eta}) - 1) \\
&\leq \frac{1}{8} \log |\mathcal{U}_\kappa| \,,
\end{aligned}$$

where (i) uses the definition of mutual information together with conditioning on $\boldsymbol{U}$, (ii) uses convexity of the KL in the second argument, (iii) uses KL tensorisation and Lemma 19, (iv) follows from the fact that $\|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_{\boldsymbol{\Sigma}^{-1}} \leq 2\kappa$ for any $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{U}_\kappa$, and (v) uses the definition of κ in (43).

Thus, plugging this into (47),

$$\sup_{\boldsymbol{\mu} \in \mathcal{U}_\kappa} \mathbb{P}_{S \sim P_{\boldsymbol{\mu}}^{\otimes n}} (\boldsymbol{\mu} \neq \widehat{\boldsymbol{\mu}}_S) \geq 1 - \frac{\frac{1}{8} \log |\mathcal{U}_\kappa| + \log 2}{\log |\mathcal{U}_\kappa|} = \frac{7}{8} - \frac{\log 2}{\log |\mathcal{U}_\kappa|}.$$

Since $\mathtt{c}_2 > 1/2$ and $N(2\widetilde{\eta}) \geq 9$, we have $\log |\mathcal{U}_\kappa| \geq \mathtt{c}_2 (N(2\widetilde{\eta}) - 1) > 1/2 \cdot 8 = 4$, and thus we arrive at

$$\sup_{\boldsymbol{\mu} \in \mathcal{U}_\kappa} \mathbb{P}_{S \sim P_{\boldsymbol{\mu}}^{\otimes n}} (\boldsymbol{\mu} \neq \widehat{\boldsymbol{\mu}}_S) \geq \frac{7}{8} - \frac{\log 2}{4} \geq \frac{3}{8},$$

and then (46) gives

$$\sup_{\boldsymbol{\mu} \in \mathcal{U}_\kappa} \mathbb{E}_{S \sim P_{\boldsymbol{\mu}}^{\otimes n}} \left[\mathcal{E}_{\boldsymbol{\mu}}(\widehat{h}_S) - \mathcal{E}_{\boldsymbol{\mu}}(\boldsymbol{w}_{\boldsymbol{\mu}}^*) \right] \geq \frac{3\mathtt{c}_1 (1 - 2p)}{256} \kappa^2.$$

Finally, as $\widehat{h}$ was chosen arbitrarily and using the definition of κ in (43),

$$\inf_{\widehat{h}} \sup_{\boldsymbol{\mu} \in \mathcal{U}_\kappa} \mathbb{E}_{S \sim P_{\boldsymbol{\mu}}^{\otimes n}} \left[\mathcal{E}_{\boldsymbol{\mu}}(\widehat{h}_S) - \mathcal{E}_{\boldsymbol{\mu}}(\boldsymbol{w}_{\boldsymbol{\mu}}^*) \right] \geq \mathtt{c} (1 - 2p) \min \left\{ \frac{2\widetilde{\eta}}{\lambda_1}, \frac{N(2\widetilde{\eta}) - 1}{n} \right\},$$

where we recall all universal constants:

$$\begin{aligned}
\mathtt{c}_1 &= \frac{\sqrt{2}}{32\pi^{3/2} e} \\
\mathtt{c}_2 &\in (1/2, 1) \\
\mathtt{c} &= \frac{3\mathtt{c}_1 \mathtt{c}_2}{4096}.
\end{aligned}$$

The conclusion of the theorem follows from the fact that $\mathcal{U}_\kappa \subseteq \{\boldsymbol{\mu} \in \mathbb{R}^d : \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}} = 1\}$. □

E Proofs for Fast and Continuously Decaying Spectrum

In this section we provide the proofs of Theorem 3 and Corollary 1.

Theorem 3 (Statistical optimality of early-stopped GD for FCD spectrum)**.** *Define*

$$\eta_n^* := \sup\{\eta > 0 : N(\eta) \geq n\eta\},$$

and assume that (FCD) *holds on* $(0, 2\eta_n^*]$*. Then,* $r_{\mathrm{eff}}(\eta) \asymp N(\eta)$ *for all* $\eta \in (0, \eta_n^*]$*.*

Moreover, if $\eta_n^* \to 0$ *as* $n \to \infty$*, and* $\log\log(n)/n \lesssim \eta_n^*$*, then the upper bound in Theorem 1 with* $\eta_n = \eta_n^*$ *matches the statistical lower bound in Theorem 2 with* $\widetilde{\eta}_n = \eta_n^*/2$*, with common rate* $\asymp \eta_n^*$*.*

Proof. We split the proof into multiple steps:

Step 1: We prove that $r_{\mathrm{eff}}(\eta) \asymp N(\eta)$ for $0 < \eta \leq \eta_n^*$. If $\lambda_i \geq \eta$, then

$$\frac{\lambda_i}{\lambda_i + \eta} \geq \frac{1}{2},$$

and so we can lower bound $r_{\mathrm{eff}}(\eta)$ as

$$r_{\mathrm{eff}}(\eta) = \sum_{i=1}^{d} \frac{\lambda_i}{\lambda_i + \eta} = \sum_{i:\lambda_i < \eta} \frac{\lambda_i}{\lambda_i + \eta} + \sum_{i:\lambda_i \geq \eta} \frac{\lambda_i}{\lambda_i + \eta} \geq 0 + \sum_{i:\lambda_i \geq \eta} \frac{1}{2} = \frac{1}{2} N(\eta),$$

i.e. $r_{\mathrm{eff}}(\eta) \gtrsim N(\eta)$. To upper bound $r_{\mathrm{eff}}(\eta)$, note that

$$r_{\mathrm{eff}}(\eta) = \sum_{i:\lambda_i \geq \eta} \frac{\lambda_i}{\lambda_i + \eta} + \sum_{i:\lambda_i < \eta} \frac{\lambda_i}{\lambda_i + \eta} \leq \sum_{i:\lambda_i \geq \eta} 1 + \frac{1}{\eta} \sum_{i:\lambda_i < \eta} \lambda_i = N(\eta) + \frac{1}{\eta} \sum_{\lambda_i < \eta} \lambda_i.$$

Since, by the (FCD) assumption, $\frac{1}{\eta} \sum_{\lambda_i < \eta} \lambda_i \lesssim N(\eta)$, we arrive at $r_{\mathrm{eff}}(\eta) \lesssim N(\eta)$.

Thus, we have $r_{\mathrm{eff}}(\eta) \asymp N(\eta)$ for all $\eta \in (0, \eta_n^*]$.

Step 2: We now show that for all $0 < \widetilde{\eta} \leq \eta_n^*$,

$$N(\widetilde{\eta}) \lesssim N(2\widetilde{\eta}). \tag{48}$$

Indeed, $N(\widetilde{\eta}) - N(2\widetilde{\eta}) = |\{i : \widetilde{\eta} \leq \lambda_i < 2\widetilde{\eta}\}|$, so that

$$\widetilde{\eta}(N(\widetilde{\eta}) - N(2\widetilde{\eta})) = \widetilde{\eta}|\{i : \widetilde{\eta} \leq \lambda_i < 2\widetilde{\eta}\}| \leq \sum_{\widetilde{\eta} \leq \lambda_i < 2\widetilde{\eta}} \lambda_i \leq \sum_{\lambda_i < 2\widetilde{\eta}} \lambda_i \lesssim \widetilde{\eta} N(2\widetilde{\eta}),$$

where the last inequality uses the (FCD) assumption for $2\widetilde{\eta} \in (0, 2\eta_n^*]$. Thus, (48) follows.

Step 3: We show that $N(\eta_n^*) \asymp n\eta_n^*$, so that $r_{\mathrm{eff}}(\eta_n^*) \asymp n\eta_n^*$ also holds by Step 1.

The definition of $\eta_n^* = \sup\{\eta > 0 : N(\eta) \geq n\eta\}$ implies that $N(2\eta_n^*) < 2n\eta_n^*$. Thus, according to (48),

$$N(\eta_n^*) \lesssim N(2\eta_n^*) \lesssim n\eta_n^*,$$

Since $\eta_n^* = \sup\{\widetilde{\eta} > 0 : N(\widetilde{\eta}) \geq n\widetilde{\eta}\}$, we have $N(\eta_n^*) \geq n\eta_n^*$, and thus $N(\eta_n^*) \asymp n\eta_n^*$.

Step 4: We first evaluate the upper bound in Theorem 1 at $\eta = \eta_n^*$, after which we will check its two conditions that η_n^* must satisfy. Note that, as $r_{\mathrm{eff}}(\eta_n^*) \asymp n\eta_n^*$ and $\log\log(n)/n \lesssim \eta_n^*$ by assumption, this upper bound is

$$\mathcal{E}(\boldsymbol{w}_\tau) - \mathcal{E}(\boldsymbol{w}^*) \lesssim \frac{r_{\mathrm{eff}}(\eta_n^*) + \log\log(n)}{n} + \eta_n^* \asymp \frac{n\eta_n^* + \log\log(n)}{n} + \eta_n^* \lesssim \eta_n^*. \tag{49}$$

Now, note that the first condition $n \gtrsim r_{\mathrm{eff}}(\eta_n)$ holds because $r_{\mathrm{eff}}(\eta_n)/n \asymp \eta_n \to 0$ as $n \to \infty$.

