Bernoulli **26**(4), 2020, 2605–2638
https://doi.org/10.3150/20-BEJ1197

Fundamental limits of exact support recovery in high dimensions

ZHENG GAO[*] and STILIAN STOEV[†]

Department of Statistics, University of Michigan, 1085 S. University Ave., Ann Arbor, MI, 48105, USA. E-mail: [*]*gaozheng@umich.edu;* [†]*sstoev@umich.edu*

We study the support recovery problem for a high-dimensional signal observed with additive noise. With suitable parametrization of the signal sparsity and magnitude of its non-zero components, we characterize a phase-transition phenomenon akin to the signal detection problem studied by Ingster in 1998. Specifically, if the signal magnitude is above the so-called *strong classification boundary*, we show that several classes of well-known procedures achieve asymptotically perfect support recovery as the dimension goes to infinity. This is so, for a very broad class of error distributions with light, rapidly varying tails which may have arbitrary dependence. Conversely, if the signal is below the boundary, then for a very broad class of error dependence structures, no thresholding estimators (including ones with data-dependent thresholds) can achieve perfect support recovery. The proofs of these results exploit a certain *concentration of maxima* phenomenon known as relative stability. We provide a complete characterization of the relative stability phenomenon for Gaussian triangular arrays in terms of their correlation structure. The proof uses classic Sudakov–Fernique and Slepian lemma arguments along with a curious application of Ramsey's coloring theorem.

We note that our study of the strong classification boundary is in a finer, point-wise, rather than minimax, sense. We also establish results on the finite-sample Bayes optimality and sub-optimality of thresholding procedures. Consequently, we obtain a minimax-type characterization of the strong classification boundary for errors with log-concave densities.

Keywords: concentration of maxima; high–dimensional inference; Ramsey theory; rapid variation; relative stability; Sudakov–Fernique inequality; support recovery

1. Introduction

Consider the canonical signal-plus-noise model where the observation x is a high-dimensional vector in $\mathbb{R}^p$,

$$x = \mu + \varepsilon. \tag{1.1}$$

The signal, $\mu = (\mu(j))_{j=1}^p$, is a vector with s non-zero components supported on the set $S = \{j : \mu(j) \neq 0\}$; the second term ε is a random error vector. The goal of high-dimensional statistics is usually two-fold:

1. To detect the presence of non-zero components in μ. That is, to test the global hypothesis $\mu = 0$, which we call the *detection* problem, and
2. To estimate the support set S, which we call the *support recovery* problem.

An archetypal application where the two problems arise is in cyber security [27]. Internet service providers routinely collect statistics of network traffic to determine if there are abnormal surges or blackouts. While this monitoring is performed over a large number of servers, only very few servers are believed to be experiencing problems at any time. The signal detection problem is then equivalent to determining if there are any anomalies among all servers; the support recovery problem is equivalent to identifying the servers with anomalies.

The same two questions are pursued in, for example, large-scale microarray experiments [13], brain imaging and fMRI analysis [32], and numerous other applications.

A common theme in such applications is that the errors are correlated, and that the signal vectors are believed to be sparse: the number of non-zero components in μ is small compared to the number of test performed. Under such dependence and sparsity assumptions, it is natural to ask if and when one can reliably (1) detect the signals, and (2) recover the support set S. In this paper, we focus on the support recovery problem, and seek minimal conditions under which the support set can be consistently estimated, as dimensionality diverges.

1.1. Set-up

We now describe the models and assumptions adopted for the rest of the paper.

1.1.1. *Assumptions on the error distributions*

We assume that the errors come from a triangular array

$$\mathcal{E} = \left\{\left(\varepsilon_p(j)\right)_{j=1}^{p}, p = 1, 2, \ldots\right\}, \tag{1.2}$$

such that the $\varepsilon_p(j)$'s have common cumulative distribution function $F(x) = \mathbb{P}[\varepsilon_p(j) \le x]$. Note that the errors are only assumed to have common marginal distributions, and may have potentially arbitrary dependence.

We shall consider light-tailed error distributions with *rapidly varying* tails (see, e.g., Definition 2.1 below). To be concrete and better convey the main ideas, we will focus on the class of asymptotically generalized Gaussian laws (see Definition 1.1), which is still a fairly general class of models commonly used in the literature on high-dimensional testing [1,7]. Extensions to other models are deferred to the supplementary material [18].

Definition 1.1. A distribution F is called asymptotic generalized Gaussian with parameter $\nu > 0$ (denoted AGG(ν)) if

1. $F(x) \in (0, 1)$ for all $x \in \mathbb{R}$, and
2. $\log \overline{F}(x) \sim -\frac{1}{\nu}x^{\nu}$ and $\log F(-x) \sim -\frac{1}{\nu}(-x)^{\nu}$,

where $\overline{F}(x) = 1 - F(x)$ is the survival function, and $a(x) \sim b(x)$ is taken to mean $\lim_{x\to\infty} a(x)/b(x) = 1$.

The AGG models include, for example, the Gaussian distribution ($\nu = 2$), and the Laplace distribution ($\nu = 1$) as special cases. Since the requirement is only placed on the tail behavior, this class encompasses a large variety of light-tailed models.

Proposition 1.1. *The* $(1/p)$*th upper quantile of* $\mathrm{AGG}(\nu)$ *is*

$$u_p := F^{\leftarrow}(1-1/p) \sim (\nu \log p)^{1/\nu}, \quad \text{as } p \to \infty, \tag{1.3}$$

where $F^{\leftarrow}(q) = \inf_x\{x : F(x) \geq q\}$, $q \in (0,1)$.

The proof of Proposition 1.1 can be found in Section C of the supplement [18].

1.1.2. *Parametrization of the signals*

We assume in model (1.1) that $\mu = (\mu(j))_{j=1}^{p}$ is a sparse signal vector with non-zero entries only at the support of the signal $S_p \subseteq \{1,\dots,p\}$. We denote the size of the support set as $s = |S_p|$, and assume that the non-zero entries of μ are positive and take values in the interval $[\underline{\Delta}, \overline{\Delta}) \subset (0,\infty)$. That is, $0 < \underline{\Delta} \leq \mu(j) < \overline{\Delta} \leq +\infty$, for all $j \in S$.

Following [1,7,12,22,25], the signal sparsity – with a few exceptions which will be explicitly stated – is parametrized as

$$s = s(p) = \lfloor p^{1-\beta} \rfloor, \tag{1.4}$$

with $0 < \beta \leq 1$ fixed. We also parametrize the signal sizes $\underline{\Delta}$ and $\overline{\Delta}$ as

$$\underline{\Delta} = \underline{\Delta}(p) = (\nu \underline{r} \log p)^{1/\nu} \quad \text{and} \quad \overline{\Delta} = \overline{\Delta}(p) = (\nu \overline{r} \log p)^{1/\nu}, \tag{1.5}$$

with parameters $0 < \underline{r} \leq \overline{r} \leq +\infty$.

1.2. Thresholding procedures

We review next four procedures, that is, measurable set-valued functions of the data, commonly used for support estimation. All of them fall under the general class of thresholding procedures, studied in this paper.

Definition 1.2 (Thresholding procedures). A thresholding procedure for estimating the support $S_p := \{j : \mu(j) \neq 0\}$ is one that takes on the form

$$\widehat{S}_p = \{j : x(j) > t_p(x)\}. \tag{1.6}$$

We note here that the threshold $t_p(x)$ may depend on the data x.

A well-known (deterministic) thresholding procedure is Bonferroni's procedure.

Definition 1.3 (Bonferroni's procedure). Suppose the errors $\varepsilon(j)$'s have a common marginal distribution F, Bonferroni's procedure with family-wise error rate (FWER) at most α is the thresholding procedure (1.6) with

$$t_p = F^{\leftarrow}(1-\alpha/p). \tag{1.7}$$

A closely related procedure is Sidák's procedure [37] which is a more aggressive (and also deterministic) thresholding procedure that uses the threshold

$$t_p = F^{\leftarrow}\big((1-\alpha)^{1/p}\big). \tag{1.8}$$

Another procedure, which is strictly more powerful (in the context of hypothesis testing) than Bonferroni's, is Holm's procedure [24]. On observing the data x, its coordinates can be ordered from largest to smallest $x(j_1) \geq x(j_2) \geq \cdots \geq x(j_p)$, where $(j_1, \ldots, j_p)$ is a permutation of $\{1, \ldots, p\}$.

Definition 1.4 (Holm's procedure). Let k be the largest index such that

$$\overline{F}\big(x(j_i)\big) \leq \alpha/(p-i+1), \quad \text{for all } i \leq k.$$

Holm's procedure with FWER controlled at α is the thresholding procedure that uses

$$t_p(x) = x(j_k), \tag{1.9}$$

In contrast to the Bonferroni procedure, Holm's procedure is data-dependent. A closely related, more aggressive (data-dependent) thresholding procedure is Hochberg's procedure [23], which replaces the index k in Holm's with the largest index i such that

$$\overline{F}\big(x(j_i)\big) \leq \alpha/(p-i+1).$$

We will analyze the performance of these thresholding procedures in Section 2. The (sub)optimality of general data-dependent thresholding procedures will be established in Section 4. We now return to the discussion of support recovery.

1.3. Prior work

Recall our goal is to establish minimal conditions under which the support set can be consistently estimated, that is,

$$\mathbb{P}[\widehat{S}_p = S_p] \longrightarrow 1 \quad \text{as } p \to \infty, \tag{1.10}$$

where $\widehat{S}_p$ is an estimate of the true set of signal support S_p. This type of support estimation problems were pursued by Zhao and Yu [43] and Wasserman and Roeder [42] in the high-dimensional regression setting (where number of samples $n \ll p$), and by Meinshausen and Bühlmann [31] in graphical models, among many others. While there is a wealth of literature focusing on methods and conditions sufficient for the consistent estimation of the support set, the study on necessary conditions is relatively scarce, with notable efforts made by Wainwright [40,41] and Comminges et al. [8] in the regression context.

Interesting sharp results have been obtained, when we switch the metric from probability of support recovery, $\mathbb{P}[\widehat{S}_p = S_p]$, to the Hamming loss, $H(\widehat{S}, S)$, defined as the number of mismatches between the estimated and true support set. In particular, Ji and Jin [26] and Genovese et al. [19] studied the problem of support recovery in linear models under the Hamming loss. Under the sparsity parametrization in (1.4) and assuming equal signal sizes of $(2r\log p)^{1/2}$, a minimax-type phase-transition result was established. Specifically, it was shown that the Hamming loss diverges to $+\infty$ when r falls below the threshold

$$g_2(\beta) = \bigl(1 + (1-\beta)^{1/2}\bigr)^2, \tag{1.11}$$

for any method of support estimation. Conversely, under orthogonal, or near-orthogonal random designs, if $r > g_2(\beta)$, their proposed method achieves vanishing Hamming loss.

Very recently, Butucea et al. [6] studied both asymptotics and non-asymptotics of the support recovery problem in the additive noise model (1.1) with Gaussian errors under the Hamming loss. It was again shown that the boundary (1.11) exists in a minimax sense. That is, when errors are *independent*, the Hamming loss cannot be made to vanish if signal sizes fall below the boundary (1.11) by any procedure. Conversely, if $r > g_2(\beta)$, the Hamming loss can be made to vanish with a particular thresholding procedure.

One might expect these sharp results to carry over to the study of the probability of support recovery. As pointed out in, for example, [6], there is a natural lower bound for the probability of support recovery by the expected Hamming loss,

$$\mathbb{P}[\widehat{S} = S] \ge 1 - \mathbb{E}\bigl[H(\widehat{S}, S)\bigr] = 1 - \sum_{j=1}^{p} \mathbb{E}\bigl|\mathbb{1}_{\widehat{S}}(j) - \mathbb{1}_S(j)\bigr|. \tag{1.12}$$

Unfortunately, vanishing Hamming loss is only sufficient, not necessary for support recovery (1.10). A key observation in Relation (1.12) is that the expected Hamming loss decouples into p terms, and dependence of the estimates $\mathbb{1}_{\widehat{S}}(j)$ among the p locations no longer plays a role in the sum. Therefore, studying support recovery problems via the expected Hamming loss is not very informative especially under severe dependence, as the bound (1.12) may become very loose.

These considerations motivated us to study directly the exact support recovery in the sense of (1.10), under general distributional and dependence assumptions. The techniques developed in this paper are entirely different from those in Ji and Jin [26] or Butucea et al. [6]. We also establish transparent characterizations of the dependence conditions under which a phase transition type result holds.

1.4. Summary of contributions

Our contribution is three-fold, briefly summarized in this section.

