Sparse classification boundaries

Yuri I. Ingster, Christophe Pouet, Alexandre B. Tsybakov

Sparse classification boundaries

Yuri I. Ingster[1], Christophe Pouet[2] and Alexandre B. Tsybakov [3]

March 27, 2009

Abstract

Given a training sample of size m from a d-dimensional population, we wish to allocate a new observation $Z \in \mathbb{R}^d$ to this population or to the noise. We suppose that the difference between the distribution of the population and that of the noise is only in a shift, which is a sparse vector. For the Gaussian noise, fixed sample size m, and the dimension d that tends to infinity, we obtain the sharp classification boundary and we propose classifiers attaining this boundary. We also give extensions of this result to the case where the sample size m depends on d and satisfies the condition $(\log m)/\log d \to \gamma$, $0 \le \gamma < 1$, and to the case of non-Gaussian noise satisfying the Cramér condition.

Keywords: Bayes risk, classification boundary, high-dimensional data, optimal classifier, sparse vectors

[1]Research partially supported by the RFBI Grant 08-01-00692-a and by Grant NSh–638.2008.1.

[2]Research partially supported by the grant ANR-07-BLAN-0234 and by PICS–2715.

[3]Research partially supported by the grant ANR-06-BLAN-0194, by the PASCAL Network of Excellence and Isaac Newton Institute for Mathematical Sciences in Cambridge (Statistical Theory and Methods for Complex, High-Dimensional Data Programme, 2008).

1 Introduction

1.1 Model and problem

Let $\mathbf{X} = (X_1, \dots, X_n)$ and $\mathbf{Y} = (Y_1, \dots, Y_m)$ be two i.i.d. samples from two different populations with probability distributions P_X and P_Y on $\mathbb{R}^d$ respectively. Here

$$X_i = (X_i^1, \dots, X_i^d), \quad Y_j = (Y_j^1, \dots, Y_j^d)$$

where X_i^k and Y_j^k are the components of X_i and Y_j. We consider the problem of discriminant analysis when the dimension of the observations d is very large (tends to $+\infty$). Assume that we observe a random vector $Z = (Z^1, \dots, Z^d)$ independent

of $(\mathbf{X}, \mathbf{Y})$ and we know that the distribution of Z is either P_X or P_Y. Our aim is to classify Z, i.e., to decide whether Z comes from the population with distribution P_X or from that with distribution P_Y.

In this paper we assume that

$$X_i^k = v_k + \xi_i^k, \qquad Y_j^k = u_k + \eta_j^k, \tag{1.1}$$

where $v = (v_1, \dots, v_d)$, $u = (u_1, \dots, u_d)$ are deterministic mean vectors and the errors $\xi_i^1, \dots, \xi_i^d$, $\eta_j^1, \dots, \eta_j^d$ are (unless other conditions are explicitly mentioned) jointly i.i.d. zero mean random variables with probability density f on $\mathbb{R}$.

Distinguishing between P_X and P_Y presents a difficulty only when the vectors v and u are close to each other. A particular type of closeness for large d can be characterized by the sparsity assumption [9, 1] that we shall adopt in this paper. As in [9, 1], we introduce the following set of sparse vectors in $\mathbb{R}^d$ characterized by a positive number a_d and a *sparsity index* $\beta \in (0, 1]$:

$$U_{\beta,a_d} = \left\{ u = (u_1, \dots, u_d) : \ u_k = a_d \varepsilon_k, \quad \varepsilon_k \in \{0, 1\}, \quad cd^{1-\beta} \le \sum_{k=1}^{d} \varepsilon_k \le Cd^{1-\beta} \right\}.$$

Here $0 < c < C < +\infty$ are two constants that are supposed to be fixed throughout the paper. The value $p = d^{-\beta}$ can be interpreted as the "probability" of occurrence of non-zero components in vector u.

In what follows we shall deal only with a special case of model (1.1) that was also considered recently by [4]. Namely, we assume:

$$v = 0, \quad u \in U_{\beta,a_d}.$$

In this paper we establish the classification boundary, i.e., we specify the necessary and sufficient conditions on β and a_d such that successful classification is possible. Let us first define the notion of successful classification. We shall need some notation. Let ψ be a decision rule, i.e., a measurable function of $\mathbf{X}, \mathbf{Y}, Z$ with values in $[0, 1]$. If $\psi = 0$ we allocate Z to the P_X-population, whereas for $\psi = 1$ we allocate Z to the P_Y-population. The rules ψ taking intermediate values in $(0, 1)$ can be interpreted as randomized decision rules. Let $P_{H_0}^{(u)}$ and $P_{H_1}^{(u)}$ denote the joint probability distributions of $\mathbf{X}, \mathbf{Y}, Z$ when $Z \sim P_X$ and $Z \sim P_Y$ respectively, and let $E_{H_0}^{(u)}$, $\mathrm{Var}_{H_0}^{(u)}$ and $E_{H_1}^{(u)}$, $\mathrm{Var}_{H_1}^{(u)}$ denote the corresponding expectation and variance operators. We shall also denote by $P^{(u)}$ the distribution of $\mathbf{Y}$ and by $E^{(u)}$, $\mathrm{Var}^{(u)}$ the corresponding expectation and variance operators. Consider the Bayes risk

$$\mathcal{R}_B(\psi) = \pi E_{H_0}^{(u)}(\psi) + (1 - \pi) E_{H_1}^{(u)}(1 - \psi),$$

where $0 < \pi < 1$ is a prior probability of the P_X-population, and the maximum risk

$$\mathcal{R}_M(\psi) = \max\left(E_{H_0}^{(u)}(\psi), E_{H_1}^{(u)}(1 - \psi) \right).$$

Let $\mathcal{R}(\psi)$ be either the Bayes risk $\mathcal{R}_B(\psi)$ or the maximum risk $\mathcal{R}_M(\psi)$.

We shall say that *successful classification is possible* if β and a_d are such that

$$\lim_{d\to+\infty} \inf_{\psi} \sup_{u\in U_{\beta,a_d}} \mathcal{R}(\psi) = 0 \tag{1.2}$$

for $\mathcal{R} = \mathcal{R}_M$ and $\mathcal{R} = \mathcal{R}_B$ with any fixed $0 < \pi < 1$. Conversely, we say that *successful classification is impossible* if β and a_d are such that

$$\liminf_{d\to+\infty} \inf_{\psi} \sup_{u\in U_{\beta,a_d}} \mathcal{R}(\psi) = \mathcal{R}_{max}, \tag{1.3}$$

where $\mathcal{R}_{max} = 1/2$ for $\mathcal{R} = \mathcal{R}_M$ and $\mathcal{R}_{max} = \min(\pi, 1-\pi)$ for $\mathcal{R} = \mathcal{R}_B$ with $0 < \pi < 1$.

We call (1.2) the *upper bound of classification* and (1.3) the *lower bound of classification*. The lower bound (1.3) for the maximum risk $\mathcal{R} = \mathcal{R}_M$ is interpreted as the fact that no decision rule is better (in a minimax sense) than the simple random guess. For the Bayes risk $\mathcal{R}_B$, the lower bound (1.3) is attained at the degenerate decision rule that does not depend on the observations: $\psi \equiv 0$ if $\pi > 1/2$ or $\psi \equiv 1$ if $\pi \leq 1/2$.

The condition on (β, a_d) corresponding to the passage from (1.2) to (1.3) is called the *classification boundary*. We shall say that a classifier $\psi = \psi_d$ is *asymptotically optimal* (or that ψ *attains the classification boundary*) if, for all β and a_d such that successful classification is possible, we have

$$\lim_{d\to+\infty} \sup_{u\in U_{\beta,a_d}} \mathcal{R}(\psi) = 0 \tag{1.4}$$

where $\mathcal{R} = \mathcal{R}_M$ or $\mathcal{R} = \mathcal{R}_B$ with any fixed $0 < \pi < 1$.

1.2 Main results

According to the value of β, we shall distinguish between *moderately sparse vectors* and *highly sparse vectors*. This division depends on the relation between m and d. For m not too large, i.e., when $\log m = o(\log d)$, moderately sparse vectors correspond to $\beta \in (0, 1/2]$ and highly sparse vectors to $\beta \in (1/2, 1)$. For large m, i.e., when $\log m \sim \gamma \log d,\ \gamma \in (0,1)$, moderate sparsity corresponds to $\beta \in (0, (1-\gamma)/2]$ and high sparsity to $\beta \in ((1-\gamma)/2, 1-\gamma)$.

The classification boundary for moderately sparse vectors is obtained in a relatively simple way (cf. Section 2). It is of the form

$$R_d \stackrel{\Delta}{=} d^{1/2-\beta} a_d \asymp 1. \tag{1.5}$$

This means that successful classification is possible if $R_d \to +\infty$, and it is impossible if $R_d \to 0$ as $d \to +\infty$. The result is valid both for $\beta \in (0, 1/2]$ and $m \geq 1$ fixed or for m depending on d such that $\log m \sim \gamma \log d,\ \gamma \in (0,1)$ as $d \to +\infty$ and $\beta \in (0, (1-\gamma)/2]$. Moreover, (1.5) holds under weak assumptions on the noise. In particular, for the upper bound of classification we only need to assume that the

noise has mean zero and finite second moment (cf. Section 2). The lower bound is proved under a mild regularity condition on the density f of the noise.

The case of highly sparse vectors is more involved. We establish the classification boundary for the following scenarios:

(A) $m \geq 1$ is a fixed integer, and the noise density f is Gaussian $\mathcal{N}(0, \sigma^2)$ with known or unknown $\sigma > 0$;

(B) $m \to +\infty$ as $d \to +\infty$, $\log m = o(\log d)$, and f is Gaussian $\mathcal{N}(0, \sigma^2)$ with known or unknown $\sigma > 0$.

(C) $\log m \sim \gamma \log d$, $\gamma \in (0, 1)$, and f is Gaussian $\mathcal{N}(0, \sigma^2)$ with known or unknown $\sigma > 0$.

The upper bounds are extended to the following additional scenario:

(D) $m \to +\infty$ as $d \to +\infty$, $\log m \sim \gamma \log d$, $0 \leq \gamma < 1$, $m/\log d \to +\infty$, and the noise satisfies the Cramér condition.

The conditions on the noise in (A)–(D) are crucial and, as we shall see later, they suggest that a special dependence of a_d on d and m of the form $a_d \asymp \sqrt{(\log d)/m}$ is meaningful in the highly sparse case. More specifically, we take

$$a_d = s\sigma\sqrt{\log d}, \quad x_1 = s\sqrt{m+1}, \tag{1.6}$$

where $x_1 > 0$ is fixed. The classification boundary in (A, B, D) is then expressed by the following condition on β, s and m:

$$x_1 = \phi(\beta) \tag{1.7}$$

where

$$\phi(\beta) = \begin{cases} \phi_1(\beta) & \text{if } 1/2 < \beta \leq 3/4, \\ \phi_2(\beta) & \text{if } 3/4 < \beta < 1, \end{cases} \tag{1.8}$$

with

$$\phi_1(\beta) = \sqrt{2\beta - 1}, \quad \phi_2(\beta) = \sqrt{2}\Big(1 - \sqrt{1-\beta}\Big). \tag{1.9}$$

In other words, successful classification is possible if $x_1 \geq \phi(\beta) + \delta$, and it is impossible if $x_1 \leq \phi(\beta) - \delta$, for any $\delta > 0$ and d large enough. This classification boundary is also extended to the case where x_1 depends on d but stays bounded.

For Scenario (C) let $a_d = \sigma x\sqrt{(\log d)/m}$ with fixed $x > 0$. We show that in this framework successful classification is impossible if $\beta > 1 - \gamma$ (cf. 1° in Section 2), and therefore we are interested in $\beta \in ((1-\gamma)/2, 1-\gamma)$. Set $\beta^* = \beta/(1-\gamma) \in (1/2, 1)$ and $x^* = x/\sqrt{1-\gamma}$. Then the classification boundary is of the form

$$x^* = \phi(\beta^*),$$

for the function $\phi(\beta)$ defined above.

Note that if f is known, the distribution P_X is also known. This means that we do not need the sample $\mathbf{X}$ to construct decision rules. Thus, in Scenarios (A), (B) and (C) when σ is known we can suppose w.l.o.g. that only the sample $\mathbf{Y}$ is available; this remark remains valid in the case of unknown σ, as we shall see it later. As to Scenario (D), we shall also treat it under the assumption that only the sample $\mathbf{Y}$ is available (w.l.o.g. if f is known), to be consistent with other results. However, if f is not known, the sample $\mathbf{X}$ contains additional information which can be used. The results for this case under Scenario (D) are similar to those that we obtain below but they are left beyond the scope of the paper.

For $m = 0$ (i.e., when there is no sample $\mathbf{Y}$) the problem that we consider here reduces to the problem of signal detection in growing dimension d, cf. [6, 7, 8, 9, 10, 1, 12], and our classification boundary coincides with the *detection boundary* established in [6]. Sharp asymptotics in the detection problem was studied in [6] (see also [9], Chapter 8) for known a_d or β. Adaptive problem (this corresponds to unknown a_d and β) was studied in [7, 8]. Various procedures attaining the detection boundary were proposed in [10, 1, 12]. Ingster and Suslina [10] introduced a method attaining the detection boundary based on the combination of three different procedures for the zones $\beta \in (0, 1/2]$, $\beta \in (1/2, 3/4]$ and $\beta \in (3/4, 1)$. Later Donoho and Jin [1] showed that a test based on the higher criticism statistic attains the detection boundary simultaneously for these zones. More recently Jager and Wellner [12] proved that the same is true for a large class of statistics including the higher criticism statistic.

The paper of Hall *et al.* [4] deals with the same classification model as the one we consider here but study a problem which is different from ours. They analyse the conditions under which some simple (for example, minimum distance) classifiers ψ satisfy

$$\lim_{d\to+\infty} E^{(u)}_{H_0}(\psi) = 0. \tag{1.10}$$

Hall *et al.* [4] conclude that for minimum distance classifiers (1.10) holds if and only if $0 < \beta < 1/2$. This implies that such classifiers cannot be optimal for $1/2 \le \beta < 1$. They also derive (1.10) for some other classifiers in the case $m = 1$.