To verify condition (3), note that its left hand side is $\lesssim \eta_n^*$ by the derivation in (49), and its right hand side is independent of n. Since $\eta_n^* \to 0$ as $n \to \infty$ by assumption, the condition (3) is also verified.

Step 5: Finally, we evaluate the statistical lower bound of Theorem 2 at $\widetilde{\eta} = \eta_n^*/2$. For its condition $N(2\eta) > 9$, note that, by the result of Step 3 and the assumption $\log\log(n)/n \lesssim \eta_n^*$,

$$N(2\eta_n^*/2) = N(\eta_n^*) \asymp n\eta_n^* \gtrsim \log\log(n),$$

so that the lower bound condition holds. At $\widetilde{\eta} = \eta_n^*/2$, since $N(\eta_n^*) \asymp n\eta_n^*$, the lower bound is

$$\inf_{\widehat{h}} \sup_{\boldsymbol{\mu}} \mathop{\mathbb{E}}_{S\sim P_{\boldsymbol{\mu}}^{\otimes n}} \left[\mathcal{E}_{\boldsymbol{\mu}}(\widehat{h}_S) - \mathcal{E}_{\boldsymbol{\mu}}(\boldsymbol{w}_{\boldsymbol{\mu}}^*)\right] \gtrsim \min\left\{\frac{\eta_n^*}{2}, \frac{N(\eta_n^*)}{n}\right\} \asymp \eta_n^*,$$

This concludes that the upper and statistical lower bounds of Theorems 1 and 2, respectively, are matching with common rate $\asymp \eta_n^*$.

□

Corollary 1 (Minimax rates for polynomially and exponentially-decaying spectra)**.** *For polynomial decay $\lambda_i \asymp i^{-2a}$ with $a > 1/2$, there exists $\boldsymbol{\mu} \in \mathbb{R}^d$ with $\left\|\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right\| \leq \sqrt{2/\lambda_1}$ such that*

$$\mathcal{E}(\boldsymbol{w}_\tau) - \mathcal{E}(\boldsymbol{w}^*) \asymp \eta_n^* \asymp n^{-\frac{2a}{2a+1}}\ .$$

For exponential decay $\lambda_i \asymp e^{-bi}$ with $b > 0$, there exists $\boldsymbol{\mu} \in \mathbb{R}^d$ with $\left\|\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right\| \leq \sqrt{2/\lambda_1}$ such that

$$\mathcal{E}(\boldsymbol{w}_\tau) - \mathcal{E}(\boldsymbol{w}^*) \asymp \eta_n^* \asymp \frac{\log n}{n}\ .$$

Proof. The proof for both spectra will verify the conditions (FCD), $\eta_n^* \to 0$ as $n \to \infty$, and $\log\log(n)/n \lesssim \eta_n^*$ of Theorem 3, and then apply its conclusion. For this, we will compute in both cases $N(\eta)$ and, as a result, deduce η_n^*. Note that the existence of an appropriate $\boldsymbol{\mu} \in \mathbb{R}^d$ with $\left\|\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right\| \leq \sqrt{2/\lambda_1}$ for which the presented rates hold is due to Theorem 2.

Polynomially decaying spectrum: Suppose $\mathtt{c}_- i^{-2a} \leq \lambda_i \leq \mathtt{c}_+ i^{-2a}$ for some constants $\mathtt{c}_+ \geq \mathtt{c}_- > 0$. Then $i \leq (\mathtt{c}_-/\widetilde{\eta})^{1/(2a)}$ implies that $\lambda_i \geq \widetilde{\eta}$, i.e. $N(\widetilde{\eta}) \gtrsim \widetilde{\eta}^{-1/(2a)}$; and $\lambda_i \geq \widetilde{\eta}$ gives $i \leq (\mathtt{c}_+/\widetilde{\eta})^{1/(2a)}$, i.e. $N(\widetilde{\eta}) \lesssim \widetilde{\eta}^{-1/(2a)}$; combining these two we get

$$N(\widetilde{\eta}) \asymp \widetilde{\eta}^{-\frac{1}{2a}}.$$

Next, if $\lambda_i < \widetilde{\eta}$, then $i > (\mathtt{c}_-/\widetilde{\eta})^{1/(2a)}$, so that

$$\sum_{i:\lambda_i<\widetilde{\eta}} \lambda_i \leq \mathtt{c}_+ \sum_{i:\lambda_i<\widetilde{\eta}} i^{-2a} \leq \mathtt{c}_+ \sum_{i>(c_-/\widetilde{\eta})^{1/(2a)}} i^{-2a} \lesssim \int_{(c_-/\widetilde{\eta})^{1/(2a)}}^{\infty} t^{-2a}\mathrm{d}t \asymp \widetilde{\eta}^{1-\frac{1}{2a}}.$$

Thus,

$$\frac{1}{\widetilde{\eta}} \sum_{i:\lambda_i<\widetilde{\eta}} \lambda_i \lesssim \widetilde{\eta}^{-\frac{1}{2a}} \asymp N(\widetilde{\eta}),$$

and the (FCD) requirement of Theorem 3 is met (for any $\widetilde{\eta} > 0$).

By Step 3 of Theorem 3, $N(\eta_n^*) \asymp n\eta_n^*$, which gives

$$(\eta_n^*)^{-\frac{1}{2a}} \asymp n\eta_n^*, \qquad \text{i.e.} \quad \eta_n^* \asymp n^{-\frac{2a}{2a+1}},$$

which is the common upper and statistical lower bound matching rate. We have $\eta_n^* \to 0$ as $n \to \infty$ and $\log\log(n)/n \lesssim \eta_n^*$.

Exponentially decaying spectrum: Suppose $\mathtt{c}_- e^{-bi} \leq \lambda_i \leq \mathtt{c}_+ e^{-bi}$ for some constants $\mathtt{c}_+ \geq \mathtt{c}_- > 0$. Then, $i \leq \frac{1}{b}\log(\mathtt{c}_-/\widetilde{\eta})$ implies that $\lambda_i \geq \widetilde{\eta}$, i.e. $N(\widetilde{\eta}) \gtrsim \log(1/\widetilde{\eta})$; and $\lambda_i \geq \widetilde{\eta}$ gives $i \leq \frac{1}{b}\log(\mathtt{c}_+/\widetilde{\eta})$, i.e. $N(\widetilde{\eta}) \lesssim \log(1/\widetilde{\eta})$; combining these two we get

$$N(\widetilde{\eta}) \asymp \log\left(\frac{1}{\widetilde{\eta}}\right) \qquad \text{for any} \quad \widetilde{\eta} > 0. \tag{50}$$

Next, if $\lambda_i < \widetilde{\eta}$, then $i > \frac{1}{b}\log(\mathtt{c}_-/\widetilde{\eta})$, so that

$$\sum_{i:\lambda_i<\widetilde{\eta}} \lambda_i \le \mathtt{c}_+ \sum_{i:\lambda_i<\widetilde{\eta}} e^{-bi} \le \mathtt{c}_+ \sum_{i>\frac{1}{b}\log(\mathtt{c}_-/\widetilde{\eta})} e^{-bi} \lesssim \int_{\frac{1}{b}\log(\mathtt{c}_-/\widetilde{\eta})}^{\infty} e^{-bt}\mathrm{d}t \asymp \widetilde{\eta}.$$

Thus,

$$\frac{1}{\widetilde{\eta}} \sum_{i:\lambda_i<\widetilde{\eta}} \lambda_i \lesssim 1 \lesssim \log\left(\frac{1}{\widetilde{\eta}}\right) \asymp N(\widetilde{\eta}),$$

and the (FCD) requirement of Theorem 3 is met (for any $\widetilde{\eta} > 0$).

Now, in order to identify η_n^* in this case, we proceed as follows. First, note that by (50) we have that for any $\widetilde{\eta} > 0$, $\mathtt{c}_1 \log(1/\widetilde{\eta}) \le N(\widetilde{\eta}) \le \mathtt{c}_2 \log(1/\widetilde{\eta})$ for some constants $\mathtt{c}_2 \ge \mathtt{c}_1 > 0$. Consider

$$\eta_n^- = \frac{\mathtt{c}_1}{4}\frac{\log(n)}{n}.$$

Then, for n larger than an appropriate universal constant, since $\mathtt{c}_1 \log(1/\widetilde{\eta}) \le N(\widetilde{\eta})$ for any $\widetilde{\eta} > 0$,

$$N(\eta_n^-) \ge \mathtt{c}_1 \log\left(\frac{n}{(\mathtt{c}_1/4)\log(n)}\right) = \mathtt{c}_1\left(\log(n) - \log\log(n) + \log\left(\frac{4}{\mathtt{c}_1}\right)\right) \ge \frac{\mathtt{c}_1}{2}\log(n) > n\eta_n^-.$$

Thus, by the definition of $\eta_n^* = \sup\{\eta > 0 : N(\eta) \ge n\eta\}$, we have $\eta_n^- \le \eta_n^*$. Similarly, take

$$\eta_n^+ = 2\mathtt{c}_2\frac{\log(n)}{n},$$

Then, for n larger than an appropriate universal constant, since $N(\widetilde{\eta}) \le \mathtt{c}_2 \log(1/\widetilde{\eta})$ for any $\widetilde{\eta} > 0$,

$$N(\eta_n^+) \le \mathtt{c}_2 \log\left(\frac{n}{2\mathtt{c}_2\log(n)}\right) = \mathtt{c}_2\left(\log(n) - \log\log(n) - \log(2\mathtt{c}_2)\right) < 2\mathtt{c}_2\log(n) = n\eta_n^+.$$

Thus, by the definition of $\eta_n^* = \sup\{\eta > 0 : N(\eta) \ge n\eta\}$, we have $\eta_n^* < \eta_n^+$. Thus,

$$\frac{\mathtt{c}_1}{4}\frac{\log(n)}{n} \le \eta_n^* < 2\mathtt{c}_2\frac{\log(n)}{n}, \qquad \text{i.e.} \quad \eta_n^* \asymp \frac{\log(n)}{n},$$

which is the common upper and statistical lower bound matching rate. Clearly we have $\eta_n^* \to 0$ as $n \to \infty$ and $\log\log(n)/n \lesssim \eta_n^*$.

