Average-case complexity and statistical inference

We give an example of the sort of question we will look at. The idea is to 'hide' some structure within a high-dimensional random object. We may then ask if we can *detect* this, i.e. whether we can distinguish the vanilla random object from the random object plus planted structure, also if we may *recover*, i.e. 'find' the planted structure from within the random object.

A favourite combinatorial random object is the Erdős–Rényi random graph, $G_{n,1/2}$, take n vertices and between each pair of vertices independently place an edge with probability 1/2. We are interested in the typical behaviour for large n. The structure we plant is a clique, i.e. between a subset S^* of the vertices of the graph, we place all the possible edges. Now, the random graph $G_{n,1/2}$ without the planted structure will naturally have some cliques by chance, and indeed the largest of these will have size approximately $2\log_2 n$ with probability tending to 1 as $n \to \infty$; which suggests it might not be possible to detect or recover a planted structure of size smaller than $2\log_2 n$. This turns out to be true, as we shall see. Interestingly, there is another phase transition. Fast algorithms finding the clique, e.g. picking the vertices of highest degrees, are only known when the planted clique has size about $n^{1/2}$ or higher; which is considerably larger than $2\log_2 n$. There is some 'evidence' that this $n^{1/2}$ threshold is fundamental: by evidence we mean rigorous statements we can prove which *suggest* that there are no polynomial time algorithms.

These ideas will be made precise in the course as we investigate what forms this evidence can take. We illustrate the techniques on three running examples, planted clique, as described above, as well a generalisation of it planted dense subgraph, and a Gaussian planted structure problem biclustering - see Appendix A for a list and the phase transitions in each model. This area is very active and many of the techniques and results presented here have been developed within the last decade and some within the last year.

1 Detection

1.1 Definitions

Problem Setup We specify a dimension n, and parameters (e.g. k size of planted structure, λ strength of 'signal', p, q probabilities of 'community' edges and 'non-community' edges respectively). For each fixed set of parameters we are interested in the behaviour for large n or as $n \to \infty$.

For a detection problem, under H_0 the *null hypothesis*, we sample from the probability space $\mathcal{Q}_n$ and under H_1 the *alternate hypothesis* we sample from the probability space $\mathcal{P}_n$. We write $\mathbb{P}_0(G = g)$ to denote the probability that a random sample G from probability distribution $\mathcal{Q}_n$ is the deterministic g (and similarly $\mathbb{P}_1(G = g)$ to denote the same for $\mathcal{P}_n$). We will try to stick to the convention of denoting random variables, random graphs or random matrices by capital letters and deterministic values, graphs and matrices by lower case letters.

A *test* is a function ϕ_n on the union of the supports of $\mathcal{Q}_n$ and $\mathcal{P}_n$, with $\phi_n(g) \in \{0, 1\}$[1]. We need a notion of how 'good' a test is at distinguishing $\mathcal{Q}_n$ and $\mathcal{P}_n$ and will use risk. The *risk* of a test ϕ, denoted $r(\phi)$ is

$$r(\phi) = \sum_{g\colon \phi(g)=1} \mathbb{P}_0(\phi(G) = g) + \sum_{g\colon \phi(g)=0} \mathbb{P}_1(\phi(G) = g) = \mathbb{P}_0(\phi(G) = 1) + \mathbb{P}_1(\phi(G) = 0)$$

[1] Suppose for now that ϕ_n is deterministic, later we may have random tests.

Observe that it is easy to design a function which achieves risk 1. We can take $\phi_{\text{guess null}}(g) = 0$ for all g in the support of $\mathcal{P}_n$ and $\mathcal{Q}_n$, or we could take the random test $\phi_p(g)$ which takes value 1 with probability p and 0 otherwise. Both of these have risk 1 for each input g.

We say a test ϕ_n achieves *strong detection* between H_0 and H_1 if $r(\phi_n) \to 0$ as $n \to \infty$. Similarly, we say a test ϕ_n achieves *weak detection* between H_0 and H_1 if there exists $\varepsilon > 0, n_0$ such that $r(\phi_n) < 1 - \varepsilon$ for all $n > n_0$.

We may now define what we mean by EASY and POSSIBLE detection. Say for $H_0 : \mathcal{Q}_n(\alpha, \beta)$ vs $H_1 : \mathcal{P}_n(\alpha, \beta)$ that *strong detection for H_0 vs H_1 is* EASY *for parameters α, β* if there exists a test ϕ_n implementable as a polynomial time algorithm such that $r(\phi_n) \to 0$ as $n \to \infty$. The definition for weak detection being EASY is similar, just replace the condition on $r(\phi_n)$ as required.

Say for $H_0 : \mathcal{Q}_n(\alpha, \beta)$ vs $H_1 : \mathcal{P}_n(\alpha, \beta)$ that *strong detection for H_0 vs H_1 is* POSSIBLE *for parameters α, β* if there exists a test ϕ_n such that $r(\phi_n) \to 0$ as $n \to \infty$. In particular ϕ_n may be a brute-force algorithm. The definition for weak detection being POSSIBLE is similar.

Later we will be able to talk about detection problems being HARD if they are POSSIBLE and we have 'evidence of hardness'. We will usually specify which evidence of hardness.

1.2 The spiked matrix model

We define a probability space BC_n over $n \times n$ matrices with real entries. The detection problem will be to distinguish between a matrix sampled from the 'null' where all entries normally distributed with mean zero and variance one, from the 'alternate' with a planted submatrix of expected size k which has entries with mean λ and variance one. See Fig 5.

