The Annals of Statistics
2010, Vol. 38, No. 5, 3063–3092
DOI: 10.1214/10-AOS817

ON COMBINATORIAL TESTING PROBLEMS[1]

BY LOUIGI ADDARIO-BERRY[2], NICOLAS BROUTIN, LUC DEVROYE[3] AND GÁBOR LUGOSI[4]

McGill University, INRIA, McGill University and ICREA and Pompeu Fabra University

We study a class of hypothesis testing problems in which, upon observing the realization of an n-dimensional Gaussian vector, one has to decide whether the vector was drawn from a standard normal distribution or, alternatively, whether there is a subset of the components belonging to a certain given class of sets whose elements have been "contaminated," that is, have a mean different from zero. We establish some general conditions under which testing is possible and others under which testing is hopeless with a small risk. The combinatorial and geometric structure of the class of sets is shown to play a crucial role. The bounds are illustrated on various examples.

1. Introduction. In this paper, we study the following hypothesis testing problem introduced by Arias-Castro et al. (2008). One observes an n-dimensional vector $\mathbf{X} = (X_1, \dots, X_n)$. The null hypothesis H_0 is that the components of $\mathbf{X}$ are independent and identically distributed (i.i.d.) standard normal random variables. We denote the probability measure and expectation under H_0 by $\mathbb{P}_0$ and $\mathbb{E}_0$, respectively.

To describe the alternative hypothesis H_1, consider a class $\mathcal{C} = \{S_1, \dots, S_N\}$ of N sets of indices such that $S_k \subset \{1, \dots, n\}$ for all $k = 1, \dots, N$. Under H_1, there exists an $S \in \mathcal{C}$ such that

$$X_i \text{ has distribution } \begin{cases} \mathcal{N}(0,1), & \text{if } i \notin S, \\ \mathcal{N}(\mu,1), & \text{if } i \in S, \end{cases}$$

where $\mu > 0$ is a positive parameter. The components of $\mathbf{X}$ are independent under H_1 as well. The probability measure of $\mathbf{X}$ defined this way by an $S \in \mathcal{C}$ is denoted by $\mathbb{P}_S$. Similarly, we write $\mathbb{E}_S$ for the expectation with respect to $\mathbb{P}_S$. Throughout, we will assume that every $S \in \mathcal{C}$ has the same cardinality $|S| = K$.

Received August 2009; revised January 2010.

[1]Parts of this work were done at the Bellairs Research Institute of McGill University and the Saturna West Island Mathematics Center.

[2]Supported by an NSERC Discovery grant.

[3]Supported by NSERC.

[4]Supported by the Spanish Ministry of Science and Technology Grant MTM2009-09063 and PASCAL2 Network of Excellence under EC Grant 216886.

AMS 2000 subject classifications. Primary 62F03; secondary 62F05.

Key words and phrases. Hypothesis testing, multiple hypotheses, Gaussian processes.

A test is a binary-valued function $f:\mathbb{R}^n \to \{0,1\}$. If $f(\mathbf{X})=0$ then we say that the test accepts the null hypothesis, otherwise H_0 is rejected. One would like to design tests such that H_0 is accepted with a large probability when $\mathbf{X}$ is distributed according to $\mathbb{P}_0$ and it is rejected when the distribution of $\mathbf{X}$ is $\mathbb{P}_S$ for some $S \in \mathcal{C}$. Following Arias-Castro et al. (2008), we consider the risk of a test f measured by

$$R(f) = \mathbb{P}_0\{f(\mathbf{X})=1\} + \frac{1}{N}\sum_{S\in\mathcal{C}} \mathbb{P}_S\{f(\mathbf{X})=0\}. \tag{1.1}$$

This measure of risk corresponds to the view that, under the alternative hypothesis, a set $S \subset \mathcal{C}$ is selected uniformly at random and the components of $\mathbf{X}$ belonging to S have mean μ. In the sequel, we refer to the first and second terms on the right-hand side of (1.1) as the type I and type II errors, respectively.

We are interested in determining, or at least estimating the value of μ under which the risk can be made small. Our aim is to understand the order of magnitude, when n is large, as a function of n, K, and the structure of $\mathcal{C}$, of the value of the smallest μ for which risk can be made small. The value of μ for which the risk of the best possible test equals $1/2$ is called *critical*.

Typically, the n components of $\mathbf{X}$ represent weights over the n edges of a given graph G and each $S \in \mathcal{C}$ is a subgraph of G. When $X_i \sim \mathcal{N}(\mu,1)$ then the edge i is "contaminated" and we wish to test whether there is a subgraph in $\mathcal{C}$ that is entirely contaminated.

In Arias-Castro et al. (2008), two examples were studied in detail. In one case, $\mathcal{C}$ contains all paths between two given vertices in a two-dimensional grid and in the other $\mathcal{C}$ is the set of paths from root to a leaf in a complete binary tree. In both cases, the order of magnitude of the critical value of μ was determined. Arias-Castro, Candès and Durand (2009) investigate another class of examples in which elements of $\mathcal{C}$ correspond to clusters in a regular grid. Both Arias-Castro et al. (2008) and Arias-Castro, Candès and Durand (2009) describe numerous practical applications of problems of this type.

Some other interesting examples are when $\mathcal{C}$ is:

- the set of all subsets $S \subset \{1,\dots,n\}$ of size K;
- the set of all cliques of a given size in a complete graph;
- the set of all bicliques (i.e., complete bipartite subgraphs) of a given size in a complete bipartite graph;
- the set of all spanning trees of a complete graph;
- the set of all perfect matchings in a complete bipartite graph;
- the set of all sub-cubes of a given size of a binary hypercube.

The first of these examples, which lacks any combinatorial structure, has been studied in the rich literature on multiple testing; see, for example, Ingster (1999), Baraud (2002), Donoho and Jin (2004) and the references therein.

As pointed out in Arias-Castro et al. (2008), regardless of what $\mathcal{C}$ is, one may determine explicitly the test f^* minimizing the risk. It follows from basic results of binary classification that for a given vector $\mathbf{x} = (x_1, \dots, x_n)$, $f^*(\mathbf{x}) = 1$, if and only if the ratio of the likelihoods of $\mathbf{x}$ under $(1/N)\sum_{S\in\mathcal{C}} \mathbb{P}_S$ and $\mathbb{P}_0$ exceeds 1. Writing

$$\phi_0(\mathbf{x}) = (2\pi)^{-n/2} e^{-\sum_{i=1}^n x_i^2/2}$$

and

$$\phi_S(\mathbf{x}) = (2\pi)^{-n/2} e^{-\sum_{i\in S}(x_i-\mu)^2/2 - \sum_{i\notin S} x_i^2/2}$$

for the probability densities of $\mathbb{P}_0$ and $\mathbb{P}_S$, respectively, the likelihood ratio at $\mathbf{x}$ is

$$L(\mathbf{x}) = \frac{1/N \sum_{S\in\mathcal{C}} \phi_S(\mathbf{x})}{\phi_0(\mathbf{x})} = \frac{1}{N}\sum_{S\in\mathcal{C}} e^{\mu x_S - K\mu^2/2},$$

where $x_S = \sum_{i\in S} x_i$. Thus, the optimal test is given by

$$f^*(\mathbf{x}) = \mathbb{1}_{\{L(\mathbf{x})>1\}} = \begin{cases} 0, & \text{if } \dfrac{1}{N}\displaystyle\sum_{S\in\mathcal{C}} e^{\mu x_S - K\mu^2/2} \le 1, \\ 1, & \text{otherwise.} \end{cases}$$

The risk of f^* (often called the Bayes risk) may then be written as

$$\begin{aligned} R^* = R^*_{\mathcal{C}}(\mu) = R(f^*) &= 1 - \frac{1}{2}\mathbb{E}_0|L(\mathbf{X}) - 1| \\ &= 1 - \frac{1}{2}\int \left|\phi_0(\mathbf{x}) - \frac{1}{N}\sum_{S\in\mathcal{C}} \phi_S(\mathbf{x})\right| d\mathbf{x}. \end{aligned}$$

We are interested in the behavior of R^* as a function of $\mathcal{C}$ and μ. Clearly, R^* is a monotone decreasing function of μ. (This fact is intuitively clear and can be proved easily by differentiating R^* with respect to μ.) For μ sufficiently large, R^* is close to zero while for very small values of μ, R^* is near its maximum value 1, indicating that testing is virtually impossible. Our aim is to understand for what values of μ the transition occurs. This depends on the combinatorial and geometric structure of the class $\mathcal{C}$. We describe various general conditions in both directions and illustrate them on examples.

REMARK (An alternative risk measure). Arias-Castro et al. (2008) also consider the risk measure

$$\overline{R}(f) = \mathbb{P}_0\{f(\mathbf{X}) = 1\} + \max_{S\in\mathcal{C}} \mathbb{P}_S\{f(\mathbf{X}) = 0\}.$$

Clearly, $\overline{R}(f) \ge R(f)$ and when there is sufficient symmetry in f and $\mathcal{C}$, we have equality. However, there are significant differences between the two measures of risk. The alternative measure $\overline{R}$ obviously satisfies the following monotonicity property: for a class $\mathcal{C}$ and parameter $\mu > 0$, let $\overline{R}^*_{\mathcal{C}}(\mu)$ denote the smallest achievable risk. If $\mathcal{A} \subset \mathcal{C}$ are two classes then for any μ, $\overline{R}^*_{\mathcal{A}}(\mu) \le \overline{R}^*_{\mathcal{C}}(\mu)$. In contrast to this, the "Bayesian" risk measure $R(f)$ does not satisfy such a monotonicity property as is shown in Section 5. In this paper, we focus on the risk measure $R(f)$.

REMARK. Throughout the paper we assume, for simplicity, that each set $S \in \mathcal{C}$ has the same cardinality K. We do this partly in order to avoid technicalities that are not difficult but make the arguments less transparent. At the same time, in many natural examples this condition is satisfied. If $\mathcal{C}$ may contain sets of different size such that all sets have approximately the same number of elements, then all arguments go through without essential changes. However, if $\mathcal{C}$ contains sets of very different size then the picture may change because large sets become much easier to detect and small sets can basically be ignored. Another approach to handle sets of different size, adopted by Arias-Castro, Candès and Durand (2009), is to change the model of the alternative hypothesis such that the level μ of contamination is appropriately scaled depending on the size of the set S.

Plan of the paper. The paper is organized as follows. In Section 2, we briefly discuss two suboptimal but simple and general testing rules (the *maximum test* and the *averaging test*) that imply sufficient conditions for testability that turn out to be useful in many examples.

In Section 3, a few general sufficient conditions are derived for the impossibility of testing under symmetry assumptions for the class.

In Section 4, we work out several concrete examples, including the class of all K-sets, the class of all cliques of a certain size in a complete graph, the class of all perfect matchings in the complete bipartite graph and the class of all spanning trees in a complete graph.

In Section 5, we show that, perhaps surprisingly, the optimal risk is not monotone in the sense that larger classes may be significantly easier to test than small ones, though monotonicity holds under certain symmetry conditions.

In the last two sections of the paper, we use techniques developed in the theory of Gaussian processes to establish upper and lower bounds related to geometrical properties of the class $\mathcal{C}$. In Section 6, general lower bounds are derived in terms of random subclasses and metric entropies of the class $\mathcal{C}$. Finally, in Section 7 we take a closer look at the type I error of the optimal test and prove an upper bound that, in certain situations, is significantly tighter than the natural bound obtained for a general-purpose maximum test.

2. Simple tests and upper bounds. As mentioned in the Introduction, the test f^* minimizing the risk is explicitly determined. However, the performance of this test is not always easy to analyze. Moreover, efficient computation of the optimal test is often a nontrivial problem though efficient algorithms are available in many interesting cases. (We discuss computational issues for the examples of Section 4.) Because of these reasons, it is often useful to consider simpler, though suboptimal, tests. In this section, we briefly discuss two simplistic tests, a test based on averaging and a test based on maxima. These are often easier to analyze and help understand the behavior of the optimal test as well. In many cases, one of these tests turn out to have a near-optimal performance.

