Detection and Classification in Latent Spaces: High-Dimensional Analysis and Validation

Enis Chenchene* **Anil S. Baslamisli*** **Paolo Braca***

*NATO STO CMRE

Abstract

Foundation models produce high-dimensional latent representations in which suitably designed statistical decision rules can be effective yet analytically tractable. We study such detectors in the high-dimensional regime, where latent dimension and training sample size grow proportionally. For linear and quadratic statistics, we derive explicit large-deviation characterizations of false-alarm and miss-detection performance under the Neyman–Pearson paradigm, and show how the binary analysis extends to multi-class classification. The resulting error probabilities and corresponding decay rates exhibit non-monotonic dependence on the shape parameter, including double-descent and, in some regimes, multiple-descent. Experiments on vision, text, and audio datasets using modern foundation models show close agreement between theoretical predictions and empirical performance. In particular, the proposed framework allows us to set a prescribed false-alarm level and accurately predict the corresponding miss-detection probability. Overall, our results provide a theoretically grounded and interpretable framework with competitive empirical performance across different modalities.

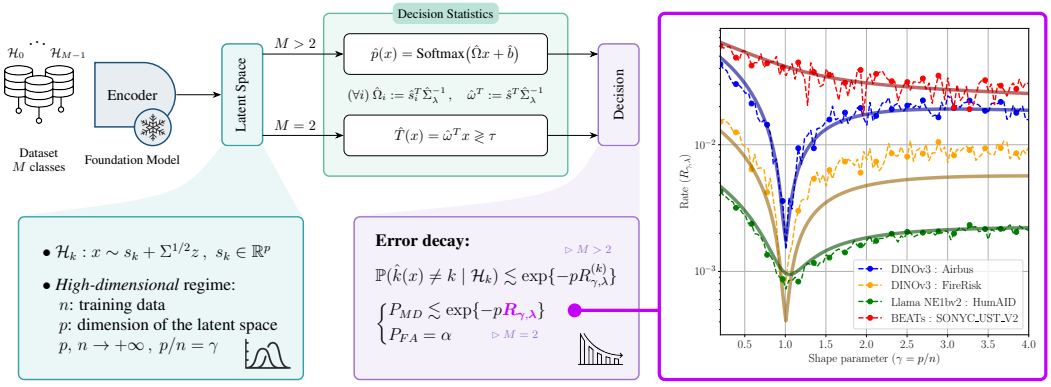


Figure 1: Leveraging foundation models as abstract feature extractors to design theoretically grounded data-driven decision statistics. **Left:** Processing pipeline for the multi-class ($M > 2$) and binary ($M = 2$) classification problems. The latent features x are processed via a linear combination layer, whose weights are obtained from the regularized sample covariance matrix $\hat{\Sigma}_\lambda := \hat{\Sigma} + \lambda I$ and the sample mean of $\mathcal{H}_i$ training data $\hat{s}_i$. The decision is taken as maximum a posteriori for $M > 2$ or with a threshold comparison for $M = 2$. **Right:** Error probability decay rate as a function of the shape parameter for $M = 2$. Comparison between empirical (dashed) and theoretical curves (solid lines) across different foundation models and datasets. We are able to predict and control the performance in terms of error probabilities when both p and n are large enough.

*Emails: {enis.chenchene,anil.baslamisli,paolo.braca}@cmre.nato.int

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

1 Introduction

Modern Machine Learning (ML) systems are increasingly driven by scale, with model size, data volume, and compute growing rapidly to improve performance, raising fundamental questions about their statistical, economic, and energy efficiency [27, 12]. Counterintuitively, large models are more sample-efficient than smaller ones, reaching the same level of performance with fewer optimization steps [27]. This points to the central role of the model-size/sample-size ratio, a key parameter governing the emergence of nontrivial performance transitions [7, 9, 8, 25, 36].

Already in the late 1980s [48, 37] (see [32] for a historical perspective) it was pointed out, both theoretically and empirically, that in several settings including random forests, neural networks and kernel methods [9, 8, 34, 25, 2], the generalization error as a function of the *shape parameter* $\gamma := p/n$, where p is the underlying model size and n is the number of data, exhibits a fascinating trend. The classical U-shaped risk curve, which initially decreases and then increases as the interpolation threshold $\gamma = 1$ is approached from below, decreases once again beyond the interpolation regime, i.e., for $\gamma > 1$—a phenomenon that is nowadays referred to as *double-descent* [7].

Over the past five years several comprehensive mathematical models of double-descent have been developed, with notable contributions by Tibshirani, Hastie, Montanari, Bach, Candès, and their collaborators [46, 25, 2, 31], providing i) a detailed characterization of this intriguing phenomenon for linear regression, an important setting due to its connections to neural networks, cf. [25, Section 1.2], and ii) the demonstration of the fundamental role played by *random matrix theory* (RMT) and the tools therein [4, 47]. While double-descent is now well understood for regression problems, the impact of high dimensionality on detection and classification error probabilities remains largely unexplored. This motivates our first research question:

Do detection methods exhibit a double-descent phenomenon?

We answer this question in the affirmative by showing that, for two standard decision statistics, the rate governing the decay of error probabilities—a natural performance criterion in this context, dating back to the seminal work of Chernoff [15] in the 1950s—exhibits double-descent and even multiple-descent behavior in the high-dimensional regime. For binary classification, our results complement an extensive body of research on high-dimensional logistic regression [46, 41, 18].

To corroborate our theoretical findings, we conduct systematic comparisons on synthetic data and, more importantly, validate our approach on multiple real-world datasets. Since real data rarely satisfies the assumptions of our statistical setting, the proposed detection methods cannot be applied directly in the original data space. Instead, to extend our statistical framework to real data, we first map the observations into a latent space using encoders from *foundation* models such as DINOv3 [45], Llama-Nemotron-Embed-1B-v2 [35], and BEATs [14]. We then consider the resulting bottleneck representations, which are naturally high-dimensional, and apply the decision statistic in this abstract feature space, leading to the data-driven detection (or classification) pipeline depicted in Figure 1. This leads to our second research question:

Can high-dimensional asymptotic analysis predict the performance of Neyman–Pearson decision statistics when applied in the latent space of foundation models?

Our focus on foundation models is motivated by recent insights from Balestriero and LeCun [5]. They show that, to minimize prediction risk uniformly across broad families of downstream tasks, the optimal embedding distribution produced by foundation models should be isotropic Gaussian, thereby offering an ideal testing ground for our statistical framework. We demonstrate that our approach yields results in close agreement with our theoretical predictions, both in terms of moment estimation and in overall detection performance, as shown in Figures 1 and 3a. To the best of our knowledge, this is the first time that high-dimensional analysis of data-driven methods is shown to closely predict performances with real-world data. Overall, our approach establishes an explainable and theoretically grounded methodology that achieves competitive empirical performance in terms of classification accuracy relative to standard baselines, such as logistic regression and linear SVM.

1.1 Problem formulation

Binary case We begin with a binary hypothesis testing problem. Given $x \in \mathbb{R}^p$, we test whether it is sampled from a distribution with zero mean (null hypothesis) or from a distribution with mean $s \in \mathbb{R}^p$. We assume that both distributions share the same non-singular covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ and are given by:

$$\mathcal{H}_0 : \; x \sim \Sigma^{1/2} z \,, \quad \text{and} \quad \mathcal{H}_1 : \; x \sim s + \Sigma^{1/2} z \,, \tag{1}$$

where z is a vector of independent and identically distributed (i.i.d.) sub-Gaussian variables with zero mean and unit variance, which for simplicity we shall assume to be Gaussian[1]. If both s and Σ are known, the log-likelihood ratio (LLR) defines the optimal detector in the sense of Neyman–Pearson, which in this setting yields $T^\dagger(x) := s^T\Sigma^{-1}x = \omega^T x$ [30, 28]. As soon as Σ and s are unknown, we resort to a strategy to estimate them. There are several possible ways to address this task, one of them being to estimate ω from a training set of size n, yielding the *data-driven decision statistic* $\hat{T}$, defined for a *regularization parameter* $\lambda > 0$ and all $x \in \mathbb{R}^p$ by:

$$\hat{T}(x) := \hat{s}^T(\hat{\Sigma} + \lambda I)^{-1}x \,, \tag{2}$$

$$\hat{T}(x) := x^T(\hat{\Sigma} + \lambda I)^{-1}x \text{ if no } \mathcal{H}_1 \text{ data is available in the training set,} \tag{3}$$

where $\hat{\Sigma} := \frac{1}{n}\hat{X}\hat{X}^T$ is the sample covariance matrix of the centered training set (stored in a matrix $\hat{X} \in \mathbb{R}^{p\times n}$), $\hat{s}$ is the sample mean of $\mathcal{H}_1$ training data and I is the identity matrix. The statistics reject the null hypothesis if $\hat{T}(x) > \tau$, for a predefined threshold $\tau \in \mathbb{R}$ and accept it otherwise.

Large deviations performance in high-dimensions In contrast to popular approaches like logistic regression, in the Neyman–Pearson setup, we aim to minimize the probability of miss detection subject to an upper bound on the probability of false alarm, which is a crucial property in safety-critical fields like healthcare or defense. As a performance criterion, we consider the rate in the sense of the *large deviation principle* (LDP) [17]. Formally, it characterizes the error probability decay when, e.g., the number of independent samples observed to take a decision diverges—a viewpoint increasingly used in recent analyses of data-driven methods [11, 24]. In the proposed work, we go beyond these fixed-dimensional settings and study the *high-dimensional regime*, i.e., we consider both the number of training samples n and model size p to grow to $+\infty$ but keeping a constant ratio:

$$n, p \to +\infty \,, \quad \text{and} \quad \gamma = p/n \geq 0 \,. \tag{4}$$

For both the decision statistics (2) and (3), we show that, in the high-dimensional regime (4), the decision statistics $\hat{T}(x)$ satisfy an LDP: For $\tau_p \in \mathbb{R}$ such that $P_{FA} := \mathbb{P}\big(\hat{T}(x) > \tau_p \mid \hat{X}, \mathcal{H}_0\big) = \alpha$ (typically small), we show that[2]

$$\lim_{p\to+\infty} \frac{1}{p} \log P_{MD} := \lim_{p\to+\infty} \frac{1}{p} \log \mathbb{P}\big(\hat{T}(x) \leq \tau_p \mid \hat{X}, \mathcal{H}_1\big) = -R_{\gamma,\lambda} \,, \tag{5}$$

for $R_{\gamma,\lambda} > 0$ named *rate*, where P_{FA} is the probability of false alarm and P_{MD} is the probability of miss detection. We also show analytically that the rate exhibits a double-descent and even a *multiple*-descent behavior in some cases, thereby giving a positive answer to our first research question. Characterizing $R_{\gamma,\lambda}$ constitutes our main technical challenge. For (2), by linearity, the rate can be deduced from the moments of $\hat{T}$. In contrast, for (3), we rely on more delicate arguments based on the Gärtner–Ellis Theorem [22, 21], and nontrivial RMT and convex analysis tools [6].

Multi-hypotheses setting The strength of our analysis for the binary case is that it immediately generalizes to a shift-in-mean *multi-hypothesis* testing setting, where we aim to discriminate among $M \geq 3$ competing distributions:

$$\mathcal{H}_i : \quad x \sim s_i + \Sigma^{1/2} z \,, \qquad s_i \in \mathbb{R}^p \,, \quad i = 0, \ldots, M-1 \,. \tag{6}$$

In this setting, the Maximum A Posteriori (MAP) rule selects the hypothesis with the largest posterior probability, which exhibits the standard linear–softmax form: $p(x) = \mathrm{Softmax}\big(\Omega x + b\big)$, where

[1] This assumption can be relaxed at several points in the analysis; we indicate explicitly whenever this is the case.

[2] To simplify notation, throughout the paper we omit the explicit dependence on p and n, whenever this dependence is clear from the context. Furthermore, we write $\lim_{p\to+\infty} a_p$ as shorthand for the high-dimensional limit $\lim_{\substack{p,n\to+\infty \\ p/n=\gamma}} a_p$.

$\Omega_i := s_i^T \Sigma^{-1}$, $b_i := \log \pi_i - \frac{1}{2} s_i^T \Sigma^{-1} s_i$, and Ω_i is the i^{th} row of Ω and π_i is the prior of $\mathcal{H}_i$, see [10, Chapter 4]. The corresponding data-driven rule picks $\hat{k}(x) := \operatorname{argmax}_{0 \le i \le M-1} \hat{p}_i(x)$, where

$$\hat{p}(x) := \operatorname{Softmax}\left(\hat{\Omega} x + \hat{b}\right), \quad \text{and} \quad \hat{\Omega}_i := \hat{s}_i^T (\hat{\Sigma} + \lambda I)^{-1}, \; \hat{b}_i := \log \pi_i - \tfrac{1}{2} \hat{s}_i^T (\hat{\Sigma} + \lambda I)^{-1} \hat{s}_i \,, \tag{7}$$

where $\hat{s}_i$ is the sample mean of $\mathcal{H}_i$ data, for all $i = 0, \dots, M-1$. As we show in detail in Appendix B.6, all error probabilities exhibit an exponential decay of the form:

$$\limsup_{p \to +\infty} \frac{1}{p} \log \mathbb{P}(\hat{k}(x) \neq k \mid \hat{X}, \mathcal{H}_k) \le -R^{(k)}_{\gamma,\lambda}, \quad \text{for all } k = 0, \dots, M-1, \tag{8}$$

where $R^{(k)}_{\gamma,\lambda}$ can be obtained by studying the rate of the binary classifiers defined in (2) between $\mathcal{H}_k$ and all competing hypotheses. Motivated by this connection, we henceforth restrict our attention to the binary case, which nonetheless provides a rich and representative setting.

1.2 Contributions

We summarize our main contributions as follows:

1. **High-dimensional LDP analysis:** We derive closed-form expressions of the rate in the sense of the LDP for $\hat{T}$ defined as in (2) and (3) in the high-dimensional regime (4), see Theorem 2.2 and Theorem 3.2. We show that these exhibit double-descent, and even multiple-descent behaviors in some cases. These formulas can be computed explicitly in isotropic regimes, and adapted swiftly to several mismatched scenarios, modeling a variety of realistic settings, where overparameterization might yield the best performance.
2. **Detection in foundation models' latent space:** We demonstrate the practical utility of our decision statistics in real-world detection tasks using the pipeline shown in Figure 1 ($M = 2$). We show that these modern architectures naturally lie within our high-dimensional regime, making our theoretical analysis fundamental for predicting empirical performance. Finally, we validate our framework across multiple data modalities (vision, text, and audio) and foundation models, confirming the *robustness* and *universality* of our approach.

1.3 Assumptions and definitions

The notation used throughout this paper follows standard conventions and is summarized in Appendix A.1. Throughout the paper we consider the following assumptions, which are in line with [2].

Assumption 1. *The following holds:*

A1) As $p \to +\infty$, the spectral measure of Σ, denoted by H_p, converges to a probability distribution H on $\mathbb{R}_+$, the so-called population spectral distribution. *Moreover, H has compact support in $\mathbb{R}_{++}$ and $aI \preceq \Sigma \preceq bI$, for $0 < a \le b$.*

A2) For every signal sequence $\theta^* \in \mathbb{R}^p$, $\|\theta^*\|^2 = \mathcal{O}(p)$ as $p \to +\infty$, and there exists some $\nu_{\theta^*} \in \mathcal{M}(\mathbb{R}_+)$, such that $\frac{1}{p} \sum_{i=1}^p (v_i^T \theta^*)^2 \delta_{\xi_i} \rightharpoonup \nu_{\theta^*}$ as $p \to +\infty$, where v_i is i^{th} eigenvector of Σ with unit norm associated to the eigenvalue ξ_i. If $\theta^* = s$ in (1)*, then $\nu_s \ll H$.*

The absolute continuity of ν_s with respect to H in Assumption A2 is not really needed in our analysis, but is imposed to ease interpretation, as it allows us to define:

Definition 1.1 (Signal-to-noise ratio)**.** *Suppose Assumption 1 holds. We say that the* asymptotic signal-to-noise ratio *and the* asymptotic directional signal-to-noise ratio *are, respectively,*

$$\mathsf{SNR} := \lim_{p \to +\infty} \frac{1}{p} s^T \Sigma^{-1} s \,, \quad \textit{and} \quad \mathsf{dSNR}(\xi) := \frac{1}{\xi} \frac{\partial \nu_s}{\partial H}(\xi) \,, \tag{9}$$

where $\frac{\partial \nu_s}{\partial H}$ is the Radon–Nikodym derivative of ν_s over H.

Formally, the asymptotic directional SNR, from now on referred to as dSNR, encodes the limiting behavior of the function that to each eigenvalue ξ_i of Σ returns $(v_i^T s)^2 / \xi_i$, which represents the power of the signal s in the direction of the eigenvector v_i associated to the eigenvalue ξ_i, weighted by the latter, hence its name. Further, it allows us to define the following *observables*.

Definition 1.2 (Observables)**.** *Under Assumption 1, let for $q \in \mathbb{N}$ and all $\kappa \geq 0$*

$$\mathfrak{m}_q(\kappa) := \int_{\mathbb{R}_+} \left[\frac{\xi}{\xi+\kappa}\right]^q \mathrm{d}H(\xi) \quad \textit{and} \quad \mathfrak{n}_q(\kappa) := \int_{\mathbb{R}_+} \left[\frac{\xi}{\xi+\kappa}\right]^q \mathsf{dSNR}(\xi)\,\mathrm{d}H(\xi)\,. \tag{10}$$

These notions, which arise naturally in our analysis as limits of trace functionals, cf. Fact B.1, admit clear formulations in the isotropic and ridgeless regimes, see Examples B.2 and B.3.

Random matrix theory The classical result from the sixties by Marchenko and Pastur [33], and later refined by Bai and Silverstein [4] states that, under Assumption 1—but not necessarily the Gaussianity assumption—in the high-dimensional regime of (4), the companion Stieltjes transform of the spectral measure of $\hat{\Sigma}$, denoted from now on $\hat{H}$, almost surely converges point-wisely to a certain φ defined as the unique solution to the *Silverstein* equation:

$$\frac{1}{\varphi(z)} + z = \gamma \int_0^{+\infty} \frac{\xi}{1+\xi\varphi(z)}\,\mathrm{d}H(\xi)\,. \tag{11}$$

See Theorem A.1, due to [40, 2]. We will need to compute in particular, for $\lambda > 0$,

$$\kappa(\lambda) := \varphi(-\lambda)^{-1}\,, \quad \text{satisfying} \quad \kappa(\lambda) = \lambda + \gamma\kappa(\lambda)\,\mathfrak{m}_1(\kappa(\lambda))\,, \tag{12}$$

which can be efficiently solved by fixed-point iterations. We refer the reader to Appendix A.2 for a more complete account on RMT.

2 Linear decision statistics

Since $\hat{T}$ is linear, it inherits the Gaussianity of x, yielding $\hat{T}(x) \sim \mathcal{N}(0, \hat{\sigma}^2)$ under $\mathcal{H}_0$, and $\hat{T}(x) \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ under $\mathcal{H}_1$, where $\hat{\mu} := \mathbb{E}[\hat{T}(x) \mid \hat{X}, \mathcal{H}_1]$, and $\hat{\sigma}^2 := \mathrm{Var}[\hat{T}(x) \mid \hat{X}, \mathcal{H}_1]$. Being $\hat{T}$ Gaussian, the probability of miss detection with fixed probability of false alarm vanishes exponentially with p, with a rate depending on the *deflection* $\hat{d} := \hat{\mu}/\hat{\sigma}$, see Example A.2:

$$\lim_{p\to+\infty} \frac{1}{p} \log P_{MD} = -R_{\gamma,\lambda}\,, \quad R_{\gamma,\lambda} := \lim_{p\to+\infty} \hat{d}^2/(2p)\,, \tag{13}$$

whenever the above almost sure limits exist. We refer to Appendix A.3 for a general account of LDPs for non-Gaussian statistics, and to Appendix B.1 for the calibration of error probabilities in this setting. Clearly, the same result holds switching the role of P_{FA} and P_{MD}. Observe in particular that $R_{\gamma,\lambda}$ can be obtained from the asymptotic behaviors of $\hat{\mu}$ and $\hat{\sigma}^2$.

Proposition 2.1. *Suppose that Assumption 1 holds, let $\kappa(\lambda)$ be defined as in* (12)*, and $\mathfrak{n}_q$, $\mathfrak{m}_q$, $q = 1, 2$ be the observables defined in Definition 1.2. Then, as $p, n \to +\infty$ with $p/n = \gamma$, $\hat{\mu}/p$ and $\hat{\sigma}^2/p$ converge almost surely to μ_∞ and σ_∞^2 defined by*

$$\mu_\infty := \frac{\kappa(\lambda)}{\lambda}\,\mathfrak{n}_1(\kappa(\lambda))\,, \quad \textit{and} \quad \sigma_\infty^2 := \frac{\kappa(\lambda)^2}{\lambda^2} \frac{\mathfrak{n}_2(\kappa(\lambda))}{\big(1-\gamma\,\mathfrak{m}_2(\kappa(\lambda))\big)}\,. \tag{14}$$

A similar result, but in a different statistical setting, has been shown by Dobriban and Wager in [19].

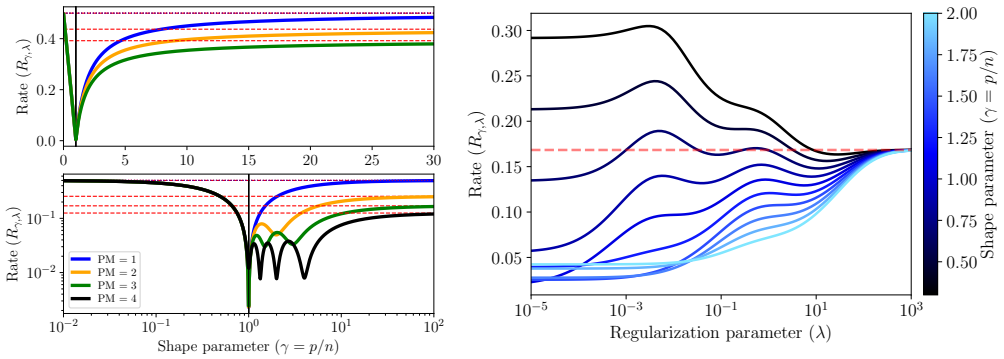
While the Gaussian assumption in (1) is not required for Proposition 2.1, cf. Remark B.5, we retain it to simplify the derivation of the LDP rate (see Appendix A.3):

Theorem 2.2. *Suppose that Assumption 1 holds, let $\kappa(\lambda)$ be defined as in* (12)*, and $\mathfrak{n}_q$, $\mathfrak{m}_q$, $q = 1, 2$ be the observables defined in Definition 1.2. Let $\hat{T}$ be defined as in* (2)*. Then,* (13) *holds with*

$$R_{\gamma,\lambda} := \frac{1}{2}\frac{\mathfrak{n}_1(\kappa(\lambda))^2}{\mathfrak{n}_2(\kappa(\lambda))}\Big(1-\gamma\,\mathfrak{m}_2(\kappa(\lambda))\Big)\,. \tag{15}$$

We show the dependence of $R_{\gamma,\lambda}$ as a function of γ and λ in Figure 2. In particular, from Figure 2a we observe that after the critical threshold $\gamma = 1$, the performances start improving again, reflecting the *double-descent*[3] phenomenon, and giving a positive answer to our first research question. Additionally,

[3]Strictly speaking, we should refer to this as *double-ascent* in this context (the higher the rate, the better for the performances). More precisely, in Figure 2a we do not strictly observe a *double*-ascent, but a single ascent after the critical threshold $\gamma = 1$. This is expected and indeed occurs even for linear regression models [25]. We can recreate the popular U-shaped curves in mismatched cases, as we discuss in Appendix B.5.



(a) Performances as a function of γ.

(b) Performances as a function of λ.