□

F Proof of the Interpolation Lower Bound

In this section, we prove Theorem 4, following a similar sequence of steps to the proof of [40, Theorem 4.2]. We show that, with high probability, every interpolating linear classifier $\boldsymbol{x} \mapsto \mathrm{sign}(\boldsymbol{x}^\top \boldsymbol{w})$ with $\boldsymbol{w} \in \mathcal{I} := \{\boldsymbol{w} \in \mathbb{R}^d : \min_{i\in[n]} Y_i \boldsymbol{X}_i^\top \boldsymbol{w} > 0\}$ incurs an excess zero-one risk bounded below.

Throughout, we use $\mathbb{S}^{m-1} := \{\boldsymbol{x} \in \mathbb{R}^m : \|\boldsymbol{x}\| = 1\}$ to denote the unit sphere in $\mathbb{R}^m$, $(\boldsymbol{a}, \boldsymbol{b}) \in \mathbb{R}^{n_a} \times \mathbb{R}^{n_b}$ to denote the vertical concatenation of $\boldsymbol{a} \in \mathbb{R}^{n_a}$ and $\boldsymbol{b} \in \mathbb{R}^{n_b}$, and ϕ and Φ to denote the PDF and CDF, respectively, of $\mathcal{N}(0, 1)$.

F.1 Setup via Whitening and Rotation

Whitening: We work in a *whitened* covariate space to simplify the presentation of the proofs. We write

$$\boldsymbol{Z} := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X} \in \mathbb{R}^d, \qquad \boldsymbol{\varepsilon}_0 := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon} \in \mathbb{R}^d, \qquad \text{and} \qquad \boldsymbol{v}^* := \boldsymbol{\Sigma}^{1/2}\boldsymbol{w}^* \in \mathbb{R}^d$$

for the whitened covariate, noise, and Bayes linear classifier, respectively. The Gaussian mixture model then takes the form

$$\boldsymbol{Z} = \widetilde{Y}\boldsymbol{v}^* + \boldsymbol{\varepsilon}_0 \quad \text{with} \quad \boldsymbol{\varepsilon}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)\,.$$

Rotation and contrast with [40]: Our interpolation lower bound is proved after *rotating* to a convenient coordinate system. Unlike [40, Theorem 4.2], this does *not* impose any structural assumption (e.g., sparsity) on $\boldsymbol{v}^*$ itself. Indeed, since the whitened Gaussian noise $\boldsymbol{\varepsilon}_0$ is rotationally invariant, for any nonzero $\boldsymbol{v}^* \in \mathbb{R}^d$ and *any* $k \in [d]$, there exists an orthogonal matrix $\boldsymbol{R} \in \mathbb{R}^{d\times d}$ such that

$$\boldsymbol{R}\boldsymbol{v}^* = \|\boldsymbol{v}^*\|\left(\tfrac{1}{\sqrt{k}}\mathbf{1}_k, \mathbf{0}_{d-k}\right).$$

Specifically, a valid choice is $\boldsymbol{R} = \boldsymbol{I}_d - \mathbb{1}\{\boldsymbol{u} \neq \mathbf{0}\}\cdot 2\boldsymbol{u}\boldsymbol{u}^\top/\|\boldsymbol{u}\|^2$ where $\boldsymbol{u} := \boldsymbol{v}^* - \|\boldsymbol{v}^*\|(\tfrac{1}{\sqrt{k}}\mathbf{1}_k, \mathbf{0}_{d-k})$. Thus, in the rotated coordinates, the signal is supported on the first k coordinates, while the remaining $d-k$ coordinates are nuisance directions. Importantly, k and $\boldsymbol{R}$ are free parameters that merely define a choice of basis that influences the behaviour of the lower bound, so this representation applies to *any* $\boldsymbol{v}^*$ rather than strictly sparse ones only. For clarity, we write

$$\boldsymbol{Z}_{\boldsymbol{R}} := \boldsymbol{R}\boldsymbol{Z} = (\boldsymbol{S}, \boldsymbol{N}) \in \mathbb{R}^k \times \mathbb{R}^{d-k} \qquad \text{and} \qquad \boldsymbol{v}^*_{\boldsymbol{R}} := \boldsymbol{R}\boldsymbol{v}^* = (\boldsymbol{\psi}^*, \mathbf{0}_{d-k}) \in \mathbb{R}^k \times \mathbb{R}^{d-k}$$

for the signal-nuisance decomposition. For the i-th sample $(\boldsymbol{X}_i, Y_i)$, we write $\boldsymbol{Z}_{\boldsymbol{R},i}$, $\boldsymbol{S}_i$ and $\boldsymbol{N}_i$ for the corresponding quantities.

Proof sketch: The key proof idea is to show that using only the k signal directions can create large negative margins on a constant fraction of samples (Section F.3), so any interpolating classifier must use the $d-k$ nuisance directions to fix those bad margins (Section F.4). This reliance on nuisance directions ultimately incurs the lower bound in Theorem 4 (Section F.5). To begin, we prove some preliminary lemmas in Section F.2.

F.2 Preliminary Lemmas

Lemma 22 (Bound on norm of signal coordinates)**.** *For every* $\delta \in (0,1)$*,*

$$\mathbb{P}\left(\max_{i\in[n]} \|\boldsymbol{S}_i\| \leq \sqrt{k} + \sqrt{2\log(2n/\delta)} + \|\boldsymbol{v}^*\|\right) \geq 1 - \tfrac{\delta}{2}\,.$$

Proof. Since $\boldsymbol{Z}_{\boldsymbol{R}} = \widetilde{Y}\boldsymbol{v}^*_{\boldsymbol{R}} + \boldsymbol{\varepsilon}_0$ with $\widetilde{Y} \in \{\pm 1\}$, $\boldsymbol{v}^*_{\boldsymbol{R}} = \boldsymbol{R}\boldsymbol{v}^* = (\boldsymbol{\psi}^*, \mathbf{0}_{d-k})$ and $\boldsymbol{\varepsilon}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$, we have

$$\|\boldsymbol{S}\| = \|\boldsymbol{Z}_{\boldsymbol{R}[1:k]}\| = \|\widetilde{Y}\boldsymbol{v}^*_{\boldsymbol{R}[1:k]} + \boldsymbol{\varepsilon}_{0[1:k]}\| \leq \|\boldsymbol{\psi}^*\| + \|\boldsymbol{\varepsilon}_{0[1:k]}\| = \|\boldsymbol{v}^*\| + \|\boldsymbol{\varepsilon}_{0[1:k]}\|,$$

where by $\boldsymbol{Z}_{\boldsymbol{R}[1:k]}$ we denote the vector obtained from the first k coordinates of $\boldsymbol{Z}_{\boldsymbol{R}}$. The result follows from standard Gaussian concentration on $\|\boldsymbol{\varepsilon}_{0[1:k]}\|$:

$$\mathbb{P}\left(\|\boldsymbol{\varepsilon}_{0[1:k]}\| > \sqrt{k} + \sqrt{2\log(2n/\delta)}\right) \leq \tfrac{\delta}{2n},$$

together with a union bound. □

The next lemma is an angular calibration result, analogous to [40, Lemma C.1], which lower bounds the excess zero-one risk in terms of the angle between the learned linear classifier and the Bayes classifier. Working in the whitened space, let $\boldsymbol{v} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{w}$ for $\boldsymbol{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ and recall $\boldsymbol{v}^* = \boldsymbol{\Sigma}^{1/2}\boldsymbol{w}^*$. The Euclidean angle $\angle(\boldsymbol{v}, \boldsymbol{v}^*)$ coincides with the $\boldsymbol{\Sigma}$-angle $\angle_{\boldsymbol{\Sigma}}(\boldsymbol{w}, \boldsymbol{w}^*)$.