Given the number of vertices n, total community size k and signal strength $\lambda > 0$, define the additive (independent) Gaussian model $\mathrm{BC}_n = \mathrm{BC}_n(k, \lambda)$ as follows. Under BC_n, independently for each $i \in [n] := \{1, 2, \ldots, n\}$, the community label σ_i is sampled such that $\sigma_i = 1$ with probability k/n and $\sigma_i = 0$ with probability $1 - k/n$, then set $X_{ij} = \sigma_i \sigma_j$ For each pair of vertices $i, j \in [n]$ the matrix entry Y_{ij} is sampled from

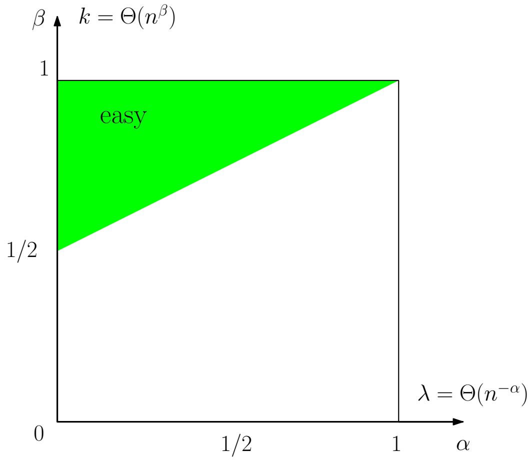
$$Y_{ij} | X_{ij} \sim \begin{cases} \mathcal{N}(\lambda, 1), & X_{ij} = 1 \quad \text{(i.e. } \sigma_i = \sigma_j = 1) \\ \mathcal{N}(0, 1), & X_{ij} = 0. \end{cases}$$

Note that if $ij \neq kl$ then $Y_{ij}|X_{ij}$ is independent from $Y_{kl}|X_{kl}$. We define the (fixed number) Gaussian model $\mathrm{BC}_n = \mathrm{BC'}_n(k, \lambda)$ as above except that we choose a set S uniformly from all subsets of $[n]$ of size k, and set $\sigma_i = 1$ if $i \in S$ and $\sigma_i = 0$ if $i \notin S$. Thus in this model we have exactly k 'community' indices.

1.3 When detection in matrix model is EASY

Lemma 1.1. *Fix α, β such that $0 < \alpha, \beta < 1$ and $\beta > \alpha/2 + 1/2$ (i.e. the green region in Fig* **??***). Let $\mathcal{Q}_n = \mathcal{Q}_n(\alpha, \beta)$ be $\mathrm{BC}(n, 0, 0)$ and $\mathcal{P}_n = \mathcal{P}_n(\alpha, \beta)$ be $\mathrm{BC}(n, k = n^{\beta}, \lambda = n^{-\alpha})$. Then the strong detection problem is* EASY *for α, β.*

end L1



(a) detection easy

Figure 1: Region considered in Lemma 1.1, see also Figure 5

Proof. (sketch)
We show the follow test distinguishes with high probability. Define $\phi_{\text{sum}} : \mathbb{R}^{n^2} \to \{0,1\}$ by

$$\phi_{\text{sum}}(x) = \begin{cases} 1, & \text{if } \sum_{i,j} x_{ij} > \frac{1}{2}\lambda k^2, \\ 0, & \text{otherwise.} \end{cases}$$

Recall that if $X_1 \sim N(\mu_1, s_1^2)$ and $X_2 \sim N(\mu_2, s_2^2)$ and X_1 and X_2 are independent then the sum is distributed $X_1 + X_2 \sim N(\mu_1 + \mu_2, s_1^2 + s_2^2)$. Note under H_0 we have n^2 variables distributed as $N(0,1)$, and under H_0 we have k^2 variables distributed as $N(\lambda, 1)$ and $n^2 - k^2$ distributed as $N(0,1)$, and thus

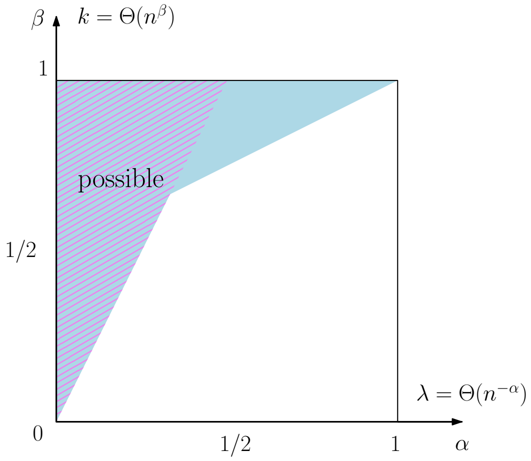
$$\sum_{i,j} X_{ij} = \begin{cases} N(0, n^2), & \text{under } H_0, \\ N(\lambda k^2, n^2), & \text{under } H_1. \end{cases}$$

The proof may now be completed using Lemma B.2[2].

□

Lemma 1.2. *Fix $\alpha, \beta \in (0,1)$ such that either of the following holds $\beta > \alpha/2 + 1/2$ or $\beta > 2\alpha$ (i.e. the shaded dashed and non-dashed regions in Fig 3). Let $\mathcal{Q}_n = \mathcal{Q}_n(\alpha,\beta)$ be $BC'(n,0,0)$ and $\mathcal{P}_n = \mathcal{P}_n(\alpha,\beta)$ be $BC'(n,k,\lambda)$. Then the strong detection problem is* POSSIBLE *for α, β.*

[2] Intuition: we succeed when the difference in means $\gg$ square-root of the variance. In this case the difference in means of the test statistic is $\lambda k^2 = n^{-\alpha+2\beta}$, and the variance is n^2 under both H_0 and H_1. So this intuition would say we succeed when $n^{-\alpha+2\beta} \gg n$, i.e. when $-\alpha + 2\beta > 1$, which is what we are aiming for.



(a) detection possible

Figure 2: Region considered in Lemma 1.2, since we already have a test for the green region in Figure 4, we need only prove there is a (brute-force) test for the dashed region. See also Figure 5

Proof. Note that by Lemma 1.1 we have a test which distinguishes whp for $\alpha, \beta \in (0,1)$ with $\beta > \alpha/2 + 1/2$ (i.e. the green region in Fig **??**). Thus it suffices to construct a test which distinguishes whp for fixed $\alpha, \beta \in (0,1)$ with $\beta > 2\alpha$ (i.e. the purple dashed area in Fig 3.

Let $\phi_{\text{search}} : \mathbb{R}^{n^2} \to \{0,1\}$ be defined by

$$\phi_{\text{search}}(x) = \begin{cases} 1, & \text{if } \max_{S\subset[n],\ |S|=k} \quad \sum_{i,j\in S} x_{ij} > \frac{1}{2}\lambda k^2, \\ 0, & \text{otherwise.} \end{cases}$$

Define the random variable $T_{\text{search}} = \max_{S\subset[n],\ |S|=k} \quad \sum_{i,j\in S} X_{ij}$, i.e. the quantity that is thresholded in ϕ_{search}.

Under H_0, T_{search} is the max of $\binom{n}{k}$ variables, $T_1, \ldots, T_{\binom{n}{k}}$ where each $T_i \sim N(0, n^2)$.