1.4.1. *A point-wise phase transition phenomena under dependence*

Under the scaling described in (1.4) and (1.5), a certain combination of signal sparsity and signal size (β, r) either leads to exact support recovery, or complete failure, for a large class of dependence structures on the errors, and a large class of error marginal distributions. Consider the function

$$g(\beta) = g_\nu(\beta) = \bigl(1 + (1-\beta)^{1/\nu}\bigr)^\nu, \quad \nu > 0, \tag{1.13}$$

which we refer to as the *strong classification boundary*. In Theorem 2.1 we show that, if the signal size is above the boundary (i.e., $\underline{r} > g(\beta)$), the procedures described in Section 1.2 with appropriately calibrated levels achieve *exact support recovery*, that is,

$$\mathbb{P}[\widehat{S}_p = S_p] \longrightarrow 1, \quad \text{as } p \to \infty. \tag{1.14}$$

Conversely, we show in Theorem 2.2, that for a surprisingly large class of dependence structures characterized by the concept of *uniform relative stability* (URS, see Definition 2.3), when the signal size is below the boundary (i.e., $r < g(\beta)$), no thresholding procedure can achieve the asymptotically perfect support recovery (1.14). In fact,

$$\mathbb{P}[\widehat{S}_p = S_p] \longrightarrow 0, \quad \text{as } p \to \infty, \tag{1.15}$$

for all $\widehat{S}_p$ in the form of (1.6).

Complementing results in Butucea et al. [6], these two results show that the thresholding procedures obey a phase transition phenomenon in a strong, *point-wise* sense over the class of URS dependence structures, and over the class of AGG(ν), $\nu > 0$ error distributions.

The strong classification boundary g characterizes a phase-transition phenomenon similar to that of the signal detection and approximate support recovery. A preview of this result is presented in Figure 1, which shows the boundaries for signal detection, approximate support recov-

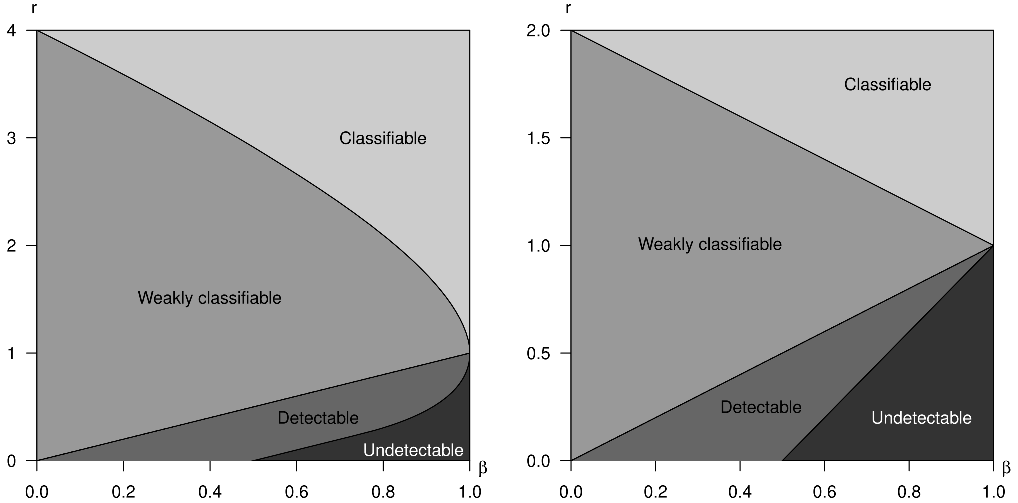

Figure 1. The phase diagrams of the detection, weak classification, and strong classification boundaries against sparse alternatives under Gaussian (left) and Laplace distributed (right) errors. Here β, $\underline{r}$, and $\overline{r}$ parametrize the signal sparsity and the lower and upper bounds of the signal sizes, respectively. We study in this paper the strong classification boundary, above which the support recovery can be achieved *exactly* in the *Classifiable* region $\{(\beta, \underline{r}) : \underline{r} > g(\beta)\}$. In a large class of dependence structures characterized by URS, when signal sizes fall below the strong classification boundary (1.13), that is, $\{(\beta, \overline{r}) : \overline{r} < g(\beta)\}$, no thresholding procedure succeeds in the exact support recovery problem. For the detection and weak classification boundaries, see, for example, [1,12,17,22,25], and Section F of the supplement.

ery, and exact support recovery, in the special case of Gaussian and Laplace errors, respectively. (For definitions of the detection and weak classification boundaries, see, for example, [1,12,18, 22,25].)

1.4.2. *Uniform relative stability and concentration of maxima*

The key probabilistic concept behind our characterization of the dependence structure under which (1.15) takes place is a certain *concentration of maxima* phenomenon, known as *relative stability* (see Section 3). We introduce and study an extension of this concept referred to as uniform relative stability (URS), which is shown to be extremely mild. Broadly speaking, the strong classification boundary phenomenon holds for dependent light-tailed errors, provided that they are uniformly relatively stable.

In the case of dependent Gaussian errors, we establish in Theorem 3.1 a complete characterization of URS in terms of a simple condition on the covariance structure, using the Sudakov–Fernique bounds and a curious application of Ramsey theory. This result may be of independent interest.

1.4.3. *Role of thresholding procedures and minimax optimality*

We show in Section 4 that data thresholding procedures are indeed optimal when the errors are independent and identically distributed (i.i.d.) with log-concave densities. In this case, no estimator can achieve perfect support recovery when the signal is below the strong classification boundary (1.13). Consequently, in the case of AGG(ν) errors with $\nu \geq 1$, the strong classification boundary is shown to hold in the minimax sense for *all* procedures. This is formalized in Theorem 4.1 and Corollary 4.2.

A final surprising result, that had only recently been taken notice of by the statistical community, is that thresholding procedures – including data-dependent ones – are not optimal in general in the support recovery problem when the errors have heavy (regularly-varying) tails. Arias-Castro et al. [2] noticed the phenomena in approximate support recovery problems. In this case, we also demonstrate the absence of a phase-transition phenomenon in exact support recovery by thresholding, in Theorem E.1 of the supplement [18].

The rest of this paper is organized as follows. The results summarized in Section 1.4 are detailed in Sections 2, 3, and 4. Main ideas of the proof of Theorem 3.1 are sketched in Section 5. Technical parts of the proof of Theorem 3.1, numerical illustrations, and extensions of the phase transition phenomenon to other light-tailed and heavy-tailed error models are discussed in the supplementary material [18].

2. Exact support recovery under dependence

We present the first result in Section 2.1, which states that exact support recovery (1.14) is achievable when signal sizes are above the boundary (1.13).

The behavior of maxima plays a key role in the study of support recovery problems. We show in Section 2.2 that the maxima of errors with rapidly varying tails can be bounded above using quantiles of their marginal distribution, regardless of their dependence structure. On the other hand, a lower bound for the error maxima can be provided for a very general class of dependence structures, as we will see in Section 2.3. This will prepare us to state the converse Theorem 2.2 in Section 2.4.

2.1. Sufficient conditions for exact support recovery

Following Butucea et al. [6], we define the parameter space for the signals μ as

$$\begin{aligned}\Theta_p^+(\beta, \underline{r}) = \big\{&\mu \in \mathbb{R}^p : \text{there exists a set } S_p \subseteq \{1, \dots, p\} \text{ such that } |S_p| \le \lfloor p^{1-\beta} \rfloor, \\ &\mu(j) \ge (\nu \underline{r} \log p)^{1/\nu} \text{ for all } j \in S_p, \text{ and } \mu(j) = 0 \text{ for all } j \notin S_p\big\}.\end{aligned} \tag{2.1}$$

Our first result states that, when $F \in \mathrm{AGG}(\nu)$ with $\nu > 0$, regardless of the error dependence structure, (asymptotic) perfect support recovery is achieved by applying Bonferroni's procedure with appropriately calibrated FWER, as long as the minimum signal size $\underline{r}$ is above the strong classification boundary (1.13).

Theorem 2.1. *Let the errors have common marginal distribution $F \in \mathrm{AGG}(\nu)$ with $\nu > 0$. Let $\widehat{S}_p$ be the Bonferroni's procedure (1.7) with vanishing FWER $\alpha = \alpha(p) \to 0$, such that $\alpha p^\delta \to \infty$ for every $\delta > 0$. If*

$$\underline{r} > g(\beta) = \big(1 + (1 - \beta)^{1/\nu}\big)^\nu, \tag{2.2}$$

then we have

$$\lim_{p\to\infty} \sup_{\mu \in \Theta_p^+(\beta, \underline{r})} \mathbb{P}[\widehat{S}_p \neq S_p] = 0. \tag{2.3}$$

Corollary 2.1 (Classes of procedures attaining the boundary). *Relation (2.3) holds for any FWER-controlling procedure that is strictly more powerful than Bonferroni's procedure. This includes Holm's procedure [24], and in the case of independent errors, Hochberg's procedure [23], and the Šidák procedure [37].*

Example 2.1. Under Gaussian errors, the particular choice of the thresholding at $t_p = \sqrt{2 \log p}$ in (1.7) corresponds to a Bonferroni's procedure with FWER decreasing at a rate of $(\log p)^{-1/2}$, and hence Theorem 2.1 applies. By Corollary 2.1, Holm's procedure – and when the errors are independent, the Šidák, and Hochberg procedures – with FWER controlled at $(\log p)^{-1/2}$ all achieve perfect support recovery provided that $\underline{r} > g(\beta)$.

The claims in Example 2.1 are verified in Section B of the supplement [18]. We now turn to the proof of Theorem 2.1.

Proof of Theorem 2.1. Throughout the proof, dependence on p will be suppressed to simplify notations when such omissions do not lead to ambiguity.

Under the AGG(ν) model, it is easy to see from equation (1.3) that the thresholds in Bonferroni's procedure are

$$t_p = F^{\leftarrow}(1-\alpha/p) = \big(\nu \log(p/\alpha)\big)^{1/\nu}\big(1+o(1)\big). \tag{2.4}$$

It is known that Bonferroni's procedure $\widehat{S}_p = \{j : x(j) > t_p\}$ controls the FWER. Indeed,

$$\begin{aligned}\mathbb{P}[\widehat{S} \subseteq S] &= 1 - \mathbb{P}\Big[\max_{j\in S^c} x(j) > t_p\Big] = 1 - \mathbb{P}\Big[\max_{j\in S^c} \varepsilon(j) > t_p\Big] \\ &\geq 1 - \sum_{j=1}^{p} \mathbb{P}\big[\varepsilon(j) > t_p\big] \geq 1 - \alpha(p) \to 1,\end{aligned} \tag{2.5}$$

where we used the union bound in the first inequality. Notice that the lower bound (2.5) is independent of the parameter μ (as well as the dependence structures), and hence holds uniformly over the parameter space, that is,

$$\lim_{p\to\infty} \inf_{\mu\in\Theta_p^+(\beta,\underline{r})} P[\widehat{S}_p \subseteq S_p] = 1. \tag{2.6}$$

On the other hand, for the probability of no missed detection, we have:

$$\mathbb{P}[\widehat{S} \supseteq S] = \mathbb{P}\Big[\min_{j\in S} x(j) > t_p\Big] = \mathbb{P}\Big[\min_{j\in S} x(j) - (\nu\underline{r}\log p)^{1/\nu} > t_p - (\nu\underline{r}\log p)^{1/\nu}\Big].$$

Since the signal sizes are no smaller than $(\nu\underline{r}\log p)^{1/\nu}$, we have

$$x(j) - (\nu\underline{r}\log p)^{1/\nu} \geq \varepsilon(j), \quad \text{for all } j \in S,$$

and hence we obtain

$$\mathbb{P}[\widehat{S} \supseteq S] \geq \mathbb{P}\Big[\min_{j\in S} \varepsilon(j) > \big(\nu\log(p/\alpha)\big)^{1/\nu}\big(1+o(1)\big) - (\nu\underline{r}\log p)^{1/\nu}\Big], \tag{2.7}$$

where we plugged in the expression for t_p in (2.4). Now, since the minimum signal size is bounded below by $\underline{r} > (1+(1-\beta)^{1/\nu})^{\nu}$, we have $\underline{r}^{1/\nu} - (1-\beta)^{1/\nu} > 1$, and so we can pick a $\delta > 0$ such that

$$\delta < \big(\underline{r}^{1/\nu} - (1-\beta)^{1/\nu}\big)^{\nu} - 1. \tag{2.8}$$

Since by assumption, for all $\delta > 0$, we have $p^{-\delta} = o(\alpha(p))$, there is an $M = M(\delta)$ such that $p/\alpha(p) < p^{1+\delta}$ for all $p \geq M$. Thus, from (2.7), we further conclude that for $p \geq M$ we have

$$\begin{aligned}\mathbb{P}[\widehat{S} \supseteq S] &\geq \mathbb{P}\Big[\min_{j\in S} \varepsilon(j) > \big((1+\delta)\nu\log p\big)^{1/\nu}\big(1+o(1)\big) - (\nu\underline{r}\log p)^{1/\nu}\Big] \\ &= \mathbb{P}\Big[\max_{j\in S}\big(-\varepsilon(j)\big) < \underbrace{\big(\underline{r}^{1/\nu} - (1+\delta)^{1/\nu}\big)(\nu\log p)^{1/\nu}\big(1+o(1)\big)}_{=:\text{A}}\Big] \\ &\geq 1 - \lfloor p^{1-\beta} \rfloor \times \overline{F}_{-}(\text{A}),\end{aligned} \tag{2.9}$$

where $\overline{F}_-(x) = \mathbb{P}[-\varepsilon(j) > x]$ is the survival function of the $(-\varepsilon(j))$'s. Notice that (2.9) follows from the union bound and the assumption that $|S_p| \le \lfloor p^{1-\beta} \rfloor$. Therefore, the lower bound does not depend on μ (nor on the error dependence structure), and holds uniformly in the parameter space. In turn, we obtain

$$\inf_{\mu\in\Theta_p^+(\beta,\underline{r})} \mathbb{P}[\widehat{S}_p \supseteq S_p] \ge 1 - \lfloor p^{1-\beta} \rfloor \times \overline{F}_-(\mathrm{A}). \tag{2.10}$$

If $\beta = 1$, we conclude that the right-hand side of (2.10) converges to 1, since $\mathrm{A} \to +\infty$. Let now $\beta \in (0, 1)$ and $u_p^- := F_-^{\leftarrow}(1 - 1/p)$. The fact that $p\overline{F}_-(u_p^-) \le 1$, implies

$$\lfloor p^{1-\beta} \rfloor \times \overline{F}_-(\mathrm{A}) \le \frac{\overline{F}_-(\mathrm{B} \times u^-_{\lfloor p^{1-\beta}\rfloor})}{\overline{F}_-(u^-_{\lfloor p^{1-\beta}\rfloor})}, \tag{2.11}$$

where $\mathrm{B} := \mathrm{A}/u^-_{\lfloor p^{1-\beta}\rfloor}$.