Figure 2: Rate $R_{\gamma,\lambda}$ according to (15) for the decision statistic defined in (2). The asymptotic values for $\gamma \to 0^+$ and $\gamma \to +\infty$ according to (59) are represented with dotted blue and dashed red lines, respectively. **Left:** In the first row, we compare the rate in one isotropic setting with the rates in two non-isotropic ones with dSNR constant equal to 1 and $\lambda = 10^{-5}$. In the second row, we compare four settings with discrete H with $\mathrm{PM} \in \mathbb{N}$ point masses spaced uniformly in logarithmic scale and again dSNR constant equal to 1 and $\lambda = 10^{-8}$, producing a *multiple-descent* behavior. **Right:** Rate $R_{\gamma,\lambda}$ as a function of λ for 10 different values of γ. The precise definitions of the corresponding population spectral distributions are detailed in Section B, and in Figure 6.

the double-descent phenomenon persists in the multi-class setting, cf. Figure 7. More surprisingly, we also observe an interesting multiple-descent behavior, which is emphasized when the limiting spectral distribution H has support with point masses spaced uniformly in logarithmic scale. In Figure 2b, we display the rate as a function of λ for 10 different values of γ ranging from 0.5 to 2 and spaced uniformly. We observe that if $\gamma > 0$ and is small enough, there are local maxima suggesting an optimal tuning of the regularization parameter λ. See also Figures 4a and 5 for further illustrations of the sensitivity of the rate to λ.

Insights and properties In Appendix B.2 and B.3 we investigate two notable settings in which computations can be performed in closed form—the case of *isotropic* data, and the *ridgeless* regime, see Proposition B.7 and Proposition B.9, respectively. Specifically, using the asymptotic behavior $\kappa(\lambda) \approx \lambda + \gamma \mathbb{E}_H(\xi)$ as $\gamma \to +\infty$, cf. Lemma B.10, we get in the isotropic regime:

$$\tfrac{1}{2}\,\mathsf{SNR} = \lim_{\gamma \to 0^+} R_{\gamma,\lambda} = \lim_{\gamma \to +\infty} R_{\gamma,\lambda}\,, \tag{16}$$

and $R_{\gamma,\lambda}$ has a minimum in $(1, \infty)$. While the first limit in (16) is well-known [28], the fact that even with $\gamma \to +\infty$ the rate tends to the optimal value is a new interesting fact. In general, however, it turns out that (16) is usually not the case:

Proposition 2.3. *Suppose Assumption 1 holds, and let $R_{\gamma,\lambda}$ be given by Theorem 2.2. Then*

$$\lim_{\gamma \to 0^+} R_{\gamma,\lambda} \geq \lim_{\gamma \to +\infty} R_{\gamma,\lambda}\,, \tag{17}$$

and equality holds if and only if $\mathsf{dSNR}\,\mathrm{d}H$ is supported on a single eigenvalue.

However, as we show in Appendix B.5, as soon as the statistical properties of the validation set differ from that of the training set, overparameterization might actually yield the best performances.

3 Quadratic decision statistic

In this section, we consider the decision statistic (3). In this case, asymptotic means and variances under $\mathcal{H}_0$ and $\mathcal{H}_1$, denoted by $\hat{\mu}_j := \mathbb{E}[\hat{T}(x) \mid \hat{X}, \mathcal{H}_j]$ and $\hat{\sigma}_j^2 := \mathbb{E}[(\hat{T}(x) - \hat{\mu}_j)^2 \mid \hat{X}, \mathcal{H}_j]$, $j = 0, 1$, can be computed as follows.

Proposition 3.1. *Let $\hat{T}$ be the decision statistic defined in* (3)*, and $\hat{\mu}_j$, $\hat{\sigma}_j^2$ be its mean and variance under $\mathcal{H}_j$, $j = 0, 1$. Then, $\hat{\mu}_j/p$ and $\hat{\sigma}_j^2/p$ converge almost surely to the following values:*

$$\begin{aligned} \mu_{0,\infty} &:= \frac{\kappa(\lambda)}{\lambda}\,\mathfrak{m}_1(\kappa(\lambda))\,, \quad \sigma_{0,\infty}^2 := \frac{\kappa(\lambda)^2}{\lambda^2}\frac{2\,\mathfrak{m}_2(\kappa(\lambda))}{1-\gamma\,\mathfrak{m}_2(\kappa(\lambda))}\,, \\ \mu_{1,\infty} &:= \mu_{0,\infty} + \frac{\kappa(\lambda)}{\lambda}\,\mathfrak{n}_1(\kappa(\lambda))\,, \quad \sigma_{1,\infty}^2 := \sigma_{0,\infty}^2 + \frac{\kappa(\lambda)^2}{\lambda^2}\frac{4\,\mathfrak{n}_2(\kappa(\lambda))}{1-\gamma\,\mathfrak{m}_2(\kappa(\lambda))}\,. \end{aligned} \tag{18}$$

The main challenge in computing the performances in the sense of the LDP in this case is that $\hat{T}$ is no longer Gaussian and the sole knowledge of (18) does not allow us to derive the rate as in Section 2. Instead, we leverage the Gärtner–Ellis Theorem [22], which states that under certain conditions, the exponential decay of error probabilities is governed by the *Fenchel–Legendre conjugate* of the limiting log-moment generating function (LMGF):

$$\Lambda^{(0,1)}(\tau) = \lim_{p\to+\infty} \frac{1}{p}\log\mathbb{E}\big[\exp\{\tau\hat{T}(x)\} \mid \hat{X}, \mathcal{H}_{0,1}\big]\,. \tag{19}$$

Using an integration trick proposed by Bai and Silverstein in [4], we show that $\Lambda^{(0)}$ can be written as an integral function with respect to the *first branch* of the Stieltjes transform of the limiting spectral distribution of $\hat{W}^{-1} := \frac{1}{n}\hat{Z}\hat{Z}^T + \lambda\Sigma^{-1}$, with $\hat{Z} := \Sigma^{-1/2}\hat{X}$, which we denote by $\nu_{\lambda,\gamma}$. Specifically, in Corollary C.4, we show that, under an additional *no-spectral-outliers* assumption (Assumption 3),

$$\Lambda^{(0)}(\tau) = \frac{1}{2}\int_0^{2\tau} \underline{\mathfrak{s}}(z)\,\mathrm{d}z\,, \quad \text{for all } \tau \neq \frac{T}{2}\,, \quad \text{where} \quad \underline{\mathfrak{s}}(\tau) := \begin{cases} \mathfrak{s}(\tau) & \text{if } \tau \leq T\,, \\ +\infty & \text{else}\,, \end{cases} \tag{20}$$

and $T := \inf\operatorname{supp}\nu_{\lambda,\gamma}$, while $\mathfrak{s}\colon \mathbb{C}_+ \to \mathbb{C}_+$ is its Stieltjes transform, characterized by [44, Theorem 1.1] as being the unique solution to the following fixed-point equation:

$$\mathfrak{s}(z) = \mathfrak{s}_{\mathsf{dRNR}}\left(z - \frac{1}{1+\gamma\mathfrak{s}(z)}\right), \quad z \in \mathbb{C}_+\,, \tag{21}$$

where $\mathfrak{s}_{\mathsf{dRNR}}$ is the Stieltjes transform of dRNR, i.e., the limiting spectral distribution of $\lambda\Sigma^{-1}$. Under $\mathcal{H}_1$, the function $\Lambda^{(1)}$ can be computed from $\Lambda^{(0)}$ with a shift, see Lemma C.5. To compute the Fenchel–Legendre conjugate $I^{(0)} := (\Lambda^{(0)})^*$, we can leverage its explicit expression and write the rate function in terms of the (pseudo) inverse Stieltjes transform of $\underline{\mathfrak{s}}$, which is defined by

$$\underline{\mathfrak{s}}^{[-1]}(u) := \inf\big\{z \in \mathbb{R} \mid \underline{\mathfrak{s}}(z) \geq u\big\}\,, \quad \text{for all } u \in \mathbb{R}\,. \tag{22}$$

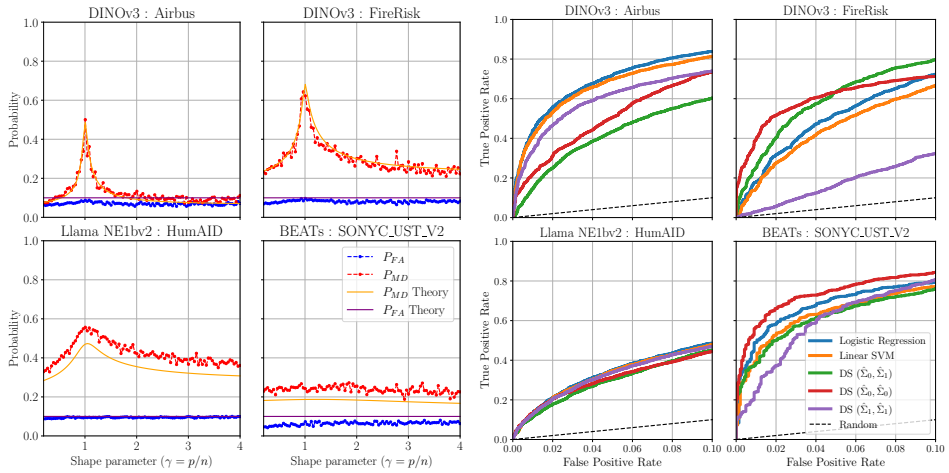
The explicit rate-function $I^{(0)}$ is given in Proposition C.6. This yields an explicit bound on the decay of the error probabilities $P_{FA} := \mathbb{P}\big(\hat{T}(x) > \tau_p \mid \hat{X}, \mathcal{H}_0\big)$ or $P_{MD} := \mathbb{P}\big(\hat{T}(x) \leq \tau_p \mid \hat{X}, \mathcal{H}_1\big)$ when the alternative is (asymptotically) fixed:

Theorem 3.2. *Suppose Assumptions 1 and 3 hold. For $\beta \in (0,1)$, choose the threshold $\tau_p = \hat{\tau}_1 := \hat{\mu}_1 + \hat{\sigma}_1 Q^{-1}(1-\beta)$. Then, almost surely, $P_{MD} \to \beta$ and*

$$\limsup_{p\to\infty} \frac{1}{p}\log P_{FA} \leq -R_{\gamma,\lambda}^{(0)}\,, \qquad R_{\gamma,\lambda}^{(0)} := I^{(0)}(\mu_{1,\infty}) = \frac{1}{2}\int_{\mu_{0,\infty}}^{\mu_{1,\infty}} \underline{\mathfrak{s}}^{[-1]}(u)\,\mathrm{d}u\,, \tag{23}$$

where $\hat{\mu}_1, \hat{\sigma}_1, \mu_{0,\infty}$ and $\mu_{1,\infty}$ are given in Proposition 3.1.

Whenever dRNR is *sharp*, i.e., $\mathfrak{s}_{\mathsf{dRNR}}(\tau) \to +\infty$ for $\tau \uparrow \inf\operatorname{supp}\mathsf{dRNR}$, which is the case, e.g., for isotropic data, we can easily compute $\underline{\mathfrak{s}}^{[-1]}$ in (22) leveraging (21), obtaining i) explicit expressions for isotropic data, and ii) accurate numerical approximations of $I^{(0)}$ for non-isotropic data. We refer to Appendix C.2.2 for further details, and to Remark C.8 for the $\mathcal{H}_1$ case, i.e., the decay of P_{MD} with fixed P_{FA}, which follows from the $\mathcal{H}_0$ case. The lower bound is notoriously trickier [17], and has to be analyzed on a case-by-case basis, see Proposition C.9 for a case where the LDP holds, i.e., the limsup is a limit and the inequalities are equalities. Interestingly, even in this case, the rate function exhibits multiple-descent behavior, cf. Figure 9.



(a) Theoretical and empirical P_{FA} and P_{MD}.

(b) ROC curves.

Figure 3: **Left:** Comparison between theoretical and empirical P_{FA} and P_{MD} with desired $P_{FA} = 0.1$. **Right:** ROC curve comparison on small P_{FA} values with benchmark classifier based on logistic regression and linear SVM. Please refer to Appendix D.2 for details on the legend and to Figure 16 for the full ROC curves both in linear and logarithmic scale.

4 Neyman–Pearson detection in foundation models' latent space

Detection techniques designed for (1) are not directly applicable to real structured data. To deal with this issue and motivated by [5], we instead map the observations into a latent space using encoders from foundation models and apply the decision statistic in the resulting representations, leading to the framework depicted in Figure 1. The key properties we establish are as follows:

- **Near-optimal detection at fixed false alarm:** As $\hat{T}$ approximates the Bayes's decision rule under Gaussian data, we show that our approach outperforms the more standard fine-tuning of logistic regression and linear SVM classifiers when we seek to maximize the probability of detection subject to a fixed probability of false alarm constraint.
- **Predictable performances:** We show that both the LDP rate as well as the probability of detection closely match their theoretical predictions, thereby providing solid theoretical foundations for our detection methodology.
- **Data and model agnostic approach:** We validate our approach across multiple foundation models and real-world datasets, showing consistent performance across vision, text, and audio modalities, and across DINOv3, Llama-Nemotron, and BEATs encoders.

4.1 Experimental design

We now summarize our main experimental pipeline. Although our analysis in Section 2 assumes the same covariance matrix under the two hypotheses, the experimental data exhibit covariance mismatch. Our theoretical predictions can be extended to account for this mismatch and are used to fit the empirical results, as detailed in Appendix D.2.1, together with further implementation details. See also Appendix D.2.4 for an alternative approach to tackle the covariance mismatch.

First, we employ a pre-trained foundation model with latent space of dimension $p \in \mathbb{N}$ and a balanced dataset for binary classification divided into a training set of size $\hat{n} \in \mathbb{N}$ and a validation set. We range across different values of γ by truncating the training set to a given parameter $n < \hat{n}$. This way, by decreasing n we reach potentially large $\gamma = p/n$, at the cost of i) losing statistical meaningfulness for very large γs and ii) accepting to not be able to inspect γ smaller than $\hat{\gamma} := p/\hat{n}$. We use the full dataset to estimate the population spectral distribution H and dSNR.

Once these are fixed, for each foundation model and each dataset, we run the following pipeline. We fix $\lambda > 0$; then, we vary $\gamma > \hat{\gamma}$ up to a fixed threshold, and, for each γ: We utilize the truncated

training set to compute the covariance matrix $\hat{\Sigma}$ and the estimate of the signal $\hat{s}$ on $\mathcal{H}_1$ training data, which we then use to form $\hat{T}(x) := \hat{s}^T \hat{\Sigma}_\lambda^{-1} x$. Next:

1. We use Proposition 2.1 to compute μ_∞, σ_∞^2 and the rate function, and compare them to the corresponding empirical ones estimated on the validation set. We adapt our formulas to account for covariance mismatch between the two hypotheses, as detailed in Appendix D.2, at essentially no additional computational cost. The results are presented in Figure 1 for the rate and Figure 13 for mean and variance.
2. We derive, for Gaussian data, the probability of detection with fixed false alarm and compare it with its empirical counterpart, which strongly depends on the underlying distribution. The results are provided in Figure 3a.
3. We compare the decision statistic (2) with logistic regression and linear SVM classifiers. We further consider the decision statistics obtained by approximating the LLR assuming different covariance matrices, as specified in Appendix D.2.4. The results are presented in Figure 3b.

For each point, we discuss the results and how they support our claims regarding the key properties introduced at the beginning of Section 4. Since the experiment is identical across models and datasets, we report aggregate results only. We fix $\lambda = 10^{-3}$ across all different cases, although the behavior is consistent with other choices.

Foundation models and data We evaluate our methodology across diverse modalities and models. First, we utilize two optical earth observation datasets: The Airbus Ship Detection Challenge (Airbus) [26] and a fire risk assessment dataset (FireRisk) [43], both paired with DINOv3 [45] for binary classification. For textual data, we employ the HumAID [1] dataset of human-annotated disaster incidents from Twitter, using Llama-Nemotron-Embed-1B-v2 [35] for feature extraction. Finally, we assess audio data through the SONYC-UST-V2 [13] dataset, utilizing the BEATs [14] model for feature extraction. We refer the reader to Appendix D.1 for a description of the datasets and models.

4.2 Results and discussion

In this section, we discuss the results of our numerical experiments listed in Section 4.1. First, in Appendix D.2, in Figure 11, we present four randomly selected marginals of the training $\mathcal{H}_0$ data mapped into the latent space. As we can observe, the data appear to perfectly match Gaussian distributions, supporting the insights of [5], and strengthening the foundations of our approach.

Point 1: In Figure 1, we show a comparison of the theoretical and empirical rate functions across all different cases. This figure demonstrates the close agreement between the theoretical formulas derived in Section 2 and the corresponding empirical behaviors, even when H and dSNR are only approximately estimated with $\gamma \neq 0$. In particular, it shows an overall good match, with only a small vertical shift observed for DINOv3 applied to the FireRisk dataset, which we believe might be due to an inaccurate estimation of H or dSNR, cf. Figure 14.

Point 2: Figure 3a illustrates that the probabilities of detection at fixed false-alarm levels—which depend strongly on the Gaussianity of the underlying distribution—closely follow the theoretical predictions. Again, the only noticeable mismatches (e.g., for Llama and BEATs) appear as vertical shifts, while the overall shape of the curves remains well captured. Overall, Figure 3a shows that this does not significantly compromise the detection probability estimate with fixed false alarm levels.

Point 3: Eventually, we conduct a systematic comparison with standard baselines. In Figure 3b, we present the ROC curves for the three approximate LLRs and logistic regression as well as linear SVM classifiers. We can observe in particular that the approximate LLRs always achieve excellent performances (with basically no fine-tuning), except when the covariance matrices under $\mathcal{H}_0$ and $\mathcal{H}_1$ are assumed equal and estimated from $\mathcal{H}_1$ training data. Notably, we can observe that the LLRs are systematically better than the baselines for small false alarm levels and achieve comparable (or better) performance for larger values.

Finally, our numerical experiments demonstrate the three key properties of our detection methodology, yielding a theoretically grounded near-optimal detector in the sense of Neyman–Pearson (Point 3), with predictable performances (Points 1 and 2), applicable to a broad range of problems.

Limitations The main limitation of our approach is the potential difficulty in estimating H and dSNR, which may be intractable in some settings. Moreover, cases beyond Assumption 1, particularly those featuring degenerate covariance matrices, are not covered by the current analysis. Extending the results to this practically relevant setting, as conventional ML architectures may produce degenerate Gaussian distributions, appears feasible provided that λ remains bounded away from zero, although further theoretical developments would be required.

5 Conclusions

Our findings bridge the gap between high-dimensional probability theory and the practical use of foundation models for detection and classification tasks. We develop a high-dimensional statistical framework that accurately predicts the performance of data-driven decision rules applied to foundation-model representations across diverse modalities, including vision, text, and audio. These results support high-dimensional asymptotics as a useful tool for understanding and engineering robust detection systems based on modern foundation models. Among other promising directions for future research, we aim to characterize the LDP rate for a broader family of non-Gaussian statistics, derive tighter error bounds in the multi-class setting, and investigate the asymptotic regime in which the number of classes grows jointly with p and n.

Acknowledgments and Disclosure of Funding

This work was supported by NATO Allied Command Transformation (ACT) under the Seabed-to-Space Situational Awareness (S3A) research programme. We would like to thank Samantha Dugelay, Rosario Di Carlo, and Davide Cortesi for fruitful discussions related to this work.

References

[1] Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda Ofli. Humaid: Human-annotated disaster incidents data from twitter. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media*, https://huggingface.co/datasets/QCRI/HumAID-all, 2021. AAAI.

[2] Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. *SIAM Journal on Mathematics of Data Science*, 6(1):26–50, 2024.

[3] Z. D. Bai and Y. Q. Yin. Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix. *The Annals of Probability*, 21(3):1275–1294, 1993.

[4] Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer, New York, 2 edition, 2010.

[5] Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. 2025. arxiv.org/abs/2511.08544.

[6] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, Cham, 2 edition, 2017.

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

[8] 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.

[9] Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. Does data interpolation contradict statistical optimality? In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1611–1619. PMLR, 16–18 Apr 2019.

[10] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, January 2006.

[11] Paolo Braca, Leonardo M. Millefiori, Augusto Aubry, Stefano Marano, Antonio De Maio, and Peter Willett. Statistical hypothesis testing based on machine learning: Large deviations analysis. *IEEE Open Journal of Signal Processing*, 3:464–495, 2022.

[12] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[13] Mark Cartwright, Ana Elisa Mendez Mendez, Jason Cramer, Vincent Lostanlen, Graham Dove, Hao-Hsiang Wu, Justin Salamon, Oded Nov, and Juan Pablo Bello. SONYC urban sound tagging (SONYC-UST): A multilabel dataset from an urban acoustic sensor network. In *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, pages 35–39, 2019.

[14] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. Beats: audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023.

[15] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

[16] Harald Cramér. Sur un nouveau théorème–limite de la théorie des probabilités. In *Actualités Scientifiques et Industrielles*, number 736, pages 5–23. Hermann, Paris, 1938. Colloque consacré à la théorie des probabilités.

[17] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Stochastic Modelling and Applied Probability. Springer, Berlin, Heidelberg, 2 edition, 2010.

[18] 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, 06 2022.

[19] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

[20] Abhimanyu Dubey et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[21] Richard S. Ellis. Large deviations for a general class of random vectors. *The Annals of Probability*, 12(1):1–12, 1984.

[22] Jürgen Gärtner. On large deviations from the invariant measure. *Theory of Probability and Its Applications*, 22(1):24–39, 1977.

[23] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017.

[24] Ryan Harvey, Paolo Braca, Leonardo M. Millefiori, and Peter Willett. Sequential hypothesis testing based on machine learning. In *2024 27th International Conference on Information Fusion (FUSION)*, pages 1–8, 2024.

[25] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

[26] Kaggle. Airbus ship detection challenge. `https://www.kaggle.com/competitions/airbus-ship-detection`, 2018. Accessed: 2025-11-10.

[27] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[28] Steven M. Kay. *Fundamentals of Statistical Signal Processing, Detection Theory*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1998.

[29] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. In *International Conference on Learning Representations (ICLR)*, 2025.

[30] Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer Cham, 4 edition, 2022.

[31] Daniel LeJeune, Pratik Patil, Hamid Javadi, Richard G. Baraniuk, and Ryan J. Tibshirani. Asymptotics of the sketched pseudoinverse. *SIAM Journal on Mathematics of Data Science*, 6(1):199–225, 2024.

[32] M. Loog, T. Viering, A. Mey, J. H. Krijthe, and D. M. J. Tax. A brief prehistory of double descent. *Proceedings of the National Academy of Sciences of the United States of America*, 117(20):10625–10626, 2020.

[33] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

[34] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

[35] Gabriel de Souza P. Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. Improving text embedding models with positive-aware hard-negative mining. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, page 2169–2178. Association for Computing Machinery, 2025.

[36] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations (ICLR)*, 2020.

[37] M. Opper, W. Kinzel, J. Kleinz, and R. Nehl. On the ability of the optimal perceptron to generalise. *Journal of Physics A: Mathematical and General*, 23:L581–L586, 1990. Letter to the Editor.

[38] Pratik Patil and Daniel LeJeune. Asymptotically free sketched ridge ensembles: Risks, cross-validation, and tuning. In *The Twelfth International Conference on Learning Representations*, 2024.

[39] Detlef Plachky and Josef Steinebach. A theorem about probabilities of large deviations with an application to queuing theory. *Periodica Mathematica Hungarica*, 6(4):343–345, 1975.

[40] Francisco Rubio and Xavier Mestre. Spectral convergence for a general class of random matrices. *Statistics & Probability Letters*, 81(5):592–602, 2011.

[41] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[42] Shayle R. Searle. *Linear Models*. Wiley Classics Library. John Wiley & Sons, 1997.

[43] Shuchang Shen, Sachith Seneviratne, Xinye Wanyan, and Michael Kirley. Firerisk: A remote sensing dataset for fire risk assessment with benchmarks using supervised and self-supervised learning. In *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 189–196, 2023.

[44] Jack W. Silverstein and Zhidong D. Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):175–192, 1995.

[45] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025.

[46] 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.

[47] Antonia Tulino and Sergio Verdú. Random matrix theory and wireless communications. *Foundations and Trends in Communications and Information Theory*, 1, 01 2004.

[48] F. Vallet, J.-G. Cailton, and Ph Refregier. Linear and nonlinear extension of the pseudo-inverse solution for learning boolean functions. *Europhysics Letters*, 9(4):315, jun 1989.