Recall that if $X_1 \sim N(\mu_1, s_1^2)$ then for constant a, we have $aX_1 \sim N(a\mu_1, a^2 s_1^2)$. Thus $n^{-1}T_i \sim N(0,1)$, and so we can apply Lemma C.1. (Note the T_i's are not independent, they are sums of overlapping parts of the matrix, but we don't need independence to apply the lemma.) By Lemma C.1 whp

$$T_{\text{search}} = \max_{i=1,\ldots,m} \frac{1}{n} T_i \leqslant \sqrt{(2+\varepsilon)\log(\binom{n}{k})} \leqslant \sqrt{(2+\varepsilon)k\log n}$$

where we used the fact that $\binom{n}{k} \leqslant n^k$.

Under H_1, again T_{search} is the max of $\binom{n}{k}$ variables, including the sum over the planted submatrix, and since it is a max it is at least as big as the sum over the planted set S^*. Let $T_{\text{planted}} = \sum_{i,j\in S^*} X_{i,j}$. Since T_{planted} is the sum of k^2 $N(\lambda, 1)$ random variables, $T_{\text{planted}} \sim N(\lambda k^2, k^2)$ and $\mathbb{E}[T_{\text{planted}}] = \lambda k^2$.

We want to show $T_{\text{planted}} > \lambda k^2/2$ whp. Since T_{planted} has expected value λk^2 it is enough to show that whp T_{planted} is at most $\lambda k^2/3$ away from its expectation - this shows whp $T_{\text{planted}} > 2\lambda k^2/3$ - see also the chat in Section B. By Lemma B.2, with $t = \lambda k^2/3$

$$\mathbb{P}(|T_{\text{planted}} - \lambda k^2| \geqslant \lambda k^2/3) \leqslant 2\exp(-(\lambda k^2/3)^2/(2k^2)) = 2\exp(-\lambda^2 k^2/6).$$

Now, note that $2\exp(-\lambda^2 k^2/6) = o(1)$ if $\lambda k \to \infty$, i.e. $n^{\beta-\alpha} \to \infty$ which is true whenever $\beta > \alpha$. Hence for $\beta > \alpha$ $T_{\text{planted}} > \lambda k^2/2$ whp. Since $T_{\text{search}} \geqslant T_{\text{planted}}$ always. Thus under H_1 whp $\phi_{search} = 1$ as required.

2 Planted Clique

Our approach for the possible and impossible regions of planted clique follows closely that of [2].

2.1 When planted clique is POSSIBLE

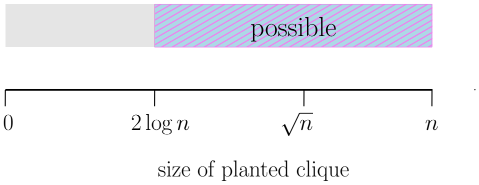


Figure 3: We show the detecting planted clique is possible in the dashed region in Lemma 2.1. We prove there is a (brute-force) test that distinguishes H_0: $G(n, 1/2)$ and H_1: $G'(n, k, 1/2)$ with high probability.

Lemma 2.1. *Let $k = k(n) > 2\log_2 n + 3$. Then for $H_0 : G(n, 1/2)$ vs $H_1 : G'(n, k, 1/2)$ strong detection is* POSSIBLE.

For graph g, define $\omega(g)$ to be the size of the largest clique in g, i.e. the size of the largest set of vertices S such that each pair of vertices in S is connected by an edge in graph g.

Proof. Our test will work by thresholding on the size of the largest clique in the graph. Let

$$\phi_n(g) = \begin{cases} 1, & \text{if } \omega(g) > 2\log_2 n + 3, \\ 0, & \text{otherwise.} \end{cases}$$

Then the risk of this test is

$$r(\phi_n) = \mathbb{P}_0(\phi_n = 1) + \mathbb{P}_1(\phi_n = 0) = \mathbb{P}_0\big(\omega(G) > 2\log_2 n + 3\big) + \mathbb{P}_1\big(\omega(G) \leqslant \log_2 n + 3\big).$$

Note that the size of the largest clique in the planted model is at least the size of the planted clique (it might be bigger if there is another vertex which happens to be connected to each vertex in the planted clique). Since, in P_n, we have planted a clique of size $2\log_2 n$, we have

$$\mathbb{P}_1(\phi_n(G) = 0) = \mathbb{P}_1(\omega(G) \leqslant \log_2 n + 3) = 0.$$

Thus the risk simplifies to consider only the size of the largest clique in $G(n, 1/2)$

$$r(\phi_n) = \mathbb{P}_0\big(\omega(G) > 2\log_2 n + 3\big).$$

Let N_m be the number of cliques of size m. Since $\mathbb{P}_0(N_m \geqslant 1) \leqslant \mathbb{E}_0[N_m]$ it suffices to bound the expected number of cliques. (This is an example of the 'first moment method'.)

Then we may calculate

$$\begin{aligned}\mathbb{E}_0[N_m] &= \sum_{S\colon |S|=m} \mathbb{P}_0\big(S \text{ is a clique in} G\big) \\ &= \binom{n}{m} 2^{-\binom{m}{2}} \\ &\leqslant n^m 2^{-m(m-1)/2} \\ &= (n2^{-(m-1)/2})^m.\end{aligned}$$

One may then check that for $m \leqslant 2\log_2 n + 3$ that $n2^{-(m-1)/2} \leqslant 1/2$. And thus for $m \leqslant 2\log_2 n + 3$

$$\mathbb{E}_0[N_m] \leqslant 2^{-m}.$$

and hence for $m \leqslant 2\log_2 n + 3$, $E_0[N_m] \to 0$ as $n \to \infty$.

Hence, since $\mathbb{P}_0(N_m \geqslant 1) \leqslant \mathbb{E}_0[N_m]$, we have $\mathbb{P}_0(\omega(G) \geqslant 2\log_2 n + 3) \to 0$ as $n \to \infty$ and thus we have that the risk of our test goes to zero as n goes to ∞ as required. □

end L2

3 Likelihood ratio and risk

begin L3

We define the likelihood ratio between discrete probability spaces $H_0 : Q$ and $H_1 : P$ by

$$L(g) = \frac{\mathbb{P}_1(G = g)}{\mathbb{P}_0(G = g)}. \tag{3.1}$$

Define $\phi^* = \phi^*(P, Q)$, the *likelihood ratio test* to be the following test

$$\phi^*(g) = \begin{cases} 1, & \text{if } L(g) \leqslant 1, \\ 0, & \text{if } L(g) > 1. \end{cases}$$

Lemma 3.1. *Suppose P and Q are discrete probability spaces. The test ϕ^* achieves minimal risk over tests to distinguish $H_0 : Q$ and $H_1 : P$.*

Proof. (in lectures). □

3.1 When planted clique is IMPOSSIBLE

(In lectures.) For further details see [2](pages 5-9) which we follow in this section.

end L3

4 Introduction to the low degree method

begin L4

We say function f *strongly*[3] *separates* $H_0 : Q_n$ and $H_1 : P_n$ if for all $\varepsilon > 0$, $\exists n_0$ such that

$$\max\{\sqrt{\mathrm{Var}_0[f]}, \sqrt{\mathrm{Var}_1[f]}\,\} \leqslant \varepsilon\big|\mathbb{E}_1[f] - \mathbb{E}_0[f]\big|.$$

The next lemma says that strong separation implies that there is a test with risk going to zero.