Notice that by assumption, the $-\varepsilon(j)$'s are also AGG(ν) distributed and by Proposition 1.1, $u_p^- := F_-^{\leftarrow}(1 - 1/p) \sim (\nu \log(p))^{1/\nu}$, as $p \to \infty$. Therefore, we have

$$u^-_{\lfloor p^{1-\beta}\rfloor} \sim \big(\nu(1-\beta)\log p\big)^{1/\nu} \tag{2.12}$$

and

$$\mathrm{B} = \frac{\mathrm{A}}{u^-_{\lfloor p^{1-\beta}\rfloor}} = \frac{\underline{r}^{1/\nu} - (1+\delta)^{1/\nu}}{(1-\beta)^{1/\nu}}\big(1 + o(1)\big) \to c > 1$$

as $p \to \infty$, by our choice of δ in (2.8).

Finally, since the distribution F_- has *rapidly varying* tails (by Definition 2.1 and Example 2.2 below), applying Proposition 2.1 (below in Section 2.2), we conclude that (2.11) vanishes. Consequently, the lower bound on the right-hand-side of (2.10) converges to 1. This, combined with (2.6), entails $\lim_{p\to\infty} \inf_{\mu\in\Theta_p^+(\beta,\underline{r})} \mathbb{P}[\widehat{S}_p = S_p] = 1$, and hence the desired conclusion (2.3), which completes the proof. □

The statements in Theorem 2.1 can be strengthened, to prepare us for a minimax result given in Section 4 below.

Remark 2.1. In the proof of Theorem 2.1, both (2.5) and (2.9) hold uniformly over all error dependence structures. Therefore, (2.6) and (2.10) may be strengthened to yield

$$\lim_{p\to\infty} \sup_{\substack{\mu\in\Theta_p^+(\beta,\underline{r})\\ \mathcal{E}\in D(F)}} P[\widehat{S}_p \ne S_p] = 0, \tag{2.13}$$

for $\underline{r} > g(\beta)$, where $D(F)$ is the collection of all arrays with common marginal F, that is,

$$D(F) = \big\{\mathcal{E} = \big(\varepsilon_p(j)\big)_p : \varepsilon_p(j) \sim F \text{ for all } j = 1, \dots, p, \text{and } p = 1, 2, \dots\big\}. \tag{2.14}$$

Remark 2.2. We emphasize that Theorem 2.1 holds for errors with *arbitrary* dependence structures. Intuitively, this is because the maxima of the errors grow at their fastest in the case of independence. Formally, the light-tailed nature of the error distribution allowed us to obtain sharp tail estimates via simple union bounds, valid under arbitrary dependence.

We turn next to the study of maxima and present the tools used in the proof of Theorem 2.1.

2.2. Rapid variation and relative stability

The behavior of the maxima of random variables has been studied extensively in the literature (see, e.g., [10,14,29,34] and the references therein). The concept of rapid variation plays an important role in the light-tailed case.

Definition 2.1 (Rapid variation). The survival function of a distribution, $\overline{F}(x) = 1 - F(x)$, is said to be rapidly varying if

$$\lim_{x\to\infty} \frac{\overline{F}(tx)}{\overline{F}(x)} = \begin{cases} 0, & t > 1, \\ 1, & t = 1, \\ \infty, & 0 < t < 1. \end{cases} \tag{2.15}$$

When $F(x) < 1$ for all finite x, Gnedenko [20] showed that the distribution F has rapidly varying tails if and only if the maxima of independent observations from F are *relatively stable* in the following sense.

Definition 2.2 (Relative stability). Let $\varepsilon_p = (\varepsilon_p(j))_{j=1}^p$ be a sequence of random variables with identical marginal distributions F. Define the sequence $(u_p)_{p=1}^\infty$ to be the $(1-1/p)$th generalized quantile of F, that is,

$$u_p = F^{\leftarrow}(1 - 1/p). \tag{2.16}$$

The triangular array $\mathcal{E} = \{\varepsilon_p, p \in \mathbb{N}\}$ is said to have relatively stable (RS) maxima if

$$\frac{1}{u_p} M_p := \frac{1}{u_p} \max_{j=1,\dots,p} \varepsilon_p(j) \xrightarrow{\mathbb{P}} 1, \quad \text{as } p \to \infty. \tag{2.17}$$

In the case of independent and identically distributed $\varepsilon_p(j)$'s, Barndorff-Nielsen [3] and Resnick and Tomkins [35] obtained necessary and sufficient conditions for the *almost sure stability* of maxima, where the convergence in (2.17) holds almost surely.

While relative stability (and almost sure stability) is well-understood in the independent case, the role of dependence has not been fully explored. We start this exploration with a small refinement of Theorem 2 in Gnedenko [20] valid under arbitrary dependence.

Proposition 2.1 (Rapid variation and relative stability). *Assume that the array $\mathcal{E}$ consists of identically distributed random variables with cumulative distribution function F, where $F(x) < 1$ for all finite $x > 0$.*

1. *If F has rapidly varying right tail, then for all $\delta > 0$,*

$$\mathbb{P}\left[\frac{1}{u_p} M_p \le 1 + \delta\right] \ge 1 - \frac{\overline{F}((1+\delta)u_p)}{\overline{F}(u_p)} \to 1. \tag{2.18}$$

2. *If, in addition, the array $\mathcal{E}$ has independent entries, then it is relatively stable if and only if F has rapidly varying tail.*

Proof of Proposition 2.1. By the union bound and the fact that $p\overline{F}(u_p) \le 1$, we have

$$\mathbb{P}\big[M_p > (1+\delta)u_p\big] \le p\overline{F}\big((1+\delta)u_p\big) \le \frac{\overline{F}((1+\delta)u_p)}{\overline{F}(u_p)}. \tag{2.19}$$

In view of (2.15) (rapid variation) and the fact that $u_p \to \infty$, as $p \to \infty$, the right-hand side of (2.19) vanishes as $p \to \infty$, for all $\delta > 0$. This completes the proof of (2.18). Part 2 is a re-statement of the classic result due to Gnedenko in [20]. □

We next demonstrate that Gaussian, Exponential, Laplace, and Gamma distributions all have rapidly varying tails.

Example 2.2 (Generalized AGG). A distribution is said to have *Generalized AGG* right tail, if $\log \overline{F}$ is regularly varying,

$$\log \overline{F}(x) = -x^{\nu} L(x), \tag{2.20}$$

where $\nu > 0$ and $L : (0, +\infty) \to (0, +\infty)$ is a slowly varying function. (A function is said to be slowly varying if $\lim_{x\to\infty} L(tx)/L(x) = 1$ for all $t > 0$.) Note that the AGG(ν) model corresponds to the special case where $L(x) \to 1/\nu$, as $x \to \infty$.

Relation (2.18) holds for all arrays $\mathcal{E}$ with *generalized* AGG marginals; if the entries are independent, the maxima are relatively stable. This follows directly from Proposition 2.1, once we show that F has rapidly varying tail. Indeed, by (2.20), we have

$$\log\big(\overline{F}(tx)/\overline{F}(x)\big) = -L(x)x^{\nu}\left(t^{\nu}\frac{L(tx)}{L(x)} - 1\right),$$

which converges to $-\infty$, 0, and $+\infty$, as $x \to \infty$, when $t > 1$, $t = 1$, and $t < 1$, respectively, since $x^{\nu}L(x) \to \infty$ as $x \to \infty$ by definition of L.

The AGG class encompasses a wide variety of rapidly varying tail models such as Gaussian and double exponential distributions. The larger class (2.20) is needed, however, for the Gamma distribution.

More generally, distributions with heavier tails (e.g., log-normal) and ligher tails (e.g., Gompertz) outside the generalized AGG class (2.20) may also possess rapidly varying tails; heavy-tailed distributions like the Pareto and t-distributions, on the other hand, do not. Strong classification boundaries for these classes of models are discussed in Sections D and E of the supplement [18]. For brevity, we focus here on the AGG(ν) models.

2.3. Dependence and uniform relative stability

An important ingredient needed for a converse of Theorem 2.1 is an appropriate characterization of the error dependence structure under which the strong classification boundary (1.13) is tight. The notion of *uniform relative stability* turns out to be the key.

Definition 2.3 (Uniform Relative Stability). Under the notations established in Definition 2.2, the triangular array $\mathcal{E}$ is said to have uniform relatively stable (URS) maxima if for *every* sequence of subsets $S_p \subseteq \{1, \dots, p\}$ such that $|S_p| \to \infty$, we have

$$\frac{1}{u_{|S_p|}} M_{S_p} := \frac{1}{u_{|S_p|}} \max_{j \in S_p} \varepsilon_p(j) \xrightarrow{\mathbb{P}} 1, \tag{2.21}$$

as $p \to \infty$, where $u_q, q \in \{1, \dots, p\}$ is the generalized quantile in (2.16). The collection of arrays $\mathcal{E} = \{\varepsilon_p(j)\}$ with URS maxima is denoted $U(F)$.

Uniform relative stability is, as its name suggests, a stronger requirement on dependence than relative stability. From the last section we see that, an array with i.i.d. components sharing a marginal distribution F with rapidly varying tails has relatively stable maxima; it is easy to see that URS also follows, by independence of the entries.

Corollary 2.2. *An independent array $\mathcal{E}$ with common marginals $F \in \mathrm{AGG}(\nu)$, $\nu > 0$, is URS; in this case, URS holds with $u_{|S_p|} \sim (\nu \log |S_p|)^{1/\nu}$.*

On the other hand, RS and URS hold under much broader dependence structures than just independent errors. The latter condition is extremely mild and can be shown to hold for many classes of error models. In Section 3 below, we will focus extensively on the Gaussian case, which is of great interest in applications and is rather challenging. Specifically, we will provide simple necessary and sufficient condition for uniform relative stability in terms of their covariance structure.

The relative stability concepts can be used to characterize dependence structures under which the maxima of error sequences *concentrate* around the quantiles (2.16) in the sense of (2.17). This concentration of maxima phenomenon, is the key to the phase-transition results in support recovery problems, discussed next.

2.4. Necessary conditions for exact support recovery

With the preparations from Section 2.3, we are ready to state the necessary conditions for exact support recovery (1.14) by thresholding procedures. It turns out that the strong classification boundary (1.13) is tight, under the general dependence structure characterized by URS (Definition 2.3).

Formally, we define the parameter space for the signals μ to be

$$\begin{aligned}\Theta_p^-(\beta, \overline{r}) = \big\{&\mu \in \mathbb{R}^p : \text{there exists a set } S_p \subseteq \{1, \dots, p\} \text{ such that } |S_p| = \lfloor p^{1-\beta} \rfloor, \\ &0 < \mu(j) \le (\nu \overline{r} \log p)^{1/\nu} \text{ for all } j \in S_p, \text{and } \mu(j) = 0 \text{ for all } j \notin S_p\big\}.\end{aligned} \tag{2.22}$$

Theorem 2.2. *Let $\mathcal{E}$ be a triangular array with common* AGG(ν) *marginal F, $\nu > 0$. Assume further that the errors $\mathcal{E}$ have uniform relatively stable maxima and minima, i.e., $\mathcal{E} \in U(F)$, and $(-\mathcal{E}) = \{-\varepsilon_p(j)\} \in U(F)$. If*

$$\overline{r} < g(\beta) = \big(1 + (1-\beta)^{1/\nu}\big)^{\nu}, \tag{2.23}$$

then

$$\lim_{p\to\infty} \inf_{\widehat{S}_p \in \mathcal{T}} \inf_{\mu \in \Theta_p^-(\beta,\overline{r})} \mathbb{P}[\widehat{S}_p \neq S_p] = 1, \tag{2.24}$$

where $\mathcal{T}$ is the class of all thresholding procedures (1.6).