The results of this paper and their extensions to the multi-class setting were summarized in [14] and presented at the Meeting "Rencontres de Statistique Mathématique" (Luminy, December 16-21, 2008) and at the Oberwolfach meeting "Sparse Recovery Problems in High Dimensions: Statistical Inference and Learning Theory" (March 15-21, 2009). In a work parallel to ours, Donoho and Jin [2, 3] and Jin [11] independently and contemporaneously have analysed a setting less general than the present one. They did not consider a minimax framework, but rather demonstrated that the higher criticism (HC) methodology can be successfully extended to the classification problem. Donoho and Jin [3] showed that, for a special case of Scenario (B), the "ideal" HC statistic attains the same upper bound of classification that we prove below. Together with our lower bound, this implies that the "ideal" HC statistic is asymptotically optimal, in the sense defined above, for the Scenario (B). Donoho and Jin announce that similar results for the HC statistic in Scenarios (A) and (C) will appear in their work in preparation.

This paper is organized as follows. Section 2 contains some preliminary remarks. In Section 3 we present the classification boundary and asymptotically optimal classifier for moderately sparse vectors under rather general conditions on the noise. In Section 4 we give the classification boundary and asymptotically optimal classifiers for highly sparse vectors under Scenarios (A), (B) and (C). Section 5 provides an extension to Scenario (D). Proofs of the lower and upper bounds of classification are given in Sections 6 and 7, respectively.

2 Preliminary remarks

In this section we collect some basic remarks on the problem assuming that f is the standard Gaussian density. As a starting point, we discuss some natural limitations for a_d.

1°. Remark that a_d cannot be too small. Indeed, assume that instead of the set U_{β,a_d} we have only one vector $u = (a_d\varepsilon_1, \dots, a_d\varepsilon_d)$ with known $\varepsilon_k \in \{0,1\}$. Then we get a familiar problem of classification with two given Gaussian populations. The notion of classification boundary can be defined here in the same terms as above, and the explicit form of the boundary can be derived from the standard textbook results. It is expressed through the behavior of $Q_d^2 \triangleq a_d^2 \sum_{k=1}^d \varepsilon_k$:

- if $Q_d \to 0$, then successful classification is impossible:

$$\liminf_{d\to+\infty} \inf_{\psi} \mathcal{R}(\psi) = \mathcal{R}_{max},$$

- if $Q_d \to +\infty$, then successful classification is realized by the maximum likelihood classifier $\psi^* = \mathbb{1}_{\{T^*>0\}}$ where

$$T^* = \sum_{k=1:\varepsilon_k=1}^{d} (Z^k - a_d/2).$$

Here and below $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function.

If we assume that $\sum_{k=1}^d \varepsilon_k \asymp d^{1-\beta}$, we immediately obtain some consequences for our model defined in Section 1. We see that successful classification in that model is impossible if a_d is so small that $d^{1-\beta}a_d^2 = o(1)$, and it makes sense to consider only such a_d that

$$d^{1-\beta}a_d^2 \to +\infty. \tag{2.1}$$

In particular, for $\gamma > 1-\beta$ successful classification is impossible under Scenario (C) with $a_d \asymp \sqrt{(\log d)/m}$.

We shall see later that (2.1) is a rough condition, which is necessary but not sufficient for successful classification in the model of Section 1. For example, in that model with $\beta \in (0, 1/2]$ and fixed m, the value a_d should be substantially larger than given by the condition $d^{1-\beta}a_d^2 \asymp 1$, cf. (1.5).

2°. Our second remark is that, on the other extreme, for sufficiently large a_d the problem is trivial. Specifically, non-trivial results can be expected only under the condition

$$x \stackrel{\Delta}{=} a_d \sqrt{m/\log d} \le 2\sqrt{2}, \tag{2.2}$$

Indeed, assume that

$$x > 2\sqrt{2}. \tag{2.3}$$

Then the problem becomes simple in the sense that successful classification is easily realisable under (2.3) and the classical condition (2.1). Indeed, take an analog of the statistic T^* where a_d and ε_k are replaced by their natural estimators:

$$T = \sum_{k=1}^{d} \left(Z^k - \frac{SY^k}{2\sqrt{m}} \right) \hat{\varepsilon}_k, \quad \text{with} \quad SY^k = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} Y_i^k, \quad \hat{\varepsilon}_k = \mathbb{1}_{\{SY^k > \sqrt{2\log d}\}}, \tag{2.4}$$

and consider the classifier $\psi = \mathbb{1}_{\{T>0\}}$. We can write $SY^k = \varepsilon_k \lambda + \zeta_k$ where ζ_k are independent standard normal random variables. It is well known that $\max_{k=1,\dots,d} |\zeta_k| \le \sqrt{2\log d}$ with probability tending to 1 as $d \to +\infty$. This and (2.3) imply that, with probability tending to 1, the vector $(\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_d)$ recovers exactly $(\varepsilon_1, \dots, \varepsilon_d)$ and the statistic T coincides with

$$\hat{T} = \sum_{k=1:\varepsilon_k=1}^{d} \left(Z^k - \frac{SY^k}{2\sqrt{m}} \right).$$

Since $E^{(u)}(SY^k/\sqrt{m}) = \varepsilon_k a_d$ and $\mathrm{Var}^{(u)}(SY^k/\sqrt{m}) = 1/m$, we find:

$$E_{H_0}^{(u)}(\hat{T}) = -\frac{a_d}{2} \sum_{k=1}^{d} \varepsilon_k, \quad E_{H_1}^{(u)}(\hat{T}) = \frac{a_d}{2} \sum_{k=1}^{d} \varepsilon_k,$$

$$\mathrm{Var}_{H_0}^{(u)}(\hat{T}) = \mathrm{Var}_{H_1}^{(u)}(\hat{T}) = \left(1 + \frac{1}{4m} \right) \sum_{k=1}^{d} \varepsilon_k.$$

It follows from Chebyshev's inequality that under (2.1) we have

$$E_{H_0}^{(u)}(\psi) = P_{H_0}^{(u)}(T > 0) \to 0, \quad E_{H_1}^{(u)}(1-\psi) = P_{H_1}^{(u)}(T \le 0) \to 0 \tag{2.5}$$

as $d \to +\infty$. Note that this argument is applicable in the general model of Section 1 (since the convergence in (2.5) is uniform in $u \in U_{\beta,a_d}$), implying successful classification by ψ under conditions (2.1) and (2.3).

3°. Let us now discuss a connection between conditions (2.1) and (2.3). First, (2.3) implies (2.1) if m is not too large:

$$m = o\left(d^{1-\beta} \log d \right). \tag{2.6}$$

On the other hand, if m is very large:

$$\exists\, b > 0 : \quad m \ge b d^{1-\beta} \log d, \tag{2.7}$$

then we have $x^2 \geq ba_d^2 d^{1-\beta}$, and condition (2.1) implies (2.3). Thus, the relation

$$d^{1-\beta} a_d^2 \asymp 1$$

determines the classification boundary in the general model of Section 1 if m is very large (satisfies (2.7)).

4°. Finally, note that we can control conditions (2.3) and (2.2) by their data-driven counterparts. In fact, $\max_{k=1,\dots,d} |\zeta_k| \leq \sqrt{2\log d}$ with probability tending to 1 as $d \to +\infty$. Hence, if (2.2) holds, then $M_Y \stackrel{\Delta}{=} \max_{1\leq k\leq d} SY^k \leq 3\sqrt{2\log d}$ with the same probability. It is therefore convenient to consider the following pre-classifier taking values in $\{0, 1, ND\}$ (ND means "No Decision", i.e., we need to switch to some other classifier):

$$\psi^{pre} = \begin{cases} 0 & \text{if} \quad T \leq 0,\ M_Y > 3\sqrt{2\log d}, \\ 1 & \text{if} \quad T > 0,\ M_Y > 3\sqrt{2\log d}, \\ ND & \text{if} \quad M_Y \leq 3\sqrt{2\log d}, \end{cases}$$

where T is given by (2.4). The argument in 2° implies that ψ^{pre} classifies successfully if ND is not chosen. Under condition (2.2) the pre-classifier chooses ND with probability tending to 1 and then we apply one of the classifiers suggested below in this paper. We prove their optimality under assumption (2.2).

The above remarks can be easily extended to the case of Gaussian errors with known variance $\sigma^2 > 0$ by using the normalization $Z^k/\sigma, SY^k/\sigma$. Moreover, they extend to the case of non-Gaussian errors under the Cramér condition and the additional assumption $m/\log d \to +\infty$ (cf. Section 5).

3 Classification boundary for moderately sparse vectors

In this section we consider the case of moderately sparse vectors. To simplify the notation, we set without loss of generality $\sigma = 1$. Assume that $R_d = d^{1/2-\beta} a_d$ satisfies:

$$\lim_{d\to+\infty} R_d = +\infty \tag{3.1}$$

and consider the classifier based on a linear statistic:

$$\psi^{lin} = \mathbb{1}_{\{T'>0\}}, \quad T' = \sum_{k=1}^{d} \left(Z^k - \frac{1}{2m} \sum_{i=1}^{m} Y_i^k \right).$$

Note that T' is similar to the statistic T defined in (2.4) with the difference that in T' we do not threshold to estimate the positions of non-zero ε_k. Indeed, here we do not necessarily assume (2.3), and thus there is no guarantee that ε_k can be correctly recovered.

Assume that η_j^k and ξ_i^k for all k, j, i are random variables with zero mean and variance 1 (we do not suppose here that η_j^k have the same distribution as ξ_i^k).

Then the means of Y_i^k and Z^k are $E^{(u)}(Y_i^k) = E_{H_1}^{(u)}(Z^k) = \varepsilon_k a_d$, $E_{H_0}^{(u)}(Z^k) = 0$, their variances are equal to 1, and we have:

$$E_{H_0}^{(u)}(T') = -\frac{a_d}{2}\sum_{k=1}^{d}\varepsilon_k, \quad E_{H_1}^{(u)}(T') = \frac{a_d}{2}\sum_{k=1}^{d}\varepsilon_k,$$
$$\mathrm{Var}_{H_0}^{(u)}(T') = \mathrm{Var}_{H_1}^{(u)}(T') = d\left(1+\frac{1}{4m}\right).$$

We consider now a vector $u \in \mathbb{R}^d$ of the form

$$u = (u_1, \ldots, u_d): \ u_k = a_d\varepsilon_k, \quad \varepsilon_k \in \{0,1\}, \quad \sum_{k=1}^{d}\varepsilon_k \geq cd^{1-\beta}. \tag{3.2}$$

By (3.2), Chebyshev's inequality and (3.1), we obtain

$$\begin{aligned} E_{H_0}^{(u)}(\psi) = P_{H_0}^{(u)}(T' > 0) &\leq P_{H_0}^{(u)}\Big(T' - E_{H_0}^{(u)}(T') > cd^{1-\beta}a_d/2\Big) \\ &\leq \frac{4d}{(cd^{1-\beta}a_d)^2}\left(1+\frac{1}{4m}\right) \to 0 \end{aligned}$$

as $d \to +\infty$. An analogous argument yields that $E_{H_1}^{(u)}(1-\psi) \to 0$. The convergence here is uniform in u satisfying (3.2), and thus uniform in $u \in U_{\beta,a_d}$. Therefore, we have the following result.

Theorem 3.1 *Let η_j^k and ξ_i^k for all k, j, i be random variables with zero mean and variance 1. If (3.1) holds, then successful classification is possible and it is realized by the classifier ψ^{lin}.*

Remark 3.1 We have proved theorem 3.1 with the set of vectors u defined by (3.2), which is larger than U_{β,a_d}. The upper bound on $\sum_k \varepsilon_k$ in the definition of U_{β,a_d} is not needed. Also the η_j^k need not have the same distribution as the ξ_i^k and their variances need not be equal to 1. It is easy to see that the result of theorem 3.1 remains valid if these random variables have unknown variances uniformly bounded by an (unknown) constant.

The corresponding lower bound is given in the next theorem. For $a > 0$, $t \in \mathbb{R}$, set

$$\ell_a(t) = f(t-a)/f(t), \quad D_a = \int \ell_a^2(t) f(t) dt,$$

and

$$D_d(m, a, \beta) = d^{1-2\beta} D_a^m (D_a - 1).$$

Theorem 3.2 *Let either $m \geq 1$ be fixed or $m = m_d \to +\infty$. If*

$$\lim_{d\to+\infty} D_d(m, a_d, \beta) = 0, \tag{3.3}$$

then successful classification is impossible.

Proof of theorem 3.2 is given in Section 6.

Corollary 3.1 *Let f be the density of standard normal distribution. If*

$$\lim_{d\to+\infty} R_d = 0, \tag{3.4}$$

then successful classification is impossible for $\beta \in (0, 1/2]$ and m fixed or for $\beta \in (0, 1/2)$ and $m = m_d \to +\infty$ such that $m = O(d^{1-2\beta})$.

