Lectures on Combinatorial Statistics

Gábor Lugosi

July 12, 2017

Contents

Chapter 1

The hidden clique problem

In order to warm up and to get a feeling of what we mean by *combinatorial statistics*, in this introductory chapter we discuss a classical problem, the so-called *hidden clique* (or *planted clique*) problem. While the question is simple to state, after a few simple observations one quickly runs into unexpected difficulties and, after decades of serious attempts, the basic question still remains far from being solved.

1.1 A hypothesis testing problem

The hidden clique problem is a simple *hypothesis testing* problem in which one observes a (labeled) graph on the set of vertices $V = \{1, \dots, n\} = [n]$. One is asked to decide between two possible ways the observed graph is generated: either the graph is an Erdős-Rényi random graph $\mathcal{G}(n, 1/2)$ (i.e., each pair of vertices is connected by an edge independently, with probability 1/2) or, alternatively, the graph is distributed as $\mathcal{G}(n, 1/2, k)$, in which a random subset of k of the vertices form a clique and all other edges are present independently, with probability 1/2.

One may formally set up the problem as a hypothesis testing problem in which the *null hypothesis* is that an observed graph g is a realization of a random graph G drawn from the distribution $\mathbb{P}_0$ of a $\mathcal{G}(n, 1/2)$ random graph. The *alternative hypothesis* is that g is drawn from the distribution $\mathbb{P}_1$ of $\mathcal{G}(n, 1/2, k)$.

In other words, $\mathbb{P}_0\{G = g\} = 2^{-\binom{n}{2}}$ for all graphs g, and

$$\mathbb{P}_1\{G = g\} = \frac{1}{\binom{n}{k}} \sum_{S \subset V : |S| = k} \mathbb{P}_S\{G = g\} \,,$$

where for each set S of k vertices, $\mathbb{P}_S$ denotes the distribution of a random graph with S as a planted clique, that is,

$$\mathbb{P}_S\{G = g\} = \begin{cases} 0 & \text{if } S \text{ is not a clique of } g \\ 2^{-\binom{n}{2}+\binom{k}{2}} & \text{otherwise.} \end{cases}$$

One may represent a graph g on the vertex set $V = \{1, \dots, n\}$ by a binary vector of length $\binom{n}{2}$ indexed by pairs (i, j) of vertices $(1 \le i < j \le n)$ such that $g_{(i,j)} = 1$ if vertices i and j are joined by an edge in g and $g_{(i,j)} = 0$ otherwise. A *test* is a function $T : \{0,1\}^{\binom{n}{2}} \to \{0,1\}$. If $T(g) = 0$, we say that the test *accepts the null hypothesis*, otherwise it rejects it.

There are various ways of measuring the performance of a test. The *type I error* is the probability that the null hypothesis is incorrectly rejected, that is,

$$\mathbb{P}_0 \{T(G) = 1\} \ .$$

Symmetrically, *type II error* is $\mathbb{P}_1 \{T(G) = 0\}$. A simple way of measuring the quality of a test T is by its *risk*, defined as the sum of the two types of errors

$$R(T) = \mathbb{P}_0 \{T(G) = 1\} + \mathbb{P}_1 \{T(G) = 0\} \ .$$

As it is well known—and easy to prove—, the test T^* that minimizes the risk is the so-called *likelihood ratio test* defined by

$$T^*(g) = 0 \qquad \text{if and only if} \qquad L(g) \le 1 \ ,$$

where

$$L(g) = \frac{\mathbb{P}_1\{G = g\}}{\mathbb{P}_0\{G = g\}}$$

is the *likelihood ratio*. The risk of the optimal test equals

$$R^* = R(T^*) = 1 - \frac{1}{2} \sum_g |\mathbb{P}_1\{G = g\} - \mathbb{P}_0\{G = g\}| = 1 - \frac{1}{2}\mathbb{E}_0|L(G) - 1| \ , \qquad (1.1)$$

where $\mathbb{E}_0$ denotes expectation with respect to the probability measure $\mathbb{P}_0$. We leave the derivation of these simple facts as exercises. Note that

$$R^* = 1 - D(\mathbb{P}_0, \mathbb{P}_1) \ ,$$

where $D(\mu, \nu) = \sup_A |\mu(A) - \nu(A)|$ denotes the *total variation distance* between the probability measures μ and ν. (The supremum is taken over all measurable sets A.)

It is intuitively clear that for large values of k the testing problem is "easy" in the sense that there exists a test with a small risk. On the other hand, small hidden cliques are difficult—or even impossible—to detect which means that R^* is close to its maximum value 1. (Note that for the trivial test that always accepts the null hypothesis, one has $R(T) = 1$.)

Our goal first is to understand for what values of k (as a function of n) is R^* close to zero or one. As it turns out, remarkably tight results may be deduced quite simply. In particular, in the next two sections we prove the following result that shows that the transition from asymptotically vanishing to maximal risk is surprisingly sharp. It all happens in an inteval of length at most two!

Theorem 1.2. *Let $\epsilon > 0$ be arbitrary and define $\omega_n = 2\log_2 n - 2\log_2\log_2 n - 1 + 2\log_2 e$. Then the minimal possible risk of any test for detecting a hidden clique of size $k = k_n$ satisfies*

$$\lim_{n\to\infty} R^* = 0 \quad \text{if } k_n > \lfloor \omega_n + \epsilon \rfloor$$
$$\lim_{n\to\infty} R^* = 1 \quad \text{if } k_n < \lfloor \omega_n - \epsilon \rfloor .$$

The theorem follows from Propositions 1.3 and 1.6 below. As in the theorem above, we are mostly interested in large values of n. Given a sequence of events (A_n), we say that A_n occurs *with high probability* if the probability of A_n tends to 1 as $n \to \infty$.

1.2 The clique number of the Erdős-Rényi random graph and a simple test

To prove upper bounds for R^*, it suffices to construct a test with a small risk. A natural candidate is a test based on the size of the largest clique of the observed graph. To analyze such a test, we need to understand the behavior of the size (i.e., the number of vertices) of the largest clique of an Erdős-Rényi random graph $G \sim \mathcal{G}(n, 1/2)$. The *clique number* of a graph is defined as the size of the largest clique in the graph.

For our purposes it suffices to derive upper bounds for the clique number $\omega(G)$, of an Erdős-Rényi random graph G. This may be done by the so-called *first moment method* as follows. For a positive integer $m \le n$, denote by N_m the number of cliques of size m in G. Then, by the linearity of expectation,

$$\mathbb{E}N_m = \binom{n}{m} 2^{-\binom{m}{2}} ,$$

as each of the $\binom{n}{m}$ subsets of V of size m forms a clique with probability $2^{-\binom{m}{2}}$. Since

$$\mathbb{P}\{\omega(G) \ge m\} = \mathbb{P}\{N_m \ge 1\} \le \mathbb{E}N_m ,$$

we see that $\omega(G) < m$ with high probability, whenever $\binom{n}{m}2^{-\binom{m}{2}} \to 0$. By a quick calculation, we see that this is the case for $m \ge 2\log_2 n + 3$. Indeed, for such values, using nothing but $\binom{n}{m} \le n^m$, we have

$$\binom{n}{m} 2^{-\binom{m}{2}} \le \left(n 2^{-(m-1)/2}\right)^m \le 2^{-m} \to 0 .$$

By a more careful bounding of the binomial coefficients using Stirling's formula (Exercise 1.2), one may show that for any $\epsilon > 0$,

$$\omega(G) \le \lfloor \omega_n + \epsilon \rfloor \quad \text{with high probability} ,$$

where $\omega_n = 2\log_2 n - 2\log_2\log_2 n - 1 + 2\log_2 e$.

Now we may use this bound to establish a lower bound for the value of k such that a hidden clique of size k is detectable in the sense that the risk of the optimal test converges to zero.

Proposition 1.3. *Let $\epsilon > 0$. Consider the test T that, upon observing the graph G accepts the null hypothesis (i.e., that $G \sim \mathcal{G}(n, 1/2)$) if and only if the largest clique of G has size less than k. Then $R(T) \to 0$ as $n \to \infty$ whenever $k > \lfloor \omega_n + \epsilon \rfloor$.*

Proof. By the argument above, under the null hypothesis, the largest clique of G is not larger than $\lfloor \omega_n + \epsilon \rfloor$, with high probability, and therefore $\mathbb{P}_0\{T(X) = 1\} \to 0$. On the other hand, under the alternative hypothesis, G contains a clique of size $k > \lfloor \omega_n + \epsilon \rfloor$ with probability one, making $\mathbb{P}_1\{T(X) = 0\} = 0$. □

The results of this section show that the simple test that checks whether the largest clique has at least k vertices has a vanishing risk whenever $k > \omega_n + 1 \approx 2\log_2 n - 2\log_2\log_2 n$. This test cannot work for smaller values of k because the largest clique in $\mathcal{G}(n, 1/2)$ is, in fact, at least $\lfloor \omega_n - \epsilon \rfloor$, with high probability, for all $\epsilon > 0$. In other words, the clique number satisfies the remarkable property that

$$\omega(G) \in (\lfloor \omega_n - \epsilon \rfloor, \lfloor \omega_n + \epsilon \rfloor) \quad \text{with high probability .}$$

Thus, $\omega(G)$ is concentrated on just two values, with high probability. The lower bound may be proven by the *second moment method* whose basic idea is that, for any integer $m \le n$,

$$\mathbb{P}\{\omega(G) < m\} = \mathbb{P}\{N_m = 0\} = \mathbb{P}\{N_m - \mathbb{E}N_m \le -\mathbb{E}N_m\} \le \frac{\mathrm{Var}(N_m)}{(\mathbb{E}N_m)^2} ,$$

by Chebyshev's inequality. As we have already seen in the previous section, $\mathbb{E}N_m = \binom{n}{m} 2^{-\binom{m}{2}}$. Moreover,

$$\begin{aligned} \mathbb{E}N_m^2 &= \mathbb{E}\left(\sum_{S \subset V : |S| = m} \mathbb{1}_{\{S \text{ is a clique in } G\}} \right)^2 \\ &= \sum_{S,T \subset V : |S| = |T| = m} \mathbb{P}\{S, T \text{ are both cliques in } G\} \\ &= \binom{n}{m} 2^{-\binom{m}{2}} \sum_{i=0}^{m} \binom{m}{i} \binom{n-m}{m-i} 2^{-\binom{m}{2} + \binom{i}{2}} . \end{aligned} \tag{1.4}$$

Thus,

$$\begin{aligned}\frac{\mathrm{Var}(N_m)}{(\mathbb{E}N_m)^2} &= \frac{\mathbb{E}N_m^2}{(\mathbb{E}N_m)^2} - 1 \\ &= \frac{\sum_{i=0}^{m}\binom{m}{i}\binom{n-m}{m-i}2^{-\binom{m}{2}+\binom{i}{2}}}{\binom{n}{m}2^{-\binom{m}{2}}} - 1 \\ &\leq \frac{\sum_{i=1}^{m}\binom{m}{i}\binom{n-m}{m-i}2^{\binom{i}{2}}}{\binom{n}{m}} . \end{aligned} \tag{1.5}$$

By careful bounding of the binomial coefficients, one may now show that the expression above converges to zero whenever $m < \lfloor \omega_n - \epsilon \rfloor$, concluding the proof of the two-point concentration of the clique number $\omega(G)$.

1.3 Lower bound for the risk

In the previous section we examined of the performance of the simple test that accepts the null hypothesis if the largest clique has at least k vertices. We showed that the test has a vanishing risk when k is at least $\lfloor \omega_n + \epsilon \rfloor$ but it has a risk tending to 1 when $k < \lfloor \omega_n - \epsilon \rfloor$.

In this section we study whether there exist other tests that work for even smaller values of k. Do there exist "smarter" tests that are able to distinguish the null hypothesis from the alternative even if k is smaller than the typical size of the largest clique in a random graph $\mathcal{G}(n, 1/2)$?

To address this question, we derive a lower bound for the minimal risk R^*. Recall the expression (1.1) for the optimal risk. Using the Cauchy-Schwarz inequality and the fact that $\mathbb{E}_0 L(G) = 1$,

$$R^* = 1 - \frac{1}{2}\mathbb{E}_0|L(G) - 1| \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}_0(L(G)-1)^2} = 1 - \frac{1}{2}\sqrt{\mathbb{E}_0\left[L(G)^2\right] - 1} .$$

Note that for any $g \in \{0,1\}^{\binom{n}{2}}$, the likelihood ratio equals

$$\begin{aligned} L(g) &= \frac{\frac{1}{\binom{n}{k}}\sum_{S\subset V:|S|=k}\mathbb{P}_S\{G=g\}}{\mathbb{P}_0\{G=g\}} \\ &= 2^{\binom{n}{2}}\frac{1}{\binom{n}{k}}\sum_{S\subset V:|S|=k}\mathbb{1}_{\{S \text{ is a clique in } g\}}2^{-\binom{n}{2}+\binom{k}{2}} \\ &= 2^{\binom{k}{2}}\frac{1}{\binom{n}{k}}\sum_{S\subset V:|S|=k}\mathbb{1}_{\{S \text{ is a clique in } g\}} . \end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}_0\left[L(G)^2\right] &= 2^{k(k-1)}\frac{1}{\binom{n}{k}^2}\sum_{S,T\subset V:|S|=|T|=k}\mathbb{P}\{S,T \text{ are both cliques in } G\}\\ &= \frac{\sum_{i=0}^{k}\binom{k}{i}\binom{n-k}{k-i}2^{\binom{i}{2}}}{\binom{n}{k}},\end{aligned}$$

where we used the expression (1.4). Now observe that

$$\mathbb{E}_0\left[L(G)^2\right]-1 \le \frac{\sum_{i=1}^{k}\binom{k}{i}\binom{n-k}{k-i}2^{\binom{i}{2}}}{\binom{n}{k}},$$

an identical expression with that of (1.5) from the previous section. Thus, we have that $\mathbb{E}_0\left[L(G)^2\right]-1 \to 0$ whenever $k < \lfloor \omega_n - \epsilon \rfloor$.

We have shown the following lower bound, showing that detection of a hidden clique of size k is impossible by *any* test if $k < \lfloor \omega_n - \epsilon \rfloor$.

Proposition 1.6. *For any $\epsilon > 0$, if $k < \lfloor \omega_n - \epsilon \rfloor$, then the minimal possible risk of any test for the hidden clique problem has $R^* \to 1$ as $n \to \infty$.*

1.4 Finding the hidden clique: detection vs. estimation

So far we have considered the problem of *testing* whether a certain random graph has a planted clique of size k. A more challenging problem is, in fact, *finding* the planted clique—if it exists. Thus, as opposed to merely detecting the presence of a hidden clique, one wishes to identify it. One may ask how large k needs to be such that, upon observing a graph, the hidden clique may be found, with high probability. This is the so-called *estimation* problem.

More formally, upon observing a graph G, drawn from the distribution of $\mathcal{G}(n, 1/2, k)$, one is required to determine a set $\widehat{S} \subset V$, of cardinality k. The goal is that $\widehat{S}$ equals the (random) subset S of vertices over which the clique is planted. We may measure the performance of the estimator by the probability of error $\mathbb{P}_1\{\widehat{S} \neq S\}$. (Recall that $\mathbb{P}_1$ denotes the distribution of $\mathcal{G}(n, 1/2, k)$.)

Here we show that in the hidden clique problem, the estimation problem is not more difficult than the testing problem. More precisely, we prove that when $k > \lfloor \omega_n + \epsilon \rfloor$, a random graph G drawn from the distribution $\mathcal{G}(n, 1/2, k)$ has a *unique* clique of size k, with high probability. Thus, the estimator that outputs any clique of size k if such a clique exists and the empty set otherwise errs with probability converging to zero.

Theorem 1.7. *If $k > \lfloor \omega_n + \epsilon \rfloor$, then, with high probability, a random graph $G \sim \mathcal{G}(n, 1/2, k)$ has a unique clique of size k.*

Proof. Once again, we use the first moment method. Indeed, by symmetry, we may fix $S = \{1, \ldots, k\}$ as the planted clique and note that

$$\begin{aligned} &\mathbb{P}_1 \{G \text{ has two cliques of size } k\} \\ &\quad = \mathbb{P}_S \{G \text{ has a clique of size } k \text{ different from } S\} \\ &\quad \leq \mathbb{E}_S N_k^{\neq S} , \end{aligned}$$

where $\mathbb{E}_S$ denotes expectation with respect to the distribution $\mathbb{P}_S$ and $N_k^{\neq S}$ denotes the number of cliques of size k in G that are different from S. Simple counting shows that

$$\mathbb{E}_S N_k^{\neq S} = \sum_{i=0}^{k-1} \binom{k}{i} \binom{n-k}{k-i} 2^{-\binom{k}{2}+\binom{i}{2}} .$$

It suffices to show that $\mathbb{E}_S N_k^{\neq S} \to 0$ whenever $k \geq \lfloor \omega_n + \epsilon \rfloor$.

First we examine the case when k is large, say $k \geq n^{1/3}$. In this case

$$\begin{aligned} \mathbb{E}_S N_k^{\neq S} &= \sum_{i=0}^{\lfloor k/2 \rfloor} \binom{k}{i} \binom{n-k}{k-i} 2^{-\binom{k}{2}+\binom{i}{2}} + \sum_{i=\lfloor k/2 \rfloor + 1}^{k-1} \binom{k}{i} \binom{n-k}{k-i} 2^{-\binom{k}{2}+\binom{i}{2}} \\ &\leq \sum_{i=0}^{\lfloor k/2 \rfloor} k^i n^{k-i} 2^{-\binom{k}{2}+\binom{i}{2}} + \sum_{i=\lfloor k/2 \rfloor + 1}^{k-1} k^{k-i} n^{k-i} 2^{-(k-i)(k+i-1)/2} \\ &\leq \left(n 2^{-k/4+1/2}\right)^k \sum_{i=0}^{\lfloor k/2 \rfloor} \left(\frac{k}{n}\right)^i + \sum_{i=\lfloor k/2 \rfloor + 1}^{k-1} \left(k n 2^{-(k+i-1)/2}\right)^{k-i} \to 0 . \end{aligned}$$

Suppose now that $k \in [\lfloor \omega_n + \epsilon \rfloor + 1, n^{1/3}]$. Again, we split the sum as

$$\mathbb{E}_S N_k^{\neq S} = \sum_{i=0}^{\ell} \binom{k}{i} \binom{n-k}{k-i} 2^{-\binom{k}{2}+\binom{i}{2}} + \sum_{i=\ell+1}^{k-1} \binom{k}{i} \binom{n-k}{k-i} 2^{-\binom{k}{2}+\binom{i}{2}}$$

for $\ell = \lfloor (2/3) \log_2(n/2) \rfloor$. Using $\binom{n-k}{k-i} \leq \binom{n}{k} (k/(n-k))^i$, we write the first term as

$$\sum_{i=0}^{\ell} \binom{k}{i} \binom{n-k}{k-i} 2^{-\binom{k}{2}+\binom{i}{2}} \leq \binom{n}{k} 2^{-\binom{k}{2}} \sum_{i=0}^{\ell} \left(\frac{k^2 2^{(i-1)/2}}{n-k}\right)^i ,$$

which tends to zero since $\binom{n}{k} 2^{-\binom{k}{2}} \to 0$ by what is shown in Section 1.2 and because $2^{(i-1)/2} \leq (n-k)/k^2$ by our choice of ℓ and since $k \leq n^{1/3}$. It remains to bound the

second term on the right-hand side. We may write

$$\begin{aligned}\sum_{i=\ell+1}^{k-1}\binom{k}{i}\binom{n-k}{k-i}2^{-\binom{k}{2}+\binom{i}{2}} &\leq \sum_{i=\ell+1}^{k-1}\frac{k^{k-i}n^{k-i}}{(k-i)!^2}2^{-(k-i)(k+i-1)/2}\\ &\leq \sum_{i=\ell+1}^{k-1}\left(\frac{kne^2 2^{-(k+i-1)/2}}{(k-i)^2}\right)^{k-i}\\ &\quad \text{(by Stirling's formula)}\end{aligned}$$

To finish the proof, it suffices to show that the expression within the parentheses goes to zero uniformly for all $i > (2/3)\log_2(n/2)$. But this follows since for $i \geq k/2$,

$$2^{(k+i-1)/2} \geq 2^{3k/2-1} \geq \frac{n^{3/2}}{2\log^{3/2} n} \gg kn$$

and for $i \in ((2/3)\log_2(n/2), k/2)$,

$$2^{(k+i-1)/2} \geq 2^{k/2+(1/3)\log_2(n/2)-1} \geq \frac{(n/2)^{4/3}}{\log^{4/3} n} \gg \frac{kn}{(k-i)^2}\,.$$

□

1.5 Computationally efficient detection

With the limits of possible detection and estimation well understood, one may ask for efficient algorithms. The optimal (likelihood-ratio) test involves computing a sum of $\binom{n}{k}$ terms, which is clearly not computationally efficient. Also, the test used to establish upper bounds for the risk requires the computation of the largest clique in a graph (or at least checking the existence of a clique of size larger than $\omega_n \approx 2\log_2 n$). One may do this by exhaustively searching over all $\omega_n + 1$-tuples of vertices, taking time proportional to $n^{O(\log n)}$. However, finding large cliques in a computationally efficient way is a notoriously difficult problem and the challenge of constructing tests computable in polynomial time has been taken up by many researchers. However, in spite of the considerable effort invested, all known computationally efficient tests have a significantly weaker performance in the sense that they are only able to detect the presence of much larger cliques than ω_n. In what follows we survey some of the basic results in this direction, with special attention to spectral techniques that prove to be a powerful tool in a wide variety of problems in combinatorial statistics.

Let us start with the possibly simplest reasonable test that one can imagine: count the number N of edges in the observed graph. Under the null hypothesis,

N is binomially distributed, with parameters $(\binom{n}{2}, 1/2)$, while in the presence of a hidden clique on k vertices, $N \sim \binom{k}{2} + \mathrm{Bin}\left(\binom{n}{2} - \binom{k}{2}, 1/2\right)$. Thus, it is natural to define the test

$$T(G) = 0 \qquad \text{if and only if} \qquad N \leq \frac{\binom{n}{2}}{2} + \frac{\binom{k}{2}}{4} \,.$$

We may bound errors of both type using Hoeffding's inequality (see Theorem A.4 in the Appendix), and conclude that the risk of this simple test is at most

$$R(T) \leq 2e^{-(k-1)^4/(16n^2)} \,.$$

In other words, for any $\delta \in (0,1)$, one may achieve a risk not larger than δ whenever $k > 2n^{1/2}\log^{1/4}(2/\delta)$. A simple argument, using the central limit theorem, shows that this bound is not improvable in the sense that no test that only uses the total edge count can perform significantly better than T.

Not surprisingly, this test is considerably weaker than the essentially optimal test studied in Section 1.2, since the edge count ignores any structure of the graph. Indeed, the gap between the logarithmic clique size detectable by the optimal test and the bound $2n^{1/2}\log^{1/4}(2/\delta)$ achieved here appears abysmal.

It is much more surprising that no computationally efficient test is known to be able to detect cliques of size $o(n^{1/2})$. By "computationally efficient" we refer to the mild notion of computability in time polynomial in n. In what follows, we describe a spectral method that is able to shave off the $\log^{1/4}(1/\delta)$ factor from the simplistic bound described above. While this may not seem to be a remarkable achievement, this is essentially the best available bound for any computationally efficient test and the spectral techniques on which the test relies are useful in a much wider context and they are worth learning.

1.6 The spectral norm of a random graph

The spectral test we study here is based on the spectral norm of the adjacency matrix of the observed graph G. It is slightly more convenient to work with the *signed* adjacency matrix $\boldsymbol{A} = (A_{i,j})$, defined by

$$A_{i,j} = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \text{ and vertices } i \text{ and } j \text{ are joined by an edge} \\ -1 & \text{otherwise} \,. \end{cases}$$

Thus, under the null hypothesis (i.e., when $G \sim \mathcal{G}(n, 1/2)$), the random variables $(A_{i,j})_{1 \leq i < j \leq n}$ are independent symmetric sign variables. Recall that the spectral norm of $\boldsymbol{A}$ equals

$$\|\boldsymbol{A}\| = \sup_{\boldsymbol{x} \in \mathbb{S}^{n-1}} |\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x} \rangle| \,,$$

where the supremum is taken over the Euclidean unit sphere $\mathbb{S}^{n-1} = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| = 1\}$.

We show that in the presence of a sufficiently large hidden clique, $\|\boldsymbol{A}\|$ is larger, with high probability, than that under the null hypothesis. In particular, we prove the following.

Theorem 1.8. *Let $\delta \in (0,1)$ and consider the test T that accepts the null hypothesis if an only if $\|\boldsymbol{A}\| \le 4\sqrt{(\log 9)n + \log(2/\delta)}$. Then the risk of the test satisfies $R(T) \le \delta$ whenever $k > 4\sqrt{(\log 9)n + \log(2/\delta)}$.*

Thus, the spectral test achieves a better performance than counting edges. If one is merely interested in a test that has a risk at most (say) $\delta = 1/10$, then the difference between the bounds $4\sqrt{(\log 9)n + \log(1/\delta)}$ and $2n^{1/2}\log^{1/4}(2/\delta)$ is not important. (In fact, we obtain better constants for the edge-counting test.) However, if one insists on exponentially small probabilities of error then the spectral test has a superior performance. Taking $\delta = e^{-n}$ the advantage becomes clear: the spectral test is able to detect the presence of cliques of size $\Omega(\sqrt{n})$ while the edge-counting bound only gives $\Omega(n^{3/4})$. In this sense the spectral test is much more robust.

Note that the spectral norm of a matrix is computable in polynomial time and hence this test is computationally feasible though not as efficient as just counting edges.