[3] weak separation is defined similarly, by replacing $\forall\varepsilon$ with $\exists C$. One may show that weak separation implies there is a test with risk strictly less than 1.

Lemma 4.1. *If f strongly separates $H_0 : Q_n$ and $H_1 : P_n$ then there exists a sequence of tests ϕ_n such that the risk of $r(\phi_n) \to 0$ as $n \to \infty$.*

Proof. Recall Chebyshev's inequality, for X a random variable with $\mathrm{Var}(X) = \sigma^2$ then

$$\mathbb{P}(|X - \mathbb{E}[X]| \geqslant t) \leqslant \frac{\sigma^2}{t^2}.$$

First note we may assume that $\mathbb{E}_1[f] \geqslant \mathbb{E}_0[f]$ because if not we may work with $-f$ instead.

Let the threshold be the midpoint of the expectations: $\tau = \frac{1}{2}(\mathbb{E}_1[f] - \mathbb{E}_0[f])$. And let the test be $\phi_f(G) = 1$ if $f(G) \geqslant \tau$ and $\phi_f(G) = 0$ otherwise.

Loosely, it is enough for the value $f(G)$ to be close $\mathbb{E}_0(f)$ to ensure $f(G)$ is small enough that the test classifies G as coming from the null, i.e. $\phi_f(G) = 0$. In particular, note that if we have $|f(G) - \mathbb{E}_0[f]| < \frac{1}{2}(\mathbb{E}_1[f] - \mathbb{E}_0[f])$ then $f(G) < \tau$ and thus $\phi_f(G) = 0$. Hence

$$\begin{aligned}
\mathbb{P}_0(\phi_f(G) = 1) &\leqslant \mathbb{P}_0\left(|f(G) - \mathbb{E}_0[f]| \geqslant \frac{1}{2}(\mathbb{E}_1[f] - \mathbb{E}_0[f])\right) \\
&\leqslant^{\mathrm{Ch}} \frac{\mathrm{Var}_0(f(G))}{\frac{1}{4}(\mathbb{E}_1[f] - \mathbb{E}_0[f])^2} \\
&\leqslant \frac{\varepsilon^2}{4} \quad \text{for all } n > n_0.
\end{aligned}$$

This shows that $\mathbb{P}_0(\phi_f(G) = 0) \to 1$ as $n \to \infty$. To show that $\mathbb{P}_1(\phi_f(G) = 1) \to 1$ as $n \to \infty$ is similar, one shows that $f(G)$ sufficiently close to $\mathbb{E}_1[f]$ implies that $\phi_f(G) = 1$ and one can thus bound the probability that $\mathbb{P}_1(\phi_f(G) = 0)$, details left to the reader. □

Given a sequence of hypothesis testing problems $H_0 : Q_n$ and $H_1 : P_n$ define the *(degree-D) advantage*, written $\mathrm{Adv}_{\leqslant \mathrm{D}}$, by

$$\mathrm{Adv}_{\leqslant D}(P_n, Q_n) = \max_{\deg f \leqslant D} \frac{\mathbb{E}_1[f]}{\sqrt{\mathbb{E}_0[f]}}. \tag{4.1}$$

Note we divide by the second moment under the null, rather than the variance under the null. Intuitively (4.1) can be thought of as the fluctuations in the planted model divided by the fluctuations in the null model. The term advantage is because it is meant to give a quantitative value for how much advantage over random guessing one can get by thresholding degree D polynomials.

We may then define notions of EASY and HARD for degree D polynomials. Say

$$\text{testing problem } H_0 : Q_n \text{ vs } H_1 : P_n \text{ is } \begin{cases} \text{EASY for degree } D \text{ if } \mathrm{Adv}_{\leqslant D}(P_n, Q_n) \to \infty \text{ as } n \to \infty \\ \text{HARD for degree } D \text{ if } \mathrm{Adv}_{\leqslant D}(P_n, Q_n) \to 0 \quad \text{as } n \to \infty. \end{cases}$$

From this we may finally define low degree polynomial hardness - basically we regard degree $\log n$ polynomials as 'low degree'. Say a testing problem $H_0 : Q_n$ vs $H_1 : P_n$ is EASY *for low degree polynomials* if there exists some constant C, such that it is easy for degree $D = C \log n$. Likewise we say a testing problem $H_0 : Q_n$ vs $H_1 : P_n$ is HARD *for low degree polynomials* (or 'low degree hard') if there exists some constant C', such that it is HARD for degree $D = C' \log n$.

4.1 Linear algebra

Suppose for now that the null $H_0 : Q_n$ is a (sequence of) discrete probability space(s). We define a linear product of functions, with respect to the null:

$$\langle f, g\rangle = \mathbb{E}_0[fg] = \sum_G P_0(G)f(G)g(G).$$

Let $\tilde{Q}_n$ be supported on vectors $Y = (Y_i)_{i=1}^N$ of length $N = N(n)$ such that each co-ordinate takes values $+1$ and -1 independently with probability a half. Then for $\alpha \subseteq [N]$, i.e. α is a set of indicies, we may define

$$h_\alpha = \prod_{i\in\alpha} Y_i \tag{4.2}$$

.

Claim. The set of functions $\{h_\alpha\}_{\alpha\subseteq[N]}$ forms an orthonormal basis for $\tilde{Q}_n$. (proof in lectures)

We will need the following lemma.

