Statistical inference on graphs: Selected Topics

Yihong Wu† Jiaming Xu‡

†Department of Statistics and Data Science, Yale University, New Haven, USA, yihong.wu@yale.edu
‡The Fuqua School of Business, Duke University, Durham, USA, jx77@duke.edu.
II Planted partition model 48

7 Detection threshold for SBM 49
7.1 Planted partition model and overview 49
7.2 Detection threshold for SBM 50
7.3 Test by counting cycles 53
  7.3.1 First moment calculation 54
  7.3.2 Variance analysis 55
7.4 Approximately counting cycles in polynomial time 56

8 Correlated recovery 59
8.1 Impossibility 59
8.2 Correlated recovery via spectral methods 61
8.3 Spectrum of non-backtracking matrices 63

9 Semidefinite programming II: general SBM 67
9.1 MLE and SDP relaxation 67
9.2 Proof of Theorem 9.2 70
  9.2.1 Proof of Lemma 9.2: Large deviation 71
9.3 Proof of Theorem 9.1 72

10 Ranking from Pairwise Comparisons (by Cheng Mao) 75
10.1 Modeling pairwise comparisons 75
  10.1.1 Models for probabilities of outcomes 75
  10.1.2 Sampling models 76
10.2 Kendall’s tau and minimax rates for noisy sorting 76
  10.2.1 Inversions and metric entropy 77
  10.2.2 Proof of the minimax upper bound 78
10.3 An efficient algorithm for noisy sorting 80
  10.3.1 Proof (sketch) of Theorem 10.2 81

11 Grothendieck Inequality and its Application in SBM 83
11.1 $\| \cdot \|_{\infty \to 1}$ norm 83
11.2 Grothendieck Inequality 84
11.3 Application to SBM 86

12 Statistical and Computational Limits 89
Part I

Clique problem
1. Introduction & Max Clique in Erdős-Rényi graphs

1. Introduction

1.1 Basic Definitions

A graph $G = (V, E)$ consists of

- A vertex set $V$. With loss of generality (WLOG), we shall assume $V = [n] \equiv \{0, 1, \ldots, n\}$ for some positive integer $n$.

- An edge set $E \subseteq \binom{V}{2}$. Each element of $E$ is an edge $e = (i, j)$ (unordered pair). We say $i$ and $j$ are connected and write $i \sim j$ if $(i, j) \in E$.

For the most part, we will be focusing on graphs that are undirected (i.e., edges do not have orientation) and simple (i.e., no multi-edges or self-loops).

Alternatively, one can also represent a graph as an adjacency matrix $A = (A_{ij})_{i,j \in [n]}$, which is an $n \times n$ symmetric binary matrix with zero diagonal. In particular, for a simple and undirected graph $G = (V, E)$, the entries $A_{ij}$ are defined as:

$$A_{ij} = \begin{cases} 1 & (i, j) \in E \\ 0 & \text{o.w.} \end{cases}.$$

Some basic concepts of graphs are defined as follows:

- The neighborhood of a given vertex $v \in V$ is defined as $N(v) = \{u \in V : u \sim v\}$, i.e., it is the set of vertices (neighbors) that are connected with $v$.

- The degree of $v$ is defined as $d_v = |N(v)|$, i.e., the number of neighbors of $v$.

- Induced subgraph: For any $S \subset V$, the subgraph induced by $S$ is defined as the graph $G[S] = (S, E_S)$, where $E_S \triangleq \{(u, v) \in E : u, v \in S\}$.

- A clique is a complete subgraph. A graph is complete iff every pair of vertices in the graph are connected.

1.1.2 Sample topics

The goal of statistical inference is to use data to make informed decisions (hypotheses testing, estimation, etc). The usual framework of statistical inference is the following:

$$\theta \in \Theta \mapsto X \mapsto \hat{\theta}.$$ 

The theoretical objectives of this class are two-fold:
1. Understand and characterize the fundamental (statistical) limits: What is possible/impossible information-theoretically?

2. Can statistical limits be attained computationally efficiently, e.g., in polynomial time? If yes, how? If not, why?

In this course,

- Data = graphs;
- Parameter = hidden (latent, or planted) structure;
- We will focus on large-graph limit (number of vertices $\to \infty$).

As a preview, we briefly describe two models that we will study below: the **Planted Clique Model** and the **Stochastic Block Model**.

**The Planted Clique Model**  Let $V$ be a vertex set and $n = |V|$, and let $k \leq n$ be a given positive integer. The edge set $E$ in a graph $G = (V, E)$ is generated in the following manner:

1. A set $S$ of $k$ vertices is selected out of $n$ vertices to form a clique (all possible edges between them are added to $E$).

2. Remaining edges are added independently with probability $\frac{1}{2}$.

Given the resulting graph $G = (V, E)$, the goal is to find the planted (hidden) clique $S$.

To start, notice that this set up follows a classical statistical framework: a sample (here, the graph $G$) is generated from a distribution (i.e., the random process described above), and we want to estimate a parameter of that distribution (here, the set $S$) via the sample (here, $G$).