Proof. For the standard normal errors we have $D_a = e^{a^2}$. Therefore, condition (3.3) can be satisfied only if $ma_d^2 = o(1)$ as $d \to +\infty$. Moreover, in this case

$$D_d(m, a_d, \beta) \asymp d^{1-2\beta} a_d^2 (1 + ma_d^2) \asymp R_d^2. \tag{3.5}$$

Thus, if $ma_d^2 = o(1)$, conditions (3.3) and (3.4) are equivalent. Now, (3.4) and the assumption $\beta \in (0, 1/2]$ imply $a_d = o(1)$. This proves the corollary for fixed m. Also, if $\beta \in (0, 1/2)$ and $m = m_d \to +\infty$ such that $m = O(d^{1-2\beta})$, then $ma_d^2 = O(R_d^2) = o(1)$. $\square$

Remark 3.2 Relation (3.5) is valid for a larger class of noise distributions, e.g., for non-Gaussian noise with finite Fisher information. Indeed, assume that $\ell_a(t)$ is $L_2(f)$-differentiable at point $a = 0$, i.e., there exists a function $\ell'(\cdot)$ such that

$$\|\ell_a(\cdot) - 1 - a\ell'(\cdot)\|_f = o(a), \quad 0 < \|\ell'(\cdot)\|_f < +\infty, \tag{3.6}$$

where $\|g(\cdot)\|_f^2 = \int_{\mathbb{R}} g^2(x)\, f(x)\, dx$. Observe that

$$\|\ell'(\cdot)\|_f^2 = \int_{\mathbb{R}} \frac{(f'(x))^2}{f(x)} dx \triangleq I(f)$$

is the Fisher information of f (with f' defined in a somewhat stronger sense than, for instance, in [5]. Under assumption (3.6) we have

$$D_a = 1 + \|\ell_a(\cdot) - 1\|_f^2, \ \|\ell_a(\cdot) - 1\|_f^2 = a^2(I(f) + o(1))$$

as $a \to 0$.

Combining remarks 3.1 and 3.2 with theorems 3.1 and 3.2 we see that relation (1.5) determines the classification boundary for $\beta \in (0, 1/2]$ and fixed m or for $\beta \in (0, 1/2)$ and $m \to +\infty$, $m = O(d^{1-2\beta})$, if the errors have zero mean, finite variance and finite Fisher information.

As corollaries of theorems 3.1 and 3.2 we can establish classification boundaries for particular choices of a_d. Recall that non-trivial results can be expected only if a_d satisfies (2.2). For instance, consider $a_d = d^{-s}$ with some $s > 0$. Then for fixed m the classification boundary in the region $\beta \in (0, 1/2]$ is given by $s = \beta - 1/2$, i.e., successful classification is possible if $s < 1/2 - \beta$, and is impossible if $s > 1/2 - \beta$. Other choices of a_d appear to be less interesting when $\beta \in (0, 1/2]$. For example,

in the next section we consider the sequence $a_d = s\sigma\sqrt{(\log d)/m}$ with some $s > 0$. If a_d is chosen in this way, successful classification is possible for all $\beta \in (0, 1/2]$ with no exception, so that there is no classification boundary in this range of β.

Finally, note that theorem 3.1 is valid for all $\beta \in (0, 1)$. However, for $\beta > 1/2$ its assumption $\lim_{d\to+\infty} R_d = +\infty$ guaranteeing successful classification is much too restrictive as compared to the correct classification boundary that we shall derive in the next section. The lower bound of theorem 3.2 is also valid for all $\beta \in (0, 1)$. However, we shall see in the next section that it is not tight for highly sparse vectors when $\beta > 3/4$ (cf. proof of theorem 4.1).

4 Classification boundary for highly sparse vectors

We now analyse the case of highly sparse vectors, i.e.,we suppose that $\beta \in (1/2, 1)$ if $\log m = o(\log d)$, and $\beta^* = \beta/(1-\gamma) \in (1/2, 1)$ if $\log m \sim \gamma \log d,\ \gamma \in (0, 1)$. We shall show that the classification boundary for this case is expressed in terms of the function

$$\phi(\beta) = \begin{cases} \phi_1(\beta) & \text{if } 1/2 < \beta \le 3/4, \\ \phi_2(\beta) & \text{if } 3/4 < \beta < 1, \end{cases}$$

where the functions ϕ_1 and ϕ_2 are defined in (1.9). Note that ϕ_1 and ϕ_2 are monotone increasing on $(1/2, 1)$, satisfy $\phi_1(\beta) \le \phi_2(\beta)$ for all $\beta \in (1/2, 1)$, and the equality $\phi_1(\beta) = \phi_2(\beta)(= 1/\sqrt{2})$ holds if and only if $\beta = 3/4$.

The following notation will be useful in the sequel:

$$T_d = \sqrt{\log d}, \quad s = s_d = a_d/\sigma T_d, \tag{4.1}$$

and

$$x = s\sqrt{m}, \quad x_0 = sm/\sqrt{m+1}, \quad x_1 = s\sqrt{m+1}, \quad x^* = \frac{x}{1-\gamma}. \tag{4.2}$$

Clearly, $x_0 < x < x_1$. We allow s, x, x_0, x_1 to depend on d but do not indicate this dependence in the notation for the sake of brevity. We shall also suppose throughout that (2.2) holds, so that $x_1 = O(1)$ as $d \to +\infty$.

4.1 Lower bound

The next theorem gives a lower bound of classification for highly sparse vectors.

Theorem 4.1 *Let the noise density f be Gaussian $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$. Assume that $\beta \in (1/2, 1)$ and $\limsup_{d\to+\infty} x_1 < \phi(\beta)$. Then successful classification is impossible for fixed m and for $m = m_d \to +\infty$.*

Proof of theorem 4.1 is given in Section 6.

Though theorem 4.1 is valid with no restriction on m, it does not provide a correct classification boundary if m is large, i.e., $\log m \sim \gamma \log d,\ \gamma \in (0, 1)$, as in Scenarios (C) and (D). The correct lower bound for large m is given in the next theorem.

Theorem 4.2 *Consider Scenario (C) with* $\beta^* = \beta/(1-\gamma) \in (1/2, 1)$ *and*

$$a_d = \sigma x \sqrt{(\log d)/m}.$$

Assume that $\limsup_{d\to+\infty} x^* < \phi(\beta^*)$. *Then successful classification is impossible.*

Proof of theorem 4.2 is given in Section 6.

Recall that, by an elementary argument, under Scenario (C) and for a_d as in theorem 4.2, successful classification is impossible if $\beta > 1-\gamma$ (cf. remark after (2.1)). This is the reason why in theorem 4.2 we consider only $\beta < 1-\gamma$.

4.2 Upper bounds for fixed m

We now propose optimal classifiers attaining the lower bound of theorem 4.1 under Scenario (A). First, we consider a procedure that attains the classification boundary only for $\beta \in [3/4, 1)$ but has a simple structure. Introduce the statistics

$$M_0 = \max_{1\le k\le d} SY^k, \quad M = \max_{1\le k\le d} SZ^k$$

where

$$SY^k = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} Y_i^k, \quad SZ^k = \frac{1}{\sqrt{m+1}} \left(Z^k + \sum_{i=1}^{m} Y_i^k \right). \tag{4.3}$$

Define

$$\Lambda_M = \frac{M}{\max(\sqrt{2}\,\sigma T_d, M_0)}.$$

Taking a small $c_0 > 0$, consider the classifier of the form:

$$\psi^{max} = \mathbb{1}_{\{\Lambda_M > 1+c_0\}}.$$

Theorem 4.3 *Consider Scenario (A). Let* $\beta \in (0,1)$ *and (2.2) hold. Then, for any* $c_0 > 0$,

$$\lim_{d\to+\infty} \sup_{u\in U_{\beta,a_d}} E_{H_0}^{(u)}(\psi^{max}) = 0. \tag{4.4}$$

If $\limsup_{d\to+\infty} x_1 < \phi_2(\beta)$, *then, for any* $c_0 > 0$,

$$\lim_{d\to+\infty} \sup_{u\in U_{\beta,a_d}} E_{H_1}^{(u)}(\psi^{max}) = 0. \tag{4.5}$$

If $\liminf_{d\to+\infty} x_1 > \phi_2(\beta)$, *then there exists* $c_0 > 0$ *such that*

$$\lim_{d\to+\infty} \sup_{u\in U_{\beta,a_d}} E_{H_1}^{(u)}(1-\psi^{max}) = 0. \tag{4.6}$$

Proof of theorem 4.3 is given in Section 7.

Theorems 4.1 and 4.3 (cf. (4.4) and (4.6) and the fact that $\phi(\beta) = \phi_2(\beta)$ for $\beta \in [3/4, 1)$) imply that ψ^{max} attains the classification boundary for $\beta \in [3/4, 1)$. On the other hand, (4.5) implies that for $\beta \in (1/2, 3/4)$ (where $\phi(\beta) = \phi_1(\beta) < \phi_2(\beta)$) the classifier ψ^{max} does not do the correct job. Its maximal risk $\mathcal{R}_{\mathcal{M}}$ is asymptotically 1, which is larger than the risk 1/2 of the simple random guess. We therefore introduce another classifier that has, however, a more involved structure. Consider the statistics

$$L_0(t) = \sum_{k=1}^{d}(\mathbb{1}_{\{SY^k > t\sigma T_d\}} - \Phi(-tT_d)), \quad \Delta_0(t) = \frac{L_0(t)}{\sqrt{d\Phi(-tT_d)}},$$

$$L(t) = \sum_{k=1}^{d}(\mathbb{1}_{\{SZ^k > t\sigma T_d\}} - \Phi(-tT_d)), \quad \Delta(t) = \frac{L(t)}{\sqrt{d\Phi(-tT_d)}}$$

where $t \in \mathbb{R}$, Φ is the standard normal cumulative distribution function and the statistics SY^k, SZ^k are defined in (4.3). Consider the grid

$$t_l = lh, \quad l = 1, ..., N, \quad t_N = \sqrt{2}\,\sigma, \tag{4.7}$$

with a step $h > 0$ depending on d and such that $h = o(1)$, $T_d h \to +\infty$. This implies that $1 \ll N \ll T_d$ as $d \to +\infty$ (here and below $v_d \ll w_d$ for $v_d > 0$ and $w_d > 0$ depending on d means that $\lim_{d\to+\infty} v_d/w_d = 0$). Set

$$\Delta_0 = \max_{1\le l\le N} \Delta_0(t_l), \quad \Delta = \max_{1\le l\le N} \Delta(t_l), \quad \Lambda^* = \frac{\Delta}{H + \Delta_0},$$

where $H = H_d$ is such that

$$d^{bh} \ll H \ll d^B \tag{4.8}$$

for any $B > 0$, $b > 0$ and any $d > d_0(B, b)$ where $d_0(B, b)$ is a constant depending only on B and b (such an H can be always determined depending on the choice of h). Consider now the classifier of the form

$$\psi_m^* = \mathbb{1}_{\{\Lambda^* > H\}}.$$

Theorem 4.4 *Consider Scenario (A) with $\beta \in (1/2, 1)$ and assume (2.2). Then*

$$\lim_{d\to+\infty} \sup_{u\in U_{\beta,a_d}} E_{H_0}^{(u)}(\psi_m^*) = 0. \tag{4.9}$$

If $\liminf_{d\to+\infty} x_1 > \phi(\beta)$ and $\limsup_{d\to+\infty} x_0 < \sqrt{2}$, then

$$\lim_{d\to+\infty} \sup_{u\in U_{\beta,a_d}} E_{H_1}^{(u)}(1 - \psi_m^*) = 0. \tag{4.10}$$

Proof of theorem 4.4 is given in Section 7.

Theorems 4.1, 4.3 and 4.4 show that the classification boundary for highly sparse vectors (i.e., for $\beta \in (1/2, 1)$) is given by (1.7). Furthermore, the classifier ψ_m^* is optimal (attains the classification boundary) for $\beta \in (1/2, 1)$, except for the case $\limsup_{d\to+\infty} x_0 \ge \sqrt{2}$, which is already covered by the classifier ψ^{max}. Indeed, $x_0 \ge \sqrt{2}$ implies that $x_1 \ge \sqrt{2}(1 + 1/m) > \phi_2(\beta)$ for all $\beta \in (1/2, 1)$.

4.3 Upper bounds for $m \to +\infty,\ \log m = o(\log d)$

In this subsection we analyse Scenario (B). Then $m = m_d \to +\infty$, $\log m = o(\log d)$ as $d \to +\infty$ and the classifier ψ_m^* is not, in general, optimal. Nevertheless, we propose another classifier ψ_∞^*, which attains essentially the same classification boundary as in Subsection 4.2 above. Introduce the statistics

$$\Delta(t) = \frac{1}{\sigma\sqrt{d\Phi(-tT_d)}} \sum_{k=1}^{d} Z^k \mathbb{1}_{\{SY^k > t\sigma T_d\}}, \quad \Delta = \max_{1\le l\le N} \Delta(t_l),$$

where the maximum is taken over the grid (4.7). Here and below we use the same notation $\Delta(t)$, Δ as previously for different ratio statistics, since it causes no ambiguity. Set also

$$\Delta_* = \sum_{k=1}^{d} \mathbb{1}_{\{SY^k > \sqrt{2}\,\sigma T_d\}}$$

and define

$$\Lambda_\infty^* = \frac{\Delta}{\sqrt{H + \Delta_*}}, \quad \psi_\infty^* = \mathbb{1}_{\{\Lambda_\infty^* > H\}}, \tag{4.11}$$

where H satisfies (4.8).

Theorem 4.5 *Consider Scenario (B). Let $\beta \in (1/2, 1)$ and let (2.2) hold. Then*

$$\lim_{d\to+\infty} \sup_{u\in U_{\beta,a_d}} E_{H_0}^{(u)}(\psi_\infty^*) = 0. \tag{4.12}$$

If $\liminf_{d\to+\infty} x > \phi(\beta)$, then

$$\lim_{d\to+\infty} \sup_{u\in U_{\beta,a_d}} E_{H_1}^{(u)}(1 - \psi_\infty^*) = 0. \tag{4.13}$$

Proof of theorem 4.5 is given in Section 7.