Proof. When a clique on the vertex set S of size k is present, then we may consider the unit vector $\boldsymbol{x}_S$ with components $(x_1, \ldots, x_n)$ such that $x_i = 1/\sqrt{k}$ if $i \in S$ and $x_i = 0$ otherwise. Then

$$\|\boldsymbol{A}\| \ge \langle \boldsymbol{x}_S, \boldsymbol{A}\boldsymbol{x}_S \rangle = \sum_{i,j \in S : i \ne j} \frac{1}{k} = k - 1 \, .$$

Thus, under the alternative hypothesis, $\mathbb{P}_1\{\|\boldsymbol{A}\| \ge k - 1\} = 1$.

It remains to examine the spectral norm of $\boldsymbol{A}$ under the null hypothesis (i.e., when $G \sim \mathcal{G}(n, 1/2)$). In order to derive an upper bound, we consider a 1/4-*net* of the unit sphere $\mathbb{S}^{n-1}$, that is, a subset $\mathcal{N}$ of $\mathbb{S}^{n-1}$ of minimal size such that for all $\boldsymbol{x} \in \mathbb{S}^{n-1}$ there exists $\boldsymbol{y} \in \mathcal{N}$ with $\|\boldsymbol{x} - \boldsymbol{y}\| \le 1/4$. A simple argument based on volumes shows that $|\mathcal{N}| \le 9^n$ (see Theorem B.2 in the Appendix).

Let $\boldsymbol{x}^* \in \mathbb{S}^{n-1}$ be such that $\|\boldsymbol{A}\| = |\langle \boldsymbol{x}^*, \boldsymbol{A}\boldsymbol{x}^* \rangle|$ and let $\boldsymbol{y} \in \mathcal{N}$ be such that $\|\boldsymbol{x}^* - \boldsymbol{y}\| \le 1/4$.

$$\begin{aligned} \|\boldsymbol{A}\| &= |\langle \boldsymbol{y}, \boldsymbol{A}\boldsymbol{y} \rangle + \langle \boldsymbol{x}^* - \boldsymbol{y}, \boldsymbol{A}\boldsymbol{x}^* \rangle + \langle \boldsymbol{y}, \boldsymbol{A}(\boldsymbol{x}^* - \boldsymbol{y}) \rangle| \\ &\le |\langle \boldsymbol{y}, \boldsymbol{A}\boldsymbol{y} \rangle| + 2\|\boldsymbol{A}\| \cdot \|\boldsymbol{x}^* - \boldsymbol{y}\| \\ &\quad \text{(by the Cauchy-Schwarz inequality)} \\ &\le |\langle \boldsymbol{y}, \boldsymbol{A}\boldsymbol{y} \rangle| + \frac{1}{2}\|\boldsymbol{A}\| \, . \end{aligned}$$

Thus, $\|\boldsymbol{A}\| \leq 2\max_{\boldsymbol{y}\in\mathcal{N}} |\langle \boldsymbol{y}, \boldsymbol{A}\boldsymbol{y}\rangle|$ and therefore it suffices to bound the maximum over the finite 1/4-net.

For any fixed $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbb{S}^{n-1}$ and $t > 0$,

$$\begin{aligned}
\mathbb{P}_0\{|\langle \boldsymbol{y}, \boldsymbol{A}\boldsymbol{y}\rangle| > t\} &= \mathbb{P}_0\left\{\left|\sum_{1\leq i<j\leq n} A_{i,j} y_i y_j\right| > \frac{t}{2}\right\} \\
&\leq 2\exp\left(\frac{-t^2}{8\sum_{1\leq i<j\leq n} y_i^2 y_j^2}\right) \quad \text{(by Hoeffding's inequality–Theorem A.4)} \\
&\leq 2\exp\left(-\frac{t^2}{4}\right),
\end{aligned}$$

where we used the fact that $\|\boldsymbol{y}\| = 1$ and therefore

$$\sum_{1\leq i<j\leq n} y_i^2 y_j^2 = \frac{1}{2}\left(\sum_{i,j\in[n]} y_i^2 y_j^2 - \sum_{i\in[n]} y_i^4\right) \leq \frac{1}{2}\sum_{i,j\in[n]} y_i^2 y_j^2 = \frac{1}{2}\,.$$

Thus, using the union bound, we obtain that

$$\mathbb{P}_0\{\|\boldsymbol{A}\| > t\} \leq 9^n \max_{\boldsymbol{y}\in\mathcal{N}} \mathbb{P}_0\{|\langle \boldsymbol{y}, \boldsymbol{A}\boldsymbol{y}\rangle| > t/2\} \leq 2\cdot 9^n e^{-t^2/16} = \delta \tag{1.9}$$

when $t = 4\sqrt{(\log 9)n + \log(2/\delta)}$, as desired. □

The constants appearing in the theorem are not optimal. In fact, under the null hypothesis, $\|\boldsymbol{A}\|$ is known to converge to $2\sqrt{n}$ by a celebrated result of Füredi and Komlós [34]. However, the elementary proof shown here provides non-asymptotic bounds with exponential inequalities and the techniques extend to a wider class of random matrices—see Vershynin [72] for a survey.

The value of the constant can always be improved at the price of increasing the computational cost. In fact, if one has an algorithm that is able to detect hidden cliques of size $c\sqrt{n + \log(1/\delta)}$ with a probability of error at most $1/\delta$, one may obtain another algorithm able to detect hidden cliques of size $c\sqrt{n/2 + \log(n/\delta)}$ by multiplying the computational complexity by $O(\sqrt{n}\log(1/\delta))$. Indeed one may pick a random vertex v. If v happens to belong to the hidden clique, then running the algorithm on the subgraph induced by the vertices adjacent to v, the clique will correctly be detected. This subgraph has about half of the size of the original graph, hence the improvement in the constant. If the algorithm fails to detect, one may repeat the procedure with another randomly drawn vertex. After $(n/k)\log(1/\delta)$ random vertices, one hits the hidden clique with probability at least $1 - \delta$.

Spectral methods may be used not only for detecting the presence of a hidden clique but also to identify it, provided that the clique is sufficiently large. Such an algorithm is discussed in the next section.

1.7 A spectral algorithm for finding the hidden clique

In this section we show that spectral methods may also be used not only to detect but also to find hidden cliques. A simple method considers the eigenvector corresponding to the largest eigenvalue of the signed adjacency matrix $\boldsymbol{A}$ introduced in the previous section. In other words, we consider the random vector $\boldsymbol{V} \in \mathbb{S}^{n-1}$ defined by

$$\boldsymbol{V} = \operatorname*{argmax}_{\boldsymbol{x} \in \mathbb{S}^{n-1}} \langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x} \rangle \ .$$

As it is pointed out in the previous section, if the vertex set of the hidden clique is S, the value of the quadratic form $\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x} \rangle$ is quite large for the vector $\boldsymbol{x} = \boldsymbol{x}_S$ (whose components are $x_i = 1/\sqrt{k}$ if $i \in S$ and $x_i = 0$ otherwise). Thus, one might expect that the components of $\boldsymbol{V}$ tend to be large within S and small outside of S. Indeed, in what follows, we confirm this intuition and establish a simple procedure based on the largest entries of the eigenvector $\boldsymbol{V}$. Since $\boldsymbol{V}$ may be computed in polynomial time, the following simple algorithm is computationally efficient:

(1) Let $T \subset \{1, \dots, n\}$ be the set of vertices corresponding to the k largest entries of $\boldsymbol{V} = (V_1, \dots, V_n)$.

(2) Let $\widehat{S} \subset \{1, \dots, n\}$ be the subset of vertices of size k that have the highest number of adjacent vertices in T.

We have the following performance guarantee.

Theorem 1.10. *Let $\delta \in (0,1)$ and let $S \subset \{1, \dots, n\}$ have cardinality k. Assume that a random graph is generated according to $\mathbb{P}_S$, that is, with S as a planted clique. If $k \geq 25\left(8\sqrt{(\log 9)n + \log(4/\delta)} + 1\right)$, then*

$$\mathbb{P}\left\{\widehat{S} \neq S\right\} \leq \delta \ .$$

The key of the proof is to show that T contains a large fraction of the vertices of the hidden clique. In order to show this, first we prove that $\boldsymbol{V}$ puts most of its "weight" on the components in S. In a second step we prove that this weight is roughly uniformly distributed in S. The following two lemmas make these statements rigorous.

Let $\boldsymbol{U}$ be the vector whose components are

$$U_i = \begin{cases} V_i & \text{if } i \in S \\ 0 & \text{otherwise.} \end{cases}$$

Lemma 1.11. *Assume* $k \geq 25\left(8\sqrt{(\log 9)n + \log(4/\delta)} + 1\right)$. *Consider the decomposition* $\boldsymbol{V} = \boldsymbol{U} + \boldsymbol{W}$. *Then, with probability at least* $1 - \delta/2$, $\|\boldsymbol{W}\| \leq 1/5$.

Proof. By the defining optimality property of $\boldsymbol{V}$,

$$\begin{aligned} k - 1 &= \langle \boldsymbol{x}_S, \boldsymbol{A}\boldsymbol{x}_S \rangle \\ &\leq \langle \boldsymbol{V}, \boldsymbol{A}\boldsymbol{V} \rangle \\ &= \langle \boldsymbol{U}, \boldsymbol{A}\boldsymbol{U} \rangle + 2\langle \boldsymbol{U}, \boldsymbol{A}\boldsymbol{W} \rangle + \langle \boldsymbol{W}, \boldsymbol{A}\boldsymbol{W} \rangle \\ &\leq k\|\boldsymbol{U}\|^2 + 2\langle \boldsymbol{U}, \boldsymbol{A}\boldsymbol{W} \rangle + \langle \boldsymbol{W}, \boldsymbol{A}\boldsymbol{W} \rangle \end{aligned} \tag{1.12}$$

where we used the fact that

$$\langle \boldsymbol{U}, \boldsymbol{A}\boldsymbol{U} \rangle = \left(\sum_{i=1}^{k} U_i \right)^2 \leq k\|\boldsymbol{U}\|^2$$

by Cauchy-Schwarz. Define the symmetric random matrix $\boldsymbol{A}^{(0)} = \left(A^{(0)}_{i,j}\right)_{n \times n}$ such that $A^{(0)}_{i,j} = A_{i,j}$ whenever at least one of i and j is not in S, $A^{(0)}_{i,i} = 0$ for $i \in S$ and all other entries $A^{(0)}_{i,j}$ for $i, j \in S$, $i < j$ are independent symmetric sign variables. Note that $\boldsymbol{A}^{(0)}$ is the signed adjacency matrix of a $\mathcal{G}(n, 1/2)$ random graph. Since the entries of $\boldsymbol{W}$ with index in S are zero, $\boldsymbol{A}\boldsymbol{W} = \boldsymbol{A}^{(0)}\boldsymbol{W}$, and therefore (1.12) equals

$$\begin{aligned} & k\|\boldsymbol{U}\|^2 + 2\left\langle \boldsymbol{U}, \boldsymbol{A}^{(0)}\boldsymbol{W} \right\rangle + \left\langle \boldsymbol{W}, \boldsymbol{A}^{(0)}\boldsymbol{W} \right\rangle \\ &\leq k\|\boldsymbol{U}\|^2 + \left\|\boldsymbol{A}^{(0)}\right\| \left(2\|\boldsymbol{U}\| \cdot \|\boldsymbol{W}\| + \|\boldsymbol{W}\|^2\right) \\ &\leq k\left(1 - \|\boldsymbol{W}\|^2\right) + 2\left\|\boldsymbol{A}^{(0)}\right\| . \end{aligned}$$

Rearranging the obtained inequality, we have

$$\|\boldsymbol{W}\|^2 \leq \frac{2\left\|\boldsymbol{A}^{(0)}\right\| + 1}{k} .$$

Recalling (1.9) from the previous section, we see that the right-hand side is at most $1/25$ with probability at least $1 - \delta/2$ whenever $k \geq 25\left(8\sqrt{(\log 9)n + \log(4/\delta)} + 1\right)$. □

The next lemma shows that the components of $\boldsymbol{V}$ that belong to S are approximately even.

Lemma 1.13. *Assume* $k \geq 25\left(8\sqrt{(\log 9)n + \log(4/\delta)} + 1\right)$. *Let* $\boldsymbol{V} = \boldsymbol{U} + \boldsymbol{W}$ *be as in Lemma 1.11 and write* $\boldsymbol{U} = \alpha \boldsymbol{x}_S + \boldsymbol{R}$ *where* $\alpha \in \mathbb{R}$ *and* $\langle \boldsymbol{x}_S, \boldsymbol{R} \rangle = 0$. *On the same event of probability at least* $1 - \delta/2$ *as in Lemma 1.11,* $\alpha^2 \geq 24/25$ *and* $\|\boldsymbol{R}\| \leq 1/5$.

Proof. Recall from the proof of Lemma 1.11 that

$$\begin{aligned}\langle \boldsymbol{U}, \boldsymbol{A}\boldsymbol{U}\rangle &\geq k - 1 - \left(2\left\langle \boldsymbol{U}, \boldsymbol{A}^{(0)}\boldsymbol{W}\right\rangle + \left\langle \boldsymbol{W}, \boldsymbol{A}^{(0)}\boldsymbol{W}\right\rangle\right) \\ &\geq k - 1 - 2\left\|\boldsymbol{A}^{(0)}\right\| \\ &\geq \frac{24}{25}k\,.\end{aligned}$$

On the other hand, by the orthogonality of $\boldsymbol{R}$ and $\boldsymbol{x}_S$, $\sum_{i=1}^k R_i = 0$ and therefore

$$\begin{aligned}\langle \boldsymbol{U}, \boldsymbol{A}\boldsymbol{U}\rangle &= \alpha^2 \langle \boldsymbol{x}_S, \boldsymbol{A}\boldsymbol{x}_S\rangle + 2\alpha \langle \boldsymbol{R}, \boldsymbol{A}\boldsymbol{x}_S\rangle + \langle \boldsymbol{R}, \boldsymbol{A}\boldsymbol{R}\rangle \\ &= \alpha^2 k + 2\alpha \frac{k-1}{\sqrt{k}} \sum_{i=1}^k R_i + \left(\sum_{i=1}^k R_i\right)^2 \\ &= \alpha^2 k\,.\end{aligned}$$

Thus, $\alpha^2 \geq 24/25$. On the other hand,

$$\|\boldsymbol{R}\|^2 = \|\boldsymbol{U}\|^2 - \alpha^2 \leq \frac{1}{25}$$

as claimed. □

Now we are prepared to prove Theorem 1.10.

Proof. Lemmas 1.11 and 1.13 imply that, with probability at least $1 - \delta/2$, $\boldsymbol{V} = \alpha \boldsymbol{x}_S + \boldsymbol{R} + \boldsymbol{W}$ where $\alpha^2 \geq 24/25$, $\|\boldsymbol{W}\| \leq 1/5$ and also $\|\boldsymbol{R}\| \leq 1/5$. This implies that the number of indices $i \in \{1, \ldots, n\}$ such that $|W_i| > 1/(3\sqrt{k})$ is at most $3k/25$ and the same holds for $\boldsymbol{R}$. Thus, $|T \cap S| \geq 19k/25$, that is, at least a fraction of $19/25$ of the vertices that correspond to the k largest components of $\boldsymbol{V}$ belong to the hidden clique S.

It remains to show that, with high probability, the subset $\widehat{S}$ of vertices that have the highest number of adjacent vertices in T equals S. To this end, simply note that on an event of probability at least $1 - \delta/2$, each vertex in S is adjacent to at least $19k/25$ vertices in T. On the other hand, if $i \notin S$, then the number of edges between i and T is at most $6k/25$ plus the number of edges between i and vertices in S. Thus, by the union bound and Hoeffding's inequality,

$$\begin{aligned}\mathbb{P}\left\{\widehat{S} \neq S\right\} &\leq \frac{\delta}{2} + n\mathbb{P}\left\{\frac{6k}{25} + \mathrm{B}in(k, 1/2) \geq \frac{19k}{25}\right\} \\ &\leq \frac{\delta}{2} + n\mathbb{P}\left\{\mathrm{B}in(k, 1/2) \geq \frac{13k}{25}\right\} \\ &\leq \frac{\delta}{2} + ne^{-k^2/1250} < \delta\,.\end{aligned}$$

□

1.8 Bibliographic notes

The systematic study of random graphs such as $\mathcal{G}(n,p)$ originates with the work of Erdős and Rényi [29, 30]. The sharp bounds for the clique number cited in the text are due to Matula [54]. We refer the reader to Palmer [59], Bollobás [13], Janson, Łuczak, and Ruciński [41], Chung and Lu [20], Durrett [27] and van der Hofstad [70] for monographs dedicated to random graphs.

The hidden clique problem dates back at least to Jerrum [42] and Kučera [48]. It was Alon, Krivelevich, and Sudakov [2] who showed how spectral methods can be used to find a hidden clique of size proportional to $n^{1/2}$. Efficient non-spectral algorithms that work in the same regime we introduced by Feige and Ron [32] and Dekel, Gurel-Gurevich and Peres [21]. The latter paper introduces a particularly simple method that reconstructs the hidden clique with computational complexity of optimal order ($O(n^2)$). Deshpande and Montanari [22] show that cliques of size $\sqrt{n/e}(1+o(1))$ can be recovered, with high probability, by an algorithm of nearly optimal complexity. This is the best known constant achievable to date.

For more on spectral algorithms, we refer to Kannan and Vempala [46] and Vershynin [72].

Quite some effort has been invested in trying to prove that it is impossible to find hidden cliques of size $o(\sqrt{n})$ with computationally efficient methods. Progress has been made in this direction by restricting the class of allowed algorithms. For a sample of such results, see Meka, Potechin, and Wigderson [55], Montanari, Reichman, and Zeitouni [57] Barak, Hopkins, Kelner, Kothari, Moitra, and Potechin [9], and Feldman, Grigorescu, Reyzin, Vempala, and Xiao [33]. For an interesting "semi-random" version of the problem, see Steinhardt [65].

1.9 Exercises

Exercise 1.1. *Consider the simple hypothesis testing problem described in Section 1.1. Prove that the likelihood ratio test minimizes the risk among all tests and prove that the minimal risk equals*

$$R^* = 1 - \frac{1}{2}\sum_g |\mathbb{P}_1\{G=g)\} - \mathbb{P}_0\{G=g)\}| = 1 - \frac{1}{2}\mathbb{E}_0|L(G)-1|\,.$$

Exercise 1.2. *Finish the calculations of the proof of Matula's theorem for the two-point concentration of the clique number of $\mathcal{G}(n,1/2)$. Namely, show that for all $\epsilon > 0$,*

$$\binom{n}{m} 2^{-\binom{m}{2}} \to 0$$

whenever $m > \lfloor \omega_n + \epsilon \rfloor$ *and that*

$$\frac{\sum_{i=1}^{m} \binom{m}{i}\binom{n-m}{m-i}2^{\binom{i}{2}}}{\binom{n}{m}} \to 0$$

whenever $m < \lfloor \omega_n - \epsilon \rfloor$*, where* $\omega_n = 2\log_2 n - 2\log_2 \log_2 n - 1 + 2\log_2 e$ *(Matula [54], Palmer [59]).*

Exercise 1.3. (HIDDEN DENSE SUBGRAPH.) *Generalize the hidden clique problem to the case when the hidden subgraph is not a clique of size* k*, but rather a random graph* $\mathcal{G}(k, q)$ *for some* $q > 1/2$*.*

Exercise 1.4. (HIDDEN CLIQUE IN $\mathcal{G}(n, p)$.) *Generalize the hidden clique problem to the case when a clique of size* k *is hidden in a random graph* $\mathcal{G}(n, p)$ *for some* $q \neq 1/2$*.*

Exercise 1.5. *Consider the hidden clique problem with* $k \geq c\sqrt{n \log n}$*. Show that the clique may be recovered, with high probability, by only looking at the degrees of the vertices.*

Chapter 2

Combinatorial testing problems

2.1 Problem formulation

In this chapter we study a general class of hypothesis testing problems with a combinatorial flavor. One observes an n-dimensional vector $\boldsymbol{X} = (X_1, \ldots, X_n)$. Under the null hypothesis the components of $\boldsymbol{X}$ are independent standard normal random variables. As usual, we denote the probability measure and expectation under the null hypothesis by $\mathbb{P}_0$ and $\mathbb{E}_0$, respectively.

To describe the alternative hypothesis, consider a class $\mathcal{C} = \{S_1, \ldots, S_N\}$ of N sets of indices such that $S_j \subset \{1, \ldots, n\}$ for all $j = 1, \ldots, N$. Under the alternative hypothesis there exists a set of indices $S \in \mathcal{C}$ such that

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

where $\mu > 0$ is a positive parameter. The components of $\boldsymbol{X}$ are independent under the alternative hypothesis as well. The probability measure of $\boldsymbol{X}$ 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 $|S| = k$.

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

$$R(T) = \mathbb{P}_0\{T(\boldsymbol{X}) = 1\} + \frac{1}{N} \sum_{S \in \mathcal{C}} \mathbb{P}_S\{T(\boldsymbol{X}) = 0\} \,. \tag{2.1}$$

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 $\boldsymbol{X}$ belonging

to S have mean μ. Similarly to Chapter 1, we refer to the first and second terms on the right-hand side of (2.1) 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, 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*.

In some interesting examples, the n components of $\boldsymbol{X}$ 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.

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

- the set of all subsets $S \subset \{1, \ldots, n\}$ of size k;
- the set of all cliques of a given size in a complete graph—this is the *Gaussian hidden clique problem*, a Gaussian version of the problem studied in detail in Chapter 1.
- 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.

Regardless of what $\mathcal{C}$ is, one may determine explicitly the test T^* minimizing the risk. This follows from a simple generalization of Exercise 1.1: for a given vector $\boldsymbol{x} = (x_1, \ldots, x_n)$, $T^*(\boldsymbol{x}) = 1$ if and only if the ratio of the likelihoods of $\boldsymbol{x}$ under $(1/N)\sum_{S \in \mathcal{C}} \mathbb{P}_S$ and $\mathbb{P}_0$ exceeds 1. Writing

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

and

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

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

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

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

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

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

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

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 (Exercise 2.1). 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.

2.2 Simple tests: averanging and scan statistics

Even though the test T^* minimizing the risk is explicitly determined, its performance is not always easy to analyze. Moreover, efficient computation of the optimal test is often a non-trivial problem though efficient algorithms are available in many interesting cases.

Here we consider two simple, though suboptimal, tests. 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 $\boldsymbol{X}$ 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*

$$T(\boldsymbol{x}) = \mathbb{1}_{\{\sum_{i=1}^n X_i > \mu k/2\}} \,.$$

Proposition 2.2. *Let $\delta > 0$. The risk of the averaging test T satisfies $R(T) \le \delta$ 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 $\mathcal{N}(\mu k, n)$. By a Gaussian tail bound (see (A.1) in the Appendix), we have $R(T) \le 2e^{-(\mu k)^2/(8n)}$. □

A test based on scan statistics

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 $\mathcal{N}(\mu k, k)$. Consider the *scan-statistic test*

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

The test statistic $\max_{S \in \mathcal{C}} X_S$ is often referred to as a *scan statistic*. Here we only need the following simple observation.

Proposition 2.3. *The risk of the maximum test T satisfies $R(T) \le \delta$ whenever*

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

Proof. 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 $\boldsymbol{X}$ with Lipschitz constant $\sqrt{k}$, by the Gaussian concentration inequality of Theorem A.11 in the Appendix, for all $t > 0$,

$$\mathbb{P}_0 \left\{ \max_{S \in \mathcal{C}} X_S \ge \mathbb{E}_0 \max_{S \in \mathcal{C}} X_S + t \right\} \le e^{-t^2/(2k)} .$$

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

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

and therefore

$$\mathbb{P}_S \left\{ \max_{S \in \mathcal{C}} X_S \le \mu k - t \right\} \le e^{-t^2/(2k)} ,$$

which completes the proof. □

The scan statistics is often easier to compute than the optimal test T^*, 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 T by any upper bound and then the same upper bound will appear in the performance bound.

Proposition 2.3 shows that the scan-statistic 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 scan-statistic 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 results available for expected maxima.

Here we simply note that if $|\mathcal{C}| = N$, by Theorem B.3 in the Appendix, one always has

$$\mathbb{E}_0 \max_{S\in\mathcal{C}} X_S \le \sqrt{2k\log N}\,.$$

2.3 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

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

It is well known (see Exercise 2.2) 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,

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

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

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

and therefore we have the following.

Proposition 2.4. *For all classes $\mathcal{C}$, $R^* \ge 1/2$ whenever $\mu \le \sqrt{(4/k)\log(4/3)}$.*

This shows that no matter what the class $\mathcal{C}$ is, detection is hopeless if μ is of the order of $k^{-1/2}$.

A lower bound based on overlapping pairs

Proposition 2.5. *Let S and S' be drawn independently, uniformly, at random from $\mathcal{C}$ and let $Z = |S \cap S'|$. Then*

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

Proof. In order to lower bound the optimal risk, we use the Cauchy–Schwarz inequality, exactly the way we did in Section 1.3:

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

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

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

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

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

But

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

and the statement follows. □

This proposition reduces the problem of obtaining lower bounds 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.

A lower bound for symmetric classes

First we derive a simple corollary of Proposition 2.5 under a general symmetry condition on the class $\mathcal{C}$. It shows that the universal bound of Proposition 2.4 may be improved by a factor of $\sqrt{\log(1 + n/k)}$ as soon as the class $\mathcal{C}$ is sufficiently symmetric.