A decision-theoretic setting is to consider the minimax framework for the worst-case analysis, in which the goal is to find an estimator $\hat{S} = \hat{S}(G)$ that correctly recover $S$ with probability close to 1, regardless of the true set $S$ used to generate the graph $G$. In other words,

$$\min_{\hat{S} \in \binom{[n]}{k}} \mathbb{P}_S \left[ \hat{S}(G) = S \right] \approx 1,$$

where $\mathbb{P}_S$ denotes the law of $G$ conditioned on the location of the planted clique $S$. Alternatively, one can consider the more relaxed Bayesian setting, assuming $S$ is drawn uniformly at random. Equivalently, this amounts to finding an $\hat{S}$ that preforms well on average:

$$\mathbb{E}_{S \sim \text{Unif}([n])} \mathbb{P}_S \left[ \hat{S}(G) = S \right] \approx 1.$$

**Remark 1.1.** For problems with symmetry, these two formulations are often equivalent, in the sense that

$$\sup_{\hat{S}} \min_{S \in \binom{[n]}{k}} \mathbb{P}_S \left[ \hat{S}(G) = S \right] = \sup_{\hat{S}} \mathbb{E}_{S \sim \text{Unif}([n])} \mathbb{P}_S \left[ \hat{S}(G) = S \right].$$

This follows from the permutation invariance of the model, which implies the least favorable prior is uniform.
The Stochastic Block Model (SBM)  
Given a vertex set $V$, suppose $V$ can be partitioned into two “communities” of equal size. Community membership is represented by a vector  
$$
\sigma = (\sigma_1, \ldots, \sigma_n) \in \{\pm 1\}^n,
$$
where $\sigma_i = \sigma_j$ means that $i$ and $j$ belong to the same community, and $\sum_{i=1}^n \sigma_i = 0$ because the size of the two communities are equal. An edge between two vertices $i, j \in V$ is added to $E$ according to the following probabilities:

$$
P[(i, j) \in E] = \begin{cases} 
p & \sigma_i = \sigma_j 
p q & \sigma_i \neq \sigma_j,
\end{cases}
$$

where $0 \leq p, q \leq 1$ (note that $p, q$ need not sum to 1). Thus, in this model, in-group ties and out-group ties have a different probability of forming. There are also several different statistical inference tasks associated with this problem that SBMs address. For example, if $p$ and $q$ are known, then our goal could be to estimate the parameter $\sigma$. Or, if $p$ and $q$ are unknown, then we may be interested in jointly estimating $p, q, \text{and } \sigma$.

1.2 Asymptotic Behavior of Max Clique in $G(n, \frac{1}{2})$

We start with the ensemble of the Erdős-Rényi graph: $G \sim G(n, p)$ is a graph on $n$ vertices where each pair of vertices is connected independently with probability $p$. Next, as a warmup, we will focus on the behavior of the maximum size of a clique in $G(n, \frac{1}{2})$.

In particular, let $G_n \sim G(n, \frac{1}{2})$. Define its clique number $\omega(G_n) \triangleq \text{size of the max clique in } G_n$. We will show that $\omega(G_n) \approx 2 \log_2 n$ for large $n$: for any $\epsilon > 0$, with high probability (whp),

$$
\omega(G_n) \leq (2 + \epsilon) \log_2 n, \quad (1.1)
$$
$$
\omega(G_n) \geq (2 - \epsilon) \log_2 n. \quad (1.2)
$$

In other words, $\frac{\omega(G_n)}{\log_2 n} \rightarrow 2$ in probability.

1.2.1 Proof of (1.1): First moment method

Let any $\epsilon > 0$ be given. We will show that $\mathbb{P}[\omega(G_n) \geq (2 + \epsilon) \log_2 n] \rightarrow 0$.

To start, consider any positive integer $k$, as well as any $S \subset [n]$ where $|S| = k$. Notice that there are $\binom{k}{2}$ possible edges that can form between the $k$ vertices in $S$, meaning that:

$$
P(G_n[S] \text{ is a } k\text{-clique}) = 2^{-\binom{k}{2}}.
$$

And, there are $\binom{n}{k}$ different sets of $k$ vertices in a graph with $n$ vertices. So, by the union bound,

$$
\mathbb{P}(\exists S \subset [n]: G_n[S] \text{ is a } k\text{-clique}) \leq \binom{n}{k} 2^{-\binom{k}{2}}.
$$
Now, let $k_0 = (2 + \epsilon) \log_2 n$. Again by the union bound, we have that:

$$
\mathbb{P}(\omega(G) \geq k_0) \leq \sum_{k=k_0}^{n} \binom{n}{k} 2^{-\binom{k}{2}}
$$

$$
\leq \sum_{k=k_0}^{n} \left(n2^{-\frac{(k_0-1)}{2}}\right)^k
$$

$$
\leq \sum_{k=k_0}^{\infty} \left(n2^{-\frac{(k_0-1)}{2}}\right)^k \leq 2(n2^{-\frac{(k_0-1)}{2}})^{k_0},
$$

where (a) follows from $\binom{n}{k} \leq n^k$ and $k_0 \leq k$, (b) follows from $n2^{-\frac{k_0-1}{2}} = \sqrt{2}n^{-\epsilon/2} < 1/2$ for sufficiently large $n$.

### 1.2.2 Proof of (1.2): Second moment method

We will now show that $\lim_{n \to \infty} \mathbb{P}[\omega(G_n) \geq k] \to 1$, where $k \triangleq (2 - \epsilon) \log_2 n$. Define:

$$
T_n \triangleq \# \text{ of cliques of size } k \text{ in } G_n = \sum_{|S|=k} 1 \{G_n[S] \text{ is a } k \text{ clique}\}. \quad (1.3)
$$

Note that if a graph contains at least one clique of size $k$, then the max clique must be of size $\geq k$, implying that $\mathbb{P}[\omega(G_n) \geq k] \geq \mathbb{P}[T_n > 0]$. So, to show that $\mathbb{P}[\omega(G_n) \geq k] \to 1$ as $n \to \infty$, it suffices to show instead that $\mathbb{P}[T_n > 0] \to 1$.