Supplementary Material

A Preliminaries **15**

A.1 Notation 15

A.2 Random matrix theory 15

A.3 Large deviation theory 16

B Missing proofs and further insights of Section 2 **17**

B.1 Calibrating error probabilities 19

B.2 Isotropic data 20

B.3 Ridgeless case 20

B.4 Asymptotic values 21

B.5 Linear decision statistic with miss-specified mean 22

B.6 Extension to the multi-hypotheses setting 24

C Missing proofs and further insights on Section 3 **26**

C.1 High-dimensional mean and variance 26

C.2 Rate in the LDP sense 27

D Further insights on Section 4 **34**

D.1 Models and data 34

D.2 Experimental design 36

D.3 Beyond foundation models 41

A Preliminaries

A.1 Notation

$\mathbb{R}$ is the real line, $\mathbb{R}_+$ and $\mathbb{R}_{++}$ are non-negative and positive real numbers. $\mathbb{C}$ is the complex plane and $\mathbb{C}_+ := \{z \in \mathbb{C} \mid \mathfrak{Im}(z) > 0\}$. We denote by $\mathbf{1} := (1, \dots, 1) \in \mathbb{R}^d$, with $d \in \mathbb{N}$ depending on the context. We use the notation $\lesssim$ to indicate that the inequality holds up to a positive constant. For a matrix $A \in \mathbb{R}^{p \times p}$ with real eigenvalues $\xi_1, \dots, \xi_p$, we call the *spectral measure of* A the probability measure defined as the sum of δ_{ξ_i}, i.e., the Dirac deltas centered on ξ_i, with uniform weights. The set $\mathcal{P}(D)$ denotes the family of probability measures on $D \subset \mathbb{R}^d$, $d \in \mathbb{N}$. The set $C_b(\mathbb{R})$ the family of continuous and bounded functions on $\mathbb{R}$. We use the notation of *asymptotic equivalents*, i.e., for two sequences $(a_n)_n$ and $(b_n)_n$ we say that $a_n \sim b_n$ as $n \to +\infty$ if $\lim_{n\to+\infty} \frac{a_n}{b_n} = 1$. We denote by $\mathbb{E}$ the *expectation* with respect to the underlying probability measure $\mathbb{P}$ defined on a suitable probability space $(\Omega, \mathfrak{F}, \mathbb{P})$ where $\mathfrak{F}$ is the σ-algebra, and by Var the *variance*. We use $\mathbb{E}_m$ to refer to the expectation taken with respect to the measure m.

A.2 Random matrix theory

We start out by showing how to solve the Silverstein equation (11).

Solving the Silverstein Equation The Silverstein equation is a nonlinear fixed-point equation that does not admit a closed-form solution in general. Let us first propose an approximation strategy based on fixed-point iterations. To solve (12) for a fixed $\lambda > 0$, we need to find a fixed point of the operator $\mathcal{F}\colon [0, +\infty) \to [0, +\infty)$ defined by

$$\mathcal{F}(\kappa) := \lambda + \gamma\kappa \int_0^{+\infty} \frac{\xi}{\kappa + \xi} \, \mathrm{d}H(\xi) \,, \quad \text{for all } \kappa \geq 0 \,, \tag{24}$$

which can be easily seen to be an increasing and concave function in $[0, +\infty)$, due to Assumption A1. Therefore, for each fixed λ, we can approximate efficiently $\kappa(\lambda)$ by κ_λ^K obtained by fixed-point iterations $\kappa_\lambda^{i+1} = \mathcal{F}(\kappa_\lambda^i)$ for all $i = 0, \dots, K-1$.

An alternative approximation strategy used, e.g., in [2], consists in taking the almost-sure limit of

$$\hat{\varphi}(z) := \frac{1}{n} \operatorname{tr}\left[\left(\frac{1}{n}\hat{X}^T\hat{X} - zI\right)^{-1}\right], \tag{25}$$

since indeed for all $z \in \mathbb{C} \setminus \mathbb{R}_+$, $\hat{\varphi}(z) \to \varphi(z)$ almost surely. However, this requires inverting high-dimensional matrices and the results are subject to randomness.

Isotropic data In the case of isotropic data, the solution to the Silverstein equation is explicit. Let $\Sigma = \sigma^2 I$, for some $\sigma > 0$, then $H = \delta_{\sigma^2}$, and the Silverstein fixed-point equation (11) turns into

$$\frac{1}{\varphi(z)} + z = \gamma \frac{\sigma^2}{1 + \sigma^2\varphi(z)} \quad \text{for } z \in \mathbb{C} \setminus \mathbb{R}_+ \,. \tag{26}$$

This admits the following closed-form solution: Setting $\gamma_\pm = \sigma^2(1 \pm \sqrt{\gamma})^2$,

$$\varphi(z) = \frac{-(z + \sigma^2 - \gamma\sigma^2) + q(z)}{2\sigma^2 z} \,, \quad q(z)^2 := (z - \gamma_-)(z - \gamma_+) \,, \tag{27}$$

where q denotes the analytic branch satisfying $q(z)/z \to 1$ as $|z| \to \infty$. In particular, $q(z) < 0$ for real $z < 0$. Plugged into (A.5) of [2], i.e.,

$$\gamma\psi(z) = \varphi(z) + \frac{1 - \gamma}{z} \,, \quad \text{for all } z \in \mathbb{C}_+ \,, \tag{28}$$

yields an explicit formula for ψ, and, by Stieltjes inversion, yields the celebrated Marchenko–Pastur distribution [33], which is given by:

$$\nu_{\mathrm{MP}} := \max\left\{0, 1 - \frac{1}{\gamma}\right\}\delta_0 + f_\gamma \, \mathrm{d}x, \quad \text{where} \quad f_\gamma(x) := \frac{1}{2\pi\sigma^2} \frac{\sqrt{(\gamma_+ - x)(x - \gamma_-)}}{\gamma x} \,. \tag{29}$$

Trace limits We recall below a technical result due to Rubio and Mestre [40] and Bai and Silverstein [4], which we use repeatedly. See [2] for a comprehensive, self-contained account.

Theorem A.1. *Let $\hat{X} = \Sigma^{1/2}\hat{Z}$ with $\hat{Z} \in \mathbb{R}^{p\times n}$ having sub-Gaussian i.i.d. components with mean zero and unit variance and suppose that Assumption 1 holds. Let A, B bounded in operator norm and such that $\sum_{i=1}^p (v_i^T A v_i)\delta_{\sigma_i}$ and $\sum_{i=1}^p (v_i^T B v_i)\delta_{\sigma_i}$ converge to measures ν_A and ν_B with bounded total variation. Then, for $z \in \mathbb{C}\setminus\mathbb{R}_+$, as $p, n \to +\infty$ with $\gamma = p/n \geq 0$ constant, we get*

$$\operatorname{tr}\left[A(\hat{\Sigma} - zI)^{-1}\right] \sim -\frac{1}{z\varphi(z)} \operatorname{tr}\left[A\left(\Sigma + \frac{1}{\varphi(z)} I\right)^{-1}\right], \tag{30}$$

$$\begin{aligned}\operatorname{tr}\left[A(\hat{\Sigma} - zI)^{-1}B(\hat{\Sigma} - zI)^{-1}\right] &\sim \frac{1}{z^2\varphi(z)^2} \operatorname{tr}\left[A\left(\Sigma + \frac{1}{\varphi(z)} I\right)^{-1} B\left(\Sigma + \frac{1}{\varphi(z)} I\right)^{-1}\right] \\ &+ \frac{1}{z^2\varphi(z)^2} \operatorname{tr}\left[A\left(\Sigma + \frac{1}{\varphi(z)} I\right)^{-2}\Sigma\right] \operatorname{tr}\left[B\left(\Sigma + \frac{1}{\varphi(z)} I\right)^{-2}\Sigma\right] \frac{1}{n - \mathrm{d}f_2(1/\varphi(z))}\,.\end{aligned} \tag{31}$$

where φ is the unique solution to (11) *and $\mathrm{d}f_2(\kappa) := \operatorname{tr}\left(\Sigma^2(\Sigma + \kappa I)^{-2}\right)$ for all $\kappa \in \mathbb{C}\setminus\mathbb{R}_-$.*

Observe that the Gaussianity assumption is not necessary to get the results of Theorem A.1, see also [31, 38] for weaker assumptions on A and B.

A.3 Large deviation theory

The theory of large deviations investigates the asymptotic behavior of the tails of sequences of probability distributions. Before delving into the general framework, we discuss a simple example:

Example A.2. *Suppose that $(\hat{T}(x))_p$ is a sequence of Gaussian random variables with $\mathbb{E}[\hat{T}(x) \mid \mathcal{H}_0] = 0$, $\mathbb{E}[\hat{T}(x) \mid \mathcal{H}_1] = \hat{\mu}$ and $\operatorname{Var}[\hat{T}(x) \mid \mathcal{H}_0] = \operatorname{Var}[\hat{T}(x) \mid \mathcal{H}_1] = \hat{\sigma}^2$. The probability of* miss detection*, and of the* false alarm*, only depend on the first two moments:*

$$\begin{aligned} P_{FA} &:= \mathbb{P}(\hat{T}(x) \geq \tau \mid \mathcal{H}_0) = Q\big(\tau/\hat{\sigma}\big) = Q\big(Q^{-1}(P_D) + \hat{d}\big)\,, \\ P_{MD} &:= \mathbb{P}(\hat{T}(x) \leq \tau \mid \mathcal{H}_1) = 1 - Q\Big(\frac{\tau - \hat{\mu}}{\hat{\sigma}}\Big) = 1 - Q\big(Q^{-1}(P_{FA}) - \hat{d}\big)\,, \end{aligned} \tag{32}$$

where Q is the standard Q-function and $\hat{d}^2 := \hat{\mu}^2/\hat{\sigma}^2$. In this setting, using the well-known asymptotic behavior of the Q function, if $\hat{d}^2 \sim p$ as $p \to +\infty$, the probability of miss detection with fixed false alarm $P_{FA} = \alpha > 0$, i.e., with $\tau = \hat{\sigma}Q^{-1}(\alpha)$, can be expressed in terms of $\hat{d}$ as:

$$\frac{1}{p}\log P_{MD} = \frac{1}{p}\log\big(1 - Q\big(Q^{-1}(\alpha) - \hat{d}\big)\big) = -\frac{\hat{d}^2}{2p} + o(1)\,, \quad \text{as } p \to +\infty\,. \tag{33}$$

Clearly, the same result holds switching the roles of P_{FA} and P_{MD}. Observe that the decay of P_{MD} is governed by the left tail of $\hat{T}(x)$ under $\mathcal{H}_1$, and is by no means a surprise that in the Gaussian setting it can indeed be controlled by only its scale parameters.

In a different yet simple setting, the large deviation theory provides the rate of convergence in the weak law of large numbers. To be precise, consider any i.i.d. random variables $z_1, \ldots, z_n$ with mean μ and LMGF $\Lambda(s) = \log\mathbb{E}(e^{sz_1}) < +\infty$ for all $s \in \mathbb{R}$, and let $S_n := n^{-1}\sum_{i=1}^n z_i$. Then for any $\tau > \mu$, we have the following decay of the tail probabilities:

$$\lim_{n\to+\infty} \frac{1}{n}\log\mathbb{P}(S_n > \tau) = -\Lambda^*(\tau)\,, \tag{34}$$

where Λ^* is the Fenchel–Legendre conjugate of Λ [16], see [17, Theorem 2.2.30]. In this context, we say that the LDP holds with rate function Λ^*. More generally, we refer to the following definition:

Definition A.3 (Large Deviation Principle)**.** *A sequence of random variables $(\mathsf{t}_p)_p$ is said to satisfy the* Large Deviation Principle *with* rate function $I\colon \mathbb{R} \to \mathbb{R}\cup\{+\infty\}$ *if for all Borel sets $\Gamma \subset \mathbb{R}$*

$$\begin{aligned} -\inf_{\tau\in\operatorname{Int}\Gamma} I(\tau) &\leq \liminf_{p\to+\infty} \frac{1}{p}\log\mathbb{P}(\mathsf{t}_p \in \Gamma) \\ &\leq \limsup_{p\to+\infty} \frac{1}{p}\log\mathbb{P}(\mathsf{t}_p \in \Gamma) \leq -\inf_{\tau\in\bar{\Gamma}} I(\tau)\,. \end{aligned} \tag{35}$$

The following results, due to Gärtner [22], Ellis [21] and Plachky and Steinebach [39], provides a broad class of settings in which the LDP holds. It requires that the following (possibly infinite) limit exists:

$$\Lambda(s) := \lim_{p\to+\infty} \frac{1}{p}\Lambda_p(ps) := \lim_{p\to+\infty} \frac{1}{p}\log \mathbb{E}(e^{ps\mathtt{t}_p}), \quad \text{for all } s \in \mathbb{R}. \tag{36}$$

We provide below a simplified statement in one dimension and refer, for example, to [17, Theorem 2.3.6] for a complete discussion. In particular, the lower bound depends on the concept of *exposed points* of a convex function, which formally captures points of strict convexity; see [17, Definition 2.3.3] for a precise definition.

Theorem A.4 (Gärtner–Ellis)**.** *Let $(\mathtt{t}_p)_p$ be a sequence of random variables such that $\Lambda\colon \mathbb{R} \to \mathbb{R}\cup\{+\infty\}$ defined in* (36) *exists and is such that $0 \in \operatorname{Int}\operatorname{Dom}\Lambda$. Then:*

1. *For every closed set $C \subset \mathbb{R}$,*

$$\limsup_{p\to+\infty} \frac{1}{p}\log \mathbb{P}(\mathtt{t}_p \in C) \le -\inf_{\tau\in C} \Lambda^*(\tau). \tag{37}$$

2. *For every open set $O \subset \mathbb{R}$,*

$$\liminf_{p\to+\infty} \frac{1}{p}\log \mathbb{P}(\mathtt{t}_p \in O) \ge -\inf_{\tau\in O\cap\mathcal{E}} \Lambda^*(\tau), \tag{38}$$

where $\mathcal{E}$ is the set of exposed points *for Λ^* whose exposing line belongs to* $\operatorname{Int}\operatorname{Dom}\Lambda$.

Suppose in addition that Λ is lower semicontinuous in $\mathbb{R}$, differentiable in $\operatorname{Int}\operatorname{Dom}\Lambda$*, and such that*

$$\lim_{s\to\bar{s}} |\Lambda'(s)| = +\infty, \quad \text{for all } \bar{s} \in \partial\operatorname{Dom}\Lambda. \tag{39}$$

Then, $(\mathtt{t}_p)_p$ satisfies the LDP with rate function Λ^*.

B Missing proofs and further insights of Section 2

The following remark is going to be useful to lighten the upcoming proofs.

Fact B.1. *Suppose that Assumption 1 holds. The observables $\mathfrak{m}_q$ and $\mathfrak{n}_q$ can be written as limits of trace functionals: Let $v_1,\dots,v_p$ be the basis of eigenvectors of Σ and $\xi_1,\dots,\xi_p$ the corresponding eigenvalues. By the properties of the trace operator, as $p \to +\infty$,*

$$\operatorname{tr}\left[\frac{ss^T}{p}\big(\Sigma+\kappa I\big)^{-1}\right] = \frac{1}{p}\sum_{i=1}^{p}\frac{\big(v_i^T s\big)^2}{\xi_i+\kappa} \to \int_{\mathbb{R}_+}\frac{1}{\xi+\kappa}\,\mathrm{d}\nu_s(\xi)$$
$$= \int_{\mathbb{R}_+}\left(\frac{\xi}{\xi+\kappa}\right)\mathsf{dSNR}(\xi)\,\mathrm{d}H(\xi) = \mathfrak{n}_1(\kappa),$$

where in the last step we have used Assumption A2 with $\theta^* = s$. Similarly,

$$\operatorname{tr}\left[\frac{ss^T}{p}\big(\Sigma+\kappa I\big)^{-2}\Sigma\right] \sim \operatorname{tr}\left[\frac{ss^T}{p}\big(\Sigma+\kappa I\big)^{-1}\Sigma\big(\Sigma+\kappa I\big)^{-1}\right] \sim \mathfrak{n}_2(\kappa).$$

While, for $\mathfrak{m}_q$ we have, using the same approach $\frac{1}{p}\operatorname{tr}\left[\Sigma\big(\Sigma+\kappa I\big)^{-1}\right] \sim \mathfrak{m}_1(\kappa)$ and

$$\frac{1}{p}\operatorname{tr}\left[\Sigma\big(\Sigma+\kappa I\big)^{-2}\Sigma\right] \sim \frac{1}{p}\operatorname{tr}\left[\Sigma\big(\Sigma+\kappa I\big)^{-1}\Sigma\big(\Sigma+\kappa I\big)^{-1}\right] \sim \mathfrak{m}_2(\kappa).$$

We now discuss how Assumption 1 as well as the observables defined in Definition 1.2 translate in two notable cases:

Example B.2 (Isotropic data)**.** *Suppose that for each $p \in \mathbb{N}$, $\Sigma = \sigma^2 I$ with $\sigma > 0$. Then Assumptions A1 and A2 are satisfied. Indeed, the spectral measure of Σ is constant δ_{σ^2}, hence converges to itself, showing Assumption A1 with $H = \delta_{\sigma^2}$ and $a = b = \sigma^2$. Moreover, pick a signal $\theta^* = (\theta_1^*,\dots,\theta_p^*) \in \mathbb{R}^p$ such that $\frac{1}{p}\|\theta^*\|^2$ converges. Then, $\|\theta^*\|^2 = \mathcal{O}(p)$ as $p \to +\infty$, and*

$$\frac{1}{p}\sum_{i=1}^{p}(v_i^T\theta^*)^2\delta_{\xi_i} = \frac{1}{p}\sum_{i=1}^{p}(\theta_i^*)^2\delta_{\sigma^2} = \frac{1}{p}\|\theta^*\|^2\delta_{\sigma^2} \rightharpoonup \Big(\lim_{p\to+\infty}\frac{1}{p}\|\theta^*\|^2\Big)\delta_{\sigma^2} =: \nu_{\theta^*}. \tag{40}$$

Furthermore, $\nu_s \ll H$, and in this case dSNR is a function defined only on σ^2 with value that indeed coincides with $\lim_{p\to+\infty} \frac{1}{p}\|s\|^2/\sigma^2$, i.e., the SNR; thus, Assumption A2 holds. Therefore, $\mathsf{dSNR}\,\mathrm{d}H = \mathsf{SNR}\,\delta_{\sigma^2}$, and

$$\mathfrak{m}_q(\kappa) = \Big(\frac{\sigma^2}{\sigma^2+\kappa}\Big)^q, \quad \textit{and} \quad \mathfrak{n}_q(\kappa) = \Big(\frac{\sigma^2}{\sigma^2+\kappa}\Big)^q \mathsf{SNR}\,. \tag{41}$$

Example B.3. *Suppose that Assumptions A1 and A2 hold. Then, $\lim_{p\to+\infty} \frac{1}{p} s^T\Sigma^{-1}s$ exists. Additionally, for all $q \in \mathbb{N}$, we have, clearly $\mathfrak{m}_q(0) = 1$, and*

$$\begin{aligned}\mathfrak{n}_q(0) = \int_{\mathbb{R}_+} \mathsf{dSNR}(\xi)\,\mathrm{d}H(\xi) &= \int_{\mathbb{R}_+} \frac{1}{\xi}\,\mathrm{d}\nu_s(\xi) \\ &= \lim_{p\to+\infty} \frac{1}{p}\sum_{i=1}^{p} \frac{(v_i^T s)^2}{\xi_i} = \lim_{p\to+\infty} \frac{1}{p} s^T\Sigma^{-1}s = \mathsf{SNR}\,.\end{aligned} \tag{42}$$

This shows that $\mathfrak{m}_q$ and $\mathfrak{n}_q$ can be seen as regularized moments of the limiting spectral distribution of Σ and the signal-to-noise ratio, respectively.

Proof of Proposition 2.1. Start from the mean and let us first consider the decision statistic $\hat{T}(x) = s^T\hat{\Sigma}_\lambda^{-1}x$. We have:

$$\begin{aligned}\frac{\hat{\mu}}{p} &= \frac{1}{p}\mathbb{E}[\hat{T}(x) \mid \hat{X}, \mathcal{H}_1] = \frac{1}{p} s^T\hat{\Sigma}_\lambda^{-1}s = \frac{1}{p}\operatorname{tr}\left[s^T\hat{\Sigma}_\lambda^{-1}s\right] = \operatorname{tr}\left[\frac{ss^T}{p}\hat{\Sigma}_\lambda^{-1}\right] \\ &= \operatorname{tr}\left[\frac{ss^T}{p}\big(\hat{\Sigma}+\lambda I\big)^{-1}\right] \sim \frac{\kappa(\lambda)}{\lambda}\operatorname{tr}\left[\frac{ss^T}{p}\big(\Sigma+\kappa(\lambda)I\big)^{-1}\right] \sim \frac{\kappa(\lambda)}{\lambda}\,\mathfrak{n}_1(\kappa(\lambda))\,,\end{aligned} \tag{43}$$

where in the last line we have used Theorem A.1 with $A := \frac{ss^T}{p}$, for which $\|A\| = \|s\|^2/p$ and is then bounded (in p), and Fact B.1. Thus the convergence of $\hat{\mu}/p$.

We argue analogously for $\hat{\sigma}^2$ applying Theorem A.1 to

$$\begin{aligned}\hat{\sigma}^2 &= \mathbb{E}[(\hat{T}(x)-\hat{\mu})^2 \mid \hat{X}, \mathcal{H}_1] = \mathbb{E}[(s^T\hat{\Sigma}_\lambda^{-1}(x-s))^2 \mid \hat{X}, \mathcal{H}_1] \\ &= s^T\hat{\Sigma}_\lambda^{-1}\mathbb{E}\big[(x-s)(x-s)^T \mid \hat{X}, \mathcal{H}_1\big]\hat{\Sigma}_\lambda^{-1}s = s^T\hat{\Sigma}_\lambda^{-1}\Sigma\hat{\Sigma}_\lambda^{-1}s = \operatorname{tr}\left[ss^T\hat{\Sigma}_\lambda^{-1}\Sigma\hat{\Sigma}_\lambda^{-1}\right].\end{aligned} \tag{44}$$

Therefore, we can now use (31) with $A := ss^T/p$ and $B := \Sigma$, to get that, almost surely, as $n, p \to +\infty$ with $p/n = \gamma \in (0,\infty)$,

$$\begin{aligned}\frac{\hat{\sigma}^2}{p} &\sim \frac{\kappa(\lambda)^2}{\lambda^2}\operatorname{tr}\left[\frac{ss^T}{p}\big(\Sigma+\kappa(\lambda)I\big)^{-1}\Sigma\big(\Sigma+\kappa(\lambda)I\big)^{-1}\right] \\ &\qquad + \frac{\kappa(\lambda)^2}{\lambda^2}\operatorname{tr}\left[\frac{ss^T}{p}\big(\Sigma+\kappa(\lambda)I\big)^{-2}\Sigma\right]\operatorname{tr}\left[\Sigma\big(\Sigma+\kappa(\lambda)I\big)^{-2}\Sigma\right]\frac{1}{n-\mathrm{d}f_2(\kappa(\lambda))} \\ &\sim \frac{\kappa(\lambda)^2}{\lambda^2}\,\mathfrak{n}_2(\kappa(\lambda)) + \frac{\kappa(\lambda)^2}{\lambda^2}\,\mathfrak{n}_2(\kappa(\lambda))\,\mathfrak{m}_2(\kappa(\lambda))\frac{p}{n-\operatorname{tr}\left(\Sigma^2(\Sigma+\kappa(\lambda)I)^{-2}\right)} \\ &\sim \frac{\kappa(\lambda)^2}{\lambda^2}\,\mathfrak{n}_2(\kappa(\lambda)) + \frac{\kappa(\lambda)^2}{\lambda^2}\,\mathfrak{n}_2(\kappa(\lambda))\,\mathfrak{m}_2(\kappa(\lambda))\frac{\gamma}{1-\gamma\,\mathfrak{m}_2(\kappa(\lambda))}\,,\end{aligned}$$

where we have used once again Fact B.1. Rearranging we obtain (14).