Lemma 4.2. *For $H_0 : \tilde{Q}_n$ (as above) and $H_1 : P_n$ a discrete probability space and for $\{h_\alpha\}_{\alpha\subseteq D}$ (defined* (4.2)*),*

$$\mathrm{Adv}_{\leqslant D}(P_n, \tilde{Q}_n)^2 = \sum_{|\alpha|\leqslant D} \big(\mathbb{E}_1[h_\alpha(Y)]\big)^2$$

end L4

4.2 When planted clique is HARD

Note if $\{i, j\} \neq \{k, \ell\}$ then

$$\mathbb{P}_1[ij \in E \text{ and } k\ell \in E \mid S^* = S] = \mathbb{P}_1[ij \in E \mid S^* = S]\mathbb{P}_1[k\ell \in E \mid S^* = S]$$

(to prove this one can check all cases, e.g. $i, j, k \in S, \ell \notin S$ and all distinct etc).
In general, for any $\alpha \subseteq \binom{[n]}{2}$ we have

$$\mathbb{E}_1[\chi_\alpha(G)] = \sum_{\varnothing\subseteq S\subseteq[n]} \prod_{ij\in\alpha} \mathbb{E}_1[2A_{ij} - 1 \mid S^* = S]\mathbb{P}_1[S^* = S]$$

Now consider $\mathbb{E}_1[A_{ij} \mid S^* = S] = \begin{cases} 1 & \text{if } i, j \in S \\ 1/2 & \text{otherwise} \end{cases}$

and thus

$$\mathbb{E}_1[2A_{ij} - 1 \mid S^* = S] = \begin{cases} 1 & \text{if } i, j \in S \\ 0 & \text{if } \{i, j\} \nsubseteq S \end{cases}$$

so finally we have the expectation of $\chi_\alpha(G)$ conditional on the planted clique S^* begin on the vertex set S

$$\mathbb{E}_1[\chi_\alpha(G) \mid S^* = S] = \begin{cases} 1 & \text{if } V(\alpha) \subseteq S \\ 0 & \text{otherwise.} \end{cases}$$

From this we can calculate the non-conditional expectation to be

$$\mathbb{E}_1[\chi_\alpha(G)] = \sum_{S\subseteq[n]} \mathbf{1}[V(\alpha) \subseteq S] \cdot \mathbb{P}_1[S^* = S] = \mathbb{P}_1[V(\alpha) \subseteq S^*] = \left(\frac{k}{n}\right)^{|V(\alpha)}.$$

□

end L5
L6

5 Reductions

5.1 Worst-case reduction

Example of clique in $\bar{G}$ and vertex-cover in G given in lectures.

5.2 Average-case complexity

We say a probabilistic algorithm $\mathcal{A}$ *succeeds in worst-case with probability* $1-\varepsilon$ if it succeeds with probability at least $1-\varepsilon$ on all inputs.

Given a probability distribution P and a (random) input $X \sim P$, we say a probabilistic algorithm $\mathcal{A}$ *succeeds in the average case with probability* $1-\varepsilon$ if

$$\sum_x \mathbb{P}(X=x)\mathbb{P}(\mathcal{A} \text{ succeeds on } x) \geqslant 1-\varepsilon.$$

Note that an algorithm $\mathcal{A}$ which *always* succeeds on $1-\varepsilon$ proportion of inputs x, and *always* fails on an ε proportion of inputs x would succeed with probability $1-\varepsilon$ in the average case, but not in the worst case.

In lecture gave an average-case to worst-case reduction, on matrix-multiplication.

5.3 Reductions in total variation

We want to be able to say statements such as : solving the hypothesis testing problem $H_0 : Q_n$ vs $H_1 : P_n$ (say this is the submatrix problem) means we can solve the hypothesis testing problem $H_0' : Q_n'$ vs $H_1' : P_n'$ (say this is the planted clique problem). This will also be a form of reduction, given an observation X, a matrix say, for the hypothesis testing problem $H_0' : Q_n'$ vs $H_1' : P_n'$ the original problem is to determine if $X \sim P_n'$ (i.e. no planted submatrix) or if $X \sim Q_n'$. We will find a reduction, i.e. a map, r which takes the the observed matrix X to a graph $r(X)$ such that if $X \sim Q_n'$ (random matrix with $N(0,1)$ entries) then the random graph $r(X)$ is distributed *pretty much* like the random graph $G(n, 1/2)$ and conversely if $X \sim P_n'$ (random matrix with planted submatrix distributed as $N(\lambda, 1)$) then the random graph $r(X)$ is distributed *pretty much* like the planted clique random graph $G(n, k, 1/2)$.

To quantify what we mean by 'pretty much' we need to define total variation distance.

Let P and Q be two discrete probability distributions on the same space. Then the total variation distance between P and Q is

$$d_{\mathrm{TV}}(P,Q) = \sum_g |\mathbb{P}_P(g) - \mathbb{P}_Q(g)| = \sup_E \; \mathbb{P}_P(E) - \mathbb{P}_Q(E) = \inf_{(X,Y),\, X\sim P,\, Y\sim Q} \mathbb{P}(X \neq Y),$$

where the infimum is over all couplings (X,Y). Similarly, let P and Q be two continuous probability distributions on the same space with density functions f and g respectively. Then the total variation distance between P and Q is

$$d_{\mathrm{TV}}(P,Q) = \int |f(x) - g(x)| = \sup_E \; \mathbb{P}_P(E) - \mathbb{P}_Q(E) = \inf_{(X,Y),\, X\sim P,\, Y\sim Q} \mathbb{P}(X \neq Y).$$

end L6
L7

We will also use the notation $\mathcal{L}(X)$, by to denote the 'law of' X, i.e. the distribution of X.

Let P and P' be probability distributions on the same space. Suppose that $X \sim P$, $\mathcal{A}$ is an map/algorithm, we write

$$P \xrightarrow{\mathcal{A}}_{\varepsilon} P' \qquad \text{if} \qquad d_{\mathrm{TV}}(\mathcal{L}\big(\mathcal{A}(X)\big), P') \leqslant \varepsilon$$

Now we are ready to define reduction in total variation distance.