Lemma 23 (Angular calibration of excess zero-one risk)**.** *Let* $\theta := \angle(\boldsymbol{v}, \boldsymbol{v}^*) \in [0, \pi]$*. Then,*

$$\mathcal{E}(\boldsymbol{w}) - \mathcal{E}(\boldsymbol{w}^*) \geq \tfrac{1}{\sqrt{2\pi}}(1-2p)(1-\cos\theta)\,\|\boldsymbol{v}^*\|\exp(-\tfrac{1}{2}\,\|\boldsymbol{v}^*\|^2).$$

Proof. To begin, recall that $\mathcal{E}(\boldsymbol{w}) - \mathcal{E}(\boldsymbol{w}^*) = (1-2p)(\mathbb{P}(\widetilde{Y}\boldsymbol{v}^\top\boldsymbol{Z} \leq 0) - \mathbb{P}(\widetilde{Y}(\boldsymbol{v}^*)^\top\boldsymbol{Z} \leq 0))$, where we have

$$\begin{aligned}\mathbb{P}(\widetilde{Y}\boldsymbol{v}^\top\boldsymbol{Z} \leq 0) &= \mathbb{P}(\widetilde{Y}\boldsymbol{v}^\top(\widetilde{Y}\boldsymbol{v}^* + \boldsymbol{\varepsilon}_0) \leq 0) = \mathbb{P}\big(\boldsymbol{v}^\top\boldsymbol{v}^* + \mathcal{N}(\mathbf{0}, \|\boldsymbol{v}\|^2) \leq \mathbf{0}\big)\\ &= \Phi\left(-\frac{\boldsymbol{v}^\top\boldsymbol{v}^*}{\|\boldsymbol{v}\|}\right) = \Phi(-\|\boldsymbol{v}^*\|\cos\theta)\,,\end{aligned}$$

and similarly, $\mathbb{P}(\widetilde{Y}(\boldsymbol{v}^*)^\top \boldsymbol{Z} \le 0) = \Phi\big(-\|\boldsymbol{v}^*\|\big)$. Hence, combining the two yields

$$\mathbb{P}(\widetilde{Y}\boldsymbol{v}^\top \boldsymbol{Z} \le 0) - \mathbb{P}(\widetilde{Y}(\boldsymbol{v}^*)^\top \boldsymbol{Z} \le 0) = \Phi(-\|\boldsymbol{v}^*\|\cos\theta) - \Phi(-\|\boldsymbol{v}^*\|) = \int_{-\|\boldsymbol{v}^*\|}^{-\|\boldsymbol{v}^*\|\cos\theta} \phi(x)\mathrm{d}x$$

$$\ge \int_{-\|\boldsymbol{v}^*\|}^{-\|\boldsymbol{v}^*\|\cos\theta} \min_{-\|\boldsymbol{v}^*\| \le x \le -\|\boldsymbol{v}^*\|\cos\theta} \phi(x)\mathrm{d}x = \big(-\|\boldsymbol{v}^*\|\cos\theta + \|\boldsymbol{v}^*\|\big)\cdot\phi(-\|\boldsymbol{v}^*\|)\,.$$

□

F.3 Signal Component Induces Large Negative Margins on Constant Fraction of Samples

Lemma 24, analogous to [40, Lemma C.2], gives a uniform lower bound on the probability of large negative margins for any fixed classifier $\bar{\boldsymbol{v}} \in \mathbb{S}^{d-1}$, where the bar $\bar{\cdot}$ indicates a unit vector.

Lemma 24 (Bad margin occurs with constant probability)**.** *For any $\bar{\boldsymbol{v}} \in \mathbb{S}^{d-1}$ and every $\mathsf{c} \in (0,1)$,*

$$\mathbb{P}\left(Y\boldsymbol{Z}_{\boldsymbol{R}}^\top \bar{\boldsymbol{v}} < -\mathsf{c}\right) \ge \varrho(\|\boldsymbol{v}^*\|) := 2\left(\Phi(-\mathsf{c}) - \Phi(-1)\right)\left(\frac{1-2p}{1+\exp\left(2\|\boldsymbol{v}^*\|(\|\boldsymbol{v}^*\|+1)\right)} + p\right).$$

Proof. Since $\bar{\boldsymbol{v}}' := \boldsymbol{R}^\top\bar{\boldsymbol{v}} \in \mathbb{S}^{d-1}$ for any $\bar{\boldsymbol{v}} \in \mathbb{S}^{d-1}$, we have $\mathbb{P}(Y\boldsymbol{Z}_{\boldsymbol{R}}^\top\bar{\boldsymbol{v}} < -\mathsf{c}) = \mathbb{P}(Y\boldsymbol{Z}^\top\bar{\boldsymbol{v}}' < -\mathsf{c})$. We bound $\mathbb{P}(Y\boldsymbol{Z}^\top\bar{\boldsymbol{v}}' < -\mathsf{c})$ in the rest of the proof. Recall $\sigma(t) = 1/(1+\exp(-t))$ denotes the sigmoid function. For the Gaussian mixture model $\boldsymbol{Z} = \widetilde{Y}\boldsymbol{v}^* + \boldsymbol{\varepsilon}_0$, we recall from Lemma 5 that

$$\mathbb{P}(Y = y \mid \boldsymbol{Z}) = (1-2p)\sigma(2y\boldsymbol{Z}^\top\boldsymbol{v}^*) + p \qquad \text{for } y \in \{\pm 1\}\,.$$

Using this we can write

$$\begin{aligned}
&\mathbb{P}(Y\boldsymbol{Z}^\top\bar{\boldsymbol{v}}' < -\mathsf{c})\\
=&\mathbb{E}\left[\mathbb{1}\{Y=1, \boldsymbol{Z}^\top\bar{\boldsymbol{v}}' < -\mathsf{c}\} + \mathbb{1}\{Y=-1, \boldsymbol{Z}^\top\bar{\boldsymbol{v}}' > \mathsf{c}\}\right]\\
=&\mathbb{E}\left[\mathbb{P}(Y=1\mid\boldsymbol{Z})\mathbb{1}\{\boldsymbol{Z}^\top\bar{\boldsymbol{v}}' < -\mathsf{c}\} + \mathbb{P}(Y=-1\mid\boldsymbol{Z})\mathbb{1}\{\boldsymbol{Z}^\top\bar{\boldsymbol{v}}' > \mathsf{c}\}\right]\\
=&\mathbb{E}\left[\big((1-2p)\sigma(2\boldsymbol{Z}^\top\boldsymbol{v}^*)+p\big)\,\mathbb{1}\{\boldsymbol{Z}^\top\bar{\boldsymbol{v}}' < -\mathsf{c}\} + \big((1-2p)\sigma(-2\boldsymbol{Z}^\top\boldsymbol{v}^*)+p\big)\,\mathbb{1}\{\boldsymbol{Z}^\top\bar{\boldsymbol{v}}' > \mathsf{c}\}\right]\\
\overset{\text{(i)}}{\ge}&\mathbb{E}\left[\big((1-2p)\sigma(-2|\boldsymbol{Z}^\top\boldsymbol{v}^*|)+p\big)\,\mathbb{1}\{|\boldsymbol{Z}^\top\bar{\boldsymbol{v}}'| > \mathsf{c}\}\right]\\
\overset{\text{(ii)}}{\ge}&\big((1-2p)\sigma(-\kappa)+p\big)\big(1 - \underbrace{\mathbb{P}(|\boldsymbol{Z}^\top\bar{\boldsymbol{v}}'| \le \mathsf{c})}_{\text{Term I}} - \underbrace{\mathbb{P}(|\boldsymbol{Z}^\top\boldsymbol{v}^*| \ge \kappa/2)}_{\text{Term II}}\big) \qquad \text{for any } \kappa \ge 0\,,
\end{aligned} \tag{51}$$

where step (i) holds because $\sigma(-|t|) \le \min\{\sigma(t), \sigma(-t)\}$ for any $t \in \mathbb{R}$, and step (ii) holds for any constant $\kappa \ge 0$. We next upper bound Term I and Term II separately and choose κ accordingly.

To this end, first note that for any vector $\boldsymbol{u} \in \mathbb{R}^d$, $\boldsymbol{Z}^\top\boldsymbol{u} \mid \widetilde{Y} \sim \mathcal{N}\big(\widetilde{Y}\boldsymbol{u}^\top\boldsymbol{v}^*, \|\boldsymbol{u}\|^2\big)$, so we can write

$$\begin{aligned}
&\mathbb{P}\big(|\boldsymbol{Z}^\top\boldsymbol{u}| \le \mathsf{c}\big)\\
&= \mathbb{E}_{\widetilde{Y}}\left[\mathbb{P}\big(|\boldsymbol{Z}^\top\boldsymbol{u}| \le \mathsf{c} \mid \widetilde{Y}\big)\right]\\
&= \mathbb{E}_{\widetilde{Y}}\left[\mathbb{P}\big(-\mathsf{c} \le \boldsymbol{Z}^\top\boldsymbol{u} \le \mathsf{c} \mid \widetilde{Y}\big)\right]\\
&= \mathbb{E}_{\widetilde{Y}}\left[\Phi\left(\frac{\mathsf{c} - \widetilde{Y}\boldsymbol{u}^\top\boldsymbol{v}^*}{\|\boldsymbol{u}\|}\right) - \Phi\left(\frac{-\mathsf{c} - \widetilde{Y}\boldsymbol{u}^\top\boldsymbol{v}^*}{\|\boldsymbol{u}\|}\right)\right]\\
&= \frac{1}{2}\left[\Phi\left(\frac{\mathsf{c} - \boldsymbol{u}^\top\boldsymbol{v}^*}{\|\boldsymbol{u}\|}\right) - \Phi\left(\frac{-\mathsf{c} - \boldsymbol{u}^\top\boldsymbol{v}^*}{\|\boldsymbol{u}\|}\right) + \Phi\left(\frac{\mathsf{c} + \boldsymbol{u}^\top\boldsymbol{v}^*}{\|\boldsymbol{u}\|}\right) - \Phi\left(\frac{-\mathsf{c} + \boldsymbol{u}^\top\boldsymbol{v}^*}{\|\boldsymbol{u}\|}\right)\right]\\
&\overset{\text{(i)}}{=} \Phi\left(\frac{\mathsf{c} - \boldsymbol{u}^\top\boldsymbol{v}^*}{\|\boldsymbol{u}\|}\right) + \Phi\left(\frac{\mathsf{c} + \boldsymbol{u}^\top\boldsymbol{v}^*}{\|\boldsymbol{u}\|}\right) - 1 \overset{\text{(ii)}}{\le} 1 - 2\Phi\left(-\frac{\mathsf{c}}{\|\boldsymbol{u}\|}\right),
\end{aligned}$$