If s is estimated from the training set, we can reduce the analysis to the case in which the mean is known by employing the following argument. As in (43) we obtain

$$\frac{\hat{\mu}}{p} = \operatorname{tr}\left[\frac{s\hat{s}^T}{p}\big(\hat{\Sigma}+\lambda I\big)^{-1}\right] = \underbrace{\operatorname{tr}\left[\frac{ss^T}{p}\big(\hat{\Sigma}+\lambda I\big)^{-1}\right]}_{:=A} + \underbrace{\operatorname{tr}\left[\frac{s(\hat{s}-s)^T}{p}\big(\hat{\Sigma}+\lambda I\big)^{-1}\right]}_{:=B}. \tag{45}$$

The asymptotic behavior of A is discussed in the point above, the only difference being that in this case $\hat{\Sigma}$ is now centered with the sample mean $\hat{s}$, which does not change the results, see [4, Chapter 3].

We now show that $B \to 0$ almost surely. Denote by $\hat{\varepsilon} := \hat{s} - s$ and note that $\hat{\varepsilon} \sim \mathcal{N}(0, \frac{1}{n_1}\Sigma)$, hence $\hat{z} := \Sigma^{-1/2}\hat{\varepsilon} \sim \mathcal{N}(0, \frac{1}{n_1}I)$, thus

$$\|\hat{z}\|^2 = \sum_{i=1}^{p} \hat{z}_i^2 = \frac{1}{n_1}\sum_{i=1}^{p}(\sqrt{n_1}\hat{z}_i)^2 = \frac{p}{n_1}\frac{1}{p}\sum_{i=1}^{p}(\sqrt{n_1}\hat{z}_i)^2 \to \gamma_1\,, \quad \text{a.s. as } p \to +\infty\,, \tag{46}$$

by the classical strong law of large numbers, where $\gamma_1 > 0$ is the limit of p/n_1. Hence, $\|\hat{\varepsilon}\|^2$ is almost surely bounded as $p \to +\infty$. Therefore, using Cauchy–Schwarz,

$$|B| := \left|\frac{1}{p}\hat{\varepsilon}^T\left(\hat{\Sigma} + \lambda I\right)^{-1}s\right| \leq \frac{1}{p}\frac{\|\hat{\varepsilon}\|\|s\|}{\lambda} = \mathcal{O}(p^{-1/2})\,, \quad \text{a.s. as } p \to +\infty\,,$$

Let us now deal with the variance. Following (44) and using the same approach as in (45), we get

$$\frac{\hat{\sigma}^2}{p} = \frac{1}{p}\operatorname{tr}\left[\hat{s}\hat{s}^T\hat{\Sigma}_\lambda^{-1}\Sigma\hat{\Sigma}_\lambda^{-1}\right] = \frac{1}{p}\hat{s}^T M\hat{s} = \underbrace{\frac{1}{p}s^T Ms}_{:=A} + \underbrace{\frac{2}{p}\hat{\varepsilon}^T Ms}_{:=B} + \underbrace{\frac{1}{p}\hat{\varepsilon}^T M\hat{\varepsilon}}_{:=C}\,, \tag{47}$$

where we have denoted by $M := \hat{\Sigma}_\lambda^{-1}\Sigma\hat{\Sigma}_\lambda^{-1}$. As above, $\|M\|$ is uniformly bounded (in p) as $p \to +\infty$. Therefore, Cauchy–Schwarz yields again $B = \mathcal{O}(p^{-1/2})$ and $C = \mathcal{O}(p^{-1})$ almost surely as $p \to +\infty$. The only term remaining is A, whose limit is computed above. □

Remark B.4. *From* (46)*, we see that if p/n_1 grows faster than $\sqrt{p}$, a bias appears in the asymptotic mean of $\hat{T}$. This highlights the role of a balanced dataset.*

Remark B.5. *Proposition 2.1 extends to non-Gaussian data with i.i.d. sub-Gaussian coordinates. In fact, Theorem A.1 does not require Gaussianity and the almost sure boundedness of $\hat{\varepsilon}$ also extends to this setting by replacing the strong law of large numbers in* (46) *with Bernstein's inequality and the Borel–Cantelli lemma.*

B.1 Calibrating error probabilities

Since $\hat{T}$ is Gaussian, the error probabilities can be calibrated directly from (32). In particular, choosing $\tau = \hat{\sigma}Q^{-1}(\alpha)$ yields $P_{FA} = \alpha$, whereas choosing $\tau = \hat{\mu} + \hat{\sigma}Q^{-1}(1-\beta)$ yields $P_{MD} = \beta$.

We can asymptotically calibrate the error probabilities beyond the Gaussian case, relying on a central limit theorem for triangular arrays under an additional condition excluding dominant coordinates. Given the practical relevance of this extension, we establish the corresponding result below.

Proposition B.6. *Suppose that the coordinates of z in* (1) *are i.i.d. sub-Gaussian random variables. Let $\hat{T}$ be defined as in* (2) *and set $\hat{w} := \Sigma^{1/2}\hat{\Sigma}_\lambda^{-1}\hat{s}$, $\hat{\mu}_0 = 0$ and $\hat{\mu}_1 = \hat{\mu}$. Assume that, almost surely,*

$$\hat{\sigma} = \|\hat{w}\| > 0\,, \quad \textit{and} \quad \lim_{p\to+\infty}\frac{\|\hat{w}\|_\infty}{\|\hat{w}\|} = 0\,. \tag{48}$$

Then, conditionally on the training data, almost surely as $p \to +\infty$

$$\frac{\hat{T}(x) - \hat{\mu}_i}{\hat{\sigma}} \xrightarrow{d} \mathcal{N}(0,1)\,, \qquad i = 0, 1\,. \tag{49}$$

Proof. For $i = 0, 1$, under $\mathcal{H}_i$ we can write $\hat{T}(x) - \hat{\mu}_i = \hat{w}^T z = \hat{w}_1 z_1 + \cdots + \hat{w}_p z_p$. The summands are independent, centered, and have total variance $\hat{\sigma}^2$. Since the coordinates of z are sub-Gaussian, $\mathbb{E}[|z_1|^3] < +\infty$. Moreover,

$$\frac{1}{\hat{\sigma}^3}\sum_{j=1}^{p}\mathbb{E}\left[|\hat{w}_j z_j|^3 \mid \hat{X}\right] = \mathbb{E}[|z_1|^3]\frac{\sum_{j=1}^{p}|\hat{w}_j|^3}{\|\hat{w}\|^3} \leq \mathbb{E}[|z_1|^3]\frac{\|\hat{w}\|_\infty}{\|\hat{w}\|} \to 0\,.$$

Thus, Lyapunov's condition is satisfied, and the claim follows from the CLT for triangular arrays. □

Consequently, for any $\alpha, \beta \in (0,1)$, choosing $\tau = \hat{\sigma}Q^{-1}(\alpha)$ yields $P_{FA} \to \alpha$, whereas choosing $\tau = \hat{\mu} + \hat{\sigma}Q^{-1}(1-\beta)$ yields $P_{MD} \to \beta$, almost surely.

B.2 Isotropic data

We now investigate a setting in which computations can be performed in closed form—the case of isotropic data, i.e., with $\Sigma = \sigma^2 I$, with $\sigma^2 > 0$, cf. Example B.2. In this case, it turns out that the rate only depends on the SNR and the *regularization-to-noise ratio* $\mathsf{RNR} := \lambda/\sigma^2$.

Proposition B.7. *Under the assumptions of Theorem 2.2, assume further that $\Sigma = \sigma^2 I$ and let $r := \mathsf{RNR} = \lambda/\sigma^2$. Then,*

$$R_{\gamma,\lambda} = \mathsf{SNR}\,\frac{\sqrt{(1+\gamma+r)^2 - 4\gamma}}{1+\gamma+r+\sqrt{(1+\gamma+r)^2-4\gamma}}\,. \tag{50}$$

Proof. Recall from Example B.2 that $H = \delta_{\sigma^2}$ and $\mathsf{dSNR}\,\mathrm{d}H = \mathsf{SNR}\,\delta_{\sigma^2}$. Thus, the observables in (10) read as in (41). Additionally, the Silverstein equation admits an explicit solution given in (27), which yields

$$\kappa(\lambda) = \frac{\lambda + \sigma^2(\gamma-1) + \sqrt{\big(\lambda+\sigma^2(\gamma-1)\big)^2 + 4\lambda\sigma^2}}{2}\,. \tag{51}$$

Altogether, (14) simplifies significantly resulting in the following:

$$\mu_\infty = \frac{\kappa(\lambda)}{\lambda}\frac{\sigma^2\,\mathsf{SNR}}{(\sigma^2+\kappa(\lambda))}\,, \quad \text{and} \quad \sigma^2_\infty = \frac{\kappa(\lambda)^2}{\lambda^2}\frac{\mathsf{SNR}\,\sigma^4}{\big((\sigma^2+\kappa(\lambda))^2 - \gamma\sigma^4\big)}\,, \tag{52}$$

By computing $R_{\gamma,\lambda} = \frac{1}{2}\mu^2_\infty/\sigma^2_\infty$, we obtain the claim after simple algebraic manipulations. □

B.3 Ridgeless case

We now investigate what we refer to as the *ridgeless regime*, i.e., the case $\lambda = 0$, yielding $\hat{T}_0(x) := \hat{s}^T\hat{\Sigma}^{-1}x$. Here, we restrict our analysis to $\gamma \in (0,1)$, since for $\gamma > 1$ the sample covariance matrix is singular. We leverage the analysis of the ridge-regularized detector by observing that $\hat{T}_0(x) = \lim_{\lambda\to 0^+}\hat{s}^T\hat{\Sigma}_\lambda^{-1}x$. To study the corresponding error probability decay as $p \to +\infty$, we compute the limit of $R_{\gamma,\lambda}$ as $\lambda \to 0^+$ and show, using the Moore–Osgood theorem, that this indeed coincides with the rate for $\hat{T}_0$.

We first recall the following consequence of the Silverstein fixed-point equation.

Lemma B.8. *Let φ be a function satisfying the Silverstein equation* (11) *and suppose that Assumption 1 holds. Then*

1. *If $\gamma > 1$, we have $\lim_{\lambda\to 0^+}\lambda\varphi(-\lambda) = 0$.*
2. *If $\gamma \in (0,1)$ it holds $\lim_{\lambda\to 0^+}\varphi(-\lambda) = +\infty$ and $\lim_{\lambda\to 0^+}\lambda\varphi(-\lambda) = 1-\gamma$.*

We use this to establish the following result. We state the result with fixed false alarm level, but the same holds by switching the roles of P_{MD} and P_{FA}.

Proposition B.9. *Suppose that Assumption 1 holds. Then, in the ridgeless regime, as $p, n \to +\infty$ with $\gamma = p/n \in (0,1)$, we get, almost surely,*

$$\lim_{p\to+\infty}\frac{1}{p}\log P_{MD} = -R_{\gamma,0}\,, \quad R_{\gamma,0} := \frac{\mathsf{SNR}}{2}(1-\gamma) \quad \textit{for } \gamma \in (0,1)\,. \tag{53}$$

Proof. First, observe that the rate for the ridgeless decision statistic T_0 coincides with

$$R_{\gamma,0} := \lim_{p\to+\infty}\lim_{\lambda\to 0^+}\hat{d}^2/(2p)\,.$$

To characterize this, we compute the limit $\lim_{\lambda\to 0^+}\lim_{p\to+\infty}\hat{d}^2/(2p) = \lim_{\lambda\to 0^+}R_{\gamma,\lambda}$, and show, using the Moore–Osgood theorem, that we can exchange the two limits. From Lemma B.8 and Example B.3, we get $\kappa(\lambda) \to 0$, and

$$\lim_{\lambda\to 0^+}\lim_{p\to+\infty}\frac{\hat{\mu}}{p} = \lim_{\lambda\to 0^+}\frac{\kappa(\lambda)}{\lambda}\,\mathfrak{n}_1(\kappa(\lambda)) = \frac{\mathsf{SNR}}{1-\gamma}\,. \tag{54}$$

For all p large enough, the convergence of $\hat{\mu}/p$ to its limit as $\lambda \to 0^+$ is uniform in p since, using (46) and standard extreme-eigenvalue results for sample covariance matrices [3] for $\gamma \in (0,1)$, we get $\|\hat{\Sigma}_\lambda^{-1}\| \leq C^{-1}$, and $\|\hat{\Sigma}_\lambda^{-2}\| \leq C^{-2}$ almost surely for C independent on p and λ, and

$$\left|\frac{d}{d\lambda}\frac{\hat{\mu}}{p}\right| = \left|\frac{d}{d\lambda}\frac{1}{p}\hat{s}^T\hat{\Sigma}_\lambda^{-1}s\right| = \left|\frac{1}{p}\hat{s}^T\hat{\Sigma}_\lambda^{-2}s\right| \leq \frac{1}{C^2}\frac{1}{p}\|s\|^2 = \mathcal{O}(1)\,, \quad \text{as } p \to +\infty\,.$$

Let us now discuss the asymptotic behavior of the variance as $p \to +\infty$ in the ridgeless regime. In this case, we get, by dominated convergence theorem,

$$\lim_{\lambda\to 0^+}\lim_{p\to+\infty}\frac{\hat{\sigma}^2}{p} = \lim_{\lambda\to 0^+}\frac{\kappa(\lambda)^2}{\lambda^2}\frac{\mathfrak{n}_2(\kappa(\lambda))}{(1-\gamma\,\mathfrak{m}_2(\kappa(\lambda)))} = \frac{\mathsf{SNR}}{(1-\gamma)^3}\,. \tag{55}$$

Similarly to the step above, we have

$$\left|\frac{d}{d\lambda}\frac{\hat{\sigma}^2}{p}\right| = \frac{1}{p}\left|\hat{s}^T\left(\hat{\Sigma}_\lambda^{-2}\Sigma\hat{\Sigma}_\lambda^{-1} + \hat{\Sigma}_\lambda^{-1}\Sigma\hat{\Sigma}_\lambda^{-2}\right)\hat{s}\right| \leq \frac{2\|\Sigma\|}{C^3}\frac{\|\hat{s}\|^2}{p} = \mathcal{O}(1)\,, \quad \text{as } p \to +\infty\,.$$

Thus, the convergence of $\hat{\sigma}^2/p$ as $\lambda \to 0^+$ is uniform in p and the Moore–Osgood allows us to exchange the two limits in (55). Taking half the ratio between (54) and (55) yields the claim. □

B.4 Asymptotic values

The following result follows from a systematic inspection of the Silverstein equation in (12).

Lemma B.10. *Suppose that Assumption 1 holds. Let $\kappa(\lambda)$ be the unique solution to the Silverstein equation* (12)*. Then, for all $\lambda > 0$, the following limits hold:*

- $\kappa(\lambda) \to \lambda$ *as* $\gamma \to 0^+$*;*
- $\kappa(\lambda) \sim \lambda + \gamma\mathbb{E}_H(\xi)$ *as* $\gamma \to +\infty$*.*

Proof. First, we start with the limit in $\gamma \to 0^+$. Let $\gamma_h \to 0^+$ as $h \to +\infty$ and κ_h the corresponding $\kappa(\lambda)$. Then, from the Silverstein equation (12) we get that, for some $\bar{\gamma} \geq \gamma_h$ for all $h \in \mathbb{N}$,

$$\kappa_h \leq \lambda + \bar{\gamma}\mathbb{E}_H(\xi) < +\infty\,. \tag{56}$$

Thus, $(\kappa_h)_h$ is bounded. Passing to a subsequence if necessary, we shall assume that $\kappa_h \to \bar{\kappa} \geq 0$. Passing to the limit in the Silverstein equation (12) we get

$$\bar{\kappa} = \lim_{h\to+\infty}\kappa_h = \lambda + \lim_{h\to+\infty}\gamma_h\int_{\mathbb{R}_+}\frac{\kappa_h\xi}{\kappa_h+\xi}\,\mathrm{d}H = \lambda\,. \tag{57}$$

Let us now show the asymptotic behavior for $\gamma \to +\infty$. Suppose now that $\gamma_h \to +\infty$. It is easy to see that assuming that κ_h has a cluster point leads to an immediate contradiction (pass to the limit in the Silverstein equation as above). Then, it must be that $\kappa_h \to +\infty$ as $h \to +\infty$. Now, from the Silverstein equation we get

$$\frac{\kappa_h}{\gamma_h} = \frac{\lambda}{\gamma_h} + \int_{\mathbb{R}_+}\frac{\kappa_h\xi}{\kappa_h+\xi}\,\mathrm{d}H \sim \frac{\lambda}{\gamma_h} + \mathbb{E}_H(\xi)\,, \quad \text{as } h \to +\infty\,, \tag{58}$$

which leads to the claim. □

From Lemma B.10 we find that the rate (15) for $\gamma \to 0$ and $\gamma \to +\infty$ has the following limits

$$\lim_{\gamma\to 0^+} R_{\gamma,\lambda} = \frac{1}{2}\frac{\mathbb{E}_{\mathsf{dSNR}\,\mathrm{d}H}\left[f_\lambda(\xi)\right]^2}{\mathbb{E}_{\mathsf{dSNR}\,\mathrm{d}H}\left[f_\lambda(\xi)^2\right]}\,, \quad \text{and} \quad \lim_{\gamma\to+\infty} R_{\gamma,\lambda} = \frac{1}{2}\frac{\mathbb{E}_{\mathsf{dSNR}\,\mathrm{d}H}\left[\xi\right]^2}{\mathbb{E}_{\mathsf{dSNR}\,\mathrm{d}H}\left[\xi^2\right]}\,. \tag{59}$$

where $f_\lambda(\xi) := \frac{\xi}{\xi+\lambda}$ for all $\xi \geq 0$. We are now ready to prove Proposition 2.3.

Proof of Proposition 2.3. Let us denote by $\mathsf{s}H := \mathsf{dSNR}\,\mathrm{d}H$ and by $R_{0,\lambda} := \lim_{\gamma\to 0^+} R_{\gamma,\lambda}$, i.e., the right-hand side of the first limit in (59). Take the derivative in λ and differentiate under the integral sign to obtain

$$\begin{aligned}\frac{d}{d\lambda}R_{0,\lambda} &= \frac{1}{2}\frac{2\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)\big]\mathbb{E}_{\mathsf{s}H}\big[\frac{d}{d\lambda}f_\lambda(\xi)\big]\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)^2\big] - \mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)\big]^2\mathbb{E}_{\mathsf{s}H}\big[2f_\lambda(\xi)\frac{d}{d\lambda}f_\lambda(\xi)\big]}{\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)^2\big]^2}\\ &= -\frac{\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)\big]}{\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)^2\big]^2}\Big(\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)\tfrac{1}{\xi+\lambda}\big]\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)^2\big] - \mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)\big]\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)^2\tfrac{1}{\xi+\lambda}\big]\Big)\,.\end{aligned}$$

Now, let us introduce the auxiliary probability measure

$$\mathbb{Q} := \frac{f_\lambda(\xi)^2}{\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)^2\big]}\,\mathsf{dSNR}(\xi)\,\mathrm{d}H(\xi)\,,$$

and denote by $C_\xi := \frac{1}{\xi+\lambda}$ for brevity. With this notation, we can write:

$$\begin{aligned}\frac{d}{d\lambda}R_{0,\lambda} &= -\frac{\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)\big]}{\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)^2\big]}\Big(\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)C_\xi\big] - \mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)\big]\frac{\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)^2C_\xi\big]}{\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)^2\big]}\Big)\\ &= -\frac{\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)\big]}{\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)^2\big]}\Big(\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)^2\big]\mathbb{E}_{\mathbb{Q}}\big[C_\xi/f_\lambda(\xi)\big] - \mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)\big]\mathbb{E}_{\mathbb{Q}}\big[C_\xi\big]\Big)\\ &= -\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)\big]\Big(\mathbb{E}_{\mathbb{Q}}\big[C_\xi/f_\lambda(\xi)\big] - \frac{\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)\big]}{\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)^2\big]}\mathbb{E}_{\mathbb{Q}}\big[C_\xi\big]\Big)\\ &= -\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)\big]\Big(\mathbb{E}_{\mathbb{Q}}\big[C_\xi/f_\lambda(\xi)\big] - \mathbb{E}_{\mathbb{Q}}\big[1/f_\lambda(\xi)\big]\mathbb{E}_{\mathbb{Q}}\big[C_\xi\big]\Big)\,.\end{aligned}$$

This yields that for all $\lambda > 0$

$$\frac{d}{d\lambda}R_{0,\lambda} = -\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)\big]\,\mathrm{Cov}_{\mathbb{Q}}\big(1/f_\lambda(\xi),\,C_\xi\big)\,. \tag{60}$$

Since $\mathbb{E}_{\mathsf{s}H}\big[f_\lambda(\xi)\big] \geq 0$ due to the fact that H is supported on $\mathbb{R}_+$, and $1/f_\lambda(\xi)$ and C_ξ are both decreasing in ξ, by Chebyshev's association inequality, we deduce that $\frac{d}{d\lambda}R_{0,\lambda} \leq 0$. In particular, $R_{0,\lambda}$ is monotonically non-increasing in λ. Thus, for each $\lambda > 0$,

$$\lim_{\gamma\to 0^+} R_{\gamma,\lambda} = R_{0,\lambda} \geq \lim_{\lambda\to+\infty} R_{0,\lambda} = \frac{1}{2}\frac{\mathbb{E}_{\mathsf{s}H}\big[\xi\big]^2}{\mathbb{E}_{\mathsf{s}H}\big[\xi^2\big]} = \lim_{\gamma\to+\infty} R_{\gamma,\lambda}\,, \tag{61}$$

where the second identity follows immediately by inspection and the first inequality is indeed an equality if and only if $\mathsf{s}H$ is supported on a single eigenvalue. □

B.5 Linear decision statistic with miss-specified mean

In this section, we assume that the signal in the training set, denoted by s, does not correspond to the signal in the test set, denoted by s—we say that it is *miss-specified*. We thus consider $\hat{T}(x) = \hat{\mathsf{s}}^T\hat{\Sigma}_\lambda^{-1}x$, for which

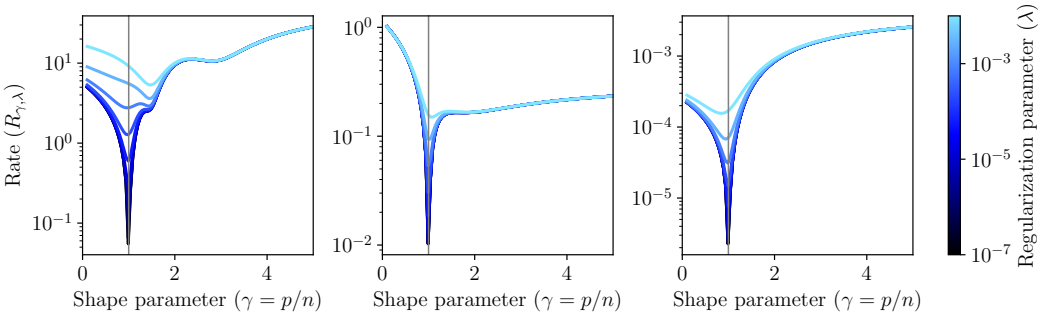
$$\begin{aligned}\hat{\mu} &= \mathbb{E}\left[\hat{T}(x) \mid \hat{X}, \mathcal{H}_1\right] = \hat{\mathsf{s}}^T\hat{\Sigma}_\lambda^{-1}s = \mathrm{tr}\left[s\hat{\mathsf{s}}^T\hat{\Sigma}_\lambda^{-1}\right],\\ \hat{\sigma}^2 &= \mathbb{E}\left[(\hat{T}(x)-\hat{\mu})^2 \mid \hat{X}, \mathcal{H}_1\right] = \hat{\mathsf{s}}^T\hat{\Sigma}_\lambda^{-1}\Sigma\hat{\Sigma}_\lambda^{-1}\hat{\mathsf{s}} = \mathrm{tr}\left[\hat{\mathsf{s}}\hat{\mathsf{s}}^T\hat{\Sigma}_\lambda^{-1}\Sigma\hat{\Sigma}_\lambda^{-1}\right].\end{aligned} \tag{62}$$

Normalizing by p and passing to the limit as $p, n \to +\infty$ with $p/n = \gamma$ we can obtain, using the same arguments of Proposition 2.1, closed form expressions for the limits. Specifically, we need to adapt Assumption A2 to this specific setting, requiring, in addition to the fact that Assumption 1 holds for both s and s, the following coupling condition.