### Intuition

But, before trying to prove that $\mathbb{P}[T_n > 0] \to 1$, let’s first build some intuition. What we computed in the union bound is in fact computing the first moment of $\mathbb{E}[T_n]$. By linearity of expectation, we have

$$
\mathbb{E}[T_n] = \binom{n}{k} 2^{-\binom{k}{2}}. \quad (1.4)
$$

Clearly, when $k = (2 + \epsilon) \log_2 n$, $\mathbb{E}[T_n] \ll 0$, which implies that $\mathbb{P}[T_n > 0] \ll 0$ since $T_n$ is integer-valued. As $T_n$ is a positive random variable, it tempting to think that a sufficient condition for $\mathbb{P}[T_n > 0] \gg 0$ is $\mathbb{E}[T_n] \gg 0$. However, this direction is generally false: a counterexample would be a distribution that places almost all of its probability mass at zero, and the remaining very small amount of probability mass at, say, $10^{100}$. Indeed, while the expected value of a random variable with this distribution would be very large, the probability that this random variable is non-zero would still be very small.

So, to show that $\mathbb{P}[T_n > 0]$ is large, it won’t be enough to show that $\mathbb{E}[T_n]$ is large. What to do? Well, one way to characterize the distribution in the counterexample above is that it has very high variance. If we can show that the variance of $T_n$ is not so large, then that would essentially show that $T_n$’s distribution does not assign low probability to extremely high valued integers, essentially ruling out counterexamples like the one previously entertained. Will this be enough?

### Second Moment Method

As it turns out, this approach works and is called the **Second Moment Method**. Briefly, suppose $X_n$ is a non-negative, integer-valued random variable. In this approach, one shows that $\mathbb{P}[X_n > 0] \to 1$
by showing that:
\[ \text{Var}[X_n] = o(\mathbb{E}^2[X_n]), \]
where Var stands for variance. Since we are going to apply the Second Moment Method to show that \( \mathbb{P}[T_n > 0] \to 1 \), let’s first take a small detour to prove it works for the general random variable \( X_n \) describe above. And, the first step in doing so will be to prove the Paley-Zygmund inequality.

**Lemma 1.1 (Paley-Zygmund Inequality).** Let \( X \geq 0 \) be a random variable with \( 0 < \mathbb{E}[X^2] < \infty \). Then for any \( 0 \leq c \leq 1 \),
\[ \mathbb{P}(X > c\mathbb{E}[X]) \geq (1 - c)^2 \frac{\mathbb{E}^2[X]}{\mathbb{E}[X]} = (1 - c)^2 \frac{\mathbb{E}^2[X]}{\mathbb{E}^2[X] + \text{Var}[X]} \quad (1.5) \]

**Proof.** First, note that:
\[ \mathbb{E}[X] = \mathbb{E}[X \mathbb{1}\{X \leq c\mathbb{E}[X]\}] + \mathbb{E}[X \mathbb{1}\{X > c\mathbb{E}[X]\}] \leq c\mathbb{E}[X] + \mathbb{E}[X \mathbb{1}\{X > c\mathbb{E}[X]\}], \]
meaning that \( (1 - c)\mathbb{E}[X] \leq \mathbb{E}[X \mathbb{1}\{X > c\mathbb{E}[X]\}] \). Next, note that by Cauchy Swartz:
\[ \mathbb{E}[X \mathbb{1}\{X > c\mathbb{E}[X]\}] \leq \sqrt{\mathbb{E}[X^2]} \sqrt{\mathbb{P}(X > c\mathbb{E}[X])}. \]
Thus:
\[ (1 - c)^2 \mathbb{E}^2[X] \leq \mathbb{E}[X^2] \mathbb{P}(X > c\mathbb{E}[X]), \]
which implies the desired inequality. \( \square \)

To show that the Second Moment Method works, notice that choosing \( c = 0 \) in the Paley Zygmund inequality gives us
\[ \mathbb{P}(X_n > 0) \geq \frac{\mathbb{E}^2[X_n]}{\mathbb{E}^2[X_n] + \text{Var}[X_n]} = \frac{1}{1 + \frac{\text{Var}[X_n]}{\mathbb{E}^2[X_n]}}, \]
so if \( \text{Var}[X_n] = o(\mathbb{E}^2[X_n]) \), then \( \mathbb{P}(X_n > 0) \to 1 \), as desired.

**Applying the Second Moment Method**

We now return to our original goal of showing that \( \mathbb{P}[T_n > 0] \to 1 \), which we shall prove via the Second Moment Method. In particular, we need to show that \( \text{Var}[T_n] = o(\mathbb{E}^2[T_n]) \). To start, notice that:

\[
\text{Var}[T_n] = \text{Var}\left[\sum_{|S| = k} \mathbb{1}\{G_n[S] \text{ is a } k \text{ clique}\}\right] = \sum_{S, S' \mid |S| = |S'| = k} \text{Cov}\left[\mathbb{1}\{G_n[S] \text{ is a } k \text{ clique}\}, \mathbb{1}\{G_n[S'] \text{ is a } k \text{ clique}\}\right] \tag{a} \\
\leq \sum_{|S \cap S'| \geq 2, |S| = |S'| = k} \mathbb{P}\left[\text{ both } G_n[S] \text{ and } G_n[S'] \text{ are } k \text{ cliques}\right],
\]

\[ 10 \]
where (a) follows from the fact that, for any two vertex sets $S$ and $S'$, if $|S \cap S'| \leq 1$ (at most one node shared between $S$ and $S'$), then the set of edges formed among nodes in $S$ are disjoint from the set of edges formed among nodes in $S'$. Thus, by independence, the covariance is zero.