Proposition 2.6. *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 2.5. 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 $S' = \{1, \ldots, k\}$ so we can write $Z = \sum_{i=1}^{k} \mathbb{1}_{\{i \in S\}}$. By Hölder's inequality,

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

Proposition 2.5 now implies the statement. □

The lower bound of Proposition 2.6 is matched, in order of magnitude, by the scan-statistic test when the class $\mathcal{C}$ is "small." Indeed, if $|\mathcal{C}| \leq n^\alpha$ for some $\alpha > 0$, then Proposition 2.3 and the fact that $\mathbb{E}_0 \max_{S \in \mathcal{C}} X_S \leq \sqrt{2k \log |\mathcal{C}|}$ imply that $R^* \leq 1/2$ for all μ with

$$\mu \geq \sqrt{\frac{2\alpha}{k} \log n} + \sqrt{\frac{8 \log 4}{k}} \,.$$

Note that for symmetric classes, Proposition 2.6 implies that $R^* \geq 1/2$ for all μ with

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

Negative association

The bound of Proposition 2.6 may be improved significantly under an additional condition of negative association that is satisfied in several interesting examples. A collection $Y_1, \ldots, Y_n$ of random variables is *negatively associated* if for any pair of

disjoint sets $I, J \subset \{1, \dots, n\}$ and (coordinate-wise) non-decreasing functions f and g,

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

Proposition 2.7. *Let $\delta \in (0,1)$ and assume that the class $\mathcal{C}$ satisfies the conditions of Proposition 2.6. Suppose that the labels are such that $S' = \{1, 2, \dots, k\} \in \mathcal{C}$. Let S be a randomly chosen element of $\mathcal{C}$. If the random variables $\mathbb{1}_{\{1 \in S\}}, \dots, \mathbb{1}_{\{k \in S\}}$ 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. We proceed similarly to the proof of Proposition 2.6. We have

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

Proposition 2.5 and the upper bound above imply that R^* is at least δ for all μ such that

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

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

Lower bounds for the maximum of a Gaussian process

We finish this section by pointing out an interesting by-product of our lower bounds.

For any class $\mathcal{C}$ of subsets of $[n]$ of size k, Proposition 2.3 implies that $R^* < 1/2$ whenever

$$\mu > \frac{\mathbb{E}_0 \max_{S \in \mathcal{C}} X_S}{k} + 4\sqrt{\frac{1}{k}} .$$

On the other hand, Proposition 2.5 implies that $R^* \geq 1/2$ whenever $\mathbb{E}\exp(\mu^2 Z) \leq 2$. Combining these two bounds, we see that

$$\mathbb{E}_0 \max_{S \in \mathcal{C}} X_S \geq k\mu_{\mathcal{C}} - 4\sqrt{k} ,$$

where $\mu_{\mathcal{C}}$ is defined by the equation $\mathbb{E}\exp(\mu_{\mathcal{C}}^2 Z) = 2$.

Thus, lower bounds for $\mathbb{E}_0 \max_{S\in\mathcal{C}} X_S$ may be proved by bounding the moment generating function of Z from above. In particular, Propositions 2.6 and 2.7 imply that, if $\mathcal{C}$ is a symmetric class (as defined in Proposition 2.6), then

$$\mathbb{E}_0 \max_{S\in\mathcal{C}} X_S \geq \sqrt{k \log\left(1 + \frac{n}{k}\right)} - 4\sqrt{k}\,.$$

Moreover, if $\mathcal{C}$ has the negative association property of Proposition 2.7, then

$$\mathbb{E}_0 \max_{S\in\mathcal{C}} X_S \geq k\sqrt{\log\left(1 + \frac{n}{k^2}\log 2\right)} - 4\sqrt{k}\,.$$

The behavior of $\mathbb{E}_0 \max_{S\in\mathcal{C}} X_S$ is well understood (at least up to a constant factor) in terms of the geometry of the underlying indexing set $\mathcal{C}$ equipped with the canonical distance

$$d(S,T) = \sqrt{\mathbb{E}\left[(X_S - X_T)^2\right]} = \sqrt{2(k - |S\cap T|)}\,,$$

see Talagrand [67, 68]. However, the general characterization is often difficult to apply in concrete examples. The lower bounds obtained here, though not always sharp, are often easy to use.

2.4 Examples

Here we list a few concrete examples and work out upper and lower bounds for the critical range of μ.

2.4.1 k-sets

Consider first the example when $\mathcal{C}$ contains all sets $S \subset \{1,\dots,n\}$ of size k, without any combinatorial structure. Thus, $N = \binom{n}{k}$.

Let $\delta \in (0,1)$. It is easy to see that the assumptions of Proposition 2.7 are satisfied (Exercise 2.3) 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 2.2 in the two different regimes. Since

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

we see from Proposition 2.3 that when $k = O\left(n^{(1-\epsilon)/2}\right)$ for some fixed $\epsilon > 0$, 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 2.7 above is of the order $\sqrt{n/k^2}$ and the averaging test provides a matching upper bound by Proposition 2.2.

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

2.4.2 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 2.2 (i.e., the averaging test), for $\delta \in (0,1)$, one has $R(T) \le \delta$ whenever $\mu \ge \sqrt{8\log(2/\delta)}$.

To show that this bound has the right order of magnitude, we may apply Proposition 2.7. 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, $R^* \ge \delta$ whenever

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

Note that in this case the optimal test T^* 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 $\{1,\ldots,m\}$) is equivalent to computing the permanent of an $m \times m$ matrix with non-negative elements. By a deep result of Jerrum, Sinclair, and Vigoda [43], this may be done by a polynomial-time randomized approximation.

2.4.3 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 $1, 2, \ldots, n = \binom{m}{2}$ represent the edges of the complete graph K_m and let $\mathcal{C}$ be the set of all spanning trees of K_m. Thus, by Cayley's formula, we have $N = m^{m-2}$ spanning trees and $K = m - 1$. By Proposition 2.2, the averaging test has risk $R(T) \le \delta$ whenever $\mu \ge \sqrt{4\log(2/\delta)}$.

This bound is indeed of the right order. To see this, we may start with Proposition 2.5. Even though Proposition 2.7 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 [31] (see also Grimmett and Winkler [36] and Benjamini, Lyons, Peres, and Schramm [11]), 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 $Z = |S \cap S'|$,

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

This, together with Proposition 2.5 shows that for any $\delta \in (0,1)$, $R^* \ge \delta$ whenever

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

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 T^*. The "partition function" $(1/N)\sum_{S\in\mathcal{C}} e^{\mu X_S}$ may be computed by an algorithm of Propp and Wilson [61], who introduced a random sampling algorithm that, given a graph with non-negative 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.

2.4.4 Gaussian hidden clique problem

An interesting example is a "Gaussian" variant of the hidden clique problem discussed in Chapter 1. Here the random variables $X_1, \ldots, X_n$ are associated with the edges of the complete graph K_m such that $\binom{m}{2} = n$ and $\mathcal{C}$ contains all cliques of size k. Thus, $k = \binom{k}{2}$ and $N = \binom{m}{k}$.

We have the following bounds for the performance of the optimal test. It shows that when k is at 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 may be adjusted to handle larger values of k as well.

Proposition 2.8. *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 2.3 and the bound $\mathbb{E}_0 \max_{S\in\mathcal{C}} X_S \le \sqrt{2k\log\binom{m}{k}}$.

To prove the lower bound (ii), by Proposition 2.5, 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 $\mathbb{E}\left[\exp(\mu^2 Z)\right] \le 2$ for the indicated values of μ.

Because of symmetry, $\mathbb{E}\left[\exp(\mu^2 Z)\right] = \mathbb{E}\left[\exp(\mu^2 Z)|S'\right]$ 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, Hoeffding's inequality for the hypergeometric distribution (see Theorem A.6 in the Appendix) implies that

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

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

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

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

$$\mathbb{E}\left[\exp\left(\mu^2 c\frac{k^2}{m}B\right)\right]\cdot\mathbb{E}\left[\exp\left(\mu^2 kB\mathbb{1}_{\left\{B>c\frac{k^2}{m}\right\}}\right)\right] \le 4\,. \tag{2.9}$$

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 (2.9) 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 $2^{1/k}-1 \ge (\log 2)/k$, this is implied by

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

To bound the second term on the left-hand side of (2.9), note that

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

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

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

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

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

On the other hand,

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

and therefore the second term on the left-hand side of (2.9) is at most 2 whenever

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

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

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

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

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

Choosing

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

(which is greater than 1 for $k \le \sqrt{m(\log 2)/e}$), 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 $a > 0$, $h(c-1) \ge c\log a - a + 1$. Choosing $a = \frac{\log(m/k)}{\log(m\log 2/k^2)}$, the first term is at least

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

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

The proposition above implies that, if, say, $\mu = 1$, then k needs to be at least of the order of $\log m$ in order to achieve a small risk, and that $\log m$ is of the optimal order. However, the test that achieves this performance is the scan statistic that needs to compute X_S for all clicks S of size k. Similarly to the hidden clique problem of Chapter 1, there remains a big gap between the performance of computationally efficient tests and the optimal test. Spectral techniques may be used here as well, in a completely analogous way to improve on the performance of the averaging test. When $\mu = 1$, a simple test based on the largest eigenvalue of the matrix of weights achieves a small risk when $k = \Omega(\sqrt{m})$. We leave the details as an exercise.

2.5 Bibliographic remarks

The general hypothesis testing problem studied in this chapter was introduced by Arias-Castro, Candès, Helgason and Zeitouni [7]. In that paper 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 [6] investigate another class of examples in which elements of $\mathcal{C}$ correspond to clusters in a regular grid. The problem when $\mathcal{C}$ contains all subsets of size k has been studied in the rich literature on multiple testing, see, for example, Ingster [40], Baraud [10], Donoho and Jin [26] and the references therein.

The study of maxima of Gaussian processes has been a central topic in probability theory. We refer to Talagrand [68] for the culmination of a long line of research.

Proposition 2.5 is due to Arias-Castro, Candès, Helgason, and Zeitouni [7].

The material presented in this chapter is mostly based on Addario-Berry, Broutin, Devroye, and Lugosi [1].

For more on the Gaussian hidden clique problem, in particular, on the limitation of spectral methods, we refer the reader to Montanari, Reichman, and Zeitouni [57]. In Section 5.3 we address a related problem introduced by Kannan and Vempala [47].

A sample of related litarature, includes Sun and Nobel [66], Butucea and Ingster [19], Balakrishnan, Kolar, Rinaldo, Singh, and Wasserman [8].

An interesting variant is the "sparse principal component detection" problem in which one tests whether a multivariate vector is isotropic or if it has a sparse principal component. Berthet and Rigollet [12] study this problem and show that computationally efficient near-optimal detection is only possible it the hidden clique problem can be solved in polynomial time—that is considered quite unlikely.

2.6 Exercises

Exercise 2.1. *Prove that for all classes $\mathcal{C}$, the Bayes risk R^* is a monotone decreasing function of μ.*

Exercise 2.2. *Consider a testing problem in which, under the null hypothesis, the observation $\mathcal{X}$ has density f_0, while under the alternative, f_1. Let $L(x) = f_1(x)/f_0(x)$ be the likelihood ratio. Let $R^* = 1 - (1/2)\mathbb{E}_0|L(X) - 1|$ be the risk of the optimal test and define*

the Bhattacharyya measure of affinity as $\rho = (1/2)\mathbb{E}_0\sqrt{L(\boldsymbol{X})}$*. Prove that*

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

(see, e.g., [23, Theorem 3.1]).

Exercise 2.3. *Prove that the class* $\mathcal{C}$ *of* k*-sets satisfies the negative association condition of Proposition 2.7.*

Exercise 2.4. *Construct and analyze a spectral test for the Gaussian hidden clique problem. Proceed in a way analogous to Section 1.7.*

Chapter 3

Detection of correlations and high-dimensional random geometric graphs

In statistics and signal processing on often faces problems in which one is asked to detect the presence of a sparse signal in a noisy environment. In this chapter we discuss a simple stylized model of such detection problems. The model naturally motivates the study of random geometric graphs in high-dimensional spaces. This leads to some intriguing mathematical questions that we study in detail.

3.1 Detection of correlations

Consider the following simple hypothesis testing problem. Upon observing random vectors $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$, each of d independent components, one wishes to test whether these vectors are independent or, alternatively, if there exists a small group of vectors that depend on each other. We write $\boldsymbol{X} = (\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)$ for the $d \times n$ matrix of observations. In remote sensing the n vectors represent the signal captured at n sensors in a noisy environment and one wishes to determine if there is a subset of the sensors that detect a common weak signal. In financial applications the n vectors represent the evolution of the price of n assets and one may be interested in the existence of a small subset that depend on each other in a certain way.

The simplest way to formalize such a hypothesis testing problem is the following. Under the *null hypothesis*, all vectors $\boldsymbol{X}_i$ are standard normal (i.e., with mean $\mathbf{0}$ and unit covariance matrix). Under the *alternative hypothesis* there is a small subset of vectors that are more correlated among themselves. This may be modeled as follows. Under the alternative hypothesis, there exists a set S of indices of a given size $|S| = k \leq n$, belonging to a class $\mathcal{C}$ of subsets of $[n]$ such that

$\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,d})$ where

$$X_{i,t} = \begin{cases} Y_{i,t} & \text{if } i \notin S, t \in [d] \\ \sqrt{\rho}\, N_t + \sqrt{1-\rho}\, Y_{i,t} & \text{if } i \in S, t \in [d] \end{cases} \tag{3.1}$$

where $(Y_{i,t})_{i\in[n],t\in[d]}, (N_t)_{t\in[d]}$ are independent standard normal sequences.

The N_t represent a common "signal" present in the vectors $\boldsymbol{X}_i$ for $i \in S$. Clearly, $X_{i,t}$ is standard normal for all i and t and $\mathbb{E}X_{i,t}X_{j,t} = 0$ if either i or j are not in S. If $i, j \in S$, then $\mathbb{E}X_{i,t}X_{j,t} = \rho$.

The problem becomes interesting when ρ is so small that calculating simply the correlation of $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ one cannot easily tell whether both i and j are in S or not. In particular, the largest empirical correlations $(\boldsymbol{X}_i, \boldsymbol{X}_j)$ do not necessarily belong to indices belonging to S. The interesting values of ρ are those for which the "signal" is covered by "noise". Clearly, if d is sufficiently large, the problem becomes easy but in the applications mentioned above it is important to keep the value of d as small as possible in order to make quick decisions.

Similarly to the previous sections, a test is a binary-valued function $T : \mathbb{R}^{nd} \to \{0, 1\}$ that, upon observing the random matrix $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, accepts the null hypothesis if and only if $T(\boldsymbol{X}) = 0$. The risk of a test T is measured by the sum of tipe I and type II errors

$$R(T) = \mathbb{P}_0\{T(\boldsymbol{X}) = 1\} + \frac{1}{|\mathcal{C}|}\sum_{S\in\mathcal{C}} \mathbb{P}_S\{T(\boldsymbol{X}) = 0\}\,,$$

where $\mathbb{P}_0$ is the probability distribution under the null hypothesis and $\mathbb{P}_S$ is the probability distribution when S is the set of indices of the correlated vectors.

As in the hypothesis testing problems discussed in Chapters 1 and 2, the test T^* minimizing the risk is the likelihood ratio test $T^* = \mathbb{1}_{\{L(\boldsymbol{X})>0\}}$, where L is the likelihood ratio, that is, the ratio of the densities of $\boldsymbol{X}$ under the alternative, and null hypotheses.

We start by describing a general lower bound for the optimal risk, analogous to Proposition 2.5 in Chapter 2.

Theorem 3.2. *Let $\boldsymbol{N} = (N_1, \ldots, N_d)$ be a standard normal vector in $\mathbb{R}^d$. For any $a > 0$,*

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

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

$$R^* \geq \frac{1}{2} - \frac{1}{4}\sqrt{\mathbb{E}\exp(d\nu Z) - 1}\,,$$

where $\nu = \rho/(1+\rho) - \frac{1}{2}\log(1-\rho^2)$. *If* $\rho \le 1/2$, *we have* $\nu \le \rho$, *and therefore*

$$R^* \ge \frac{1}{2} - \frac{1}{4}\sqrt{\mathbb{E}\exp(d\rho Z) - 1}\,.$$

The strategy one may try to bound the optimal risk R^* from below is the one that we successfully applied in Chapters 1 and 2. There we bounded R^* using the Cauchy-Schwarz inequality as

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

However, this strategy fails in the present case as $L(\boldsymbol{X})$ is not square integrable under the distribution of the null hypothesis. The proof of the theorem applies a simple conditioning trick before the Cauchy-Schwarz inequality.

Proof. Fix a value of the "signal" vector $\boldsymbol{N}$ at $\boldsymbol{N} = \boldsymbol{u} = (u_1, \ldots, u_d) \in \mathbb{R}^d$. We consider now the alternative hypothesis with this value fixed. Let $R(T)$, L, T^* and $R_{\boldsymbol{u}}(T)$, $L_{\boldsymbol{u}}$, $T^*_{\boldsymbol{u}}$ be the risk of a test T, the likelihood ratio, and the optimal test, for the original and "conditional" hypothesis testing problems. For any $\boldsymbol{u} \in \mathbb{R}^d$, $R_{\boldsymbol{u}}(T^*_{\boldsymbol{u}}) \le R_{\boldsymbol{u}}(T^*)$ by the optimality of $T^*_{\boldsymbol{u}}$. Therefore, conditioning on $\boldsymbol{N} = \boldsymbol{u}$,

$$\begin{aligned} R^* &= R(T^*) \\ &= \mathbb{E}_{\boldsymbol{N}} R_{\boldsymbol{N}}(T^*) \\ &\ge \mathbb{E}_{\boldsymbol{N}} R_{\boldsymbol{N}}(T^*_{\boldsymbol{N}}) \\ &= 1 - \frac{1}{2}\mathbb{E}_{\boldsymbol{N}}\mathbb{E}_0|L_{\boldsymbol{N}}(\boldsymbol{X}) - 1|\,. \end{aligned}$$

($\mathbb{E}_{\boldsymbol{N}}$ denotes expectation with respect to $\boldsymbol{N} \sim \mathcal{N}(0, \boldsymbol{I})$.) Using the fact that $\mathbb{E}_0|L_{\boldsymbol{u}}(\boldsymbol{X}) - 1| \le 2$ for all $\boldsymbol{u} \in \mathbb{R}^d$, we have (with $B(0,a)$ being the Euclidean ball centered at the origin and of radius a in $\mathbb{R}^d$),

$$\mathbb{E}_{\boldsymbol{N}}\mathbb{E}_0|L_{\boldsymbol{N}}(\boldsymbol{X}) - 1| \le 2\mathbb{P}\{\|\boldsymbol{N}\| > a\} + \mathbb{P}\{\|\boldsymbol{N}\| \le a\} \max_{\boldsymbol{u} \in B(0,a)} \mathbb{E}_0|L_{\boldsymbol{u}}(\boldsymbol{X}) - 1|\,,$$

and therefore, using the Cauchy-Schwarz inequality,

$$\begin{aligned} 1 - \frac{1}{2}\mathbb{E}_{\boldsymbol{N}}\mathbb{E}_0|L_{\boldsymbol{N}}(\boldsymbol{X}) - 1| &\ge \mathbb{P}\{\|\boldsymbol{N}\| \le a\}\left(1 - \frac{1}{2}\max_{\boldsymbol{u} \in B(0,a)} \mathbb{E}_0|L_{\boldsymbol{u}}(\boldsymbol{X}) - 1|\right) \\ &\ge \mathbb{P}\{\|\boldsymbol{N}\| \le a\}\left(1 - \frac{1}{2}\max_{\boldsymbol{u} \in B(0,a)} \sqrt{\mathbb{E}_0 L^2_{\boldsymbol{u}}(\boldsymbol{X}) - 1}\right). \end{aligned}$$

Since

$$\begin{aligned}
&L_{\boldsymbol{u}}(\boldsymbol{x}) \\
&\quad = \frac{1}{|\mathcal{C}|}\sum_{S\in\mathcal{C}}\frac{1}{(1-\rho)^{dk/2}}\exp\left(-\sum_{t=1}^{d}\sum_{i\in S}\frac{(x_{i,t}-\sqrt{\rho}u_t)^2}{2(1-\rho)}-\sum_{t=1}^{d}\sum_{i\notin S}\frac{x_{i,t}^2}{2}\right)\exp\left(\sum_{t=1}^{d}\sum_{i=1}^{n}\frac{x_{i,t}^2}{2}\right) \\
&\quad = \frac{1}{|\mathcal{C}|}\sum_{S\in\mathcal{C}}\frac{1}{(1-\rho)^{dk/2}}\exp\left(\sum_{t=1}^{d}\sum_{i\in S}\frac{x_{i,t}^2}{2}-\frac{(x_{i,t}-\sqrt{\rho}u_t)^2}{2(1-\rho)}\right),
\end{aligned}$$

we get

$$\begin{aligned}
&\mathbb{E}_0 L_{\boldsymbol{u}}^2(\boldsymbol{X}) \\
&= \frac{1}{|\mathcal{C}|^2}\sum_{S,S'\in\mathcal{C}}\frac{1}{(1-\rho)^{dk}}\mathbb{E}_0\exp\left(\sum_{t=1}^{d}\sum_{i\in S\cap S'}X_{i,t}^2-\frac{(X_{i,t}-\sqrt{\rho}u_t)^2}{1-\rho}+\sum_{t=1}^{d}\sum_{i\in S\Delta S'}\frac{X_{i,t}^2}{2}-\frac{(X_{i,t}-\sqrt{\rho}u_t)^2}{2(1-\rho)}\right) \\
&= \frac{1}{|\mathcal{C}|^2}\sum_{S,S'\in\mathcal{C}}\frac{1}{(1-\rho)^{dk}(2\pi)^{dn/2}} \\
&\qquad \times\int_{-\infty}^{+\infty}\exp\left(\sum_{t=1}^{d}\sum_{i\in S\cap S'}\frac{x_{i,t}^2}{2}-\frac{(x_{i,t}-\sqrt{\rho}u_t)^2}{1-\rho}-\sum_{t=1}^{d}\sum_{i\in S\Delta S'}\frac{(x_{i,t}-\sqrt{\rho}u_t)^2}{2(1-\rho)}-\sum_{t=1}^{d}\sum_{i\notin S\cup S'}\frac{x_{i,t}^2}{2}\right)d\boldsymbol{x}\,.
\end{aligned}$$

It is easy to check that

$$\frac{x_{i,t}^2}{2}-\frac{(x_{i,t}-\sqrt{\rho}u_t)^2}{1-\rho}=\frac{\rho u_t^2}{1+\rho}-\frac{1+\rho}{2(1-\rho)}\left(x_{i,t}-\frac{2\sqrt{\rho}u_t}{1+\rho}\right)^2,$$

which implies

$$\begin{aligned}
&\mathbb{E}_0 L_{\boldsymbol{u}}^2(\boldsymbol{X}) \\
&= \frac{1}{|\mathcal{C}|^2}\sum_{S,S'\in\mathcal{C}}\frac{\exp\left(\sum_{t=1}^{d}\frac{\rho u_t^2}{1+\rho}|S\cap S'|\right)}{(1-\rho)^{dk}(2\pi)^{dn/2}} \\
&\quad \times\int_{-\infty}^{+\infty}\exp\left(-\sum_{t=1}^{d}\sum_{i\in S\cap S'}\frac{1+\rho}{2(1-\rho)}\left(x_{i,t}-\frac{2\sqrt{\rho}u_t}{1+\rho}\right)^2-\sum_{t=1}^{d}\sum_{i\in S\Delta S'}\frac{(x_{i,t}-\sqrt{\rho}u_t)^2}{2(1-\rho)}-\sum_{t=1}^{d}\sum_{i\notin S\cup S'}\frac{x_{i,t}^2}{2}\right)d\boldsymbol{x} \\
&= \frac{1}{|\mathcal{C}|^2}\sum_{S,S'\in\mathcal{C}}\frac{\exp\left(\sum_{t=1}^{d}\frac{\rho u_t^2}{1+\rho}|S\cap S'|\right)}{(1-\rho)^{dk}}\left(\frac{1-\rho}{1+\rho}\right)^{d|S\cap S'|/2}(1-\rho)^{d(k-|S\cap S'|)} \\
&\le \frac{1}{|\mathcal{C}|^2}\sum_{S,S'\in\mathcal{C}}\exp\left(\left(\sum_{t=1}^{d}\frac{\rho u_t^2}{1+\rho}-\frac{d}{2}\log(1-\rho^2)\right)|S\cap S'|\right),
\end{aligned}$$

concluding the proof of the first statement. Finally, observe that since $\log(1-\rho^2) \geq -\rho^2/(1-\rho^2)$,

$$\nu \leq \rho \frac{1-\rho/2}{1-\rho^2} \leq \rho \quad \text{when } \rho \leq 1/2 \,.$$

□

Interestingly, the moment generating function of the same random variable $Z = |S \cap S'|$ appears in the lower bound of Theorem 3.2 as in Proposition 2.5. Thus, all the work done in Chapter 2 for bounding the moment generating function of Z in various examples may be re-used in the context of this chapter. In order to avoid repetitions, here and in the rest of this chapter we only consider the case when $\mathcal{C}$ is the class of all k-sets, that is, all subsets of $[n]$ of cardinality k. As it is pointed out in Section 2.4, in this case

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

and therefore, for $\rho \leq 1/2$, we have

$$R^* \geq \frac{1}{2} - \frac{1}{4} \sqrt{\exp\left(\left(\left(e^{d\rho} - 1\right) \frac{k^2}{n} \right) \right) - 1} \,. \tag{3.3}$$

One may read off several interesting corollaries from this bound. For example, it is instructive to set $d = (c_n/\rho) \log n$ for some $c_n \to 0$. In that case we see that it is impossible to have $R^* \to 0$ unless k is at least of the order of $n^{1/2 - o(1)}$.

The situation dramatically changes for just slightly larger values of d. In fact, if $d \geq (9/\rho) \log n$, then a simple scan-statistic test has a risk converging to zero as soon as $k > 8/\rho + 1$. This test is based on the test statistic

$$\max_{S \in \mathcal{C}} \sum_{i,j \in S : i \neq j} \left\langle \boldsymbol{X}_i, \boldsymbol{X}_j \right\rangle \,. \tag{3.4}$$

Observe that, under the null hypothesis, for all $S \in \mathcal{C}$, $\mathbb{E}_0 \sum_{i,j \in S : i \neq j} \left\langle \boldsymbol{X}_i, \boldsymbol{X}_j \right\rangle = 0$, while under the alternative hypothesis $\mathbb{E}_S \sum_{i,j \in S : i \neq j} \left\langle \boldsymbol{X}_i, \boldsymbol{X}_j \right\rangle = dk(k-1)\rho/2$. Thus, it is natural to define a test that accepts the null hypothesis if and only if the scan statistic define above is less than $dk(k-1)\rho/4$. One may analyze this test by establishing concentration inequalities for random quadratic forms of the type $\sum_{i,j \in S : i \neq j} \left\langle \boldsymbol{X}_i, \boldsymbol{X}_j \right\rangle$. The details are left as an exercise.