where step (i) is due to $\Phi(t) + \Phi(-t) = 1$ and step (ii) uses the fact that for any $a \ge 0$, $t \mapsto \Phi(a-t) + \Phi(a+t)$ is maximised at $t = 0$ for $t \in \mathbb{R}$. Then taking $\boldsymbol{u} = \bar{\boldsymbol{v}}'$ in the last display and using $\|\bar{\boldsymbol{v}}'\| = 1$ gives

$$\text{Term I} = \mathbb{P}(|\boldsymbol{Z}^\top\bar{\boldsymbol{v}}'| \le \mathsf{c}) \le 1 - 2\Phi(-\mathsf{c})\,. \tag{52}$$

To control Term II, first note that by choosing $\kappa = 2\|\boldsymbol{v}^*\|(\|\boldsymbol{v}^*\|+1)$, we have

$$\text{Term II} = \mathbb{P}\big(|\boldsymbol{Z}^\top \boldsymbol{v}^*| \geq \kappa/2\big) = \mathbb{P}\big(|\boldsymbol{Z}^\top \boldsymbol{v}^*| \geq \|\boldsymbol{v}^*\|(\|\boldsymbol{v}^*\|+1)\big) = \mathbb{P}\left(\frac{|\boldsymbol{Z}^\top \boldsymbol{v}^*|}{\|\boldsymbol{v}^*\|} \geq \|\boldsymbol{v}^*\|+1\right),$$

where $\boldsymbol{Z}^\top \boldsymbol{v}^*/\|\boldsymbol{v}^*\| = \widetilde{Y}\|\boldsymbol{v}^*\| + G$ with $G \sim \mathcal{N}(0,1)$, which by the triangle inequality implies $|\boldsymbol{Z}^\top \boldsymbol{v}^*|/\|\boldsymbol{v}^*\| \leq |\widetilde{Y}|\|\boldsymbol{v}^*\| + |G| = \|\boldsymbol{v}^*\| + |G|$. Therefore, we have

$$\text{Term II} \leq \mathbb{P}(\|\boldsymbol{v}^*\| + |G| \geq \|\boldsymbol{v}^*\|+1) = \mathbb{P}(|G| \geq 1) = 2\Phi(-1)\,. \tag{53}$$

Finally, substituting (52) and (53) into (51) along with $\kappa = 2\|\boldsymbol{v}^*\|(\|\boldsymbol{v}^*\|+1)$, we obtain

$$\mathbb{P}(Y\boldsymbol{Z}^\top \bar{\boldsymbol{v}}' < -\mathsf{c}) \geq \left(\frac{1-2p}{1+\exp\big(2\|\boldsymbol{v}^*\|(\|\boldsymbol{v}^*\|+1)\big)} + p\right) \cdot 2\left(\Phi(-\mathsf{c}) - \Phi(-1)\right),$$

where $2\left(\Phi(-\mathsf{c}) - \Phi(-1)\right) > 0$ given $\mathsf{c} \in (0,1)$.

□

Lemma 24, analogous to [40, Lemma C.2], gives a *population* guarantee for any *fixed* classifier $\bar{\boldsymbol{v}} \in \mathbb{S}^{d-1}$, Lemma 25 below lifts this to a *finite-sample* statement that holds uniformly over *all* signal components that an interpolator might learn. To isolate the role of the signal component, we consider classifiers of the form $\bar{\boldsymbol{v}} = (\bar{\boldsymbol{\psi}}, \mathbf{0}_{d-k}) \in \mathbb{S}^{d-1}$ with $\bar{\boldsymbol{\psi}} \in \mathbb{S}^{k-1}$. Then $\boldsymbol{Z}_{\boldsymbol{R}}^\top \bar{\boldsymbol{v}} = \boldsymbol{S}^\top \bar{\boldsymbol{\psi}}$, implying the margin depends only on the signal component $\bar{\boldsymbol{\psi}}$. We show that, for every candidate signal direction $\bar{\boldsymbol{\psi}} \in \mathbb{S}^{k-1}$, a constant fraction of samples have large negative margins. As before, the bar $\bar{\cdot}$ denotes a unit vector.

Lemma 25 (Bad margins occur on a constant fraction of samples)**.** *Fix $\mathsf{c} \in (0,1)$ and define $\varrho(\|\boldsymbol{v}^*\|) := 2\left(\Phi(-\mathsf{c}) - \Phi(-1)\right)\big\{(1-2p)/\big[1+\exp\big(2\|\boldsymbol{v}^*\|(\|\boldsymbol{v}^*\|+1)\big)\big] + p\big\} > 0$. For every $\delta \in (0,1)$, if for a sufficiently large constant $\mathsf{c}_1 > 0$,*

$$n \geq \frac{\mathsf{c}_1}{\varrho(\|\boldsymbol{v}^*\|)} \cdot \left(k \log\left(\frac{\sqrt{k}+\|\boldsymbol{v}^*\|}{\mathsf{c}}\right) + k\log\log\left(\frac{\mathsf{c}_1 k}{\varrho(\|\boldsymbol{v}^*\|)\delta}\right) + \log\tfrac{2}{\delta}\right), \tag{54}$$

then the following holds:

$$\mathbb{P}\left(\forall \bar{\boldsymbol{\psi}} \in \mathbb{S}^{k-1},\ \sum_{i=1}^n \mathbb{1}\left\{Y_i \boldsymbol{S}_i^\top \bar{\boldsymbol{\psi}} \leq -\tfrac{\mathsf{c}}{2}\right\} \geq \tfrac{n}{2}\varrho(\|\boldsymbol{v}^*\|)\right) \geq 1-\delta\,.$$

Proof. Define the binary random variables

$$\xi_i(\bar{\boldsymbol{\psi}}) := \mathbb{1}\{Y_i \boldsymbol{S}_i^\top \bar{\boldsymbol{\psi}} \leq -\mathsf{c}\}, \quad \text{for } i \in [n]\,.$$

By Lemma 24, $\mathbb{E}[\xi_i(\bar{\boldsymbol{\psi}})] \geq \varrho(\|\boldsymbol{v}^*\|)$. So by the Chernoff bound, we have

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n \xi_i(\bar{\boldsymbol{\psi}}) < \tfrac{n}{2}\varrho(\|\boldsymbol{v}^*\|)\right) &\leq \mathbb{P}\left(\sum_{i=1}^n \xi_i(\bar{\boldsymbol{\psi}}) < \tfrac{n}{2}\mathbb{E}[\xi_1(\bar{\boldsymbol{\psi}})]\right) \leq \exp\left(-\tfrac{n}{8}\mathbb{E}[\xi_1(\bar{\boldsymbol{\psi}})]\right) \\ &\leq \exp\left(-\tfrac{n}{8}\varrho(\|\boldsymbol{v}^*\|)\right).\end{aligned} \tag{55}$$

We now use a covering argument to turn a *pointwise* probability bound which holds for a fixed $\bar{\boldsymbol{v}} = (\bar{\boldsymbol{\psi}}, \mathbf{0}_{d-k})$ into a *uniform* bound over all $\bar{\boldsymbol{\psi}} \in \mathbb{S}^{k-1}$. Since $|\mathbb{S}^{k-1}| = \infty$, we discretise $\mathbb{S}^{k-1}$ with a finite ϵ-ℓ_2-covering $\mathcal{C}$, so that a union bound can be applied. Since $|\mathcal{C}| \leq \left(\frac{3}{\epsilon}\right)^k$, applying a union bound over $\mathcal{C}$ and using (55), we obtain

$$\begin{aligned}\mathbb{P}\left(\exists \bar{\boldsymbol{\psi}}' \in \mathcal{C} : \sum_{i=1}^n \xi_i(\bar{\boldsymbol{\psi}}') < \tfrac{n}{2}\varrho(\|\boldsymbol{v}^*\|)\right) &\leq |\mathcal{C}| \exp\left(-\tfrac{n}{8}\varrho(\|\boldsymbol{v}^*\|)\right) \\ &\leq \exp\left(-\tfrac{n}{8}\varrho(\|\boldsymbol{v}^*\|) + k\log\tfrac{3}{\epsilon}\right) \leq \tfrac{\delta}{2}\,,\end{aligned} \tag{56}$$

where the last inequality holds provided $n \geq 8\varrho(\|\boldsymbol{v}^*\|)^{-1} \cdot \left(k \log \frac{3}{\epsilon} + \log \frac{2}{\delta}\right)$, which is ensured by the sample size n condition (54) for a sufficiently large constant $\mathsf{c}_1 > 0$.