Our first comment is on the signal sizes, and in particular, the gap between the sufficient conditions (Theorem 2.1) and the necessary conditions (Theorem 2.2).

Remark 2.3. The sufficient condition in Theorem 2.1 requires that *all* signals be larger than the strong classification boundary $g(\beta)$ in order to achieve exact support recovery (1.14), while Theorem 2.2 states that exact support recovery fails (in the sense of (1.15)) when *all* signal sizes are below the boundary – the two conditions are *not* complements of each other. This gap between the sufficient and necessary conditions on signal sizes, however, may be difficult to bridge. Indeed, in general, when signal sizes straddle the boundary $g(\beta)$, either outcome is possible, as we demonstrate in Example 2.3 below.

Example 2.3 (Signals straddling the boundary). Let the signal μ have $|S_p| = \lfloor p^{(1-\beta)} \rfloor$ non-zero entries, composed of two disjoint sets $S_p = S_p^{(1)} \cup S_p^{(2)}$. Let also the magnitude of the signals be equal within the two sets, that is, $\mu(j) = \sqrt{2r^{(k)} \log p}$ if $j \in S_p^{(k)}$ for some constants $r^{(k)} > 0$ for $k = 1, 2$. For simplicity, assume that the errors are i.i.d. standard Gaussians.

Consider two scenarios

1. $r^{(1)} = (1+\delta)g(\beta)$, $r^{(2)} = (1+\delta)$ with $|S_p^{(1)}| = |S_p| - 1$, $|S_p^{(2)}| = 1$,
2. $r^{(1)} = (1+\delta)g(\beta)$, $r^{(2)} = (1-\delta)g(\beta)$ with $|S_p^{(1)}| = \lfloor |S_p|/2 \rfloor$, $|S_p^{(2)}| = |S_p| - |S_p^{(1)}|$,

for some constants $0 < \delta < 1 - \beta < 1$. In both cases, signals in $S_p^{(1)}$ (respectively, $S_p^{(2)}$) are above (respectively, below) the strong classification boundary (1.13). However, in the first scenario, we have $\mathbb{P}[\widehat{S}_p^{\mathrm{Bonf}} = S_p] \to 1$ where $\widehat{S}_p^{\mathrm{Bonf}}$ is the Bonferroni's procedure described in Theorem 2.1, while in the second scenario, we have $\mathbb{P}[\widehat{S}_p = S_p] \to 0$ for *all* thresholding procedures $\widehat{S}_p$.

The claims in Example 2.3 are verified in Section B of the supplement [18].

Our second comment is on the interplay between thresholding procedures and the dependence class characterized by URS.

Remark 2.4. Paraphrasing Theorems 2.1 and 2.2: if we consider only thresholding procedures, then for a very large class of dependence structures, we cannot improve upon the Bonferroni procedure $\widehat{S}_p^{\mathrm{Bonf}}$. Specifically, for all $\mathcal{E} \in U(F)$ and $-\mathcal{E} \in U(F)$, and for all $S_p \in \mathcal{S}$, where $\mathcal{S} = \{S \subseteq \{1, \dots, p\}; |S| = \lfloor p^{1-\beta} \rfloor\}$, we have

$$\lim_{p\to\infty} \mathbb{P}\big[\widehat{S}_p^{\mathrm{Bonf}} \neq S_p\big] = \begin{cases} \limsup\limits_{p\to\infty} \inf\limits_{\widehat{S}_p\in\mathcal{T}} \mathbb{P}[\widehat{S}_p \neq S_p] = 0, & \text{if } \underline{r} > g(\beta), \\ \liminf\limits_{p\to\infty} \inf\limits_{\widehat{S}_p\in\mathcal{T}} \mathbb{P}[\widehat{S}_p \neq S_p] = 1, & \text{if } \overline{r} < g(\beta), \end{cases} \tag{2.25}$$

where $\mathcal{T}$ is the set of all thresholding procedures (1.6).

Theorem 2.2 also yields an answer the question raised in Butucea et al. [6]. In particular, the authors of [6] commented that independent error is the ‘least favorable model’ in the problem of support recovery, and conjectured that the support recovery problem may be easier to solve under dependence, similar to how the problem of signal detection is easier under dependent errors (see [21]). Surprisingly, our results here state that asymptotically, *all* error dependence structures in the large URS class are equally difficult for *thresholding procedures*. Therefore, the phase-transition behavior is universal in the class of dependence structures characterized by URS.

To facilitate comparison with results in existing literature, we will formulate explicit minimax statements in Section 4.

We must emphasize that the restriction to the URS dependence class is *not* an assumption of convenience. The condition on dependence characterized by uniform relative stability is, in fact, the weakest of its kind in the literature; see Section 3 below.

Our third comment is on the practical implications of the phase transitions.

Remark 2.5. As pointed out by an anonymous referee, the error distributions are seldom known precisely in practice, and are often only known *approximately* through large sample properties of the observations/statistics x. It would be interesting to improve upon the current work, and study the non-asymptotic properties of the phenomena.

Nevertheless, asymptotic results such as Theorems 2.1 and 2.2 have already led to powerful insights into applications. For example, it has long been observed, empirically, that there is a sharp transition in the statistical power of estimating the set of relevant genetic locations in genome-wide association studies [5]. Using theory established in this paper, we have been able to show rigorously that such empirically observed “sharp power curves” are simply manifestations of the phase transition phenomena; we refer readers to the recent work [17] for a dedicated treatment.

Section F of the supplement [18] offers some further discussions on the statistical implications of the results in this section.

We conclude with the proof of Theorem 2.2.

Proof of Theorem 2.2. To avoid cumbersome double subscript notations, we will sometimes suppress dependence on p of the set sequences $\widehat{S}_p$ and S_p in the proof.

Since the estimator $\widehat{S}_p = \{x(j) \geq t_p(x)\}$ is thresholding, exact support recovery takes place if and only if the threshold separates the signals and null part, that is,

$$\mathbb{P}[\widehat{S}_p = S_p] = \mathbb{P}\Big[\max_{j\in S^c} x(j) < t_p(x) \leq \min_{j\in S} x(j)\Big] \leq \mathbb{P}\Big[\max_{j\in S^c} x(j) < \min_{j\in S} x(j)\Big].$$

Since the right-hand side does not depend on the procedure $\widehat{S}_p$, we also have

$$\sup_{\widehat{S}_p\in\mathcal{T}} \mathbb{P}[\widehat{S}_p = S_p] \leq \mathbb{P}\Big[\max_{j\in S^c} x(j) < \min_{j\in S} x(j)\Big] \leq \mathbb{P}\Big[\max_{j\in S^c} \varepsilon(j) < \overline{\Delta} + \min_{j\in S} \varepsilon(j)\Big], \tag{2.26}$$

where we used the assumption that the signal sizes are no greater than $\overline{\Delta}$. Let $S^* = S_p^*$ be a sequence of support sets that maximize the right-hand side of (2.26), that is, let

$$S_p^* \in \underset{S\subseteq\{1,\dots,p\}:|S|=\lfloor p^{1-\beta}\rfloor}{\arg\max} \mathbb{P}\Big[\max_{j\in S^c} \varepsilon(j) < \overline{\Delta} + \min_{j\in S} \varepsilon(j)\Big].$$

Then, we obtain the following bound which only depends on $\overline{r}$ and the distribution of $\mathcal{E}$,

$$\begin{aligned}\sup_{\widehat{S}_p\in\mathcal{T}} \sup_{\mu\in\Theta_p^-(\beta,\overline{r})} \mathbb{P}[\widehat{S}_p = S_p] &\leq \mathbb{P}\Big[\max_{j\in S^{*c}} \varepsilon(j) < \overline{\Delta} + \min_{j\in S^*} \varepsilon(j)\Big] \\ &= \mathbb{P}\left[\frac{M_{S^{*c}}}{u_p} < \frac{\overline{\Delta} - m_{S^*}}{u_p}\right],\end{aligned} \tag{2.27}$$

where $M_{S^{*c}} = \max_{j\in S^{*c}} \varepsilon(j)$ and $m_{S^*} = \max_{j\in S^*}(-\varepsilon(j))$. Since the error arrays $\mathcal{E}$ and $(-\mathcal{E})$ are URS by assumption, using the expression for the AGG quantiles (1.3), we have

$$\frac{M_{S^{*c}}}{u_p} = \frac{M_{S^{*c}}}{u_{|S^{*c}|}}\frac{u_{|S^{*c}|}}{u_p} \xrightarrow{\mathbb{P}} 1 \quad \text{and} \quad \frac{m_{S^*}}{u_p} = \frac{m_{S^*}}{u_{|S^*|}}\frac{u_{|S^*|}}{u_p} \xrightarrow{\mathbb{P}} (1-\beta)^{1/\nu}, \tag{2.28}$$

so that the two random terms in probability (2.27) converge to constants. Notice that the second relation in (2.28) holds by URS for any $\beta \in (0, 1)$; when $\beta = 1$, the relation holds since $u_{|S^*|}/u_p$ vanishes while $\{m_{S^*}/u_{|S^*|}\}$ is tight.

Since signal sizes are bounded above by $\overline{r} < (1 + (1-\beta)^{1/\nu})^\nu$, we can write $\overline{r}^{1/\nu} = 1 + (1-\beta)^{1/\nu} - d$ for some $d > 0$. By our parametrization of $\overline{\Delta}$, we have

$$\frac{\overline{\Delta}}{u_p} = \big(1 + (1-\beta)^{1/\nu} - d\big)\big(1 + o(1)\big). \tag{2.29}$$

Combining (2.28) and (2.29), we conclude that the right-hand side of the probability (2.27) converges in probability to a constant strictly less than 1, that is,

$$\frac{\overline{\Delta} - m_S}{u_p} \xrightarrow{\mathbb{P}} 1 - d, \tag{2.30}$$

while $M_{S^{*c}}/u_p \xrightarrow{\mathbb{P}} 1$. Therefore, the probability in (2.27) must go to 0. □

2.5. Dense signals

We treat briefly the case of dense signals, where the size of the support set is proportional to the problem dimension, that is, $s \sim cp$ for some constant $c \in (0,1)$. We show that in this case, a phase-transition-type result still holds, independently of the value of c. Analogous to the set-up of Theorems 2.1 and 2.2, let

$$\begin{aligned}\Theta_p^{\mathrm{d}+}(c,\underline{r}) = \big\{\mu \in \mathbb{R}^p : &\text{there exists a set } S_p \subseteq \{1,\ldots,p\} \text{ such that } |S_p| \le \lfloor cp \rfloor,\\ &\mu(j) \ge (\nu \underline{r} \log p)^{1/\nu} \text{ for all } j \in S_p, \text{ and } \mu(j) = 0 \text{ for all } j \notin S_p\big\},\end{aligned} \tag{2.31}$$

where "d" in the notation $\Theta_p^{\mathrm{d}+}$ stands for "dense". Similarly, define

$$\begin{aligned}\Theta_p^{\mathrm{d}-}(c,\overline{r}) = \big\{\mu \in \mathbb{R}^p : &\text{there exists a set } S_p \subseteq \{1,\ldots,p\} \text{ such that } |S_p| = \lfloor cp \rfloor,\\ &0 < \mu(j) \le (\nu \overline{r} \log p)^{1/\nu} \text{ for all } j \in S_p, \text{ and } \mu(j) = 0 \text{ for all } j \notin S_p\big\}.\end{aligned} \tag{2.32}$$

Theorem 2.3. *Let $c \in (0,1)$ be a fixed constant. In the context of Theorem 2.1, if $\underline{r} > 1$, then we have*

$$\lim_{p\to\infty} \sup_{\mu \in \Theta_p^{d+}(c,\underline{r})} \mathbb{P}[\widehat{S}_p \neq S_p] = 0. \tag{2.33}$$

While in the context of Theorem 2.2, if $\overline{r} < 1$, then

$$\lim_{p\to\infty} \inf_{\widehat{S}_p \in \mathcal{T}} \inf_{\mu \in \Theta_p^{d-}(c,\overline{r})} \mathbb{P}[\widehat{S}_p \neq S_p] = 1, \tag{2.34}$$

where $\mathcal{T}$ is the class of all thresholding procedures (1.6).

Remark 2.6. Notice that the boundary for the signal size parameter is identically 1 in this dense regime. Therefore, if we interpret $\beta = 0$ of the parametrization (1.4) as $s \sim cp$, where $c \in (0,1)$, then the strong classification boundary (1.13) may be continuously extended to the left-end point where $g(0) = 1$.