4.4 Upper bound for Scenario (C)

We now suggest an asymptotically optimal classifier for Scenario (C). For $t \ge 0$ we introduce the statistics

$$L^1(t) = \sum_{k=1}^{d} Z^k \mathbb{1}_{\{SY^k > \sigma t T_d\}}, \quad L^0(t) = \sum_{k=1}^{d} \mathbb{1}_{\{SY^k > \sigma t T_d\}}, \quad \Delta(t) = \frac{L^1(t)}{\sigma\sqrt{N^2 + L^0(t)}},$$

Take a grid $t_1, \dots, t_N$ of the form (4.7) and define the classifier

$$\psi_\infty = \mathbb{1}_{\{\Delta > 4\,N\}}, \quad \text{where} \quad \Delta = \max_{1\le l\le N} \Delta(t_l).$$

Theorem 4.6 *Consider Scenario (C) with* $a_d = \sigma x\sqrt{(\log d)/m}$. *Let* $\beta^* = \beta/(1-\gamma) \in (1/2, 1)$ *and let (2.2) hold. Then*

$$\lim_{d\to+\infty} \sup_{u\in U_{\beta,a_d}} E^{(u)}_{H_0}(\psi_\infty) = 0. \tag{4.14}$$

If $\liminf_{d\to+\infty} x^* > \phi(\beta^*)$, *then*

$$\lim_{d\to+\infty} \sup_{u\in U_{\beta,a_d}} E^{(u)}_{H_1}(1-\psi_\infty) = 0. \tag{4.15}$$

Proof of theorem 4.6 is given in Section 7.

5 Extensions

5.1 Unknown variances

The classifiers proposed in the previous section can be easily extended to the model with unknown variance σ^2, so that the results of theorems 4.3, 4.4, 4.5 and 4.6 remain valid. We present here the general lines of such a modification without going into the details of the proofs that do not differ much from those in Section 7.

First, note that there exists an estimator $\hat\sigma^2_d$ satisfying

$$\hat\sigma^2_d \;=\; \sigma^2 + \eta_d, \tag{5.1}$$

where $\eta_d \to 0$ in $P^{(u)}_{H_0}$-probability, and

$$\hat\sigma^2_d \;=\; \sigma^2 + O(d^{-\beta}a_d^2) + (1 + d^{-\beta/2}a_d)^{1/2}\eta'_d, \tag{5.2}$$

where $\eta'_d \to 0$ in $P^{(u)}_{H_1}$-probability, uniformly in $u \in U_{\beta,a_d}$, as $d \to +\infty$.

For example, we can take the standard sample variance

$$\hat\sigma^2_d = \frac{1}{d}\sum_{k=1}^{d}(Z^k)^2.$$

Assume that η^k_j are i.i.d. $\mathcal{N}(0,\sigma^2)$ random variables with unknown σ. Then (5.1) and (5.2) are satisfied. In fact, it is easy to see that

$$E^{(u)}_{H_0}(\hat\sigma^2_d) = \sigma^2, \quad E^{(u)}_{H_1}(\hat\sigma^2_d) = \sigma^2 + \frac{1}{d}\sum_{k=1}^{d} u_k^2 = \sigma^2 + O(a_d^2 d^{-\beta}),$$

and analogously

$$\mathrm{Var}^{(u)}_{H_0}(\hat\sigma^2_d) = \frac{2\sigma^4}{d}, \quad \mathrm{Var}^{(u)}_{H_1}(\hat\sigma^2_d) = \frac{1}{d}\left(2\sigma^4 + O(d^{-\beta}a_d^2)\right) = o(1 + d^{-\beta}a_d^2)$$

as $d \to +\infty$. Applying Chebyshev's inequality, we get (5.1) and (5.2). We also note that these relations hold under much weaker assumptions than the normality

of η_j^k. It suffices to have, for example, independent random variables η_j^k such that $E(\eta_j^k) = 0$, $E[(\eta_j^k)^2] = \sigma^2$ and $\max_{j,k} E[(\eta_j^k)^4] < +\infty$.

We now discuss how to modify the proposed classifiers using $\hat{\sigma}_d$. For ψ^{pre} and ψ^{max}, we replace the unknown σ in their definitions by $\hat{\sigma}_d$ and change $\sqrt{2\log d}$ into $\sqrt{b\log d}$, $b > 2$ for ψ^{pre}. If $R_d = O(1)$ (which is the case for highly sparse vectors under (4.1)), then $d^{-\beta}a_d^2 = o(1)$ and (5.1) implies that the ratio $\hat{\sigma}_d/\sigma$ is close to 1 in $P_{H_1}^{(u)}$-probability as well. Therefore, for the study of the variance modified versions of classifiers ψ^{pre}, ψ^{max}, we can use not only (5.1) but also the fact that $\hat{\sigma}_d^2 = \sigma^2 + \bar{\eta}_d$ where $\bar{\eta}_d \to 0$ in $P_{H_1}^{(u)}$-probability. Thus, the desired upper bounds for these classifiers follow in an easy way from the results in Section 4.2.

For the classifier ψ_m^*, we replace the statistics $L_0(t), L(t), \Delta_0(t), \Delta(t)$ by

$$\begin{aligned} L_0(t) &= \sum_{k=1}^{d}(\mathbb{1}_{\{SY^k > tT_d\}} - \mathbb{1}_{\{Z^k > tT_d\}}), \quad \Delta_0(t) = \frac{L_0(t)}{\sqrt{d\Phi(-tT_d/\hat{\sigma}_d)}}, \\ L(t) &= \sum_{k=1}^{d}(\mathbb{1}_{\{SZ^k > tT_d\}} - \mathbb{1}_{\{Z^k > tT_d\}}), \quad \Delta(t) = \frac{L(t)}{\sqrt{d\Phi(-tT_d/\hat{\sigma}_d)}}, \end{aligned}$$

and we take a grid

$$t_l = lh,\ l = 1, ..., N, t_N = \sqrt{2}\,\hat{\sigma}_d + O(h),$$

with step h as in (4.7). The cardinality N of the grid thus becomes a random variable. However, the relation $N = O(T_d)$ holds true in probability under (5.1). Note that the modified statistics $\Delta_0(t)$ and $\Delta(t)$ contain the additional factor $A(t) = \sqrt{\Phi(-tT_d/\sigma)/\Phi(-tT_d/\hat{\sigma}_d)}$ as compared to the original ones. If (5.1) holds, these factors are (in probability) of the form $\exp(o(T_d))$ uniformly in $t = O(1)$. Under $P_{H_s}^{(u)}$, $s = 0, 1$, the expectations of the summands with $\varepsilon_k = 0$ in $L_0(t)$ and $L(t)$ vanish. The other elements of the proof for the modified statistics are similar to those in Section 7.

For the classifier ψ_∞^*, we replace the statistics $\Delta(t)$ and Δ_0 by

$$\Delta(t) = \frac{1}{\hat{\sigma}_d\sqrt{d\Phi(-tT_d/\hat{\sigma}_d)}}\sum_{k=1}^{d} Z^k \mathbb{1}_{\{SY^k > tT_d\}}, \quad \Delta = \max_{1\le l\le N}\Delta(t_l),$$

with the same grid as above, and

$$\Delta_0 = \sum_{k=1}^{d}\mathbb{1}_{\{SY^k > \hat{\sigma}_d\sqrt{2}T_d\}}.$$

We make similar modifications for the classifier ψ_∞. The arguments above are enough for the proof of Sections 7.3, 7.4 to hold through.

5.2 Non-Gaussian noise

We now discuss an extension of our results to Scenario (D). The remarks on the pre-classifier ψ^{pre} in Section 2 and the proofs of the upper bounds in Section 7 are

only based on the constraint (2.2) and the following property of the tails of the Gaussian distribution:

$$\log P(S\zeta^m > \sigma t) \sim -\frac{t^2}{2}, \quad t \in [U_0, U_1] \quad \text{for } U_0 \to +\infty \text{ and } U_1 = O(T_d). \tag{5.3}$$

Here $S\zeta^m = \frac{1}{\sqrt{m}} \sum_{i=1}^m \zeta_i$, and ζ_i are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables.

Indeed, in Subsection 7.4 we can write $\Phi(-tT_d)$ as $P(S\zeta^m > \sigma t T_d)$. From (5.3) we deduce

$$P(S\zeta^m > tT_d/\sigma) = A_d d^{-(t_+)^2/2}, \quad t_+ = \max(0, t),$$

where A_d satisfies (7.4) for $t_+ = O(1)$. This is exactly the relation (7.3), which is also the only property of the noise distribution needed for the proofs in Subsection 7.4.

If m is large enough, relation (5.3) holds not only for the Gaussian ζ_i. It suffices to have the i.i.d. ζ_i with $E\zeta_i = 0$, $E(\zeta_i^2) = \sigma^2 > 0$ satisfying the Cramér condition:

$$\exists \quad h_0 > 0 : \quad E\left(e^{h\zeta_i}\right) < +\infty, \quad \forall\ h \in (-h_0, h_0).$$

and $m \gg \log d$. In fact, using theorem 5.23 in [13] we get that, under the Cramér condition and for $t = o(\sqrt{m})$,

$$P(S\zeta^m > \sigma t) = \Phi(-t) \exp\left(\frac{t^3}{\sqrt{m}} \lambda\left(\frac{t}{\sqrt{m}}\right)\right) \left\{1 + O\left(\frac{t+1}{m}\right)\right\},$$

where $\lambda(t)$ is the Cramér series. Inserting here the expression for the Cramér series and the relation $\log \Phi(-t) = -t^2/2 - \log t + O(1)$ as $t \to +\infty$, we obtain

$$\log P(S\zeta^m > \sigma t) = -\frac{t^2}{2}\left(1 + O\left(\frac{t}{\sqrt{m}}\right) + o(1)\right) \sim -\frac{t^2}{2}$$

as $t \to +\infty$, $t = o(\sqrt{m})$. These remarks allow us to follow the proof of theorem 4.6 in Section 7 leading to the next result.

Theorem 5.1 *Consider Scenario (D) with $a_d = \sigma x \sqrt{(\log d)/m}$. Let $\beta^* = b/(1-\gamma) \in (1/2, 1)$ and let (2.2) hold. Then*

$$\lim_{d \to +\infty} \sup_{u \in U_{\beta, a_d}} E_{H_0}^{(u)}(\psi_\infty) = 0. \tag{5.4}$$

If $\liminf_{d \to +\infty} x^* > \phi(\beta^*)$, then

$$\lim_{d \to +\infty} \sup_{u \in U_{\beta, a_d}} E_{H_1}^{(u)}(1 - \psi_\infty) = 0. \tag{5.5}$$

5.3 Adaptive procedures

We have proposed several classifiers, which attain the classification boundary under various conditions on m, a_d, β. In order to obtain an adaptive procedure that attains this boundary simultaneously for several domains of m, a_d, β, it suffices to combine the classifiers in the following way. We start with the pre-classifier ψ^{pre}. If it outputs "No Decision", then we combine the classifiers ψ^{lin}, ψ^{max} and ψ^*_m using the Bonferroni device, i.e., our classifier will be $\max(\psi^{lin}, \psi^{max}, \psi^*_m)$. This means that we allocate Z to the P_Y-population iff it is allocated to P_Y by at least one of the three classifiers. Analogously, if $m \to +\infty$, then we classify by $\max(\psi^{lin}, \psi^*_\infty)$ or by $\max(\psi^{lin}, \psi_\infty)$.

6 Proof of the lower bounds

In this section we prove theorems 3.2, 4.1 and 4.2. Without loss of generality we consider only the case $\mathcal{R}_{\max} = 1/2$ (cf. (1.3)). Observe that if a probability measure μ^d on $\mathbb{R}^d$ is such that

$$\mu^d(U_{\beta,a}) = 1 + o(1) \tag{6.1}$$

as $d \to +\infty$, then

$$\begin{aligned}
\sup_{u \in U_{\beta,a_d}} \mathcal{R}_M(\psi) &\geq \max\Big(\int E^{(u)}_{H_0}(\psi)\mu^d(du), \int E^{(u)}_{H_1}(1-\psi)\mu^d(du)\Big) + o(1) \\
&\geq \frac{1}{2}\Big(\int E^{(u)}_{H_0}(\psi)\mu^d(du) + \int E^{(u)}_{H_1}(1-\psi)\mu^d(du)\Big) + o(1) \\
&= \frac{1}{2}\int \left(\psi + (1-\psi)\frac{d\mathsf{P}_{H_1}}{d\mathsf{P}_{H_0}}\right) d\mathsf{P}_{H_0} + o(1)
\end{aligned} \tag{6.2}$$

where P_{H_s}, $s = 0, 1$, are the "posterior" probability measures defined by

$$\mathsf{P}_{H_s}(A) = \int P^{(u)}_{H_s}(A)\mu^d(du)$$

for any Borel set A of $\big(\mathbb{R}^d\big)^m \times \mathbb{R}^d$. In view of (6.2) , if the likelihood ratio

$$L(\mathbf{Y}, Z) = \frac{d\mathsf{P}_{H_1}}{d\mathsf{P}_{H_0}}$$

satisfies

$$L(\mathbf{Y}, Z) \to 1 \quad \text{in } \mathsf{P}_{H_0}\text{-probability} \tag{6.3}$$

as $d \to +\infty$, then the left-hand side of (1.3) is greater than or equal to $1/2$. This immediately entails (1.3) because the risk of the simple random guess classifier equals $1/2$.

Since $\mathsf{E}_{H_0}(L(\mathbf{Y}, Z) - 1)^2 = \mathsf{E}_{H_0}L^2(\mathbf{Y}, Z) - 1$, relation (6.3) holds if

$$\limsup_{d \to +\infty} \mathsf{E}_{H_0}L^2(\mathbf{Y}, Z) = 1. \tag{6.4}$$

Based on these remarks, the proofs of theorems 3.2 and 4.1 will proceed by constructing a prior measure μ^d satisfying (6.1) and proving (6.4).

In this section we assume without loss of generality that $\sigma = 1$ and that the constants c, C in the definition of U_{β,a_d} are such that $c < 1 < C$.

The prior measure that we choose here is of the form $\mu^d(du) = \prod_{k=1}^d \mu(du_k)$ where $\mu = (1-p)\delta_0 + p\delta_{a_d}$, $p = d^{-\beta}$, and δ_t is the Dirac mass at point $t \in \mathbb{R}$. In other words, the prior measure corresponds to $u_k = a_d\varepsilon_k$ with i.i.d Bernoulli random entries ε_k that take value 1 with probability $p = d^{-\beta}$ and value 0 with probability $1 - d^{-\beta}$.