Define $\mathsf{A} := \sqrt{k} + \sqrt{2\log(2n/\delta)} + \|\boldsymbol{v}^*\|$. On the high-probability event $\{\max_{i\in[n]} \|\boldsymbol{S}_i\| \leq \mathsf{A}\}$ from Lemma 22, we choose $\epsilon = \mathsf{c}/(2\mathsf{A})$. Then, for every $\bar{\boldsymbol{\psi}} \in \mathbb{S}^{k-1}$, there exists $\bar{\boldsymbol{\psi}}' \in \mathcal{C}$ such that for all $i \in [n]$,

$$\left|\boldsymbol{S}_i^\top(\bar{\boldsymbol{\psi}} - \bar{\boldsymbol{\psi}}')\right| \leq \|\boldsymbol{S}_i\|\|\bar{\boldsymbol{\psi}} - \bar{\boldsymbol{\psi}}'\| \leq \mathsf{A}\epsilon \leq \tfrac{\mathsf{c}}{2}\,,$$

which implies $-\frac{\mathsf{c}}{2} \leq Y_i\boldsymbol{S}_i^\top\bar{\boldsymbol{\psi}} - Y_i\boldsymbol{S}_i^\top\bar{\boldsymbol{\psi}}' \leq \frac{\mathsf{c}}{2}$, so whenever $Y_i\boldsymbol{S}_i^\top\bar{\boldsymbol{\psi}}' \leq -\mathsf{c}$, we must have $Y_i\boldsymbol{S}^\top\bar{\boldsymbol{\psi}} \leq \frac{\mathsf{c}}{2} - \mathsf{c} = -\frac{\mathsf{c}}{2}$. This means that on the event $\{\max_{i\in[n]} \|\boldsymbol{S}_i\| \leq \mathsf{A}\}$,

$$\sum_{i=1}^n \xi_i(\bar{\boldsymbol{\psi}}') = \sum_{i=1}^n \mathbb{1}\{Y_i\boldsymbol{S}_i^\top\bar{\boldsymbol{\psi}}' \leq -\mathsf{c}\} \leq \sum_{i=1}^n \mathbb{1}\{Y_i\boldsymbol{S}_i^\top\bar{\boldsymbol{\psi}} \leq -\tfrac{\mathsf{c}}{2}\}\,. \tag{57}$$

Therefore, using (57) and (56) we can write

$$\begin{aligned}
&\mathbb{P}\left(\forall\bar{\boldsymbol{\psi}} \in \mathbb{S}^{k-1},\ \sum_{i=1}^n \mathbb{1}\{Y_i\boldsymbol{S}_i^\top\bar{\boldsymbol{\psi}} \leq -\tfrac{\mathsf{c}}{2}\} \geq \tfrac{n}{2}\varrho(\|\boldsymbol{v}^*\|)\right)\\
&= 1 - \mathbb{P}\left(\exists\bar{\boldsymbol{\psi}} \in \mathbb{S}^{k-1} : \sum_{i=1}^n \mathbb{1}\{Y_i\boldsymbol{S}_i^\top\bar{\boldsymbol{\psi}} \leq -\tfrac{\mathsf{c}}{2}\} < \tfrac{n}{2}\varrho(\|\boldsymbol{v}^*\|)\right)\\
&\geq 1 - \mathbb{P}\left(\max_{i\in[n]}\|\boldsymbol{S}_i\| > \mathsf{A}\right) - \mathbb{P}\left(\exists\bar{\boldsymbol{\psi}}' \in \mathcal{C} : \sum_{i=1}^n \xi_i(\bar{\boldsymbol{\psi}}') < \tfrac{n}{2}\varrho(\|\boldsymbol{v}^*\|)\right)\\
&\geq 1 - \tfrac{\delta}{2} - \tfrac{\delta}{2} = 1 - \delta\,.
\end{aligned}$$

□

F.4 Nuisance Component Cannot Fix Large Negative Margins for Affected Samples

Recall the rotated Bayes classifier admits a *signal-nuisance* decomposition $\boldsymbol{v}_{\boldsymbol{R}}^* = (\boldsymbol{\psi}^*, \mathbf{0}_{d-k}) \in \mathbb{R}^d$. Lemma 25 shows that a learned classifier using only its signal component, i.e., $\bar{\boldsymbol{v}} = (\bar{\boldsymbol{\psi}}, \mathbf{0}_{d-k}) \in \mathbb{S}^{d-1}$, produces large negative margins on a constant fraction of samples. To interpolate, a *general* classifier $(\boldsymbol{\psi}, \boldsymbol{\nu}) \in \mathbb{R}^k \times \mathbb{R}^{d-k}$ must therefore use its nuisance component $\boldsymbol{\nu}$ to *correct* these bad margins. Lemma 26 shows that this correction is limited: even the best nuisance direction $\bar{\boldsymbol{\nu}} \in \mathbb{S}^{d-k-1}$ can only achieve margins *no larger* than $\mathfrak{m}$ over the affected subset of samples $\mathcal{J}(\bar{\boldsymbol{\psi}})$.

Lemma 26 (Nuisance component cannot fix bad margins)**.** *For each $\bar{\boldsymbol{\psi}} \in \mathbb{S}^{k-1}$, let $\mathcal{J}(\bar{\boldsymbol{\psi}}) \subset [n]$ be a random subset independent of $(\boldsymbol{N}_i)_{i=1}^n$. Suppose there exists a function $\widetilde{\varrho}(n, \|\boldsymbol{v}^*\|) \geq 0$ such that $|\mathcal{J}(\bar{\boldsymbol{\psi}})| \geq \widetilde{\varrho}(n, \|\boldsymbol{v}^*\|)$ for all $\bar{\boldsymbol{\psi}} \in \mathbb{S}^{k-1}$, then for every $\delta \in (0,1)$,*

$$\mathbb{P}\left(\forall\bar{\boldsymbol{\psi}} \in \mathbb{S}^{k-1},\ \sup_{\bar{\boldsymbol{\nu}}\in\mathbb{S}^{d-k-1}} \min_{i\in\mathcal{J}(\bar{\boldsymbol{\psi}})} Y_i\boldsymbol{N}_i^\top\bar{\boldsymbol{\nu}} \leq \mathfrak{m}\right) \geq 1 - \delta\,.$$

where

$$\mathfrak{m}(n, d, k, \|\boldsymbol{v}^*\|) = \frac{\sqrt{n} + \sqrt{d-k} + \sqrt{2\log(1/\delta)}}{\sqrt{\widetilde{\varrho}(n, \|\boldsymbol{v}^*\|)}}\,.$$

Proof. Recall we have $\boldsymbol{Z}_{\boldsymbol{R},i} = \widetilde{Y}_i\boldsymbol{v}_{\boldsymbol{R}}^* + \boldsymbol{\varepsilon}_{0,i} = (\boldsymbol{S}_i, \boldsymbol{N}_i)$ with $\boldsymbol{\varepsilon}_{0,i} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$, so

$$(\boldsymbol{N}_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d-k})\,.$$

Collect $\boldsymbol{N}_1, \ldots, \boldsymbol{N}_n$ into a matrix $\mathbf{N}_{1:n} \in \mathbb{R}^{n\times(d-k)}$, then $\mathbf{N}_{1:n}$ is a standard Gaussian matrix, with the standard spectral-norm upper bound

$$\mathbb{P}\left(\|\mathbf{N}_{1:n}\| \leq \sqrt{n} + \sqrt{d-k} + \sqrt{2\log(1/\delta)}\right) \geq 1 - \delta\,.$$

In the rest of the proof, we work on this event.

Fix $\bar{\boldsymbol{\psi}} \in \mathbb{S}^{k-1}$ and write $\mathcal{J} \equiv \mathcal{J}(\bar{\boldsymbol{\psi}}) \subset [n]$ for simplicity. Fix $\bar{\boldsymbol{\nu}} \in \mathbb{S}^{d-k-1}$, and stack $(Y_i \boldsymbol{N}_i^\top \bar{\boldsymbol{\nu}})_{i\in\mathcal{J}}$ into a vector $\boldsymbol{A} \in \mathbb{R}^{|\mathcal{J}|}$. If $\min_{i\in\mathcal{J}} \boldsymbol{A}_i \le 0$, then since $\mathfrak{m} \ge 0$, it holds that $\min_{i\in\mathcal{J}} \boldsymbol{A}_i \le \mathfrak{m}$. If $\min_{i\in\mathcal{J}} \boldsymbol{A}_i > 0$, then by Cauchy-Schwarz

$$\|\boldsymbol{A}\| \ge \sqrt{|\mathcal{J}|} \min_{i\in\mathcal{J}} \boldsymbol{A}_i, \tag{58}$$

and note also since $Y_1, \dots, Y_n \in \{\pm 1\}$ does not change Euclidean norms, $\|\boldsymbol{A}\| = \|\mathbf{N}_{\mathcal{J}} \bar{\boldsymbol{\nu}}\|$ where $\mathbf{N}_{\mathcal{J}} \in \mathbb{R}^{|\mathcal{J}|\times(d-k)}$ is the submatrix of $\mathbf{N}_{1:n}$ formed by the rows indexed by $\mathcal{J}$, a subset that does not itself depend on $\mathbf{N}_{1:n}$. Since (58) holds for every $\bar{\boldsymbol{\nu}} \in \mathbb{S}^{d-k-1}$, we have