We say that $\mathcal{A}$ is a *reduction in total variation* ε from the hypothesis testing problem $H_0 : Q_n$ vs $H_1 : P_n$ to the hypothesis testing problem $H_0' : Q_n'$ vs $H_1' : P_n'$ if

$$P_n \xrightarrow{\mathcal{A}}_{\varepsilon} P_n'$$

and

$$Q_n \xrightarrow{\mathcal{A}}_{\varepsilon} Q_n'.$$

Lemma 5.1. *Suppose $\mathcal{A}$ is a reduction in total variation ε from the hypothesis testing problem $H_0 : Q_n$ vs $H_1 : P_n$ to the hypothesis testing problem $H_0' : Q_n'$ vs $H_1' : P_n'$. If test ϕ_n distinguishes $H_0 : Q_n$ vs $H_1 : P_n$ with risk $r(\phi_n) \leqslant \delta$ then the test $\phi \circ \mathcal{A}$ distinguishes $H_0' : Q_n'$ vs $H_1' : P_n'$ with risk $r(\phi_n \circ \mathcal{A}) \leqslant \delta + \varepsilon$.*

We now show two lemmas illustrating that reductions in total variation and d_{TV} interplay nicely.

Lemma 5.2. *Suppose we have probability spaces $P, P', \tilde{P}$ and algorithm $\mathcal{A}$ such that*

$$d_{\mathrm{TV}}(P, \tilde{P}) \leqslant \delta \quad \textit{and} \quad P \xrightarrow{\mathcal{A}}_{\varepsilon} P'.$$

Then

$$\tilde{P} \xrightarrow{\mathcal{A}}_{\delta+\varepsilon} P'.$$

Proof. Let $X \sim P$, $Y \sim \tilde{P}$ and $Z \sim P'$. By definition of d_{TV} and since $d_{\mathrm{TV}}(P, \tilde{P}) \leqslant \delta$ there exists a coupling (X, Y) such that $\mathbb{P}(X = Y) \geqslant 1 - \delta$. Again by definition of d_{TV} and since $P \xrightarrow{\mathcal{A}}_{\varepsilon} P'$ there must be a coupling $(Z, \mathcal{A}(X))$ such that $\mathbb{P}(Z = \mathcal{A}(X)) \leqslant \varepsilon$.

Now note that the coupling of (X, Y) implies we have a coupling $(\mathcal{A}(X), \mathcal{A}(Y))$ such that $\mathbb{P}(\mathcal{A}(X) = \mathcal{A}(Y)) \geqslant 1 - \delta$. Thus we have a distribution on $(\mathcal{A}(X), \mathcal{A}(Y), Z)$ such that with probability at least $1 - \varepsilon - \delta$, all entries are equal, i.e. we have $\mathcal{A}(X) = \mathcal{A}(Y) = Z$. Hence forgetting the first component of the joint distribution we have a coupling $(\mathcal{A}(Y), Z)$ which has equality with probability at least $1 - \varepsilon - \delta$ and thus $\tilde{P} \xrightarrow{\mathcal{A}}_{\delta+\varepsilon} P'$ as required.

□

Lemma 5.3. *Let P, P_1 and P_2 be three probability spaces, and $\mathcal{A}_1$ and $\mathcal{A}_2$ algorithms such that*

$$P \xrightarrow{\mathcal{A}_1}_{\varepsilon_1} P_1 \quad \textit{and} \quad P_1 \xrightarrow{\mathcal{A}_2}_{\varepsilon_2} P_2.$$

Then

$$P \xrightarrow{\mathcal{A}_2 \circ \mathcal{A}_1}_{\varepsilon_1+\varepsilon_2} P_2.$$

Let f_0 be the density of $N(0,1)$ and f_μ be the density of $N(\mu,1)$, i.e.

$$f_0 = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2) \qquad f_\mu = \frac{1}{\sqrt{2\pi}}\exp(-(x-\mu)^2/2)$$

Define the distribution ν via its density function as follows

$$f_\nu(x) \ \propto\ \big(2f_0(x) - f_\mu(x)\big)\mathbf{1}[2f_0(x) \geqslant f_\mu(x)] \tag{5.1}$$

(we write $\propto$ since one would need to normalise by the integral of the RHS of (5.1) in order to get a probability density function).

For two probability distributions P and P' denote by $\frac{1}{2}P + \frac{1}{2}P'$ the distribution where with probability $\frac{1}{2}$ one samples from P and with probability $\frac{1}{2}$ one samples from P', this is called the half-half mixture of P and P'.

Lemma 5.4 (follows from Lemma 14 of [1])**.** *Let ν be the probability distribution as defined in* (5.1)*. If $\mu \leqslant 1/(6\sqrt{\log n})$ then*

$$d_{\mathrm{TV}}(\frac{1}{2}\nu + \frac{1}{2}N(\mu,1), N(0,1)) = o(n^{-3}).$$

Lemma 5.5. *Suppose X_1 is independent of X_2 and Y_1 is independent of Y_2. Then*

$$d_{\mathrm{TV}}((X_1,X_2),(Y_1,Y_2)) \leqslant d_{\mathrm{TV}}(X_1,Y_1) + d_{\mathrm{TV}}(X_2,Y_2).$$

Lemma 5.6. *Suppose P and P' are two probability distributions on the same space and let $X \sim P$ and $Y \sim P'$. Suppose that there exist events $E_1, \dots, E_m$ which partition the space such that for all i we have*

$$\mathbb{P}_P(E_i) = \mathbb{P}_{P'}(E_i).$$

Then

$$d_{\mathrm{TV}}(X,Y) \leqslant \max_i\ d_{\mathrm{TV}}((X|E_i),(Y|E_i)).$$

Proof sketch. This lemma follows from the coupling definition of total variation distance. Define $\delta_i = d_{\mathrm{TV}}((X|E_i),(Y|E_i))$. Let c_i be a coupling of $(X|E_i)$ and $(Y|E_i)$ - we know that such a coupling c_i exists such that $(X|E_i) \neq (Y|E_i)$, 'c_i fails', with probability at most δ_i. Now note that since $\mathbb{P}_P(E_i) = \mathbb{P}_{P'}(E_i)$ for each i, we may build a coupling c from the set of couplings $\{c_i\}_i$. Now the probability that c fails is

$$\mathbb{P}(\text{`}c \text{ fails}') = \sum_i \mathbb{P}(\text{`}c_i \text{ fails}'|E_i)\mathbb{P}(E_i).$$

Now note $\sum_i \mathbb{P}(E_i)$, and thus the RHS is a weighted sum of the $\mathbb{P}(\text{`}c_i \text{ fails}'|E_i)$ with total weight 1, and thus is at most the max of the terms. Thus we have

$$\mathbb{P}(\text{`}c \text{ fails}') \leqslant \max_i \mathbb{P}(\text{`}c_i \text{ fails}'|E_i) = \max_i d_{\mathrm{TV}}((X|E_i),(Y|E_i)).$$