Now, for any given pair of sets $S, S'$, let $\ell = |S \cap S'|$. In order for $S$ and $S'$ to both be $k$-cliques, there are a total of $2\binom{k}{2} - \binom{\ell}{2}$ possible edges that must be formed (think: inclusion-exclusion principle), so we have

$$\text{Var}[T_n] = \sum_{\ell=2}^{k} \left\{ \binom{k}{\ell} \binom{n-k}{k-\ell} \right\} \cdot 2^{-2\binom{\ell}{2} + \binom{\ell}{2}} \quad (1.6)$$

$$= \sum_{\ell=2}^{k} \binom{\ell}{k} \binom{n-k}{k-\ell} \cdot 2^{-2\binom{\ell}{2} + \binom{\ell}{2}} \quad (1.7)$$

where the last step follows from the following reasoning: there are $\binom{n}{k}$ ways of picking a set $S$ of $k$ vertices from a graph on $n$ vertices. And, for each such set $S$, there are exactly $\binom{k}{\ell}$ ways to pick $\ell$ nodes from $S$ that will also be part of another set $S'$. Once $S$ and the nodes of $S$ that will be shared with $S'$ have been determined, it remains to pick from $S'$ the remaining $k - \ell$ nodes of $S'$, and there are exactly $\binom{n-k}{k-\ell}$ ways of doing that.

At this point, one can analyze the above sum by brute force, focusing on the exponent of each term. Next we present a more “statistician’s approach”. Note that the counting step in (1.6) is precisely how hypergeometric distribution (sampling without replacement) arises. Indeed, if we have an urn of $n$ balls among which $k$ balls are red, let $H$ denote the number of red balls if we draw $k$ balls from the urn uniformly at random without replacements. Then $H \sim \text{Hypergeometric}(n, k, k)$. Thus, we can express the same quantity in terms of $H$ as follows:

$$\text{Var}[T_n] = \sum_{\ell=2}^{k} \binom{n-k}{k-\ell} \cdot 2^{\binom{\ell}{2}} \leq \sum_{\ell=2}^{k} \binom{\ell}{2} \binom{n-k}{k-\ell} \cdot 2^{\binom{\ell}{2}}$$

$$= \mathbb{E}[2^{H/2}1 \{H \geq 2\}] \leq \mathbb{E}[2^{H/2}] - \mathbb{P}[H = 0]. \quad (1.8)$$

Next we will show that both $\mathbb{P}[H = 0] \to 1$ and $\mathbb{E}[2^{H/2}] \to 1$. Indeed,

$$\mathbb{P}[H = 0] = \binom{n-k}{k} \frac{1}{\binom{n}{k}} = \left(1 - \frac{k}{n}\right) \left(1 - \frac{k}{n-1}\right) \cdots \left(1 - \frac{k}{n-k+1}\right) \to 1,$$

since $k = (2 - \epsilon) \log_2 n = o(\sqrt{n})$.

To bound the generating function, we use the comparison between sampling with replacements (binomial) and sampling without replacements (hypergeometric). The following result of Hoeffding (proved in the homework) will be useful in several places in this course:

**Lemma 1.2** (Hoeffding’s lemma). Binom$(k, \frac{k}{n})$ dominates Hypergeometric$(n, k, k)$ in the order of convex functions. In other words, if $B \sim \text{Binomial}(k, \frac{k}{n})$, then $\mathbb{E}[f(H)] \leq \mathbb{E}[f(B)]$ for all convex functions $f$.

Using this lemma, we have

$$\mathbb{E}[2^{k/2}] \leq \mathbb{E}[2^{kB/2}] = \left(1 + \frac{k}{n} \left(2^{\frac{k}{2}} - 1\right)\right)^k \leq \exp\left(\frac{k^2}{n} \left(2^{\frac{k}{2}} - 1\right)\right) \to 1,$$

since $k^2 2^{k/2} \ll n$ by the assumption that $k = (2 - \epsilon) \log_2 n$. 

11
To summarize, we have shown that $\frac{\text{Var}[T_n]}{\mathbb{E}^2[T_n]} \to 0$. By Paley-Zygmund (Lemma 1.1), it follows that $\mathbb{P}[T_n > 0] \to 1$, i.e., $\mathbb{P}[\omega(G_n) \geq (2 - \epsilon) \log_2 n] \to 1$, so we’ve proven the desiderata.

**Remark 1.2.** Note that in computing the second moment, (1.8) can be equivalently written as

$$\frac{\text{Var}[T_n]}{\mathbb{E}^2[T_n]} = \mathbb{E}[2^k|S \cap S'|/2 \mathbb{I}(|S \cap S'| \geq 2)],$$

where $S$ and $S'$ are independent random $k$-sets drawn uniformly. This is something we will frequently encounter in computing the second moment, which typically involves *two independent copies* of the same randomness and their overlap $|S \cap S'|$.

