Combinatorial Hypothesis Testing

April 16, 2023

1 Introduction ❖

Suppose we observe an n-dimensional vector $\mathbf{X} = (X_1, ..., X_n)$. 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.

Combinatorics kicks in as we consider the alternative hypotheses, by introducing a class $\mathcal{C}$ with some combinatorial structure: consider a class $\mathcal{C} = \{S_1, \ldots, S_N\}$ of N sets of indices such that $S_k \subset \{1, \ldots, n\}$ for all $k = 1, \ldots, N$. Under H_1, there exists an $S \in \mathcal{C}$ such that X_i has a distribution determined by whether i is in S:

Alternative 1. [Detection of Means] In its simplest form, as discussed in [1, 3, 4], we consider

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

where $\mu > 0$ is a positive parameter and components of $\mathbf{X}$ are independent.

Alternative 2. [Detection of Correlations] In testing correlations [2], we consider

$$\mathrm{Cov}(X_i, X_j) = \begin{cases} 1, & \text{if } i = j \\ \rho, & \text{if } i \neq j \text{ with } i, j \in S \\ 0, & \text{otherwise} \end{cases}$$

For each $S \in \mathcal{C}$, we denote the probability measure and expectation by $\mathbb{P}_S$ and $\mathbb{E}_S$, respectively. Many interesting examples of $\mathcal{C}$ arises for this scenario: subsets of size K, cliques, perfect matchings, spanning trees, and clusters.

A *test* is a binary-valued function $f : \mathbb{R}^n \to \{0, 1\}$. If $f(X) = 0$, then the test accepts the null hypothesis H_0; otherwise H_0 is rejected by f. We measure the performance of a test based on the *minimax risk*:

$$R_*^{\max} := \inf_f R^{\max}(f).$$

where $R^{\max}(f)$ is the worst-case risk over the class of interest $\mathcal{C}$, formally defined by

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

In this report, we discuss the techniques introduced in [1–3] to derive the asymptotic upper and lower bounds of $R_*^{\max}$, as well as more recent extensions.

2 Lower Bounds ❖

A standard way of obtaining lower bounds for the minimax risk is by putting a prior on the class $\mathcal{C}$ and obtaining a lower bound on the corresponding *Bayesian risk*, which never exceeds the worst-case risk. Because this is true for any prior, the idea is to find one that is hardest (often called *least favorable*). Consider the uniform prior on $\mathcal{C}$, giving rise to the following *average risk*:

$$R(f) = \mathbb{P}_0\{f(X) = 1\} + \mathbb{P}_1\{f(X) = 0\},$$

where

$$\mathbb{P}_1\{f(X) = 0\} := \frac{1}{N} \sum_{S \in \mathcal{C}} \mathbb{P}_S\{f(X) = 0\},$$

and $N := |\mathcal{C}|$ is the cardinality of $\mathcal{C}$. The advantage of considering the average risk over the worst-case risk is that we know an optimal test for the former, which, by the Neyman–Pearson fundamental lemma, is the likelihood ratio test, denoted f^*. Introducing $L(X)$, the likelihood ratio between H_0 and H_1, the optimal test becomes

$$f^*(x) = 0 \quad \text{if and only if} \quad L(x) \leq 1.$$

The (average) risk $R^* = R(f^*)$ of the optimal test is called the *Bayes risk* and it satisfies

$$R^* = 1 - \frac{1}{2}\mathbb{E}_0|L(X) - 1|$$

2.1 Detection of Means ❄

In this section we focus on the first alternative hypothesis 1. In this case, if we write

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

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 $x_S = \sum_{i\in S} x_i$. The Bayes risk can then be written as

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

Via Jensen's inequality, we observe that

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

because for any $S \in \mathcal{C}$,

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

Combining this inequality with $R^* \geq 1 - \sqrt{1 - (\mathbb{E}_0\sqrt{L(\mathbf{X})})^2}$, we see that for all classes $\mathcal{C}$, $R^* \geq 1/2$ whenever $\mu \leq \sqrt{(4/K)} \times \sqrt{\log(4/3)}$, i.e. small risk cannot be achieved unless μ is substantially large compared to $K^{-1/2}$.