A simple test based on averaging. Perhaps the simplest possible test is based on the fact that the sum of the components of $\mathbf{X}$ is zero-mean normal under $\mathbb{P}_0$ and has mean μK under the alternative hypothesis. Thus, it is natural to consider the *averaging test*

$$f(\mathbf{x}) = \mathbb{1}_{\{\sum_{i=1}^n X_i > \mu K/2\}}.$$

PROPOSITION 2.1. *Let $\delta > 0$. The risk of the averaging test f satisfies $R(f) \le \delta$ whenever*

$$\mu \ge \sqrt{\frac{8n}{K^2} \log \frac{2}{\delta}}.$$

PROOF. Observe that under $\mathbb{P}_0$, the statistic $\sum_{i=1}^n X_i$ has normal $\mathcal{N}(0, n)$ distribution while for each $S \in \mathcal{C}$, under $\mathbb{P}_S$, it is distributed as $\mathcal{N}(\mu K, n)$. Thus, $R(f) \le 2e^{-(\mu K)^2/(8n)}$. $\square$

A test based on maxima. Another natural test is based on the fact that under the alternative hypothesis for some $S \in \mathcal{C}$, $X_S = \sum_{i \in S} X_i$ is normal $(\mu K, K)$. Consider the *maximum test*

$$f(\mathbf{x}) = 1 \quad \text{if and only if} \quad \max_{S \in \mathcal{C}} X_S \ge \frac{\mu K + \mathbb{E}_0 \max_{S \in \mathcal{C}} X_S}{2}.$$

The test statistic $\max_{S \in \mathcal{C}} X_S$ is often referred to as a *scan statistic* and has been thoroughly studied for a wide range of applications; see Glaz, Naus and Wallenstein (2001). Here, we only need the following simple observation.

PROPOSITION 2.2. *The risk of the maximum test f satisfies $R(f) \le \delta$ whenever*

$$\mu \ge \frac{\mathbb{E}_0 \max_{S \in \mathcal{C}} X_S}{K} + 2\sqrt{\frac{2}{K} \log \frac{2}{\delta}}.$$

In the analysis, it is convenient to use the following simple Gaussian concentration inequality; see Tsirelson, Ibragimov and Sudakov (1976).

LEMMA 2.1 (Tsirelson's inequality). *Let $X = (X_1, \dots, X_n)$ be an vector of n independent standard normal random variables. Let $f : \mathbb{R}^n \to \mathbb{R}$ denote a Lipschitz function with Lipschitz constant L (with respect to the Euclidean distance). Then for all $t > 0$,*

$$\mathbb{P}\{f(X) - \mathbb{E} f(X) \ge t\} \le e^{-t^2/(2L^2)}.$$

PROOF OF PROPOSITION 2.2. Simply note that under the null hypothesis, for each $S \in \mathcal{C}$, X_S is a zero-mean normally distributed random variable with variance $K = |S|$. Since $\max_{S\in\mathcal{C}} X_S$ is a Lipschitz function of $\mathbf{X}$ with Lipschitz constant $\sqrt{K}$, by Tsirelson's inequality, for all $t > 0$,

$$\mathbb{P}_0\Big\{\max_{S\in\mathcal{C}} X_S \geq \mathbb{E}_0 \max_{S\in\mathcal{C}} X_S + t\Big\} \leq e^{-t^2/(2K)}.$$

On the other hand, under $\mathbb{P}_S$ for a fixed $S \in \mathcal{C}$,

$$\max_{S'\in\mathcal{C}} X_{S'} \geq X_S \sim \mathcal{N}(\mu K, K)$$

and therefore

$$\mathbb{P}_S\Big\{\max_{S\in\mathcal{C}} X_S \leq \mu K - t\Big\} \leq e^{-t^2/(2K)},$$

which completes the proof. $\square$

The maximum test is often easier to compute than the optimal test f^*, though maximization is not always possible in polynomial time. If the value of $\mathbb{E}_0 \max_{S\in\mathcal{C}} X_S$ is not exactly known, one may replace it in the definition of f by any upper bound and then the same upper bound will appear in the performance bound.

Proposition 2.2 shows that the maximum test is guaranteed to work whenever μ is at least $\mathbb{E}_0 \max_{S\in\mathcal{C}} X_S/K + \text{const.}/\sqrt{K}$. Thus, in order to better understand the behavior of the maximum test (and thus obtain sufficient conditions for the optimal test to have a low risk), one needs to understand the expected value of $\max_{S\in\mathcal{C}} X_S$ (under $\mathbb{P}_0$). As the maximum of Gaussian processes have been studied extensively, there are plenty of directly applicable results available for expected maxima. The textbook of Talagrand (2005) is dedicated to this topic. Here, we only recall some of the basic facts.

First, note that one always has $\mathbb{E}_0 \max_{S\in\mathcal{C}} X_S \leq \sqrt{2K \log N}$ but sharper bounds can be derived by chaining arguments; see Talagrand (2005) for an elegant and advanced treatment. The classical chaining bound of Dudley (1978) works as follows. Introduce a metric on $\mathcal{C}$ by

$$d(S,T) = \sqrt{\mathbb{E}_0(X_S - X_T)^2} = \sqrt{d_H(S,T)}, \qquad S, T \in \mathcal{C},$$

where $d_H(S,T) = \sum_{i=1}^n \mathbb{1}_{\{\mathbb{1}_{\{i\in S\}} \neq \mathbb{1}_{\{i\in T\}}\}}$ denotes the Hamming distance. For $t > 0$, let $N(t)$ denote the t-covering number of $\mathcal{C}$ with respect to the metric d, that is, the smallest number of open balls of radius t that cover $\mathcal{C}$. By Dudley's theorem, there exists a numerical constant C such that

$$\mathbb{E}_0 \max_{S\in\mathcal{C}} X_S \leq C \int_0^{\mathrm{diam}(\mathcal{C})} \sqrt{\log N(t)}\, dt,$$

where $\mathrm{diam}(\mathcal{C}) = \max_{S,T\in\mathcal{C}} d(S,T)$ denotes the diameter of the metric space $\mathcal{C}$. Note that since $|S| = K$ for all $S \in \mathcal{C}$, $\mathrm{diam}(\mathcal{C}) \leq \sqrt{2K}$. Dudley's theorem is not optimal but it is relatively easy to use. Dudley's theorem has been refined, based on "majorizing measures," or "generic chaining" which gives sharp bounds; see, for example, Talagrand (2005).

REMARK (The VC dimension). In certain cases, it is convenient to further bound Dudley's inequality in terms of the VC dimension; see Vapnik and Chervonenkis (1971). Recall that the VC dimension $V(\mathcal{C})$ of $\mathcal{C}$ is the largest positive integer m such that there exists an m-element set $\{i_1, \dots, i_m\} \subset \{1, \dots, n\}$ such that for all 2^m subsets $A \subset \{i_1, \dots, i_m\}$ there exists an $S \in \mathcal{C}$ such that $S \cap \{i_1, \dots, i_m\} = A$. Haussler (1995) proved that the covering numbers of $\mathcal{C}$ may be bounded as

$$N(t) \le e \cdot \big(V(\mathcal{C}) + 1\big)\left(\frac{2en}{t^2}\right)^{V(\mathcal{C})},$$

so by Dudley's bound,

$$\mathbb{E}_0 \max_{S \in \mathcal{C}} X_S \le C\sqrt{V(\mathcal{C}) K \log n}.$$

REMARK (Tests based on symmetrization). An interesting alternative to the maximum test, proposed and investigated by Durot and Rozenholc (2006) and Arlot, Blanchard and Roquain (2010a), is based on the idea that under the null hypothesis the distribution of the vector $\mathbf{X}$ does not change if the sign of each component is changed randomly, while under the alternative hypothesis the distribution changes. In Durot and Rozenholc (2006) and Arlot, Blanchard and Roquain (2010a), methods based on symmetrization and bootstrap are suggested and analyzed. Such tests are meaningful and interesting in the setup of the present paper as well and it would be interesting to analyze their behavior.

3. Lower bounds.

In this section, we investigate conditions under which the risk of any test is large. We start with a simple universal bound that implies that regardless of what the class $\mathcal{C}$ is, small risk cannot be achieved unless μ is substantially large compared to $K^{-1/2}$.

A universal lower bound. An often convenient way of bounding the Bayes risk R^* is in terms of the Bhattacharyya measure of affinity [Bhattacharyya (1946)]

$$\rho = \rho_{\mathcal{C}}(\mu) = \tfrac{1}{2}\mathbb{E}_0\sqrt{L(\mathbf{X})}.$$

It is well known [see, e.g., Devroye, Györfi and Lugosi (1996), Theorem 3.1] that

$$1 - \sqrt{1 - 4\rho^2} \le R^* \le 2\rho.$$

Thus, 2ρ essentially behaves as the Bayes error in the sense that R^* is near 1 when 2ρ is near 1, and is small when 2ρ is small. Observe that, by Jensen's inequality,

$$2\rho = \mathbb{E}_0\sqrt{L(\mathbf{X})} = \int \sqrt{\frac{1}{N}\sum_{S \in \mathcal{C}} \phi_S(\mathbf{x})\phi_0(\mathbf{x})}\, d\mathbf{x} \ge \frac{1}{N}\sum_{S \in \mathcal{C}} \int \sqrt{\phi_S(\mathbf{x})\phi_0(\mathbf{x})}\, d\mathbf{x}.$$

Straightforward calculation shows that for any $S \in \mathcal{C}$,

$$\int \sqrt{\phi_S(\mathbf{x})\phi_0(\mathbf{x})}\, d\mathbf{x} = e^{-\mu^2 K/8}$$

and therefore we have the following.

PROPOSITION 3.1. *For all classes $\mathcal{C}$, $R^* \geq 1/2$ whenever $\mu \leq \sqrt{(4/K)} \times \sqrt{\log(4/3)}$.*

This shows that no matter what the class $\mathcal{C}$ is, detection is hopeless if μ is of the order of $K^{-1/2}$. This classical fact goes back to Le Cam (1970).

A lower bound based on overlapping pairs. The next lemma is due to Arias-Castro et al. (2008). For completeness, we recall their proof.

PROPOSITION 3.2. *Let S and S' be drawn independently, uniformly, at random from $\mathcal{C}$ and let $Z = |S \cap S'|$. Then*

$$R^* \geq 1 - \tfrac{1}{2}\sqrt{\mathbb{E}e^{\mu^2 Z} - 1}.$$

PROOF. As noted above, by the Cauchy–Schwarz inequality,

$$R^* = 1 - \tfrac{1}{2}\mathbb{E}_0|L(\mathbf{X}) - 1| \geq 1 - \tfrac{1}{2}\sqrt{\mathbb{E}_0|L(\mathbf{X}) - 1|^2}.$$

Since $\mathbb{E}_0 L(\mathbf{X}) = 1$,

$$\mathbb{E}_0|L(\mathbf{X}) - 1|^2 = \mathrm{Var}_0(L(\mathbf{X})) = \mathbb{E}_0[L(\mathbf{X})^2] - 1.$$

However, by definition $L(\mathbf{X}) = \frac{1}{N}\sum_{S\in\mathcal{C}} e^{\mu X_S - K\mu^2/2}$, so we have

$$\mathbb{E}_0[L(\mathbf{X})^2] = \frac{1}{N^2}\sum_{S,S'\in\mathcal{C}} e^{-K\mu^2}\mathbb{E}_0 e^{\mu(X_S + X_{S'})}.$$

But

$$\begin{aligned}
\mathbb{E}_0 e^{\mu(X_S + X_{S'})} &= \mathbb{E}_0[e^{\mu\sum_{i\in S\setminus S'} X_i} e^{\mu\sum_{i\in S'\setminus S} X_i} e^{2\mu\sum_{i\in S\cap S'} X_i}] \\
&= (\mathbb{E}_0 e^{\mu X})^{2(K - |S\cap S'|)}(\mathbb{E}_0 e^{2\mu X})^{|S\cap S'|} \\
&= e^{\mu^2(K - |S\cap S'|) + 2\mu^2|S\cap S'|},
\end{aligned}$$

and the statement follows. $\square$

The beauty of this proposition is that it reduces the problem to studying a purely combinatorial quantity. By deriving upper bounds for the moment generating function of the overlap $|S \cap S'|$ between two elements of $\mathcal{C}$ drawn independently and uniformly at random, one obtains lower bounds for the critical value of μ. This simple proposition turns out to be surprisingly powerful as it will be illustrated in various applications below.