3.2 A high-dimensional random geometric graph

A natural approach of constructing tests for the correlation-detection problem defined in the previous section is based on the simple observation that, while under the null hypothesis, we have $\mathbb{E}_0\left\langle \boldsymbol{X}_i, \boldsymbol{X}_j \right\rangle = 0$, under the alternative hypothesis, the empirical correlations $\left\langle \boldsymbol{X}_i, \boldsymbol{X}_j \right\rangle$ tend to be larger if both i and j are in the "contaminated" set S. Thus, one may construct a graph on the vertex set $[n]$ in which vertices i and j are connected by an edge if and only if their empirical correlations are large. In such a graph one expects that a large clique appears under the alternative hypothesis.

One way of formalizing this idea is by considering the random geometric graph defined by the normalized vectors $\boldsymbol{Z}_i = \boldsymbol{X}_i/\|\boldsymbol{X}_i\|$. Fix some $p \in (0, 1/2)$ and define the random geometric graph based on the points $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$, connecting vertex i and vertex j if and only if $\left\langle \boldsymbol{Z}_i, \boldsymbol{Z}_j \right\rangle \geq t_{p,d}$, where $t_{p,d}$ is a threshold value chosen such that, under the null hypothesis (i.e., when the $\boldsymbol{X}_i$ are independent standard normal vectors),

$$\mathbb{P}\left\{\left\langle \boldsymbol{Z}_i, \boldsymbol{Z}_j \right\rangle \geq t_{p,d}\right\} = p \, .$$

In other words, vertices i and j are connected if and only if the empirical correlation $\left\langle \boldsymbol{Z}_i, \boldsymbol{Z}_j \right\rangle$ of the observed vectors $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ exceeds the threshold $t_{p,d}$.

A possible test is based on computing the clique number of the obtained graph. One expects that, under the alternative hypothesis, for sufficiently large values of ρ, vertices belonging to S form a clique.

In order to understand the behavior of such a test, we need to examine the behavior of the clique number of the random graph just defined. In the remainder of this chapter we consider the null hypothesis.

Recall that the unit sphere in $\mathbb{R}^d$ is denoted by $\mathbb{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$ where $\|\cdot\|$ stands for the Euclidean norm. Under the null hypothesis, $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ are independent random vectors, uniformly distributed in $\mathbb{S}^{d-1}$. We denote the components of $\boldsymbol{Z}_i$ by $(Z_{i,1}, \ldots, Z_{i,d})$.

For a given value of $p \in (0, 1)$ (possibly depending on n and d) the *random geometric graph* $\overline{\mathcal{G}}(n, d, p)$ is defined on the vertex set $[n]$ as above: vertex i and vertex j are connected by an edge if an only if

$$\left\langle \boldsymbol{Z}_i, \boldsymbol{Z}_j \right\rangle \geq t_{p,d} \, .$$

Equivalently, vertex i and vertex j are connected if and only if $\|\boldsymbol{Z}_i - \boldsymbol{Z}_j\| \leq \sqrt{2(1 - t_{p,d})}$.

For example, for $p = 1/2$, $t_{p,d} = 0$. To understand the behavior of $t_{p,d}$ as a function of p, we introduce some notation. Let μ_{d-1} denote the uniform probability

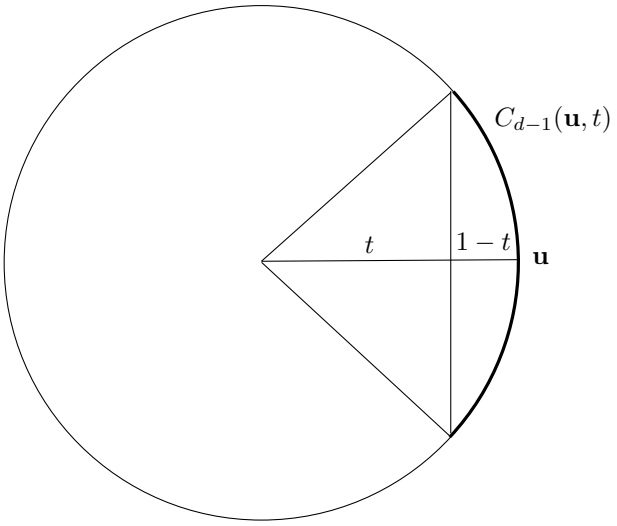

Figure 3.1: A spherical cap of height $1-t$.

measure over $\mathbb{S}^{d-1}$. For a unit vector $\boldsymbol{u} \in \mathbb{S}^{d-1}$ and real number $0 \le t \le 1$, let $C_{d-1}(\boldsymbol{u}, t) = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x} \in \mathbb{S}^{d-1}, (\boldsymbol{x}, \boldsymbol{u}) \ge t\}$ denote a spherical cap of height $1-t$ around $\boldsymbol{u}$ (see Figure 3.1). The *angle* of a spherical cap $C_{d-1}(\boldsymbol{u}, t)$ is defined by $\arccos(t)$.

Then $p = \mu_{d-1}(C_{d-1}(\mathbf{1}, t_{p,d}))$ is the normalized surface area of a spherical cap of height $1 - t_{p,d}$ centered at (say) the first standard basis vector $\mathbf{1} = (1, 0, 0, \ldots, 0)$.

Often it is useful to think about random points on $\mathbb{S}^{d-1}$ as projections of Gaussian vectors on the unit sphere. In particular, if $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent standard normal vectors, then the vectors

$$\boldsymbol{Z}_i = \frac{\boldsymbol{X}_i}{\|\boldsymbol{X}_i\|}, \quad i \in [n]$$

are independent and uniformly distributed on $\mathbb{S}^{d-1}$. This representation may be used to determine the asymptotic value of $t_{p,d}$. Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be a standard Gaussian vector and let $\boldsymbol{Z} = \boldsymbol{X}/\|\boldsymbol{X}\| = (Z_1, \ldots, Z_d)$. Observe that $\mathbb{E}\|\boldsymbol{X}\|^2 = d$. Also, by the law of large numbers, $\|\boldsymbol{X}\|/\sqrt{d} \to 1$ in probability. This implies that $Z_1\sqrt{d}$ converges, in distribution, to a standard normal random variable. In fact, for any fixed k, the joint distribution of $\sqrt{d}(Z_1, \ldots, Z_k)$ is asymptotically standard normal. One consequence of this is that for any $s > 0$,

$$\mu_{d-1}(C_{d-1}(\mathbf{1}, s/\sqrt{d})) = \mathbb{P}\{Z_1 > s/\sqrt{d}\} = \mathbb{P}\{X_1/\|\boldsymbol{X}\| > s/\sqrt{d}\} \to 1 - \Phi(s)$$

as $d \to \infty$ where $\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-t^2/2} dt$. This implies that $t_{p,d}$ satisfies, for any

fixed $p \in (0,1)$,

$$\lim_{d\to\infty} t_{p,d}\sqrt{d} = \Phi^{-1}(1-p)\,. \tag{3.5}$$

Thus, if $p < 1/2$ is fixed and d is large, $t_{p,d}$ is of the order of $1/\sqrt{d}$.

Of course, this asymptotic result may be sharpened. For example, we have, for $\sqrt{2/d} \le t_{p,d} \le 1$,

$$\frac{1}{6t_{p,d}\sqrt{d}}(1-t_{p,d}^2)^{\frac{d-1}{2}} \le p \le \frac{1}{2t_{p,d}\sqrt{d}}(1-t_{p,d}^2)^{\frac{d-1}{2}} \tag{3.6}$$

(see Brieden et al. [15]).

To simplify the presentation, from now on we fix $p = 1/2$. In this case, $t_{p,d} = 0$.

3.3 The clique number

In this section we study the clique number $\omega(n,d)$ of the d-dimensional random geometric graph $\overline{\mathcal{G}}(n,d,1/2)$. In particular, we are interested in the dependence of $\omega(n,d)$ on the dimension d.

If d is fixed as we let n grow, the largest clique has linear size. This follows from observing that if k points fall in any spherical cap C of height $1-1/\sqrt{2}$, then they are mutually connected and therefore form a clique. The expected number of points that fall in any such fixed cap C is $n\mu_{d-1}(C)$ which, by (3.6) is at least

$$\frac{n}{6}\sqrt{\frac{2}{d}}2^{-\frac{d-1}{2}}\,. \tag{3.7}$$

On the other hand, when n is fixed and d grows, the clique number is logarithmic in n. Indeed, one may show that, when n and p are fixed, then the d-dimensional random geometric graph $\overline{\mathcal{G}}(n,d,p)$ converges, in distribution, to the Erdős-Rényi random graph $\mathcal{G}(n,p)$, as $d\to\infty$. This may be seen, for example, by using the fact that, by the multivariate central limit theorem,

$$\left(\frac{1}{\sqrt{d}}\left\langle \boldsymbol{X}_i, \boldsymbol{X}_j\right\rangle\right)_{1\le i<j\le n} \Longrightarrow \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}_{\binom{n}{2}}\right) \qquad \text{as } d\to\infty, \text{ in distribution.}$$

(The details are left to the reader. We prove a qualitative version of this statement in Chapter 4.)

As a consequence of this, we have that $\omega(n,d)$ converges, in distribution, to the clique number of an Erdős-Rényi random graph $\mathcal{G}(n,1/2)$. Recall from Section 1.2 that this clique number is sharply concentrated around $\omega_n = 2\log_2 n - 2\log_2\log_2 n - 1 + 2\log_2 e$.

Thus, as d grows, from a constant value to infinity, the clique number $\omega(n,d)$ decreases from $\Omega(n)$ to roughly $2\log_2 n$.

A natural question—in particular, having the correlation detection problem in mind—is how the clique number behaves when d grows as a function of n. In what follows we present an upper bound and a lower bound. The upper bound implies that, when $d \gg \log^3 n$, then the clique number is $\omega(n,d) = (2+o_p(1))\log_2 n$, that is, it is essentially at its asymptotic value. On the other hand, we prove that if $d \geq 2\log((4\log 2)n)$, then, with probability at least 1/2,

$$\omega(n,d) \geq (1/32)\exp\left(\log^2(n)/(5d)\right),$$

and therefore, if d is proportional to $\log n$, the clique number is at least some positive power of n and for $d \sim \log^{2-\epsilon} n$, the clique number is still much larger than any power of $\log n$. Note also that, when $d = o(\log n)$, then (3.7) implies that the expected clique number is $n^{1-o(1)}$, that is, it grows almost linearly in n.

The proof of the upper bound is based on the first-moment method. All we need is a good upper bound for the expected number of cliques of size k.

Denote the number of cliques of size k in $\overline{\mathcal{G}}(n,d,1/2)$ by $N_k = N_k(n,d)$. Since

$$\mathbb{E}N_k = \binom{n}{k}\mathbb{P}\{\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_k \text{ form a clique}\},$$

it suffices to study the probability that k points are all connected with each other. Let $p_k = \mathbb{P}\{\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_k \text{ form a clique}\}$ denote this probability.

The heart of the argument is the following lemma.

Lemma 3.8. *Let $k \geq 2$ be a positive integer, let $\delta_n > 0$, and assume*

$$d \geq \frac{8(k+1)^2\log 2}{\delta_n^2}\left(k\log 8 + \log\frac{k-1}{2}\right).$$

Then

$$p_k \leq e\cdot\Phi(\delta_n)^{\binom{k}{2}}.$$

Recall that in a $\mathcal{G}(n,1/2)$ Erdős-Rényi graph, the probability that k vertices form a clique equals $2^{-\binom{k}{2}}$. Since $\Phi(0) = 1/2$, the lemma shows that, when δ_n is small, p_k takes a similar form.

Proof. Fix a $\ell \leq k$. We use the Gaussian representation of the $\boldsymbol{Z}_i$, writing $\boldsymbol{Z}_i = \boldsymbol{X}_i/\|\boldsymbol{X}_i\|$ where $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ are independent standard normal vectors in $\mathbb{R}^d$. First we perform Gram-Schmidt orthogonalization for $\boldsymbol{X}_1^{\ell-1} = \boldsymbol{X}_1,\ldots,\boldsymbol{X}_{\ell-1}$. In other words, let

$$\boldsymbol{v}_1 = \frac{\boldsymbol{X}_1}{\|\boldsymbol{X}_1\|}$$

and define $\boldsymbol{r}_1 = \mathbf{0}$ (the d-dimensional zero vector). For $j = 2, \ldots, \ell - 1$, introduce, recursively,

$$\boldsymbol{r}_j = \sum_{i=1}^{j-1} \left\langle \boldsymbol{X}_j, \boldsymbol{v}_i \right\rangle \boldsymbol{v}_i \quad \text{and} \quad \boldsymbol{v}_j = \frac{\boldsymbol{X}_j - \boldsymbol{r}_j}{\|\boldsymbol{X}_j - \boldsymbol{r}_j\|} .$$

Then $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{\ell-1}$ are orthonormal vectors, depending on $\boldsymbol{X}_1^{\ell-1}$ only.

Introduce the "bad" event

$$B_{\ell-1} = \left\{ \exists j \le \ell - 1 : \|\boldsymbol{r}_j\|^2 > 2(\ell+1)^2 \log 2 \text{ or } \exists j \le \ell - 1 : \|\boldsymbol{X}_j - \boldsymbol{r}_j\|^2 < \frac{d}{2} \right\} .$$

and write

$$\begin{aligned} p_\ell &\le \mathbb{P}\{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_\ell \text{ form a clique}, B_{\ell-1}^c\} + \mathbb{P}\{B_{\ell-1}\} \\ &= \mathbb{E}\left[\mathbb{P}\left\{ \left\langle \boldsymbol{X}_\ell, \boldsymbol{X}_j \right\rangle \ge 0 \text{ for all } j \le \ell - 1 \middle| \boldsymbol{X}_1^{\ell-1} \right\} \mathbb{1}_{\{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{\ell-1} \text{ form a clique}\}} \mathbb{1}_{\{B_{\ell-1}^c\}} \right] \\ &\quad + \mathbb{P}\{B_{k-1}\} . \end{aligned} \tag{3.9}$$

Now fix $\boldsymbol{X}_1^{\ell-1}$ such that $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{\ell-1}$ form a clique and $B_{\ell-1}$ does not occur. Then, using $\boldsymbol{X}_j = \boldsymbol{v}_j \|\boldsymbol{X}_j - \boldsymbol{r}_j\| + \boldsymbol{r}_j$ and the union bound, we have, for any $\delta_n > 0$,

$$\begin{aligned} &\mathbb{P}\left\{ \left\langle \boldsymbol{X}_\ell, \boldsymbol{X}_j \right\rangle \ge 0 \text{ for all } j \le \ell - 1 \middle| \boldsymbol{X}_1^{\ell-1} \right\} \\ &\le \mathbb{P}\left\{ \left\langle \boldsymbol{X}_\ell, \boldsymbol{v}_j \right\rangle \ge -\delta_n \text{ for all } j \le \ell - 1 \middle| \boldsymbol{X}_1^{\ell-1} \right\} + \sum_{j=1}^{\ell-1} \mathbb{P}\left\{ \left\langle \boldsymbol{X}_\ell, \frac{\boldsymbol{r}_j}{\|\boldsymbol{X}_j - \boldsymbol{r}_j\|} \right\rangle > \delta_n \middle| \boldsymbol{X}_1^{\ell-1} \right\} . \end{aligned}$$

Since on $B_{\ell-1}^c$, we have $\|\boldsymbol{X}_j - \boldsymbol{r}_j\| \ge \sqrt{d/2}$, for any $1 \le j \le \ell - 1$, on this event we have

$$\begin{aligned} \mathbb{P}\left\{ \left\langle \boldsymbol{X}_\ell, \frac{\boldsymbol{r}_j}{\|\boldsymbol{X}_j - \boldsymbol{r}_j\|} \right\rangle > \delta_n \middle| \boldsymbol{X}_1^{\ell-1} \right\} &\le \mathbb{P}\left\{ \left\langle \boldsymbol{X}_\ell, \boldsymbol{r}_j \right\rangle > \delta_n \sqrt{d/2} \middle| \boldsymbol{X}_1^{\ell-1} \right\} \\ &\le \frac{1}{2} e^{-\frac{\delta_n^2 d}{4\|\boldsymbol{r}_j\|^2}} \le \frac{1}{2} e^{-\frac{\delta_n^2 d}{8(\ell+1)^2 \log 2}} , \end{aligned} \tag{3.10}$$

where we used the fact that, conditionally on $\boldsymbol{X}_1^{\ell-1}$, $\left\langle \boldsymbol{X}_\ell, \boldsymbol{r}_j \right\rangle$ has centered normal distribution with variance $\|\boldsymbol{r}_j\|^2 \le 2(\ell+1)^2 \log 2$. Furthermore,

$$\mathbb{P}\left\{ \left\langle \boldsymbol{X}_\ell, \boldsymbol{v}_j \right\rangle \ge -\delta_n \text{ for all } j \le \ell - 1 \middle| \boldsymbol{X}_1^{\ell-1} \right\} = \Phi(\delta_n)^{\ell-1} ,$$

where we used the fact that by rotational invariance of the multivariate standard normal distribution, the $(\boldsymbol{X}_\ell, \boldsymbol{v}_1), \ldots, (\boldsymbol{X}_\ell, \boldsymbol{v}_{\ell-1})$ are independent standard normal random variables. Therefore, the first term in (3.9) may be bounded as

$$\begin{aligned} &\mathbb{E}\left[\mathbb{P}\left\{ \left\langle \boldsymbol{X}_\ell, \boldsymbol{X}_j \right\rangle \ge 0 \text{ for all } j \le \ell - 1 \middle| \boldsymbol{X}_1^{\ell-1} \right\} \mathbb{1}_{\{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{\ell-1} \text{ form a clique }\}} \mathbb{1}_{\{B_{\ell-1}^c\}} \right] \\ &\qquad \le p_{\ell-1} \left(\Phi(\delta_n)^{\ell-1} + \frac{\ell - 1}{2} e^{-\frac{\delta_n^2 d}{8(\ell+1)^2 \log 2}} \right) . \end{aligned} \tag{3.11}$$

The last term within the parentheses above may be bounded by 8^{-k} using

$$\delta_n^2 \geq \frac{8(k+1)^2 \log 2}{d} \left(k \log 8 + \log \frac{k-1}{2} \right) .$$

Thus, (3.11) is bounded from above by

$$p_{\ell-1} \left(\Phi(\delta_n)^{\ell-1} + 8^{-k} \right) \leq p_{\ell-1} \left(1 + 2^{-3k+\ell} \right) \Phi(\delta_n)^{\ell-1} . \tag{3.12}$$

To bound the probability of the "bad" event $B_{\ell-1}$, note that, since $\boldsymbol{r}_j$ is a projection of $\boldsymbol{X}_j$ onto the subspace spanned by $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{j-1}$, $\|\boldsymbol{X}_j - \boldsymbol{r}_j\| \leq \|\boldsymbol{X}_j\|$, and therefore

$$\begin{aligned} \mathbb{P}\{B_{\ell-1}\} &\leq \mathbb{P}\left\{ \exists j \leq \ell - 1 : \|\boldsymbol{r}_j\|^2 > 2(\ell+1)^2 \log 2 \right\} + \mathbb{P}\left\{ \exists j \leq \ell - 1 : \|\boldsymbol{Z}_j\|^2 < \frac{d}{2} \right\} \\ &\leq (\ell-1)\mathbb{P}\left\{ \chi^2_{\ell-1} > 2(\ell+1)^2 \log 2 \right\} + (\ell-1)\mathbb{P}\left\{ \chi^2_d < \frac{d}{2} \right\} , \end{aligned}$$

where χ^2_d denotes a random variable with χ^2 distribution with d degrees of freedom. Both terms may be bounded by standard tail bounds for the χ^2 distribution, see Theorem A.7 in the Appendix. In particular, for the second term we obtain $\mathbb{P}\{\chi^2_d < d/2\} \leq e^{-d/16}$. The first term can be bounded as

$$\mathbb{P}\{\chi^2_{\ell-1} > 2(\ell+1)^2 \log 2\} \leq e^{-2(\ell+1)^2(\log 2)/4} = 2^{-(\ell+1)^2/2} .$$

Thus

$$\mathbb{P}\{B_{\ell-1}\} \leq (\ell-1)\left(2^{-(\ell+1)^2/2} + e^{-d/16} \right) .$$

Since by assumption $d \geq 8(\ell+1)^2 \log 2$, we obtain

$$\mathbb{P}\{B_{\ell-1}\} \leq 2(\ell-1)2^{-(\ell+1)^2/2} ,$$

and so, summarizing, we have

$$p_\ell \leq p_{\ell-1} \left(1 + 2^{-3k+\ell} \right) \Phi(\delta_n))^{\ell-1} + 2(\ell-1)^{-(\ell+1)^2/2} . \tag{3.13}$$

From this, we deduce, by induction, that

$$p_\ell \leq \Phi(\delta_n)^{\binom{\ell}{2}} \prod_{j=1}^{\ell-1} (1 + 2^{-j-1/2}) . \tag{3.14}$$

This concludes the proof since $\prod_{j=1}^{\ell} (1 + 2^{-j-1/2}) \leq e^{\sum_{j=1}^{\ell} 2^{-j-1/2}} < e$.

To prove (3.14), note first that it trivially holds for $\ell = 1$. Assuming it holds for $\ell - 1$ for some $\ell \geq 2$, from (3.13) we obtain

$$\begin{aligned} p_\ell &\leq \Phi(\delta_n)^{\binom{\ell-1}{2}} \left(\prod_{j=1}^{\ell-2} (1 + 2^{-j-1/2}) \right) \left(1 + 2^{-3k+\ell}\right) \Phi(\delta_n)^{\ell-1} + 2(\ell-1)2^{-(\ell+1)^2/2} \\ &\leq \Phi(\delta_n)^{\binom{\ell}{2}} \left(\prod_{j=1}^{\ell-2} (1 + 2^{-j-1/2}) \right) \left(1 + 2^{-3k+\ell} + 2(\ell-1)2^{-\frac{3\ell+1}{2}}\right) \\ &\leq \Phi(\delta_n)^{\binom{\ell}{2}} \prod_{j=1}^{\ell-1} (1 + 2^{-j-1/2}) \end{aligned}$$

where we used $2^{-3k+\ell} + 2(\ell-1)2^{-\frac{3\ell+1}{2}} < 2^{-\ell+1/2}$ for $k \geq 2$ since $2(\ell-1)2^{-\ell/2} \leq 3/2$ for all ℓ. This completes the proof of (3.14). □

Now it is easy to deduce upper bounds for the clique number by a simple application of the first-moment method. As expected, as the dimension grows, the clique number behaves more and more like in the case of the Erdős-Rényi graph. When $d \sim \log^5 n$, the difference by the clique numbers is bounded by a constant, with high probability.

Theorem 3.15. *The following statements hold with high probability.*

$$\begin{array}{llll} \textit{If} & d \geq 12500 \log^3 n, & \textit{then} & \omega(n,d) \leq 5\log_2 n + 1 \ ; \\ \textit{if} & d/\log^3 n \to \infty, & \textit{then} & \omega(n,d) \leq (2+o(1))\log_2 n + 1 \ ; \\ \textit{if} & \liminf d/\log^5 n > 0, & \textit{then} & \omega(n,d) \leq 2\log_2 n - 2\log_2\log_2 n + O(1) \ . \end{array}$$

Proof. Denote the (random) number of cliques of size k by N_k. Then

$$\mathbb{P}\{\omega(n,d) \geq k\} = \mathbb{P}\{N_k \geq 1\} \leq \mathbb{E}N_k \ .$$

Now we may use Lemma 3.8.

For example, if $\delta_n = 1/2$, the the lemma implies that for $d \geq 100k^3$,

$$\mathbb{E}N_k \leq e \cdot \binom{n}{k} \Phi(1/2)^{\binom{k}{2}} \leq e \cdot \binom{n}{k} (7/10)^{\binom{k}{2}} \leq e \cdot \left(n \left(\frac{7}{10}\right)^{\frac{k-1}{2}} \right)^k \ ,$$

which converges to zero when $k > 2\log_{10/7} n + 1 \leq 5\log n + 1$.

The second and third statements follow from the same inequality. We leave the calculations as an exercise. □

Finally, we derive a lower bound for the clique number. The rough idea is the following. If the clique number of $\overline{\mathcal{G}}(n,d,1/2)$ was small, then we could construct a test in the correlation detection problem of Section 3.1 based on computing the clique number. Since under the alternative hypothesis, large cliques are likely to appear, this would lead to a test with a small risk. However, we may contradict this by invoking the lower bound of Theorem 3.2 for the risk on *any* test.

To formalize these ideas, first we show that, under the alternative hypothesis (i.e., when there is a set S of k indices such that the pairwise correlations within the set equal to $\mathbb{E}_S X_{i,t}, X_{j,t} = \rho$ for all $i, j \in S$, $i \neq j$ and $t \in [d]$), then a clique of size k is likely to appear in the random geometric graph when ρ is sufficiently large. Recall the definition of (3.1).

Lemma 3.16. *Let $\delta \in (0,1)$ and consider the random geometric graph with $p = 1/2$ defined by the points $\boldsymbol{Z}_i = \boldsymbol{X}_i/\|\boldsymbol{X}_i\|$, $i = 1,\dots,n$. Suppose $0 < \rho \leq 1/2$. Under the alternative hypothesis, with probability at least $1-\delta$, the graph contains a clique of size k whenever*

$$\binom{k}{2} \leq \delta \exp\left(\frac{d\sigma^4}{10}\right),$$

where $\sigma^2 = \rho/(1-\rho)$.