Hence we have exhibited a coupling between X and Y with failure probability at most the RHS above, and so $d_{\mathrm{TV}}(X,Y)$ is bounded above by the RHS above, as required. □

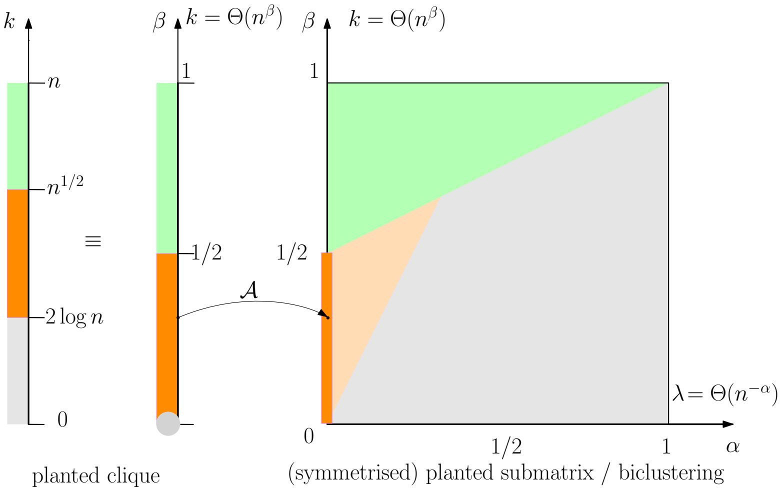


Figure 4: Reduction from planted clique to (symmetrised) planted submatrix biclustering, as in Lemma 5.7. The map $\mathcal{A}$ is a reduction in total variation distance from $G'(n, k = n^\beta)$ to $\widetilde{\mathrm{BC}}'(n, k = n^\beta, \lambda = n^{0.01})$.

5.4 Reduction from planted clique to a symmetrised planted submatrix

We define a symmetric variant $\widetilde{\mathrm{BC}}$ of the usual spiked matrix model BC introduced in Section 1.2. Define a probability space $\widetilde{\mathrm{BC}}_n$ over $n \times n$ matrices with real entries. Given the number of vertices n, total community size k and signal strength $\lambda > 0$, define $\widetilde{\mathrm{BC}}_n = \widetilde{\mathrm{BC}}_n(k, \lambda)$ as follows. Under $\widetilde{\mathrm{BC}}_n$, independently for each $i \in [n] := \{1, 2, \ldots, n\}$, the community label σ_i is sampled such that $\sigma_i = 1$ with probability k/n and $\sigma_i = 0$ with probability $1 - k/n$, then set $X_{ij} = \sigma_i \sigma_j$ For each pair of vertices $i < j$ the matrix entry Y_{ij} is sampled from

$$Y_{ij}|X_{ij} \sim \begin{cases} \mathcal{N}(\lambda, 1), & X_{ij} = 1 \quad \text{(i.e. } \sigma_i = \sigma_j = 1) \\ \mathcal{N}(0, 1), & X_{ij} = 0. \end{cases}$$

and we set $Y_{ji} = Y_{ij}$ and the diagonal entries to zero.

We define the (fixed number) symmetric Gaussian model $\widetilde{\mathrm{BC}'}_n = \widetilde{\mathrm{BC}'}_n(k, \lambda)$ as above except that we choose a set S uniformly from all subsets of $[n]$ of size k, and set $\sigma_i = 1$ if $i \in S$ and $\sigma_i = 0$ if $i \notin S$. Thus in this model we have exactly k 'community' indices.

Lemma 5.7.

end L7
L8

A List of Planted problems

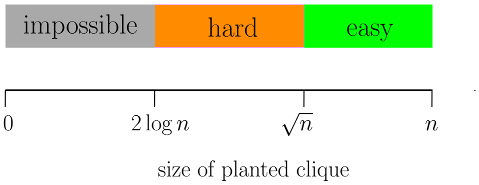
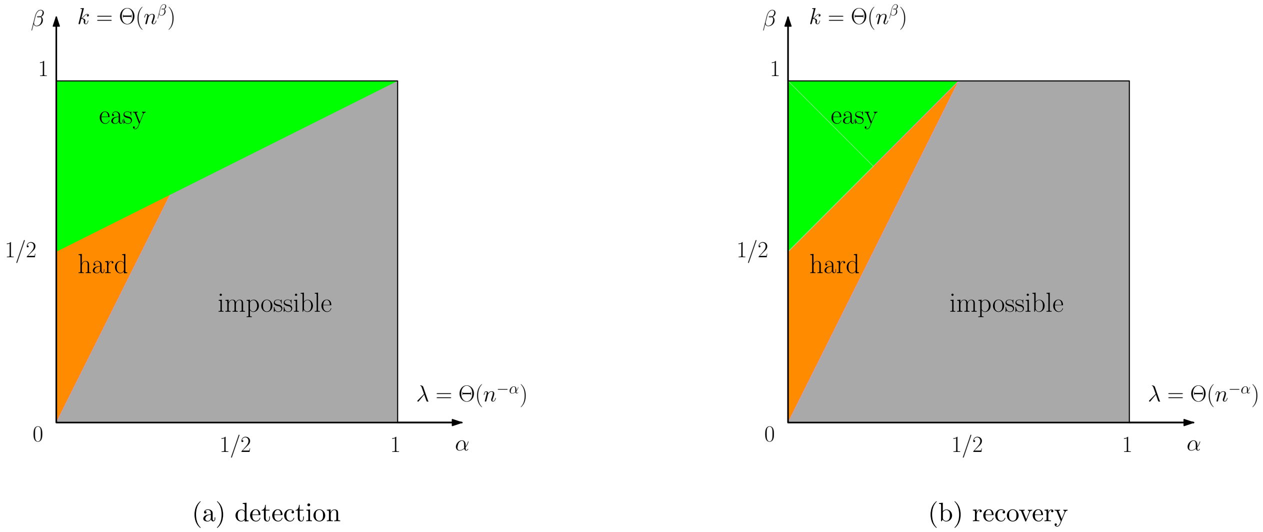