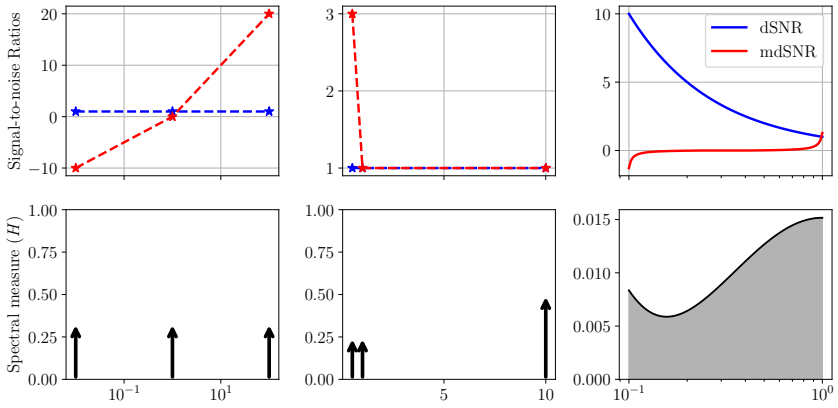
Assumption 2 (Mismatched means)**.** *The norm of the signals $s, \mathsf{s} \in \mathbb{R}^p$ grow as $\mathcal{O}(p^{1/2})$ as $p \to +\infty$ and there exists a signed measure $\nu_{s,\mathsf{s}}$ such that*

$$\frac{1}{p}\sum_{i=1}^{p}(v_i^T s)(v_i^T\mathsf{s})\delta_{\xi_i} \rightharpoonup \nu_{s,\mathsf{s}}\,, \quad \text{as } p \to +\infty\,, \tag{63}$$

where v_i is the eigenvector of Σ with unit norm associated to ξ_i. Additionally, $\nu_{s,\mathsf{s}} \ll H$ with Radon–Nikodym derivative $\xi \mapsto \xi\,\mathsf{mdSNR}(\xi)$.



(a) Performances as a function of γ.



(b) Geometric setting considered to produce the results in Figure 4a. First row: signal-to-noise ratios dSNR and its miss-specified variant mdSNR. Second row: Population spectral distribution $\mathrm{d}H$.

Figure 4: Rate for the linear decision statistic defined in (2) under miss-specification of the mean.

Similarly to Definition 1.1, we refer to $\mathsf{mdSNR}(\xi) = \frac{1}{\xi}\frac{\partial \nu_{s,\mathsf{s}}}{\partial H}(\xi)$ as the *mismatched directional signal-to-noise* ratio, which is a new not necessarily positive function that models the limiting behavior of $(v_i^T s)(v_i^T \mathsf{s})/\xi_i$. Obviously, the further away from dSNR, the more the model is mismatched.

Non-isotropic setting Using the same techniques as in the proof of Proposition 2.1 we can see that the asymptotic variance of $\hat{T}$ coincides with σ_∞^2 with dSNR computed with respect to s, while the mean reduces to

$$\begin{aligned}\frac{\hat{\mu}}{p} &= \operatorname{tr}\left[\frac{s\hat{\mathsf{s}}^T + \hat{\mathsf{s}}s^T}{2p}\hat{\Sigma}_\lambda^{-1}\right] = \frac{\kappa(\lambda)}{\lambda}\int_{\mathbb{R}_+}\frac{\xi}{\xi+\kappa(\lambda)}\frac{(v_i^T s)(v_i^T\hat{\mathsf{s}})}{\xi}\,\mathrm{d}H_p(\xi) + o(1)\\ &= \frac{\kappa(\lambda)}{\lambda}\int_{\mathbb{R}_+}\frac{\xi}{\xi+\kappa(\lambda)}\,\mathsf{mdSNR}(\xi)\,\mathrm{d}H(\xi) + o(1)\,,\end{aligned} \tag{64}$$

almost surely as $p \to +\infty$. Since μ_∞ is not necessarily positive in this setting, from the asymptotic values μ_∞ and σ_∞^2, we compute the rate as $R_{\gamma,\lambda} = \frac{1}{2}\max\{\mu_\infty, 0\}^2/\sigma_\infty^2$. We show in Figure 4 three notable examples. In these cases, we can see that, as soon as $s \neq \mathsf{s}$ and Σ is non-isotropic, the best rate can be achieved in the overparameterized regime.

Isotropic case In the isotropic case, mdSNR is defined only on σ^2 and thus can be identified with a number, which is $\mathsf{mSNR} := \lim_{p\to+\infty}\frac{1}{p}\mathsf{s}^T s/\sigma^2$, which exists by Assumption 2. Therefore, mean and variance now read as

$$\mu_\infty = \frac{\mathsf{mSNR}\,\kappa(\lambda)}{\lambda}\frac{\sigma^2}{\sigma^2+\kappa(\lambda)}\,, \quad \text{and} \quad \sigma_\infty^2 = \frac{\tilde{\mathsf{SNR}}\kappa(\lambda)^2}{\lambda^2}\frac{\sigma^4}{(\sigma^2+\kappa(\lambda))^2 - \gamma\sigma^4}\,, \tag{65}$$

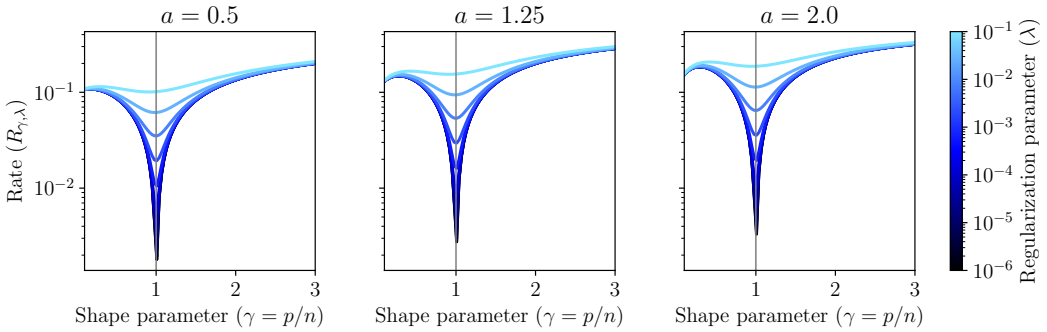


Figure 5: Isotropic case: Rate for the linear decision statistic defined in (2) assuming a miss-specified mean. The *artificial miss-specification* degree is polynomially decreasing in γ as in (68) with $c = 0.9$ and three different choices of a. Confer Appendix B.5 for details.

where $\tilde{\mathsf{SNR}} := \lim_{p\to+\infty} \frac{1}{p}\|\mathsf{s}\|^2/\sigma^2$, which exists by Assumption 1 for s. We deduce that the asymptotic rate has the same shape as in (50), but with a different constant

$$R_{\gamma,\lambda} = \frac{\mathsf{mSNR}^2}{2\tilde{\mathsf{SNR}}}\left(1 - \frac{\gamma\sigma^4}{(\sigma^2+\kappa(\lambda))^2}\right) = \mathcal{K}\frac{\mathsf{SNR}}{2}\left(1 - \frac{\gamma\sigma^4}{(\sigma^2+\kappa(\lambda))^2}\right), \tag{66}$$

where we have denoted by $\mathcal{K}$ the constant

$$\mathcal{K} := \frac{\mathsf{mSNR}^2}{\mathsf{SNR}\cdot\tilde{\mathsf{SNR}}} = \lim_{p\to+\infty}\left[\frac{\mathsf{s}^T s}{\|\mathsf{s}\|\|s\|}\right]^2 \in [0,1]\,. \tag{67}$$

First, observe that $\mathcal{K}$ coincides with the squared limiting *cosine similarity* between s and s, quantifying the degree of miss-specification. Secondly, compared to (50), we now have a loss in performances quantifiable by $\mathcal{K}$.

Following [25], if we assume that the degree of dissimilarity decays as the resolution of the model increases, e.g., supposing that $\mathcal{K}$ is a function of γ as

$$\mathcal{K}(\gamma) := 1 - \frac{c}{(1+\gamma)^a}\,, \quad \text{for all } \gamma > 0\,, \tag{68}$$

and two constants $a > 0$ and $c \in (0,1)$, we obtain the curves shown in Figure 5. In Figure 5, we observe a full *double-descent*-type behavior, where $R_{\gamma,\lambda}$ exhibits a local maximum in $(0,1)$ and increases again after $\gamma = 1$.

To conclude this section, let us make the population spectral distribution H and the dSNR of Figure 2 explicit in Figure 6.

B.6 Extension to the multi-hypotheses setting

Let us now consider the multi-hypotheses setting of (6) and the decision statistic

$$\hat{k}(x) := \underset{0\le i\le M-1}{\operatorname{argmax}}\ \hat{p}_i(x)\,, \quad \text{where} \quad \hat{p}(x) := \operatorname{Softmax}(\hat{\Omega}x + \hat{b})\,, \tag{69}$$

and $\hat{\Omega}_i := \hat{s}_i^T\hat{\Sigma}_\lambda^{-1}$, $\hat{b}_i := \log\pi_i - \frac{1}{2}\hat{s}_i^T\hat{\Sigma}_\lambda^{-1}\hat{s}_i$ for all $i = 0,\dots,M-1$.

Error probability decay Assume for simplicity that $\pi_i = 1/M$ for all $i = 0,\dots,M-1$. In this case, the probability of error for class $\mathcal{H}_k$ can be bounded as follows. Let $\hat{T}_i(x) := \hat{s}_i^T\hat{\Sigma}_\lambda^{-1}x + \hat{b}_i$ for

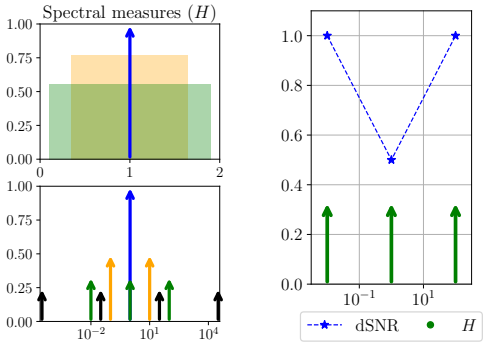


Figure 6: Population spectral distribution H and dSNR for the three examples shown in Figure 2. **Left:** Spectral measures for the two cases shown on the left of Figure 2. **Right:** Spectral measure for the case on the right of Figure 2. Vertical arrows represent Dirac masses, while shaded rectangles indicate uniform distributions. Color-coding is consistent with the corresponding curves in Figure 2.

all $i = 0, \dots, M-1$. Then, using the union bound and that $x \sim s_k + \Sigma^{1/2} z$ under $\mathcal{H}_k$,

$$\begin{aligned}
&\mathbb{P}(\hat{k}(x) \neq k \mid \hat{X}, \mathcal{H}_k) = \mathbb{P}(\text{There exists } j \neq k : \hat{T}_j(x) > \hat{T}_k(x) \mid \hat{X}, \mathcal{H}_k) \\
&= \mathbb{P}\Big(\bigcup_{j \neq k} \{\hat{T}_j(x) \geq \hat{T}_k(x)\} \mid \hat{X}, \mathcal{H}_k \Big) \leq \sum_{j \neq k} \mathbb{P}\big(\hat{T}_j(x) \geq \hat{T}_k(x) \mid \hat{X}, \mathcal{H}_k\big) \\
&= \sum_{j \neq k} \mathbb{P}\Big(\hat{s}_j^T \hat{\Sigma}_\lambda^{-1}(s_k + \Sigma^{1/2} z) - \frac{1}{2} \hat{s}_j^T \hat{\Sigma}_\lambda^{-1} \hat{s}_j \geq \hat{s}_k^T \hat{\Sigma}_\lambda^{-1}(s_k + \Sigma^{1/2} z) - \frac{1}{2} \hat{s}_k^T \hat{\Sigma}_\lambda^{-1} \hat{s}_k \mid \hat{X}, \mathcal{H}_k \Big) \\
&\leq \sum_{j \neq k} \mathbb{P}\Big(\hat{s}_{jk}^T \hat{\Sigma}_\lambda^{-1} \Sigma^{1/2} z \geq \frac{1}{2} \hat{s}_{jk}^T \hat{\Sigma}_\lambda^{-1} \hat{s}_{jk} + \hat{s}_{jk}^T \hat{\Sigma}_\lambda^{-1}(\hat{s}_k - s_k) \mid \hat{X}, \mathcal{H}_k \Big),
\end{aligned}$$

where $\hat{s}_{jk} := \hat{s}_j - \hat{s}_k$ for all $j \neq k$. Using now that $\frac{1}{p} \hat{s}_{jk}^T \hat{\Sigma}_\lambda^{-1}(\hat{s}_k - s_k) \to 0$ almost surely, and the same bounds as in Section 2 with $\hat{s}$ replaced by $\hat{s}_{jk}$, we get as $p \to +\infty$

$$\begin{aligned}
&\mathbb{P}(\hat{k}(x) \neq k \mid \hat{X}, \mathcal{H}_k) \\
&\leq \sum_{j \neq k} \mathbb{P}\Big(\frac{1}{p} \hat{s}_{jk}^T \hat{\Sigma}_\lambda^{-1} \Sigma^{1/2} z \geq \frac{1}{2p} \hat{s}_{jk}^T \hat{\Sigma}_\lambda^{-1} \hat{s}_{jk} + \frac{1}{p} \hat{s}_{jk}^T \hat{\Sigma}_\lambda^{-1}(\hat{s}_k - s_k) \mid \hat{X}, \mathcal{H}_k \Big) \\
&= \sum_{j \neq k} \exp\Big\{ -\frac{p}{2} \Big(\frac{\mu_{jk,\infty}}{2\sigma_{jk,\infty}} \Big)^2 + o(p) \Big\} \lesssim \exp\Big\{ -p \Big(R_{\gamma,\lambda}^{(k)} + o(1) \Big) \Big\},
\end{aligned}$$

where we have denoted by

$$R_{\gamma,\lambda}^{(k)} := \frac{1}{8} \min_{j \neq k} \Big(\frac{\mu_{jk,\infty}}{\sigma_{jk,\infty}} \Big)^2, \tag{70}$$

and $\mu_{jk,\infty}$ and $\sigma_{jk,\infty}^2$ are the almost sure limits of mean and variance (normalized by p) of the decision statistic $\hat{T}_{jk}(x) := \hat{s}_{jk}^T \hat{\Sigma}_\lambda^{-1} x$ with $x \sim \mathcal{N}(0, \Sigma)$ under $\mathcal{H}_0$ and $x \sim \mathcal{N}(s_{jk}, \Sigma)$ under $\mathcal{H}_1$.

Comparison with baseline We evaluate the classification performance of the proposed multi-hypothesis decision rule against a fully connected network (FCN) baseline. Using the entire FireRisk dataset with 7 distinct classes, we first perform feature extraction using the DINOv3 model. To investigate model behavior across varying complexity levels, we systematically downsample the training set to span shape parameter values $\gamma = p/n$ from 0.1 to 4.0 to cover both the underparameterized and overparameterized regimes.

The FCN baseline consists of a single linear layer with softmax output, equivalent to multi-class logistic regression. The proposed decision rule is based on a multi-class Gaussian, and consists of a linear layer (different weight compared to the FCN) and softmax output. The decision for both the classifiers is the MAP rule, which selects the hypothesis with the largest posterior probability.

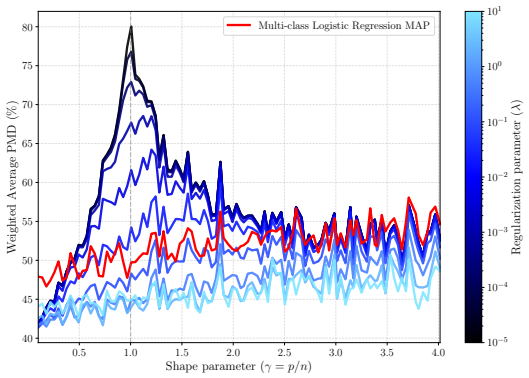


Figure 7: Classification performance on multi-hypotheses setting. We compare the weighted probability of miss detection of the proposed multi-class Gaussian MAP framework against the multi-class logistic regression baseline as a function of the shape parameter $\gamma = p/n$ on the FireRisk dataset ($M = 7$). The Gaussian MAP framework with regularization parameters $\lambda \in \{0.1, 10, 100\}$ (bold solid lines) achieves significantly lower error rates. The results demonstrate the robustness of the multi-hypothesis approach compared to standard linear baseline.

The FCN is trained with stochastic gradient descent (SGD) and a weighted cross-entropy loss to address class imbalance. Training is performed in full-batch mode for 300 epochs with a learning rate of 0.01 and ridge regularization ($\lambda_{L2} = 10^{-3}$). We observed that the FCN performance was insensitive to the choice of λ_{L2} within the tested range ($10^{-5}, 10^{-3}, 10^{-1}$).

Performance is evaluated using the weighted probability of miss-classification, which accounts for class imbalances by weighting the error rate of each class by its respective prior probability We show the results in Figure 7. Much like in Figure 3b, compared to the baseline, our approach exhibits comparable or better performances across all values of $\gamma > 0$ yet requiring essentially no fine-tuning.

C Missing proofs and further insights on Section 3

In this section, we consider the case in which no $\mathcal{H}_1$ data is available in the training set, and we aim solving (1) by means of the decision statistic defined in (3).

C.1 High-dimensional mean and variance

The following classical identities for Gaussian quadratic forms can be found in [42, Chapter 2].

Lemma C.1. *Let $x \sim \mathcal{N}(\mu, \Sigma)$ where $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p\times p}$ is symmetric positive definite. Let $A \in \mathbb{R}^{p\times p}$ be a symmetric positive semidefinite matrix. Then*

$$\mathbb{E}[x^T Ax] = \operatorname{tr}(A\Sigma) + \mu^T A\mu\,, \quad \textit{and} \quad \operatorname{Var}(x^T Ax) = 2\,\operatorname{tr}(A\Sigma A\Sigma) + 4\mu^T A\Sigma A\mu\,. \tag{71}$$

We employ these formulas to establish Proposition 3.1 below:

Proof of Proposition 3.1. Using Lemma C.1, we obtain

$$\hat{\mu}_{0,1} = \operatorname{tr}(\hat{\Sigma}_\lambda^{-1}\Sigma) + \mu_{0,1}^T \hat{\Sigma}_\lambda^{-1}\mu_{0,1}\,, \quad \text{and} \quad \hat{\sigma}_{0,1}^2 = 2\,\operatorname{tr}(\hat{\Sigma}_\lambda^{-1}\Sigma\hat{\Sigma}_\lambda^{-1}\Sigma) + 4\mu_{0,1}^T \hat{\Sigma}_\lambda^{-1}\Sigma\hat{\Sigma}_\lambda^{-1}\mu_{0,1}\,, \tag{72}$$

with $\mu_0 = 0$ under $\mathcal{H}_0$ and $\mu_1 = s$ under $\mathcal{H}_1$. Arguing as in Proposition 2.1 and using systematically Fact B.1, Theorem A.1 yields the following. For the mean under $\mathcal{H}_0$:

$$\frac{\hat{\mu}_0}{p} \sim \frac{\kappa(\lambda)}{\lambda}\frac{1}{p}\operatorname{tr}\left[\Sigma\big(\Sigma + \kappa(\lambda)I\big)^{-1}\right] \sim \frac{\kappa(\lambda)}{\lambda}\,\mathfrak{m}_1(\kappa(\lambda))\,. \tag{73}$$

For the variance under $\mathcal{H}_0$, instead, we get

$$\begin{aligned}\frac{\hat{\sigma}_0^2}{p} &\sim \frac{2\kappa(\lambda)^2}{\lambda^2}\frac{1}{p}\operatorname{tr}\left[\Sigma\big(\Sigma+\kappa(\lambda)I\big)^{-1}\Sigma\big(\Sigma+\kappa(\lambda)I\big)^{-1}\right] \\ &\quad+\frac{2\kappa(\lambda)^2}{\lambda^2}\frac{1}{p}\operatorname{tr}\left[\Sigma\big(\Sigma+\kappa(\lambda)I\big)^{-2}\Sigma\right]^2\frac{1}{n-\mathrm{d}f_2(\kappa(\lambda))} \\ &\sim \frac{2\kappa(\lambda)^2}{\lambda^2}\,\mathfrak{m}_2(\kappa(\lambda))+\frac{2\kappa(\lambda)^2}{\lambda^2}\,\mathfrak{m}_2(\kappa(\lambda))^2\frac{p}{n-\operatorname{tr}\left(\Sigma^2(\Sigma+\kappa(\lambda))^{-2}\right)} \\ &\sim \frac{2\kappa(\lambda)^2}{\lambda^2}\,\mathfrak{m}_2(\kappa(\lambda))+\frac{2\kappa(\lambda)^2}{\lambda^2}\,\mathfrak{m}_2(\kappa(\lambda))^2\frac{\gamma}{1-\gamma\,\mathfrak{m}_2(\kappa(\lambda))}\,,\end{aligned} \tag{74}$$

which eventually yields (18). While under $\mathcal{H}_1$, applying once again Theorem A.1, we get

$$\frac{\hat{\mu}_1}{p}\sim\frac{\hat{\mu}_0}{p}+\frac{\kappa(\lambda)}{\lambda}\operatorname{tr}\left[\frac{ss^T}{p}\big(\Sigma+\kappa(\lambda)I\big)^{-1}\right]\sim\frac{\kappa(\lambda)}{\lambda}\,\mathfrak{m}_1(\kappa(\lambda))+\frac{\kappa(\lambda)}{\lambda}\,\mathfrak{n}_1(\kappa(\lambda))\,. \tag{75}$$

Regarding the variance, we get,

$$\begin{aligned}\frac{\hat{\sigma}_1^2}{p} &\sim \frac{\hat{\sigma}_0^2}{p}+\frac{4\kappa(\lambda)^2}{\lambda^2}\operatorname{tr}\left[\frac{ss^T}{p}\big(\Sigma+\kappa(\lambda)I\big)^{-1}\Sigma\big(\Sigma+\kappa(\lambda)I\big)^{-1}\right] \\ &\quad+\frac{4\kappa(\lambda)^2}{\lambda^2}\operatorname{tr}\left[\frac{ss^T}{p}\big(\Sigma+\kappa(\lambda)I\big)^{-2}\Sigma\right]\operatorname{tr}\left[\Sigma\big(\Sigma+\kappa(\lambda)I\big)^{-2}\Sigma\right]\frac{1}{n-\mathrm{d}f_2(\kappa(\lambda))} \\ &\sim \frac{\hat{\sigma}_0^2}{p}+\frac{4\kappa(\lambda)^2}{\lambda^2}\,\mathfrak{n}_2(\kappa(\lambda))+\frac{4\kappa(\lambda)^2}{\lambda^2}\,\mathfrak{n}_2(\kappa(\lambda))\,\mathfrak{m}_2(\kappa(\lambda))\frac{p}{n-\operatorname{tr}\left(\Sigma^2(\Sigma+\kappa(\lambda))^{-2}\right)} \\ &\sim \frac{\hat{\sigma}_0^2}{p}+\frac{4\kappa(\lambda)^2}{\lambda^2}\,\mathfrak{n}_2(\kappa(\lambda))+\frac{4\kappa(\lambda)^2}{\lambda^2}\,\mathfrak{n}_2(\kappa(\lambda))\,\mathfrak{m}_2(\kappa(\lambda))\frac{\gamma}{1-\gamma\,\mathfrak{m}_2(\kappa(\lambda))}\,.\end{aligned} \tag{76}$$

Rearranging everything and plugging in the limits in the $\mathcal{H}_0$ case, we obtain the claim. □

C.2 Rate in the LDP sense

Since $\hat{T}$ is not Gaussian in this quadratic setting, the performances of our detection mechanism cannot be explained simply by the ratio between asymptotic mean and asymptotic variance. We shall resort on the Gärtner–Ellis Theorem [22, 21]. See Appendix A.3 for non-exhaustive introduction to the topic, which nonetheless contains the most important tools for our analysis.

C.2.1 Computing the limiting LMGF

We first discuss the $\mathcal{H}_0$ case, as the $\mathcal{H}_1$ case can be deduced from the former. In this setting, it turns out that the ratio between regularization and noise plays a central role in determining the performances in the sense of LDP. We introduce the measure dRNR as the push-forward of H under $f(x)=\frac{\lambda}{x}$, which, in relation to Assumption 1, can be understood as the limiting spectral distribution of $\lambda\Sigma^{-1}$. Clearly in the isotropic case, $\mathsf{dRNR}=\delta_{\lambda/\sigma^2}$, in which case we denote $\mathsf{RNR}:=\lambda/\sigma^2$.