**Remark 1.3.** As a small aside, we can further show that not only there exists a clique of size $k = (2 - \epsilon) \log_2 n$, there are an *abundance* of them. Indeed, by (1.4) and using $\binom{n}{k} \geq \left(\frac{n}{k}\right)^k$, we have

$$\mathbb{E}[T_n] = \binom{n}{k} 2^{-\binom{k}{2}} \geq \left(\frac{n}{k}\right)^k 2^{-\binom{k}{2}} = n^{\Omega(\log n)} \to \infty.$$ 

By Lemma 1.1, we have $T_n > o(\mathbb{E}[T_n])$ with probability $1 - o(1)$. This shows that there exists super polynomially many cliques of size $(2 - \epsilon) \log_2 n$. Unfortunately, the best polynomial-time algorithm can only guarantee to find a clique of size $(1 - \epsilon) \log_2 n$ with high probability. We will discuss this next time.
Let $\omega(G(n, \frac{1}{2}))$ denotes the maximum size of cliques in $G(n, \frac{1}{2})$ which is a random variable. Recall that $\omega(G(n, \frac{1}{2}))$ concentrates around $2\log_2 n$ as shown in the previous lecture. In fact, not only there exists a clique of size $(2-\epsilon)\log_2 n$, there exist an abundance of them. The reason is that

$$\mathbb{E}[\# \text{ of cliques of size } k] = \binom{n}{k}2^{(k^2)} \rightarrow \begin{cases} 0 & \text{if } k = (2 + \epsilon)\log_2 n \\ +\infty & \text{if } k = (2 - \epsilon)\log_2 n \end{cases}.$$

Now that the statistical aspect of the problem has been understood, what about the computational aspect? The complexity of the exhaustive search is

$$\left(\frac{n}{\log_2 n}\right)^{\log n} \approx n^\log n$$

and grows superpolynomially in $n$. What if we limit ourselves to “efficient” algorithms that run in time polynomial in the size of the graph, say, $n^C$ for some constant $C$. In the following section, we are going to present a greedy algorithm that runs in polynomial time (in fact, sublinear time) and is able to find cliques of size $$(1-\epsilon)\log_2 n$$ (a factor-of-two approximation of the maximum clique). In contrast, for the **max clique** problem (finding the maximum clique in a given graph or deciding whether a clique of a given size exists) in the worst case is impossible to approximate even within a factor of $n^{1-\epsilon}$, unless P=NP [Hås99]. This shows the drastic difference between worst-case analysis and average-case analysis, due to the atypicality of the hard instances. Nevertheless, for $G(n, \frac{1}{2})$, it remains open whether there exists an efficient algorithm that finds a clique of size bigger than this threshold, say, $1.01\log_2 n$.

### 2.1 Grimmett-McDiarmid’s greedy algorithm to find cliques of size $(1-\epsilon)\log_2 n$

Before we present a greedy algorithm that provably works, let us start with another greedy algorithm which is intuitive but might be difficult to analyze.

**Algorithm 1**: Greedy algorithm I

- Start from an arbitrary vertex
- Given a clique of size, repeat:
  - Add a vertex randomly from the common neighbors of the existing clique
  - If there is no common neighbors, stop and return the clique

The justification to this algorithm is the following: given that we have found an $m$-clique, for a given vertex $v$ outside, the probability that $v$ is connected to all $m$ vertices in the clique is, assuming
that each edge happens with probability $\frac{1}{2}$ independently,

$$P(v \text{ is connected to all } m \text{ vertices}) = 2^{-m}.$$ 

Therefore, the probability that there exists a $v$ that is connected to all $m$ vertices in the existing clique is

$$P(\exists v \text{ connected to all } m) = 1 - (1 - 2^{-m})^n \to 1 \text{ if } 2^{-m} \ll 1/n, \text{ e.g., } m = (1 - \epsilon) \log_2 n.$$ 

However, the reasoning is flawed because given the information that we have an existing clique of size $m$, the probabilities of $v$ connected to each of them is no longer $\frac{1}{2}$. Furthermore, the edges are not conditionally independent either. Therefore it is unclear how to analyze Algorithm 1.

Next we present a variant of the greedy algorithm, due to Grimmett-McDiarmid [GM75],\(^1\) which is easily analyzable.

**Algorithm 2:** Greedy algorithm II

Label the vertices $v_1, \ldots, v_n$ arbitrarily

for $t = 1$ to $n$ do

Given the current clique, if $v_t$ is connected to all vertices in the current clique, add $v_t$ to the clique.

end for

**Theorem 2.1.** Fix any $\epsilon > 0$. With probability tending to one as $n \to \infty$, the output of Algorithm 2 is a clique of size at least $(1 - \epsilon) \log_2 n$.

**Proof.** It is obvious that the output of the above algorithm, denoted as $S_n$, is a clique. It remains to show that with high probability, the size of the clique is at least $(1 - \epsilon) \log_2 n$.

Let $T_i$ be the time for the size of the clique to grow from $i - 1$ to $i$. By the design of the algorithm we can see that $T_i$’s are independent and geometrically distributed as

$$T_i \overset{\text{ind}}{\sim} \text{Geo}(2^{-i}) \Rightarrow \mathbb{E}T_i = 2^i.$$ 

Therefore,

$$P(|S_n| \geq k) = P(T_1 + T_2 + \cdots + T_k \leq n)$$

$$\geq \prod_{i=1}^{k} P\left(T_i \leq \frac{n}{k}\right)$$

$$= \prod_{i=1}^{k} \left(1 - (1 - 2^{-i})^{n/k}\right)$$

$$\geq \left(1 - (1 - 2^{-k})^{n/k}\right)^k$$

$$\geq 1 - k(1 - 2^{-k})^{n/k}$$

$$\to 1 \text{ if } k = (1 - \epsilon) \log_2 n.$$ 

\(^1\)The original paper [GM75] deals with the number of vertex coloring (so that no adjacent vertices are colored the same) and independent set (subset of vertices that induce an empty graph). Note that cliques in the original graphs correspond to independent sets in the complementary graph, and the complement of $G(n, \frac{1}{2})$ is still $G(n, \frac{1}{2})$.