$$\begin{aligned}\sup_{\bar{\boldsymbol{\nu}}\in\mathbb{S}^{d-k-1}} \min_{i\in\mathcal{J}} \boldsymbol{A}_i \le \sup_{\bar{\boldsymbol{\nu}}\in\mathbb{S}^{d-k-1}} \frac{\|\mathbf{N}_{\mathcal{J}}\bar{\boldsymbol{\nu}}\|}{\sqrt{|\mathcal{J}|}} &\overset{\text{(i)}}{\le} \frac{\|\mathbf{N}_{\mathcal{J}}\|}{\sqrt{|\mathcal{J}|}} \overset{\text{(ii)}}{\le} \frac{\|\mathbf{N}_{1:n}\|}{\sqrt{|\mathcal{J}|}} \\ &\overset{\text{(iii)}}{\le} \frac{\sqrt{n}+\sqrt{d-k}+\sqrt{2\log(1/\delta)}}{\sqrt{\widetilde{\varrho}(n,\|\boldsymbol{v}^*\|)}} =: \mathfrak{m}\,,\end{aligned}$$

where step (i) is due to $\|\bar{\boldsymbol{\nu}}\| = 1$, step (ii) holds because $\|\mathbf{N}_{\mathcal{J}}\| \le \|\mathbf{N}_{1:n}\|$ for any random subset $\mathcal{J}(\bar{\boldsymbol{\psi}}) \subset [n]$ that is itself independent of $\mathbf{N}_{1:n}$, and step (iii) is by assumption $|\mathcal{J}| \ge \widetilde{\varrho}(n, \|\boldsymbol{v}^*\|)$. Hence, the bound in the last display applies simultaneously to all $\mathcal{J}(\bar{\boldsymbol{\psi}}) \subset [n]$, and hence to all $\bar{\boldsymbol{\psi}} \in \mathbb{S}^{k-1}$. □

F.5 Proof of Theorem 4

Combining the lemmas above, we prove the main theorem.

Theorem 4 (Interpolation lower bound)**.** *Suppose $\|\boldsymbol{w}^*\|_{\boldsymbol{\Sigma}} \asymp 1$. Fix any $\mathsf{c} \in (0,1)$ and $\delta \in (0,1)$, and let $k \in \{1,\dots,n\}$ be arbitrary. There exist constants $\mathsf{c}_1, \mathsf{c}_2, \mathsf{c}_3 > 0$ such that if*

$$n \ge \mathsf{c}_1 \cdot \left(k \log \tfrac{\sqrt{k}}{\mathsf{c}} + k \log\log \tfrac{2k}{\delta} + \log \tfrac{4}{\delta}\right) \quad \text{and} \quad d - k \ge \mathsf{c}_2 n\,, \tag{6}$$

then, defining the set of interpolators $\mathcal{I} := \{\boldsymbol{w} \in \mathbb{R}^d : \min_i Y_i \boldsymbol{X}_i^\top \boldsymbol{w} > 0\}$, with probability at least $1-\delta$,

$$\forall \boldsymbol{w} \in \mathcal{I},\ \mathcal{E}(\boldsymbol{w}) - \mathcal{E}(\boldsymbol{w}^*) \ge \frac{\mathsf{c}_3 \cdot (1-2p)n}{\left(\sqrt{n}+\sqrt{d-k}+\sqrt{2\log(2/\delta)}\right)^2}\,. \tag{7}$$

Proof. Fix $\delta \in (0,1)$. The proof follows four steps.

Step 1: We first recall the key consequences of Lemmas 25 and 26. Fix $\mathsf{c} \in (0,1)$. For each $\bar{\boldsymbol{\psi}} \in \mathbb{S}^{k-1}$, define $\mathcal{J}(\bar{\boldsymbol{\psi}}) := \{i \in [n] : Y_i \boldsymbol{S}_i^\top \bar{\boldsymbol{\psi}} \le -\frac{\mathsf{c}}{2}\}$. Under the sample size condition (6) and since $\|\boldsymbol{v}^*\| = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{w}^*\| = \|\boldsymbol{w}^*\|_{\boldsymbol{\Sigma}} \asymp 1$, by Lemma 25 applied with $\delta/2$ instead of δ,

$$\mathbb{P}\Big(\forall \bar{\boldsymbol{\psi}} \in \mathbb{S}^{k-1},\ |\mathcal{J}(\bar{\boldsymbol{\psi}})| \ge \tfrac{n}{2}\varrho(\|\boldsymbol{v}^*\|)\Big) \ge 1 - \delta/2\,, \tag{59}$$

where $\varrho(\|\boldsymbol{v}^*\|)$ is defined as in Lemma 25 and $\varrho(\|\boldsymbol{v}^*\|) \asymp 1$ given $\|\boldsymbol{v}^*\| \asymp 1$. Define

$$\mathfrak{m}(n,d,k,\|\boldsymbol{v}^*\|) = \frac{\sqrt{n}+\sqrt{d-k}+\sqrt{2\log(2/\delta)}}{\sqrt{\frac{n}{2}\varrho(\|\boldsymbol{v}^*\|)}}\,.$$

Let $E_2 := \{\|\mathbf{N}_{1:n}\| \le \sqrt{n}+\sqrt{d-k}+\sqrt{2\log(2/\delta)}\}$, the spectral-norm event in the proof of Lemma 26 with δ replaced with $\delta/2$, so $\mathbb{P}(E_2) \ge 1-\delta/2$. The proof of Lemma 26 shows that on E_2 the bound below holds for every $\bar{\boldsymbol{\psi}}$ with $|\mathcal{J}(\bar{\boldsymbol{\psi}})| \ge \frac{n}{2}\varrho(\|\boldsymbol{v}^*\|)$. Hence, on $E_1 \cap E_2$, where E_1 is the event inside the probability in (59),

$$\forall \bar{\boldsymbol{\psi}} \in \mathbb{S}^{k-1},\quad \sup_{\bar{\boldsymbol{\nu}}\in\mathbb{S}^{d-k-1}} \min_{i\in\mathcal{J}(\bar{\boldsymbol{\psi}})} Y_i \boldsymbol{N}_i^\top \bar{\boldsymbol{\nu}} \le \mathfrak{m}\,. \tag{60}$$

Step 2: Define $\mathcal{I}_0 := \{\bar{\boldsymbol{v}} \in \mathbb{S}^{d-1} : \min_{i\in[n]} Y_i \boldsymbol{Z}_{\boldsymbol{R},i}^\top \bar{\boldsymbol{v}} > 0\}$. The second step of the proof is to establish

$$\mathbb{P}\left(\forall \bar{\boldsymbol{v}} \in \mathcal{I}_0,\ \frac{\bar{\boldsymbol{v}}^\top \boldsymbol{v}_{\boldsymbol{R}}^*}{\|\boldsymbol{v}_{\boldsymbol{R}}^*\|} \le 1/\sqrt{1+\tfrac{\mathtt{c}^2}{4\mathtt{m}^2}}\right) \ge 1-\delta\,. \tag{61}$$

Let E_1 and E_2 denote the events defined above. We work on $E_1 \cap E_2$ with $\mathbb{P}(E_1 \cap E_2) \ge 1-\delta$. Fix any $\bar{\boldsymbol{v}} \in \mathcal{I}_0$ and write its signal-nuisance decomposition as $\bar{\boldsymbol{v}} = (\boldsymbol{\psi}, \boldsymbol{\nu})$ with $\boldsymbol{\psi} \in \mathbb{R}^k$ and $\boldsymbol{\nu} \in \mathbb{R}^{d-k}$. If $\boldsymbol{\psi} = \mathbf{0}$, then $\bar{\boldsymbol{v}}^\top \boldsymbol{v}_{\boldsymbol{R}}^*/\|\boldsymbol{v}_{\boldsymbol{R}}^*\| = 0$ so the claimed upper bound in (61) is trivial. Thus, we focus on $\boldsymbol{\psi} \neq \mathbf{0}$ in the rest of the proof, and denote $\bar{\boldsymbol{\psi}} = \boldsymbol{\psi}/\|\boldsymbol{\psi}\| \in \mathbb{S}^{k-1}$.