A lower bound for symmetric classes. We begin by deriving some simple consequences of Proposition 3.2 under some general symmetry conditions on the class $\mathcal{C}$. The following proposition shows that the universal bound of Proposition 3.1 can be improved by a factor of $\sqrt{\log(1+n/K)}$ for all sufficiently symmetric classes.

PROPOSITION 3.3. *Let $\delta \in (0,1)$. Assume that $\mathcal{C}$ satisfies the following conditions of symmetry. Let S, S' be drawn independently and uniformly at random from $\mathcal{C}$. Assume that*: (i) *the conditional distribution of $Z = |S \cap S'|$ given S' is identical for all values of S'*; (ii) *for any fixed $S_0 \in \mathcal{C}$ and $i \in S_0$, $\mathbb{P}\{i \in S\} = K/n$. Then $R^* \geq \delta$ for all μ with*

$$\mu \leq \sqrt{\frac{1}{K}\log\left(1+\frac{4n(1-\delta)^2}{K}\right)}.$$

PROOF. We apply Proposition 3.2. By the first symmetry assumption, it suffices to derive a suitable upper bound for $\mathbb{E}[e^{\mu^2 Z}] = \mathbb{E}[e^{\mu^2 Z}|S']$ for an arbitrary $S' \in \mathcal{C}$. After a possible relabeling, we may assume that $S' = \{1,\dots,K\}$ so we can write $Z = \sum_{i=1}^K \mathbb{1}_{\{i\in S\}}$. By Hölder's inequality,

$$\begin{aligned}
\mathbb{E}[e^{\mu^2 Z}] &= \mathbb{E}\left[\prod_{i=1}^K e^{\mu^2 \mathbb{1}_{\{i\in S\}}}\right] \\
&\leq \prod_{i=1}^K (\mathbb{E}[e^{K\mu^2 \mathbb{1}_{\{i\in S\}}}])^{1/K} \\
&= \mathbb{E}[e^{K\mu^2 \mathbb{1}_{\{1\in S\}}}] \qquad \text{[by assumption (ii)]} \\
&= (e^{\mu^2 K} - 1)\frac{K}{n} + 1.
\end{aligned}$$

Proposition 3.2 now implies the statement. □

Surprisingly, the lower bound of Proposition 3.3 is close to optimal in many cases. This is true, in particular when the class $\mathcal{C}$ is "small," made precise in the following statement.

COROLLARY 3.1. *Assume that $\mathcal{C}$ is symmetric in the sense of Proposition 3.3 and that it contains at most n^α elements where $\alpha > 0$. Then $R^* \geq 1/2$ for all μ with*

$$\mu \leq \sqrt{\frac{1}{K}\log\left(1+\frac{n}{K}\right)}$$

and $R^* \le 1/2$ *for all* μ *with*

$$\mu \ge \sqrt{\frac{2\alpha}{K}\log n} + \sqrt{\frac{8\log 4}{K}}.$$

PROOF. The first statement follows from Proposition 3.3 while the second from Proposition 2.2 and the fact that $\mathbb{E}_0 \max_{S\in\mathcal{C}} X_S \le \sqrt{2K\log|\mathcal{C}|}$. □

The proposition above shows that for any small and sufficiently symmetric class, the critical value of μ is of the order of $\sqrt{(\log n)/K}$, at least if $K \le n^\beta$ for some $\beta \in (0,1)$. Later, we will see examples of "large" classes for which Proposition 3.3 also gives a bound of the correct order of magnitude.

Negative association. The bound of Proposition 3.3 may be improved significantly under an additional condition of negative association that is satisfied in several interesting examples (see Section 4 below). Recall that a collection $Y_1,\ldots,Y_n$ of random variables is *negatively associated* if for any pair of disjoint sets $I, J \subset \{1,\ldots,n\}$ and (coordinate-wise) nondecreasing functions f and g,

$$\mathbb{E}[f(Y_i, i\in I)g(Y_j, j\in J)] \le \mathbb{E}[f(Y_i, i\in I)]\mathbb{E}[g(Y_j, j\in J)].$$

PROPOSITION 3.4. *Let* $\delta \in (0,1)$ *and assume that the class* $\mathcal{C}$ *satisfies the conditions of Proposition* 3.3. *Suppose that the labels are such that* $S' = \{1,2,\ldots,K\} \in \mathcal{C}$. *Let* S *be a randomly chosen element of* $\mathcal{C}$. *If the random variables* $\mathbb{1}_{\{1\in S\}},\ldots,\mathbb{1}_{\{K\in S\}}$ *are negatively associated, then* $R^* \ge \delta$ *for all* μ *with*

$$\mu \le \sqrt{\log\left(1 + \frac{n\log(1+4(1-\delta)^2)}{K^2}\right)}.$$

PROOF. We proceed similarly to the proof of Proposition 3.3. We have

$$\begin{aligned}\mathbb{E}[e^{\mu^2 Z}] &= \mathbb{E}\left[\prod_{i=1}^K e^{\mu^2 \mathbb{1}_{\{i\in S\}}}\right]\\ &\le \prod_{i=1}^K \mathbb{E}[e^{\mu^2\mathbb{1}_{\{i\in S\}}}] \qquad \text{(by negative association)}\\ &= \left((e^{\mu^2}-1)\frac{K}{n}+1\right)^K.\end{aligned}$$

Proposition 3.2 and the upper bound above imply that R^* at least δ for all μ such that

$$\mu \le \sqrt{\log\left(1 + \frac{n((1+4(1-\delta)^2)^{1/K}-1)}{K}\right)}.$$

The result follows by using $e^y \ge 1+y$ with $y = K^{-1}\log(1+4(1-\delta)^2)$. □

4. Examples. In this section, we consider various concrete examples and work out upper and lower bounds for the critical range of μ.

4.1. *Disjoint sets.* We start with the simplest possible case, that is, when all $S \in \mathcal{C}$ are disjoint (and therefore $KN \le n$). Fix $\delta \in (0, 1)$. Then, under $\mathbb{P}_0$, the X_S are independent normal $(0, K)$ random variables and the bound $\mathbb{E}_0 \max_{S\in\mathcal{C}} X_S \le \sqrt{2K \log N}$ is close to being tight. By applying the maximum test f, we see that $R^* \le R(f) \le \delta$ whenever

$$\mu \ge \sqrt{\frac{2\log N}{K}} + 2\sqrt{\frac{2\log(2/\delta)}{K}}.$$

To see that this bound gives the correct order of magnitude, we may simply apply Proposition 3.2. Here Z may take two values:

$$Z = K \qquad \text{with probability } 1/N \quad \text{and} \quad Z = 0 \qquad \text{with probability } 1 - 1/N.$$

Thus,

$$\mathbb{E}e^{\mu^2 Z} - 1 = \frac{1}{N}(e^{\mu^2 K} - 1) \le \frac{1}{N}e^{\mu^2 K}$$

and therefore $R^* \ge \delta$ whenever

$$\mu \le \sqrt{\frac{\log(4N(1-\delta)^2)}{K}}.$$

So in this case the critical transition occurs when μ is of the order of $\sqrt{(1/K)\log N}$. In Section 6, we use this simple lower bound to establish lower bounds for general classes $\mathcal{C}$ of sets. Note that in this simple case one may directly analyze the risk of the optimal test and obtain sharper bounds. In particular, the leading constant in the lower bound is suboptimal. However, in this paper our aim is to understand some general phenomena and we focus on orders of magnitude rather than on nailing down sharp constants.

REMARK (Multiple hypothesis testing). Taking $S = \{\{1\}, \ldots, \{n\}\}$, $K = 1$, and $N = n \ge 2$ in the above example, we obtain a connection with multiple hypothesis testing. In the latter, one tests "$m_i = 0$" against "$m_i = \mu$" for every $1 \le i \le n$, and traditionally uses as test statistics X_i, $1 \le i \le n$, to build a multiple testing procedure (often rejecting all the hypotheses corresponding to large X_i), see, for instance, Romano and Wolf (2005), Arlot, Blanchard and Roquain (2010b). Such a procedure will reject the global null hypothesis "$\forall i \in S, m_i = 0$" if at least one of the alternatives "$m_i = \mu$" is preferred. The main difference with the approach taken in this paper concerns the error rate. A multiple testing procedure is generally calibrated to control measures of the type I error like the family wise error rate or the false discovery rate, while the tests defined in this paper are designed to control the entire risk. Finally, in this example, the subsets are disjoint which is the traditional framework in multiple testing.

4.2. *K-sets.* Consider the example when $\mathcal{C}$ contains all sets $S \subset \{1, \ldots, n\}$ of size K. Thus, $N = \binom{n}{K}$. As mentioned in the Introduction, this problem is very well understood as sharp bounds and sophisticated tests are available; see, for example, Ingster (1999), Baraud (2002), Donoho and Jin (2004). We include it for illustration purposes only and we warn the reader that the obtained bounds are not sharpest possible.

Let $\delta \in (0, 1)$. It is easy to see that the assumptions of Proposition 3.4 are satisfied [this follows, e.g., from Proposition 11 of Dubdashi and Ranjan (1998)] and therefore $R^* \geq \delta$ for all

$$\mu \leq \sqrt{\log\left(1 + \frac{n \log(1 + 4(1-\delta)^2)}{K^2}\right)}.$$

This simple bound turns out to have the correct order of magnitude both when $n \gg K^2$ [in which case it is of the order of $\sqrt{\log(n/K^2)}$] and when $n \ll K^2$ (when it is of the order of $\sqrt{n/K^2}$).

This may be seen by considering the two simple tests described in Section 2 in the two different regimes. Since

$$\frac{\mathbb{E}_0 \max_{S \in \mathcal{C}} X_S}{K} \leq \frac{\sqrt{2K \log \binom{n}{K}}}{K} \leq \sqrt{2 \log\left(\frac{ne}{K}\right)},$$

we see from Proposition 2.2 that when $K = O(n^{(1-\varepsilon)/2})$ for some fixed $\varepsilon > 0$, then the threshold value is of the order of $\sqrt{\log n}$. On the other hand, when K^2/n is bounded away from zero, then the lower bound implied by Proposition 3.4 above is of the order $\sqrt{n/K^2}$ and the averaging test provides a matching upper bound by Proposition 2.1.

Note that in this example the maximum test is easy to compute since it suffices to find the K largest values among $X_1, \ldots, X_n$.

4.3. *Perfect matchings.* Let $\mathcal{C}$ be the set of all perfect matchings of the complete bipartite graph $K_{m,m}$. Thus, we have $n = m^2$ edges and $N = m!$, and $K = m$. By Proposition 2.1 (i.e., the averaging test), for $\delta \in (0, 1)$, one has $R(f) \leq \delta$ whenever $\mu \geq \sqrt{8 \log(2/\delta)}$.

To show that this bound has the right order of magnitude, we may apply Proposition 3.4. The symmetry assumptions hold obviously and the negative association property follows from the fact that $Z = |S \cap S'|$ has the same distribution as the number of fixed points in a random permutation. The proposition implies that for all m, $R^* \geq \delta$ whenever

$$\mu \leq \sqrt{\log(1 + \log(1 + 4(1-\delta)^2))}.$$

Note that in this case the optimal test f^* can be approximated in a computationally efficient way. To this end, observe that computing

$$\frac{1}{N}\sum_{S\in\mathcal{C}} e^{\mu X_S} = \frac{1}{m!}\sum_{\sigma}\prod_{j=1}^{m} e^{\mu X_{(j,\sigma(j))}}$$

(where the summation is over all permutations of $\{1,\dots,m\}$) is equivalent to computing the permanent of an $m \times m$ matrix with nonnegative elements. By a deep result of Jerrum, Sinclair and Vigoda (2004), this may be done by a polynomial-time randomized approximation.