The $\mathcal{H}_0$ Case Our objective is to identify the Fenchel–Legendre conjugate of $\Lambda^{(0)}$ according to (19), which requires studying the (possibly infinite) almost sure limit of

$$\hat{\Lambda}^{(0)}(\tau):=\frac{1}{p}\log\mathbb{E}\big[\exp\{\tau\hat{T}(x)\}\mid\hat{X},\mathcal{H}_0\big]\,,\quad\tau\in\mathbb{R}\,. \tag{77}$$

In contrast to the linear setting of Section 2, to obtain a closed form solution for the rate in the sense of the LDP, we need a *no-spectral-outliers* assumption on the *whitened decision matrix*

$$\hat{W}^{-1}:=\Sigma^{-\frac{1}{2}}(\hat{\Sigma}+\lambda I)\Sigma^{-\frac{1}{2}}=\tfrac{1}{n}\hat{Z}\hat{Z}^T+\lambda\Sigma^{-1}\,, \tag{78}$$

where we have denoted $\hat{Z}:=\Sigma^{-\frac{1}{2}}\hat{X}$, which is a collection of i.i.d. random variables with zero mean and unit variance.

Assumption 3 (No spectral outliers)**.** *Let $\hat{W}^{-1}$ be defined as in* (78)*. Assume that its empirical spectral distribution $\hat{\nu}_{\gamma,\lambda}$ converges weakly to a probability distribution $\nu_{\gamma,\lambda}$ with*

$$\lambda_{\min}(\hat{W}^{-1}) \to T := \inf \operatorname{supp} \nu_{\gamma,\lambda}\,, \quad \text{a.s. as } p \to +\infty\,. \tag{79}$$

Note that Assumption 3 holds for isotropic data. In terms of the underlying population geometry, (79) is strictly linked to $\|\Sigma\| \to \sup \operatorname{supp} H$. We formulate it as is to simplify the derivation of the following:

Lemma C.2. *Suppose that Assumptions 1 and 3 hold. Then, for every $\tau \neq T/2$ the conditional almost-sure limit defining $\Lambda^{(0)}(\tau)$ exists and is given by:*

$$\Lambda^{(0)}(\tau) = -\frac{1}{2}\int_{\mathbb{R}_+} \log\left(1 - \frac{2\tau}{\xi}\right) \mathrm{d}\nu_{\gamma,\lambda}(\xi)\,, \quad \text{for all } \tau < \frac{T}{2}\,, \tag{80}$$

and $\Lambda^{(0)}(\tau) = +\infty$ for $\tau > \frac{T}{2}$, where $T := \inf \operatorname{supp} \nu_{\gamma,\lambda}$, $\nu_{\gamma,\lambda}$ is the probability measure whose Stieltjes transform $\mathfrak{s}\colon \mathbb{C}_+ \to \mathbb{C}_+$ is the unique solution to the following fixed-point equation:

$$\mathfrak{s}(z) = \mathfrak{s}_{\mathsf{dRNR}}\left(z - \frac{1}{1+\gamma\mathfrak{s}(z)}\right), \quad z \in \mathbb{C}_+\,, \tag{81}$$

and $\mathfrak{s}_{\mathsf{dRNR}}$ is the Stieltjes transform of dRNR*.*

Proof. Using the definition of $\hat{T}$, we get, under $\mathcal{H}_0$,

$$\hat{T}(x) = x^T\hat{\Sigma}_\lambda^{-1}x = x^T\Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}\hat{\Sigma}_\lambda^{-1}\Sigma^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}x = z^T\hat{W}z\,,$$

Therefore, denoting by $\tilde{v} = Oz$, where O is the (orthogonal) diagonalizing matrix of $\hat{W}$, we shall write, using that under $\mathcal{H}_0$, the components $|\tilde{v}_i|^2$ are i.i.d. χ^2 variables with LMGF $t \mapsto -\frac{1}{2}\log(1-2t)$ for all $t < \frac{1}{2}$,

$$\begin{aligned}\hat{\Lambda}^{(0)}(\tau) &= \frac{1}{p}\log\mathbb{E}\big[\exp\{\tau z^T\hat{W}z\} \mid \hat{X}, \mathcal{H}_0\big] = \frac{1}{p}\log\mathbb{E}\left[\exp\left\{\sum_{i=1}^p \frac{\tau}{\xi_i}|\tilde{v}_i|^2\right\}\middle|\hat{X}, \mathcal{H}_0\right] \\ &= \frac{1}{p}\sum_{i=1}^p \log\mathbb{E}\Big[\exp\Big\{\frac{\tau}{\xi_i}|\tilde{v}_i|^2\Big\}\Big|\hat{X}, \mathcal{H}_0\Big] = -\frac{1}{2}\int \log\left(1 - \frac{2\tau}{\xi}\right)\mathrm{d}\hat{\nu}_{\gamma,\lambda}(\xi)\,,\end{aligned} \tag{82}$$

where $\hat{\nu}_{\gamma,\lambda}$ is the spectral empirical distribution of $\hat{W}^{-1}$. We can thus invoke [44, Theorem 1.1], which yields the weak convergence of $\hat{\nu}_{\gamma,\lambda}$ to the measure $\nu_{\gamma,\lambda}$ whose Stieltjes transform is the unique solution to (81). Under Assumption 3, for each $\tau < T/2$, the empirical spectral measures $\hat{\nu}_{\gamma,\lambda}$ are almost surely eventually supported uniformly away from the singularity of the integrand, allowing us to pass to the limit in (82) as $p \to +\infty$. Conversely, for each $\tau > T/2$, the corresponding finite-dimensional LMGF is almost surely eventually equal to $+\infty$. □

Remark C.3 (Boundary case)**.** *We assume that $\hat{\Lambda}^{(0)}(T/2)$ converges almost surely in the extended real line. Its limiting value depends on the spectral behavior at the boundary and must be determined on a case-by-case basis. However, this value does not affect the remainder of our analysis.*

We have obtained a second fixed-point equation (81), that this time characterizes the limiting measure $\nu_{\gamma,\lambda}$. Note that $\mathfrak{s}$ is defined on $\mathbb{C}_+$, but can be obviously extended by continuity to $\mathbb{R} \setminus \operatorname{supp}\nu_{\gamma,\lambda}$, where by dominated convergence theorem still admits the integral representation:

$$\mathfrak{s}(z) = \int_{\mathbb{R}_+} \frac{1}{\xi - z}\,\mathrm{d}\nu_{\gamma,\lambda}(\xi)\,, \quad \text{for all } z < T\,. \tag{83}$$

For notational convenience, we do not rename $\mathfrak{s}$.

Although the integral in (80) is somewhat challenging to compute, in practice, we can resort to an integration trick proposed by Silverstein and Bai, see e.g., [4]. It consist in writing the integrand as an integral in ξ and applying Fubini's theorem.

Corollary C.4. *Let $\mathfrak{s}\colon \mathbb{C}_+ \to \mathbb{C}_+$ be the unique solution to* (81)*. Then,*

$$\Lambda^{(0)}(\tau) = \frac{1}{2}\int_0^{2\tau} \mathfrak{s}(z)\,\mathrm{d}z\,, \quad \textit{for all } \tau < \frac{T}{2}\,. \tag{84}$$

Proof. Observe that

$$-\frac{1}{2}\log\left(1-\frac{2\tau}{\xi}\right) = \frac{1}{2}\int_0^{2\tau}\frac{1}{\xi - z}\,\mathrm{d}z\,, \quad \text{for all } \tau < \frac{\xi}{2}\,, \tag{85}$$

and is $+\infty$ otherwise. We now plug this expression into (80) and apply Fubini's theorem, which yields for all $\tau < \frac{T}{2}$

$$\Lambda^{(0)}(\tau) = \frac{1}{2}\int_{\mathbb{R}_+}\int_0^{2\tau}\frac{1}{\xi - z}\,\mathrm{d}z\,\mathrm{d}\nu_{\gamma,\lambda}(\xi) = \frac{1}{2}\int_0^{2\tau}\int_{\mathbb{R}_+}\frac{1}{\xi - z}\,\mathrm{d}\nu_{\gamma,\lambda}(\xi)\,\mathrm{d}z\,. \tag{86}$$

Recalling (83) we get the claim. □

Since the fixed-point equation (81) can be solved efficiently, this expression can be leveraged to compute $\Lambda^{(0)}$ with high accuracy, yet not in closed-form in general.

The $\mathcal{H}_1$ Case Computing the limiting LMGF under the $\mathcal{H}_1$ hypothesis is more delicate. We have the following result:

Lemma C.5. *Suppose that Assumptions 1 and 3 hold and let b be a uniform upper bound for $\|\Sigma\|$ given by Assumption A1. Then, almost surely,*

$$\Lambda^{(1)}(\tau) = \Lambda^{(0)}(\tau) + \tau\Delta(\tau)\,, \quad \textit{for all } \tau < \tfrac{1}{2}\lambda/b\,, \tag{87}$$

where $\Lambda^{(0)}$ is defined as in (80) *and $\Delta(\tau)$ is defined by*

$$\Delta(\tau) := \frac{\kappa_\tau(\lambda)}{\lambda}\int_{\mathbb{R}_+}\frac{\xi}{\kappa_\tau(\lambda) + (1 - 2\tau\kappa_\tau(\lambda)/\lambda)\xi}\,\mathsf{dSNR}(\xi)\,\mathrm{d}H(\xi)\,, \tag{88}$$

where $\kappa_\tau(\lambda)$ is the unique solution to

$$\kappa_\tau(\lambda) = \lambda + \gamma\kappa_\tau(\lambda)\int_{\mathbb{R}_+}\frac{\xi}{\kappa_\tau(\lambda) + (1 - 2\tau\kappa_\tau(\lambda)/\lambda)\xi}\,\mathrm{d}H(\xi)\,. \tag{89}$$

Proof. Let $\hat{\xi}_1, \ldots, \hat{\xi}_p$ the eigenvalues of $\hat{W}^{-1}$ and $\hat{\xi}_{\min}$ the minimum. First, note that from (82),

$$\hat{\Lambda}^{(0)}(\tau) = -\frac{1}{2p}\log\det(I - 2\tau\hat{W})\,, \tag{90}$$

for all $\tau < \hat{\xi}_{\min}/2$ and $\hat{\Lambda}^{(0)}(\tau) = +\infty$ otherwise. Under $\mathcal{H}_1$, we have

$$\begin{aligned}
\hat{\Lambda}^{(1)}(\tau) &= \frac{1}{p}\log\mathbb{E}\big[\exp\{\tau x^T\hat{\Sigma}_\lambda^{-1}x\} \mid \hat{X}, \mathcal{H}_1\big] \\
&= \frac{1}{p}\log\mathbb{E}\big[\exp\{\tau(x-s)^T\hat{\Sigma}_\lambda^{-1}(x-s)\}\exp\{2\tau(x-s)^T\hat{\Sigma}_\lambda^{-1}s\}\exp\{\tau s^T\hat{\Sigma}_\lambda^{-1}s\} \mid \hat{X}, \mathcal{H}_1\big] \\
&= \frac{1}{p}\log\mathbb{E}\big[\exp\{\tau(x-s)^T\hat{\Sigma}_\lambda^{-1}(x-s)\}\exp\{2\tau(x-s)^T\hat{\Sigma}_\lambda^{-1}s\} \mid \hat{X}, \mathcal{H}_1\big] + \frac{\tau}{p}s^T\hat{\Sigma}_\lambda^{-1}s\,.
\end{aligned}$$

We apply Gaussian integration, see, e.g., [10, Eq. (2.84)], which gives for $\tau < \hat{\xi}_{\min}/2$ that

$$\begin{aligned}
\hat{\Lambda}^{(1)}(\tau) &= \hat{\Lambda}^{(0)}(\tau) + \frac{\tau}{p}\,s^T\hat{\Sigma}_\lambda^{-1}s + \frac{2\tau^2}{p}\,s^T\hat{\Sigma}_\lambda^{-1}\Sigma^{1/2}\big(I - 2\tau\hat{W}\big)^{-1}\Sigma^{1/2}\hat{\Sigma}_\lambda^{-1}s \\
&= \hat{\Lambda}^{(0)}(\tau) + \frac{\tau}{p}\,s^T\big(\hat{\Sigma}_\lambda - 2\tau\Sigma\big)^{-1}s\,,
\end{aligned} \tag{91}$$

and $+\infty$ otherwise.

To pass to the limit, we restrict ourselves to $\tau < \tau_0 := \frac{1}{2}\lambda/b$ and turn the quadratic form in the format of Theorem A.1. First, we define $\Sigma_\tau := \big(\Sigma^{-1} - \frac{2\tau}{\lambda} I\big)^{-1}$, $\hat{\Sigma}_\tau := \Sigma_\tau^{1/2}\big(\frac{1}{n}\hat{Z}\hat{Z}^T\big)\Sigma_\tau^{1/2}$, and $s_\tau := \Sigma_\tau^{1/2}\Sigma^{-1/2}s$. Now, we have

$$\begin{aligned}\hat{\Sigma}_\lambda - 2\tau\Sigma &= \Sigma^{1/2}\big(\hat{W}^{-1} - 2\tau I\big)\Sigma^{1/2} \\ &= \Sigma^{1/2}\Big(\frac{1}{n}\hat{Z}\hat{Z}^T + \lambda\Sigma_\tau^{-1}\Big)\Sigma^{1/2} = \Sigma^{1/2}\Sigma_\tau^{-1/2}\big(\hat{\Sigma}_\tau + \lambda I\big)\Sigma_\tau^{-1/2}\Sigma^{1/2}\,.\end{aligned} \tag{92}$$

Therefore, the quadratic form in the right-hand side of (91) reads as

$$\begin{aligned}\frac{\tau}{p}s^T\big(\hat{\Sigma}_\lambda - 2\tau\Sigma\big)^{-1}s &= \frac{\tau}{p}s^T\Sigma^{-1/2}\Sigma_\tau^{1/2}\big(\hat{\Sigma}_\tau + \lambda I\big)^{-1}\Sigma_\tau^{1/2}\Sigma^{-1/2}s \\ &= \frac{\tau}{p}s_\tau^T\big(\hat{\Sigma}_\tau + \lambda I\big)^{-1}s_\tau\,.\end{aligned} \tag{93}$$

Observe that $\hat{\Sigma}_\tau$ is sample covariance matrix of the training set $\hat{X}_\tau := \Sigma_\tau^{1/2}\hat{Z}$ with population covariance matrix Σ_τ, and s_τ is a deterministic vector. For $\tau < \tau_0$, we are in the setting of Theorem A.1, which yields as $p \to +\infty$, almost surely,

$$\frac{\tau}{p}s_\tau^T\big(\hat{\Sigma}_\tau + \lambda I\big)^{-1}s_\tau = \tau\,\mathrm{tr}\Big[\frac{s_\tau s_\tau^T}{p}\big(\hat{\Sigma}_\tau + \lambda I\big)^{-1}\Big] \sim \frac{\tau\kappa_\tau(\lambda)}{\lambda}\,\mathrm{tr}\Big[\frac{s_\tau s_\tau^T}{p}\big(\Sigma_\tau + \kappa_\tau(\lambda)I\big)^{-1}\Big]\,, \tag{94}$$

where $\kappa_\tau(\lambda)$ is the unique solution to the corresponding Silverstein equation (89). Passing to the limit, after omitted elementary manipulations to write the limit with respect to the original problem geometry, we get as $p \to +\infty$, that the quadratic form in (94) converges almost surely to $\tau\Delta(\tau)$, which yields the claim. □

Note that, while (91) is exact and holds for all $\tau < \hat{\xi}_{\min}/2$, the limit in Lemma C.5 is established only for $\tau < \tau_0 := \lambda/(2b)$. Although these two bounds may coincide, e.g., for isotropic data with $\gamma > 1$, extending the argument beyond τ_0 would make Σ_τ indefinite (or even singular), so that it can no longer define a covariance matrix. Nevertheless, as we show in Appendix C.2.3, this result is sufficient to establish the rate function in the Neyman–Pearson setup.

C.2.2 Computing the rate function

According to Gärtner–Ellis, Theorem A.4, as soon as $0 \in \mathrm{Int}\,\mathrm{Dom}\,\Lambda^{(0)}$, the rate in the sense of the LDP is governed by the Fenchel–Legendre conjugate function [6] of $\Lambda^{(0)}$, which is defined by

$$I^{(0)}(\omega) := (\Lambda^{(0)})^*(\omega) := \sup_{\tau\in\mathbb{R}}\ \big(\omega\tau - \Lambda^{(0)}(\tau)\big)\,. \tag{95}$$

In our case, $\Lambda^{(0)}$ can be written as an integral function with respect to the *first branch* of the Stieltjes transform of $\nu_{\gamma,\lambda}$, cf. (20). Leveraging this specific structure, we derive an explicit formula for the rate function. Once again, the $\mathcal{H}_1$ case can be deduced from $\mathcal{H}_0$, so we focus on $\mathcal{H}_0$ first.

Proposition C.6. *$I^{(0)}$ coincides with*

$$I^{(0)}(\omega) = \frac{1}{2}\int_{\mu_{0,\infty}}^{\omega} \underline{\mathfrak{s}}^{[-1]}(u)\,\mathrm{d}u\,, \quad \textit{for all } \omega > 0 \textit{ and } I^{(0)}(\omega) = +\infty \textit{ else}\,. \tag{96}$$

Proof. We denote by $\underline{\Lambda}^{(0)}$ the lower semicontinuous envelope of $\Lambda^{(0)}$, which might differ from $\Lambda^{(0)}$ only at the boundary of its domain $\tau = T/2 > 0$, and in particular has the same subgradient of $\Lambda^{(0)}$. We heavily rely on [6, Theorem 16.29], which yields $\partial I^{(0)} = (\partial\Lambda^{(0)})^{-1}$. In our case, we have $(\Lambda^{(0)})'(\tau) = \underline{\mathfrak{s}}(2\tau)$ for all $\tau < \frac{T}{2}$. Since $\underline{\mathfrak{s}}$ is increasing, the following limit exists

$$S := \lim_{\tau\to T^-/2}(\Lambda^{(0)})'(\tau) = \lim_{\tau\to T^-}\underline{\mathfrak{s}}(\tau) \in \mathbb{R}\cup\{+\infty\}\,. \tag{97}$$

We divide the analysis into two cases:

Suppose first that $S = +\infty$. By [6, Theorem 16.29], $(I^{(0)})'(\omega) = \frac{1}{2}\underline{\mathfrak{s}}^{-1}(\omega)$ for all $\omega \in (0, +\infty)$. Therefore, we get (96) by integrating with starting point $\mu_{0,\infty}$, since for this choice we have

$I^{(0)}(\mu_{0,\infty}) = 0$. Indeed, in this case $I^{(0)}$ has a unique stationary point, which, once again applying [6, Theorem 16.29], can be seen to be exactly $\mu_{0,\infty} = (\Lambda^{(0)})'(0)$. Thus, it is the unique minimizer of $I^{(0)}$, which has value

$$\inf_{\omega\in\mathbb{R}} I^{(0)}(\omega) = -\sup_{\omega\in\mathbb{R}} -I^{(0)}(\omega) = \underline{\Lambda}^{(0)}(0) = \Lambda^{(0)}(0) = 0\,.$$

Suppose now that $S < +\infty$. It must also hold that $\lim_{\tau\to T/2}\Lambda^{(0)}(\tau) < +\infty$. Therefore, we have

$$\partial\Lambda^{(0)}(\tau) = \begin{cases} \{\underline{\mathfrak{s}}(2\tau)\} & \text{if } \tau < \frac{T}{2}\,, \\ \left[S, +\infty\right) & \text{if } \tau = \frac{T}{2}\,, \\ \emptyset & \text{if } \tau > \frac{T}{2}\,. \end{cases} \tag{98}$$

Inverting the subgradient of $\Lambda^{(0)}$ we obtain that of $I^{(0)}$, which turns out to be single-valued and indeed exactly equal to $\frac{1}{2}\underline{\mathfrak{s}}^{[-1]}$. Integrating once again with starting point $\mu_{0,\infty}$ we obtain (96).

To conclude, we need to show that $I^{(0)}(0) = +\infty$. In this edge case, we have, using the fixed-point equation (81),

$$\begin{aligned} I^{(0)}(0) &= -\inf_{\tau\in\mathbb{R}}\Lambda^{(0)}(\tau) = \int_{(-\infty,0]} \underline{\mathfrak{s}}(z)\,\mathrm{d}z \\ &= \int_{(-\infty,0]}\int_{\mathbb{R}_+}\left(\lambda t^{-1} - z + \frac{1}{1+\gamma\underline{\mathfrak{s}}(z)}\right)^{-1}\mathrm{d}H(t)\,\mathrm{d}z \\ &\geq \int_{(-\infty,0]}\int_{\mathbb{R}_+}\frac{1}{\lambda t^{-1} - z + 1}\,\mathrm{d}H(t)\,\mathrm{d}z = \int_{\mathbb{R}_+}\int_{(-\infty,0]}\frac{1}{\lambda t^{-1} - z + 1}\,\mathrm{d}z\,\mathrm{d}H(t) = +\infty\,, \end{aligned}$$

where in the last identity we have used Tonelli, which is applicable since the integrand is positive. □

Remark C.7 (Computing $I^{(0)}$)**.** *While computing $\mathfrak{s}(z)$ requires solving the fixed-point equation* (81)*, identifying the pseudo inverse $\underline{\mathfrak{s}}^{[-1]}$ is indeed much simpler, provided that* dRNR *is* sharp*, i.e., $\mathfrak{s}_{\mathsf{dRNR}}(\tau) \to +\infty$ as $\tau \uparrow \inf\operatorname{supp}\mathsf{dRNR}$. In fact, starting from* (81) *it is easy to see that, in this case, we must have*

$$\underline{\mathfrak{s}}^{[-1]}(u) = \underline{\mathfrak{s}}^{[-1]}_{\mathsf{dRNR}}(u) + \frac{1}{1+\gamma u}\,, \quad \textit{for all } u \in (0, S)\,, \tag{99}$$

and is extended by a constant for all $u \geq S$. We can leverage this explicit expression to i) Compute exact rates for isotropic data, and ii) Build accurate numerical approximation of the rate function for non-isotropic data. In the latter, we only need to approximate the function $\underline{\mathfrak{s}}^{[-1]}_{\mathsf{dRNR}}$, which does not admit a closed form expression in general: We show an example in Figure 8. This function as well as its integral admit closed form expressions in the isotropic regime, cf. Proposition C.9.

C.2.3 Exponential decay in the Neyman–Pearson setup

Now that the explicit expression for $I^{(0)}$ has been derived, we can use it to bound or characterize the probability of error decay under one hypothesis, while the probability of error under the alternative hypothesis is kept fixed. This is exactly where we utilize the Gärtner–Ellis Theorem, Theorem A.4.

Proof of Theorem 3.2. We divide the proof into two steps. First, we show that with the selected threshold $\lim_{p\to+\infty} P_{MD} = \beta$, then the exponential decay of P_{FA}.

Observe that by the CLT for triangular arrays, when suitably normalized, $\hat{T}$ under $\mathcal{H}_1$ converges in distribution to a unit Gaussian since, conditionally on the training data, we have

$$\hat{T}(x) = \sum_{i=1}^{p}\frac{(v_i + a_i)^2}{\hat{\xi}_i}, \qquad v_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)\,, \tag{100}$$

where $a = U^T\Sigma^{-1/2}s$ and $\hat{W} = U\,\mathrm{diag}(\hat{\xi}_1^{-1}, \ldots, \hat{\xi}_p^{-1})U^T$. Thus, $\hat{T}$ is a sum of independent noncentral chi-square variables. Thus, setting $\tau_p = \hat{\tau}_1 := \hat{\mu}_1 + \hat{\sigma}_1 Q^{-1}(1-\beta)$, we obtain $\lim_{p\to+\infty} P_{MD} = \beta$ almost surely.