Proof. It suffices to show that, if i and j both belong to S then

$$\mathbb{P}\left\{\left\langle \boldsymbol{X}_i, \boldsymbol{X}_j \right\rangle < 0\right\} \leq e^{-d\sigma^4/10}. \tag{3.17}$$

The lemma then follows by the union bound applied for the $\binom{k}{2}$ pairs or vertices of S. Since

$$\begin{aligned}
&\mathbb{P}\left\{\left\langle \boldsymbol{X}_i, \boldsymbol{X}_j \right\rangle < 0\right\} \\
&\quad = \mathbb{P}\left\{\frac{1}{d}\sum_{t=1}^{d}(Y_{i,t}+\sigma N_t)(Y_{j,t}+\sigma N_t) < 0\right\} \\
&\quad = \mathbb{P}\left\{\frac{1}{d}\sum_{t=1}^{d}\Big((Y_{i,t}+\sigma N_t)(Y_{j,t}+\sigma N_t) - \mathbb{E}(Y_{i,t}+\sigma N_t)(Y_{j,t}+\sigma Y_t)\Big) < -\sigma^2\right\},
\end{aligned}$$

the problem boils down to finding appropriate left-tail bounds for independent sums of products of correlated normal random variables.

To this end, we proceed by the Chernoff bound. By a general formula for the cumulant generating function of the product of dependent normal random variables (Exercise 3.3), we have

$$\begin{aligned}
F(\lambda) &\overset{\text{def.}}{=} \ln \mathbb{E}\Big[\exp(\lambda(Y_{i,t}+\sigma N_t)(Y_{j,t}+\sigma N_t))\Big] \\
&= \frac{1}{2}\ln\frac{1-\rho^2}{1-(\rho+(1+\rho)\lambda)^2}
\end{aligned}$$

for all λ such that $|\rho+(1+\rho)\lambda| < 1$. Since we are interested in lower tail probabilities, we consider negative values of λ. Then $F(\lambda)$ is well defined for $\lambda \in (-1, 0]$. By Taylor's theorem, for every such λ there exists $y \in (\lambda, 0)$ such that

$$F(\lambda) = F(0) + \lambda F'(0) + \frac{\lambda^2}{2} F''(y) \,.$$

By straightforward calculation, $F(0) = 0$, $F'(0) = \sigma^2$, and

$$F''(y) = (1+\rho)^2 \frac{1 + (\rho + (1+\rho)y)^2}{\left(1 - (\rho + (1+\rho)y)^2\right)^2}$$

which is monotone increasing for $y \in (-\rho/(1+\rho), 0]$ and therefore

$$F(\lambda) \le \lambda\sigma^2 + \frac{\lambda^2}{2} F''(0) = \lambda\sigma^2 + \frac{\lambda^2}{2} \frac{1+\rho^2}{(1-\rho)^2} \quad \text{for all } \lambda \in (-\rho/(1+\rho), 0] \,.$$

Thus, by the Chernoff bound (Section A.1 in the Appendix), for all $\lambda \in (-\rho/(1+\rho), 0]$,

$$\mathbb{P}\left\{\left\langle \boldsymbol{X}_i, \boldsymbol{X}_j \right\rangle < 0\right\} \le \exp(dF(\lambda)) \le \exp\left(d\lambda\sigma^2 + \frac{d\lambda^2}{2} \frac{1+\rho^2}{(1-\rho)^2}\right) .$$

The upper bound is minimized for $\lambda = -\sigma^2(1-\rho)^2/(1+\rho^2)$ which is a legal choice since $\sigma^2(1-\rho)^2/(1+\rho^2) < \rho/(1+\rho)$. The upper bound becomes

$$\mathbb{P}\left\{\left\langle \boldsymbol{X}_i, \boldsymbol{X}_j \right\rangle < 0\right\} \le \exp\left(-\frac{d\sigma^4(1-\rho)^2}{2(1+\rho^2)}\right) .$$

Since $\sigma^2 \le 1$, we have $\rho \le 1/2$ and we obtain (3.17). □

Now we are ready to state the lower bound for the clique number of $\overline{\mathcal{G}}(n, d, 1/2)$.

Theorem 3.18. *There exist universal constants $c_1, c_2, c_3, c_4 > 0$ such that for all n, d such that $d \ge c_1 \log(c_2 n)$, the median of the clique number $\omega(n, d)$ of $\overline{\mathcal{G}}(n, d, 1/2)$ satisfies*

$$\mathbb{M}\omega(n, d) \ge c_3 \exp\left(\frac{c_4 \log^2(c_2 n)}{d}\right) .$$

One may take $c_1 = 2$, $c_2 = 4\log 2$, $c_3 = 1/32$, and $c_4 = 1/5$. In particular,

$$d \le c_4 \log^{2-\epsilon} n \quad \textit{implies} \qquad \mathbb{M}\omega(n, d) = \Omega(\exp(\log^{\epsilon} n)) \,.$$

Proof. Let $\omega_0 = \mathbb{M}\omega(n,d)$ be the median of the clique number. Consider the hypothesis testing problem of Section 3.1 with $k = 16\omega_0$. Define the random geometric graph on the vertex set $[n]$, connecting vertices i and j whenever $\left\langle \boldsymbol{Z}_i, \boldsymbol{Z}_j \right\rangle \geq 0$. The test statistic we consider is the clique number of the resulting graph, denoted by ω. In particular, consider the test T_n that accepts the null hypothesis if and only if $\omega < k$.

Under the null hypothesis, the $\boldsymbol{Z}_i$'s are i.i.d. uniform on the sphere $\mathbb{S}^{d-1}$ and, consequently, ω has the same distribution as $\omega(n,d)$. By Lemma 3.16, under the alternative hypothesis, with probability at least 7/8, the graph contains a clique of size k whenever

$$\binom{k}{2} < (1/8)e^{d\rho^2/10} .$$

When this is the case, the type II error is bounded as $\mathbb{P}_1\{T_n = 0\} \leq 1/8$. To bound the probability of type I error of T_n, we first prove that $\mathbb{E}_0\omega < 2\omega_0$ for any d and n sufficiently large. We start with

$$\mathbb{E}_0\omega \geq 2\omega_0 \quad \Leftrightarrow \quad \mathbb{E}_0\omega - \omega_0 \geq \frac{1}{2}\mathbb{E}_0\omega \quad \Rightarrow \quad \frac{1}{2}\mathbb{E}_0\omega \leq \mathbb{E}_0\omega - \omega_0 \leq \sqrt{\mathrm{Var}(\omega)} ,$$

where in the last step we used the well-known fact that the difference between the mean and the median of any random variable is bounded by its standard deviation. Now observe that ω, as a function of the independent random variables $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$, is a self-bounding function which implies, by the Efron-Stein inequality, that $\mathrm{Var}(\omega) \leq \mathbb{E}_0\omega$, (see Theorem A.9 in the Appendix). We arrive at

$$\mathbb{E}_0\omega \geq 2\omega_0 \quad \Rightarrow \quad \frac{1}{2}\mathbb{E}_0\omega \leq \sqrt{\mathbb{E}_0\omega} \quad \Leftrightarrow \quad \mathbb{E}_0\omega \leq 4 .$$

However, it is a simple matter to show that $\mathbb{E}_0\omega > 4$ for all d if n is sufficiently large. (To see this it suffices to show that the probability that 5 random points form a clique is bounded away from zero.) We then bound the probability of type I error as follows

$$\mathbb{P}_0\{\omega \geq k\} = \mathbb{P}_0\{\omega \geq 16\omega_0\} \leq \mathbb{P}_0\{\omega \geq 8\mathbb{E}_0\omega\} \leq \frac{1}{8} ,$$

where we used Markov's inequality in the last line.

Combining the bounds on the probabilities of type I and type II errors, we conclude that $R^* \leq 1/4$. Put it another way,

$$R^* > 1/4 \quad \Longrightarrow \quad \binom{16\omega_0}{2} \geq (1/8)e^{d\rho^2/10} .$$

Now, by Theorem 3.2—and in particular by (3.3)—, we see that

$$(16\omega_0)^2 < e^{-\rho d} n \ln 2 \quad \Longrightarrow \quad R^* > 1/4 .$$

We conclude that, for any $\rho \in (0,1)$,

$$(16\omega_0)^2 < e^{-\rho d} n \ln 2 \Longrightarrow (16\omega_0)^2 \geq (1/4)e^{d\rho^2/10} .$$

Therefore, if ρ is such that $e^{-\rho d} n \ln 2 > (1/4)e^{d\rho^2/10}$, then $(16\omega_0)^2 \geq (1/4)e^{d\rho^2/10}$. Choosing $\rho = (1/d)\log((4\log 2)n)$—which is possible since $d \geq 2\log((4\log 2)n)$—clearly satisfies the required inequality and this choice gives rise to the announced lower bound. □

3.4 Bibliographic notes

Various variants of the correlation-detection problem discussed in this chapter has been discussed in the literature. Our general framework is based on Arias-Castro, Bubeck, and Lugosi [4, 5]. The lower bound of Theorem 3.2 appears in [5].

The high-dimensional random geometric graph model was introduced in Devroye, György, Lugosi, and Udina [24] where the dependence of the clique number on the dimension is investigated, see also [5]. We refer to these papers for further bounds for the clique number.

The study of random geometric graphs (on the plane) was initiated by Gilbert [35]. Penrose [60] is a standard text for asymptotic properties of random geometric graphs in fixed dimensions.

3.5 Exercises

Exercise 3.1. *Consider the test hypothesis testing problem defined in Section 3.1 with $\mathcal{C}$ being the class of all subsets of $[n]$ of size k. Define a test T_n that accepts the null hypothesis if and only if the scan-statistic defined in (3.4) does not exceed $dk(k-1)\rho/4$. Prove that*

$$\begin{aligned} \text{if } k > 8/\rho + 1 \text{ and } d \geq (9/\rho)\log n, \quad \text{then} \quad R(T_n) \to 0 \\ \text{if } k \leq 8/\rho + 1 \text{ and } d \geq 9\sqrt{(\log n)/(d(k-1))}, \quad \text{then} \quad R(T_n) \to 0 . \end{aligned}$$

Exercise 3.2. *Finish the calculations for the proof of the second and third statements of Theorem 3.15.*

Exercise 3.3. *Suppose that (ξ, ζ) are jointly normal zero-mean random variables with variances s_ξ^2 and s_ζ^2, respectively, and correlation $r = \mathbb{E}[\xi\zeta]/(s_\xi s_\zeta)$. Prove that the cumulant generating function of their product equals*

$$\ln \mathbb{E}[\exp(\lambda \xi \zeta)] = \frac{1}{2} \ln \frac{1 - r^2}{1 - \Big(r + (1-r^2) s_\xi s_\zeta \lambda\Big)^2}$$

for all λ *such that* $\left|r + (1 - r^2)s_\xi s_\zeta \lambda\right| < 1$.

Chapter 4

Dimension estimation of random geometric graphs

4.1 Detecting underlying geometry in random graphs

In this section we set up and address some basic questions about inferring the geometric structure underlying certain observed graphs. In particular, we consider a hypothesis testing problem in which one observes a graph on n vertices. Under the null hypothesis, the observed graph is a realization of a $\mathcal{G}(n, 1/2)$ Erdős-Rényi graph. Under the alternative hypothesis, the graph is a $\overline{\mathcal{G}}(n, d, 1/2)$ random geometric graph for a certain (known) value of d.

Recall from Chapter 3 that a $\overline{\mathcal{G}}(n, d, 1/2)$ random graph is obtained by drawing n independent random vectors $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$, uniformly distributed on the unit sphere $\mathbb{S}^{d-1}$. Vertices i and j are connected by an edge if and only if $\left\langle \boldsymbol{Z}_i, \boldsymbol{Z}_j \right\rangle \geq 0$.

We are interested in the behavior of the risk R^* of the optimal test. As mentioned in Chapter 3, it follows from the multivariate central limit theorem that, for fixed n, as $d \to \infty$, the random geometric graph $\overline{\mathcal{G}}(n, d, 1/2)$ converges, in distribution, to $\mathcal{G}(n, 1/2)$. This implies that $R^* \to 1$ when n is fixed and $d \to \infty$. Thus, for large values of d, $\overline{\mathcal{G}}(n, d, 1/2)$ and $\mathcal{G}(n, 1/2)$ become essentially indistinguishable and "geometry disappears". On the other hand, when d is fixed and $n \to \infty$, then clearly $R^* \to 0$ (e.g., by comparing clique numbers, as seen in Section 1.2).

Thus, it is of interest to determine for what values of d the optimal risk becomes large—as a function of n. Recall from Chapter 1 that R^* equals 1 minus the total variation distance between the distributions of $\overline{\mathcal{G}}(n, d, 1/2)$ and $\mathcal{G}(n, 1/2)$.

The following theorem shows that the "critical" dimension is about $d \sim n^3$.

Theorem 4.1. *Let R^* be the risk of the optimal test in the hypothesis testing problem*

$\mathcal{G}(n, 1/2)$ *vs.* $\overline{\mathcal{G}}(n, d, 1/2)$ *described above. Then, as* $n \to \infty$,

$$R^* \to 0 \quad \text{if} \quad \tfrac{d}{n^3} \to 0$$
$$R^* \to 1 \quad \text{if} \quad \tfrac{d}{n^3} \to \infty\,.$$

To prove the first half of the theorem, it suffices to exhibit a test with vanishing risk for $d = o(n^3)$. The proof of the second statement follows from an upper bound for the total variation distance between the distribution of a random Wishard matrix and a Gaussian Orthogonal Ensemble. These arguments are sketched below.

4.1.1 The triangle test

Faced with the problem of testing whether an observed graph is a geometric graph, a natural idea is to count triangles (i.e., cliques of size 3) in the graph. The intuition is that, while under the null hypothesis each triangle is present with probability 1/8—and hence the expected number of triangles in $\mathcal{G}(n, 1/2)$ is $\binom{n}{3}/8$—, the probability that three points form a triangle is larger in the geometric model $\overline{\mathcal{G}}(n, d, 1/2)$. This is because, conditionally on the event that two vertices are connected by an edge, the probability that a third vertex is connected to both is larger than 1/4 since, given that it is close to one of them, it is more likely to be close to the other one as well.

To quantify this statement, one may prove that, if $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3$ are independent random vectors uniformly distributed on $\mathbb{S}^{d-1}$, then

$$\mathbb{P}_1\left\{\langle \boldsymbol{Z}_1, \boldsymbol{Z}_2\rangle \geq 0, \langle \boldsymbol{Z}_1, \boldsymbol{Z}_3\rangle \geq 0, \langle \boldsymbol{Z}_2, \boldsymbol{Z}_3\rangle \geq 0\right\} \geq 2^{-3}\left(1 + \frac{C}{\sqrt{d}}\right) \tag{4.2}$$

for some constant $C > 0$, and therefore, under the alternative hypothesis, the number N of triangles satisfies

$$\mathbb{E}_1 N \geq \binom{n}{3} 2^{-3}\left(1 + \frac{C}{\sqrt{d}}\right).$$

Showing (4.2) is somewhat technical. We only sketch here the basic intuition behind the proof. First note that

$$\begin{aligned}
&\mathbb{P}_1\{\text{vertices } 1, 2, 3 \text{ form a triangle}\} \\
&\quad = \mathbb{P}\{\langle \boldsymbol{Z}_1, \boldsymbol{Z}_2\rangle \geq 0, \langle \boldsymbol{Z}_1, \boldsymbol{Z}_3\rangle \geq 0, \langle \boldsymbol{Z}_2, \boldsymbol{Z}_3\rangle \geq 0\} \\
&\quad = \mathbb{P}\Big\{\langle \boldsymbol{Z}_2, \boldsymbol{Z}_3\rangle \geq 0 \Big| \langle \boldsymbol{Z}_1, \boldsymbol{Z}_2\rangle \geq 0, \langle \boldsymbol{Z}_1, \boldsymbol{Z}_3\rangle \geq 0\Big\}\mathbb{P}_1\left\{\langle \boldsymbol{Z}_1, \boldsymbol{Z}_2\rangle \geq 0, \langle \boldsymbol{Z}_1, \boldsymbol{Z}_3\rangle \geq 0\right\} \\
&\quad = \frac{1}{4}\mathbb{P}\Big\{\langle \boldsymbol{Z}_2, \boldsymbol{Z}_3\rangle \geq 0 \Big| \langle \boldsymbol{Z}_1, \boldsymbol{Z}_2\rangle \geq 0, \langle \boldsymbol{Z}_1, \boldsymbol{Z}_3\rangle \geq 0\Big\}\,,
\end{aligned}$$

since the events $\langle \boldsymbol{Z}_1, \boldsymbol{Z}_2 \rangle \geq 0$ and $\langle \boldsymbol{Z}_1, \boldsymbol{Z}_3 \rangle \geq 0$ are independent. To show that the latter conditional probability is at least $1/2 + C/\sqrt{d}$, we need to understand the typical angle is between $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$. By rotational invariance we may assume that $\boldsymbol{Z}_1 = (1, 0, 0, \ldots, 0)$, and therefore $\langle \boldsymbol{Z}_1, \boldsymbol{Z}_2 \rangle = Z_{2,1}$, the first component of the vector $\boldsymbol{Z}_2$. By the argument outlined in Section 3.2, $\sqrt{d} Z_{2,1}$ is approximately standard normal when d is large. Thus, conditioned on the event $Z_{2,1} \geq 0$, $Z_{2,1}$ this gives the boost in the conditional probability that we see.

Thus, the expected number of triangles under the alternative hypothesis is larger by at least $\Theta\left(n^3/\sqrt{d}\right)$ than under the null hypothesis:

$$\mathbb{E}_1 N - \mathbb{E}_0 N \geq \binom{n}{3} \frac{C}{\sqrt{d}}$$

for some positive constant C. Hence, the test T that accepts the null hypothesis if and only if $N < \mathbb{E}_0 N + \binom{n}{3}(C/2)d^{-1/2}$ has a risk tending to zero if the standard deviation of N is $o(n^3 d^{-1/2})$.

One may show that $\mathrm{Var}_0(N) = \binom{n}{3}(7/64) + \binom{n}{4}\binom{4}{2}(1/32)$ and $\mathrm{Var}_1(N) \leq n^4$ (Exercis 4.2). Thus, the standard deviations are of the order of n^2 which is $o(n^3/\sqrt{d})$ whenever $d = o(n^2)$. This falls short of proving the first half of Theorem 4.1.

To prove that one may obtain a small risk for $d = o(n^3)$, one may construct a similar test statistic that has a smaller variance. The trick is to count, instead of triangles, *signed* triangles. Let $A_{i,j} = \mathbb{1}_{\{i \sim j\}}$ be the indicator that vertices i and j are connected by an edge. Consider the test statistic

$$N_{\pm} = \sum_{1 \leq i < j < k \leq n} \left(A_{i,j} - \frac{1}{2}\right)\left(A_{j,k} - \frac{1}{2}\right)\left(A_{k,i} - \frac{1}{2}\right).$$

A simple calculation shows that $\mathbb{E}_0 N_{\pm} = 0$ and

$$\mathrm{Var}_0(N_{\pm}) = \frac{1}{64}\binom{n}{3}.$$

On the other hand, a similar argument as in the case of the triangle count shows that there exists a positive constant c such that

$$\mathbb{E}_1 N_{\pm} \geq \frac{cn^3}{\sqrt{d}} \quad \text{and} \quad \mathrm{Var}_1(N_{\pm}) \leq n^3 + \frac{3n^4}{d},$$

which implies that the test that accepts the null hypothesis if and only if $N_{\pm} \leq cn^3/(2\sqrt{d})$ has a regret converging to zero whenever $n^3/\sqrt{d} = o\left(\sqrt{n^3 + n^4/d}\right)$. This happens if and only if $d = o(n^3)$.

4.1.2 Geometry disappears in high dimensions

Here we sketch the proof of the second statement of Theorem 4.1. We need to prove that when the dimension d is much larger than n^3 then there is no test that is able to distinguish the random geometric graph $\overline{\mathcal{G}}(n,d,1/2)$ from the Erdős-Rényi random graph $\mathcal{G}(n,1/2)$. This is equivalent to saying that the *total variation distance* between the two distributions satisfies

$$D(\overline{\mathcal{G}}(n,d,1/2),\mathcal{G}(n,1/2)) \to 0 \quad \text{when} \quad d^3/n \to \infty\,.$$

The total variation distance between the two random graph distributions is difficult to handle analytically. In order to circumvent this difficulty, we represent both random objects as functions of certain random matrices that are more manageable.

In the case of the random geometric graph, it is natural to represent it as a function of the *Wishart matrix* defined by the symmetric square matrix $\boldsymbol{W} = (W_{i,j})_{n\times n}$ whose entries are

$$W_{i,j} = \left\langle \boldsymbol{X}_i, \boldsymbol{X}_j \right\rangle\,, \quad i,j \in [n]\,.$$

Recall that the $\boldsymbol{X}_i$ are independent standard normal vectors in $\mathbb{R}^d$. Indeed, the adjacency matrix $\overline{\boldsymbol{A}} = (\overline{A}_{i,j})_{n\times n}$ of $\overline{\mathcal{G}}(n,d,1/2)$ may be defined by

$$\overline{A}_{i,j} = \begin{cases} \mathbb{1}_{\{W_{i,j}\geq 0\}} & \text{if } i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

One may similarly represent an Erdős-Rényi random graph. To this end, let $N_{i,j}$ be independent standard normal random variables for $1 \leq i \leq j \leq n$. The random graph in which vertices i and j are connected by an edge if and only if $N_{i,j} \geq 0$ is clearly $\mathcal{G}(n,1/2)$. Thus, the adjacency matrix $\boldsymbol{A}$ of a $\mathcal{G}(n,1/2)$ may be generated as a function of the $N_{i,j}$. In order to make it comparable to the Wishart matrix $\boldsymbol{W}$, we rescale the $N_{i,j}$ and add appropriate diagonal entries (that are irrelevant for the definition of the random graph) and write

$$\overline{A}_{i,j} = \begin{cases} \mathbb{1}_{\{V_{i,j}\geq 0\}} & \text{if } i \neq j \\ 0 & \text{otherwise,} \end{cases}$$

where the entries of the symmetric random matrix $\boldsymbol{V}$ are defined by

$$V_{i,j} = \begin{cases} \sqrt{d}N_{i,j} & \text{if } i < j \\ V_{j,i} & \text{if } i > j \\ d + \sqrt{2d}N_{i,i} & \text{otherwise.} \end{cases}$$

Having written the random graphs as functions of the random matrices $\boldsymbol{W}$ and $\boldsymbol{V}$, by Exercise 4.3 it suffices to prove that the total variation distance $D(\boldsymbol{W},\boldsymbol{V}) \to 0$ whenever $d^3/n \to \infty$.

The matrix $d^{-1/2}(\boldsymbol{V} - d\boldsymbol{I})$ is called a *Gaussian orthogonal ensemble* and it is a well-studied random matrix. In particular, the densities of both $\boldsymbol{W}$ and $\boldsymbol{V}$ are explicitly known over the cone $\mathcal{C} \subset \mathbb{R}^{n^2}$ of symmetric positive semidefinite matrices (with respect to the Lebesgue measure). In particular, the density of the Wishart matrix $\boldsymbol{W}$ is

$$f_{\boldsymbol{W}}(\boldsymbol{M}) := \frac{(\det(\boldsymbol{M}))^{\frac{1}{2}(d-n-1)} \exp\left(-\frac{1}{2}\mathrm{Tr}(\boldsymbol{M})\right)}{2^{\frac{1}{2}dn} \pi^{\frac{1}{4}n(n-1)} \prod_{i=1}^{n} \Gamma\left(\frac{1}{2}(d+1-i)\right)}, \quad \boldsymbol{M} \in \mathcal{C},$$

where $\mathrm{Tr}(\boldsymbol{M})$ denotes the trace of the matrix $\boldsymbol{M}$. Similarly, from the known formula of the density of a Gaussian orthogonal ensemble, the density of $\boldsymbol{V}$ may be derived:

$$f_{\boldsymbol{V}}(\boldsymbol{M}) := \frac{\exp\left(-\frac{1}{4d}\mathrm{Tr}\left((\boldsymbol{M} - d\boldsymbol{I})^2\right)\right)}{(2\pi d)^{\frac{1}{4}n(n+1)} 2^{\frac{n}{2}}}, \quad \boldsymbol{M} \in \mathcal{C}.$$

Since the total variation distance of $\boldsymbol{W}$ and $\boldsymbol{V}$ equals

$$D(\boldsymbol{W}, \boldsymbol{V}) = \frac{1}{2} \int_{\mathcal{C}} |f_{\boldsymbol{W}}(\boldsymbol{M}) - f_{\boldsymbol{V}}(\boldsymbol{M})| \, d\boldsymbol{M},$$

the explicit formulas of the densities may be used to estimate $D(\boldsymbol{W}, \boldsymbol{V})$. In particular, one may show the following result.

Theorem 4.3. *If $d/n^3 \to \infty$, then $D(\boldsymbol{W}, \boldsymbol{V}) \to 0$.*

This theorem allows us to complete the proof of Theorem 4.1.

4.2 Bibliographic notes

The material of Section 4.1 is based on Bubeck, Ding, Eldan, and Rácz [17]. The full proofs of the results may be found there. In particular, our presentation was largely inspired by Rácz and Bubeck [62].

Theorem 4.3 was proven independently by Bubeck, Ding, Eldan, and Rácz [17] and Jiang and Li [45]. This theorem has been extended to Wishart matrices with distributions more general than Gaussian by Bubeck and Ganguly [18].