4.4. *Stars*. Consider a network of m nodes in which each pair of nodes interacts. One wishes to test if there is a corrupted node in the network whose interactions slightly differ from the rest. This situation may be modeled by considering the class of *stars*.

A star is a subgraph of the complete graph K_m which contains all $K = m - 1$ edges containing a fixed vertex (see Figure 1). Consider the set $\mathcal{C}$ of all stars. In this setting, $n = \binom{m}{2}$ and $N = m$.

In this case, we are in the situation of Corollary 3.1 and Propositions 3.3 and 2.2 imply that if $\mathcal{C}$ is the class of all stars in K_m then for any $\varepsilon > 0$,

$$\lim_{m\to\infty} R^* = \begin{cases} 0, & \text{if } \mu \ge (\sqrt{2}+\varepsilon)\sqrt{\dfrac{\log m}{m}}, \\ 1, & \text{if } \mu \le (1-\varepsilon)\sqrt{\dfrac{\log m}{m}}. \end{cases}$$

4.5. *Spanning trees*. Consider again a network of m nodes in which each pair of nodes interact. One may wish to test if there exists a corrupted connected subgraph containing each node. This leads us to considering the class of all spanning trees as follows.

Let $1, 2, \dots, n = \binom{m}{2}$ represent the edges of the complete graph K_m and let $\mathcal{C}$ be the set of all spanning trees of K_m. Thus, we have $N = m^{m-2}$ spanning trees and $K = m - 1$. [See, e.g., Moon (1970).] By Proposition 2.1, the averaging test has risk $R(f) \le \delta$ whenever $\mu \ge \sqrt{4\log(2/\delta)}$.

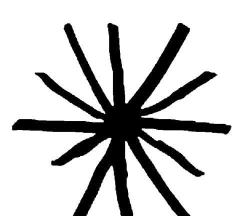

FIG. 1. *A star* [*Vonnegut* (1973)].

This bound is indeed of the right order. To see this, we may start with Proposition 3.2. There are (at least) two ways of proceeding. One is based on negative association. Even though Proposition 3.4 is not applicable because of the lack of symmetry in $\mathcal{C}$, negative association still holds. In particular, by a result of Feder and Mihail (1992) [see also Grimmett and Winkler (2004) and Benjamini et al. (2001)], if S is a random uniform spanning tree of K_m, then the indicators $\mathbb{1}_{\{1\in S\}}, \ldots, \mathbb{1}_{\{n\in S\}}$ are negatively associated. This means that, if S and S' are independent uniform spanning trees and $Z = |S \cap S'|$,

$$\begin{aligned}
\mathbb{E}[e^{\mu^2 Z}] &= \mathbb{E}\mathbb{E}[e^{\mu^2 |S\cap S'|}|S'] \\
&= \mathbb{E}\mathbb{E}[e^{\mu^2 \sum_{i\in S'} \mathbb{1}_{\{i\in S\}}}|S'] \\
&\le \mathbb{E}\prod_{i\in S'} \mathbb{E}[e^{\mu^2 \mathbb{1}_{\{i\in S\}}}|S'] \qquad \text{(by negative association)} \\
&\le \mathbb{E}\prod_{i\in S'} \left(\frac{2}{m} e^{\mu^2} + 1\right) \\
&= \left(\frac{2}{m} e^{\mu^2} + 1\right)^{m-1} \\
&\le \exp(2e^{\mu^2}).
\end{aligned}$$

This, together with Proposition 3.2 shows that for any $\delta \in (0, 1)$, $R^* \ge \delta$ whenever

$$\mu \le \sqrt{\log(1 + \tfrac{1}{2}\log(1 + 4(1-\delta)^2))}.$$

We note here that the same bound can be proved by a completely different way that does not use negative association. The key is to note that we may generate the two random spanning trees based on $2(m-1)$ independent random variables $X_1, \ldots, X_{2(m-1)}$ taking values in $\{1, \ldots, m-1\}$ as in Aldous (1990) [see also Broder (1989)]. The key property we need is that if Z_i denotes the number of common edges in the two spanning trees when X_i is replaced by an independent copy X_i' while keeping all other X_j's fixed, then

$$\sum_{i=1}^{2(m-1)} (Z - Z_i)_+ \le Z$$

(the details are omitted). For random variables satisfying this last property, an inequality of Boucheron, Lugosi and Massart (2000) implies the sub-Poissonian bound

$$\mathbb{E}\exp(\mu^2 Z) \le \exp(\mathbb{E}Z(e^{\mu^2} - 1)).$$

Clearly, $\mathbb{E}Z = 2(n-1)/n \le 2$, so essentially the same bound as above is obtained.

As the bounds above show, the computationally trivial average test has a close-to-optimal performance. In spite of this, one may wish to use the optimal test f^*. The "partition function" $(1/N)\sum_{S\in\mathcal{C}} e^{\mu X_S}$ may be computed by an algorithm of Propp and Wilson (1998), who introduced a random sampling algorithm that, given a graph with nonnegative weights w_i over the edges, samples a random spanning tree from a distribution such that the probability of any spanning tree S is proportional to $\prod_{i\in S} w_i$. The expected running time of the algorithm is bounded by the cover time of an associated Markov chain that is defined as a random walk over the graph in which the transition probabilities are proportional to the edge weights. If μ is of the order of a constant (as in the critical range) then the cover time is easily shown to be polynomial (with high probability) as all edge weights $w_i = e^{\mu^2 X_i}$ are roughly of the same order both under the null and under the alternative hypotheses.

4.6. *Cliques.* Another natural application is the class of all cliques of a certain size in a complete graph. More precisely, the random variables $X_1, \dots, X_n$ are associated with the edges of the complete graph K_m such that $\binom{m}{2} = n$ and let $\mathcal{C}$ contain all cliques of size k. Thus, $K = \binom{k}{2}$ and $N = \binom{m}{k}$. This case is more difficult than the class of K-sets discussed above because negative association does not hold anymore. (This may be easily seen by considering the indicator variables of two adjacent edges both being in the randomly chosen clique.) Also, computationally the class of cliques is much more complex. A related, well-studied model starts with the subgraph K_m containing each edge independently with probability $1/2$, as null hypothesis. The alternative hypothesis is the same as the null hypothesis, except that there is a clique of size k on which each edge is independently present with probability $p > 1/2$. This is called the "hidden clique" problem (usually only the special case $p = 1$ is considered). Despite substantial interest in the hidden clique problem, polynomial time detection algorithms are only known when $k = \Omega(\sqrt{n})$ [Alon, Krivelevich and Sudakov (1999), Feige and Krauthgamer (2000)]. We may obtain the hidden clique model from our model by thresholding at weight zero (retaining only edges whose normal random variable is positive), and so our model is easier for testing than the hidden clique model. However, it seems likely that designing an efficient test in the normal setting will be as difficult as it has proved for hidden cliques. It would be of interest to construct near-optimal tests that are computable in polynomial time for larger values of k.

We have the following bounds for the performance of the optimal test. It shows that when k is a most of the order of $\sqrt{m}$, the critical value of μ is of the order of $\sqrt{(1/k)\log(m/k)}$. The proof below may be adjusted to handle larger values of k as well but we prefer to keep the calculations more transparent.

PROPOSITION 4.1. *Let $\mathcal{C}$ represent the class of all $N = \binom{m}{k}$ cliques of a complete graph K_m and assume that $k \le \sqrt{m(\log 2)/e}$. Then*:

(i) *for all* $\delta \in (0,1)$, $R^* \le \delta$ *whenever*

$$\mu \ge 2\sqrt{\frac{1}{k-1}\log\left(\frac{me}{k}\right)} + 4\sqrt{\frac{\log(2/\delta)}{k(k-1)}};$$

(ii) $R^* \ge 1/2$ *whenever*

$$\mu \le \sqrt{\frac{1}{k}\log\left(\frac{m}{2k}\right)}.$$

PROOF. (i) follows simply by a straightforward application of Proposition 2.2 and the bound $\mathbb{E}_0 \max_{S\in\mathcal{C}} X_S \le \sqrt{2K\log N}$.

To prove the lower bound (ii), by Proposition 3.2, it suffices to show that if S, S' are k-cliques drawn randomly and independently from $\mathcal{C}$ and Z denotes the number of *edges* in the intersection of S and S', then $\mathbb{E}[\exp(\mu^2 Z)] \le 2$ for the indicated values of μ.

Because of symmetry, $\mathbb{E}[\exp(\mu^2 Z)] = \mathbb{E}[\exp(\mu^2 Z)|S']$ for all S' and therefore we might as well fix an arbitrary clique S'. If Y denotes the number of vertices in the clique $S \cap S'$, then $Z = \binom{Y}{2}$. Moreover, the distribution of Y is hypergeometrical with parameters m and k. If B is a binomial random variable with parameters k and k/m, then since $\exp(\mu^2 x^2/2)$ is a convex function of x, an inequality of Hoeffding (1963) implies that

$$\mathbb{E}[e^{\mu^2 Z}] = \mathbb{E}[e^{\mu^2 Y^2/2}] \le \mathbb{E}[e^{\mu^2 B^2/2}].$$

Thus, it remains to derive an appropriate upper bound for the moment generating function of the squared binomial. To this end, let $c > 1$ be a parameter whose value will be specified later. Using

$$B^2 \le B\left(k\mathbb{1}_{\{B>ck^2/m\}} + c\frac{k^2}{m}\right)$$

and the Cauchy–Schwarz inequality, it suffices to show that

$$\mathbb{E}\left[\exp\left(\mu^2 c\frac{k^2}{m}B\right)\right] \cdot \mathbb{E}\big[\exp\big(\mu^2 kB\mathbb{1}_{\{B>ck^2/m\}}\big)\big] \le 4. \tag{4.1}$$

We show that, if μ satisfies the condition of (ii), for an appropriate choice of c, both terms on the left-hand side are at most 2.

The first term on the left-hand side of (4.1) is

$$\mathbb{E}\left[\exp\left(\mu^2 c\frac{k^2}{m}B\right)\right] = \left(1 + \frac{k}{m}\left(\exp\left(\mu^2 c\frac{k^2}{m}\right) - 1\right)\right)^k,$$

which is at most 2 if and only if

$$\frac{k}{m}\left(\exp\left(\mu^2 c\frac{k^2}{m}\right) - 1\right) \le 2^{1/k} - 1.$$

Since $2^{1/k} - 1 \geq (\log 2)/k$, this is implied by

$$\mu \leq \sqrt{\frac{m}{ck^2} \log\left(1 + \frac{m \log 2}{k^2}\right)}.$$

To bound the second term on the left-hand side of (4.1), note that

$$\begin{aligned}\mathbb{E}[\exp(\mu^2 kB \mathbb{1}_{\{B > ck^2/m\}})] &\leq 1 + \mathbb{E}[\mathbb{1}_{\{B > ck^2/m\}} \exp(\mu^2 kB)] \\ &\leq 1 + \left(\mathbb{P}\left\{B > c\frac{k^2}{m}\right\}\right)^{1/2} (\mathbb{E}[\exp(\mu^2 kB)])^{1/2},\end{aligned}$$

by the Cauchy–Schwarz inequality, so it suffices to show that

$$\mathbb{P}\left\{B > c\frac{k^2}{m}\right\} \cdot \mathbb{E}[\exp(\mu^2 kB)] \leq 1.$$

Denoting $h(x) = (1 + x)\log(1 + x) - x$, Chernoff's bound implies

$$\mathbb{P}\left\{B > c\frac{k^2}{m}\right\} \leq \exp\left(-\frac{k^2}{m} h(c - 1)\right).$$