\(^2\)This is reminiscent of the coupon collector problem, where it becomes increasingly more difficult to collect the last few uncollected coupons, although here the situation is more drastic.
Remark 2.1. The time complexity of Algorithm 2 is, in expectation,
\[ \sum_{i=1}^{\log_2 n} 2^i \times i = O(n \log_2^2 n). \]
This is sublinear in the size of the graph (Θ(n^2) edges).

2.2 Planted Clique Model

2.2.1 Overview

The planted clique model is a random graph model which can be described as follows: First choose a subset \( K \) of size \( k \) uniformly at random from all \( n \) vertices and form a clique. The remaining vertex pairs are connected independently with probability \( \frac{1}{2} \). In other words,
\[ \forall i, j \quad \mathbb{P}[i \sim j] = \begin{cases} 1 & \text{if both } i, j \in K \\ \frac{1}{2} & \text{otherwise} \end{cases}. \]
Denote the resulting graph \( G \sim G(n, \frac{1}{2}, k) \). Note that \( G(n, \frac{1}{2}, 0) \) is the usual Erdős-Rényi graph \( G(n, \frac{1}{2}) \).

Ignoring the computation cost, we can use exhaustive search to recover the planted clique with high probability if \( k \geq (2 + \epsilon) \log_2 n \), because the maximum clique in \( G(n, \frac{1}{2}) \) is approximately \( 2 \log_2 n \), as shown in Section 1.2. For the same reason, if \( k \leq (2 - \epsilon) \log_2 n \), recovering the planted clique is information‐theoretically impossible, as there is an abundance of cliques of such size in \( G(n, \frac{1}{2}) \) (cf. Remark 1.3). However, what if we only consider efficient algorithms? It turns out the state of the art can only find planted cliques of size \( k = \Omega(\sqrt{n}) \).

In the following sections, we will discuss a number of efficient algorithms that is able to recover the planted clique with high probability if \( k = \Theta(\sqrt{n}) \).

- **Degree test** method, which works for \( k = \Omega(\sqrt{n \log n}) \). We will discuss an iterative version that works for \( k = C \sqrt{n} \) for some \( C > 0 \), that runs in linear time.\(^3\)
- **Spectral method**, which works for \( k = C \sqrt{n} \) for some \( C \). This can be improved to arbitrarily small \( C \) at the price of time complexity \( n^{O(1/C^2)} \).
- **Semi-definite programming** approach, which also work for \( k = C \sqrt{n} \). The added advantage is robustness, which the previous methods lack.

Remark 2.2. We see that the gap between the information-theoretical limit and what computationally efficient algorithms can achieve is significantly larger for the planted clique problem (\( \log_2 n \) versus \( \sqrt{n} \)) than the counterpart for the maximum clique problem in \( G(n, \frac{1}{2}) \) (\( 2 \log_2 n \) versus \( \log_2 n \)).

Exercise (Slightly smarter exhaustive search). To find the hidden \( k \)-clique in \( G(n, \frac{1}{2}, k) \), exhaustive search takes \( \binom{n}{k} \sim n^k \) time. Here is an \( n^{\Theta(\log n)} \)-time algorithm for all \( k = \Omega(\log n) \).

1. By exhaustive search we can find a clique \( T \) of size \( C \log_2 n \) for some large constant \( C \). Then \( |T \cap K| \geq (C - 2 - \epsilon) \log_2 n \). (Why?)\(^3\)

\(^3\)Since the graph is dense, here linear time means \( O(n^2) \).
2. Let $S$ denote the set of all vertices that has at least $C/2 \log n$ neighbors in $T$. Then one can show that $S \subseteq K$ with high probability.

3. Now $S$ might also contain some non-clique vertices, which requires some cleanup. So we report the $k$ highest-degree vertices in the induced subgraph $G[S]$.

### 2.2.2 Degree Test

By degree test we meant declaring the $k$ vertices with the largest degrees as the estimated clique. The motivation is that the vertices in the clique tend to have a higher degree than those outside. To analyze this method, let $d_i$ denote the degree of a vertex $i$. If $i \notin K$, then

$$d_i \sim \text{Binom} \left( n - 1, \frac{1}{2} \right), \quad \mathbb{E}d_i \approx \frac{n}{2}.$$  

If $i \in K$, then

$$d_i \sim k - 1 + \text{Binom} \left( n - k, \frac{1}{2} \right), \quad \mathbb{E}d_i \approx \frac{n+k}{2}.$$  

Therefore, the separation in mean is proportional to $k$. Usually, we need the separation to be at least as large as the standard deviation. Therefore, $k = \Omega(\sqrt{n})$ is clearly necessary. Furthermore, for the degree test method to work, we need an additional $\sqrt{\log n}$ factor to accommodate for the possibility that some of the $n - k$ non-clique vertices will have an atypically high degree, and some of the $k$ vertices in the clique will have an atypically low degree. This will be carried out by an union bound, as we shall see later. In fact, although the degrees are not mutually independent ($d_i$ and $d_j$ are positively correlated through $1 \{i \sim j\}$), they are almost independent so $\sqrt{\log n}$ factor is necessary in order for the degrees to fully separate.

Formally, the degree-based estimator is just

$$\hat{K} = \text{set of } k \text{ vertices with the highest degree}. \quad (2.1)$$

We will show that with high probability, the degree based estimator can recover the true planted clique of size $k = \Omega(\sqrt{n \log n})$. This simple observation is usually attributed to [Kuc95].