2.1.1 Moment Methods

The moment method applies the following insight to move beyond the lower bound we obtained earlier: by the Cauchy–Schwarz inequality,

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

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

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

We are now ready to prove the following lower bound based on overlapping pairs, which reduces the problem to studying a purely combinatorial quantity [1, 4]:

Proposition 2.1 ([1], 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. Because $L(\mathbf{X}) = \frac{1}{N}\sum_{S\in\mathcal{C}} e^{\mu X_S - K\mu^2/2}$,

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

Meanwhile,

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

□

Example 2.2 (Disjoint Sets, [1], Section 4.1). Suppose all $S \in \mathcal{C}$ are disjoint (and therefore $KN \leq n$). Fix $\delta \in (0, 1)$. Let $Z = K$ with probability $1/N$ and $Z = 0$ otherwise. Thus,

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

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

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

Example 2.3 (Spanning Trees, [1], Section 4.5). Let $1, 2, \ldots, n = \binom{m}{2}$ 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 $K = m - 1$. With the fact$\mathbb{E}[e^{\mu^2 Z}] \leq \exp(2e^{\mu^2})$, we obtain that for any $\delta \in (0, 1)$, $R^* \geq \delta$ whenever

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

Example 2.4 (Cliques, [1], Section 4.6). Consider the random variables $X_1, \ldots, X_n$ 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 $N = \binom{m}{k}$. With some technical work, one can show that $\mathbb{E}[\exp(\mu^2 Z)] \leq 2$. This gives us $R^* \geq 1/2$ whenever

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

Thus, 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, we can obtain lower bounds for the critical value of μ. This allows us to exploit special combinatorial structures of the class $\mathcal{C}$; one such combinatorial property is symmetry:

Definition 2.5. We say that the class $\mathcal{C}$ is *symmetric* if it satisfies the following conditions. Let S, S' be drawn independently and uniformly at random from $\mathcal{C}$. Then,

1. the conditional distribution of $Z = |S \cap S'|$ given S' is identical for all values of S';
2. for any fixed $S_0 \in \mathcal{C}$ and $i \in S_0$, $\mathbb{P}\{i \in S\} = K/n$.

Via Hölder's inequality, we can obtain the folloiwng improvement of the universal lower bound obtained earlier.

Proposition 2.6 ([1], Proposition 3.3)**.** *Let $\delta \in (0,1)$. Assume that $\mathcal{C}$ is symmetric. 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. Integrating Hölder's inequality and symmetry, we obtain

$$\mathbb{E}[e^{\mu^2 Z}] \leq (e^{\mu^2 K} - 1)\frac{K}{n} + 1.$$

Then we can apply Proposition 2.1. We omit the details here. □

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 $K \leq n^\beta$ for some $\beta \in (0,1)$.

Example 2.7 (Stars, [1], Section 4.4)**.** A star is a subgraph of the complete graph K_m which contains all $K = m-1$ edges incident to a fixed vertex. Consider the set $\mathcal{C}$ of all stars in K_m. In this setting, $n = \binom{m}{2}$ and $N = m$. Hence, for any $\varepsilon > 0$, we have $\lim_{m\to\infty} R^* = 1$ if

$$\mu \leq (1-\varepsilon)\sqrt{\frac{\log m}{m}}$$

Another interesting property is negative association, which allow us to improve the previous lower bound further.

Definition 2.8. 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,

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

Proposition 2.9 ([1], Proposition 3.4)**.** *Let $\delta \in (0,1)$ and assume that the class $\mathcal{C}$ is symmetric. 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 $\mathbf{1}_{\{1\in S\}}, \ldots, \mathbf{1}_{\{K\in S\}}$ are negatively associated, then $R^* \geq \delta$ for all μ with*

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

Proof. Negative association gives us

$$\mathbb{E}[e^{\mu^2 Z}] \leq \left((e^{\mu^2} - 1)\frac{K}{n} + 1\right)^K.$$

Then we can apply Proposition 2.1. We omit the details here. □

Example 2.10 (K-sets, [1], Section 4.2)**.** Consider the example when $\mathcal{C}$ contains all sets $S \subset \{1, \ldots, n\}$ of size K. Note $N = \binom{n}{K}$. This class is symmetric and satisfies the condition in the previous proposition.