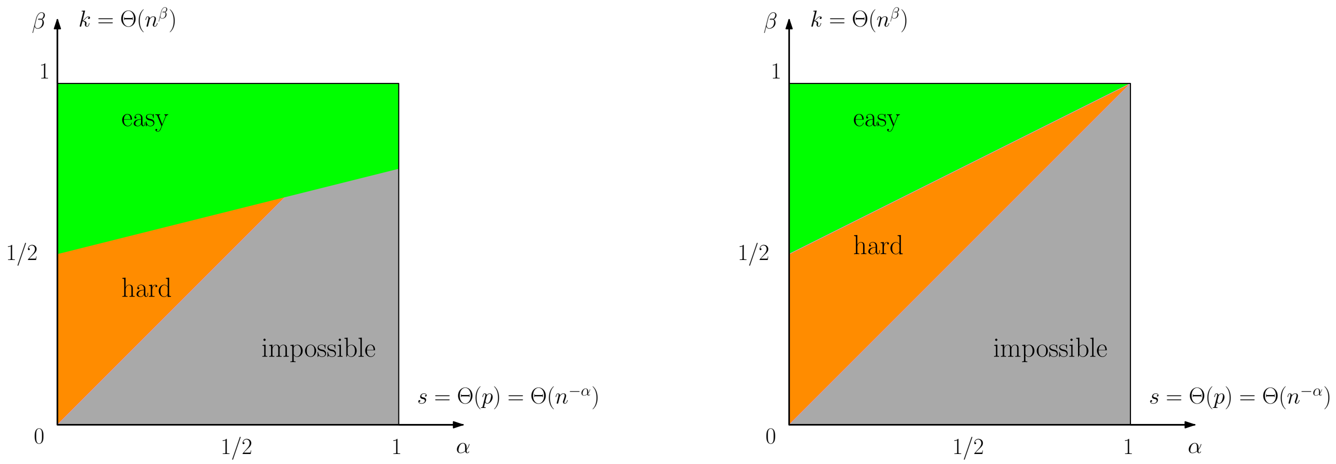
Figure 4: **Planted clique.**
H_0: $G(n, \frac{1}{2})$ random graph on n vertices where each edge is present independently with probability 1/2.
H_1: $G(n, k, \frac{1}{2})$, random graph on n vertices where each vertex is part of 'community' S independently with probability k/n. Each edge ij is present independently either with probability 1 if $i, j \in S$ or with probability 1/2 otherwise. We sometimes take $H_1 : G'(n, k\frac{1}{2})$ where S^* is chosen uniformly at random from all subsets of vertices of size k.



(a) detection

(b) recovery

Figure 5: **Spiked Matrix Model** (planted submatrix with elevated mean).
H_0: a random $n \times n$ matrix with each entry independent with distribution $N(0, 1)$.
H_1: $\mathrm{BC}(n, k, \lambda)$, an $n \times n$ matrix with each index in set S independently with probability k/n. Each entry independent with distribution $N(\lambda, 1)$ if $i, j \in S$ and with distribution $N(0, 1)$ otherwise.



(a) detection

(b) recovery

Figure 6: **Planted dense subgraph**.
H_0: $G(n, q)$ random graph on n vertices where each edge is present independently with probability q.
H_1: $G(n, k, q, s)$ with $s > 0$, random graph on n vertices where each vertex is part of 'community' S independently with probability k/n. Each edge ij is present independently either with probability $q + s$ if $i, j \in S$ or with probability q otherwise.

B Concentration Inequalities

Sometimes we are interested in a random variable X_n which is very likely to fall within some interval $[a_n, b_n]$ (and this can be very useful for us!). Often we can prove this statement in two steps. First we calculate the expected value. Let $c_n = \mathbb{E}[X_n]$ and suppose for simplicity that $c_n = (a_n + b_n)/2$. The second step is to show it is unlikely that X_n is far from its expected value c_n; i.e. to show $\mathbb{P}(|X_n - c_n| > (a_n - b_n)/2) \to 0$ as $n \to \infty$. Note these two steps together prove that X_n lies in $[a_n, b_n]$ whp, i.e. that $\mathbb{P}(a_n \leqslant X_n \leqslant b_n) \to 1$ as $n \to \infty$.

We say in this case (i.e. when the second step works), that a random variable is *concentrated about its mean* and refer to the bounds below as *concentration inequalities*. We will use these often so collect them in this section of the appendix for easy reference.

Lemma B.1 (Hoeffding's inequality)**.** *Let $S = X_1 + \ldots + X_n$ where $X_1, \ldots, X_n$ are independent and $a \leqslant X_i \leqslant b$ for all i. Then*

$$\mathbb{P}\big(\,|\,S - \mathbb{E}[S]\,|\, \geqslant t\big) \leqslant 2\exp\left(-\frac{2t^2}{n(a-b)^2}\right).$$

Lemma B.2. *Let $X \sim N(\mu, \sigma^2)$. Then*

$$\mathbb{P}\big(\,|\,X - \mathbb{E}[X]\,|\, \geqslant t\big) \leqslant 2\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

C Probability Background

We will use many properties of the distributions, we collate these here for reference while reading the proofs or doing exercises.

C.1 General Probability

We say a sequence of events E_n holds whp 'with high probability' if $\mathbb{P}(E_n) \to 1$ as $n \to \infty$.

C.2 Normal Distribution

The following lemma shows the max of m $N(0,1)$ variables is not too big. Note the variables $X_1, \ldots, X_m$ need not be independent.

Lemma C.1. *Let $\varepsilon > 0$. Suppose $X_1, \ldots, X_m \sim N(0,1)$. Then*

$$X_{\max} = \max_{i \in 1,\ldots,m} X_i \leqslant \sqrt{(2+\varepsilon)\log m}$$

with probability tending to 1 as $m \to \infty$.

D Helpful Combinatorial notation and inequalities

The notation $\binom{n}{k}$ read 'n choose k' is the number of ways to pick a set of k items from a set of n items,

$$\binom{n}{k} = \frac{n(n-1)\ldots(n-k+1)}{k(k-1)\ldots 1} = \frac{n!}{(n-k)!k!}$$

and

$$\frac{(n-k+1)^k}{k^k} \leqslant \binom{n}{k} \leqslant n^k.$$

D.1 Big 'o' notation

We will use notation $O(.)$, $o(.)$, $\omega(.)$ and $\Omega(.)$.

Index

References

[1] Matthew Brennan, Guy Bresler, and Wasim Huleihel. “Reducibility and Computational Lower Bounds for Problems with Planted Sparse Structure”. In: *Proceedings of the 31st Conference On Learning Theory.* Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 48–166. URL: `https://proceedings.mlr.press/v75/brennan18a.html`.

[2] Gábor Lugosi. “Lectures on combinatorial statistics”. In: *47th Probability Summer School, Saint-Flour* (2017), pp. 1–91.