Lemma 6.1 *Let $0 < c < 1 < C < +\infty$. Then the prior measure μ^d defined above satisfies (6.1).*

Proof of Lemma 6.1. Set $G(u) = \sum_{k=1}^d u_k$. We have to check that

$$\mu^d(G(u) > a_d C d^{1-\beta}) \to 0, \quad \mu^d(G(u) < a_d c d^{1-\beta}) \to 0.$$

Since

$$E_{\mu^d}(G(u)) = a_d dp = a_d d^{1-\beta}, \quad \mathrm{Var}_{\mu^d}(G(u)) = a_d^2 dp(1-p) \sim a_d^2 d^{1-\beta} \quad \forall\ \beta \in (0,1),$$

it follows from Chebyshev's inequality that

$$\begin{aligned} \mu^d(G(u) > a_d C d^{1-\beta}) &\leq \frac{1}{d^{1-\beta}(C-1)^2} \to 0, \\ \mu^d(G(u) < a_d c d^{1-\beta}) &\leq \frac{1}{d^{1-\beta}(1-c)^2} \to 0 \end{aligned}$$

as $d \to +\infty$. □

It remains now to prove that (6.4) holds under the assumptions of theorems 3.2 and 4.1.

We shall need some notation. For $a \in \mathbb{R}$ define the probability densities

$$f_a(\mathbf{Y}^k) = \prod_{i=1}^m f(Y_i^k - a), \quad f_a(\mathbf{Y}^k, Z^k) = f_a(\mathbf{Y}^k) f(Z^k - a). \tag{6.5}$$

Let $P_0 = \prod_{k=1}^d P_{0,k}$ be the probability measure that corresponds to the pure noise. Here the measure $P_{0,k}$ has the density $f_0(\mathbf{Y}^k, Z^k) = f_0(\mathbf{Y}^k) f(Z^k)$ and $E_{0,k}(\cdot)$ denotes the expectation under $P_{0,k}$.

Next, write $\mathsf{P}_{H_s} = \prod_{k=1}^d P_{H_s,k}$, $s = 0, 1$, where the probability measures $P_{H_s,k}$ $s = 0, 1$, have the densities

$$f_{H_s,k}(\mathbf{Y}^k, Z^k) = (1-p) f_0(\mathbf{Y}^k) f(Z^k) + p f_{a_d}(\mathbf{Y}^k) f(Z^k - s a_d).$$

We denote by $E_{H_s,k}, s = 0, 1$, the corresponding expectations. The measures P_{H_s} have the following densities :

$$f_{H_s}(\mathbf{Y}, Z) = \prod_{k=1}^d f_{H_s,k}(\mathbf{Y}^k, Z^k).$$

The likelihood ratio is of the form

$$L(\mathbf{Y}, Z) = \frac{d\mathsf{P}_{H_1}}{d\mathsf{P}_{H_0}} = \prod_{k=1}^{d} L_k(\mathbf{Y}^k, Z^k),$$

where

$$L_k(\mathbf{Y}^k, Z^k) = \frac{(1-p) + pL(\mathbf{Y}^k, Z^k)}{(1-p) + pL(\mathbf{Y}^k)} = \frac{1 + p(L(\mathbf{Y}^k, Z^k) - 1)}{1 + p(L(\mathbf{Y}^k) - 1)},$$

and we set

$$L(\mathbf{Y}^k) = \prod_{j=1}^{m} \ell_{a_d}(Y_j^k), \quad L(\mathbf{Y}^k, Z^k) = L(\mathbf{Y}^k)\ell_{a_d}(Z^k)$$

where $\ell_{a_d}(t) = f(t - a_d)/f(t)$. It will be convenient to write L_k in the form

$$L_k(\mathbf{Y}^k, Z^k) = 1 + \Delta_k, \quad \Delta_k = \frac{pL(\mathbf{Y}^k)(\ell_{a_d}(Z^k) - 1)}{1 + p(L(\mathbf{Y}^k) - 1)}. \tag{6.6}$$

6.1 Proof of theorem 3.2

Recall that $\frac{dP_{H_0,k}}{dP_{0,k}} = 1 + p(L(\mathbf{Y}^k) - 1)$. Since $E_{H_0,k}(\Delta_k) = 0$, we obtain

$$\begin{aligned}
\mathsf{E}_{H_0}(L^2(\mathbf{Y}, Z)) &= \prod_{k=1}^{d} \left(1 + E_{H_0,k}\Delta_k^2\right) \leq \exp\left(\sum_{k=1}^{d} E_{H_0,k}\Delta_k^2\right) \\
&= \exp\left(\sum_{k=1}^{d} E_{0,k}\left(\Delta_k^2 \frac{dP_{H_0,k}}{dP_{0,k}}\right)\right) \\
&\leq \exp\left(\frac{p^2}{1-p} \sum_{k=1}^{d} E_{0,k}L^2(\mathbf{Y}^k)\, E_{0,k}(\ell_{a_d}(Z^k) - 1)^2\right) \\
&= \exp\left(\frac{dp^2 D_{a_d}^m (D_{a_d} - 1)}{1-p}\right), \qquad (6.7)
\end{aligned}$$

where

$$D_{a_d} = \int_{\mathbb{R}} \ell_{a_d}^2(t) f(t) dt = \int_{\mathbb{R}} \frac{f^2(t - a_d)}{f(t)} dt.$$

Since $p = d^{-\beta} \to 0$, relation (6.4) holds if

$$D_d(m, a_d, \beta) = d^{1-2\beta} D_{a_d}^m (D_{a_d} - 1) \to 0. \tag{6.8}$$

This completes the proof of theorem 3.2.

6.2 Proof of theorem 4.1

Assume w.l.o.g. that $\sigma = 1$. Then f is the standard normal density, and thus $D_a = e^{a^2}$. We shall assume that x_1 is fixed; the general case can be treated in a

similar way by passing to subsequences $x_{1,d} \to x_1 \geq 0$. By (1.6), the condition (6.8) takes the form

$$d^{1-2\beta} \exp((m+1)a_d^2) = d^{1-2\beta+x_1^2} \to 0, \tag{6.9}$$

In other terms, the proof of theorem 3.2 implies that successful classification is impossible if

$$x_1^2 - 2\beta + 1 < 0. \tag{6.10}$$

This bound applies for any $\beta \in (1/2, 1)$, and it yields the result of theorem 4.1 for $\beta \in (1/2, 3/4]$. It remains to show that a bound better than (6.10) can be obtained for $\beta \in (3/4, 1)$, namely

$$x_1^2 - 2\left(1 - \sqrt{1-\beta}\right)^2 < 0. \tag{6.11}$$

In order to prove this, set

$$SY_k = \sum_{j=1}^{m} Y_j^k, \quad SZ_k = SY_k + Z^k, \quad k = 1, ..., d, \quad T_{l,d} = \sqrt{2l \log d},$$

and introduce the events

$$\mathcal{A}_{SY,k} = \{SY_k < T_{m,d}\}, \quad \mathcal{A}_{SZ,k} = \{SZ_k < T_{m+1,d}\},$$
$$\mathcal{A}_{SY} = \bigcap_{k=1}^{d} \mathcal{A}_{SY,k}, \quad \mathcal{A}_{SZ} = \bigcap_{k=1}^{d} \mathcal{A}_{SZ,k}.$$

Observe that since

$$P_{0,k}(SY_k \geq T_{m,d}) = P_{0,k}(SZ_k \geq T_{m+1,d}) = \Phi\left(-\sqrt{2\log d}\right) = o(d^{-1}),$$

we have

$$P_0(\mathcal{A}_{SY}) \to 1, \quad P_0(\mathcal{A}_{SZ}) \to 1.$$

Moreover

$$\begin{aligned}
P_{H_0,k}(SY_k \geq T_{m,d}) &= P_{H_1,k}(SY_k \geq T_{m,d}) \\
&= (1-p)P_{0,k}(SY_k \geq T_{m,d}) + pP_{0,k}(SY_k \geq T_{m,d} - ma_d), \\
pP_{0,k}(SY_k \geq T_{m,d} - ma_d) &= d^{-\beta}\Phi\left(a_d\sqrt{m} - \sqrt{2\log d}\right) \\
&< d^{-\beta}\Phi\left(a_d\sqrt{m+1} - \sqrt{2\log d}\right) \asymp \frac{d^{-g}}{\sqrt{\log d}} = o(d^{-1}),
\end{aligned}$$

where $g \stackrel{\Delta}{=} \beta + \left(\sqrt{2} - x_1\right)^2/2 \geq 1$ in view of (6.11). Analogously, we have for $s = 0, 1$:

$$\begin{aligned}
P_{H_s,k}(SZ_k \geq T_{m+1,d}) &= (1-p)P_{0,k}(SZ_k \geq T_{m+1,d}) \\
&\quad + pP_{0,k}(SZ_k \geq T_{m+1,d} - (m+s)a_d), \\
pP_{0,k}(SZ_k \geq T_{m+1,d} - ma_d) &\leq pP_{0,k}(SZ_k \geq T_{m+1,d} - (m+1)a_d) \\
&= d^{-\beta}\Phi\left(a_d\sqrt{m+1} - \sqrt{2\log d}\right) \\
&\asymp \frac{d^{-g}}{\sqrt{\log d}} = o(d^{-1}).
\end{aligned}$$

Thus,

$$\mathsf{P}_{H_s}(\mathcal{A}_{SY}) \to 1, \quad \mathsf{P}_{H_s}(\mathcal{A}_{SZ}) \to 1, \quad s = 0, 1, \tag{6.12}$$

as $d \to +\infty$. Set $\hat{L}_k(\mathbf{Y}^k, Z^k) = L_k(\mathbf{Y}^k, Z^k)\mathbb{1}_{\{\mathcal{A}_{SY,k} \bigcap \mathcal{A}_{SZ,k}\}}$, $\hat{\Delta}_k = \Delta_k \mathbb{1}_{\{\mathcal{A}_{SY,k} \bigcap \mathcal{A}_{SZ,k}\}}$, where Δ_k is defined by (6.6), and $\hat{L}(\mathbf{Y}, Z) = \prod_{k=1}^{d} \hat{L}_k(\mathbf{Y}, Z)$. Using (6.12) we get that the main term in (6.2) satisfies

$$\begin{aligned}
\int (\psi + (1-\psi)L(\mathbf{Y}, Z))\, d\mathsf{P}_{H_0} &= \int \left(\psi + (1-\psi)\hat{L}(\mathbf{Y}, Z)\right) d\mathsf{P}_{H_0} \\
&\quad + \int (1-\psi)\mathbb{1}_{\{\bar{\mathcal{A}}_{SY} \bigcup \bar{\mathcal{A}}_{SZ}\}} d\mathsf{P}_{H_1} \\
&= \int \left(\psi + (1-\psi)\hat{L}(\mathbf{Y}, Z)\right) d\mathsf{P}_{H_0} + o(1)
\end{aligned}$$

as $d \to +\infty$. Repeating the argument after (6.2) we see that to prove the theorem it suffices to show that

$$\hat{L}(\mathbf{Y}, Z) \to 1 \quad \text{in } \mathsf{P}_{H_0}\text{-probability.} \tag{6.13}$$

Using (6.12) we obtain that $\mathsf{E}_{H_0}\hat{L}(\mathbf{Y}, Z) = \mathsf{P}_{H_1}(\mathcal{A}_{SY} \cap \mathcal{A}_{SZ}) \to 1$. Therefore, to show (6.13) it suffices to prove that (cf. (6.4)):

$$\limsup_{d \to +\infty} \mathsf{E}_{H_0}\hat{L}^2(\mathbf{Y}, Z) = 1. \tag{6.14}$$

We now prove (6.14). First note that, as follows from the displays preceding (6.12),

$$E_{H_0,k}(\hat{\Delta}_k) = P_{H_1,k}(\mathcal{A}_{SY,k} \cap \mathcal{A}_{SZ,k}) - P_{H_0,k}(\mathcal{A}_{SY,k} \cap \mathcal{A}_{SZ,k}) = o(d^{-1}),$$

and

$$0 \le \hat{L}_k(\mathbf{Y}^k, Z^k) \le 1 + \hat{\Delta}_k.$$

Therefore, arguing as in (6.7) we obtain

$$\begin{aligned}
\mathsf{E}_{H_0}(\hat{L}^2(\mathbf{Y}, Z)) &= \prod_{k=1}^{d} E_{H_0,k}(\hat{L}_k^2(\mathbf{Y}^k, Z^k)) \le \prod_{k=1}^{d} \left(1 + E_{H_0,k}\hat{\Delta}_k^2 + 2E_{H_0,k}\hat{\Delta}_k\right) \\
&\le \exp\left(\sum_{k=1}^{d} E_{H_0,k}\hat{\Delta}_k^2 + 2\sum_{k=1}^{d} E_{H_0,k}\hat{\Delta}_k\right) \\
&= \exp\left(\sum_{k=1}^{d} E_{0,k}\left(\hat{\Delta}_k^2 \frac{dP_{H_0,k}}{dP_{0,k}}\right) + o(1)\right) \\
&\le \exp\left(\frac{p^2}{1-p}\sum_{k=1}^{d} E_{0,k}\left[L^2(\mathbf{Y}^k)(\ell_{a_d}(Z^k) - 1)^2 \mathbb{1}_{\{\mathcal{A}_{SY,k} \bigcap \mathcal{A}_{SZ,k}\}}\right] + o(1)\right) \\
&= \exp\left(\frac{dp^2 A}{1-p} + o(1)\right),
\end{aligned} \tag{6.15}$$

where $A = E_{0,1}\left[L^2(\mathbf{Y}^1)(\ell_{a_d}(Z^1)-1)^2 \mathbb{1}_{\{\mathcal{A}_{SY,1}\bigcap \mathcal{A}_{SZ,1}\}}\right]$. Observe that

$$A \le B + C, \quad B = E_{0,1}\left(L^2(\mathbf{Y}^1)\ell^2_{a_d}(Z^1)\mathbb{1}_{\{\mathcal{A}_{SZ,1}\}}\right), \quad C = E_{0,1}\left(L^2(\mathbf{Y}^1)\mathbb{1}_{\{\mathcal{A}_{SY,1}\}}\right).$$

Setting $b_l = a_d\sqrt{l}$ with $l = m$ or $m+1$, $T_d = \sqrt{2\log d}$, we can write

$$B = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{T_d} e^{-b^2_{m+1}+2b_{m+1}t-t^2/2}dt = e^{b^2_{m+1}}\Phi(T_d - 2b_{m+1}),$$

and analogously,

$$C = e^{b^2_m}\Phi(T_d - 2b_m).$$

Recall that we consider $\beta \in (3/4, 1)$ under assumption (6.11). Thus,

$$1/2 < 2\beta - 1 \le (m+1)s^2 \le 2(1-\sqrt{1-\beta})^2. \tag{6.16}$$

Next, by(1.6),

$$-T_d + 2b_{m+1} = \sqrt{2\log d}\,\left(\sqrt{2}\ s\sqrt{m+1} - 1\right) = \sqrt{2\log d}\,\left(\sqrt{2}x_1 - 1\right).$$

Thus, for $1/\sqrt{2} < x_1 \le \sqrt{2}$ we have

$$dp^2 B = dp^2 e^{b^2_{m+1}}\Phi(T_d - 2b_{m+1}) \asymp \frac{d^{-2\beta+2\sqrt{2}x_1 - x_1^2}}{\sqrt{\log d}} = \frac{d^{2(1-\beta)-(\sqrt{2}-x_1)^2}}{\sqrt{\log d}}.$$

Here the exponent is $2(1-\beta) - \left(\sqrt{2}-x_1\right)^2 \le 0$ in view of the last inequality in (6.16). Therefore $dp^2B = o(1)$ as $d \to +\infty$. In order to control dp^2C observe that the function $b \mapsto e^{b^2}\Phi(T-2b)$ is increasing for b large enough and $T > b$. Therefore $C \le B$ for d large enough and $dp^2C = o(1)$ as well. Thus $dp^2A = o(1)$ as $d \to +\infty$, and (6.14) follows. This completes the proof of theorem 4.1.