Proof of Theorem 2.3. The proof is entirely analogous to that of Theorems 2.1 and 2.2. Specifically, (2.33) follows by replacing $\lfloor p^{1-\beta} \rfloor$ with $\lfloor cp \rfloor$ in Relation (2.9) onward, and replacing (2.12) with

$$u_s^- \sim (\nu \log cp)^{1/\nu} \sim (\nu \log p)^{1/\nu}.$$

in the proof of Theorem 2.1. Similarly, (2.34) follows the proof of Theorem 2.2. Indeed, by using the fact that

$$\frac{u_{|S^{*c}|}}{u_p} \sim \frac{(\nu \log (1-c)p)^{1/\nu}}{(\nu \log p)^{1/\nu}} \to 1$$

and $u_{|S^*|}/u_p \to 1$ for all $c \in (0,1)$, we see that Relation (2.28) holds with $\beta = 0$, and the rest of Theorem 2.2 applies. □

3. Characterization of URS for Gaussian arrays

As promised, we characterize in this section the URS class of dependence structures in the case of Gaussian errors with a transparent necessary and sufficient condition in terms of their covariances.

Definition 3.1 (Uniformly decreasing dependence (UDD)). Consider a triangular array of jointly Gaussian distributed errors $\mathcal{E} = \{(\varepsilon_p(j))_{j=1}^p, p = 1, 2, \ldots\}$ with unit variances,

$$\varepsilon_p \sim \mathrm{N}(0, \Sigma_p), \quad p = 1, 2, \ldots.$$

The array $\mathcal{E}$ is said to be uniform decreasingly dependent (UDD) if for every $\delta > 0$ there exists a finite $N(\delta) < \infty$, such that for every $j \in \{1, \ldots, p\}$, and $p \in \mathbb{N}$, we have

$$\left|\left\{k \in \{1, \ldots, p\} : \Sigma_p(j, k) > \delta\right\}\right| \le N(\delta) \quad \text{for all } \delta > 0. \tag{3.1}$$

That is, for every coordinate j, the number of elements which are more than δ-correlated with $\varepsilon_p(j)$ does not exceed $N(\delta)$.

Note that the bound in (3.1) holds uniformly in j and p, and only depends on δ. Also observe that on the left-hand side of (3.1), we merely count in each row of Σ_p the number of exceedances of covariances (not their absolute values!) over level δ.

Remark 3.1. Without loss of generality, we may require that $N(\delta)$ be a monotone non-increasing function of δ, for we can take

$$N(\delta) = \sup_{p,j} \left|\left\{k : \Sigma_p(j, k) > \delta\right\}\right|,$$

which is non-increasing in δ. Definition 3.1 therefore states that the array is UDD when $N(\delta) < \infty$ for all $\delta > 0$.

Observe that the UDD condition does not depend on the order of the coordinates in the error vector $\varepsilon_p = (\varepsilon_p(j))_{j=1}^p$. Often times, however, the errors are thought of coming from a stochastic process indexed by time or space. To illustrate the generality of the UDD condition, we formulate next a simple sufficient condition (UDD′) that is easier to check in a time-series context.

Definition 3.2 (UDD′). For $\varepsilon_p \sim \mathrm{N}(0, \Sigma_p)$ with unit variances, an array $\mathcal{E} = (\varepsilon_p(j))_{j=1}^p$ is said to satisfy the UDD′ condition if there exist:

(i) permutations $l = l_p$ of $\{1, \ldots, p\}$, for all $p \in \mathbb{N}$, and
(ii) a non-negative sequence $(r_n)_{n=1}^\infty$ converging to zero $r_n \to 0$, as $n \to \infty$,

such that

$$\sup_{p \in \mathbb{N}} \left|\Sigma_p(i', j')\right| \le r_{|i-j|}, \tag{3.2}$$

where $i' = l(i)$, $j' = l(j)$, for all $i, j \in \{1, \ldots, p\}$.

Remark 3.2. Without loss of generality, we may also require that r_n be non-increasing in n, for we can replace r_n with the non-increasing sequence $r_n' = \sup_{m \geq n} r_m$.

Proposition 3.1. *UDD′ implies UDD.*

Proof. Since $r_n \to 0$, for any $\delta > 0$, there exists an integer $M = M(\delta) < \infty$ such that $r_n \leq \delta$, for all $n \geq M$. Thus, by (3.2), for every fixed $j' \in \{1, \dots, p\}$, we can have $|\operatorname{Cov}(\varepsilon_p(k'), \varepsilon_p(j'))| > \delta$, only if k' belongs to the set:

$$\bigl\{k' \in \{1, \dots, p\} : j - M \leq k := l_p^{-1}\bigl(k'\bigr) \leq j + M\bigr\},$$

where $j := l_p^{-1}(j')$. That is, there are at most $2M + 1 < \infty$ indices $k' \in \{1, \dots, p\}$, whose covariances with $\varepsilon(j')$ may exceed δ. Since this holds uniformly in $j' \in \{1, \dots, p\}$, Condition UDD follows with $N(\delta) = 2M + 1$. □

We now state the main result of this section: a Gaussian sequence is URS if and only if it is UDD. The URS condition essentially requires that the dependencies decay in a uniform fashion, the rate at which dependence decay does *not* matter.

Theorem 3.1. *Let $\mathcal{E}$ be a Gaussian triangular array with standard normal marginals. The array $\mathcal{E}$ has uniformly relatively stable* (*URS*) *maxima if and only if it is uniformly decreasing dependent* (*UDD*).

The proof of Theorem 3.1 is given in Section 5.

Returning again to the study of support recovery problems, Theorems 3.1 and 2.2 yields the following corollary.

Corollary 3.1. *For UDD Gaussian errors, the result in Theorem 2.2 holds.*

As a counterpart to Remark 2.4, we demonstrate the tightness of the dependence conditions in Theorem 2.2. Specifically, we demonstrate that if the URS dependence condition is violated, then it may be possible to recover the support of weaker signals below the boundary.

Example 3.1. Suppose $\mathcal{E} = (\varepsilon_p(j))_{j=1}^p$ is Gaussian, and is comprised of $\lfloor p^{1-\beta} \rfloor$ blocks, each of size at least $\lfloor p^\beta \rfloor$; let the elements of each block have correlation 1, and let elements from different blocks be independent. If $\underline{r} \geq 4(1 - \beta)$, then the procedure $\widehat{S} = \{j : x(j) > \sqrt{2(1 - \beta)\log p}\}$ yields $\mathbb{P}[\widehat{S} = S] \to 1$. This requirement on signal size is strictly weaker than that of the strong classification boundary, since $4(1 - \beta) < (1 + \sqrt{1 - \beta})^2$ on $\beta \in (0, 1)$.

The above example shows that if the correlations of the Gaussian errors do not decay in a uniform fashion (UDD fails), then we can do substantially better in terms of support recovery. The claims in the example are verified in Section B of the supplementary material [18], while numerical simulations of this example can be found in Section A therein.

We conclude this section with a brief discussion on the relationships between UDD and other dependence conditions in the context of extreme value theory.

Suppose that the array of errors $\mathcal{E}$ comes from a stationary Gaussian time series $\varepsilon(j)$, $j \in \mathbb{N}$, with auto-covariance $r_p = \mathrm{Cov}(\varepsilon(j+p), \varepsilon(j))$. One is interested in the asymptotic behavior of the maxima $M_p := \max_{j=1,\dots,p} \varepsilon(j)$.

In this setting, the Berman's condition, introduced in [4], requires that

$$r_p \log p \to 0, \quad \text{as } p \to \infty. \tag{3.3}$$

This condition entails that

$$a_p(M_p - b_p) \xrightarrow{d} Z, \quad \text{as } p \to \infty, \tag{3.4}$$

with the Gumbel limit distribution $\mathbb{P}[Z \le x] = \exp\{-e^{-x}\}$, $x \in \mathbb{R}$, where

$$a_p = \sqrt{2\log p}, \qquad b_p = \sqrt{2\log p} - \frac{1}{2}(\sqrt{2\log p})^{-1}\big(\log\log(p) + \log(4\pi)\big),$$

are *the same* centering and normalization sequences as in the case of iid $\varepsilon(j)$'s. Berman's condition is one of the weakest dependence conditions in the literature for which this result holds. See, for example, Theorem 4.4.8 in [14], where (3.3) is described as "very weak".

For dependence conditions weaker than (3.3), the sequences of normalizing and centering constants in (3.4) are *different* from the i.i.d. case, and the corresponding limit is no longer Gumbel; see, for example, Theorems 6.5.1 and 6.6.4 in [29], and [30].

On the other hand, in our high dimensional support estimation context, the notion of relative stability is sufficient and more natural than the finer notions of distributional convergence. If one is merely interested in the asymptotic relative stability of the Gaussian maxima, then Berman's condition can be relaxed significantly (see also, Theorem 4.1 of [4]). Observe that by Proposition 3.1, the Berman condition (3.3) implies UDD and hence relative stability (Theorem 3.1), that is,

$$\frac{1}{b_p} M_p \xrightarrow{\mathbb{P}} 1, \quad \text{as } p \to \infty. \tag{3.5}$$

This *concentration of maxima* property can be readily deduced from (3.4), since $a_p b_p \sim 2\log(p) \to \infty$ as $p \to \infty$. Theorem 3.1 shows that (3.5) holds if the much weaker uniform dependence condition UDD holds. Note that our condition is coordinate free – neither monotonicity of the sequence r_p nor stationarity of the underlying array is required.

4. Minimax optimality, and Bayes (sub)optimality of thresholding procedures

To facilitate comparison with the literature, we establish in this section minimax versions of our results from Section 2. Specifically, if we restrict ourselves to the class of thresholding procedures $\mathcal{T}$ (defined in (1.6)), then Bonferroni's procedure is minimax optimal, for *any* fixed dependence structures in the URS class. This is formalized in Corollary 4.1 in Section 4.1. We refer to this result as *point-wise* minimax, to emphasize the fact that this optimality holds for every fixed URS array.

Meanwhile, if we search over *all procedures*, but expand the parameter space to include all dependence structures, then a different minimax optimality statement holds for Bonferroni's procedure. With a careful study of the finite-sample Bayes optimality of thresholding procedures in Sections 4.2 and 4.3, we establish this minimax statement in Section 4.4.

Finally, we offer some insights into the support recovery problem in the case when errors have heavier-than-exponential tails in Section 4.5.

4.1. Point-wise minimax optimality

Theorems 2.1 and 2.2 can be cast in the form of an asymptotic minimax statement.

Corollary 4.1 (Point-wise minimax). *Let $\widehat{S}^{\mathrm{Bonf}}$ be the sequence of Bonferroni's procedure described in Theorem 2.1. Let also the errors have common* AGG(ν) *distribution F with parameter $\nu > 0$, and Θ_p^+ be as defined in (2.1). If $\underline{r} > g(\beta)$, then we have*

$$\limsup_{p\to\infty} \sup_{\mu\in\Theta_p^+(\beta,\underline{r})} \mathbb{P}\big(\widehat{S}_p^{\mathrm{Bonf}} \neq S_p\big) = 0, \tag{4.1}$$

for arbitrary dependence structure of the error array $\mathcal{E} = \{\varepsilon_p(j)\}_p$. Let $\mathcal{T}$ be the class of thresholding procedures (1.6). If $\underline{r} < g(\beta)$, then we have

$$\liminf_{p\to\infty} \inf_{\widehat{S}_p\in\mathcal{T}} \sup_{\mu\in\Theta_p^+(\beta,\underline{r})} \mathbb{P}(\widehat{S}_p \neq S_p) = 1, \tag{4.2}$$

for any error dependence structure such that $\mathcal{E} \in U(F)$ and $(-\mathcal{E}) \in U(F)$.

Proof of Corollary 4.1. The first conclusion (4.1) is a restatement of Theorem 2.1.

For the second statement (4.2), since $\underline{r} < g(\beta)$, we can pick a sequence $\mu^* \in \Theta_p^+(\beta,\underline{r})$ such that $|S_p| = \lfloor p^{1-\beta}\rfloor$, with signals having the same signal size $\mu(j) = (2r\log p)^{1/\nu}$ for all $j \in S_p$, where $\underline{r} < r < g(\beta)$. For this particular choice of μ^* we have $\mu^* \in \Theta_p^-(\beta,\overline{r})$ where $r < \overline{r} < g(\beta)$, and according to Theorem 2.2, we obtain $\lim_{p\to\infty} \inf_{\widehat{S}_p\in\mathcal{T}} \mathbb{P}[\widehat{S}_p \neq S_p] = 1$, for all dependence structures in the URS class. □

Remark 4.1. Theorem 2.2 is a stronger result than the traditional minimax claim in Relation (4.2). Indeed, (2.24) involves an infimum (over the class Θ_p^-) while (4.2) has a supremum (over the class Θ_p^+).