4.3 Exercises

Exercise 4.1. *Extend Theorem 4.1 to the problem of testing $\mathcal{G}(n,p)$ versus $\overline{\mathcal{G}}(n,d,p)$ for a fixed $p \in (0,1)$.*

Exercise 4.2. *Show that if N is the number of triangles in $\mathcal{G}(n,p)$, then*

$$\mathrm{Var}(N) = \binom{n}{3}\left(p^3 - p^6\right) + \binom{n}{4}\binom{4}{2}\left(p^5 - p^6\right)$$

and that if N is the number of triangles in $\overline{\mathcal{G}}(n,d,p)$, then $\mathrm{Var}(N) \leq n^4$.

Exercise 4.3. *Let X, Y be random variables taking values in some measurable space $\mathcal{X}$. Let $f, g : \mathcal{X} \to \mathcal{Y}$ be measurable functions taking values in a set $\mathcal{Y}$. Prove that*

$$D(X,Y) \geq D(f(X), g(Y)) \,,$$

where $D(X,Y) = \sup_A |\mathbb{P}\{X \in A\} - \mathbb{P}\{Y \in A\}|$ is the total variation distance between the distributions of X and Y. (The supremum is taken over all measurable subsets of $\mathcal{X}$.)

Chapter 5

Mean estimation

In this chapter we examine the classical problem of estimating the mean of a random variable. Let $X_1, \ldots, X_n$ be independent, identically distributed real random variables with mean $\mu = \mathbb{E}X_1$. Upon observing these random variables, one would like to estimate μ. An estimator $\widehat{\mu}_n = \widehat{\mu}_n(X_1, \ldots, X_n)$ is simply a function of $X_1, \ldots, X_n$.

The quality of an estimator may be measured in various different ways. We are interested in the smallest possible value $a = a(n, \delta)$ such that

$$\mathbb{P}\left\{\left|\widehat{\mu}_n - \mu\right| > a\right\} \leq \delta \,.$$

We emphasize the non-asymptotic nature of this requirement.

The most natural choice of a mean estimator is the standard empirical mean

$$\overline{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \,.$$

The behavior of the empirical mean is well understood. For example, if the X_i have a finite second moment, then the central limit theorem guarantees that this estimator has Gaussian tails, asymptotically, when $n \to \infty$. Indeed,

$$\mathbb{P}\left\{\left|\overline{\mu}_n - \mu\right| > \frac{\sigma\Phi^{-1}(1-\delta/2)}{\sqrt{n}}\right\} \to \delta \,,$$

where $\sigma^2 > 0$ is the variance of the X_i and $\Phi(x) = \mathbb{P}\{N \leq x\}$ is the cumulative distribution function of a standard normal random variable N. By the Chernoff bound (see Appendix), for all $x \geq 0$,

$$1 - \Phi(x) \leq e^{-x^2/2} \,.$$

This implies that $\Phi^{-1}(1-\delta/2) \le \sqrt{2\log(2/\delta)}$, and the central limit theorem asserts that

$$\lim_{n\to\infty} \mathbb{P}\left\{ \left|\overline{\mu}_n - \mu\right| > \frac{\sigma\sqrt{2\log(2/\delta)}}{\sqrt{n}} \right\} \le \delta \,.$$

However, the asymptotic nature of this property limits the usefulness of this performance bound. We seek non-asymptotic performance bounds of the same form. In particular, we say that a mean estimator $\widehat{\mu}_n$ is *sub-Gaussian* if for some constant $L > 0$, for all sample sizes n, with probability at least $1-\delta$,

$$\left|\widehat{\mu}_n - \mu\right| \le \frac{L\sigma\sqrt{\log(2/\delta)}}{\sqrt{n}} \,. \tag{5.1}$$

It is easy to see that, under certain conditions on the distribution of the X_i, the empirical mean has the desired performance. Indeed, suppose that the X_i have a sub-Gaussian distribution in the sense that, for all $\lambda > 0$,

$$\mathbb{E}e^{\lambda(X_i-\mu)} \le e^{\sigma^2\lambda^2/2} \,.$$

Then, by the Chernoff bound, $\widehat{\mu}_n$ is sub-Gaussian for all $\delta \in (0,1)$ with $L = \sqrt{2}$. However, the sub-Gaussian assumption is quite restrictive and imposes a strong condition on the decay of the tail probabilities of the X_i. If one only assumes that σ exists (i.e., the variance of the X_i is finite), then one still has, by Chebyshev's inequality, that, with probability at least $1-\delta$,

$$\left|\overline{\mu}_n - \mu\right| \le \sigma\sqrt{\frac{1}{n\delta}} \,.$$

While the bound decays at the optimal $O(n^{-1/2})$ rate as a function of the sample size, the dependence on the confidence parameter δ is exponentially worse than in (5.1).

Perhaps surprisingly, there exist mean estimators that achieve a sub-Gaussian performance for all distributions with a finite variance. A simple such estimator is presented and analyzed in the next section.

5.1 The median-of-means estimator

Roughly speaking, the median-of-means estimator partitions the data into k groups of roughly equal size, computes the empirical mean in each group, and finally takes the median of the obtained values.

Formally, recall that the median of m real numbers $x_1, \dots, x_m \in \mathbb{R}$ is defined as $M(x_1, \dots, x_m) = x_i$ where x_i is such that

$$\left|\{j \in [m] : x_j \le x_i\}\right| \ge \frac{m}{2} \quad \text{and} \quad \left|\{j \in [m] : x_j \ge x_i\}\right| \ge \frac{m}{2} \,.$$

(If several i fit the above description, we take the smallest one.)

Now let $1 \le k \le n$ and partition $[n] = \{1,\dots,n\}$ into k blocks $B_1,\dots,B_k$, each of size $|B_i| \ge \lfloor n/k \rfloor \ge 2$.

Given $X_1,\dots,X_n$, compute the sample mean in each block

$$Y_j = \frac{1}{|B_j|} \sum_{i \in B_j} X_i$$

and define the median-of-means estimator by $\widehat{\mu}_n = M(Y_1,\dots,Y_m)$.

The performance of the estimator is established next. For simplicity, assume that n is divisible by k so that each block has $m = n/k$ elements.

Theorem 5.2. *Let $X_1,\dots,X_n$ be independent, identically distributed random variables with mean μ and variance σ^2. Let m,k be positive integers assume that $n = mk$. Then the median-of-means estimator with k blocks satisfies*

$$\mathbb{P}\left\{\left|\widehat{\mu}_n - \mu\right| > \sigma\sqrt{4/m}\right\} \le e^{-k/8} .$$

In particular, for any $\delta \in (0,1)$, if $k = \lceil 8\log(1/\delta)\rceil$, then

$$\mathbb{P}\left\{\left|\widehat{\mu}_n - \mu\right| > \sigma\sqrt{\frac{32\log(1/\delta)}{n}}\right\} \le \delta ,$$

Proof. By Chebyshev's inequality, for each $j = 1,\dots,k$, with probability at least 3/4,

$$\left|Y_j - \mu\right| \le \sigma\sqrt{\frac{4}{m}} .$$

Thus, $\left|\widehat{\mu}_n - \mu\right| > \sigma\sqrt{4/m}$ implies that at least $k/2$ of the means Y_j are such that $\left|Y_j - \mu\right| > \sigma\sqrt{4/m}$. Hence,

$$\begin{aligned}
\mathbb{P}\left\{\left|\widehat{\mu}_n - \mu\right| > \sigma\sqrt{4/m}\right\} &\le \mathbb{P}\left\{\mathrm{B}in(k,1/4) \ge \frac{k}{2}\right\} \\
&\quad \text{(where } \mathrm{B}in(k,1/4) \text{ is a binomial } (k,1/4) \text{ random variable)} \\
&= \mathbb{P}\left\{\mathrm{B}in(k,1/4) - \mathbb{E}\mathrm{B}in(k,1/4) \ge \frac{k}{4}\right\} \\
&\le e^{-k/8} \quad \text{(by Hoeffding's inequality–Theorem A.4)}
\end{aligned}$$

□

The theorem shows that the median-of-means estimator has a sub-Gaussian performance with $L = 8$. It is remarkable that this is possible for all distributions with a finite variance. On the negative side, it is important to point out that the estimator $\widehat{\mu}_n$ depends on the confidence level δ as the number of blocks k is chosen as a function of δ.

Next we show that no estimator can have a significantly better performance.

Theorem 5.3. *Let $n > 5$ be a positive integer. Let $\mu \in \mathbb{R}$, $\sigma > 0$ and $\delta \in (2e^{-n/4}, 1/2)$. Then for any mean estimator $\widehat{\mu}_n$, there exists a distribution with mean μ and variance σ^2 such that*

$$\mathbb{P}\left\{\left|\widehat{\mu}_n - \mu\right| > \sigma\sqrt{\frac{\log(1/\delta)}{n}}\right\} \geq \delta\,.$$

Proof. To derive the announced "minimax" lower bound, it suffices to consider two distributions, P_+, P_-, both concentrated on two points, defined by

$$P_+(\{0\}) = P_-(\{0\}) = 1 - p\,, \qquad P_+(\{c\}) = P_-(\{-c\}) = p\,,$$

where $p \in [0,1]$ and $c > 0$. Note that the means of the two distributions are $\mu_{P_+} = pc$ and $\mu_{P_-} = -pc$ and both have variance $\sigma^2 = c^2 p(1-p)$.

For $i = 1, \ldots, n$, let (X_i, Y_i) be independent pairs of real-valued random variables such that

$$\mathbb{P}\{X_i = Y_i = 0\} = 1 - p \quad \text{and} \quad \mathbb{P}\{X_i = c, Y_i = -c\} = p\,.$$

Note that X_i is distributed as P_+ and Y_i is distributed as P_-. Let $\delta \in (0, 1/2)$. If $\delta \geq 2e^{-n/4}$ and $p = (1/(2n))\log(2/\delta)$, then (using $1 - p \geq \exp(-p/(1-p))$),

$$\mathbb{P}\{X_1^n = Y_1^n\} = (1-p)^n \geq 2\delta\,.$$

Let $\widehat{\mu}_n$ be any mean estimator, possibly depending on δ. Then

$$\begin{aligned}
&\max\left(\mathbb{P}\left\{\left|\widehat{\mu}_n(X_1^n) - \mu_{P_+}\right| > cp\right\}, \mathbb{P}\left\{\left|\widehat{\mu}_n(Y_1^n) - \mu_{P_-}\right| > cp\right\}\right) \\
&\quad \geq \frac{1}{2}\mathbb{P}\left\{\left|\widehat{\mu}_n(X_1, \ldots, X_n) - \mu_{P_+}\right| > cp \quad \text{or} \quad \left|\widehat{\mu}_n(Y_1, \ldots, Y_n) - \mu_{P_-}\right| > cp\right\} \\
&\quad \geq \frac{1}{2}\mathbb{P}\{\widehat{\mu}_n(X_1, \ldots, X_n) = \widehat{\mu}_n(Y_1, \ldots, Y_n)\} \\
&\quad \geq \frac{1}{2}\mathbb{P}\{X_1, \ldots, X_n = Y_1, \ldots, Y_n\} \geq \delta\,.
\end{aligned}$$

From $\sigma^2 = c^2 p(1-p)$ and $p \leq 1/2$ we have that $cp \geq \sigma\sqrt{p/2}$, and therefore

$$\max\left(\mathbb{P}\left\{\left|\widehat{\mu}_n(X_1, \ldots, X_n) - \mu_{P_+}\right| > \sigma\sqrt{\frac{\log\frac{2}{\delta}}{n}}\right\}, \mathbb{P}\left\{\left|\widehat{\mu}_n(Y_1, \ldots, Y_n) - \mu_{P_-}\right| > \sigma\sqrt{\frac{\log\frac{2}{\delta}}{n}}\right\}\right) \geq \delta\,.$$

Theorem 5.3 follows. □

The median-of-means estimator may also used even if the distribution of the X_i has an infinite variance but has a finite moment of order $1+\alpha$ for some $\alpha \in (0,1)$. The rate of convergence of the estimator deteriorates—the error is of the order of $n^{-\alpha/(1+\alpha)}$ but the median-of-means estimator exhibits a similar optimality property as in the case of finite variance. The following result summarizes this extension. We leave the proof as an exercise.

Theorem 5.4. *Let $\alpha \in (0,1]$. $X_1, \dots, X_n$ be independent, identically distributed random variables with mean μ and $(1+\alpha)$-th central moment $M = \mathbb{E}\left[|X_i - \mu|^{1+\alpha}\right]$. Let m, k be positive integers assume that $n = mk$. Then the median-of-means estimator with $k = \lceil 8\log(2/\delta)\rceil$ blocks satisfies*

$$\mathbb{P}\left\{\left|\widehat{\mu}_n - \mu\right| > 8\left(\frac{12M^{1/\alpha}\log(1/\delta)}{n}\right)^{\alpha/(1+\alpha)}\right\} \le \delta\,.$$

Moreover, for any mean estimator $\widehat{\mu}_n$, there exists a distribution with mean μ and $(1+\alpha)$-th central moment M such that

$$\mathbb{P}\left\{\left|\widehat{\mu}_n - \mu\right| > \left(\frac{M^{1/\alpha}\log(2/\delta)}{n}\right)^{\alpha/(1+\alpha)}\right\} \ge \delta\,.$$

5.2 Estimating the mean of a random vector

In this section we discuss extensions of the mean estimation problem to the multivariate setting, that is, when one is interested in estimating the mean of a random vector.

Let X be a random vector taking values in $\mathbb{R}^d$. Assume that the mean vector $\mu = \mathbb{E}X$ and covariance matrix $\Sigma = \mathbb{E}(X-\mu)(X-\mu)^T$ exist. Given n independent, identically distributed samples $X_1, \dots, X_n$ drawn from the distribution of X, one wishes to estimate the mean vector.

Just like in the univariate case, a natural choice is the sample mean $(1/n)\sum_{i=1}^n X_i$. The sample mean has a near-optimal behavior whenever the distribution is sufficiently light tailed. However, whenever heavy tails are a concern, the sample mean is to be avoided as it may have a sub-optimal performance.

5.2.1 Sub-Gaussian property

In the previous section, for the univariate problem, we constructed an mean estimator with a sub-Gaussian performance. In order to properly set up our goal for the d-dimensional case, first we need to understand what "sub-Gaussian performance" means.

If X has a multivariate normal distribution with mean vector μ and covariance matrix Σ, then $\overline{\mu}_n$ is also multivariate normal with mean μ and covariance matrix $(1/n)\Sigma$. Thus, for all $t > 0$,

$$\mathbb{P}\left\{\|\overline{\mu}_n - \mu\| \geq \mathbb{E}\|\overline{\mu}_n - \mu\| + t\right\} = \mathbb{P}\left\{\|\overline{X}\| - \mathbb{E}\|\overline{X}\| \geq t\sqrt{n}\right\} ,$$

where $\overline{X}$ be a Gaussian vector in $\mathbb{R}^d$ with zero mean and covariance matrix Σ. A key property of Gaussian vectors is that $\overline{X}$ has the same distribution as $\Sigma^{1/2}Y$ where Y is a standard normal vector (i.e., with zero-mean and identity covariance matrix) and $\Sigma^{1/2}$ is the positive semidefinite square root of Σ. Also, observe that for all $y, y' \in \mathbb{R}^d$,

$$\left|\|\Sigma^{1/2}y\| - \|\Sigma^{1/2}y'\|\right| \leq \|\Sigma^{1/2}(y - y')\| \leq \|\Sigma^{1/2}\| \cdot \|y - y'\| ,$$

where $\|\Sigma^{1/2}\|$ is the spectral norm of $\Sigma^{1/2}$. Thus, $\Sigma^{1/2}y$ is a Lipschitz function of $y \in \mathbb{R}^d$ with Lipschitz constant $\|\Sigma^{1/2}\| = \sqrt{\lambda_{\max}}$, with $\lambda_{\max} = \lambda_{\max}(\Sigma)$ denoting the largest eigenvalue of the covariance matrix Σ. Now it follows from the Gaussian concentration inequality (Theorem A.11 in the Appendix) that

$$\mathbb{P}\left\{\|\overline{X}\| - \mathbb{E}\|\overline{X}\| \geq t\sqrt{n}\right\} \leq e^{-nt^2/(2\lambda_{\max})} .$$

Noting that

$$\mathbb{E}\|\overline{X}\| \leq \sqrt{\mathbb{E}\|\overline{X}\|^2} = \sqrt{\mathrm{Tr}(\Sigma)} ,$$

the trace of the covariance matrix Σ, we have that, for $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\|\overline{\mu}_n - \mu\| \leq \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{2\lambda_{\max}\log(1/\delta)}{n}} . \tag{5.5}$$

5.2.2 Multivariate median-of-means

For non-Gaussian and possibly heavy-tailed distributions, one cannot expect a sub-Gaussian behavior of the sample mean similar to (5.5).

One may try to extend the median-of-means estimator to the multivariate case. To this end, just like in the univariate case, we partition $[n] = \{1, \ldots, n\}$ into k blocks $B_1, \ldots, B_k$, each of size $|B_i| \geq \lfloor n/k \rfloor \geq 2$. Here k is a parameter of the estimator to be chosen later. For simplicity, we assume that $km = n$ for some positive integer m. Just like before, we compute the sample mean of the random vectors within each block: for $j = 1, \ldots, k$, let

$$Y_j = \frac{1}{m}\sum_{i \in B_j} X_i .$$

Since $\mathbb{E}\|Y_j - \mu\|^2 = \mathrm{Tr}(\Sigma)/m$, by Chebyshev's inequality, $\|Y_j - \mu\| \le r \stackrel{\text{def.}}{=} 2\sqrt{\mathrm{Tr}(\Sigma)/m}$ with probability at least 3/4. Thus, by choosing $k = \lceil 8\log(1/\delta)\rceil$, we have that, with probability at least $1-\delta$, more than half of the points Y_j satisfy

$$\|Y_j - \mu\| \le r\,.$$

Denote this event by E. (Thus, $\mathbb{P}\{E\} \ge 1-\delta$.) Now choose $\widehat{\mu}_n$ to be the point in $\mathbb{R}^d$ with the property that the Euclidean ball centered at $\widehat{\mu}_n$ that contains more than $k/2$ of the points Y_j has minimal radius. On the event E, this radius is at most r. Hence, at least one of the Y_j is within distance r to both μ and $\widehat{\mu}_n$. Thus, by the triangle inequality, $\|\widehat{\mu}_n - \mu\| \le 2r$. We have obtained the following proposition.

Proposition 5.6. *Let $X_1,\dots,X_n$ be i.i.d. random vectors in $\mathbb{R}^d$ with mean μ and covariance matrix Σ. Let $\delta \in (0,1)$ and let $\widehat{\mu}_n$ be the estimator defined above with $k = \lceil 8\log(1/\delta)\rceil$. Then, with probability at least $1-\delta$,*

$$\left\|\widehat{\mu}_n - \mu\right\| \le 4\sqrt{\frac{\mathrm{Tr}(\Sigma)(8\log(1/\delta)+1)}{n}}\,.$$

Given n points $x_1,\dots,x_n \in \mathbb{R}^d$, the center of the smallest ball that contains at least half of the points may be considered as a notion of a multivariate median. Computing such a median is a nontrivial problem. One may replace it by the so-called *geometric median* defined as

$$m = \operatorname*{argmin}_{y\in\mathbb{R}^d} \sum_{i=1}^{n} \|x_i - y\|\,.$$

The multivariate median-of-means estimator may be defined as the geometric median of the sample means $Y_1,\dots,Y_k$ of the k blocks. This estimator has a similar performance, see Exercise 5.3.

While the bound is quite remarkable–note that no assumption other than the existence of the covariance matrix is made–, it does not quite achieve a sub-Gaussian performance bound that resembles (5.5).

An instructive example is when all eigenvalues are identical and equal to $\lambda_{\max}$. If the dimension d is large, (5.5) is of the order of $\sqrt{(\lambda_{\max}/n)(d+\log(\delta^{-1}))}$ while the bound above only gives the order $\sqrt{(\lambda_{\max}/n)(d\log(\delta^{-1}))}$.

In order to achieve a truly sub-Gaussian performance, we need to define a new estimator.

5.2.3 Median of means tournament

Here we introduce a mean estimator with a sub-Gaussian performance for all distributions whose covariance matrix exists.

Recall that we are given an i.i.d. sample $X_1, \ldots, X_n$ of random vectors in $\mathbb{R}^d$. As in the case of the median-of-means estimator, we start by partitioning the set $\{1, \ldots, n\}$ into k blocks $B_1, \ldots, B_k$, each of size $|B_j| \geq m \stackrel{\text{def.}}{=} \lfloor n/k \rfloor$, where k is a parameter of the estimator whose value depends on the desired confidence level, as specified below. In order to simplify the presentation, we assume that n is divisible by k and therefore $|B_j| = m$ for all $j = 1, \ldots, k$.

Define the sample mean within each block by

$$Y_j = \frac{1}{m} \sum_{i \in B_j} X_i \, .$$

For each $a \in \mathbb{R}^d$, let

$$T_a = \left\{ x \in \mathbb{R}^d : \exists J \subset [k] : |J| \geq k/2 \text{ such that for all } j \in J, \ \|Y_j - x\| \leq \|Y_j - a\| \right\} \tag{5.7}$$

and define the mean estimator by

$$\widehat{\mu}_n \in \operatorname*{argmin}_{a \in \mathbb{R}^d} \operatorname{diam}(T_a) \, .$$

Thus, $\widehat{\mu}_n$ is chosen to minimize, over all $a \in \mathbb{R}^d$, the diameter of the set T_a defined as the set of points $x \in \mathbb{R}^d$ for which $\|Y_j - x\| \leq \|Y_j - a\|$ for the majority of the blocks. (If there are several minimizers, one may pick any one of them.)

Note that the minimum is always achieved. This follows from the fact that $\operatorname{diam}(T_a)$ is a continuous function of a (since, for each a, T_a is the intersection of a finite union of closed balls, and the centers and radii of the closed balls are continuous in a).

One may interpret $\operatorname{argmin}_{a \in \mathbb{R}^d} \operatorname{diam}(T_a)$ as a multivariate notion of the median of $Y_1, \ldots, Y_k$. Indeed, when $d = 1$, it is a particular choice of the median and the estimator coincides with the median-of-means estimator.

The following performance bound shows that the estimator has the desired sub-Gaussian performance.

Theorem 5.8. *Let $\delta \in (0,1)$ and consider the mean estimator $\widehat{\mu}_n$ with parameter $k = \lceil 200 \log(2/\delta) \rceil$. If $X_1, \ldots, X_n$ are i.i.d. random vectors in $\mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$ and covariance matrix Σ, then for all n, with probability at least $1 - \delta$,*

$$\left\| \widehat{\mu}_n - \mu \right\| \leq 2 \max \left(800 \sqrt{\frac{\operatorname{Tr}(\Sigma)}{n}}, 240 \sqrt{\frac{\lambda_{max} \log(2/\delta)}{n}} \right) .$$

Just like, the performance bound of Proposition 5.6, Theorem 5.8 is "infinite-dimensional" in the sense that the bound does not depend on the dimension d

explicitly. Indeed, the same estimator may be defined for Hilbert-space valued random vectors and Theorem 5.8 remains valid as long as $\mathrm{Tr}(\Sigma) = \mathbb{E}\|X-\mu\|^2$ is finite.

Theorem 5.8 is an outcome of the following observation.

Theorem 5.9. *Using the same notation as above and setting*

$$r = \max\left(800\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}, 240\sqrt{\frac{\lambda_{max}\log(2/\delta)}{n}}\right),$$

with probability at least $1-\delta$*, for any* $a \in \mathbb{R}^d$ *such that* $\|a-\mu\| \geq r$*, one has* $\|Y_j - a\| > \|Y_j - \mu\|$ *for more than* $k/2$ *indices* j*.*

Theorem 5.9 implies that for a 'typical' collection $X_1,\ldots,X_n$, μ is closer to a majority of the Y_j's when compared to any $a \in \mathbb{R}^d$ that is sufficiently far from μ. Obviously, for an arbitrary collection $x_1,\ldots,x_n \subset \mathbb{R}^d$ such a point need not exist, and it is surprising that for a typical i.i.d. configuration, this property is satisfied by μ.

The fact that Theorem 5.9 implies Theorem 5.8 is straightforward. Indeed, Theorem 5.9 implies that $\mathrm{diam}(T_\mu) \leq 2r$ and that if $\|a-\mu\| \geq r$, then $\mu \in T_a$. By the definition of T_a, one always has $a \in T_a$, and thus if $\|a-\mu\| > 2r$ then $\mathrm{diam}(T_a) > 2r$. Therefore, the minimizer $\widehat{\mu}$ must satisfy that $\|\widehat{\mu}-\mu\| \leq 2r$, as required.

The constants appearing in Theorem 5.8 are certainly not optimal. They were obtained with the goal of making the proof transparent.

The proof of Theorem 5.9 is based on the following idea. The mean μ is the minimizer of the function $f(x) = \mathbb{E}\|X-x\|^2$. A possible approach is to use the available data to guess, for any pair $a,b \in \mathbb{R}^d$, whether $f(a) < f(b)$. To this end, we may set up a "tournament" as follows.

Recall that $[n]$ is partitioned into k disjoint blocks $B_1,\ldots,B_k$ of size $m = n/k$. For $a,b \in \mathbb{R}^d$, we say that *a defeats b* if

$$\frac{1}{m}\sum_{i\in B_j}\left(\|X_i-b\|^2 - \|X_i-a\|^2\right) > 0$$

on more than $k/2$ blocks B_j. The main technical lemma is the following.

Lemma 5.10. *Let* $\delta \in (0,1)$*,* $k = \lceil 200\log(2/\delta)\rceil$*, and define*

$$r = \max\left(800\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}, 240\sqrt{\frac{\lambda_{max}\log(2/\delta)}{n}}\right).$$

With probability at least $1-\delta$*,* μ *defeats all* $b \in \mathbb{R}^d$ *such that* $\|b-\mu\| \geq r$*.*

Proof. Note that

$$\|X_i - b\|^2 - \|X_i - \mu\|^2 = \|X_i - \mu + \mu - b\|^2 - \|X_i - \mu\|^2 = -2\langle X_i - \mu, b - \mu\rangle + \|b - \mu\|^2 ,$$

set $\overline{X} = X - \mu$ and put $v = b - \mu$. Thus, for a fixed b that satisfies $\|b - \mu\| \geq r$, μ defeats b if

$$-\frac{2}{m}\sum_{i \in B_j}\left\langle \overline{X}_i, v\right\rangle + \|v\|^2 > 0$$

on the majority of blocks B_j.