6.3 Proof of theorem 4.2

Assume w.l.o.g. that $\sigma = 1$. By assumptions of the theorem, $\log m \sim \gamma \log d$, $\gamma \in (0,1)$ and

$$\beta \in ((1-\gamma)/2, 1-\gamma), \quad a = a_d = x\sqrt{\log(d)/m}, \quad x = O(1). \tag{6.17}$$

In view of the first two lines of (6.7), it suffices to show that

$$\sum_{k=1}^{d} E_{H_{0,k}}\Delta_k^2 = dE_{H_{0,1}}\Delta_1^2 = o(1). \tag{6.18}$$

Set

$$\Delta_{\mathbf{Y}^1} = \frac{pL(\mathbf{Y}^1)}{1-p+pL(\mathbf{Y}^1)}$$

and observe that

$$\Delta_{\mathbf{Y}^1} \le (1-p)^{-1} \min(1, pL(\mathbf{Y}^1)). \tag{6.19}$$

Next, by definition,

$$\begin{aligned} L(\mathbf{Y}^1) &= \exp(-ma^2/2 + \sqrt{m}aSY^1) = d^{-x^2/2 + xSY^1/\sqrt{\log d}}, \\ SY^1 &\stackrel{\Delta}{=} m^{-1/2} \sum_{i=1}^{m} Y_i^1. \end{aligned}$$

Take a threshold $H_* = t\sqrt{\log d}$ such that $pL_1(H_*) = 1$, i.e.,

$$-\beta - x^2/2 + xt = 0, \quad t = x/2 + \beta/x.$$

Then $pL(\mathbf{Y}^1) < 1$ (respectively, $pL(\mathbf{Y}^1) > 1$) is equivalent to $SY^1 < H_*$ (respectively, $SY^1 > H_*$).

Since Z^1, Y_i^1 are independent and, by the condition $\limsup_{d\to+\infty} x^* < \phi(\beta^*)$, the values a_d are bounded uniformly in d we have $E_{H_{0,1}}(\ell_{a_d}(Z^1) - 1)^2 = e^{a_d^2} - 1 \le c_0 a_d^2$ where c_0 is a constant. Therefore, using (6.19) we find

$$\begin{aligned} E_{H_{0,1}}\Delta_1^2 &= E_{H_{0,1}}(\ell_{a_d}(Z^1) - 1)^2 E_P \Delta_{\mathbf{Y}^1}^2 \le c_0 a_d^2 E_P \Delta_{\mathbf{Y}^1}^2 \\ &\le \frac{c_0 a^2}{1-p} \left(p^2 E_P \left(L^2(\mathbf{Y}^1) \mathbb{1}_{\{pL(\mathbf{Y}^1)\le 1\}} \right) + P(pL(\mathbf{Y}^1) > 1) \right), \end{aligned}$$

where $P = (1-p)P_0 + pP_a$, $a = a_d$ for brevity, P_a is the Gaussian measure with the density $f_a(\cdot)$, cf. (6.5). Note that $SY^1 \sim \mathcal{N}(0,1)$ under P_0 and $SY^1 \sim \mathcal{N}(0, \sqrt{m}a)$ under P_a. Therefore

$$\begin{aligned} E\left(L^2(\mathbf{Y}^1) \mathbb{1}_{\{pL(\mathbf{Y}^1)\le 1\}} \right) &= E_{P_0}\left(L^2(\mathbf{Y}^1) \mathbb{1}_{\{pL(\mathbf{Y}^1)\le 1\}} \right) + pE_{P_a}\left(L^2(\mathbf{Y}^1) \mathbb{1}_{\{pL(\mathbf{Y}^1)\le 1\}} \right), \\ P(pL(\mathbf{Y}^1) > 1) &= (1-p)P_0(SY^1 > H_*) + pP_a(SY^1 > H_*) \\ &= (1-p)c + \lambda(1-p), \end{aligned}$$

where

$$c = \Phi(-H_*) = Ad^{-t^2/2}, \quad \lambda = p\Phi(\sqrt{m}a - H_*) = Ad^{-\beta - (t-x)_+^2/2},$$

and A is a logarithmic factor: $b(\log d)^{-1/2} \le A \le B(\log d)^{1/2}$ for some positive constants b, B. It is easy to see that $c \le \lambda$ and $c = A\lambda$ as $\sqrt{m}a \le H_*$.

Since $L(\mathbf{Y}^1) = \frac{dP_a}{dP_0}(\mathbf{Y}^1)$ we get

$$E_{P_0}\left(L^2(\mathbf{Y}^1) \mathbb{1}_{\{pL(\mathbf{Y}^1)\le 1\}} \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{H_*} \exp(-ma^2 + 2xz)dz = e^{ma^2}\Phi(H_* - 2\sqrt{m}a),$$

and

$$pE_{P_a}\left(L^2(\mathbf{Y}^1) \mathbb{1}_{\{pL(\mathbf{Y}^1)\le 1\}} \right) = pE_{P_0}\left(L^3(\mathbf{Y}^1) \mathbb{1}_{\{pL(\mathbf{Y}^1)\le 1\}} \right) \le E_{P_0}\left(L^2(\mathbf{Y}^1) \mathbb{1}_{\{pL(\mathbf{Y}^1)\le 1\}} \right).$$

Therefore

$$E_{H_{0,1}}\Delta_1^2 \le \frac{2a^2(1+o(1))}{1-p}(u+\lambda),$$

where

$$u = p^2 e^{ma^2} \Phi(H_* - 2\sqrt{m}a) = Ad^{-2\beta + x^2 - (2x-t)_+^2/2}.$$

It is easily seen that $u = O(\lambda)$ for $H_* \le \sqrt{m}a$, i.e., for $t \le x$, which is equivalent to $x^2 \ge 2\beta$. Also $\lambda = O(u)$ for $H_* \ge 2\sqrt{m}a$, i.e., for $t \ge 2x$, which is equivalent to $x^2 \le 2\beta/3$. If $\sqrt{m}a < H_* < 2\sqrt{m}a$, i.e., if $x < t < 2x$, then $u = A\lambda = Ac$; cf. [9], pp. 295-296. The conditions $x < t < 2x$ are equivalent to $2\beta/3 \le x^2 \le 2\beta$. Therefore we get

$$dE_{H_{0,1}} \Delta_1^2 = Ad^{\nu_d}, \quad \nu_d = -\gamma + 1 + \begin{cases} -\beta & x^2 \ge 2\beta, \\ -t^2/2 & 2\beta/3 \le x^2 \le 2\beta, \\ -2\beta + x^2 & 0 < x^2 < 2\beta/3. \end{cases} \tag{6.20}$$

Thus the relation (6.18) holds true as

$$\liminf_{d \to +\infty} \nu_d < 0. \tag{6.21}$$

Set

$$\beta^* = \beta/(1-\gamma), \quad x^* = x/\sqrt{1-\gamma}, \quad t^* = x^*/2 + \beta^*/x^*. \tag{6.22}$$

Then the condition (6.21) is equivalent to $\liminf_{d \to +\infty} \nu_d^* < 0$ where

$$\nu_d^* = \nu_d/(1-\gamma) = 1 + \begin{cases} -\beta^* & \text{as } (x^*)^2 \ge 2\beta^*, \\ -(t^*)^2/2 & \text{as } 2\beta/3 \le (x^*)^2 \le 2\beta^*, \\ -2\beta^* + (x^*)^2 & \text{as } 0 < (x^*)^2 < 2\beta^*/3. \end{cases} \tag{6.23}$$

The relations (6.23) imply that successful classification is impossible as $\limsup_{d \to +\infty} x^* - \phi(\beta^*) < 0$ where $\phi(\beta^*)$ is defined by (1.8) for $\beta^* \in (1/2, 1)$.

7 Proof of the upper bounds

In this section we prove theorems 4.3 – 4.6. Without loss of generality, we shall assume throughout that $\sigma = 1$. We shall consider that s is fixed in theorems 4.3 and 4.4 and that x is fixed in theorems 4.5, 4.6. The general case can be treated in a similar way by passing to subsequences $s_d \to s > 0$, $x_d \to x > 0$. Sometimes we shall set for brevity (and without loss of generality) $c = 1$ or $C = 1$ where c and C are the constants in the definition of U_{β, a_d}.

7.1 Proof of theorem 4.3

Note first that, for any $\delta > 0$, uniformly in $u \in U_{\beta, a_d}$,

$$P^{(u)}(|M_0 - h(x)\sqrt{\log d}| > \delta) \to 0, \tag{7.1}$$

$$P_{H_s}^{(u)}(|M - h(x_s)\sqrt{\log d}| > \delta) \to 0, \quad s = 0, 1, \tag{7.2}$$

as $d \to +\infty$, where $P^{(u)}$ denotes the distribution of $\mathbf{Y}$, the notation x, x_0, x_1 is defined in (4.2) and $h(t) = \max(\sqrt{2}, t + \sqrt{2(1-\beta)})$. Indeed, setting $T(x) = h(x)\sqrt{\log d} \geq \sqrt{2\log d}$, for any $\delta > 0$ we obtain

$$\begin{aligned} P^{(u)}(M_0 > T(x) + \delta) &\leq \sum_{k=1}^{d} P^{(u)}(SY^k > T(x) + \delta) \leq d\Phi(-T(x) - \delta) \\ &\quad + \sum_{k=1}^{d} \varepsilon_k \Phi(a_d\sqrt{m} - T(x) - \delta) \\ &\leq o(1) + Cd^{1-\beta}\Phi(-\sqrt{2(1-\beta)\log d} - \delta) = o(1) \end{aligned}$$

as $d \to +\infty$. Next,

$$\begin{aligned} P^{(u)}(M_0 < T(x) - \delta) &= \prod_{k=1}^{d}(1 - P^{(u)}(SY^k \geq T(x) - \delta)) \\ &\leq \exp\left(-\sum_{k=1}^{d} P^{(u)}(SY^k \geq T(x) - \delta)\right), \end{aligned}$$

and

$$\sum_{k=1}^{d} P^{(u)}(SY^k \geq T(x) - \delta) \geq (d - Cd^{1-\beta})\Phi(-T(x) + \delta) + cd^{1-\beta}\Phi(a\sqrt{m} - T(x) + \delta).$$

If $h(x) = \sqrt{2}$, then $(d - Cd^{1-\beta})\Phi(-T(x) + \delta)$ tends to $+\infty$ as $d \to +\infty$. If $h(x) > \sqrt{2}$, then

$$cd^{1-\beta}\Phi(a\sqrt{m} - T(x) + \delta) = cd^{1-\beta}\Phi(-\sqrt{2(1-\beta)\log d} + \delta) \to +\infty$$

as $d \to +\infty$. This proves (7.1). The proof of (7.2) is analogous.

It follows from (7.1)-(7.2) that if $x_1 \leq \phi_2(\beta)$ (which is the same as $h(x_1) = \sqrt{2}$, implying $h(x) = h(x_0) = \sqrt{2}$), then $\Lambda_M < 1 + \delta$, for any $\delta > 0$ with both $P_{H_0}^{(u)}$ and $P_{H_1}^{(u)}$ probabilities tending to 1 as $d \to +\infty$. Next, let $x_1 > \phi_2(\beta)$. Then $h(x_1) > h(x) \geq h(x_0)$. This yields that $\Lambda_M < 1 + \delta$ for any $\delta > 0$ with $P_{H_0}^{(u)}$ probability tending to 1 as $d \to +\infty$. Therefore, (4.4) and (4.5) follow. We finally prove (4.6). Using (7.1) and (7.2) we get that, with $P_{H_1}^{(u)}$ probability tending to 1 as $d \to +\infty$,

$$\Lambda_M \geq \frac{h(x_1)\sqrt{\log d} - \delta}{h(x)\sqrt{\log d} + \delta} > \frac{h(x_1)}{h(x)}(1 - \delta) - \delta$$

for any $0 < \delta < 1$, where the last inequality is satisfied for any $d \geq 2$. Then (4.6) holds, since we can always choose a small c_0 in the definition of ψ^{max} and a small δ such that

$$\frac{h(x_1)}{h(x)}(1 - \delta) - \delta > 1 + c_0.$$

Finally, note that all the bounds on the probabilities above are independent of u and thus the convergence of the probabilities is uniform in $u \in U_{\beta,a_d}$ and in (a_d, β) such that $h(x_1)/h(x)$ is bounded away from 1. This completes the proof.