**Theorem 2.2.** $\mathbb{P}(\hat{K} = K) \to 1$, provided $k = C \sqrt{n \log n}$ for some absolute constant $C$.  

It is obvious from the definition of the degree-based estimator (2.1) that a sufficient condition for $\hat{K} = K$ is that

$$\min_{i \in K} d_i > \max_{i \notin K} d_i. \quad (2.2)$$

Before we prove Theorem 2.2, we first review some basic facts about Gaussian approximation and concentration inequalities.

**Lemma 2.1.** Suppose that $X_1, \ldots, X_n \sim N(0,1)$, then

$$X_{\max} = \max_{i \in 1, \ldots, n} X_i \leq \sqrt{(2 + \epsilon) \log n} \quad \text{w.h.p.}$$

**Proof.**

$$\mathbb{P}(X_{\max} > t) \leq n \mathbb{P}(X_1 > t) \leq ne^{-t^2/2} \to 0 \text{ if } t > \sqrt{(2 + \epsilon) \log n}. \quad \square$$
Note that Lemma 2.1 is tight if $X_1,\ldots,X_n$ are independent.

Using the heuristic of approximating binomials by Gaussians, we can analyze the degree test as follows: If $i \notin K$, then
\[
d_i \sim \text{Binom} \left( n - 1, \frac{1}{2} \right) \approx N(n/2, n/4)\]
Using Lemma 2.1, the right hand side of Equation (2.2) is approximately
\[
\max_{i \notin K} d_i \leq \sqrt{2 \log(n - k)} \frac{n}{4} + \frac{n}{2} \approx \frac{n}{2} + \frac{1}{2} \sqrt{2n \log n}.
\]
Similarly, since for $i \in K$ we have $d_i \sim k + \text{Binom}(n - k, \frac{1}{2})$,
\[
\min_{i \in K} d_i \geq k + \frac{n - k}{2} + \sqrt{2 \log k} \frac{n - k}{2} \approx \frac{n + k}{2} + \frac{1}{2} \sqrt{2n \log k}.
\]
Therefore (2.2) holds with high probability if $k \geq \sqrt{Cn \log n}$ for some large constant $C$.

To justify the above Gaussian intuition, we use Hoeffding’s inequality, which is a very useful concentration inequality. We will prove it in Lecture 4.

**Lemma 2.2** (Hoeffding’s inequality). Let $S = X_1 + \cdots + X_n$ where $X_1,\ldots,X_n$ are independent and $a \leq X_i \leq b, \forall i$. Then
\[
\Pr(|S - ES| \geq t) \leq 2 \exp \left( -\frac{2t^2}{n(b - a)^2} \right).
\]
If we apply the Hoeffding inequality to binomial distribution, we get the following corollary:

**Corollary 2.1.** If $S \sim \text{Binom}(n,p)$,
\[
\Pr(|S - np| \geq t) \leq 2 \exp \left( -\frac{2t^2}{n} \right).
\]

Two remarks are in order.

**Remark 2.3.** In the large-deviation regime where $t = \alpha n$, we have
\[
\Pr(|S - np| \geq t) = \exp \left( -\Theta(n) \right).
\]
In the moderate-deviation regime where $O(\sqrt{n}) \leq t = o(n)$, if $t = n^{1/2+\epsilon}$, we have
\[
\Pr(|S - np| \geq t) = \exp \left( -\Theta(n^{2\epsilon}) \right).
\]
When $t = o(\sqrt{n})$, the inequality becomes meaningless.

**Remark 2.4.** The bound is good if $p$ is a constant like $\frac{1}{2}$. If $p = o(1)$, i.e., for sparse graphs, then the variance is $np \ll n$ and we should aim for tail bound like $\exp(-\frac{2t^2}{np})$ instead if $p \gg \frac{1}{n}$. However, Hoeffding’s inequality does not capture this.

Using the Hoeffling’s inequality, we obtain the following lemma about the degree test:
Lemma 2.3. Suppose $G \sim G(n, \frac{1}{2}, k)$ with hidden $K$. Let $\hat{K}$ be the set of the $k$ highest degree vertices in $G$. Then,
\[
P[\hat{K} = K] \geq 1 - cn \exp\left(-\frac{k^2}{8n}\right).
\]
Therefore, if $k = \Omega(\sqrt{n \log n})$, then $\hat{K} = K$ with high probability.

For the next algorithm, we need the following lemma that quantifies the convergence rate in the central limit theorem, in terms of the Kolmogorov distance, that is, the sup norm between CDFs; this result is due to Berry-Esseen. We state without proof the version for the normal approximation for binomial distributions.

Lemma 2.4 (Berry-Esseen theorem). There exists $C = C(p)$, such that
\[
\sup_{x \in \mathbb{R}} |\Pr(\text{Binom}(n, p) \leq x) - \Pr(N(np, np(1-p)) \leq x)| \leq \frac{C(p)}{\sqrt{n}}.
\]
The constant $C(p)$ can be bounded by $\frac{C}{\sqrt{p}}$ for some absolute constant $C$.

2.2.3 Iterating the degree test

The degree test works for $k = \Omega(\sqrt{n \log n})$. Now we will present an iterative algorithm by Dekel–Gurel-Gurevich–Peres [DGGP11] that is able to find $K$ in $O(n^2)$ times where $k = |K| = C \sqrt{n}$ for some $C > 0$.