On the other hand,

$$\mathbb{E}[\exp(\mu^2 kB)] = \left(1 + \frac{k}{m}\exp(\mu^2 k)\right)^k,$$

and therefore the second term on the left-hand side of (4.1) is at most 2 whenever

$$1 + \frac{k}{m}\exp(\mu^2 k) \leq \exp\left(\frac{k}{m} h(c - 1)\right).$$

Using $\exp(\frac{k}{m} h(c - 1)) \geq 1 + \frac{k}{m} h(c - 1)$, we obtain the sufficient condition

$$\mu \leq \sqrt{\frac{1}{k} \log h(c - 1)}.$$

Summarizing, we have shown that $R^* \geq 1/2$ for all μ satisfying

$$\mu \leq 2 \cdot \min\left(\sqrt{\frac{1}{k} \log h(c - 1)}, \sqrt{\frac{m}{ck^2} \log\left(1 + \frac{m \log 2}{k^2}\right)}\right).$$

Choosing

$$c = \frac{m}{k} \frac{\log(m/k)}{\log(m \log 2/k^2)}$$

[which is greater than 1 for $k \leq \sqrt{m(\log 2)/e}$], the second term on the right-hand side is at most $\sqrt{(1/k)\log(m/k)}$. Now observe that since $h(c-1) = c \log c - c + 1$

is convex, for any $a > 0$, $h(c-1) \geq c \log a - a + 1$. Choosing $a = \frac{\log(m/k)}{\log(m \log 2/k^2)}$, the first term is at least

$$\sqrt{\frac{1}{k}\log\left(\frac{m}{k} - \frac{\log(m/k)}{\log(m\log 2/k^2)}\right)} \geq \sqrt{\frac{1}{k}\log\left(\frac{m}{2k}\right)},$$

where we used the condition that $m \log 2/k^2 \geq e$ and that $x \geq 2 \log x$ for all $x > 0$. □

REMARK (A related problem). A closely related problem arising in the exploratory analysis of microarray data [see Shabalin et al. (2009)] is when each member of $\mathcal{C}$ represents the K edges of a $\sqrt{K} \times \sqrt{K}$ biclique of the complete bipartite graph $K_{m,m}$ where $m = \sqrt{n}$. (A biclique is a complete bipartite subgraph of $K_{m,m}$.) The analysis and the bounds are completely analogous to the one worked out above, the details are omitted.

5. On the monotonicity of the risk. Intuitively, one would expect that the testing problem becomes harder as the class $\mathcal{C}$ gets larger. More precisely, one may expect that if $\mathcal{A} \subset \mathcal{C}$ are two classes of subsets of $\{1,\dots,n\}$, then $R^*_{\mathcal{A}}(\mu) \leq R^*_{\mathcal{C}}(\mu)$ holds for all μ. The purpose of this section is to show that this intuition is wrong in quite a strong sense as not only such general monotonicity property does not hold for the risk, but there are classes $\mathcal{A} \subset \mathcal{C}$ for which $R^*_{\mathcal{A}}(\mu)$ is arbitrary close to 1 and $R^*_{\mathcal{C}}(\mu)$ is arbitrary close to 0 for the same value of μ.

However, monotonicity does hold if the class $\mathcal{C}$ is sufficiently symmetric. Call a class $\mathcal{C}$ *symmetric* if for the optimal test

$$f^*_{\mathcal{C}}(\mathbf{x}) = \mathbb{1}_{\{(1/N)\sum_{S\in\mathcal{C}} \exp(\mu \sum_{i\in S} x_i) \geq \exp(K\mu^2/2)\}},$$

the value of $\mathbb{P}_T\{f^*_{\mathcal{C}}(\mathbf{X}) = 0\}$ is the same for all $T \in \mathcal{C}$. Note that several of the examples discussed in Section 4 satisfy the symmetry assumption, such as the classes of K-sets, stars, perfect matchings, and cliques. However, the class of spanning trees is not symmetric in the required sense.

THEOREM 5.1. *Let $\mathcal{C}$ be a symmetric class of subsets of $\{1,\dots,n\}$. If $\mathcal{A}$ is an arbitrary subclass of $\mathcal{C}$, then for all $\mu > 0$, $R^*_{\mathcal{A}}(\mu) \leq R^*_{\mathcal{C}}(\mu)$.*

PROOF. In this proof, we fix the value of $\mu > 0$ and suppress it in the notation. Recall the definition of the alternative risk measure

$$\overline{R}_{\mathcal{C}}(f) = \mathbb{P}_0\{f(\mathbf{X}) = 1\} + \max_{S\in\mathcal{C}} \mathbb{P}_S\{f(\mathbf{X}) = 0\},$$

which is to be contrasted with our main risk measure

$$R_{\mathcal{C}}(f) = \mathbb{P}_0\{f(\mathbf{X}) = 1\} + \frac{1}{N}\sum_{S\in\mathcal{C}} \mathbb{P}_S\{f(\mathbf{X}) = 0\}.$$

The risk $\overline{R}$ is obviously monotone in the sense that if $\mathcal{A} \subset \mathcal{C}$ then for every f, $\overline{R}_{\mathcal{A}}(f) \le \overline{R}_{\mathcal{C}}(f)$. Let $\overline{f}^*_{\mathcal{C}}$ and $f^*_{\mathcal{C}}$ denote the optimal tests with respect to both measures of risk.

First, observe that if $\mathcal{C}$ is symmetric, then $\overline{R}_{\mathcal{C}}(f^*_{\mathcal{C}}) = R_{\mathcal{C}}(f^*_{\mathcal{C}})$. But since $R_{\mathcal{C}}(f) \le \overline{R}_{\mathcal{C}}(f)$ for every f, we have

$$\overline{R}_{\mathcal{C}}(\overline{f}^*_{\mathcal{C}}) \le \overline{R}_{\mathcal{C}}(f^*_{\mathcal{C}}) = R_{\mathcal{C}}(f^*_{\mathcal{C}}) \le R_{\mathcal{C}}(\overline{f}^*_{\mathcal{C}}) \le \overline{R}_{\mathcal{C}}(\overline{f}^*_{\mathcal{C}}).$$

This means that all inequalities are equalities and, in particular, $\overline{f}^*_{\mathcal{C}} = f^*_{\mathcal{C}}$.

Now if $\mathcal{A}$ is an arbitrary subclass of $\mathcal{C}$, then

$$R^*_{\mathcal{C}} = R_{\mathcal{C}}(f^*_{\mathcal{C}}) = \overline{R}_{\mathcal{C}}(\overline{f}^*_{\mathcal{C}}) \ge \overline{R}_{\mathcal{A}}(\overline{f}^*_{\mathcal{C}}) \ge R_{\mathcal{A}}(\overline{f}^*_{\mathcal{C}}) \ge R_{\mathcal{A}}(f^*_{\mathcal{A}}) = R^*_{\mathcal{A}},$$

which completes the proof. □

THEOREM 5.2. *For every* $\varepsilon \in (0, 1)$ *there exist* n, μ, *and classes* $\mathcal{A} \subset \mathcal{C} \subset \{1, \dots, n\}$ *such that* $R^*_{\mathcal{A}}(\mu) \ge 1 - \varepsilon$ *and* $R^*_{\mathcal{C}}(\mu) \le 2\varepsilon$.

PROOF. We work with L_1 distances. For any class $\mathcal{L}$, denote $\phi_{\mathcal{L}}(\mathbf{x}) = \frac{1}{N} \times \sum_{S \in \mathcal{L}} \phi_S(\mathbf{x})$. Recall that

$$R^*_{\mathcal{L}}(\mu) = 1 - \frac{1}{2} \int |\phi_0(\mathbf{x}) - \phi_{\mathcal{L}}(\mathbf{x})| \, d\mathbf{x}.$$

Given ε, we fix an integer $K = K(\varepsilon)$ large enough that $K + 1 \ge 1/\varepsilon$ and that

$$\sqrt{\frac{\log(4(K+1)\varepsilon^2)}{K+1}} \ge \sqrt{\frac{8}{K} \log\left(\frac{2}{\varepsilon}\right)},$$

and let $n = n(\varepsilon) = (K + 1)^2$. We let $\mathcal{A}$ consist of $K + 1$ disjoint subsets of $\{1, \dots, n\}$, each of size $K + 1$. We let $\mathcal{B}$ consist of all sets of the form $\{1, \dots, K, i\}$, where i ranges from $K + 1$ to n, and assume $\mathcal{A}$ has been chosen so that $\mathcal{A} \cap \mathcal{B} = \varnothing$. We then let $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$. We take

$$\mu = \sqrt{\frac{\log(4(K+1)\varepsilon^2)}{K+1}},$$

so that, as seen in Section 4.1, we have $R^*_{\mathcal{A}}(\mu) \ge 1 - \varepsilon$. We will require an upper bound on $R^*_{\mathcal{B}}(\mu)$, which we obtain by considering the averaging test on variables $1, \dots, K$,

$$f(\mathbf{x}) = \mathbb{1}_{\{\sum_{i=1}^K x_i \ge (\mu K)/2\}}.$$

Just as in Proposition 2.1, we have $R(f) \le \varepsilon$ whenever $\mu \ge \sqrt{\frac{8}{K} \log(\frac{2}{\varepsilon})}$, which is indeed the case by our choices of μ and K. It follows that $R^*_{\mathcal{B}}(\mu) \le \varepsilon$. We remark that

$$\int |\phi - \phi_{\mathcal{A}}| = 2 - 2R^*_{\mathcal{A}}(\mu) \le 2\varepsilon.$$

We let $M = |\mathcal{B}| = (K+1)^2 - K$; then $N = |\mathcal{C}| = M + K + 1 = (K+1)^2 + 1$, and note

$$\begin{aligned}\int |\phi - \phi_{\mathcal{C}}| &= \int \left|\phi - \frac{(K+1)\phi_{\mathcal{A}} + M\phi_{\mathcal{B}}}{N}\right| \\ &= \int \left|\frac{(K+1)(\phi - \phi_{\mathcal{A}}) + M(\phi - \phi_{\mathcal{B}})}{N}\right| \\ &\geq \frac{M}{N}\int |\phi - \phi_{\mathcal{B}}| - \frac{(K+1)}{N}\int |\phi - \phi_{\mathcal{A}}| \\ &\geq (1-\varepsilon)\int |\phi - \phi_{\mathcal{B}}| - 2\varepsilon^2 \\ &= (1-\varepsilon)\big(2 - 2R^*_{\mathcal{B}}(\mu)\big) - 2\varepsilon^2 \\ &\geq 2 - 4\varepsilon.\end{aligned}$$

Thus, $R^*_{\mathcal{C}}(\mu) \leq 2\varepsilon$. □

Observe that nonmonotonicity of the Bhattacharyya affinity also follows from the same argument. To this end, we may express $\rho_{\mathcal{C}}(\mu) = \frac{1}{2}\int \sqrt{\phi_0(\mathbf{x})\phi_S(\mathbf{x})}\,d\mathbf{x}$ in function of the Hellinger distance

$$H(\phi_0, \phi_{\mathcal{C}}) = \sqrt{\int \left(\sqrt{\phi_0(\mathbf{x})} - \sqrt{\phi_{\mathcal{C}}(\mathbf{x})}\right)^2 d\mathbf{x}}$$

as $\rho_{\mathcal{C}}(\mathbf{x}) = \frac{1}{2} - \frac{1}{4}H(\phi_0, \phi_{\mathcal{C}})^2$. Recalling [see, e.g., Devroye and Györfi (1985), page 225] that

$$H(\phi_0, \phi_{\mathcal{C}})^2 \leq \int |\phi_0(\mathbf{x}) - \phi_{\mathcal{C}}(\mathbf{x})| \leq 2H(\phi_0, \phi_{\mathcal{C}}),$$

we see that the same example as in the proof above, for n large enough, shows the nonmonotonicity of the Bhattacharyya affinity as well.

6. Lower bounds on based random subclasses and metric entropy. In this section, we derive lower bounds for the Bayes risk $R^* = R^*_{\mathcal{C}}(\mu)$. The bounds are in terms of some geometric features of the class $\mathcal{C}$. Again, we treat $\mathcal{C}$ as a metric space equipped with the canonical distance $d(S,T) = \sqrt{\mathbb{E}_0(X_S - X_T)^2}$ [i.e., the square root of the Hamming distance $d_H(S,T)$].