Example 2.11 (Perfect Matchings, [1], Section 4.3)**.** 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 $K = m$. 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. Hence for all m, $R^* \geq \delta$ whenever

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

2.2 Detection of Correlations ❄

First, we note that we can rewrite the hypotheses as

$$H_0 : \mathbf{X} \sim \mathcal{N}(0, \mathbf{I}) \quad \text{vs.} \quad H_1 : \mathbf{X} \sim \mathcal{N}(0, \mathbf{A}_S) \qquad \text{for some } S \in \mathcal{C},$$

where $\mathbf{I}$ denotes the $n \times n$ identity matrix and

$$(\mathbf{A}_S)_{i,j} = \begin{cases} 1, & \text{i = j}, \\ \rho, & i \neq j, i, j \in S, \\ 0, & \text{otherwise}. \end{cases}$$

Introducing

$$Z_S = \exp\bigl(\tfrac{1}{2} X^T (\mathbf{I} - \mathbf{A}_S^{-1}) X\bigr)$$

for all $S \in \mathcal{C}$, the likelihood ratio between H_0 and H_1 may be written as

$$L(X) = \frac{1}{N} \sum_{S \in \mathcal{C}} \frac{Z_S}{\mathbb{E}_0 Z_S}$$

Thus the Bayes risk satisfies

$$R^* = 1 - \frac{1}{2} \mathbb{E}_0 |L(X) - 1| = 1 - \frac{1}{2} \mathbb{E}_0 \left| \frac{1}{N} \sum_{S \in \mathcal{C}} \frac{Z_S}{\mathbb{E}_0 Z_S} - 1 \right|.$$

The next representation theorem of Gaussain random variables plays a key role in analysing this test:

Lemma 2.12 ([5]; [2], Lemma 1.1). *Let $X_1, \dots, X_k$ be standard normal with* $\mathrm{Cov}(X_i, X_j) = \rho$ *for $i \neq j$. Then there are i.i.d. standard normal random variables, denoted $U, U_1, \dots, U_k$, such that $X_i = \sqrt{\rho} U + \sqrt{1 - \rho} U_i$ for all i.*

Thus, given U, the problem becomes that of detecting a subset of variables with nonzero mean (equal to $\sqrt{\rho} U$) and with a variance equal to $1 - \rho$ (instead of 1). This simple observation will be very useful to us later on.

When $\mathcal{C}$ contains just one set $S = \{1, \dots, k\}$, we can leverage the following lemma and the fact that $\mathbb{E}_0 Z_S = \sqrt{\det(\mathbf{A}_S)}$ to analyse the Bayes risk directly.

Lemma 2.13 ([2], Lemma 2.1). *Under $\mathbb{P}_0$, $X^T (\mathbf{I} - \mathbf{A}_S^{-1}) X$ is distributed as*

$$-\frac{\rho}{1 - \rho} \chi^2_{k-1} + \frac{\rho(k-1)}{1 + \rho(k-1)} \chi^2_1,$$

and under the alternative $\mathbb{P}_S$, it has the same distribution as

$$-\rho \chi^2_{k-1} + \rho(k-1) \chi^2_1,$$

where χ^2_1 and χ^2_{k-1} denote independent χ^2 random variables with degrees of freedom 1 and $k - 1$, respectively.

Proposition 2.14 ([2], Proposition 2.1). $\lim_{k \to \infty} R^* = 0$ *if and only if $\rho k \to \infty$. Similarly,* $\lim_{k \to \infty} R^* = 1$ *if and only if $\rho k \to 0$.*

Proof. Suppose $\rho k \to \infty$. It suffices to show that there exists a threshold τ_k such that $\mathbb{P}_0\{X^T (\mathbf{I} - \mathbf{A}_S^{-1}) X \geq \tau_k\} \to 0$ and $\mathbb{P}_S\{X^T (\mathbf{I} - \mathbf{A}_S^{-1}) X < \tau_k\} \to 0$. We use Lemma 2.13 and the fact that, by Chebyshev's inequality,

$$\mathbf{P}\bigl\{|\chi^2_k - k| > t_k \sqrt{k}\bigr\} \to 0, \qquad k \to \infty,$$

for any sequence $t_k \to \infty$, and the fact that

$$\mathbf{P}\{t_k^{-1} < \chi^2_1 < t_k\} \to 1 \qquad \text{as } k \to \infty.$$

We choose $t_k = \log k$ and define $\tau_k := -\rho k + \rho t_k \sqrt{k} + t_k$. Then under the null,

$$\mathbb{P}_0\{X^T(\mathbf{I} - \mathbf{A}_S^{-1})X \geq \tau_k\} \to 0,$$

and under the alternative, setting $\eta_k := -\rho k - \rho t_k \sqrt{k} + \rho k t_k^{-1}$,

$$\mathbb{P}_S\{X^T(\mathbf{I} - \mathbf{A}_S^{-1})X < \eta_k\} \to 0.$$

We then conclude with the fact that, for k large enough, $\tau_k < \eta_k$.