Turning to the decay of P_{FA}, using the continuity of $\hat{T}$, we get that

$$\limsup_{p\to+\infty} \frac{1}{p} \log P_{FA} = \limsup_{p\to+\infty} \frac{1}{p} \log \mathbb{P}\Big(\tfrac{1}{p}(\hat{T}(x) - \hat{\tau}_1) \geq 0 \;\Big|\; \hat{X}, \mathcal{H}_0 \Big) . \tag{101}$$

Now, since $\hat{\mu}_1/p \to \mu_{1,\infty}$ and $\hat{\sigma}_1/p \to 0$ almost surely by Proposition 3.1, the asymptotic LMGF of $\hat{T}(x) - \hat{\tau}_1$ under $\mathcal{H}_0$ in the sense of (19) is exactly $\tau \mapsto \Lambda^{(0)}(\tau) - \mu_{1,\infty}\tau$ for all $\tau \in \mathbb{R}$, whose Fenchel–Legendre conjugate is simply $\omega \mapsto I^{(0)}(\omega + \mu_{1,\infty})$. Since 0 belongs to the interior of the domain of the limiting LMGF, cf. Lemma C.2, we get using the Gärtner–Ellis Theorem (cf. Theorem A.4 applied to $\mathtt{t}_p := \frac{1}{p}(\hat{T}(x) - \hat{\tau}_1)$) that

$$\limsup_{p\to+\infty} \frac{1}{p} \log P_{FA} \leq - \inf_{\omega\in[0,+\infty)} I^{(0)}(\omega + \mu_{1,\infty}) = -I^{(0)}(\mu_{1,\infty}) =: -R^{(0)}_{\gamma,\lambda} , \tag{102}$$

where in the last step we have used that $I^{(0)}$ is convex and attains its minimum on $\mu_{0,\infty}$, which is smaller than $\mu_{1,\infty}$, which makes $I^{(0)}$ non-decreasing on $[\mu_{1,\infty}, +\infty)$. □

Remark C.8 (Decay under $\mathcal{H}_1$)**.** *Since Lemma C.5 does not establish the existence of $\Lambda^{(1)}(\tau)$ for $\tau > \tau_0$, we cannot directly apply Theorem A.4, which requires the existence of $\Lambda^{(1)}(\tau)$ for all $\tau \in \mathbb{R}$. Although local variants of the Gärtner–Ellis theorem could be invoked, the upper bound can be established directly, even in our setting.*

Similarly to the $\mathcal{H}_1$ case, if we pick $\tau_p = \hat{\tau}_0 := \hat{\mu}_0 + \hat{\sigma}_0 Q^{-1}(\alpha)$, we get $\lim_{p\to+\infty} P_{FA} = \alpha$. Now, for every $\tau < 0$ (not to be confused with the threshold), Markov inequality yields

$$P_{MD} = \mathbb{P}(\hat{T}(x) \leq \hat{\tau}_0 \mid \hat{X}, \mathcal{H}_1) \leq \mathbb{E}\big[\exp\big\{\tau \hat{T}(x)\big\} \;\big|\; \hat{X}, \mathcal{H}_1 \big] e^{-\tau\hat{\tau}_0} . \tag{103}$$

Since, almost surely, $\hat{\Lambda}^{(1)}(\tau) \to \Lambda^{(1)}(\tau)$ for all $\tau < 0 \leq \tau_0$ by Lemma C.5, continuity, and $\hat{\tau}_0/p \to \mu_{0,\infty}$ almost surely, we obtain

$$\begin{aligned} \limsup_{p\to\infty} \frac{1}{p} \log P_{MD} &\leq \inf_{\tau\leq 0} \Big\{ \Lambda^{(1)}(\tau) - \tau\mu_{0,\infty} \Big\} \\ &= \inf_{\tau\leq 0} \Big\{ \Lambda^{(0)}(\tau) + \tau\Delta(\tau) - \tau\mu_{0,\infty} \Big\} =: -R^{(1)}_{\gamma,\lambda} , \end{aligned} \tag{104}$$

Thus, $R^{(1)}_{\gamma,\lambda}$ is the infimal convolution *between $I^{(0)}$ and the sum of the Fenchel–Legendre conjugate of $g\colon \tau \mapsto \tau\Delta(\tau)$ and the indicator function of the positive line at $\mu_{0,\infty}$, see, e.g., [6, Proposition 15.2].*

In general, $R^{(0)}_{\gamma,\lambda}$ and $R^{(1)}_{\gamma,\lambda}$ only represent upper bounds for the decay rates of the corresponding error probabilities. Next, we show that in some cases the LDP indeed holds.

Isotropic regime Suppose we have isotropic data, i.e., $\Sigma = \sigma^2 I$. Then, $\mathsf{dRNR} = \delta_{\mathsf{RNR}}$, with $\mathsf{RNR} := \lambda/\sigma^2$. In this case, $\mathfrak{s}_{\mathsf{dRNR}}(z) = \frac{1}{\mathsf{RNR} - z}$, hence dRNR is sharp. Furthermore, the no-spectral-outlier assumption, Assumption 3, is satisfied. We can thus utilize (99) to turn Proposition C.6 into an explicit expression for the rate function, at least under $\mathcal{H}_0$:

Proposition C.9. *Let $\Sigma = \sigma^2 I$, $r = \mathsf{RNR} = \lambda/\sigma^2 > 0$. Let $\hat{T}$ be defined as in* (3)*. Define*

$$F_\gamma(\omega) = \frac{r}{2}(\omega - \mu_{0,\infty}) + \frac{1}{2\gamma} \log \frac{1 + \gamma\omega}{1 + \gamma\mu_{0,\infty}} - \frac{1}{2} \log \frac{\omega}{\mu_{0,\infty}} , \quad \textit{for } \omega > 0 . \tag{105}$$

Then, the following hold:

- *If $\gamma \in (0,1)$, then $T = r + (1 - \sqrt{\gamma})^2$, $S = (\sqrt{\gamma} - \gamma)^{-1}$ and*

$$I^{(0)}(\omega) = \begin{cases} F_\gamma(\omega) , & \textit{for } 0 < \omega \leq S , \\ F_\gamma(S) + \frac{T}{2}(\omega - S) , & \textit{for } \omega > S . \end{cases} \tag{106}$$

- *If $\gamma > 1$, then $T = r$, $S = +\infty$, and*

$$I^{(0)}(\omega) = F_\gamma(\omega) , \quad \textit{for all } \omega > 0 . \tag{107}$$

  and the LDP holds.

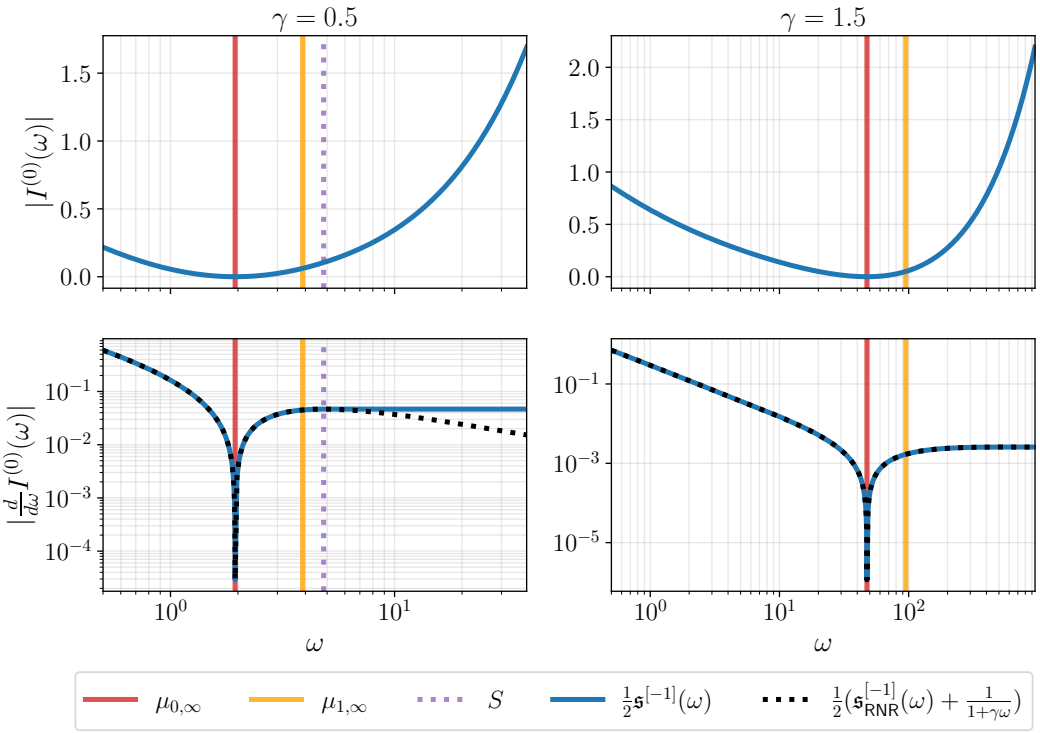


Figure 8: Computing the inverse Stieltjes transform $\underline{\mathfrak{s}}^{[-1]}$ according to (22) and the function $I^{(0)}$ given in Proposition C.6, with $\gamma < 1$ and $\gamma > 1$. Here, $\lambda = 10^{-2}$, $H = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2$ and dSNR is constant equal to 1.

Proof. First, observe that $\underline{\mathfrak{s}}^{[-1]}_{\mathsf{dRNR}}(u) := r - \frac{1}{u}$ and from (99) that $S = +\infty$ if and only if $\gamma > 1$. To see this, note that by a simple calculation

$$u \mapsto \underline{\mathfrak{s}}^{[-1]}_{\mathsf{dRNR}}(u) + \frac{1}{1+\gamma u} = r - \frac{1}{u} + \frac{1}{1+\gamma u}\,, \tag{108}$$

has a stationary point (a maximum indeed) in $\mathbb{R}_+$ if and only if $\gamma \in (0,1)$. Therefore, if $\gamma > 1$, we must have that $\underline{\mathfrak{s}}^{[-1]}$ can be inverted on $(0,+\infty)$, yielding $S = +\infty$ and, by inspection, $T = r$. If $\gamma \in (0,1)$, this maximum $\frac{1}{\sqrt{\gamma}-\gamma}$, should correspond exactly to S, otherwise we can either extend $\underline{\mathfrak{s}}$ beyond T, or have infinite derivative at $z = \frac{1}{\sqrt{\gamma}-\gamma}$, both of which are not possible, so $S = \frac{1}{\sqrt{\gamma}-\gamma} < +\infty$, and the value there is $T = r + (1-\sqrt{\gamma})^2$. Moreover,

$$0 = (I^{(0)})'(\mu_{0,\infty}) = \frac{1}{2}\underline{\mathfrak{s}}^{[-1]}(\mu_{0,\infty})\,. \tag{109}$$

Thus, $\underline{\mathfrak{s}}^{[-1]}(\mu_{0,\infty}) = 0$, which implies $\mu_{0,\infty} \leq S$. Altogether, we conclude that $\underline{\mathfrak{s}}^{[-1]}$ coincides with (108) in $(0,S)$, passes through zero at $\mu_{0,\infty} \leq S$ and is defined as its unique constant continuation for $u \geq S$.

Once the exact formula of the integrand is identified in both the cases $\gamma \in (0,1)$ and $\gamma > 1$, we can now derive (106) and (107) by integrating it. As $S = +\infty$ in the case $\gamma > 1$, relying on Theorem A.4, the LDP must hold since

$$\lim_{\tau \to T/2} (\Lambda^{(0)})'(\tau) = \lim_{\tau \to T/2} \underline{\mathfrak{s}}(2\tau) = S = +\infty\,, \tag{110}$$

and thus the steepness condition of Theorem A.4 is satisfied. □

Remark C.10. *Combining the explicit expression of the rates given by Proposition C.9 with the explicit expression of the moments given by Proposition 3.1 and* (41)*, we obtain an explicit expression of the rates in the Neyman–Pearson setup as detailed in Appendix C.2.3.*

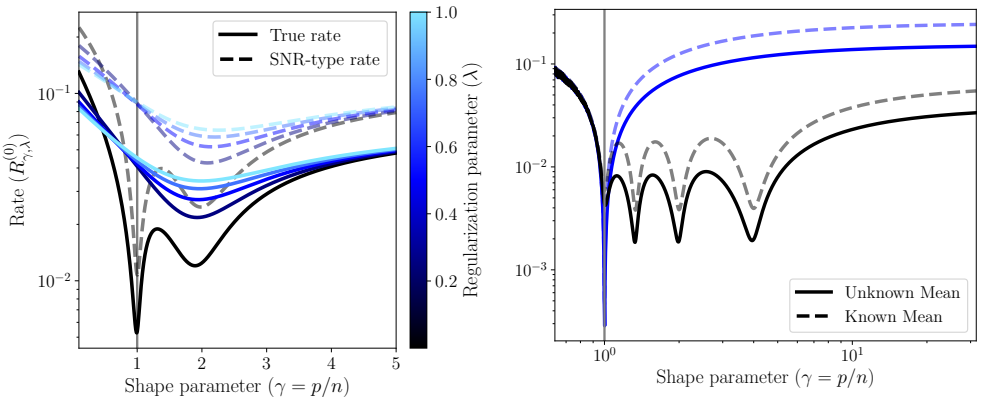


Figure 9: Multiple-descent in the quadratic case. **Right:** False alarm rate function of the quadratic decision statistic compared to the same for the linear decision statistic (2) in the geometric setting of Figure 2a. **Left:** comparing, for different choices of λ, the true rate function with the *SNR-type rate* that one obtains by assuming $\hat{T}$ to be Gaussian.

We are now ready to compute the rate function in this quadratic case and compare it with its linear counterpart discussed in Section 2. To avoid repetitions, we consider the same geometric setting of Section 2 and show the results in Figure 9. Specifically, we also compare the rate function $R^{(0)}_{\gamma,\lambda}$ with a naive approximation that one gets assuming $\hat{T}$ to be Gaussian, yielding

$$\text{SNR-type rate} \approx (\mu_{1,\infty} - \mu_{0,\infty})^2/(2\sigma^2_{0,\infty})\,.$$

We refer to this approximation as the *SNR-type rate*. From Figure 9, we can observe an expected loss in performances compared to the case where the mean is known or approximable from the training set, with an overall shapes that exhibit the same behaviors. Moreover, we observe that approximating the rate of $\hat{T}$ with its SNR-type rate would incur into a severe mistake.

D Further insights on Section 4

D.1 Models and data

We now describe the foundation models as well as the datasets considered for this work. The experiments use an NVIDIA A100 (80GB) GPU and 216GB of RAM.

D.1.1 Optical images

We first consider optical Earth observation imagery, where detection under controlled false-alarm constraints is critical for operational monitoring and surveillance. Satellite imagery plays a central role in maritime domain awareness, as well as critical undersea infrastructure monitoring and strategic intelligence. It provides a representative high-dimensional regime in which pre-trained visual foundation models generate semantically structured latent representations. We evaluate whether Neyman–Pearson detection applied directly to these latent features achieves reliable performance across distinct remote sensing tasks that vary in semantic content, spatial structure, image resolution, and sensor characteristics.

Feature extraction model For the task, we employ the DINOv3 model, which is a vision foundation model based on a self-supervised Vision Transformer architecture, capable of generating high-quality dense features [45]. Specifically, we use the publicly available DINOv3 ViT-L architecture with 300M parameters that is pre-trained on a dataset of 493 million Maxar satellite images of 512×512 resolution (SAT-493M). The CLS token, which is the final 1024-dimensional representation generated by the model, is utilized as the global image feature to apply the decision statistic. We evaluate the decision statistics of foundation model representations across two Earth observation tasks: ship detection and fire risk assessment.

Airbus ship detection challenge dataset First, we use the publicly available Airbus Ship Detection Challenge dataset released as part of a public competition on Kaggle aimed at automated ship detection in high resolution satellite imagery [26]. It provides around 192,000 optical satellite images at 768×768 resolution. Each image is paired with a ground-truth ship segmentation mask. Among these, approximately 42,000 contain at least one ship with a variety of sizes, orientations, and environmental conditions. The remaining ≈150,000 images contain only background scenes of ocean, shoreline, or land without vessels. From the dataset, we consider images containing at least one ship as positive samples ($\mathcal{H}_1$) and background-only images as negative samples ($\mathcal{H}_0$). Figure 10a shows example images for each hypothesis. We construct a balanced dataset with 37,376 training images (≈81%) and 8,000 validation images (≈19%) per class. The dataset occupies approximately 13 GB of storage, and feature extraction takes about 4 minutes with a batch size of 32.

A remote sensing dataset for fire risk assessment The FireRisk dataset includes seven fire risk classes with a total of 91,872 labeled images of 320×320 resolution derived from high-resolution aerial imagery provided by the U.S. Department of Agriculture National Agriculture Imagery Program for fire risk assessment. The labels are determined by the U.S. Department of Agriculture based on several factors such as the wildfire likelihood, potential intensity and landscape conditions [43]. From the dataset, we consider images with *very low risk* label as positive samples ($\mathcal{H}_1$), and images with *non-burnable* label as negative samples ($\mathcal{H}_0$). Figure 10c presents a number of scenes for each hypothesis. We sample 23,050 images per class to have a balanced dataset, where 17,959 (≈78%) of the images are used for training, and 5,091 (≈22%) are for testing. The dataset occupies approximately 7 GB of storage, and feature extraction takes about 2 minutes and 20 seconds with a batch size of 32.

D.1.2 Text

We next evaluate our framework on textual data, which serves as a critical source of actionable information for operational monitoring and anomaly detection. Textual reports, alerts, and communications can provide early indications of incidents or anomalies that may not be directly observable via physical sensors. Effectively extracting high-level semantic features from such text is therefore critical for timely automated decision-making. Semantic abstraction in language is inherently high-dimensional, and foundation models induce richly structured embedding spaces to capture these nuances. Unlike vision representations learned through self-supervised invariance objectives, text embeddings are explicitly optimized for semantic retrieval via contrastive query–document supervision. This training paradigm directly shapes the geometry of the latent space to reflect semantic similarity and ranking structure rather than task-agnostic feature invariance. We evaluate whether Neyman–Pearson detection applied directly to these latent features achieves reliable performance across textual inputs that vary in semantic content, linguistic style, length, and contextual origin.

Feature extraction model For the text modality, we employ the Llama-Nemotron-Embed-1B-v2 model, a text foundation model optimized for multilingual semantic retrieval [35]. Its backbone is based on the Llama 3.2 1B decoder-only architecture [20]. To adapt the model for embedding extraction, the standard causal attention mask is replaced with a bidirectional attention mechanism to repurpose the transformer blocks as a non-causal encoder [29]. The model was developed by NVIDIA using a two-stage contrastive training process: an initial pre-training phase on a proprietary mixture of 12 million document–passage pairs curated from public datasets, followed by supervised fine-tuning on 1 million high-fidelity, instruction-wrapped query–document pairs. To obtain a global semantic representation, we perform masked mean pooling across the final hidden states by excluding padding tokens from the average, which yields a 2048-dimensional feature vector that serves as the global text feature for applying the decision statistic.

Disaster incidents data from Twitter We use the publicly available HumAID: Human-Annotated Disaster Incidents Data from Twitter dataset, comprising annotated tweets collected from multiple real-world disaster events [1]. Each tweet is labeled with one of 10 humanitarian categories based on its context. We consider tweets reporting injured or dead people, missing or found people, infrastructure and utility damage, requests or urgent needs, and rescue, volunteering or donation efforts as positive samples ($\mathcal{H}_1$), which represents urgent or actionable information. The remaining tweets with labels caution and advice, displaced people and evacuations, other relevant information, sympathy and support, and non-humanitarian content are considered negative samples ($\mathcal{H}_0$). Table 1

presents a number of example tweets and their corresponding humanitarian categories for each hypothesis. To construct a balanced binary dataset, we first combine the validation and test splits to form a larger test set. Then, we sample 36,764 tweets per class to create a balanced dataset, where 25,732 (≈70%) are used for training, and 11,032 (≈30%) are used for testing. The dataset occupies approximately 120 MB of storage, and feature extraction takes about 8 minutes and 40 seconds with a batch size of 32.

D.1.3 Audio

Finally, we investigate acoustic event detection in audio signals, which is fundamental for surveillance and early-warning systems. Reliable detection under false-alarm constraints is particularly important in safety-critical acoustic sensing applications. Audio representations differ from vision and text as they encode temporal dependencies and spectral structure inherent to raw waveforms. Pre-trained audio foundation models produce compact global embeddings that summarize long temporal contexts. In contrast to retrieval-optimized text embeddings and invariance-driven visual features, foundation audio representations are shaped by self-supervised predictive objectives that model masked acoustic units. This training paradigm emphasizes temporal consistency and contextual prediction rather than semantic ranking. The audio modality therefore provides an additional test of whether Neyman–Pearson detection in latent space remains effective when the underlying geometry is induced by sequence prediction objectives rather than semantic ranking or invariance-driven training.

Feature extraction model For this task, BEATs is used as the backbone architecture to extract features. As a self-supervised audio foundation model built on a Transformer encoder, it learns high-level acoustic representations directly from raw waveforms [14]. We use the publicly available BEATs_iter3+ model having 90M parameters and pre-trained on 2 million 10-second audio clips sampled at 16 kHz from YouTube videos (AudioSet-2M) [23]. The frame-level embeddings produced by the final Transformer layer are mean-pooled over the temporal dimension to obtain a 768-dimensional global representation of the audio signal, which is then used to compute the decision statistic.

SONYC urban sound tagging dataset We use the SONYC Urban Sound Tagging v2 dataset (SONYC-UST-V2), a multilabel urban acoustic dataset collected from the SONYC sensor network deployed across New York City [13]. The dataset provides 10-second recordings annotated across 23 fine-grained sound classes grouped into 8 coarse categories, including engine noise, machinery impact, powered saws, alert signals, music, human voice, and dog sounds. We consider the human voice coarse class as the target and formulate a one-versus-rest binary detection task. Recordings in which human voice is present form positive samples ($\mathcal{H}_1$), while recordings in which human voice is entirely absent form negative samples ($\mathcal{H}_0$). Since the dataset is multilabel, negative samples may contain other sound classes, but contain no human voice activity. Annotations from all available annotators and all dataset splits are pooled prior to class assignment. We prioritize the verified ground-truth labels provided by the SONYC team when available; otherwise, we consider the class present if at least one crowd annotator labeled it as such. Figure 10b illustrates representative mel spectrograms for each hypothesis. We sample an equal number of audio recordings per class to obtain a balanced dataset, where 11,520 (80%) are used for training and 2,880 (20%) are used for testing. The dataset occupies approximately 17 GB of storage, and feature extraction takes about 14 minutes with a batch size of 32.

D.2 Experimental design

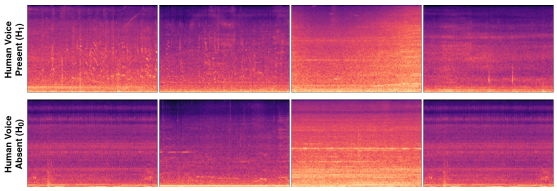
In this section, we give further details on the implementation aspects of Section 4.1.

As a preliminary check, we inspect randomly selected marginal distributions of each training data mapped into the latent space of the considered foundation models and fit them with Gaussian distributions. The results are displayed in Figure 11.

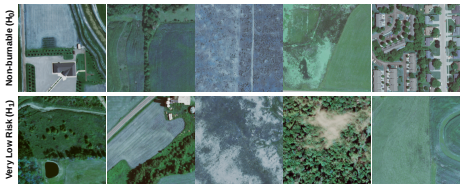
Second, Figure 12 extends the theoretical rate curves of Figure 1 to larger values of γ, for which empirical validation is not possible due to limited data availability. Interestingly, the curves exhibit non-monotonic behavior, reminiscent of the multiple-descent phenomenon observed in Figure 2.



(a) Airbus Ship Detection dataset.



(b) SONYC-UST-V2 dataset.



(c) FireRisk dataset.

Figure 10: Samples extracted from three different image datasets, including satellite images 10a and 10c, and a spectrogram 10b.

Hypothesis	**Tweet**	**Humanitarian Category**
$\mathcal{H}_0$	Flash Flooding Threatens Towns Devastated by Hurricane Maria. #Weather #PuertoRico: via @YouTube	Caution and advice
$\mathcal{H}_0$	Louisiana – Hurricane categories tell only part of the story — here's the damage storms like Irma and Harvey can do	Other relevant information
$\mathcal{H}_1$	Community Blood Center needs donations to prepare for Hurricane Dorian via @ndn	Requests or urgent needs
$\mathcal{H}_1$	Drone video: St. Augustine condos destroyed by Hurricane Irma - | Droneblog - #drones	Infrastructure and utility damage

Table 1: Examples from the HumAID dataset.