For an integer $M \leq N$, we define a real-valued parameter $t_{\mathcal{C}}(M) > 0$ of the class $\mathcal{C}$ as follows. Let $\mathcal{A} \subset \mathcal{C}$ be obtained by choosing M elements of $\mathcal{C}$ at random, without replacement. Let the random variable τ denote the smallest distance between elements of $\mathcal{A}$ and let $t_{\mathcal{C}}(M)$ be a median of τ.

THEOREM 6.1. *Let $M \leq N$ be an integer. Then for any class $\mathcal{C}$,*

$$R^*_{\mathcal{C}} \geq 1/4,$$

whenever

$$\mu \le \min\left(\sqrt{\frac{\log(M/16)}{K}}, \frac{8\log(\sqrt{3}/8)}{\sqrt{K - t_{\mathcal{C}}(M)^2/2}}\right).$$

To interpret the statement of the theorem, note that

$$K - \tau^2/2 = \max_{\substack{S,T \in \mathcal{A} \\ S \neq T}} |S \cap T|$$

is the largest overlap between any pair of elements of $\mathcal{A}$. Thus, just like in Proposition 3.2, the distribution of the overlap between random elements of $\mathcal{C}$ plays a key role in establishing lower bounds for the optimal risk. However, while in Proposition 3.2 the moment generating function $\mathbb{E}\exp(\mu^2|S \cap T|)$ of the overlap between two random elements determines an upper bound for the critical value of μ, here it is the median of the largest overlap between many random elements that counts. The latter seems to carry more information about the fine geometry of the class. In fact, invoking a simple union bound, upper bounds for $\mathbb{E}\exp(\mu^2|S \cap T|)$ may be used together with Theorem 6.1.

In applications, often it suffices to consider the following special case.

COROLLARY 6.1. *Let $M \le N$ be the largest integer for which zero is a median of* $\max_{S,T\in\mathcal{A},S\neq T}|S \cap T|$ *where $\mathcal{A}$ is a random subset of $\mathcal{C}$ of size M* [*i.e.*, $t_{\mathcal{C}}(M)^2 = 2K$]. *Then $R^*_{\mathcal{C}}(\mu) \ge 1/4$ for all $\mu \le \sqrt{\log(M/16)/K}$.*

EXAMPLE (Sub-squares of a grid). To illustrate the corollary, consider the following example which is the simplest in a family of problems investigated by Arias-Castro, Candès and Durand (2009): assume that n and K are both perfect squares and that the indices $\{1, \dots, n\}$ are arranged in a $\sqrt{n} \times \sqrt{n}$ grid. The class $\mathcal{C}$ contains all $\sqrt{K} \times \sqrt{K}$ sub-squares. Now if S and T are randomly chosen elements of $\mathcal{C}$ (with or without replacement) then, if $(K+1)^2 \le 2\sqrt{n}$,

$$\mathbb{P}\{|S \cap T| \neq 0\} \ge \frac{(\sqrt{n} - 2K)^2}{(\sqrt{n} - K + 1)^2} \cdot \frac{K}{(\sqrt{n} - K + 1)^2} \ge \frac{K}{n}$$

and therefore

$$\mathbb{P}\Big\{\max_{\substack{S,T \in \mathcal{A} \\ S \neq T}} |S \cap T| = 0\Big\} = 1 - \mathbb{P}\Big\{\max_{\substack{S,T \in \mathcal{A} \\ S \neq T}} |S \cap T| > 0\Big\} \ge 1 - M^2\frac{K}{n},$$

which is at least $1/2$ if $M \le \sqrt{n/(2K)}$ in which case $t_{\mathcal{C}}(M)^2 = 2K$. Thus, by Corollary 6.1, $R^*_{\mathcal{C}}(\mu) \ge 1/4$ for all $\mu \le \sqrt{\log(n/(512K))/(2K)}$. This bound is of the optimal order of magnitude as it is easily seen by an application of Proposition 2.2.

In some other applications, a better bound is obtained if some overlap is allowed. A case in point is the example of stars from Section 4.4. In that case, any two elements of $\mathcal{C}$ overlap but by taking $M = N(= m)$, we have $K - t_{\mathcal{C}}(M)^2/2 = 1$, so Theorem 6.1 still implies $R^*_{\mathcal{C}}(\mu) \ge 1/4$ whenever $\mu \le \sqrt{(1/K)\log(m/16)}$.

The main tool of the proof of Theorem 6.1 is Slepian's lemma which we recall here [Slepian (1962)]. [For this version, see Ledoux and Talagrand (1991), Theorem 3.11.]

LEMMA 6.1 (Slepian's lemma). *Let $\boldsymbol{\xi} = (\xi_1, \dots, \xi_N), \boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_N) \in \mathbb{R}^N$ be zero-mean Gaussian vectors such that for each $i, j = 1, \dots, N$,*

$$\mathbb{E}\xi_i^2 = \mathbb{E}\zeta_i^2 \qquad \textit{for each } i = 1, \dots, N \quad \textit{and} \quad \mathbb{E}\xi_i\xi_j \le \mathbb{E}\zeta_i\zeta_j \qquad \textit{for all } i \ne j.$$

Let $F : \mathbb{R}^N \to \mathbb{R}$ be such that for all $\mathbf{x} \in \mathbb{R}^N$ and $i \ne j$,

$$\frac{\partial^2 F}{\partial x_i\,\partial x_j}(\mathbf{x}) \le 0.$$

Then $\mathbb{E}F(\boldsymbol{\xi}) \ge \mathbb{E}F(\boldsymbol{\zeta})$.

PROOF OF THEOREM 6.1. Let $M \le N$ be fixed and choose M sets from $\mathcal{C}$ uniformly at random (without replacement). Let $\mathcal{A}$ denote the random subclass of $\mathcal{C}$ obtained this way. Denote the likelihood ratio associated to this class by

$$L_{\mathcal{A}}(\mathbf{X}) = \frac{1/M \sum_{S \in A} \phi_S(\mathbf{X})}{\phi_0(\mathbf{X})} = \frac{1}{M} \sum_{S \in \mathcal{A}} V_S,$$

where $V_S = e^{\mu X_S - K\mu^2/2}$. Then the optimal risk of the class $\mathcal{C}$ may be lower bounded by

$$R^*_{\mathcal{C}}(\mu) - R^*_{\mathcal{A}}(\mu) = \tfrac{1}{2}\big(\mathbb{E}_0|L_{\mathcal{A}}(\mathbf{X}) - 1| - \mathbb{E}_0|L_{\mathcal{C}}(\mathbf{X}) - 1|\big) \ge -\tfrac{1}{2}\mathbb{E}_0|L_{\mathcal{A}}(\mathbf{X}) - L_{\mathcal{C}}(\mathbf{X})|.$$

Denoting by $\widehat{\mathbb{E}}$ expectation with respect to the random choice of $\mathcal{A}$, we have

$$\begin{aligned}
R^*_{\mathcal{C}}(\mu) &\ge \widehat{\mathbb{E}} R^*_{\mathcal{A}}(\mu) - \frac{1}{2}\mathbb{E}_0\widehat{\mathbb{E}}\left|\frac{1}{M}\sum_{S\in\mathcal{A}} V_S - \frac{1}{N}\sum_{S\in\mathcal{C}} V_S\right| \\
&\ge \widehat{\mathbb{E}} R^*_{\mathcal{A}}(\mu) - \frac{1}{2}\sqrt{\mathbb{E}_0\widehat{\mathbb{E}}\left(\frac{1}{M}\sum_{S\in\mathcal{A}} V_S - \frac{1}{N}\sum_{S\in\mathcal{C}} V_S\right)^2} \\
&\ge \widehat{\mathbb{E}} R^*_{\mathcal{A}}(\mu) - \frac{1}{2}\sqrt{\mathbb{E}_0\left[\frac{1}{M}\cdot\frac{1}{N}\sum_{T\in\mathcal{C}}\left(V_T - \frac{1}{N}\sum_{S\in\mathcal{C}} V_S\right)^2\right]} \\
&\qquad \text{(since the variance of a sample without replacement} \\
&\qquad \text{is less than that with replacement)} \\
&= \widehat{\mathbb{E}} R^*_{\mathcal{A}}(\mu) - \frac{1}{2\sqrt{M}}\sqrt{\frac{1}{N}\sum_{T\in\mathcal{C}}\mathbb{E}_0\left(V_T - \frac{1}{N}\sum_{S\in\mathcal{C}} V_S\right)^2}.
\end{aligned}$$

An easy way to bound the right-hand side is by writing

$$\begin{aligned}
&\mathbb{E}_0\left(V_T - \frac{1}{N}\sum_{S\in\mathcal{C}} V_S\right)^2 \\
&\quad \le 2\mathbb{E}_0(V_T-1)^2 + 2\mathbb{E}_0\left(1 - \frac{1}{N}\sum_{S\in\mathcal{C}} V_S\right)^2 \\
&\quad \le 2\mathbb{E}_0(V_T-1)^2 + \frac{2}{N}\sum_{S\in\mathcal{C}} \mathbb{E}_0(1-V_S)^2 \\
&\quad = 4\,\mathrm{Var}(V_T) = 4(e^{\mu^2 K} - 1).
\end{aligned}$$

Summarizing, we have

$$R^*_{\mathcal{C}}(\mu) \ge \widehat{\mathbb{E}} R^*_{\mathcal{A}}(\mu) - \sqrt{\frac{e^{\mu^2 K} - 1}{M}} \ge \widehat{\mathbb{E}} R^*_{\mathcal{A}}(\mu) - \frac{1}{4},$$

where we used the assumption that $\mu \le \sqrt{(1/K)\log(M/16)}$. Thus, it suffices to prove that $\widehat{\mathbb{E}} R^*_{\mathcal{A}}(\mu) \ge 1/2$.

We bound the optimal risk associated with $\mathcal{A}$ in terms of the Bhattacharyya affinity

$$\rho_{\mathcal{A}}(\mu) = \frac{1}{2}\mathbb{E}_0\sqrt{\frac{(1/M)\sum_{S\in\mathcal{A}}\phi_S(\mathbf{X})}{\phi_0(\mathbf{X})}} = \frac{1}{2}\mathbb{E}_0\sqrt{\frac{1}{|\mathcal{A}|}\sum_{S\in\mathcal{A}} V_S}.$$

Recalling from Section 3 that $R^*_{\mathcal{A}}(\mu) \ge 1 - \sqrt{1 - 4\rho_{\mathcal{A}}(\mu)^2}$ and using that $\sqrt{1-4x^2}$ is concave, we have

$$\widehat{\mathbb{E}} R^*_{\mathcal{A}}(\mu) \ge 1 - \sqrt{1 - 4(\widehat{\mathbb{E}}\rho_{\mathcal{A}}(\mu))^2}.$$

Therefore, it suffices to show that the expected Bhattacharyya affinity $\widehat{\mathbb{E}}\rho_{\mathcal{A}}(\mu)$ corresponding to the random class $\mathcal{A}$ satisfies

$$\widehat{\mathbb{E}}\rho_{\mathcal{A}}(\mu) = \frac{1}{2}\widehat{\mathbb{E}}\mathbb{E}_0\sqrt{\frac{1}{|\mathcal{A}|}\sum_{S\in\mathcal{A}} V_S} \ge \frac{\sqrt{3}}{4}.$$

In the argument below, we fix the random class $\mathcal{A}$, relabel the elements so that $\mathcal{A} = \{1, 2, \dots, |\mathcal{A}|\}$, and bound $\rho_{\mathcal{A}}(\mu)$ from below. Denote the minimum distance between any two elements of $\mathcal{A}$ by τ. To bound $\rho_{\mathcal{A}}(\mu)$, we apply Slepian's lemma with the function

$$F(\mathbf{x}) = \sqrt{\frac{1}{|\mathcal{A}|}\sum_{i=1}^{|\mathcal{A}|} e^{\mu x_i - K\mu^2/2}},$$

where $\mathbf{x} = (x_1, \dots, x_{|\mathcal{A}|})$. Simple calculation shows that the mixed second partial derivatives of F are negative, so Slepian's lemma is indeed applicable.