If ρk is bounded, the densities of the test statistic under both hypotheses have a significant overlap and the risk cannot converge to 0.

The proof of the second statement is similar. □

2.2.1 Generalised Moment Method

When $\mathcal{C} > 1$, an direct application of the momont method discussed earlier does not yield very promising lower-bounds; instead, we leverage the insight from the Representation Lemma 2.12.

Proposition 2.15 ([2], Theorem 2.1)**.** *For any class $\mathcal{C}$ and any $a > 0$,*

$$R^* \geq \mathbf{P}\{|\mathcal{N}(0,1)| \leq a\}\left(1 - \tfrac{1}{2}\sqrt{\mathbb{E}\exp(\nu_a\ Z) - 1}\right),$$

where $\nu_a := \rho a^2/(1+\rho) - \frac{1}{2}\log(1-\rho^2)$ and $Z = |S \cap S'|$, with S, S' drawn independently, uniformly at random from $\mathcal{C}$. In particular, taking $a = 1$,

$$R^* \geq 0.6 - 0.3\sqrt{\mathbb{E}\exp(\nu_1 Z) - 1},$$

where $\nu_1 = \nu(\rho) := \rho/(1+\rho) - \frac{1}{2}\log(1-\rho^2)$.

Proof. Via Lemma 2.12, we can write

$$X_i = \begin{cases} U_i, & \text{if } i \notin S, \\ \sqrt{\rho}U + \sqrt{1-\rho}U_i, & \text{if } i \in S \end{cases}$$

where $U, U_1, \ldots, U_n$ are independent standard normal random variables. We consider now the alternative $H_1(u)$, defined as the alternative H_1 given $U = u$. Let $R(f)$, L, f^* [resp., $R_u(f)$, L_u, f_u^*] be the risk of a test f, the likelihood ratio, and the optimal (likelihood ratio) test, for H_0 versus H_1 [resp., H_0 versus $H_1(u)$]. For any $u \in \mathbb{R}$, $R_u(f_u^*) \leq R_u(f^*)$, by the optimality of f_u^* for H_0 versus $H_1(u)$. Therefore, conditioning on U,

$$R^* = R(f^*) = \mathbb{E}_U R_U(f^*) \geq \mathbb{E}_U R_U(f_U^*) = 1 - \tfrac{1}{2}\mathbb{E}_U\mathbb{E}_0|L_U(X) - 1|$$

Using the fact that $\mathbb{E}_0|L_u(X) - 1| \leq 2$ for all u, we have

$$\mathbb{E}_U\mathbb{E}_0|L_U(X) - 1| \leq 2\mathbb{P}\{|U| > a\} + \mathbb{P}\{|U| \leq a\} \max_{u\in[-a,a]} \mathbb{E}_0|L_u(X) - 1|$$

and therefore, using the Cauchy–Schwarz inequality,

$$\begin{aligned} 1 - \frac{1}{2}\mathbb{E}_U\mathbb{E}_0|L_U(X) - 1| &\geq \mathbb{P}\{|U| \leq a\}\left(1 - \frac{1}{2}\max_{u\in[-a,a]} \mathbb{E}_0|L_u(X) - 1|\right) \\ &\geq \mathbb{P}\{|U| \leq a\}\left(1 - \frac{1}{2}\max_{u\in[-a,a]} \sqrt{\mathbb{E}_0 L_u^2(X) - 1}\right). \end{aligned}$$

After some computation, we obtain

$$\mathbb{E}_0 L_u^2(X) \leq \frac{1}{N^2}\sum_{S,S'\in\mathcal{C}} \exp\left(\left(\frac{\rho u^2}{1+\rho} - \frac{1}{2}\log(1-\rho^2)\right)|S \cap S'|\right)$$

□

Again, we reduce the problem to studying the purely combinatorial quantity $Z = |S \cap S'|$. We demonstrate the implications of this proposition via a few examples.