On E_1, the classifier $\bar{\boldsymbol{v}} = (\bar{\boldsymbol{\psi}}, \mathbf{0}) \in \mathbb{S}^{d-1}$ incorrectly classifies many samples with large negative margins. Therefore, for a different, general classifier $\bar{\boldsymbol{v}} = (\boldsymbol{\psi}, \boldsymbol{\nu}) \in \mathbb{S}^{d-1}$ to interpolate, we must use $\boldsymbol{\nu}$ to fix the negative margins caused by $\boldsymbol{\psi}$ (or $\bar{\boldsymbol{\psi}}$), that is, we must have for every $i \in \mathcal{J}(\bar{\boldsymbol{\psi}})$,

$$0 < Y_i \boldsymbol{Z}_{\boldsymbol{R},i}^\top \bar{\boldsymbol{v}} = Y_i \boldsymbol{S}_i^\top \boldsymbol{\psi} + Y_i \boldsymbol{N}_i^\top \boldsymbol{\nu} = Y_i \boldsymbol{S}_i^\top \bar{\boldsymbol{\psi}} \|\boldsymbol{\psi}\| + Y_i \boldsymbol{N}_i^\top \boldsymbol{\nu}.$$

On E_1, the last display implies $\frac{\mathtt{c}}{2}\|\boldsymbol{\psi}\| < \min_{i\in\mathcal{J}(\bar{\boldsymbol{\psi}})} Y_i \boldsymbol{N}_i^\top \boldsymbol{\nu}$. If $\boldsymbol{\nu} = \mathbf{0}$, this forces $\boldsymbol{\psi} = \mathbf{0}$, contradicting our assumption that $\boldsymbol{\psi} \neq \mathbf{0}$. So $\boldsymbol{\nu} \neq \mathbf{0}$, and we denote $\bar{\boldsymbol{\nu}} = \boldsymbol{\nu}/\|\boldsymbol{\nu}\| \in \mathbb{S}^{d-k-1}$. Now given E_2 holds, we can write

$$\frac{\mathtt{c}}{2}\frac{\|\boldsymbol{\psi}\|}{\|\boldsymbol{\nu}\|} < \min_{i\in\mathcal{J}(\bar{\boldsymbol{\psi}})} Y_i \boldsymbol{N}_i^\top \bar{\boldsymbol{\nu}} \le \mathtt{m}.$$

Together with $\|\bar{\boldsymbol{v}}\|^2 = \|\boldsymbol{\psi}\|^2 + \|\boldsymbol{\nu}\|^2 = 1$, we can solve for $\|\boldsymbol{\psi}\|$ to obtain $\|\boldsymbol{\psi}\|^2 \le \frac{4\mathtt{m}^2}{\mathtt{c}^2}(1 - \|\boldsymbol{\psi}\|^2) \implies \|\boldsymbol{\psi}\| \le 1/\sqrt{1+\frac{\mathtt{c}^2}{4\mathtt{m}^2}}$. We complete the proof of (61) by noticing

$$\frac{\bar{\boldsymbol{v}}^\top \boldsymbol{v}_{\boldsymbol{R}}^*}{\|\boldsymbol{v}_{\boldsymbol{R}}^*\|} = \boldsymbol{\psi}^\top \frac{\boldsymbol{\psi}^*}{\|\boldsymbol{\psi}^*\|} \le \|\boldsymbol{\psi}\|.$$

In the rest of the proof, we work on the event in (61).

Step 3: The next step is to transform from the rotated-whitened coordinates to the original coordinates. To this end, for each $\bar{\boldsymbol{v}} \in \mathcal{I}_0$ define $\boldsymbol{w} := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}^\top \bar{\boldsymbol{v}}$. Then for every sample i, we have

$$Y_i \boldsymbol{Z}_{\boldsymbol{R},i}^\top \bar{\boldsymbol{v}} = Y_i (\boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i)^\top \bar{\boldsymbol{v}} = Y_i \boldsymbol{X}_i^\top (\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}^\top \bar{\boldsymbol{v}}) = Y_i \boldsymbol{X}_i^\top \boldsymbol{w}\,,$$

which implies that the two interpolator sets $\mathcal{I}_0$ and $\mathcal{I}$ have a one-to-one correspondence (up to scaling) where

$$\bar{\boldsymbol{v}} \in \mathcal{I}_0 \iff \boldsymbol{w} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}^\top \bar{\boldsymbol{v}} \in \mathcal{I}\,.$$

Moreover, note that $\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}} = \|\bar{\boldsymbol{v}}\| = 1$, and so

$$\frac{\bar{\boldsymbol{v}}^\top \boldsymbol{v}_{\boldsymbol{R}}^*}{\|\boldsymbol{v}_{\boldsymbol{R}}^*\|} = \frac{(\boldsymbol{R}\boldsymbol{\Sigma}^{1/2}\boldsymbol{w})^\top (\boldsymbol{R}\boldsymbol{\Sigma}^{1/2}\boldsymbol{w}^*)}{\|\boldsymbol{R}\boldsymbol{\Sigma}^{1/2}\boldsymbol{w}^*\|} = \frac{\boldsymbol{w}^\top \boldsymbol{\Sigma}\boldsymbol{w}^*}{\|\boldsymbol{w}^*\|_{\boldsymbol{\Sigma}}} = \frac{\boldsymbol{w}^\top \boldsymbol{\Sigma}\boldsymbol{w}^*}{\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}}\|\boldsymbol{w}^*\|_{\boldsymbol{\Sigma}}} = \cos\theta\,,$$

where $\theta := \angle_{\boldsymbol{\Sigma}}(\boldsymbol{w}, \boldsymbol{w}^*)$. This implies that the following events are equivalent

$$\left\{\forall \bar{\boldsymbol{v}} \in \mathcal{I}_0,\ \frac{\bar{\boldsymbol{v}}^\top \boldsymbol{v}_{\boldsymbol{R}}^*}{\|\boldsymbol{v}_{\boldsymbol{R}}^*\|} \le 1/\sqrt{1+\tfrac{\mathtt{c}^2}{4\mathtt{m}^2}}\right\} = \left\{\forall \boldsymbol{w} \in \mathcal{I},\ \cos\theta \le 1/\sqrt{1+\tfrac{\mathtt{c}^2}{4\mathtt{m}^2}}\right\}.$$

Step 4: The final step is to apply the calibration result from Lemma 23, which lower bounds the excess zero-one risk in terms of $1-\cos\theta$. First, note that if $d-k \ge \mathtt{c}_2 n$ for a sufficiently large $\mathtt{c}_2 > 0$, then $\mathtt{m} = \frac{\sqrt{n}+\sqrt{d-k}+\sqrt{2\log(2/\delta)}}{\sqrt{\frac{n}{2}\varrho(\|\boldsymbol{v}^*\|)}} \ge \frac{1+\sqrt{\mathtt{c}_2}+\sqrt{\frac{2}{n}\log(2/\delta)}}{\sqrt{\frac{1}{2}\varrho(\|\boldsymbol{v}^*\|)}} \ge 3$. Using the inequality $1 - 1/\sqrt{1+\frac{1}{a}} \ge \frac{3}{8a}$ for all $a \ge 3$, we obtain that

$$1-\cos\theta \ge 1 - 1/\sqrt{1+\tfrac{\mathtt{c}^2}{4\mathtt{m}^2}} \ge \tfrac{3}{32}\tfrac{\mathtt{c}^2}{\mathtt{m}^2} = \tfrac{3\mathtt{c}^2}{64} \cdot \frac{n\varrho(\|\boldsymbol{v}^*\|)}{\left(\sqrt{n}+\sqrt{d-k}+\sqrt{2\log(2/\delta)}\right)^2}\,.$$

Applying Lemma 23, we have

$$\begin{aligned}\mathcal{E}(\boldsymbol{w}) - \mathcal{E}(\boldsymbol{w}^*) &\geq \tfrac{1}{\sqrt{2\pi}}(1-2p)(1-\cos\theta)\left\|\boldsymbol{v}^*\right\| \exp(-\tfrac{1}{2}\left\|\boldsymbol{v}^*\right\|^2) \\ &\geq \tfrac{3\mathsf{c}^2}{64\sqrt{2\pi}}(1-2p)\frac{n\varrho(\|\boldsymbol{v}^*\|)}{\left(\sqrt{n}+\sqrt{d-k}+\sqrt{2\log(2/\delta)}\right)^2}\left\|\boldsymbol{v}^*\right\| \exp(-\tfrac{1}{2}\left\|\boldsymbol{v}^*\right\|^2).\end{aligned}$$

Since $\|\boldsymbol{v}^*\| = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{w}^*\| = \|\boldsymbol{w}^*\|_{\boldsymbol{\Sigma}} \asymp 1$ and so $\varrho(\|\boldsymbol{v}^*\|) \asymp 1$, we obtain that there exists a constant $\mathsf{c}_3 > 0$ such that

$$\mathcal{E}(\boldsymbol{w}) - \mathcal{E}(\boldsymbol{w}^*) \geq \mathsf{c}_3 \cdot (1-2p) \cdot \frac{n}{\left(\sqrt{n}+\sqrt{d-k}+\sqrt{2\log(2/\delta)}\right)^2}.$$

□

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Our results are outlined in Section 1.1 and presented formally in Sections 3 and 4.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Limitations are discussed in Section 5.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: Each theorem in the Main is fully formal and has corresponding proofs in the Appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Our experiments are mainly used to validate empirically the studied phenomena and the rates we prove theoretically. The captions of Figures 1 and 2 contain all relevant details.

   Guidelines:

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

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

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

Answer: [N/A]

Justification: Our experiments are mainly used to validate empirically the studied phenomena and the rates we prove theoretically. The captions of Figures 1 and 2 contain all relevant details.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: Our experiments are mainly used to validate empirically the studied phenomena and the rates we prove theoretically. The captions of Figures 1 and 2 contain all relevant details.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: Our experiments are mainly used to validate empirically the studied phenomena and the rates we prove theoretically. The captions of Figures 1 and 2 contain all relevant details.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [N/A]

   Justification: Our experiments are mainly used to validate empirically the studied phenomena and the rates we prove theoretically.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Our work is theoretical and we reviewed the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: Our work is theoretical.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: Our work is theoretical.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

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