Therefore, to prove our claim we need that, with probability at least $1 - \delta$, for every $v \in \mathbb{R}^d$ with $\|v\| \geq r$,

$$-\frac{2}{m}\sum_{i \in B_j}\left\langle \overline{X}_i, v\right\rangle + \|v\|^2 > 0 \tag{5.11}$$

for more than $k/2$ blocks B_j. Clearly, it suffices to show that (5.11) holds when $\|v\| = r$.

Consider a fixed $v \in \mathbb{R}^d$ with $\|v\| = r$. By Chebyshev's inequality, with probability at least 9/10,

$$\left|\frac{1}{m}\sum_{i \in B_j}\left\langle \overline{X}_i, v\right\rangle\right| \leq \sqrt{10}\sqrt{\frac{\mathbb{E}\left\langle \overline{X}, v\right\rangle^2}{m}} \leq \sqrt{10}\|v\|\sqrt{\frac{\lambda_{\max}}{m}} ,$$

where recall that $\lambda_{\max}$ is the largest eigenvalue of the covariance matrix of X. Thus, if

$$r = \|v\| \geq 4\sqrt{10}\sqrt{\frac{\lambda_{\max}}{m}} \tag{5.12}$$

then with probability at least 9/10,

$$-\frac{2}{m}\sum_{i \in B_j}\left\langle \overline{X}_i, v\right\rangle \geq \frac{-r^2}{2}. \tag{5.13}$$

Applying a Hoeffding's inequality (Theorem A.4 in the Appendix), we see that (5.13) holds for a single v with probability at least $1 - \exp(-k/50)$ on at least 8/10 of the blocks B_j.

Now we need to extend the above from a fixed vector v to all vectors with norm r. In order to show that (5.13) holds simultaneously for all $v \in r \cdot S^{d-1}$ on at least 7/10 of the blocks B_j, we first consider a maximal ϵ-separated set $V_1 \subset r \cdot S^{d-1}$ with respect to the $L_2(X)$ norm. In other words, V_1 is a subset of $r \cdot S^{d-1}$ of maximal

cardinality such that for all $v_1, v_2 \in V_1$, $\|v_1 - v_2\|_{L_2(X)} = \langle v_1 - v_2, \Sigma(v_1 - v_2)\rangle^{1/2} \geq \epsilon$. We may estimate this cardinality by the "dual Sudakov" inequality (see [50] and also [71] for a version with the specified constant) which implies that the cardinality of V_1 is bounded by

$$\log|V_1| \leq \left(\frac{\mathbb{E}\left[\langle G, \Sigma G\rangle^{1/2}\right]}{4\epsilon/r}\right)^2 ,$$

where G is a standard normal vector in $\mathbb{R}^d$. Notice that for any $a \in \mathbb{R}^d$, $\mathbb{E}_X \langle a, X\rangle^2 = \langle a, \Sigma a\rangle$, and therefore,

$$\begin{aligned}\mathbb{E}\left[\langle G, \Sigma G\rangle^{1/2}\right] &= \mathbb{E}_G\left[\left(\mathbb{E}_X\left[\left\langle G, \overline{X}\right\rangle^2\right]\right)^{1/2}\right] \leq \left(\mathbb{E}_X\mathbb{E}_G\left[\left\langle G, \overline{X}\right\rangle^2\right]\right)^{1/2} \\ &= \left(\mathbb{E}\left[\left\|\overline{X}\right\|^2\right]\right)^{1/2} = \sqrt{\mathrm{Tr}(\Sigma)}\,.\end{aligned}$$

Hence, by setting

$$\epsilon = (5/2)r\left(\frac{1}{k}\mathrm{Tr}(\Sigma)\right)^{1/2} , \tag{5.14}$$

we have $|V_1| \leq e^{k/100}$ and thus, by the union bound, with probability at least $1 - e^{-k/100} \geq 1 - \delta/2$, (5.13) holds for all $v \in V_1$ on at least 8/10 of the blocks B_j.

Next we check that property (5.11) holds simultaneously for all x with $\|x\| = r$ on at least 7/10 of the blocks B_j.

For every $x \in r \cdot S^{d-1}$, let v_x be the nearest element to x in V_1 with respect to the $L_2(X)$ norm. It suffices to show that, with probability at least $1 - \exp(-k/200) \geq 1 - \delta/2$,

$$\sup_{x \in r\cdot S^{d-1}} \frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{\left\{|m^{-1}\sum_{i\in B_j}\langle \overline{X}_i, x - v_x\rangle| \geq r^2/4\right\}} \leq \frac{1}{10}\,. \tag{5.15}$$

Indeed, on that event it follows that for every $x \in r \cdot S^{d-1}$, on at least 7/10 of the coordinate blocks B_j, both

$$-\frac{2}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, v_x\right\rangle \geq \frac{-r^2}{2} \quad \text{and} \quad 2\left|\frac{1}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, x\right\rangle - \frac{1}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, v_x\right\rangle\right| < \frac{r^2}{2}$$

hold and hence, on those blocks, $-\frac{2}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, x\right\rangle + r^2 > 0$ as required.

It remains to prove (5.15). Observe that

$$\frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{\left\{|m^{-1}\sum_{i\in B_j}\langle \overline{X}_i, x - v_x\rangle| \geq r^2/4\right\}} \leq \frac{4}{r^2}\frac{1}{k}\sum_{j=1}^{k}\left|\frac{1}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, x - v_x\right\rangle\right| .$$

Since $\|x-v_x\|_{L_2(X)} = (\mathbb{E}\left\langle \overline{X}, x-v_x\right\rangle^2)^{1/2} \le \epsilon$ it follows that for every j

$$\mathbb{E}\left|\frac{1}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, x-v_x\right\rangle\right| \le \sqrt{\frac{\mathbb{E}\left[\left\langle \overline{X}, x-v_x\right\rangle^2\right]}{m}} \le \frac{\epsilon}{\sqrt{m}}\,,$$

and therefore,

$$\begin{aligned}
&\mathbb{E}\sup_{x\in r\cdot S^{d-1}}\frac{1}{k}\sum_{j=1}^{k}\mathbb{1}_{\left\{|m^{-1}\sum_{i\in B_j}\langle \overline{X}_i, x-v_x\rangle|\ge r^2/4\right\}} \\
&\quad\le \quad \frac{4}{r^2}\mathbb{E}\sup_{x\in r\cdot S^{d-1}}\frac{1}{k}\sum_{j=1}^{k}\left(\left|\frac{1}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, x-v_x\right\rangle\right| - \mathbb{E}\left|\frac{1}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, x-v_x\right\rangle\right|\right) + \frac{4\epsilon}{r^2\sqrt{m}} \\
&\quad\stackrel{\text{def.}}{=}\quad (A)+(B)\,.
\end{aligned}$$

To bound (B), note that, by (5.14),

$$\frac{4\epsilon}{r^2\sqrt{m}} = 20\left(\frac{\mathrm{Tr}(\Sigma)}{n}\right)^{1/2}\cdot\frac{1}{r} \le \frac{1}{40}$$

provided that

$$r \ge 800\left(\frac{\mathrm{Tr}(\Sigma)}{n}\right)^{1/2}.$$

To bound (A), we may use standard techinques of emprical process theory, such as symmetrization, contraction for Bernoulli processes, and de-symmetrization (see, e.g., [50]) to show that

$$\begin{aligned}
(A) &\le \frac{8}{r^2}\mathbb{E}\sup_{x\in r\cdot S^{d-1}}\left|\frac{1}{n}\sum_{i=1}^{n}\left\langle \overline{X}_i, x-v_x\right\rangle\right| \le \frac{16}{r}\mathbb{E}\sup_{\{t:\|t\|\le 1\}}\left|\frac{1}{n}\sum_{i=1}^{n}\left\langle \overline{X}_i, t\right\rangle\right| \\
&\le \frac{16}{r}\cdot\frac{\mathbb{E}\left\|\overline{X}\right\|}{\sqrt{n}} = \frac{16}{r}\left(\frac{\mathrm{Tr}(\Sigma)}{n}\right)^{1/2} \le \frac{1}{40}
\end{aligned}$$

provided that $r \ge 640\left(\frac{\mathrm{Tr}(\Sigma)}{n}\right)^{1/2}$.

Thus, for

$$Z = \sup_{x\in r\cdot S^{d-1}}\frac{1}{k}\sum_{j=1}^{k}\mathbb{1}_{\left\{|m^{-1}\sum_{i\in B_j}\langle \overline{X}_i, x-v_x\rangle|\ge r^2/4\right\}}\,,$$

we have proved that $\mathbb{E}Z \le 1/20$. Finally, in order to prove (5.15), it suffices to prove that, $\mathbb{P}\{Z > \mathbb{E}Z + 1/20\} \le e^{-k/200}$, which follows from the bounded differences inequality (see, Theorem A.10 in the Appendix). □

Proof of Theorem 5.9

Theorem 5.9 is easily derived from Lemma 5.10. Fix a block B_j, and recall that $Y_j = \frac{1}{m}\sum_{i\in B_j} X_i$. Let $a, b \in \mathbb{R}^d$. Then

$$\begin{aligned}\frac{1}{m}\sum_{i\in B_j}\Big(\|X_i - a\|^2 - \|X_i - b\|^2\Big) &= \frac{1}{m}\sum_{i\in B_j}\Big(\|X_i - b - (a-b)\|^2 - \|X_i - b\|^2\Big)\\ &= -\frac{2}{m}\sum_{i\in B_j}\langle X_i - b, a - b\rangle + \|a-b\|^2 = (*)\end{aligned}$$

Observe that $-\frac{2}{m}\sum_{i\in B_j}\langle X_i - b, a-b\rangle = -2\Big\langle \frac{1}{m}\sum_{i\in B_j} X_i - b, a-b\Big\rangle = -2\Big\langle Y_j - b, a-b\Big\rangle$, and thus

$$\begin{aligned}(*) &= -2\Big\langle Y_j - b, a-b\Big\rangle + \|a-b\|^2\\ &= -2\Big\langle Y_j - b, a-b\Big\rangle + \|a-b\|^2 + \|Y_j - b\|^2 - \|Y_j - b\|^2\\ &= \|Y_j - b - (a-b)\|^2 - \|Y_j - b\|^2 = \|Y_j - a\|^2 - \|Y_j - b\|^2\,.\end{aligned}$$

Therefore, $(*) > 0$ (i.e., b defeats a on block B_j) if and only if $\|Y_j - a\| > \|Y_j - b\|$.

Recall that Lemma 5.10 states that, with probability at least $1-\delta$, if $\|a-\mu\| \geq r$ then on more than $k/2$ blocks B_j, $\frac{1}{m}\sum_{i\in B_j}\Big(\|X_i - a\|^2 - \|X_i - \mu\|^2\Big) > 0$, which, by the above argument, is the same as saying that for at least $k/2$ indices j, $\|Y_j - a\| > \|Y_j - \mu\|$.

5.3 The hidden hubs problem

To illustrate the use of the robust mean estimation techniques described in the previous section, we study a variant of the hidden clique problem. Consider the hypothesis testing problem in which one observes an $n \times n$ matrix $\boldsymbol{X}$ with independent entries $X_{i,j}$, $i, j \in [n]$. Under the null hypothesis, all entries are standard normal random variables. Under the alternative hypothesis, there exist $k < n$ rows of the matrix that have about k entries that are normally distributed with mean 0 and variance $\sigma^2 > 1$. More precisely, under the alternative hypothesis, there exists a set $S \subset [n]$ with $|S| = k$ such that, for all $i \in S$, $X_{i,j}$ is standard normal with probability $1 - k/n$ and normal $\mathcal{N}(0, \sigma^2)$ with probability k/n. If $i \notin S$, the $X_{i,j}$ have the standard normal distribution. To define the null and alternative hypotheses formally, introduce the notation, for $u \in \mathbb{R}$,

$$\phi_0(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \quad \text{and} \quad \phi_1(u) = \frac{1}{\sigma\sqrt{2\pi}} e^{-u^2/(2\sigma^2)}\,.$$

Under the null hypothesis, the joint density of $\boldsymbol{X} = (X_{i,j})_{n\times n}$ is

$$f_0(x) = \prod_{i,j\in[n]} \phi_0(x_{i,j}) \quad \text{for } x \in \mathbb{R}^{n^2} ,$$

while under the alternative hypothesis, $\boldsymbol{X}$ has density f_S for some set S of size k, where

$$f_S(x) = \prod_{i\in S}\prod_{j=1}^{n}\left(\left(1-\frac{k}{n}\right)\phi_0(x_{i,j}) + \frac{k}{n}\phi_1(x_{i,j})\right)\prod_{i\notin S}\prod_{j=1}^{n}\phi_0(x_{i,j}) .$$

There are various differences between this model and the Gaussian hidden clique problem discussed in Chapter 2. First, while "typical" entries of the matrix are standard normal $\mathcal{N}(0,1)$ in both models, "atypical" entries are normal $\mathcal{N}(0,\sigma)$ instead of $N(\mu,1)$. Thus, instead of a shifted mean, the variance is increased. Also, the atypical values are not forced to form a $k\times k$ submatrix but rather they are placed in k rows at arbitrary positions. The fact that the number of atypical entries per row is random $\sim \mathrm{B}in(n,k/n)$ instead of k is a minor difference assumed only for convenience.

In this illustrative example we only address the case when $\sigma^2 = 2$. In a sense this is a threshold case mostly due to the fact that the expected value $\mathbb{E}_S[\phi_1(X_{i,j})/\phi_0(X_{i,j})]$ of the likelihood ratio is infinite for $i \in S$ if and only if $\sigma^2 \geq 2$. Indeed, Kannan and Vempala [47] prove that the problem becomes much easier when $\sigma^2 \geq 2$. Here we show a simplified analysis of the test of Kannan and Vempala, made possible by the robust mean estimation techniques presented in this chapter.

To introduce the proposed test, consider the "likelihood ratio"

$$L(X_{i,j}) = \frac{\phi_1(X_{i,j})}{\phi_0(X_{i,j})}$$

for each entry $X_{i,j}$ of the matrix $\boldsymbol{X}$. Under the null hypothesis, we clearly have

$$\mathbb{E}_0 L(X_{i,j}) = \int L(u)\phi_0(u)du = 1 ,$$

while under the alterative, if $i \in S$,

$$\mathbb{E}_S L(X_{i,j}) = \int L(u)\phi_1(u)du = 1 - \frac{k}{n} + \frac{k}{n}\int L(u)\phi_1(u)du .$$

For $i \notin S$, $\mathbb{E}_S L(X_{i,j}) = \mathbb{E}_0 L(X_{i,j}) = 1$. Since

$$\int L(u)\phi_1(u)du = \int \frac{1}{\sigma^2\sqrt{2\pi}} e^{-u^2/4} e^{u^2/2} e^{-u^2/4} du = \infty ,$$

$\mathbb{E}_S L(X_{i,j}) = \infty$ whenever $i \in S$. (Note that the same holds not only when $\sigma^2 = 2$ but also whenever $\sigma^2 \geq 2$. Thus, it is a natural idea to base a test on estimating the expected value of L. In order to make sure that the expected value is well defined and to control the fluctuations of $L(X_{i,j})$, we truncate its value appropriately.

Let M be a positive number whose value is chosen below and introduce

$$B_{i,j} = \exp\left(\frac{\min\left(X_{i,j}^2, M^2\right)}{4}\right).$$

The proposed test T_n works as follows. For each row $i \in [n]$ of the matrix $\boldsymbol{X}$, compute the median-of-means estimator $\widehat{\mu_i}$ of the values $B_{i,1}, \ldots, B_{i,n}$, with confidence parameter $\delta \in (0,1)$. Accept the null hypothesis if and only if less than $k/2$ rows have a large estimated mean. More precisely, the test $T_n : \mathbb{R}^{n^2} \to \{0,1\}$ is defined by

$$T_n(\boldsymbol{X}) = 0 \quad \text{if and only if} \quad \left|\left\{i \in [n] : \widehat{\mu_i} > \sqrt{2} + 8(2/\pi)^{1/4}\sqrt{\frac{M\log(1/\delta)}{n}}\right\}\right| < \frac{k}{2}.$$

The test has two parameters: M is the level of truncation and δ is the confidence level in the definition of the median-of-means estimator. The next result shows that in the "critical" case when $\sigma^2 = 2$, the test is correct with high probability provided that k is at least $cn^{1/2}\log^{1/4} n$. We refer to Kannan and Vempala [47] for a thorough analysis on the entire range of values of σ^2.

Theorem 5.16. *Consider the test defined above with parameters $M = \sqrt{2\log n}$ and $\delta = 1/n$. Then for any $S \subset [n]$ of cardinality k,*

$$\mathbb{P}_0\{T_n(\boldsymbol{X}) = 1\} + \mathbb{P}_S\{T_n(\boldsymbol{X}) = 0\} \leq 2e^{-3k/16},$$

whenever $k \geq 32n^{1/2}\log^{1/4} n$.

Proof. In order to analyze our test, we first bound the type I error $\mathbb{P}_0\{T_n(\boldsymbol{X}) = 1\}$. To this end, we first estimate the expectation and variance of $B_{i,j}$ under both hypotheses. Under the null, for all $i, j \in [n]$,

$$\mathbb{E}_0 B_{i,j} = \int \frac{1}{\sqrt{2\pi}}\exp\left(\frac{\min(u^2, M^2)}{4} - \frac{u^2}{2}\right)du \leq \int \frac{1}{\sqrt{2\pi}}\exp\left(\frac{u^2}{4} - \frac{u^2}{2}\right)du = \sqrt{2}.$$

Also,

$$\begin{aligned}\mathbb{E}_0 B_{i,j} &\geq \int_{-M}^{M} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{4}\right)du \\ &= \sqrt{2}\left(1 - \mathbb{P}\{\mathcal{N}(0,1) \geq M/\sqrt{2}\right) \geq \sqrt{2}\left(1 - e^{-M^4/4}\right) = \sqrt{2}\left(1 - \frac{1}{n}\right).\end{aligned}$$

On the other hand,

$$\begin{aligned}\mathbb{E}_0 B_{i,j}^2 &= \int \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\min(u^2, M^2)}{2} - \frac{u^2}{2}\right) du \\ &= \sqrt{\frac{2}{\pi}} M + e^{M^2/2} \mathbb{P}\{|\mathcal{N}(0,1)| > M^2\} \le \sqrt{\frac{2}{\pi}} M + 2 \,,\end{aligned}$$

where we used the standard tail estimate of the standard normal distribution $\mathbb{P}\{\mathcal{N}(0,1) > M^2\} \le e^{-M^2/2}$. (This follows easily from the Chernoff bound.) Hence,

$$\mathrm{Var}_0(B_{i,j}) = \mathbb{E}_0 B_{i,j}^2 - \left(\mathbb{E}_0 B_{i,j}\right)^2 \le \sqrt{\frac{2}{\pi}} M \,.$$

Thus, the performance bound Theorem 5.2 of the median-of-means estimator guarantees that, for all $i \in [n]$, the probability that

$$\widehat{\mu_i} > \sqrt{2} + 8(2/\pi)^{1/4} \sqrt{\frac{M \log n}{n}}$$

is at most $1/n$. Thus, the number of indices i with this property is dominated by a binomial random variable $\mathrm{B}in(n, 1/n)$. Since, by Bernstein's inequality (Theorem A.5 in the Appendix)

$$\mathbb{P}\left\{\mathrm{B}in(n,1/n) > \frac{k}{2}\right\} \le \mathbb{P}\left\{\mathrm{B}in(n,1/n)) - \mathbb{E}\mathrm{B}in(n,1/n) > \frac{k}{4}\right\} \le e^{-3k/16} \,,$$

we have

$$\mathbb{P}_0\{T_n(\boldsymbol{X}) = 1\} \le e^{-3k/16} \,.$$

It remains to examine the behavior of the test under the alternative hypothesis.

Naturally, for $i \notin S$, we have $\mathbb{E}_S B_{i,j} = \mathbb{E}_0 B_{i,j}$ and $\mathrm{Var}_S(B_{i,j}) = \mathrm{Var}_0(B_{i,j})$ for all $j \in [n]$.

On the other hand, if $i \in S$, then for all $j \in [n]$, by a computation similar to the above,

$$\begin{aligned}\mathbb{E}_S B_{i,j} &= \left(1 - \frac{k}{n}\right) \mathbb{E}_0 B_{i,j} + \frac{k}{n} \int \frac{1}{2\sqrt{\pi}} \exp\left(\frac{\min(u^2, M^2)}{4} - \frac{u^2}{4}\right) du \\ &= \left(1 - \frac{k}{n}\right) \mathbb{E}_0 B_{i,j} + \frac{k}{n} \left(\frac{M}{\sqrt{\pi}} + e^{M^2/4} \mathbb{P}\{|\mathcal{N}(0,2)| > M\}\right) \\ &\ge \left(1 - \frac{k}{n}\right)\left(1 - \frac{1}{\sqrt{n}}\right)\sqrt{2} + \frac{M}{\sqrt{\pi}} \frac{k}{n} \\ &\ge \sqrt{2} + \frac{M - 3}{\sqrt{\pi}} \frac{k}{n} \,.\end{aligned}$$

Also,

$$\mathbb{E}_S B_{i,j}^2 \;=\; \left(1-\frac{k}{n}\right)\mathbb{E}_0 B_{i,j}^2 + \frac{k}{n}\int \frac{1}{2\sqrt{\pi}}\exp\left(\frac{\min(u^2,M^2)}{2}-\frac{u^2}{4}\right)du\ .$$

Since

$$\begin{aligned}
&\int \frac{1}{2\sqrt{\pi}}\exp\left(\frac{\min(u^2,M^2)}{2}-\frac{u^2}{4}\right)du \\
&= \int_{|u|\le M}\frac{1}{2\sqrt{\pi}}\exp\left(\frac{u^2}{4}\right)du + \exp\left(\frac{M^2}{2}\right)\int_{|u|\ge M}\frac{1}{2\sqrt{\pi}}\exp\left(-\frac{u^2}{4}\right)du \\
&\le \frac{1}{\sqrt{\pi}}\int_0^M \exp\left(\frac{Mu}{4}\right)du + \frac{1}{M\sqrt{\pi}}\exp\left(\frac{M^2}{2}\right)\int_M^\infty u\exp\left(-\frac{u^2}{4}\right)du \\
&\le \frac{4}{M\sqrt{\pi}}e^{M^2/4} + \frac{1}{M\sqrt{\pi}}e^{M^2/4} \\
&= \frac{5}{M\sqrt{\pi}}e^{M^2/4}\ ,
\end{aligned}$$

we have, for all $i \in S$ and $j \in [n]$,

$$\mathrm{Var}_S(B_{i,j}) \le \left(1-\frac{k}{n}\right)\left(\sqrt{\frac{2}{\pi}}M+2\right) + \frac{k}{n}\frac{5}{M\sqrt{\pi}}e^{M^2/4}\ .$$

Thus, invoking Theorem 5.2 again, we have that, if $i \in S$, then with probability at least $1-1/n$,

$$\widehat{\mu_i} \;>\; \sqrt{2} + \frac{M-3}{\sqrt{\pi}}\frac{k}{n} - 8\sqrt{\frac{\log n}{n}}\sqrt{\left(\sqrt{\frac{2}{\pi}}M+2\right)+\frac{k}{n}\frac{5}{M\sqrt{\pi}}e^{M^2/4}}$$

The right-hand side of the inequality is grater than

$$\sqrt{2} + 8(2/\pi)^{1/4}\sqrt{\frac{M\log n}{n}}$$

whenever

$$\frac{Mk}{n} > 32\sqrt{\frac{M}{n}\log n} \quad \text{and} \quad \frac{Mk}{n} > 36\sqrt{\frac{k}{Mn^2}e^{M^2/4}\log n}$$

(with a generous bounding of the constants for convenience). Both inequalities are easily seen to hold with $M = \sqrt{2\log n}$ if $k \ge 32n^{1/2}\log^{1/4} n$. Hence, under the alternative hypothesis, the number of indices i for which $\widehat{\mu}_n > \sqrt{2} + 8(2/\pi)^{1/4}\sqrt{\frac{M\log n}{n}}$ is at least as large as a binomial $\mathrm{B}in(k, 1-1/n)$ random variable. Once again, Bernstein's inequality may be invoked to conclude that the probability that this number is smaller than $k/2$ is at most $e^{-3k/16}$, concluding the proof of $\mathbb{P}_S\{T_n(\boldsymbol{X})=0\} \le e^{-3k/16}$. □

5.4 Bibliographic notes

The median-of-means estimator has been proposed in different forms in various papers, see Nemirovsky and Yudin [58], Hsu [38], Jerrum, Valiant, and Vazirani [44], Alon, Matias, and Szegedy [3].

Theorem 5.3 and the lower bound of Theorem 5.4 is from Devroye, Lerasle, Lugosi, and Oliveira [25] The first inequality of Theorem 5.4 appears in Bubeck, Cesa-Bianchi, and Lugosi [16].

The problem of constructing estimators with sub-Gaussian performance that do not depend on the confidence lever δ was studied in Devroye, Lerasle, Lugosi, and Oliveira [25].

Lerasle and Oliveira [51], Hsu and Sabato [39], and Minsker [56] extend the median-of-means estimator to more general spaces.

The median-of-means tournament estimator and the Theorem 5.8 appear in Lugosi and Mendelson [52].

The hidden hubs problem was introduced by Kannan and Vempala [47]. The solution presented here is a simplified version of their techniques.