7.2 Proof of theorem 4.4

Fix $u \in U_{\beta,a_d}$. We first analyse the expectations and the variances of the statistics $L_0(t)$ and $L(t)$. Recall that $\Phi(z) \asymp e^{-z^2/2}/z, \quad z \to +\infty$, which implies

$$\Phi(-tT_d) = A_d d^{-(t_+)^2/2}, \quad t_+ = \max(t, 0). \tag{7.3}$$

Here A_d is a positive factor satisfying $A_d = O(1)$ and $A_d^{-1} = O(\sqrt{\log d})$ for $t = O(1)$ and $d \to +\infty$. In this proof and the proof of theorem 4.5 below we assume a weaker condition: A_d is a quantity depending on d (maybe different on different occasions) such that

$$|\log A_d| = o(\log d) \tag{7.4}$$

as $d \to +\infty$. Arguing under this weaker condition will allow us to get the proof of theorem 5.1 in parallel with that of theorem 4.5.

The expectations of $L_0(t)$ and $L(t)$ for any fixed $u \in U_{\beta,a_d}$ and $h \le t \le \sqrt{2}$ satisfy

$$\begin{aligned} E^{(u)}L_0(t) &= d^{1-\beta}(\Phi(a_d\sqrt{m} - tT_d) - \Phi(-tT_d)) \\ &= A_d d^{1-\beta}\left(d^{-((t-x)_+)^2/2} - d^{-t^2/2}\right), \\ E^{(u)}_{H_s}L(t) &= d^{1-\beta}(\Phi((m+s)a_d/\sqrt{m+1} - tT_d) - \Phi(-tT_d)) \\ &= A_d d^{1-\beta}\left(d^{-((t-x_s)_+)^2/2} - d^{-t^2/2}\right), \end{aligned}$$

where $s = 0, 1$ (recall that $h \le t_l \le \sqrt{2}$ for all t_l in the considered grid). Note that if $x > b$ for some constant $b > 0$, then in view of our assumptions on h we have $d^{-t^2/2} \le d^{-(t-x)_+^2/2}/2$ for $t \ge h$ and all d large enough. Therefore, for $x, x_s > b$ and all d large enough,

$$\begin{aligned} E^{(u)}\Delta_0(t) &= A_d d^{1/2-\beta} d^{t^2/4-(t-x)_+^2/2}, \\ E^{(u)}_{H_s}\Delta(t) &= A_d d^{1/2-\beta} d^{t^2/4-(t-x_s)_+^2/2}, \end{aligned}$$

where $s = 0, 1$. Since the maximum of $t^2/4 - ((t-x)_+)^2/2$ in $0 \le t \le \sqrt{2}$ is attained either at $t = 2x$ when $0 < x \le 1/\sqrt{2}$, or at $t = \sqrt{2}$ when $x > 1/\sqrt{2}$, we have, for $x > b$ and all d large enough,

$$\max_{h \le t \le \sqrt{2}} E^{(u)}\Delta_0(t) = \begin{cases} A_d d^{1/2-\beta+x^2/2}, & x \le 1/\sqrt{2}, \\ A_d d^{1-\beta-((\sqrt{2}-x)_+)^2/2}, & x > 1/\sqrt{2}. \end{cases} \tag{7.5}$$

Analogously, we have for $s = 0, 1$, $x_s > b$ and all d large enough:

$$\max_{h \le t \le \sqrt{2}} E^{(u)}_{H_s}\Delta(t) = \begin{cases} A_d d^{1/2-\beta+x_s^2/2}, & x_s \le 1/\sqrt{2}, \\ A_d d^{1-\beta-((\sqrt{2}-x_s)_+)^2/2}, & x_s > 1/\sqrt{2}. \end{cases} \tag{7.6}$$

We shall need the exact asymptotics (7.5) and (7.6) only when $x > b$ and $x_s > b$ for some constants $b > 0$. For small x and x_s it will be enough for our purposes to

use the fact that the right-hand sides of (7.5) and (7.6) constitute upper bounds for the corresponding left-hand sides for all $x, x_s > 0$.

We now consider bounds for the corresponding variances:

$$\begin{aligned}
\mathrm{Var}^{(u)}(L_0(t)) &\leq d\Phi(tT_d)\Phi(-tT_d) + d^{1-\beta}\Phi(tT_d - a_d\sqrt{m})\Phi(-tT_d + a_d\sqrt{m}) \\
&\leq A_d(d^{1-t^2/2} + d^{1-\beta-(t-x)^2/2}), \\
\mathrm{Var}^{(u)}_{H_s}(L(t)) &\leq d\Phi(tT_d)\Phi(-tT_d) \\
&\quad + d^{1-\beta}\Phi(tT_d - (m+s)a_d/\sqrt{m+1})\Phi(-tT_d + (m+s)a_d/\sqrt{m+1}) \\
&\leq A_d(d^{1-t^2/2} + d^{1-\beta-(t-x_s)^2/2}),
\end{aligned}$$

where $s = 0, 1$. Since for $x > 0$ the maximum of $t^2 - (t-x)^2$ in $0 \leq t \leq \sqrt{2}$ is attained at $t = \sqrt{2}$,

$$\begin{aligned}
\mathrm{Var}^{(u)}(\Delta_0(t)) &\leq A_d(1 + d^{-\beta+(t^2-(t-x)^2)/2}) \leq A_d(1 + d^{1-\beta-(\sqrt{2}-x)^2/2}) \\
&\leq \begin{cases} A_d, & x \leq \phi_2(\beta), \\ A_d d^{1-\beta-(\sqrt{2}-x)^2/2}, & x > \phi_2(\beta), \end{cases} \\
\mathrm{Var}^{(u)}_{H_s}(\Delta(t)) &\leq A_d(1 + d^{-\beta+(t^2-(t-x_s)^2)/2}) \leq A_d(1 + d^{1-\beta-(\sqrt{2}-x_s)^2/2}) \\
&\leq \begin{cases} A_d, & x_s \leq \phi_2(\beta), \\ A_d d^{1-\beta-(\sqrt{2}-x_s)^2/2}, & x_s > \phi_2(\beta), \end{cases}
\end{aligned}$$

where $s = 0, 1$. Take $N_0 > 0$ such that $N_0^2 \asymp T_d \gg N$. By Chebyshev's inequality, for each $l = 1, ..., N$, each $u \in U_{\beta,a_d}$ and $s = 0, 1$, with $P^{(u)}_{H_s}$-probability greater than $1 - 1/N_0^2$ we have

$$E^{(u)}_{H_s}\Delta_0(t_l) - N_0 \max_{0\leq t\leq\sqrt{2}} \sqrt{\mathrm{Var}^{(u)}_{H_s}\Delta_0(t)} \leq \Delta_0(t_l) \leq E^{(u)}_{H_s}\Delta_0(t_l) + N_0 \max_{0\leq t\leq\sqrt{2}} \sqrt{\mathrm{Var}^{(u)}_{H_s}\Delta_0(t)}$$

and these inequalities also valid for $\Delta(\cdot)$ instead of $\Delta_0(\cdot)$. All these inequalities (with $\Delta(\cdot)$ and $\Delta_0(\cdot)$) simultaneously hold with probability greater than $1 - 2N_0^{-2}N \to 1$ (uniformly in $u \in U_{\beta,a_d}$). On this event of high probability we can evaluate Δ_0 and Δ by taking the maxima of the expectations and comparing them with the maxima of the square root of the variances. Proceeding in this way and using the bounds obtained above we find:

$$\Delta_0 = \begin{cases} O(A_d), & x \leq \phi_1(\beta),\ \beta \leq 3/4 \text{ or } x < \phi_2(\beta),\ \beta > 3/4, \\ d^{1/2-\beta+x^2/2+O(h)}, & 1/\sqrt{2} \geq x > \phi_1(\beta),\ \beta \leq 3/4, \\ d^{1-\beta-((\sqrt{2}-x)_+)^2/2+O(h)}, & x > 1/\sqrt{2},\ \beta \leq 3/4 \text{ or } x \geq \phi_2(\beta),\ \beta > 3/4 \end{cases}$$

with $P^{(u)}$-probability tending to 1 as $d \to +\infty$, and, for $s = 0, 1$:

$$\Delta = \begin{cases} O(A_d), & x_s \leq \phi_1(\beta),\ \beta \leq 3/4 \text{ or } x_s < \phi_2(\beta),\ \beta > 3/4, \\ d^{1/2-\beta+x_s^2/2+O(h)}, & 1/\sqrt{2} \geq x_s > \phi_1(\beta),\ \beta \leq 3/4, \\ d^{1-\beta-((\sqrt{2}-x_s)_+)^2/2+O(h)}, & x_s > 1/\sqrt{2},\ \beta \leq 3/4 \text{ or } x_s \geq \phi_2(\beta),\ \beta > 3/4 \end{cases}$$

with $P_{H_s}^{(u)}$-probability tending to 1 as $d \to +\infty$ (the convergence of all the probabilities is uniform in $u \in U_{\beta,a_d}$). Using these relations we get the following results. First, $\Lambda^* = o(H)$ with $P_{H_0}^{(u)}$-probability tending to 1. Next, $\Lambda^* = o(H)$ with $P_{H_s}^{(u)}$-probability tending to 1 (for $s = 0, 1$) if either $x_1 \le \phi_1(\beta),\ \beta \le 3/4$ or $x_1 < \phi_2(\beta),\ \beta > 3/4$. Furthermore, if $x_0 < \sqrt{2} - \tau$ for some small $\tau > 0$, and either $x_1 > \phi_1(\beta) + \tau,\ \beta \le 3/4$ or $x_1 > \phi_2(\beta) + \tau,\ \beta > 3/4$, then with $P_{H_1}^{(u)}$-probability tending to 1 we have $\Lambda^* \ge d^{c\tau} \gg H$ for some $c > 0$. Clearly, the convergence of all the probabilities here is uniform in $u \in U_{\beta,a_d}$. Thus, the theorem follows.

7.3 Proof of theorem 4.5

Fix $u \in U_{\beta,a_d}$. Let $m = m_d \to +\infty$ such that $\log m = o(\log d)$. Observe that a_d cannot be "too large" in view of (2.2). Also a_d cannot be "too small" since $x \stackrel{\Delta}{=} a_d\sqrt{m/\log d} > \phi(\beta) \ge b$ for some $b > 0$. In particular, $a_d d^\delta \to +\infty$, for any $\delta > 0$, so that a_d satisfies a condition similar to (7.4):

$$|\log a_d| = o(\log d). \tag{7.7}$$

We first analyse the statistic $\Delta(t)$. Clearly, $E_{H_0}^{(u)}\Delta(t) = 0$, since $E_{H_0}^{(u)} Z^k = 0$ and Z^k and SY^k are independent. We also have $E_{H_0}^{(u)}(Z^k)^2 = 1$. Recalling (7.3), we obtain

$$\begin{aligned}
\mathrm{Var}_{H_0}^{(u)}\Delta(tT_d) &= \frac{1}{d\Phi(-tT_d)}\left(\sum_{k=1:\varepsilon_k=0}^{d}\Phi(-tT_d) + \sum_{k=1:\varepsilon_k=1}^{d}\Phi(-(t-x)T_d)\right)\\
&= 1 + A_d d^{-\beta+t^2/2-((t-x)_+)^2/2},\\
D_0^2(x,\beta) &\stackrel{\Delta}{=} \max_{0<t\le\sqrt{2}}\mathrm{Var}_{H_0}^{(u)}\Delta(t) = 1 + A_d d^{1-\beta-((\sqrt{2}-x)_+)^2/2}\\
&= 1 + A_d\begin{cases} O(1), & x \le \phi_2(\beta),\\ d^{1-\beta-((\sqrt{2}-x)_+)^2/2}, & x \ge \phi_2(\beta).\end{cases}
\end{aligned}$$

Here and below A_d is a factor satisfying (7.4). Next,

$$E_{H_1}^{(u)}\Delta(t) = \frac{a_d}{\sqrt{d\Phi(-tT_d)}}\sum_{k=1:\varepsilon_k=1}^{d}\Phi(-(t-x)T_d) = a_d A_d d^{1/2-\beta+t^2/4-((t-x)_+)^2/2},$$

which yields

$$\begin{aligned}
E(x,\beta) &\stackrel{\Delta}{=} \max_{0\le t\le\sqrt{2}} E_{H_1}^{(u)}\Delta(t) = a_d A_d\begin{cases} d^{1/2-\beta+x^2/2}, & x \le 1/\sqrt{2},\\ d^{1-\beta-((\sqrt{2}-x)_+)^2/2}, & x \ge 1/\sqrt{2}\end{cases}\\
&= a_d A_d\begin{cases} O(1), & x \le \phi_1(\beta),\ \beta \le 3/4 \text{ or } x \le \phi_2(\beta),\ \beta \ge 3/4,\\ d^{1/2-\beta+x^2/2}, & 1/\sqrt{2} \ge x > \phi_1(\beta),\ \beta \le 3/4,\\ d^{1-\beta-((\sqrt{2}-x)_+)^2/2}, & x > \phi_2(\beta),\ \beta \ge 3/4 \text{ or } x \ge 1/\sqrt{2},\ \beta \le 3/4.\end{cases}
\end{aligned}$$