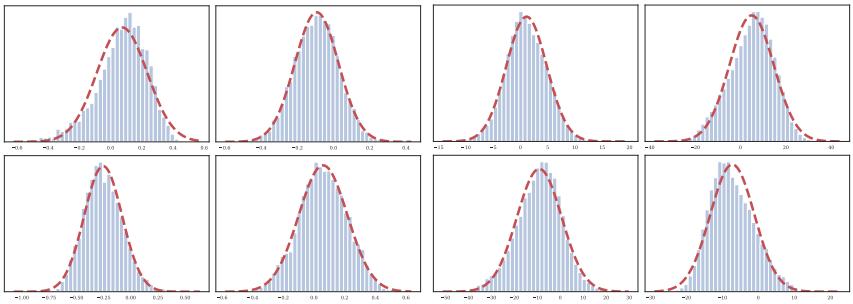
D.2.1 Mismatched covariances

In Point 1 we adapt our theoretical formula to the case in which in the test set $\hat{x} \sim \mathcal{N}(s, \Sigma_1)$, with $\Sigma_1 \neq \Sigma_0 := \Sigma$, i.e., where the covariance matrix under $\mathcal{H}_1$ does not coincide with the covariance matrix under $\mathcal{H}_0$. In this case, we denote by $\hat{\Sigma}_{\lambda,0} := (\hat{\Sigma}_0 + \lambda I)$, where $\hat{\Sigma}_0$ is the centered sample covariance matrix from $\mathcal{H}_0$ data and shall not be confused with $\hat{\Sigma}_\lambda$ from (2), with $\lambda = 0$. First, the mean now reads as

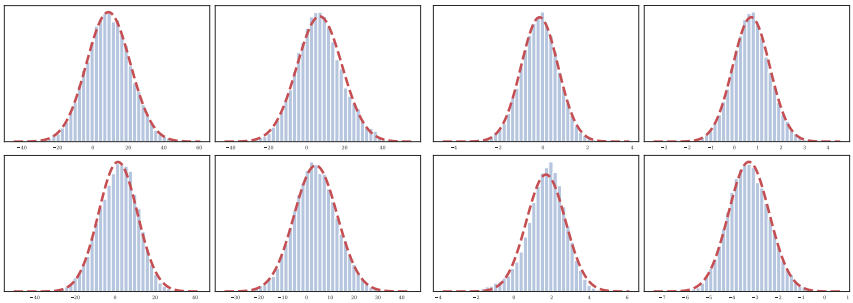
$$\frac{\hat{\mu}}{p} = \operatorname{tr}\left[\frac{s\hat{s}^T}{p}\left(\hat{\Sigma}_0 + \lambda I\right)^{-1}\right] \sim \frac{\kappa(\lambda)}{\lambda}\operatorname{tr}\left[\frac{ss^T}{p}\left(\Sigma_0 + \kappa(\lambda) I\right)^{-1}\right] \sim \frac{\kappa(\lambda)}{\lambda}\,\mathfrak{n}_1(\kappa(\lambda))\,,$$

where $\mathfrak{n}_1$ is computed entirely from $\mathcal{H}_0$. Next, note that the variances of $\hat{T}$ under $\mathcal{H}_0$ and $\mathcal{H}_1$ are no longer equal. We have:

$$\begin{aligned}\frac{\hat{\sigma}_0^2}{p} &= \frac{1}{p}\mathbb{E}[\hat{T}(x)^2 \mid \hat{X}, \mathcal{H}_0] = \frac{1}{p}\mathbb{E}[(\hat{s}^T\hat{\Sigma}_{\lambda,0}^{-1}x)^2 \mid \hat{X}, \mathcal{H}_0] \\ &= \frac{1}{p}\hat{s}^T\hat{\Sigma}_{\lambda,0}^{-1}\mathbb{E}\left[xx^T \mid \hat{X}, \mathcal{H}_0\right]\hat{\Sigma}_{\lambda,0}^{-1}\hat{s} = \frac{1}{p}\hat{s}^T\hat{\Sigma}_{\lambda,0}^{-1}\Sigma_0\hat{\Sigma}_{\lambda,0}^{-1}\hat{s} = \frac{1}{p}\operatorname{tr}\left[\hat{s}\hat{s}^T\hat{\Sigma}_{\lambda,0}^{-1}\Sigma_0\hat{\Sigma}_{\lambda,0}^{-1}\right]. \\ &\sim \frac{\kappa(\lambda)^2}{\lambda^2}\frac{\mathfrak{n}_2(\kappa(\lambda))}{\left(1 - \gamma\,\mathfrak{m}_2(\kappa(\lambda))\right)}\,,\end{aligned}$$



(a) BEATs : SONYC UST V2. (b) DINOv3 : Airbus.



(c) DINOv3 : FireRisk. (d) Llama NE1bv2 : HumAID.

Figure 11: Distribution of different training data in the latent space of different foundation models. Showing four randomly selected marginals and their Gaussian fit.

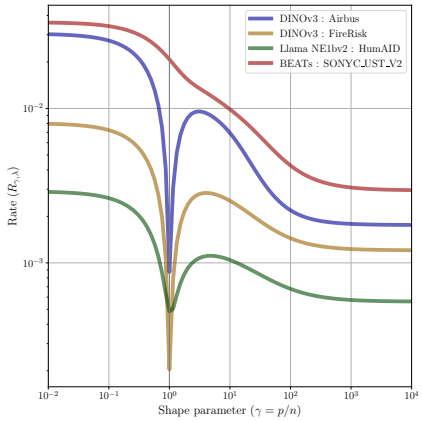


Figure 12: Theoretical rate as a function of the shape parameter γ across different foundation models and datasets, extending Figure 1 to larger values of γ on a logarithmic scale.

entirely computed on $\mathcal{H}_0$, so once again the limit of $\hat{\sigma}_0^2$ is not affected by the mismatch. For $\hat{\sigma}_1^2$ things change, we have, again from Theorem A.1 and neglecting infinitesimal terms,

$$\begin{aligned}
\frac{\hat{\sigma}_1^2}{p} &= \frac{1}{p}\mathbb{E}[(\hat{T}(x)-\hat{\mu})^2 \mid \hat{X}, \mathcal{H}_1] = \frac{1}{p}\mathbb{E}[(\hat{s}^T\hat{\Sigma}_{\lambda,0}^{-1}(x-s))^2 \mid \hat{X}, \mathcal{H}_1] \\
&= \frac{1}{p}\hat{s}^T\hat{\Sigma}_{\lambda,0}^{-1}\mathbb{E}\big[(x-s)(x-s)^T \mid \hat{X}, \mathcal{H}_1\big]\hat{\Sigma}_{\lambda,0}^{-1}\hat{s} = \frac{1}{p}\hat{s}^T\hat{\Sigma}_{\lambda,0}^{-1}\Sigma_1\hat{\Sigma}_{\lambda,0}^{-1}\hat{s} = \operatorname{tr}\left[\frac{\hat{s}\hat{s}^T}{p}\hat{\Sigma}_{\lambda,0}^{-1}\Sigma_1\hat{\Sigma}_{\lambda,0}^{-1}\right]. \\
&\sim \frac{\kappa(\lambda)^2}{\lambda^2}\operatorname{tr}\left[\frac{ss^T}{p}\big(\Sigma_0+\kappa(\lambda)I\big)^{-1}\Sigma_1\big(\Sigma_0+\kappa(\lambda)I\big)^{-1}\right] \\
&\quad + \frac{\kappa(\lambda)^2}{\lambda^2}\operatorname{tr}\left[\frac{ss^T}{p}\big(\Sigma_0+\kappa(\lambda)I\big)^{-2}\Sigma_0\right]\operatorname{tr}\left[\Sigma_1\big(\Sigma_0+\kappa(\lambda)I\big)^{-2}\Sigma_0\right]\frac{1}{n-\mathrm{d}f_2(\kappa(\lambda))}.
\end{aligned}$$

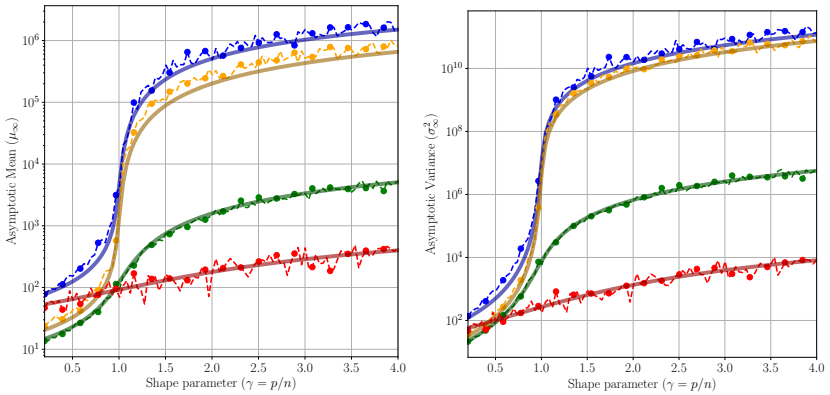


Figure 13: Mean and variance of the decision statistic $\hat{T}$ across the different models and datasets. Comparison between theoretical and empirical curves. Refer to Figure 1 for the legend.

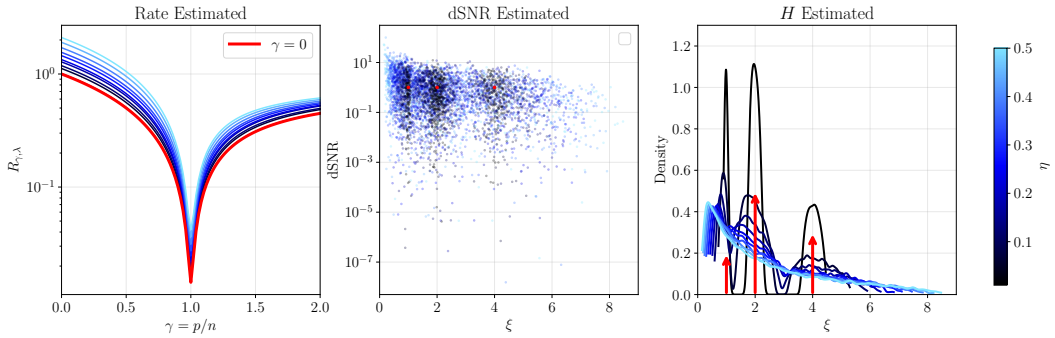


Figure 14: Stability of the estimated rate and population geometry under dataset truncation.

While we can easily derive a formula in the spirit of (14), we do not display it for brevity. We compare these theoretical formulas with their corresponding empirical ones in the four considered modalities in Figure 13. Please refer to Figure 1 for the corresponding legend.

D.2.2 Estimating H and dSNR from truncated datasets

In our experimental pipeline, we estimate H and dSNR from the full training set. Although large, this dataset may still yield estimates affected by high-dimensional effects: The estimation aspect ratio is close to zero, but remains positive. A natural question is therefore whether replacing H and dSNR by estimates obtained at a positive aspect ratio yields a good approximation of $R_{\gamma,\lambda}$. Although we can establish a quantitative convergence result, we limit ourselves here to a numerical check using simulated Gaussian data.

We consider a non-isotropic population spectral distribution $H = 0.2\,\delta_1 + 0.5\,\delta_2 + 0.3\,\delta_4$ and unit dSNR. For each $\eta = p/n \in [0.01, 0.5]$, we estimate H and dSNR with a truncated dataset of independent Gaussian observations of size $\hat{n} := p/\eta$. We then used these estimates to compute the approximate rate curve over $\gamma \in [0, 2]$. For small values of η, the estimated curves remain uniformly close to the rate curve obtained using the true population geometry, although they exhibit a systematic vertical shift. For values of η larger than 0.5, the mismatch is more significant, so we neglected these values. We show the results in Figure 14.

The vertical shift—analogous to the mismatch observed in Figure 1 for "Llama NE1bv2: HumAID"—is not surprising. As η increases, the empirical spectral estimate becomes increasingly biased toward zero, as the sample covariance matrix is indeed shifting closer to singularity. This effect is also visible in the third panel of Figure 14. This distortion of the estimated spectral geometry can, in turn, lead to an overestimation of the rate function.

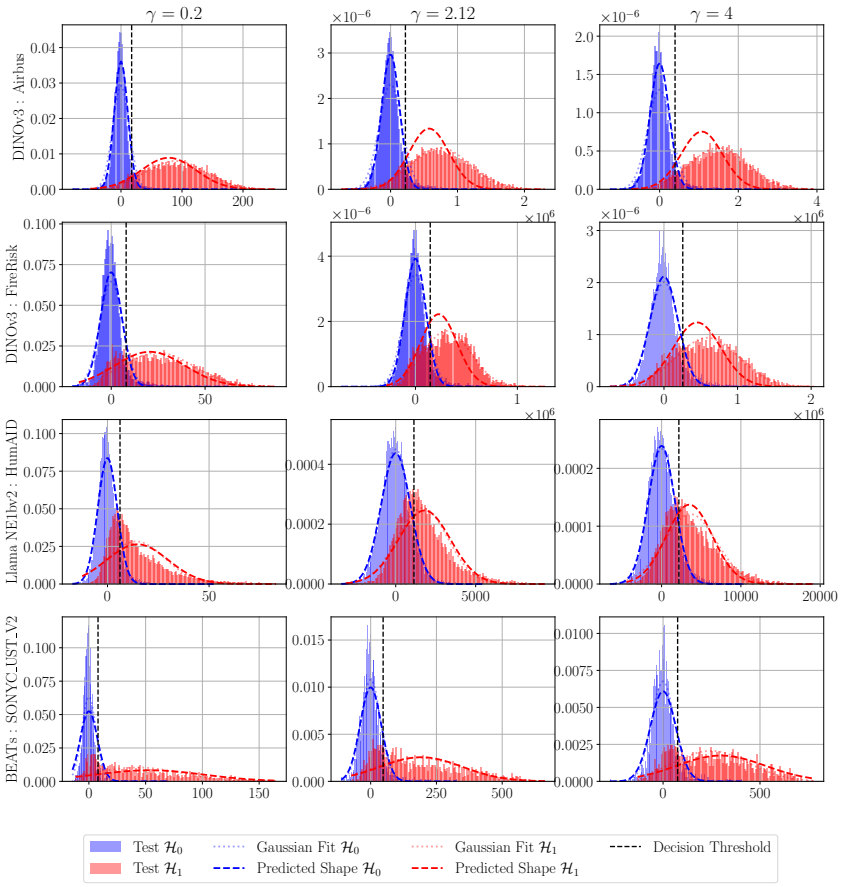


Figure 15: Sample distribution of $\hat{T}$ under $\mathcal{H}_0$ and $\mathcal{H}_1$.

D.2.3 Computing the theoretical P_{MD} with fixed P_{FA}

Once mean and variances for $\hat{T}$ are computed both under $\mathcal{H}_0$ and under $\mathcal{H}_1$, we can easily extend (32) to the mismatched case considered in this numerical example. The formula for P_{MD} as a function of P_{FA} simply becomes:

$$P_{MD} = 1 - Q\left(\frac{\hat{\sigma}_0 Q^{-1}(P_{FA}) - \hat{\mu}}{\hat{\sigma}_1}\right). \tag{111}$$

Plotting the theoretical prediction for $\hat{\sigma}_0$, $\hat{\sigma}_1$ and $\hat{\mu}$ in the high-dimensional regime, we obtain the results discussed in Figure 3a. Further, in Figure 15 we show the agreement between the predicted distributions of $\hat{T}$ under $\mathcal{H}_0$ and $\mathcal{H}_1$ test data for different values of γ. In particular, we observe that while for $\mathcal{H}_0$ data the predicted distribution is very close to the empirical one, for $\mathcal{H}_1$ data the prediction is less accurate, yet still yielding a good match.

D.2.4 LLR in different statistical settings

In the case featuring different covariance matrices under $\mathcal{H}_0$ and $\mathcal{H}_1$, the data-driven log-likelihood ratio shall be replaced by

$$\mathrm{DS}(H, K) \; : \quad \hat{T}(x) = -\frac{1}{2}(x - \hat{s})^T H (x - \hat{s}) + \frac{1}{2} x^T K x \,, \tag{112}$$

with H, K being $\hat{\Sigma}_{\lambda,0}^{-1}$ or $\hat{\Sigma}_{\lambda,1}^{-1}$, yielding three possible decision statistics. Note that while using $H = K$ yield a linear decision statistic, using $H \neq K$ yields a quadratic form. For the linear decision statistics the analysis can be performed using the same arguments introduced Section 2. In contrast, for the quadratic setup, we can employ the approach presented in Section 3—leveraging RMT and the Gärtner–Ellis Theorem.

For completeness, we also display the results of Figure 3b displayed in the full range of false positive range both in linear and log space in Figure 16.

D.3 Beyond foundation models

To further assess the generality of our framework beyond modern foundation models, we extended our experimental evaluation to earlier and more compact neural architectures. In particular, for the text modality, we considered BERT-base (110M parameters) and ALBERT-base (11.8M parameters) on the HumAID dataset. Interestingly, despite their substantially different architectures and model sizes, both produce latent representations that closely resemble isotropic Gaussian distributions. Moreover, as illustrated in Figure 17, the empirical error decay rates closely follow our theoretical predictions across different values of the shape parameter γ. These findings provide further evidence that the applicability of our high-dimensional framework is not restricted to recent large-scale foundation models, but extends to a broader family of representation-learning architectures.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main contributions are clearly articulated through two central research questions outlined in the introduction, with the corresponding results summarized in Section 1.2. Furthermore, the numerical experiments are organized into four parts, each designed to directly address and substantiate the claims presented in the introduction.

   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: The main limitations of our approach are clearly emphasized in Section 4.2.

   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: The full set of assumptions and proofs for all results are provided in the supplementary material. For brevity, some elementary arguments are presented in a condensed form.

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: The experimental pipeline is fully specified in Section 4.1, including data generation, model configuration, and evaluation protocols, with additional implementation details provided in the supplementary material. This level of description is sufficient for an interested reader to reproduce the experiments and verify the main claims independently of code availability.

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: [No]

   Justification: Due to authorization constraints, we are unable to release the code at this time. Nevertheless, the paper and supplementary material provide detailed descriptions of the experimental setup, data generation process, and evaluation procedures, offering a clear pathway for reproducing the main results independently.

   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: The paper specifies all relevant experimental details. Additional implementation details are provided in the supplementary material, ensuring clarity and completeness.

   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: [No]

   Justification: We do not report error bars, as our theoretical results provide strong guarantees (in an almost sure sense), which support the stability of the observed outcomes. As a result, point-wise evaluations are representative of the method's behavior.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The experiments use an NVIDIA A100 (80GB) GPU and 216GB of RAM. We provide additional details in Appendix E.

   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: The research has been conducted in accordance with the NeurIPS Code of Ethics, and no ethical concerns or violations have been identified.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [N/A]

Justification: The work is primarily theoretical and does not target specific applications or deployment scenarios. As such, no direct societal impacts (positive or negative) have been identified beyond the general influence of advances in AI and statistical methodology.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [N/A]

Justification: The paper does not involve the release of data, models, or systems with potential for misuse. The contribution is purely theoretical, and therefore no specific safeguards are required.

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: [Yes]

Justification: All external assets are properly credited, with appropriate citations to the original sources. Their usage complies with the respective licenses and terms of use.

- 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**

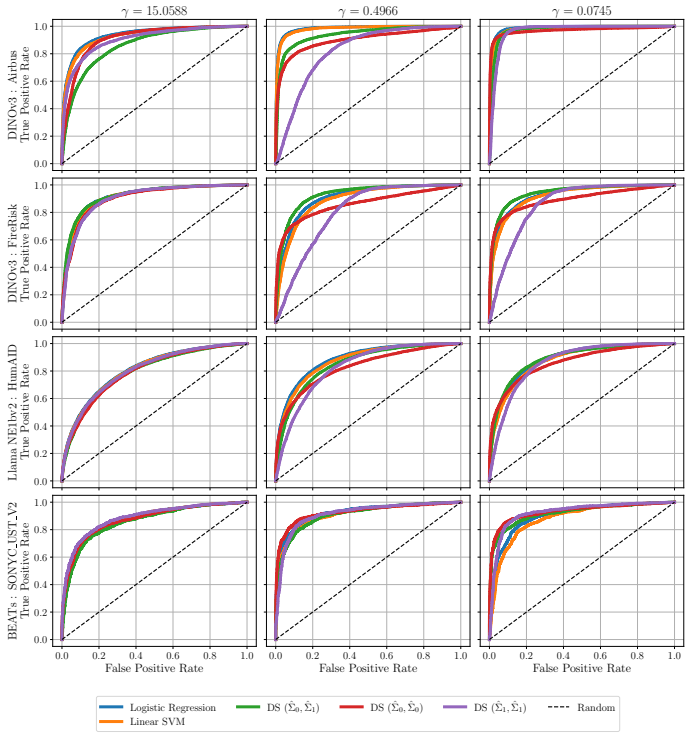
    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality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

    Answer: [N/A]

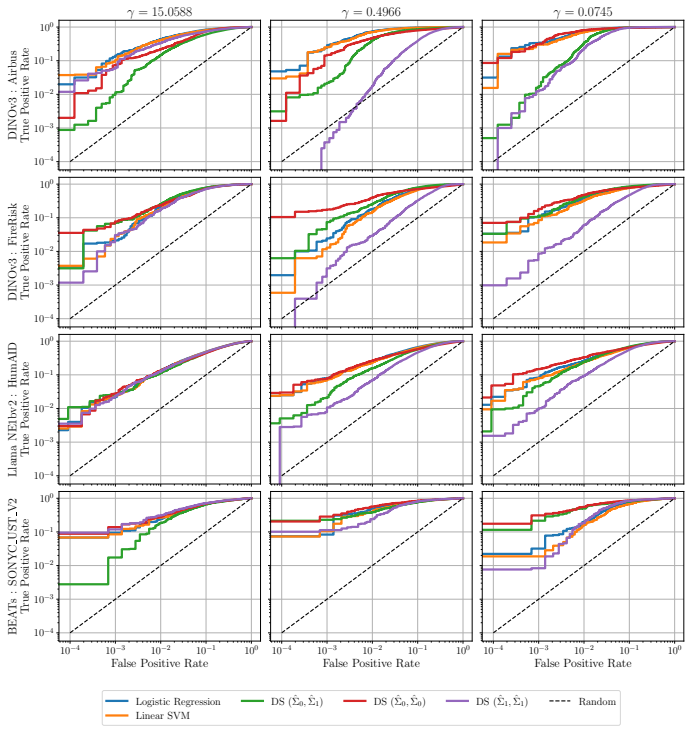
    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    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.



(a) Linear scale.



(b) Logarithmic scale.

Figure 16: ROC curves: Comparison with benchmark classifiers in linear and log-scale.

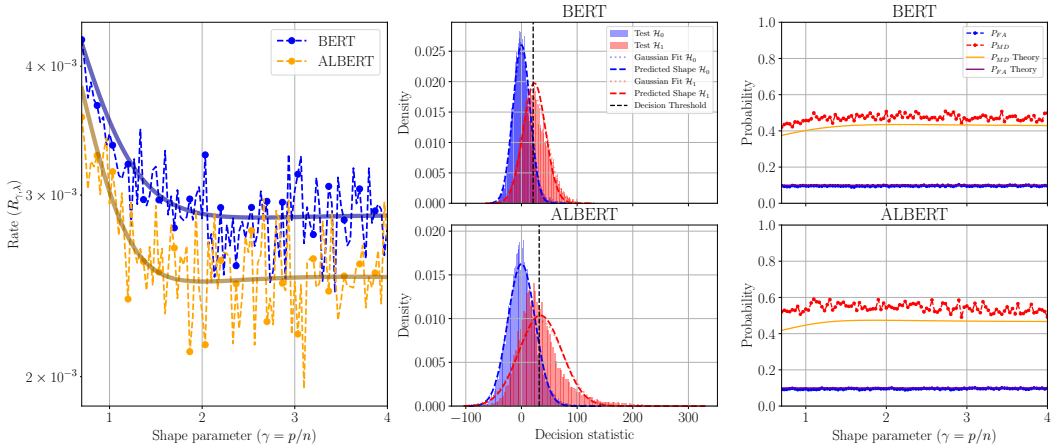


Figure 17: Beyond foundation models: Comparison between theoretical and empirical error decay rates as a function of the shape parameter $\gamma = p/n$ for BERT-base and ALBERT-base representations of the HumAID dataset.