Next, we introduce the random vectors $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$. Let the components of $\boldsymbol{\xi}$ be indexed by elements $S \in \mathcal{A}$ and define $\xi_S = X_S = \sum_{i \in S} X_i$. Thus, under $\mathbb{P}_0$, each ξ_S is normal $(0, K)$ and $\mathbb{E}F(\boldsymbol{\xi})$ is just the Bhattacharyya affinity $\rho_{\mathcal{A}}(\mu)$. To define the random vector $\boldsymbol{\zeta}$, introduce $M+1$ independent standard normal random variables: one variable G_S for each $S \in \mathcal{A}$ and an extra variable G_0. Recall that the definition of τ guarantees that the minimal distance between any two elements of $\mathcal{A}$ as at least τ. Now let

$$\zeta_S = G_S \frac{\tau}{\sqrt{2}} + G_0 \sqrt{K - \frac{\tau^2}{2}}.$$

Then clearly for each $S, T \in \mathcal{A}$, $\mathbb{E}\zeta_S^2 = K$ and $\mathbb{E}\zeta_S\zeta_T = K - \tau^2/2$ $(S \neq T)$. On the other hand, $\mathbb{E}\xi_S^2 = K$ and

$$\mathbb{E}\xi_S\xi_T = |S \cap T| = K - \frac{d(S,T)^2}{2} \le K - \frac{\tau^2}{2} = \mathbb{E}\zeta_S\zeta_T.$$

Therefore, by Slepian's lemma, $\rho_{\mathcal{A}}(\mu) = \mathbb{E}F(\boldsymbol{\xi}) \ge \mathbb{E}F(\boldsymbol{\zeta})$. However,

$$\begin{aligned}
\mathbb{E}F(\boldsymbol{\zeta}) &= \mathbb{E}\sqrt{\frac{1}{|\mathcal{A}|}\sum_{S\in\mathcal{A}} e^{\mu\zeta_S - K\mu^2/2}} \\
&= \mathbb{E}\sqrt{e^{\mu\sqrt{K-\tau^2/2}G_0 - (K-\tau^2/2)\mu^2/2}\frac{1}{|\mathcal{A}|}\sum_{S\in\mathcal{A}} e^{\mu\tau G_S/\sqrt{2} - \tau^2\mu^2/4}} \\
&= \mathbb{E}e^{\mu\sqrt{K-\tau^2/2}G_0/2 - (K-\tau^2/2)\mu^2/4}\,\mathbb{E}\sqrt{\frac{1}{|\mathcal{A}|}\sum_{S\in\mathcal{A}} e^{\mu\tau G_S/\sqrt{2} - \tau^2\mu^2/4}} \\
&= e^{-\mu^2(K-\tau^2/2)/8}\,\mathbb{E}\sqrt{\frac{1}{|\mathcal{A}|}\sum_{S\in\mathcal{A}} e^{\mu\tau G_S/\sqrt{2} - \tau^2\mu^2/4}}.
\end{aligned}$$

To finish the proof, it suffices to observe that the last expression is the Bhattacharyya affinity corresponding to a class of disjoint sets, all of size $\tau^2/2$, of cardinality $|\mathcal{A}| = M$. This case has been handled in the first example of Section 4 where we showed that

$$\mathbb{E}\sqrt{\frac{1}{|\mathcal{A}|}\sum_{S\in\mathcal{A}} e^{\mu\tau G_S/\sqrt{2} - \tau^2\mu^2/4}} \ge R^*_{\mathcal{A}} \ge 1 - \frac{1}{2}\sqrt{\frac{1}{M}e^{\mu^2\tau^2/2}} \ge \frac{3}{4},$$

where again we used the condition $\mu \le \sqrt{\log(M/16)/K}$ and the fact that $\tau^2/2 \le K$.

Therefore, under this condition on μ, we have that for any fixed $\mathcal{A}$,

$$\rho_{\mathcal{A}}(\mu) = \tfrac{1}{2}\mathbb{E}F(\boldsymbol{\zeta}) \ge \tfrac{3}{8}e^{-\mu^2(K-\tau^2/2)/8}$$

and therefore

$$\widehat{\mathbb{E}}\rho_{\mathcal{A}}(\mu) \ge \tfrac{3}{16}e^{-\mu^2(K-t_{\mathcal{C}}(M)^2/2)/8},$$

where $t_{\mathcal{C}}(M)$ is the median of τ. This concludes the proof. $\square$

REMARK (An improvement). At the risk of losing a constant factor in the statement of Theorem 6.1, one may replace the parameter $t_{\mathcal{C}}(M)$ by a larger quantity. The idea is that by thinning the random subclass $\mathcal{A}$ one may consider a subset of $\mathcal{A}$ that has better separation properties. More precisely, for an even integer $M \le N$ we may define a real-valued parameter $\bar{t}_{\mathcal{C}}(M) > 0$ of the class $\mathcal{C}$ as follows. Let $\mathcal{A} \subset \mathcal{C}$ be obtained by choosing M elements of $\mathcal{C}$ at random, without replacement. Order the elements $S_1, \ldots, S_M$ of $\mathcal{A}$ such that

$$\min_{i\neq 1} d(S_1, S_i) \ge \min_{i\neq 2} d(S_2, S_i) \ge \cdots \ge \min_{i\neq M} d(S_M, S_i)$$

and define the subset $\widehat{\mathcal{A}} \subset \mathcal{A}$ by $\widehat{\mathcal{A}} = \{A_1, \ldots, A_{M/2}\}$. Let the random variable $\bar{\tau}$ denote the smallest distance between elements of $\widehat{\mathcal{A}}$ and let $\bar{t}_{\mathcal{C}}(M)$ be the median of τ. It is easy to see that the proof of Theorem 6.1 goes through, and one may replace $t_{\mathcal{C}}(M)$ by $\bar{t}_{\mathcal{C}}(M)$ (by adjusting the constants appropriately). One simply needs to observe that since each V_S is nonnegative,

$$\rho_{\mathcal{A}}(\mu) = \frac{1}{2}\mathbb{E}_0\sqrt{\frac{1}{|\mathcal{A}|}\sum_{S\in\mathcal{A}} V_S} \ge \frac{1}{2}\mathbb{E}_0\sqrt{\frac{1}{|\mathcal{A}|}\sum_{S\in\widehat{\mathcal{A}}} V_S} = \frac{1}{\sqrt{2}}\rho_{\widehat{\mathcal{A}}}(\mu).$$

If $\bar{t}_{\mathcal{C}}(M)$ is significantly larger than $t_{\mathcal{C}}(M)$, the gain may be substantial.

If the class $\mathcal{C}$ is symmetric then thanks to Theorem 5.1, the theorem above can be improved and simplified. If the class is symmetric, instead of having to work with randomly chosen subclasses, one may optimally choose a separated subset. Then the bounds can be expressed in terms of the metric entropy of $\mathcal{C}$, more precisely, by its *packing numbers* with respect to the canonical distance $d(S,T) = \sqrt{\mathbb{E}_0(X_S - X_T)^2}$.

We say that $\mathcal{A} \subset \mathcal{C}$ is a t-separated set (or t-packing) if for any $S, T \in \mathcal{A}$, $d(S,T) \ge t$. For $t < \sqrt{2K}$, define the *packing number* $M(t)$ as the size of a maximal t-separated subset $\mathcal{A}$ of $\mathcal{C}$. It is a simple well-known fact that packing numbers are closely related to the covering numbers introduced in Section 2 by the inequalities $N(t) \le M(t) \le N(t/2)$.

THEOREM 6.2. *Let $\mathcal{C}$ be symmetric in the sense of Theorem 5.1 and let $t \le \sqrt{2K}$. Then*

$$R^*_{\mathcal{C}} \ge 1/2,$$

whenever

$$\mu \le \min\left(\sqrt{\frac{\log(M(t)/16)}{K}}, \frac{8\log(\sqrt{3}/2)}{\sqrt{K - t^2/2}}\right).$$

PROOF. Let $\mathcal{A} \subset \mathcal{C}$ be a maximal t-separated subclass. Since $\mathcal{C}$ is symmetric, by Theorem 5.1, $R^*_{\mathcal{C}} \ge R^*_{\mathcal{A}}$ so it suffices to show that $R^*_{\mathcal{A}} \ge 1/2$ for the indicated values of μ. The rest of the proof is identical to that of Theorem 6.1. □

To interpret this result, take $t = \sqrt{2K(1-\varepsilon)}$ for some $\varepsilon \in (0, 1/2)$. Then, by the theorem, $R^* \ge 1/2$ if

$$\mu \le \frac{1}{\sqrt{K}} \min\left(\frac{8\log(\sqrt{3}/2)}{\sqrt{\varepsilon}}, \sqrt{\log\bigl(M\bigl(\sqrt{2K(1-\varepsilon)}\bigr)/16\bigr)}\right).$$

As an example, suppose that the class $\mathcal{C}$ is such that there exists a constant $V > 0$ such that $M(t) \sim (n/t^2)^V$. (Recall that all classes with VC dimension V have an upper bound of this form for the packing numbers, see remark on page 3069.) In this case, one may choose $\varepsilon^{-1} \sim V \log(n/K)$ and obtain that $R^* \ge 1/2$ whenever $\mu \le c\sqrt{(V/K)\log(n/K)}$ (for some constant c). This closely matches the bound obtained for the maximum test by Dudley's chaining bound.

7. Optimal versus maximum test: An analysis of the type I error.

In all examples considered above, upper bounds for the optimal risk R^* are derived by analyzing either the maximum test or the averaging test. As the examples show, very often these simple tests have a near-optimal performance. The optimal test f^* is generally more difficult to study. In this section, we analyze directly the performance of the optimal test. More precisely, we derive general upper bounds for the type I error (i.e., the probability that the null hypothesis is rejected under $\mathbb{P}_0$) of f^*. The upper bound involves the expected value of the maximum of a Gaussian process indexed by a sparse subset of $\mathcal{C}$ and can be significantly smaller than the maximum over the whole class that appears in the performance bound of the maximum test in Proposition 2.2. Unfortunately, we do not have an analogous bound for the type II error.

We consider the type I error of the optimal test f^*

$$\mathbb{P}_0\{f^*(\mathbf{X}) = 1\} = \mathbb{P}_0\{L(\mathbf{X}) > 1\} = \mathbb{P}_0\left\{\frac{1}{N}\sum_{S \in \mathcal{C}} e^{\mu X_S} > e^{K\mu^2/2}\right\}.$$

An easy bound is $\frac{1}{N}\sum_{S\in\mathcal{C}} e^{\mu X_S} \le e^{\mu \max_{S\in\mathcal{C}} X_S}$ so

$$\mathbb{P}_0\{L(\mathbf{X}) > 1\} \le \mathbb{P}_0\Big\{\max_S X_S > K\mu/2\Big\}.$$

Thus, $\mathbb{P}_0\{L(\mathbf{X}) > 1\} \le \delta$ whenever $\mu \ge (1/K)\mathbb{E}_0 \max_S X_S + \sqrt{(2/K)\log(1/\delta)}$. Of course, we already know this from Proposition 2.2 where this bound was derived for the (suboptimal) test based on maxima.

In order to understand the difference between the performance of the optimal test f^* and the maximum test, one needs to compare the random variables $(1/\mu)\log \frac{1}{N}\sum_{S\in\mathcal{C}} e^{\mu X_S}$ and $\max_{S\in\mathcal{C}} X_S$.

PROPOSITION 7.1. *For any $\delta \in (0,1)$, the type* I *error of the optimal test f^* satisfies*

$$\mathbb{P}_0\{f^*(\mathbf{X}) = 1\} \le \delta,$$

whenever

$$\mu \ge \frac{2}{K}\mathbb{E}_0 \max_{S\in\mathcal{A}} X_S + \sqrt{\frac{32\log(2/\delta)}{K}},$$

where $\mathcal{A}$ is any $\sqrt{K}/2$-cover of $\mathcal{C}$.