Example 2.16 (Disjoint Sets, [2], Section 2.3.1)**.** Suppose all $S \in \mathcal{C}$ are disjoint (and therefore $KN \leq n$). Let $Z = K$ with probability $1/N$ and $Z = 0$ otherwise. Thus,

$$\mathbb{E}e^{\nu Z} - 1 = \frac{1}{N}(e^{\nu K} - 1) \leq \frac{1}{N}e^{\nu K}$$

which is bounded by 1 if $\nu \leq \log(N)/k$, in which case $R^* \geq 0.3$.

Example 2.17 (k-intervals, [2], Section 2.3.2)**.** Suppose $\mathcal{C}$ is the class of all intervals of size k of the form $\{i, \ldots, i+k-1\}$ modulo n. Then $N \leq n$. For two k-intervals chosen independently and uniformly at random,

$$\mathbb{P}\{|S \cap S'| = \ell\} = \frac{2}{N} \qquad \forall \ell = 1, \ldots, k.$$

Thus,

$$\mathbb{E}e^{\nu Z} - 1 = \frac{2}{N}\left(\sum_{\ell=1}^{k} e^{\nu \ell} - k\right) \leq \frac{2k}{N}e^{\nu k},$$

which is bounded by 1 if

$$\nu \leq \frac{\log(n/2k)}{k}$$

in which case $R^* \geq 0.3$.

Example 2.18 (k-sets, [2], Section 2.3.3)**.** Suppose $\mathcal{C}$ is the class of all sets of size k. By negative association, (see Proposition 2.9)

$$\mathbb{E}e^{\nu Z} \leq \left((e^{\nu} - 1)\frac{k}{n} + 1\right)^k \leq \exp\left((e^{\nu} - 1)\frac{k^2}{n}\right),$$

which is bounded by 2 when

$$\frac{k^2}{n} \leq \frac{\ln 2}{\exp(\nu(\rho)) - 1}$$

in which case $R^* \geq 0.3$.

Example 2.19 (Perfect Matchings, [2], Section 2.3.4)**.** Suppose $\mathcal{C}$ is the class of all perfect matchings of size $k = \sqrt{n}$. Using the same Z as in Example 2.11,

$$\mathbb{E}e^{\nu Z} \leq \left((e^{\nu} - 1)\frac{k}{n} + 1\right)^k \leq \exp\left((e^{\nu} - 1)\frac{k^2}{n}\right),$$

which is bounded by 2 when

$$\frac{k^2}{n} \leq \frac{\ln 2}{\exp(\nu(\rho)) - 1}$$

in which case $R^* \geq 0.3$.

Example 2.20 (Spanning Trees, [2], Section 2.3.5)**.** Suppose $\mathcal{C}$ is the class of all spanning trees of a complete graph with $k+1$ vertices. Similar to Example 2.3, notice

$$\mathbb{E}e^{\nu Z} \leq \exp 2(e^{\nu} - 1),$$

which is bounded by $13/4$ when $\nu \leq 1 + \ln((\ln(13/4))/2)$, in which case $R^* \geq 0.15$.

3 Clusters ❖

4 Extension ❖

4 References ❖

[1] Louigi Addario-Berry, Nicolas Broutin, Luc Devroye, and Gábor Lugosi. On combinatorial testing problems. *The Annals of Statistics*, pages 3063–3092, 2010.

[2] ERY ARIAS-CASTRO, SÉBASTIEN BUBECK, and GÁBOR LUGOSI. Detection of correlations. *The Annals of Statistics*, 40(1):412–435, 2012.

[3] Ery Arias-Castro, Emmanuel J Candes, and Arnaud Durand. Detection of an anomalous cluster in a network. *The Annals of Statistics*, pages 278–304, 2011.

[4] Ery Arias-Castro, Emmanuel J Candès, Hannes Helgason, and Ofer Zeitouni. Searching for a trail of evidence in a maze. *Ann. Statist.*, 36(1):1726–1757, 2008.

[5] Simeon M Berman. Equally correlated random variables. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 155–156, 1962.