On the other hand, Corollary 4.1 is more informative than many minimax-type statements, since it applies "point-wise" to any fixed error dependence structure in the URS class.

Remark 4.2. Corollary 4.1 echoes Remark 2.4: for a very large class of dependence structures, we cannot improve upon Bonferroni's procedure in exact support recovery problems (asymptotically), unless we look beyond thresholding procedures.

4.2. Bayes optimality in support recovery problems

In studying support recovery problems (e.g., Arias-Castro and Chen [1]), restrictions to the thresholding procedures are sometimes justified by arguing that such procedures are the "reasonable" choice for estimating the support set. We show in this section that, perhaps surprisingly, for general error models, thresholding procedures are not always optimal, even when the observations are independent.

We shall study the optimal procedure for support recovery problems under a Bayesian setting with general distributional assumptions (including but not limited to additive models (1.1)). Specifically, we assume that there is an ordered set $P = (j_1, \ldots, j_s)$, $j_i \in \{1, \ldots, p\}$, and s (not necessarily equal) densities $f_1, \ldots, f_s$, such that the observations indexed by P have corresponding densities. That is,

$$x(j_i) \sim f_i, \quad i = 1, \ldots, s. \tag{4.3}$$

Let also the rest $(p - s)$ observations have common density f_0, that is, $x(j) \sim f_0$ for $j \notin S$. We further assume that the observations x are mutually independent.

We adopt here a Bayesian framework to measure statistical risks. Let the ordered support $P = (j_1, \ldots, j_s)$ have prior

$$\pi\big((j_1, \ldots, j_s)\big) := \mathbb{P}\big[P = (j_1, \ldots, j_s)\big] = (p - s)!/p!, \tag{4.4}$$

for all distinct $1 \le j_1, \ldots, j_s \le p$. Consequently, the unordered support $S = \{j_1, \ldots, j_s\}$ is distributed uniformly in the collection of all set of size s, with the unordered uniform distribution π^{u}. That is, for all for all $S \in \mathcal{S} := \{S \subseteq \{1, \ldots, p\}; |S| = s\}$, we have

$$\pi^{\mathrm{u}}\big(\{j_1, \ldots, j_s\}\big) := \mathbb{P}\big[S = \{j_1, \ldots, j_s\}\big] = (p - s)!s!/p!. \tag{4.5}$$

Consider the loss function,

$$\ell(\widehat{S}, S) := \mathbb{P}[\widehat{S} \neq S] = \mathbb{P}_P[\widehat{S} \neq S],$$

where the probability is taken over the randomness in the observations for a fixed configuration P, the Bayes optimal procedures should minimize

$$\mathbb{E}_\pi \mathbb{P}[\widehat{S} \neq S], \tag{4.6}$$

where the expectation is taken over the random configurations P, with a uniform distribution π as specified in (4.4).

If, however, the sparsity $s = |S|$ of the problem is known, then a natural estimator for S would be based on the set of top s order statistics.

Definition 4.1 (Oracle data thresholding). We call $\widehat{S}^* = \{j \mid x(j) \ge x_{[s]}\}$ the oracle data thresholding procedure, where $x_{[1]} \ge \cdots \ge x_{[p]}$ are the order statistics of x.

The finite-sample optimality of the oracle thresholding procedure $\widehat{S}^*$ is intimately linked with the *monotone likelihood ratio* (MLR) property.

Definition 4.2 (Monotone Likelihood Ratio). A family of positive densities on $\mathbb{R}$, $\{f_\delta, \delta \in U\}$, is said to have the MLR property if, for all $\delta_0, \delta_1 \in U \subseteq \mathbb{R}$ such that $\delta_0 < \delta_1$, the likelihood ratio $(f_{\delta_1}(x)/f_{\delta_0}(x))$ is an increasing function of x.

Their relationship is summarized in the following lemma.

Proposition 4.1. *Let the observations $x(j)$, $j = 1, \ldots, p$ be as prescribed as in* (4.3) *through* (4.4). *If each of $\{f_0, f_1\}, \ldots, \{f_0, f_s\}$ form an MLR family, then the oracle data thresholding procedure $\widehat{S}^* = \{j \mid x(j) \geq x_{[s]}\}$ is finite-sample optimal in terms of Bayes risk $\mathbb{E}_\pi \mathbb{P}[\widehat{S} \neq S]$. That is,*

$$\widehat{S}^* \in \arg\min_{\widehat{S}} \mathbb{E}_\pi \mathbb{P}[\widehat{S} \neq S]. \tag{4.7}$$

for all s and p.

The proof of Proposition 4.1 is found in Section B of the supplement [18].

We emphasize that the oracle thresholding procedures are in fact *finite-sample optimal* in the above Bayesian context. Further, our setup allows for different alternative distributions, and relaxes the assumptions of Butucea et al. [6] when studying distributional generalizations, where the alternatives are assumed to be identically distributed.

It remains to understand when the key MLR property holds. We elaborate on this question next.

4.3. Bayes optimality under sub-exponential errors

Returning to the more concrete signal-plus-noise model (1.1), it turns out that the error tail behavior is what determines the optimality of data thresholding procedures. In this setting, log-concavity of the error densities is *equivalent* to the MLR property (Lemma 4.1). This, in turn, yields the finite-sample optimality of data thresholding procedures (Proposition 4.2).

Lemma 4.1. *Let δ be the magnitude of the non-zero signals in the signal-plus-noise model* (1.1) *with positive error density f_0, and let $f_\delta(x) = f_0(x - \delta)$. The family $\{f_\delta, \delta \in \mathbb{R}\}$ has the MLR property if and only if the error density f_0 is log-concave.*

Proof of Lemma 4.1. Suppose MLR holds, we will show that $f_0(t) = \exp\{\phi(t)\}$ for some concave function ϕ. By the assumption of MLR, for any $x_1 < x_2$, setting $\delta_0 = 0$, and $\delta_1 = (x_2 - x_1)/2 > 0$, we have

$$\log \frac{f_{\delta_1}(x_2)}{f_{\delta_0}(x_2)} = \phi\left(\frac{(x_1 + x_2)}{2}\right) - \phi(x_2) \geq \phi(x_1) - \phi\left(\frac{(x_1 + x_2)}{2}\right) = \log \frac{f_{\delta_1}(x_1)}{f_{\delta_0}(x_1)}.$$

This implies that the log-density $\phi(t)$ is midpoint-concave, i.e., for all x_1 and x_2, we have,

$$\phi\left(\frac{(x_1 + x_2)}{2}\right) \geq \frac{1}{2}\phi(x_1) + \frac{1}{2}\phi(x_2). \tag{4.8}$$

For Lebesgue measurable functions, midpoint concavity is equivalent to concavity by the Sierpinki Theorem (see, e.g., Sec I.3 in [11]). This proves the 'only-if' part.

For the 'if' part, when $\phi(t) = \log(f_0(t))$ is log-concave, then for any $\delta_0 < \delta_1$, and any $x < y$, we have

$$\log \frac{f_{\delta_1}(y)}{f_{\delta_0}(y)} - \log \frac{f_{\delta_1}(x)}{f_{\delta_0}(x)} = \phi(y - \delta_1) - \phi(y - \delta_0) - \phi(x - \delta_1) + \phi(x - \delta_0) \geq 0, \tag{4.9}$$

where the last inequality is a simple consequence of concavity (see Lemma B.4 in the supplement [18]). This proves the 'if' part. □

Proposition 4.1 and Lemma 4.1 yield immediately the following.

Proposition 4.2. *Consider the additive error model* (1.1). *Let the errors ε be independent with common distribution F. Let the signal μ have s positive entries with magnitudes $0 < \delta_1 \leq \cdots \leq \delta_s$, located on $\{1, \ldots, p\}$ as prescribed in* (4.4). *If F has a positive, log-concave density f, then the oracle thresholding procedure $\widehat{S}^* = \{j; x(j) \geq x_{[s]}\}$ is finite-sample optimal in terms of Bayes risk in the sense of* (4.7).

Notice that under MLR (or equivalently, log-concavity of the errors in additive models), the oracle thresholding procedure is finite-sample optimal even in the case where the signals have different (positive) sizes.

The assumption of log-concavity of the densities is compatible with the AGG model when $\nu \geq 1$, as demonstrated in the next example.

Example 4.1. The generalized Gaussian density $f(x) \propto \exp\{-|x|^\nu/\nu\}$ is log-concave for all $\nu \geq 1$. Therefore in the additive error model (1.1), according to Proposition 4.2, the oracle thresholding procedure is Bayes optimal in the sense of (4.7).

Consider the asymptotic Bayes risk as defined in (4.6), the statement for the necessary condition of support recovery, with the help of Proposition 4.2, can be strengthened to include all procedures (in the Bayesian context), regardless of whether they are thresholding.

Theorem 4.1. *Consider the additive model* (1.1) *where the $\varepsilon_p(j)$'s are independent and identically distributed with log-concave densities in the AGG class. Let the signals be as prescribed in Proposition* 4.2. *If the signal sizes fall below the strong classification boundary* (1.13), *that is, $\overline{r} < g(\beta)$, then we have*

$$\liminf_{p\to\infty} \inf_{\widehat{S}_p} \mathbb{E}_\pi \mathbb{P}[\widehat{S}_p \neq S_p] = 1, \tag{4.10}$$

where the infimum on $\widehat{S}_p$ is taken over all procedures.

Proof of Theorem 4.1. When errors are independent with log-concave density, the oracle thresholding procedure $\widehat{S}_p^*$, by Proposition 4.2, minimizes the Bayes risk (4.6) among *all* procedures.

That is,

$$\liminf_{p\to\infty}\inf_{\widehat{S}_p}\mathbb{E}_\pi\mathbb{P}[\widehat{S}_p \neq S_p] \geq \liminf_{p\to\infty}\mathbb{E}_\pi\mathbb{P}\big[\widehat{S}_p^* \neq S_p\big].$$

Since $\widehat{S}_p^*$ belongs to the class of all thresholding procedures, we have

$$\begin{aligned}\liminf_{p\to\infty}\mathbb{E}_\pi\mathbb{P}\big[\widehat{S}_p^* \neq S_p\big] &\geq \liminf_{p\to\infty}\inf_{\widehat{S}_p\in\mathcal{T}}\mathbb{E}_\pi\mathbb{P}[\widehat{S}_p \neq S_p] \\ &\geq \liminf_{p\to\infty}\inf_{\widehat{S}_p\in\mathcal{T}}\inf_{S_p}\mathbb{P}[\widehat{S}_p \neq S_p] = 1,\end{aligned}$$

when $\overline{r} < g(\beta)$, where the last line follows from Theorem 2.2. □

4.4. Minimax optimality over all procedures

Theorem 4.1 allows us to state another minimax conclusion – one in which we search over *all procedures*, by allowing the supremum in the minimax statement to be taken over the dependence structures.

Corollary 4.2. *Let $D(F)$ be the the collection of error arrays with common marginal F as defined in* (2.14) *where F is an* AGG(ν) *distribution. Let also $\widehat{S}_p^{\text{Bonf}}$ be Bonferroni's procedure as described in Theorem* 2.1. *If $\underline{r} > g(\beta)$, then we have*

$$\limsup_{p\to\infty}\sup_{\substack{\mu\in\Theta_p^+(\beta,\underline{r})\\ \mathcal{E}\in D(F)}}\mathbb{P}\big(\widehat{S}_p^{\text{Bonf}} \neq S_p\big) = 0. \tag{4.11}$$

Further, when $\underline{r} < g(\beta)$, and F has a positive log-concave density f, we have

$$\liminf_{p\to\infty}\inf_{\widehat{S}_p}\sup_{\substack{\mu\in\Theta_p^+(\beta,\underline{r})\\ \mathcal{E}\in D(F)}}\mathbb{P}(\widehat{S}_p \neq S_p) = 1, \tag{4.12}$$

where the infimum on $\widehat{S}_p$ is taken over all procedures.

Remark 4.3. Since the class AGG(ν), $\nu \geq 1$ contains distributions with log-concave densities (Example 4.1), the minimax statement (4.12) continues to hold if the supremum is taken over the entire class $F \in$ AGG(ν), $\nu \geq 1$. We opted for a more informative formulation which emphasizes the log-concavity condition on the density of F.

Remark 4.4. Corollary 4.2 is no stronger than Corollary 4.1. In Corollary 4.1 we search over only the class of thresholding procedures, but offer a tight, point-wise lower bound on the asymptotic risk over the class of URS dependence structures. On the other hand, Corollary 4.2 provides a uniform lower bound for the asymptotic risk over all dependence structures, which may not be tight except in the case of independent errors.

Proof of Corollary 4.2. Relation (4.11) is a re-statement of Remark 2.1.