If $\mathcal{A}$ is a minimal $\sqrt{K}/2$-cover of $\mathcal{C}$, then

$$(1/K)\mathbb{E}_0 \max_{S\in\mathcal{A}} X_S \le \sqrt{\frac{2\log N(\sqrt{K}/2)}{K}}.$$

By "Sudakov's minoration" [see Ledoux and Talagrand (1991), Theorem 3.18] this upper bound is sharp up to a constant factor.

It is instructive to compare this bound with that of Proposition 2.2 for the performance of the maximum test. In Proposition 7.1, we were able to replace the expected maximum $\mathbb{E}_0 \max_{S\in\mathcal{C}} X_S$ by $\mathbb{E}_0 \max_{S\in\mathcal{A}} X_S$ where now the maximum is taken over a potentially much smaller subset $\mathcal{A} \subset \mathcal{C}$. It is not difficult to construct examples when there is a substantial difference, even in the order of magnitude, between the two expected maxima so we have a genuine gain over the simple upper bound of Proposition 2.2. Unfortunately, we do not know if an analog upper bound holds for the type II error $(1/N)\sum_{S\in\mathcal{C}} \mathbb{P}_S\{f^*(\mathbf{X}) = 0\}$ of the optimal test f^*. In cases when $\mathbb{E}_0 \max_{S\in\mathcal{A}} X_S \ll \mathbb{E}_0 \max_{S\in\mathcal{C}} X_S$, we suspect that the maximum test is far from optimal. However, to verify this conjecture, one would need a similar analysis for the type II error as well.

PROOF OF PROPOSITION 7.1. Introduce the notation

$$M_{\mathcal{C}}(\mu) = \mathbb{E}_0 \frac{1}{\mu}\log\bigg(\frac{1}{N}\sum_{S\in\mathcal{C}} e^{\mu X_S}\bigg).$$

Then

$$\begin{aligned}
&\mathbb{P}_0\left\{\frac{1}{N}\sum_{S\in\mathcal{C}} e^{\mu X_S} > e^{K\mu^2/2}\right\}\\
&\quad = \mathbb{P}_0\left\{\frac{1}{\mu}\log\left(\frac{1}{N}\sum_{S\in\mathcal{C}} e^{\mu X_S}\right) > \frac{K\mu}{2}\right\}\\
&\quad = \mathbb{P}_0\left\{\frac{1}{\mu}\log\left(\frac{1}{N}\sum_{S\in\mathcal{C}} e^{\mu X_S}\right) - M_{\mathcal{C}}(\mu) > \frac{K\mu}{2} - M_{\mathcal{C}}(\mu)\right\}.
\end{aligned}$$

We use Tsirelson's inequality (Lemma 2.1) to bound this probability. To this end, we need to show that the function $h:\mathbb{R}^N \to \mathbb{R}$ defined by

$$h(\mathbf{x}) = \frac{1}{\mu}\log\left(\frac{1}{N}\sum_{S\in\mathcal{C}} e^{\mu\sum_{i\in S} x_i}\right)$$

is Lipschitz [where $\mathbf{x} = (x_1,\dots,x_N)$]. Observing that

$$\frac{\partial h}{\partial x_j}(\mathbf{x}) = \frac{1/N\sum_{S\in\mathcal{C}} \mathbb{1}_{\{j\in S\}} e^{\mu x_S}}{1/N\sum_{S\in\mathcal{C}} e^{\mu x_S}} \in (0,1),$$

we have

$$\|\nabla h(\mathbf{x})\|^2 = \sum_{j=1}^n \left(\frac{\partial h}{\partial x_j}(\mathbf{x})\right)^2 \le \sum_{j=1}^n \frac{\partial h}{\partial x_j}(\mathbf{x}) = K$$

and therefore h is indeed Lipschitz $\sqrt{K}$. By Tsirelson's inequality, we have

$$\mathbb{P}_0\{f^*(\mathbf{X}) = 1\} \le \exp\left(-\frac{(K\mu/2 - M_{\mathcal{C}}(\mu))^2}{2K}\right).$$

Thus, the type I error is bounded by δ if

$$\mu \ge \frac{2M_{\mathcal{C}}(\mu)}{K} + \sqrt{\frac{8}{K}\log\frac{1}{\delta}}.$$

It remains to bound $M_{\mathcal{C}}(\mu)$.

Let $t \le \sqrt{2K}$ be a positive integer and consider a minimal t-cover of the set $\mathcal{C}$, that is, a set $\mathcal{A} \subset \mathcal{C}$ with cardinality $|\mathcal{A}| = N(t)$ such that, if $\pi(S)$ denotes an element in $\mathcal{A}$ whose distance to $S \in \mathcal{C}$ is minimal then $d(S,\pi(S)) \le t$ for all $S \in \mathcal{C}$. Then clearly,

$$M_{\mathcal{C}}(\mu) \le \mathbb{E}_0\frac{1}{\mu}\log\left(\frac{1}{N}\sum_{S\in\mathcal{C}} e^{\mu(X_S - X_{\pi(S)})}\right) + \mathbb{E}_0\max_{S\in\mathcal{A}} X_S.$$

To bound the first term on the right-hand side, note that, by Jensen's inequality,

$$\mathbb{E}_0\frac{1}{\mu}\log\left(\frac{1}{N}\sum_{S\in\mathcal{C}} e^{\mu(X_S - X_{\pi(S)})}\right) \le \frac{1}{\mu}\log\left(\frac{1}{N}\sum_{S\in\mathcal{C}} \mathbb{E}_0 e^{\mu(X_S - X_{\pi(S)})}\right) \le \frac{\mu t^2}{2}$$

since for each S, $d_H(X_S, X_{\pi(S)}) \le t^2$ and therefore $X_S - X_{\pi(S)}$ is a centered normal random variable with variance $d_H(X_S, X_{\pi(S)})$. For the second term, we have

$$\mathbb{E}_0 \max_{S \in \mathcal{A}} X_S \le \sqrt{2K \log N(t)}.$$

Choosing $t^2 = K/4$, we obtain the proposition. □

Acknowledgments. We thank Ery Arias-Castro and Emmanuel Candès for discussions on the topic of the paper. We also thank the referees for their valuable remarks.

REFERENCES

ALDOUS, D. J. (1990). The random walk construction of uniform spanning trees and uniform labelled trees. *SIAM J. Discrete Math.* **3** 450–465. MR1069105

ALON, N., KRIVELEVICH, M. and SUDAKOV, B. (1999). Finding a large hidden clique in a random graph. *Randoms Structures Algorithms* **13** 457–466. MR1662795

ARIAS-CASTRO, E., CANDÈS, E. J., HELGASON, H. and ZEITOUNI, O. (2008). Searching for a trail of evidence in a maze. *Ann. Statist.* **36** 1726–1757. MR2435454

ARIAS-CASTRO, E., CANDÈS, E. and DURAND, A. (2009). Detection of abnormal clusters in a network. Technical report, Univ. California, San Diego.

ARLOT, S., BLANCHARD, G. and ROQUAIN, E. (2010a). Some nonasymptotic results on resampling in high dimension, I: Confidence regions. *Ann. Statist.* **38** 51–82. MR2589316

ARLOT, S., BLANCHARD, G. and ROQUAIN, E. (2010b). Some nonasymptotic results on resampling in high dimension. II. Multiple tests. *Ann. Statist.* **38** 83–99. MR2589317

BARAUD, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli* **8** 577–606. MR1935648

BENJAMINI, I., LYONS, R., PERES, Y. and SCHRAMM, O. (2001). Uniform spanning forests. *Ann. Probab.* **29** 1–65. MR1825141

BHATTACHARYYA, A. (1946). On a measure of divergence between two multinomial populations. *Sankhyā* **7** 401–406. MR0018387

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2000). A sharp concentration inequality with applications. *Random Structures Algorithms* **16** 277–292. MR1749290

BRODER, A. (1989). Generating random spanning trees. In *30th Annual Symposium on Foundations of Computer Science* 442–447. IEEE Press, Research Triangle Park, NC.

DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation: The L_1 View*. Wiley, New York. MR0780746

DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York. MR1383093

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195

DUBDASHI, D. and RANJAN, D. (1998). Balls and bins: A study in negative dependence. *Random Structures Algorithms* **13** 99–124. MR1642566

DUDLEY, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6** 899–929. MR0512411

DUROT, C. and ROZENHOLC, Y. (2006). An adaptive test for zero mean. *Math. Methods Statist.* **15** 26–60. MR2225429

FEDER, T. and MIHAIL, M. (1992). Balanced matroids. In *STOC'92: Proceedings of the Twenty-Fourth Annual ACM Symposium on Theory of Computing* 26–38. ACM, New York.

FEIGE, U. and KRAUTHGAMER, R. (2000). Finding and certifying a large hidden clique in a semi-random graph. *Random Structures Algorithms* **16** 195–208. MR1742351

GLAZ, J., NAUS, J. and WALLENSTEIN, S. (2001). *Scan Statistics*. Springer, New York. MR1869112

GRIMMETT, G. R. and WINKLER, S. N. (2004). Negative association in uniform forests and connected graphs. *Random Structures Algorithms* **24** 444–460. MR2060630

HAUSSLER, D. (1995). Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik–Chervonenkis dimension. *J. Combin. Theory Ser. A* **69** 217–232. MR1313896

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

INGSTER, Y. I. (1999). Minimax detection of a signal for l_p^n-balls. *Math. Methods Statist.* **7** 401–428. MR1680087

JERRUM, M., SINCLAIR, A. and VIGODA, E. (2004). A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *J. ACM* **51** 671–697. MR2147852

LE CAM, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.* **41** 802–828. MR0267676

LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces*. Springer, New York. MR1102015

MOON, J. W. (1970). *Counting Labelled Trees*. *Canadian Mathematical Monographs* **1**. Canadian Mathematical Congress, Montreal. MR0274333

PROPP, J. G. and WILSON, D. B. (1998). How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph. *J. Algorithms* **27** 170–217. MR1622393

ROMANO, J. P. and WOLF, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.* **100** 94–108. MR2156821

SHABALIN, A. A., WEIGMAN, V. J., PEROU, C. M. and NOBEL, A. B. (2009). Finding large average submatrices in high dimensional data. *Ann. Appl. Statist.* **3** 985–1012.

SLEPIAN, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell System Tech. J.* **41** 463–501. MR0133183

TALAGRAND, M. (2005). *The Generic Chaining*. Springer, New York. MR2133757

TSIRELSON, B. S., IBRAGIMOV, I. A. and SUDAKOV, V. N. (1976). Norm of Gaussian sample function. In *Proceedings of the 3rd Japan–U.S.S.R. Symposium on Probability Theory*. *Lecture Notes in Math.* **550** 20–41. Springer, Berlin. MR0458556

VAPNIK, V. N. and CHERVONENKIS, A. Y. (1971). On the uniform convergence of relative frequences of events to their probabilities. *Theory Probab. Appl.* **16** 264–280.

VONNEGUT, K. (1973). *Breakfast of Champions*. Delacorte Press, New York.

L. ADDARIO-BERRY
DEPARTMENT OF MATHEMATICS AND STATISTICS
MCGILL UNIVERSITY
MONTREAL, H3A 2K6
CANADA
E-MAIL: louigi@gmail.com

N. BROUTIN
INRIA ROCQUENCOURT
78153 LE CHESNAY
FRANCE
E-MAIL: nicolas.broutin@inria.fr

L. DEVROYE
SCHOOL OF COMPUTER SCIENCE
MCGILL UNIVERSITY
MONTREAL, H3A 2A7
CANADA
E-MAIL: lucdevroye@gmail.com

G. LUGOSI
ICREA
AND DEPARTMENT OF ECONOMICS
POMPEU FABRA UNIVERSITY
BARCELONA
SPAIN
E-MAIL: gabor.lugosi@gmail.com