Analogously,

$$\begin{aligned}
\mathrm{Var}^{(u)}_{H_1}\Delta(t) &= \frac{1}{d\Phi(-tT_d)}\Big(\sum_{k=1:\varepsilon_k=0}^{d}\Phi(-tT_d)+\sum_{k=1:\varepsilon_k=1}^{d}\big((\Phi(-(t-x)T_d)\\
&\quad +a_d^2\Phi(-(t-x)T_d)\Phi((t-x)T_d)\big)\Big)\\
&= 1+A_d(1+a_d^2)d^{-\beta+t^2/2-((t-x)_+)^2/2},\\
D_1^2(x,\beta) &\stackrel{\Delta}{=} \max_{0<t\le\sqrt{2}}\mathrm{Var}^{(u)}_{H_1}\Delta(t) = 1+A_d(1+a_d^2)d^{1-\beta-((\sqrt{2}-x)_+)^2/2}\\
&= 1+A_d(1+a_d^2)\begin{cases}O(1), & x\le\phi_2(\beta),\\ d^{1-\beta-((\sqrt{2}-x)_+)^2/2}, & x\ge\phi_2(\beta).\end{cases}
\end{aligned}$$

Suppose that, for some small $\tau>0$,

$$\beta\in[1/2+\tau,1-\tau],\quad x\ge\phi(\beta)+\tau. \tag{7.8}$$

These relations and the inequality $1/2-\beta+x^2/2\ge 1-\beta-(\sqrt{2}-x)^2/2$ imply that under (7.8) and (7.7) we have, for some $\tau_1>0,\ \tau_2>0$ depending on τ in (7.8),

$$D_s(x,\beta)\le d^{-\tau_1}E(x,\beta),\quad s=0,1\quad\text{and}\quad E(x,\beta)\ge d^{\tau_2}.$$

Arguing as in the proof of theorem 4.4 above we obtain the following facts. First,

$$|\Delta|\le A_d D_0(x,\beta) \tag{7.9}$$

with $P^{(u)}_{H_0}$-probability tending to 1 as $d\to+\infty$. Second, if $x\le\phi(\beta)$, then

$$|\Delta|\le A_d D_1(x,\beta) \tag{7.10}$$

with $P^{(u)}_{H_1}$-probability tending to 1 as $d\to+\infty$. Finally, if (7.8) holds, then

$$\Delta\ge A_d E(x,\beta) \tag{7.11}$$

with $P^{(u)}_{H_1}$-probability tending to 1 as $d\to+\infty$. Thus, with $P^{(u)}_{H_0}$-probability tending to 1, the ratio

$$\tilde\Lambda(x,\beta)=\frac{\Delta}{\sqrt{H+D_0^2(x,\beta)}}$$

is small. The same holds with $P^{(u)}_{H_1}$-probability tending to 1 if $x<\phi(\beta)$. To finish the proof, we show that these properties hold also for Λ^*_∞ which differs from $\tilde\Lambda(x,\beta)$ only in that we replace $D_0^2(x,\beta)$ by Δ_* (note that $\tilde\Lambda(x,\beta)$ is not a statistic, since $D_0(x,\beta)$ depends on the unknown parameters x,β). The distribution of Δ_* is the same under $P^{(u)}_{H_0}$ and $P^{(u)}_{H_1}$, and depends only on the parameter u. We have

$$\begin{aligned}
E^{(u)}(\Delta_*) &= \sum_{k=1:\varepsilon_k=0}^{d}\Phi(-\sqrt{2}T_d)+\sum_{k=1:\varepsilon_k=1}^{d}\Phi(-(\sqrt{2}-x)T_d)\\
&= o(1)+A_d d^{1-\beta-(\sqrt{2}-x)_+^2/2},\\
\mathrm{Var}^{(u)}(\Delta_*) &\le E^{(u)}(\Delta_*).
\end{aligned}$$

These inequalities yield that $H+\Delta_* = H + A_d D_0^2(x,\beta)$ with probability tending to 1, and the statistic Λ^*_∞ has the properties that we have proved for $\tilde{\Lambda}(x,\beta)$. Finally, note that the convergence of all the probabilities in the above argument is uniform in $u \in U_{\beta,a_d}$. Thus, the theorem follows.

7.4 Proof of theorem 4.6

For the statistics $L^1(t)$ we have

$$E^{(u)}_{H_0} L^1(t) = 0, \quad E(t) \stackrel{\Delta}{=} E^{(u)}_{H_1} L^1(t) = a_d d^{1-\beta} \Phi(a_d\sqrt{m} - tT_d) = A a_d d^{1-\beta-(t-x)^2_+/2};$$

$$\mathrm{Var}^{(u)}_{H_0} L^1(t) = d(1-p)\Phi(-tT_d) + dp\Phi(a_d\sqrt{m} - tT_d),$$
$$\mathrm{Var}^{(u)}_{H_1} L^1(t) = d(1-p)\Phi(-tT_d) + dp(1+a_d^2)\Phi(a_d\sqrt{m} - tT_d),$$

which yields

$$\mathrm{Var}^{(u)}_{H_0} L^1(t) \le 2R(t), \quad \mathrm{Var}^{(u)}_{H_1} L^1(t) \le 3R(t)$$

with

$$R(t) = \max\left(d\Phi(-tT_d),\ dp\Phi(a_d\sqrt{m} - tT_d)\right).$$

Thus, for all $l = 1, \ldots, N$, with $P^{(u)}_{H_0}$-probability tending to 1 the statistics $L^1(t_l)$ belong to the intervals $[-N\sqrt{2R(t_l)},\ +N\sqrt{2R(t_l)}]$ and with $P^{(u)}_{H_1}$-probability tending to 1 they belong to the intervals $[E(t_l) - N\sqrt{3R(t_l)},\ E(t_l) + N\sqrt{3R(t_l)}]$.

Consider the ratios $\Delta(t_l),\ l = 1, ..., N$. First, let $R(t_l) \le 4N^2$. Then for all $l = 1, ..., N$, with $P^{(u)}_{H_s}$-probability tending to 1 ($s = 0, 1$), we have the inequalities

$$\begin{aligned} N^2 \le N^2 + L^0(t_l) \le N^2 + 2R(t_l) + N\sqrt{R(t_l)} &\le 11N^2, \\ L^1(t_l) \le N\sqrt{2R(t_l)} &< 3N^2 \quad \text{for } s = 0, \\ L^1(t_l) \ge E(t_l) - N\sqrt{2R(t_l)} &\ge E(t_l) - 3N^2 \quad \text{for } s = 1. \end{aligned}$$

Therefore, we get for all $l = 1, ..., N$ such that $R(t_l) \le 4N^2$, with $P^{(u)}_{H_s}$-probability tending to 1,

$$\begin{aligned} \Delta(t_l) &< 4N \quad \text{for } s = 0, \\ \Delta(t_l) &\ge E(t_l)/4N - N \ge E(t_l)/(2\sqrt{R(t_l)}\,) - N \quad \text{for } s = 1. \end{aligned}$$

Next, let $R(t_l) > 4N^2$. Then analogously, with $P^{(u)}_{H_s}$-probability tending to 1,

$$\begin{aligned} N^2 + L^0(t_l) \le N^2 + 2R(t_l) + N\sqrt{R(t_l)} &< 3R(t_l), \\ N^2 + L^0(t_l) \ge R(t_l) - N\sqrt{R(t_l)} &> R(t_l)/2, \\ L^1(t_l) \le N\sqrt{2R(t_l)} \quad &\text{for } s = 0, \\ L^1(t_l) \ge E(t_l) - N\sqrt{2R(t_l)} \quad &\text{for } s = 1. \end{aligned}$$

Hence, we get for all $l = 1, ..., N$ such that $R(t_l) > 4N^2$, with $P^{(u)}_{H_s}$-probability tending to 1,

$$\begin{aligned} \Delta(t_l) &< 4N \quad \text{for } s = 0, \\ \Delta(t_l) &\geq E(t_l)/(2\sqrt{R(t_l)}) - N \quad \text{for } s = 1. \end{aligned}$$

Thus uniformly over $u \in U_{\beta,a_d}$,

$$E^{(u)}_{H_0}\psi_\infty = P^{(u)}_{H_0}(\Delta > 4N) \to 0.$$

Recalling (6.17), (6.22) let us show that under the condition

$$x^* > \phi(\beta^*) \tag{7.12}$$

we have, for some $\eta > 0$,

$$\max_{1\leq l\leq N} E(t_l)/\sqrt{R(t_l)} > d^\eta. \tag{7.13}$$

This implies that uniformly over $u \in U_{\beta,a_d}$,

$$E^{(u)}_{H_1}(1 - \psi^*_m) = P^{(u)}_{H_1}(\Delta \leq 4N) \to 0.$$

In order to verify (7.13), let us study the ratio $E(t)/\sqrt{R(t)}$. We have, with a logarithmic factor A,

$$\frac{E(t)}{\sqrt{R(t)}} = Ad^{s(t)}, \quad s(t) = -\gamma/2 + 1/2 - \beta + \frac{1}{2}\min(t^2/2 - (t-x)^2_+,\ \beta - (t-x)^2_+/2).$$

Set $t_0 = x/2 + \beta/x$. Let us check that

$$s^* \triangleq \max_{0\leq t\leq\sqrt{2}} s(t) \geq -\frac{\gamma}{2} + \frac{1}{2} + \begin{cases} -\beta/2, & \text{if } \ x \geq t_0, \\ -t_0^2/4, & \text{if } \ x \leq t_0 \leq 2x, \\ -\beta + x^2/2, & \text{if } \ 2x \leq t_0. \end{cases} \tag{7.14}$$

Indeed, the relation $x \geq t_0$ is equivalent to $x^2 \geq 2\beta$. So, $s^* \geq s(\sqrt{2\beta}) = -\gamma/2 + (1-\beta)/2$, which implies the first relation (7.14).

The relation $2x \leq t_0$ is equivalent to $x^2 \leq 2\beta/3$ and if $2x \leq \sqrt{2}$, then $s^* \geq s(2x) = -\gamma/2 + 1/2 - \beta + x^2/2$, which implies the third relation (7.14). Let us show that the case $2x > \sqrt{2}$, $x^2 \leq 2\beta/3$ is impossible under (7.12). In fact, we have

$$\sqrt{1-\gamma}\,\phi(\beta^*) \geq \phi(\beta), \quad 0 \leq \beta \leq 1 - \gamma. \tag{7.15}$$

Combining (7.15) and (7.12) we find $x > \phi(\beta)$. It is easy to see that $x > \phi(\beta)$ and $x^2 \leq 2\beta/3$ only if $\beta \leq 3/4$. This implies $2x \leq \sqrt{2}$.

The relation $x \leq t_0 \leq 2x$ is equivalent to $2\beta/3 \leq x^2 \leq 2\beta$. If $t_0 \leq \sqrt{2}$, then $s^* \geq s(\sqrt{2}) = -\gamma/2 + 1/2 - t_0^2/4$, which implies the second relation (7.14). Let us show that the case $t_0 > \sqrt{2}$, $2\beta/3 \leq x^2$ is impossible if $x^* > \phi(\beta^*)$. In

fact, it is easy to check that these inequalities are simultaneously satisfied only if $x \le \phi_2(\beta),\ \beta \ge 3/4$. However, (7.12) and (7.15) imply

$$x > \phi(\beta) = \phi_2(\beta), \quad \text{for } \beta \ge 3/4,$$

a contradiction. By comparing (7.14) with (6.23) and repeating the argument from the end of Subsection 6.2 we see that (7.12) implies $\liminf s^* > 0$. Since $s(\cdot)$ is a Lipschitz function, we can replace the maximum over the interval $[0, \sqrt{2}]$ by the maximum over our grid with step δ, inducing the error of order $O(\delta)$. This yields (7.13). $\square$

References

[1] Donoho, D. and Jin, J. (2004) Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32**, 962–994.

[2] Donoho, D. and Jin, J. (2008) Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Nat. Acad. Sci.* **105**, 14790–14795.

[3] Donoho, D. and Jin, J. (2008) Feature selection by higher criticism thresholding: Optimal phase diagram. *Manucript*, available at arXiv:0812.2263.

[4] Hall, P., Pittelkow, Y. and Ghosh, M. (2008) Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes.*J. R. Stat. Soc.* B,**70**, Part 1, 159–173.

[5] Ibragimov, I. A. and Has'minski, R. Z. (1981) *Statistical Estimation. Asymptotic Theory.* Springer, New York.

[6] Ingster, Yu.I. 1997 Some problems of hypothesis testing leading to infinitely divisible distributions. *Math. Methods of Stat.* **6**, 47–69.

[7] Ingster, Yu. I. and Suslina, I. A. (2001) Adaptive detection of a signal of growing dimension. I. *Math. Methods of Stat.* **10**, 395–421.

[8] Ingster, Yu. I. and Suslina, I. A. (2001) Adaptive detection of a signal of growing dimension. II. *Math. Methods of Stat.* **11**, 37–68.

[9] Ingster, Yu. I. and Suslina, I. A. (2002) *Nonparametric Goodness-of-Fit Testing under Gaussian Model*. Springer Lectures Notes in Statistics. Vol. **169**, Springer, New York.

[10] Ingster, Yu. I. and Suslina, I .A. (2002) On a detection of a signal of known shape in multichannel system. *Zapiski Nauchn. Sem. POMI*, **294**, 88–112 (in Russian, translation in *J. Math. Sci.* **127** (2005), 1, 1723–1736).

[11] Jin, J. (2009) Impossibility of successful classification when useful features are rare and weak. Manuscript.

[12] Jager, L. and Wellner, J. A. (2007) Goodness-of-fit tests via phi-divergences. *Ann. Statist.* **35**, 2018–2053.

[13] Petrov, V.V. (1995) *Limit Theorems of Probability Theory*. Oxford University Press, Oxford.

[14] Pouet, C. (2008) *Quelques contributions à la théorie des tests*. Mémoire d'habilitation à diriger des recherches, Université Aix-Marseille 1.

Yu.I. Ingster: St.Petersburg State Electrotechnical University,
5, Prof. Popov str., 197376 St.Petersburg, Russia

Ch. Pouet: LATP, University of Provence
39, rue F. Joliot-Curie, 13453 Marseille cedex 13, France

A.B. Tsybakov: Laboratoire de Statistique, CREST, Timbre J340
3, av. Pierre Larousse, 92240 Malakoff cedex, France
and LPMA, University of Paris 6
4, Place Jussieu, 75252 Paris cedex 05, France