5.5 Exercises

Exercise 5.1. *Prove Theorem 5.4.*

Exercise 5.2. *Let Y be a standard normal random variable. Show that, for all $\lambda < 1/2$,*

$$\log \mathbb{E} e^{\lambda Y^2} = \frac{1}{2}\left(-\log(1-2\lambda)\right) \leq \lambda + \frac{\lambda^2}{1-2\lambda} \, .$$

Exercise 5.3. (GEOMETRIC MEDIAN-OF-MEANS ESTIMATOR). *Let $X_1, \ldots, X_n$ be i.i.d. random vectors in $\mathbb{R}^d$ with mean μ and covariance matrix Σ. Partition $[n] = \{1, \ldots, n\}$ into k blocks $B_1, \ldots, B_k$ of size $|B_i| \geq \lfloor n/k \rfloor \geq 2$. For $j = 1, \ldots, k$, let*

$$Y_j = \frac{1}{m} \sum_{i \in B_j} X_i \, .$$

Let $\widehat{\mu}_n$ be the geometric median of $Y_1, \ldots, Y_k$. Show that $\widehat{\mu}_n$ satisfies an inequality similar to the estimator of Proposition 5.6 (Minsker [56]).

Appendices

Appendix A

Probability inequalities

Here we gather some of the probabilistic tools used in the text.

A.1 Chernoff bounds: concentration of sums of independent random variables

First of all, recall *Markov's inequality*: for any nonnegative random variable X, and $t > 0$,

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}X}{t} .$$

An easy application of this is *Chebyshev's inequality*: if X is an arbitrary random variable and $t > 0$, then

$$\mathbb{P}\{|X - \mathbb{E}X| \geq t\} = \mathbb{P}\left\{|X - \mathbb{E}X|^2 \geq t^2\right\} \leq \frac{\mathbb{E}\left[|X - \mathbb{E}X|^2\right]}{t^2} = \frac{\mathrm{Var}(X)}{t^2}.$$

One often obtains sharper bounds by a more clever use of Markov's inequality. The idea is simple: by Markov's inequality, if s is an arbitrary positive number, then for any random variable X, and any $t > 0$,

$$\mathbb{P}\{X \geq t\} = \mathbb{P}\{e^{sX} \geq e^{st}\} \leq \frac{\mathbb{E}e^{sX}}{e^{st}} .$$

One may now pick the value of s that yields the sharpest bound. The obtained inequality is often called the *Chernoff bound*:

$$\mathbb{P}\{X \geq t\} \leq \inf_{s>0} \frac{\mathbb{E}e^{sX}}{e^{st}} .$$

As an illustration, consider the case when X is a standard normal random variable. Then $\mathbb{E}e^{sX} = e^{s^2/2}$ and the Chernoff bound implies that

$$\mathbb{P}\{X \geq t\} \leq \inf_{s>0} e^{s^2/2 - st} = e^{-t^2/2} . \tag{A.1}$$

This bound is quite sharp, though it may slightly be improved by noticing that

$$\mathbb{P}\{X \geq t\} = \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \leq \frac{1}{t\sqrt{2\pi}} \int_t^\infty x e^{-x^2/2} dx = \frac{1}{t\sqrt{2\pi}} \left[-e^{-x^2/2}\right]_t^\infty = \frac{1}{t\sqrt{2\pi}} e^{-t^2/2} .$$

The Chernoff bound is especially convenient for obtaining tail bounds for sums of independent random variables. Define $S_n = \sum_{i=1}^n X_i$, where $X_1, \ldots, X_n$ are independent real-valued random variables. Then

$$\begin{aligned} \mathbb{P}\{S_n - \mathbb{E}S_n \geq t\} &\leq e^{-st} \mathbb{E}\left[\exp\left(s \sum_{i=1}^n (X_i - \mathbb{E}X_i)\right)\right] \\ &= e^{-st} \prod_{i=1}^n \mathbb{E}\left[e^{s(X_i - \mathbb{E}X_i)}\right] \quad \text{(by independence).} \end{aligned} \tag{A.2}$$

Hence, the problem of finding tight tail bounds for S_n boils down to finding good upper bounds for the moment generating function of the random variables $X_i - \mathbb{E}X_i$. For bounded random variables perhaps the most elegant version is due to Hoeffding:

Lemma A.3. *Let X be a random variable with $\mathbb{E}X = 0$, $a \leq X \leq b$. Then for $s > 0$,*

$$\mathbb{E}\left[e^{sX}\right] \leq e^{s^2(b-a)^2/8} .$$

Proof. Note that by convexity of the exponential function

$$e^{sx} \leq \frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa} \quad \text{for } a \leq x \leq b.$$

Exploiting $\mathbb{E}X = 0$, and introducing the notation $p = -a/(b-a)$ we get

$$\begin{aligned} \mathbb{E}e^{sX} &\leq \frac{b}{b-a} e^{sa} - \frac{a}{b-a} e^{sb} \\ &= \left(1 - p + pe^{s(b-a)}\right) e^{-ps(b-a)} \\ &\stackrel{\text{def.}}{=} e^{\phi(u)}, \end{aligned}$$

where $u = s(b-a)$, and $\phi(u) = -pu + \log(1 - p + pe^u)$. But by straightforward calculation it is easy to see that the derivative of ϕ is

$$\phi'(u) = -p + \frac{p}{p + (1-p)e^{-u}} ,$$

therefore $\phi(0) = \phi'(0) = 0$. Moreover,

$$\phi''(u) = \frac{p(1-p)e^{-u}}{(p + (1-p)e^{-u})^2} \leq \frac{1}{4} .$$

Thus, by Taylor's theorem, for some $\theta \in [0,u]$,

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8} .$$

Now we may directly plug this lemma into (A.2):

$$\begin{aligned}
&\mathbb{P}\{S_n - \mathbb{E}S_n \geq t\} \\
&\quad \leq \quad e^{-st} \prod_{i=1}^{n} e^{s^2(b_i-a_i)^2/8} \quad \text{(by Lemma A.3)} \\
&\quad = \quad e^{-st} e^{s^2 \sum_{i=1}^n (b_i-a_i)^2/8} \\
&\quad = \quad e^{-2t^2/\sum_{i=1}^n (b_i-a_i)^2} \quad \text{(by choosing } s = 4t/\textstyle\sum_{i=1}^n (b_i-a_i)^2) .
\end{aligned}$$

The result we have just derived is generally known as *Hoeffding's inequality* ([37]):

Theorem A.4. (HOEFFDING'S INEQUALITY). *Let $X_1, \ldots, X_n$ be independent bounded random variables such that X_i falls in the interval $[a_i, b_i]$ with probability one. Then for any $t > 0$ we have*

$$\mathbb{P}\{S_n - \mathbb{E}S_n \geq t\} \leq e^{-2t^2/\sum_{i=1}^n (b_i-a_i)^2}$$

and

$$\mathbb{P}\{S_n - \mathbb{E}S_n \leq -t\} \leq e^{-2t^2/\sum_{i=1}^n (b_i-a_i)^2} .$$

A version that is often useful when the variance of the summands is small is the following inequality, often called *Bernstein's inequality*.

Theorem A.5. (BENNETT'S INEQUALITY.) *Let $X_1, \ldots, X_n$ be independent random variables with finite variance such that $X_i \leq b$ for some $b > 0$ almost surely for all $i \leq n$. Let*

$$S = \sum_{i=1}^{n} (X_i - \mathbb{E}X_i)$$

and $v = \sum_{i=1}^n \mathbb{E}\left[X_i^2\right]$. Then for all $t > 0$,

$$\mathbb{P}\{S \geq t\} \leq \exp\left(-\frac{t^2}{2(v + bt/3)}\right) .$$

Let N, b, and n be positive integers with $N > n$ and $N > b$. A random variable X taking values on the integers $0, 1, \ldots, b$ is *hypergeometric* with parameters N, b and n, if

$$\mathbb{P}\{X = k\} = \frac{\binom{b}{k}\binom{N-b}{n-k}}{\binom{N}{n}}, \quad k = 1, \ldots, b.$$

X models the number of blue balls in a sample of n balls drawn without replacement from an urn containing b blue and $N-b$ red balls. The next tail bound is due to Hoeffding [37]:

Theorem A.6. *Let the set A consist of N numbers $a_1,\dots,a_N$. Let $Z_1,\dots,Z_n$ denote a random sample taken without replacement from A, where $n \le N$. Denote*

$$m = \frac{1}{N}\sum_{i=1}^{N} a_i \quad \text{and} \quad c = \max_{i,j\le N} |a_i - a_j|.$$

Then for any $\epsilon > 0$ we have

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - m\right| \ge \epsilon\right\} \le 2e^{-2n\epsilon^2/c^2}.$$

Specifically, if X is hypergeometrically distributed with parameters N, b, and n, then

$$\mathbb{P}\{|X - b| \ge n\epsilon\} \le 2e^{-2n\epsilon^2}.$$

For more inequalities of this type, see Hoeffding [37] and Serfling [63].

Theorem A.7. *Let χ^2_d denote a random variable with χ^2 distribution with d degrees of freedom. Then*

$$\mathbb{P}\{\chi^2_d < d - 2\sqrt{dt}\} \le e^{-t}$$

and

$$\mathbb{P}\{\chi^2_d > d + 2\sqrt{dt} + 2t\} \le e^{-t}$$

(see, e.g., Massart [53]).

A.2 Concentration inequalities for functions of independent random variables

A.2.1 Efron-Stein inequality

Theorem A.8. (EFRON-STEIN INEQUALITY.) *Suppose $X_1,\dots,X_n$ are independent random variables taking values in some set $\mathcal{X}$ and let $f : \mathcal{X}^n \to \mathbb{R}$. Denote $X = (X_1,\dots,X_n)$. Let $X' = (X'_1,\dots,X'_n)$ be an independent copy of X. Denoting $X^{(i)} = (X_1,\dots,X_{i-1},X'_i,X_{i+1},\dots,X_n)$,*

$$\mathrm{Var}(f(X)) \le \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\left[\left(f(X) - f(X^{(i)})\right)^2\right].$$

The theorem is due to Steele [64], building on earlier work of Efron and Stein [28]. The beautiful proof presented here is due to Sourav Chatterjee.

Proof. Introduce the notation

$$X^{[i]} = (X_1', \dots, X_i', X_{i+1}, \dots, X_n) .$$

In particular, $X^{[0]} = X$ and $X^{[n]} = X'$. Let $f : \mathcal{X}^n \to \mathbb{R}$ be such that $\mathbb{E}f(X) = 0$ and let $g : \mathcal{X}^n \to \mathbb{R}$. Then

$$\mathbb{E}[g(X)f(X)] = \mathbb{E}[g(X)(f(X) - f(X'))] = \sum_{i=1}^{n} \mathbb{E}\Big[g(X)\Big(f(X^{[i-1]}) - f(X^{[i]})\Big)\Big]$$

Notice that $g(X)(f(X^{[i-1]}) - f(X^{[i]}))$ has the same distribution as that of $-g(X^{(i)})(f(X^{[i-1]}) - f(X^{[i]}))$. Thus,

$$\mathbb{E}[g(X)f(X)] = \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\Big[\Big(g(X) - g(X^{(i)})\Big)\Big(f(X^{[i-1]}) - f(X^{[i]})\Big)\Big]$$

By taking $g = f$, we obtain the variance fomula

$$\mathrm{Var}(f(X)) = \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\Big[\Big(f(X) - f(X^{(i)})\Big)\Big(f(X^{[i-1]}) - f(X^{[i]})\Big)\Big] .$$

The theorem now follows from the Cauchy-Schwarz inequality. □

A simple but useful corollary is the following upper bound for the variance of "self-bounding" functions. For more information on such inequalities, see Boucheron, Lugosi, and Massart [14].

We say that $f : \mathcal{X}^n \to [0, \infty)$ is *self bounding* if there is a function $f_{n-1} : \mathcal{X}^{n-1} \to [0, \infty)$ such that, for all $x \in \mathcal{X}^n$,

$$f(x) - f_{n-1}(x^{(-i)}) \in [0, 1] \quad \text{and} \quad \sum_{i=1}^{n} \Big(f(x) - f_{n-1}(x^{(-i)})\Big) \le f(x) ,$$

where $x^{(-i)} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$.

Theorem A.9. (SELF-BOUNDING FUNCTION INEQUALITY.) *Let $X_1, \dots, X_n$ be independent random variables taking values in some set $\mathcal{X}$ and let $f : \mathcal{X}^n \to \mathbb{R}$ be a self-bounding function. Then*

$$\mathrm{Var}(f(X)) \le \mathbb{E}f(X) .$$

A.2.2 Bounded differences inequality

Let $\mathcal{X}$ be a measurable set. We say that a function $f : \mathcal{X}^n \to \mathbb{R}$ has the *bounded differences property* if for some nonnegative constants $c_1, \ldots, c_n$,

$$\sup_{\substack{x_1,\ldots,x_n,\\ x_i' \in \mathcal{X}}} |f(x_1,\ldots,x_n) - f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)| \leq c_i \ , \ 1 \leq i \leq n \ .$$

The following inequality is one of the most basic and important concentration inequalities for functions of independent random variables, see, for example [14, Theorem 6.2].

Theorem A.10. (BOUNDED DIFFERENCES INEQUALITY.) *Assume that the function f satisfies the bounded differences assumption with constants $c_1, \ldots, c_n$ and denote*

$$v = \sum_{i=1}^{n} c_i^2 \ .$$

Let $X_1, \ldots, X_n$ be independent random variables taking values in $\mathcal{X}$ and let $Z = f(X_1, \ldots, X_n)$. Then

$$\mathbb{P}\{Z - \mathbb{E}Z > t\} \leq e^{-2t^2/v} \ .$$

A.2.3 Gaussian concentration inequality

The following Gaussian concentration inequality is due to by Tsirelson, Ibragimov, and Sudakov [69] who proved it using arguments based on stochastic calculus. See Ledoux [49] and Boucheron, Lugosi, and Massart [14] for more information.

Theorem A.11. (GAUSSIAN CONCENTRATION INEQUALITY.) *Let $X = (X_1, \ldots, X_n)$ be a vector of n independent standard normal random variables. Let $L > 0$ and let $f : \mathbb{R}^n \to \mathbb{R}$ denote an L-Lipschitz function, that is, a f is such that for all $x, y \in \mathbb{R}^n$,*

$$|f(x) - f(y)| \leq L\|x - y\| \ .$$

Then, for all $t > 0$,

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

Appendix B

Empirical process techniques

B.1 Covering numbers

Let $(\mathcal{X}, d)$ be a metric space and let $\epsilon > 0$. A set $A \subset \mathcal{X}$ is an *ϵ-net* of $\mathcal{X}$ if for all $x \in \mathcal{X}$ there exists an $y \in A$ such that $d(x, y) \leq \epsilon$.

If $\mathcal{X}$ has a finite ϵ-net, then one may define the *ϵ-covering number* $N(\mathcal{X}, \epsilon)$ of $\mathcal{X}$ as the cardinality ϵ-net.

A set $A \subset \mathcal{X}$ is an *ϵ-packing* (or ϵ-separated set) if for all $x \neq y \in A$, $d(x, y) \geq \epsilon$. The *ϵ-packing number* $M(\mathcal{X}, \epsilon)$ is the number of points in the ϵ-packing with largest cardinality.

The next lemma follows immediately from the definitions.

Lemma B.1. *For all $\mathcal{X}$ and $\epsilon > 0$,*

$$M(\mathcal{X}, \epsilon/2) \leq N(\mathcal{X}, \epsilon) \leq M(\mathcal{X}, \epsilon) .$$

Next we estimate the packing numbers of the Euclidean sphere $\mathbb{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$.

Theorem B.2. *For every $\epsilon > 0$,*

$$M\left(\mathbb{S}^{d-1}, \epsilon\right) \leq \left(1 + \frac{2}{\epsilon}\right)^d .$$

Proof. Let $A \subset \mathbb{S}^{d-1}$ be a maximal ϵ-packing. Then A is an ϵ-covering of $\mathbb{S}^{d-1}$ because otherwise one could add a point to A while keeping it ϵ-separated. Then the balls of radius $\epsilon/2$, centered at points of A are pairwise disjoint. Since all these balls are inside the ball centered at the origin, of radius $1 + \epsilon/2$, denoting the volume of the unit ball in $\mathbb{R}^d$ by V, we have

$$M\left(\mathbb{S}^{d-1}, \epsilon\right) \cdot \left(\frac{\epsilon}{2}\right)^d V \leq \left(1 + \frac{\epsilon}{2}\right)^d V .$$

Rearranging, we obtain the announced bound. □

B.2 A maximal inequality

We start with a simple maximal inequality for sub-Gaussian random variables.

Theorem B.3. *Let $X_1, \dots, X_n$ be random variables such that for all i and $s > 0$, $\mathbb{E}e^{sX} \leq e^{s^2/2}$. Then*

$$\mathbb{E} \max_{i=1,\dots,n} X_i \leq \sqrt{2 \log n} \, .$$

Proof. Let $s > 0$. Then

$$e^{s\mathbb{E}\max_{i=1,\dots,n} X_i} \leq \mathbb{E}e^{s \max_{i=1,\dots,n} X_i} \leq \sum_{i=1}^{n} \mathbb{E}e^{sX_i} \leq ne^{s^2/2} \, .$$

Taking logarithms of both sides and optimizing the value of s yields the announced bound. □

Bibliography

[1] Louigi Addario-Berry, Nicolas Broutin, Luc Devroye, Gábor Lugosi, et al. On combinatorial testing problems. *The Annals of Statistics*, 38(5):3063–3092, 2010.

[2] N. Alon, M. Krivelevich, and B. Sudakov. Finding a large hidden clique in a random graph. *Random Structures and Algorithms*, 13:457–466, 1999.

[3] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58:137–147, 2002.

[4] E. Arias-Castro, S. Bubeck, and G. Lugosi. Detection of correlations. *Annals of Statistics*, 40:412–435, 2012.

[5] E. Arias-Castro, S. Bubeck, and G. Lugosi. Detecting positive correlations in a multivariate sample. *Bernoulli*, 21:209–241, 2015.

[6] E. Arias-Castro, E. Candès, and A. Durand. Detection of an anomalous cluster in a network. *The Annals of Statistics*, pages 278–304, 2011.

[7] Ery Arias-Castro, Emmanuel J Candès, Hannes Helgason, and Ofer Zeitouni. Searching for a trail of evidence in a maze. *The Annals of Statistics*, pages 1726–1757, 2008.

[8] S. Balakrishnan, M. Kolar, A. Rinaldo, A. Singh, and L. Wasserman. Statistical and computational tradeoffs in biclustering. In *NIPS 2011 Workshop on Computational Trade-offs in Statistical Learning*, volume 4, 2011.

[9] B. Barak, S.B. Hopkins, J. Kelner, P. Kothari, A. Moitra, and A. Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 428–437. IEEE, 2016.

[10] Y. Baraud. Non asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8:577–606, 2002.

[11] I. Benjamini, R. Lyons, Y. Peres, and O. Schramm. Uniform spanning forests. *The Annals of Probability*, 29:1–65, 2001.

[12] Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *COLT*, pages 1046–1066, 2013.

[13] B. Bollobás. *Random Graphs*. Cambridge University Press, Cambridge, UK, 2001.

[14] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities:A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[15] A. Brieden, P. Gritzmann, R. Kannan, V. Klee, L. Lovász, and M. Simonovits. Deterministic and randomized polynomial-time approximation of radii. *Mathematika. A Journal of Pure and Applied Mathematics*, 48(1-2):63–105, 2001.

[16] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59:7711–7717, 2013.

[17] S. Bubeck, J. Ding, R. Eldan, and M. Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms*, 49:503–532, 2016.

[18] Sébastien Bubeck and Shirshendu Ganguly. Entropic clt and phase transition in high-dimensional wishart matrices. *International Mathematics Research Notices*, page rnw243, 2016.

[19] C. Butucea and Y. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688, 2013.

[20] Fan RK Chung and Linyuan Lu. *Complex graphs and networks*. American mathematical society, Providence, 2006.

[21] Y. Dekel, O. Gurel-Gurevich, and Y. Peres. Finding hidden cliques in linear time with high probability. In *Proceedings of the Meeting on Analytic Algorithmics and Combinatorics*, pages 67–75. Society for Industrial and Applied Mathematics, 2011.

[22] Y. Deshpande and A. Montanari. Finding hidden cliques of size\ sqrt {N/e} in nearly linear time. *Foundations of Computational Mathematics*, 15(4):1069–1128, 2015.

[23] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

[24] L. Devroye, A. György, G. Lugosi, and F. Udina. High-dimensional random geometric graphs and their clique number. *Electron. J. Probab.*, 16:2481–2508, 2011.

[25] L. Devroye, M. Lerasle, G. Lugosi, and R.I. Oliveira. Sub-Gausssian mean estimators. *Annals of Statistics*, 2016.

[26] D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32:962–994, 2004.

[27] Richard Durrett. *Random graph dynamics*. Cambridge university press Cambridge, 2007.

[28] B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9:586–596, 1981.

[29] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[30] P. Erdős and A. Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

[31] T. Feder and M. Mihail. Balanced matroids. In *STOC '92: Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*, pages 26–38, New York, NY, USA, 1992. ACM.

[32] U. Feige and D. Ron. Finding hidden cliques in linear time. In *21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'10)*, pages 189–204. Discrete Mathematics and Theoretical Computer Science, 2010.

[33] V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *Journal of the ACM (JACM)*, 64(2):8, 2017.

[34] Z. Füredi and J. Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1:233–241, 1981.

[35] E.N. Gilbert. Random plane networks. *Journal of the Society for Industrial and Applied Mathematics*, 9:533–543, 1961.

[36] G.R. Grimmett and S.N. Winkler. Negative association in uniform forests and connected graphs. *Random Structures & Algorithms*, 24:444–460, 2004.

[37] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[38] D. Hsu. Robust statistics. http://www.inherentuncertainty.org/2010/12/robust-statistics.html, 2010.

[39] D. Hsu and S. Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17:1–40, 2016.

[40] Y.I. Ingster. Minimax detection of a signal for l_p^n-balls. *Mathematical Methods of Statistics*, 7:401–428, 1999.

[41] S. Janson, T. Łuczak, and A. Ruciński. *Random Graphs*. John Wiley, New York, 2000.

[42] M. Jerrum. Large cliques elude the metropolis process. *Random Structures & Algorithms*, 3(4), 1992.

[43] M. Jerrum, A. Sinclair, and E. Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *Journal of the ACM*, 51:671–697, 2004.

[44] M. Jerrum, L. Valiant, and V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:186–188, 1986.

[45] T. Jiang and D. Li. Approximation of rectangular beta-laguerre ensembles and large deviations. *Journal of Theoretical Probability*, 28(3):804–847, 2015.

[46] R. Kannan and S. Vempala. Spectral algorithms. *Foundations and Trends in Theoretical Computer Science*, 4(3–4):157–288, 2009.

[47] R. Kannan and S. Vempala. Chi-squared amplification: Identifying hidden hubs. *arXiv preprint arXiv:1608.03643*, 2016.

[48] L. Kučera. Expected complexity of graph partitioning problems. *Discrete Applied Mathematics*, 57(2):193–212, 1995.

[49] M. Ledoux. *The concentration of measure phenomenon*. American Mathematical Society, Providence, RI, 2001.

[50] M. Ledoux and M. Talagrand. *Probability in Banach Space*. Springer-Verlag, New York, 1991.

[51] M. Lerasle and R. I. Oliveira. Robust empirical mean estimators. arXiv:1112.3914, 2012.

[52] G. Lugosi and S. Mendelson. Sub-Gaussian estimators of the mean of a random vector. *preprint*, 2016.

[53] P. Massart. *Concentration inequalities and model selection*. Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics. Springer, 2006.

[54] D.W. Matula. Employee party problem. In *Notices of the American Mathematical Society*, volume 19, pages A382–A382, 1972.

[55] R. Meka, A. Potechin, and A. Wigderson. Sum-of-squares lower bounds for planted clique. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 87–96. ACM, 2015.

[56] S. Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21:23082335, 2015.

[57] A. Montanari, D. Reichman, and O. Zeitouni. On the limitation of spectral methods: From the gaussian hidden clique problem to rank-one perturbations of gaussian tensors. In *Advances in Neural Information Processing Systems*, pages 217–225, 2015.

[58] A.S. Nemirovsky and D.B. Yudin. Problem complexity and method efficiency in optimization. 1983.

[59] E.M. Palmer. *Graphical Evolution*. John Wiley & Sons, New York, 1985.

[60] M. Penrose. *Random geometric graphs*, volume 5 of *Oxford Studies in Probability*. Oxford University Press, Oxford, 2003.

[61] J.G. Propp and D.B. Wilson. How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph. *Journal of Algorithms*, 27:170–217, 1998.

[62] M. Rácz and S. Bubeck. Basic models and questions in statistical network analysis. *arXiv preprint arXiv:1609.03511*, 2016.

[63] R.J. Serfling. Probability inequalities for the sum in sampling without replacement. *Annals of Statistics*, 2:39–48, 1974.

[64] J.M. Steele. An Efron-Stein inequality for nonsymmetric statistics. *The Annals of Statistics*, 14:753–758, 1986.

[65] J. Steinhardt. Does robustness imply tractability? a lower bound for planted clique in the semi-random model. *arXiv preprint arXiv:1704.05120*, 2017.

[66] X. Sun and A.B. Nobel. On the maximal size of large-average and anova-fit submatrices in a gaussian random matrix. *Bernoulli*, 19(1):275, 2013.

[67] M. Talagrand. *The generic chaining*. Springer, New York, 2005.

[68] M. Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

[69] B.S. Tsirelson, I.A. Ibragimov, and V.N. Sudakov. Norm of Gaussian sample function. In *Proceedings of the 3rd Japan-U.S.S.R. Symposium on Probability Theory*, volume 550 of *Lecture Notes in Mathematics*, pages 20–41. Springer-Verlag, Berlin, 1976.

[70] R. van der Hofstad. *Random graphs and complex networks*. Cambridge university press, 2016.

[71] R. Vershynin. Lectures in geometric functional analysis. 2009.

[72] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge University Press, 2012.