For any distribution π (with a slight abuse of notation) over the parameter space $\Theta_p^+ \times D(F)$, we have

$$\liminf_{p\to\infty} \inf_{\widehat{S}_p} \sup_{\substack{\mu\in\Theta_p^+(\beta,\underline{r})\\ \mathcal{E}\in D(F)}} \mathbb{P}(\widehat{S}_p \neq S_p) \geq \liminf_{p\to\infty} \inf_{\widehat{S}_p} \mathbb{E}_\pi \mathbb{P}(\widehat{S}_p \neq S_p), \tag{4.13}$$

since the supremum is bounded from below by expectations. In particular, define π to be the uniform distribution over the configurations $\Theta_p^* \times I(f)$, where

$$\Theta_p^* = \big\{\mu \in \mathbb{R}^d : |S_p| = \lfloor p^{1-\beta} \rfloor, \mu(j) = 0 \text{ for all } j \notin S, \text{and}$$
$$\mu(j) = (\nu r \log p)^{1/\nu} \text{ for all } j \in S, \text{where } \underline{r} < r < g(\beta)\big\},$$

and

$$I(f) = \big\{\mathcal{E} = \big(\varepsilon_p(j)\big)_p : \varepsilon_p(j) \text{ are independently and identically distributed}$$
$$\text{with density } f(x) \propto \exp\big\{-|x|^\nu/\nu\big\}\big\}.$$

Since the density f of F is log-concave, the distribution of the signal configurations satisfies the conditions in Theorem 4.1. Thus, the desired conclusion (4.12) follows from Theorem 4.1 and (4.13). □

4.5. Bayes optimality of likelihood ratio thresholding

The following result provides the general form of finite-sample Bayes optimal procedures. It turns out that in general, *likelihood ratio thresholding* is optimal.

Proposition 4.3. *Let the observations $x(j)$, $j = 1, \ldots, p$ have s signals as prescribed in (4.4) having common density f_a, and let the rest $(p - s)$ locations have common density f_0. Define the likelihood ratios*

$$L(j) := f_a\big(x(j)\big)/f_0\big(x(j)\big),$$

and let $L_{[1]} \geq L_{[2]} \geq \cdots \geq L_{[p]}$ be the order statistics of the $L(j)$'s. Then the procedure $\widehat{S}_{opt} = \{j \mid L(j) \geq L_{[s]}\}$ is finite-sample optimal in terms of Bayes risk. That is,

$$\widehat{S}_{opt} \in \arg\min_{\widehat{S}} \mathbb{E}_\pi \mathbb{P}[\widehat{S} \neq S]. \tag{4.14}$$

for all s and p, where the infimum on $\widehat{S}_p$ is taken over all procedures.

The proof of Proposition 4.3 is found in Section B of the supplement [18].

The characterization of optimal likelihood ratio thresholding procedures in Proposition 4.3 may not always yield practical estimators, as the density of alternatives, and number of signals are typically unknown. Nevertheless, some insights can still be gained by virtue of Proposition 4.3. In particular, when MLR fails (or equivalently, when the errors in model (1.1) do not have log-concave densities), data thresholding is sub-optimal.

Example 4.2 (Sub-optimality of data thresholding). Let the errors have i.i.d. generalized Gaussian density with $\nu = 1/2$, that is, $\log f_0(x) \propto -x^{1/2}$. Let dimension $p = 2$, sparsity $s = 1$ with uniform prior, and signal size $\delta = 1$. That is, $\mathbb{P}[\mu = (0, 1)^{\mathrm{T}}] = \mathbb{P}[\mu = (1, 0)^{\mathrm{T}}] = 1/2$. If the observations take on values $x = (x_1, x_2)^{\mathrm{T}} = (1, 2)^{\mathrm{T}}$, we see from a comparison of the likelihoods (and hence, the posteriors),

$$\log \frac{f(x|\{1\})}{f(x|\{2\})} = 2x_1^{1/2} + 2(x_2 - 1)^{1/2} - 2x_2^{1/2} - 2(x_1 - 1)^{1/2} = 4 - 2\sqrt{2} > 0,$$

that even though $x_1 < x_2$, the set $\{1\}$ is a better estimate of support than $\{2\}$, i.e., $\mathbb{P}[S = \{1\} \mid x] > \mathbb{P}[S = \{2\} \mid x]$.

This simple example shows that, in the case when the errors have super-exponential tails, the optimal procedures are in general *not* data thresholding. A slightly more general conclusion can be found in Corollary B.2 in the supplement [18].

Remark 4.5. Consider the model (1.1) with independent errors, Proposition 4.3, and indeed, Example 4.2 demonstrate that thresholding procedures are in fact *sub-optimal* for AGG(ν) models with $\nu < 1$. Therefore, the optimality of thresholding procedures (specifically, Bonferroni's procedure) only applies to AGG(ν) models with $\nu \geq 1$.

If we restrict the space of methods to only thresholding procedures, then results in Section 4.1 state that the phase-transition phenomenon – the 0-1 law in the sense of Corollary 4.1 – is universal in all error models with rapidly varying tails. This includes AGG(ν) models *for all* $\nu > 0$. In contrast, models with heavy (regularly varying) tailed errors do not exhibit this phenomenon (see Theorem E.1 in the supplementary material [18]). We summarize the properties of thresholding procedures in Table 1.

Table 1. Properties of thresholding procedures under different error distributions when errors are independent. Properties of the error distributions are listed in brackets

Thresholding procedure (Error distributions)	Bayes optimality (Log-concave density)	Phase-transition (Rapidly-varying tails)
AGG(ν), $\nu \geq 1$	Yes (Yes)	Yes (Yes)
AGG(ν), $0 < \nu < 1$	No (No)	Yes (Yes)
Power laws	No (No)	No (No)

5. Proof of Theorem 3.1

We now prove the main result in Section 3. We first introduce a key lemma regarding the structure of correlation matrix of high-dimensional random variables. The proof uses a surprising, yet elegant application of Ramsey's theorem from the study of combinatorics. The 'only if' part of Theorem 3.1 follows from this lemma, in Section 5.2.

The proof of the 'if' part is postponed until Section B of the supplement [18]. The arguments there was recently extended to establish bounds on the rate of concentration of maxima in [28]; see also, [39] and references therein for related work on this topic.

5.1. Ramsey's coloring theorem and structure of correlation matrices

Given any integer $k \geq 1$, there is always an integer $R(k,k)$ called the *Ramsey number*:

$$k \leq R(k,k) \leq \binom{2k-2}{k-1} \tag{5.1}$$

such that the following property holds: every undirected graph with at least $R(k,k)$ vertices will contain *either* a clique of size k, or an *independent set* of k nodes. Recall that a clique is a complete sub-graph where all pairs of nodes are connected, and an independent set is a set of nodes where no two nodes are connected.

This result is a consequence of the celebrated work of Ramsey [33], which gave birth to Ramsey Theory (see, e.g., Conlon, Fox and Sudakov [9]). The Ramsey theorem and the upper bound (5.1) (established first in [15]) are at the heart of the proof of the following result.

Proposition 5.1. *Fix $\gamma \in (0,1)$ and let $P = (\rho(i,j))_{n\times n}$ be an arbitrary correlation matrix. If*

$$k := \lfloor \log_2(n)/2 \rfloor \geq \lceil 1/\gamma \rceil + 1, \tag{5.2}$$

then there is a set of k indices $K = \{l_1, \ldots, l_k\} \subseteq \{1, \ldots, n\}$ such that

$$\rho(i,j) \geq -\gamma, \quad \text{for all } i,j \in K. \tag{5.3}$$

Proof of Proposition 5.1. By using (5.1) and a refinement of the Stirling's formula, we will show at the end of the proof that for $k \leq \log_2(n)/2$, we have

$$R(k,k) \leq n, \tag{5.4}$$

where $R(k,k)$ is the Ramsey number.

Now, construct a graph with vertices $\{1, \ldots, n\}$ such that there is an edge between nodes i and j if and only if $\rho(i,j) > -\gamma$. In view of (5.4) and Ramsey's theorem (see, e.g., Theorem 1 in [16] or [9] for a recent survey on Ramsey theory), there is a subset of k nodes $K = \{l_1, \ldots, l_k\}$, which is either a *complete graph* or an *independent set*.

If K is a complete graph, then by our construction of the graph, Relation (5.3) holds.

Now, suppose that K is a set of independent nodes. This means, again by the construction of our graph, that

$$\rho(i,j) < -\gamma, \quad \text{for all } i \neq j \in K.$$

Let $Z_i, i \in K$ be zero-mean random variables such that $\rho(i,j) = \mathbb{E}[Z_i Z_j]$. Observe that

$$\operatorname{Var}\Bigg(\sum_{i\in K} Z_i\Bigg) = \sum_{i\in K} \operatorname{Var}(Z_i) + \sum_{\substack{i\neq j\\ i,j\in K}} \operatorname{Cov}(Z_i, Z_j) < k - k(k-1)\gamma, \tag{5.5}$$

since $\operatorname{Var}(Z_i) = 1$ and $\rho(i,j) < -\gamma$ for $i \neq j$. By our assumption, $k \geq (\lceil 1/\gamma \rceil + 1)$, or equivalently, $(k-1) \geq 1/\gamma$, the variance in (5.5) is negative. This is a contradiction showing that there are no independent sets K with cardinality k.

To complete the proof, it remains to show that Relation (5.4) holds. In view of the upper bound on the Ramsey numbers (5.1), it is enough to show that $k \leq \log_2(\sqrt{n})$ implies

$$\binom{2k-2}{k-1} \leq n.$$

This follows from a refinement of the Stirling formula, due to Robbins [36]:

$$\sqrt{2\pi} m^{m+1/2} e^{-m} e^{\frac{1}{(12m+1)}} \leq m! \leq \sqrt{2\pi} m^{m+1/2} e^{-m} e^{\frac{1}{12m}}.$$

Indeed, letting $\widetilde{k} := k-1$, and applying the above upper and lower bounds to the terms $(2\widetilde{k})!$ and $\widetilde{k}!$, respectively, we obtain:

$$\binom{2k-2}{k-1} \equiv \frac{(2\widetilde{k})!}{(\widetilde{k}!)^2} \leq \frac{2^{2\widetilde{k}}}{\sqrt{\pi \widetilde{k}}} \exp\left\{ \frac{1}{24\widetilde{k}} - \frac{2}{12\widetilde{k}+1} \right\} < 2^{2k},$$

where the last two inequalities follow by simply dropping positive factors less than 1. Since $2k \leq \log_2(n)$, the above bound implies Relation (5.4) and the proof is complete. □

Using Proposition 5.1, we establish the key lemma used in the proof of Theorem 3.1.

Lemma 5.1. *Let $c \in (0,1)$, and $P = (\rho(i,j))_{(n+1)\times(n+1)}$ be a correlation matrix such that*

$$\rho(1,j) > c \quad \textit{for all } j = 1, \ldots, n+1. \tag{5.6}$$

If $n \geq 2^{2\lceil 2/c^2 \rceil + 4}$, then there is a set of indices $K = \{l_1, \ldots, l_k\} \subseteq \{2, \ldots, n+1\}$ of cardinality $k = |K| = \lfloor \log_2 \sqrt{n} \rfloor$, such that

$$\rho(i,j) > \frac{c^2}{2} \quad \textit{for all } i, j \in K. \tag{5.7}$$

That is, all entries of the $k \times k$ sub-correlation matrix $P_K := (\rho(i,j))_{i,j\in K}$ are larger than $c^2/2$.

Proof of Lemma 5.1. Let $Z_1, \dots, Z_{n+1}$ be random variables with covariance matrix P. Denote $\rho_j = \rho(1, j)$ and define

$$R(j) = \begin{cases} \dfrac{1}{\sqrt{1-\rho_j^2}}\big(Z(j) - \rho_j Z(1)\big), & \text{if } \rho_j < 1, \\ R^*, & \text{if } \rho_j = 1, \end{cases} \tag{5.8}$$

where R^* is an arbitrary zero-mean, unit-variance random variable. It is easy to see that $\operatorname{Var}(R(j)) = 1$, and

$$\begin{aligned} \operatorname{Cov}\big(Z(i), Z(j)\big) &= \operatorname{Cov}\Big(\rho_i Z(1) + \sqrt{1-\rho_i^2} R(i), \rho_j Z(1) + \sqrt{1-\rho_j^2} R(j)\Big) \\ &= \rho_i \rho_j + \sqrt{1-\rho_i^2}\sqrt{1-\rho_j^2} \operatorname{Cov}\big(R(i), R(j)\big) \\ &\geq c^2 + \min\big\{\operatorname{Cov}\big(R(i), R(j)\big), 0\big\}. \end{aligned}$$

Therefore, Relation (5.7) would hold if we can find a set of indices $K = \{l_1, \dots, l_k\}$ such that $\operatorname{Cov}(R(i), R(j)) > -c^2/2$ for all $i, j \in K$, where $k = |K| = \lfloor \log_2 \sqrt{n} \rfloor$. This, however, follows from Proposition 5.1 applied to $(R(j))_{j=2}^{n+1}$ with $\gamma = c^2/2$, provided that

$$k = \lfloor \log_2 \sqrt{n} \rfloor \geq \lceil 2/c^2 \rceil + 1.$$

The last inequality indeed follows form the assumption that $n \geq 2^{2\lceil 2/c^2 \rceil + 4}$. □

5.2. URS implies UDD ('only if' part of Theorem 3.1)

In view of Remark 3.1, UDD is equivalent to the requirement that $N(\delta) := 1 + \sup_p N_p(\delta) < \infty$ for all $\delta \in (0, 1)$, where

$$N_p(\delta) := \max_{j \in \{1, \dots, p\}} \big|\big\{i : i \neq j, \Sigma_p(j, i) > \delta\big\}\big|. \tag{5.9}$$

Therefore, if $\mathcal{E}$ is not UDD, then there must exist a constant $c \in (0, 1)$ for which $N(c)$ is infinite, that is, there is a subsequence $\widetilde{p} \to \infty$ such that $N_{\widetilde{p}}(c) \to \infty$. Without loss of generality, we may assume that $\widetilde{p} = p$.

Let $j_p(c)$ be the maximizers of (5.9), and let

$$S_p(c) := \big\{i \in \{1, \dots, p\} : \Sigma_p\big(j_p(c), i\big) > c\big\}. \tag{5.10}$$

Observe that $|S_p(c)| = N_p(c) + 1 \to \infty$, as $p \to \infty$ (note $j_p(c) \in S_p(c)$).

Applying Lemma 5.1 to the set of random variables indexed by $S_p(c)$, we conclude, for $N_p(c) \geq 2^{2\lceil 2/c^2 \rceil + 4}$, there must be a further subset

$$K_p(c) \subseteq S_p(c), \tag{5.11}$$

of cardinality

$$k_p(c) := \left|K_p(c)\right| \geq \log_2 \sqrt{N_p(c)}, \tag{5.12}$$

such that all pairwise correlations of the random variables indexed by $K_p(c)$ are greater than $c^2/2$. Since the sequence $N_p(c) \to \infty$, by (5.12), we have $k_p(c) \to \infty$ as $p \to \infty$.

Therefore, we have identified a sequence of subsets $K_p(c) \subseteq \{1, \dots, p\}$ with the following two properties:

1. $k_p(c) := |K_p(c)| \to \infty$, as $p \to \infty$, and
2. For all $i, j \in K_p(c)$, we have

$$\Sigma_p(i, j) > c^2/2. \tag{5.13}$$

Without loss of generality, we may assume $K_p(c) = \{1, \dots, k_p(c)\} \subseteq \{1, \dots, p\}$, upon relabeling of the coordinates.

Now consider a Gaussian sequence $\varepsilon^* = \{\varepsilon^*(j), j = 1, 2, \dots\}$, independent of $\mathcal{E}$, defined as follows:

$$\varepsilon^*(j) := Z(c/\sqrt{2}) + Z(j)\sqrt{1 - c^2/2}, \quad j = 1, 2, \dots,$$

where Z and $Z(j)$, $j = 1, 2, \dots$ are independent standard normal random variables. Hence,

$$\operatorname{Var}\big(\varepsilon^*(j)\big) = 1 = \operatorname{Var}\big(\varepsilon_p(j)\big), \tag{5.14}$$

and

$$\operatorname{Cov}\big(\varepsilon^*(i), \varepsilon^*(j)\big) = \frac{c^2}{2} \leq \operatorname{Cov}\big(\varepsilon_p(i), \varepsilon_p(j)\big), \tag{5.15}$$

for all p, and all $i \neq j$, $i, j \in K_p(c)$. Thus we have, as $p \to \infty$,

$$\frac{1}{u_{k_p(c)}} \max_{j \in K_p(c)} \varepsilon^*(j) = \frac{c/\sqrt{2}}{u_{k_p(c)}} Z + \frac{\sqrt{1 - c^2/2}}{u_{k_p(c)}} \max_{j \in K_p(c)} Z(j) \xrightarrow{\mathbb{P}} \sqrt{1 - \frac{c^2}{2}}, \tag{5.16}$$

where the convergence in probability follows from Proposition 2.1 part 2.

Relations (5.14) and (5.15), by Slepian's lemma [38], also imply,

$$\frac{1}{u_{k_p(c)}} \max_{j \in K_p(c)} \varepsilon^*(j) \overset{d}{\geq} \frac{1}{u_{k_p(c)}} \max_{j \in K_p(c)} \varepsilon_p(j). \tag{5.17}$$

Therefore, by (5.17) and (5.16), for all $\sqrt{1 - c^2/2} \leq \delta < 1$, we have,

$$\mathbb{P}\left[\frac{1}{u_{k_p(c)}} \max_{j \in K_p(c)} \varepsilon_p(j) < \delta\right] \to 1 \quad \text{as } p \to \infty.$$

This contradicts the definition of URS (with the particular choice of $S_p := K_p(c)$), and the proof of the 'only if' part is complete.

Acknowledgements

We thank the Associate Editor and two anonymous referees for their constructive feedback, which helped improve the manuscript, and inspired the study on minimax optimality in Section 4.

This work was partially supported by the NSF ATD grant DMS-1830293.

Supplementary Material

Supplement to "Fundamental limits of exact support recovery in high dimensions" (DOI: 10.3150/20-BEJ1197SUPP; .pdf). Supplementary material [18] contains extensive numerical simulations of the main results (Section A), additional proofs (Section B), auxiliary results (Section C), generalizations of the phase transition phenomena to other classes of error distributions (Sections D), analysis of thresholding procedures under heavy, regularly varying tails (Section E), and further discussions on the statistical implications (Section F).

References

[1] Arias-Castro, E. and Chen, S. (2017). Distribution-free multiple testing. *Electron. J. Stat.* **11** 1983–2001. MR3651021 https://doi.org/10.1214/17-EJS1277

[2] Arias-Castro, E. and Ying, A. (2019). Detection of sparse mixtures: Higher criticism and scan statistic. *Electron. J. Stat.* **13** 208–230. MR3899951 https://doi.org/10.1214/18-ejs1512

[3] Barndorff-Nielsen, O. (1963). On the limit behaviour of extreme order statistics. *Ann. Math. Stat.* **34** 992–1002. MR0150889 https://doi.org/10.1214/aoms/1177704022

[4] Berman, S.M. (1964). Limit theorems for the maximum term in stationary sequences. *Ann. Math. Stat.* **35** 502–516. MR0161365 https://doi.org/10.1214/aoms/1177703551

[5] Bush, W.S. and Moore, J.H. (2012). Genome-wide association studies. *PLoS Comput. Biol.* **8** e1002822. https://doi.org/10.1371/journal.pcbi.1002822

[6] Butucea, C., Ndaoud, M., Stepanova, N.A. and Tsybakov, A.B. (2018). Variable selection with Hamming loss. *Ann. Statist.* **46** 1837–1875. MR3845003 https://doi.org/10.1214/17-AOS1572

[7] Cai, T.T., Jin, J. and Low, M.G. (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.* **35** 2421–2449. MR2382653 https://doi.org/10.1214/009053607000000334

[8] Comminges, L. and Dalalyan, A.S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Statist.* **40** 2667–2696. MR3097616 https://doi.org/10.1214/12-AOS1046

[9] Conlon, D., Fox, J. and Sudakov, B. (2015). Recent developments in graph Ramsey theory. In *Surveys in Combinatorics* 2015. *London Mathematical Society Lecture Note Series* **424** 49–118. Cambridge: Cambridge Univ. Press. MR3497267

[10] de Haan, L. and Ferreira, A. (2007). *Extreme Value Theory: An Introduction. Springer Series in Operations Research and Financial Engineering*. New York: Springer. MR2234156 https://doi.org/10.1007/0-387-34471-3

[11] Donoghue, W.F. Jr. (2014). *Distributions and Fourier Transforms. Pure and Applied Mathematics* **32**. New York: Academic Press. MR3363413

[12] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195 https://doi.org/10.1214/009053604000000265

[13] Dudoit, S., Shaffer, J.P. and Boldrick, J.C. (2003). Multiple hypothesis testing in microarray experiments. *Statist*. *Sci*. **18** 71–103. MR1997066 https://doi.org/10.1214/ss/1056397487

[14] Embrechts, P., Klüppelberg, C. and Mikosch, T. (2013). *Modelling Extremal Events*: *For Insurance and Finance*. *Applications of Mathematics* (*New York*) **33**. Berlin: Springer. MR1458613 https://doi.org/10.1007/978-3-642-33483-2

[15] Erdös, P. and Szekeres, G. (1935). A combinatorial problem in geometry. *Compos*. *Math*. **2** 463–470. MR1556929

[16] Fox, J. (2009). Lecture 5: Ramsey theory. In *MAT* 307: *Combinatorics* (*Spring* 2009), *MIT Lecture Notes*. Available at http://math.mit.edu/~Fox/MAT307.html.

[17] Gao, Z. (2019). Five shades of grey: Phase transitions in high-dimensional multiple testing. Preprint. Available at arXiv:1910.05701.

[18] Gao, Z. and Stoev, S. (2020). Supplement to "Fundamental limits of exact support recovery in high dimensions." https://doi.org/10.3150/20-BEJ1197SUPP

[19] Genovese, C.R., Jin, J., Wasserman, L. and Yao, Z. (2012). A comparison of the lasso and marginal regression. *J*. *Mach*. *Learn*. *Res*. **13** 2107–2143. MR2956354

[20] Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Ann*. *of Math*. (2) **44** 423–453. MR0008655 https://doi.org/10.2307/1968974

[21] Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann*. *Statist*. **38** 1686–1732. MR2662357 https://doi.org/10.1214/09-AOS764

[22] Haupt, J., Castro, R.M. and Nowak, R. (2011). Distilled sensing: Adaptive sampling for sparse detection and estimation. *IEEE Trans*. *Inf*. *Theory* **57** 6222–6235. MR2857969 https://doi.org/10.1109/TIT.2011.2162269

[23] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75** 800–802. MR0995126 https://doi.org/10.1093/biomet/75.4.800

[24] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand*. *J*. *Stat*. **6** 65–70. MR0538597

[25] Ingster, Y.I. (1998). Minimax detection of a signal for l^n-balls. *Math*. *Methods Statist*. **7** 401–428. MR1680087

[26] Ji, P. and Jin, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *Ann*. *Statist*. **40** 73–103. MR3013180 https://doi.org/10.1214/11-AOS947

[27] Kallitsis, M., Stoev, S.A., Bhattacharya, S. and Michailidis, G. (2016). AMON: An open source architecture for online monitoring, statistical analysis, and forensics of multi-gigabit streams. *IEEE J*. *Sel*. *Areas Commun*. **34** 1834–1848.

[28] Kartsioukas, R., Gao, Z. and Stoev, S. (2019). On the rate of concentration of maxima in Gaussian arrays. Preprint. Available at arXiv:1910.04259.

[29] Leadbetter, M.R., Lindgren, G. and Rootzén, H. (2012). *Extremes and Related Properties of Random Sequences and Processes*. New York: Springer.

[30] McCormick, W. and Mittal, Y. (1976). *On Weak Convergence of the Maximum*. Stanford Univ. Department of Statistics.

[31] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann*. *Statist*. **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

[32] Nichols, T. and Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: A comparative review. *Stat*. *Methods Med*. *Res*. **12** 419–446. MR2005445 https://doi.org/10.1191/0962280203sm341ra

[33] Ramsey, F.P. (2009). On a problem of formal logic. In *Classic Papers in Combinatorics* 1–24. New York: Springer.

[34] Resnick, S.I. (2013). *Extreme Values*, *Regular Variation and Point Processes*. *Springer Series in Operations Research and Financial Engineering*. New York: Springer. MR2364939

[35] Resnick, S.I. and Tomkins, R.J. (1973). Almost sure stability of maxima. *J. Appl. Probab.* **10** 387–401. MR0350828 https://doi.org/10.2307/3212355

[36] Robbins, H. (1955). A remark on Stirling's formula. *Amer. Math. Monthly* **62** 26–29. MR0069328 https://doi.org/10.2307/2308012

[37] Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J. Amer. Statist. Assoc.* **62** 626–633. MR0216666

[38] Slepian, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell Syst. Tech. J.* **41** 463–501. MR0133183 https://doi.org/10.1002/j.1538-7305.1962.tb02419.x

[39] Tanguy, K. (2015). Some superconcentration inequalities for extrema of stationary Gaussian processes. *Statist. Probab. Lett.* **106** 239–246. MR3389997 https://doi.org/10.1016/j.spl.2015.07.028

[40] Wainwright, M.J. (2009). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inf. Theory* **55** 5728–5741. MR2597190 https://doi.org/10.1109/TIT.2009.2032816

[41] Wainwright, M.J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory* **55** 2183–2202. MR2729873 https://doi.org/10.1109/TIT.2009.2016018

[42] Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689 https://doi.org/10.1214/08-AOS646

[43] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

Received April 2019 and revised September 2019