There are essentially two ideas:

1. The first idea is this: if we define the “relative size” as $\frac{k^2}{n}$, then we know from Lemma 2.3 that it needs to exceed $\log n$ in order for the degree test to succeed. If we can subsample the graph cleverly, then we might be able to gradually increase the relative size until it reaches this threshold. However, blindly subsample each vertex independently with probability $\tau$ clearly does not work, as $n \to n\tau$ and $k \to k\tau$ and this will only decrease the relative size. Instead, we are going to subsample in such a way that prefers clique vertices (e.g., based on degrees!), so that $n \to \tau n$ and $k \to \rho k$ where $\rho > \sqrt{\tau}$. This way, the relative size will grow by a constant factor in each iteration, and it takes $\log \log n$ rounds to pass the log $n$ threshold.

Specifically, we will generate a sequence of graphs $G = G_0 \supset G_1 \supset \cdots \supset G_T$, so that each $G_t$ is an instance of the planted clique model $G(n_t, \frac{1}{2}, k_t)$, with
\[
n_t \approx n'_t \triangleq \tau^t n, \quad k_t \approx k'_t \triangleq \rho^t k \quad \text{(2.3)}
\]
where
\[
\tau = (1 - \alpha)Q(\beta), \quad \rho = (1 - \alpha)Q(\beta - C \sqrt{\alpha}),
\]
and $Q(t) \triangleq \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ denotes the complementary CDF of the standard normal distribution.

2. The catch with the subsampling idea is that we ended up with recovering just a subset of the hidden clique. Nevertheless, provided this seed set is not too small, it is not hard to blow it up to the entire clique. Indeed, suppose that we have successfully found a seed set $S \subseteq K$, we can recover the planted clique by first taking the union of the seed set and its common neighbors and then finding the highest $k$ degree vertices therein.
Algorithm 3: Iterative degree testing algorithm [DGGP11]

Step 1: Given parameters $\alpha, \beta, T$, we will generate $G = G_0 \supseteq G_1 \supseteq \cdots \supseteq G_T$ as follows

for $t = 0$ to $T - 1$ do

  Given the current $G_t = (V_t, E_t)$, define $V_{t+1}$ as follows and denote the induced subgraph $G_{t+1} = G[V_{t+1}]

  Pick a subset $S_t \subseteq V_t$ by keeping each vertex with probability $\alpha$

  Let $V_{t+1} \triangleq \{v \in V_t \setminus S_t: \; d_{S_t}(v) \geq \frac{1}{2} |S_t| + \frac{\beta}{2} \sqrt{|S_t|}\}$

  $t \leftarrow t + 1$

end for

Step 2: Let $\tilde{K} = \{\frac{k_T}{2}\}$ highest-degree vertices on $G_T$, where $k_T'$ is defined in (2.3).

Step 3: Let $K'$ be the union of $\tilde{K}$ and its common neighbors. Report $\tilde{K}$, the $k$ highest-degree vertices in $G' = G[K']$.

Theorem 2.3. If $k = C\sqrt{n}$ for some $C > 0$, then $\tilde{K} = K$ with high probability.

Remark 2.5. Before proving the theorem, let me explain two steps in Algorithm 3:

- In Step 1, the vertex $V_t$ is chosen based on its degree, since vertices in the clique tend to have a higher degree. Thus, intuitively, we could have chosen $V_t$ as the vertices in $G_{t-1}$ whose degree exceeds a given threshold. However, in this way we created a lot of dependency and we cannot ensure each $G_t$ is still an instance of the planted clique model (because we want to apply the degree test and invoke Lemma 2.3). Instead, what we did is to choose a “test set” $S_t$ and compute the degree by withholding this test set. This is a commonly used trick for type of problems, which we will revisit later in stochastic block models.

- In Step 2, since $G_T \sim G(nT, \frac{1}{2}, k_T)$, according to the degree test we should choose $\tilde{K}$ as the $k_T$ highest-degree vertices. However, since $k_T$ is not observed, we use $\frac{k_T'}{2}$ as a conservative proxy, which is a high-probability lower bound for $k_T$.

To prove the theorem, we will prove a series of claims, upon which the proof of the theorem becomes straightforward. In order to focus on the main ideas, we will be sloppy with notations like “$\approx$” and “whp”.

Claim 2.1. Define $n_t = |V_t|$ and $k_t = |K \cap V_t|$. For all $t$, $G_t$ is an instance of $G(n_t, \frac{1}{2}, k_t)$. In other words, conditioned on $(n_t, k_t)$, $G_t \sim G(n_t, \frac{1}{2}, k_t)$.

Proof. This claim is true because the vertex set $V_t$ is chosen without exposing any edges in $E_t$. Indeed, by induction, it suffices to consider $t = 0$, i.e., $G_0 \to G_1$. Note that each vertex $v \in S_0$ is included in $V_1$ if its degree in the test set $S_0$, namely, $d_{S_0}(v)$, exceeds a given threshold. Therefore $G_t = (V_1, E_t)$ is distributed as $G(n_1, \frac{1}{2}, k_1)$, where $n_1 = |V_1|$ and $k_1 = |V_1 \cap K|$. This can also been seen from the perspective of the adjacency matrix: $V_1$ is determined based on the submatrix $A_{S_0,S_0}$ and hence independent of $A_{S_0,S_0}$ (see Fig. 2.1).

Claim 2.2. Let $G_t = (V_t, E_t)$, then

$$n_t \approx n_t' \triangleq \tau^n t n$$

and

$$k_t \approx k_t' \triangleq \rho^k k$$

whp.,

where $\tau$ and $\rho$ are defined in (2.4).
Figure 2.1: The vertex set $V_1$ is chosen by withholding the test set $S_0$.

Proof. Again, by induction, it suffices to consider $t = 0$, i.e., $G_0 \rightarrow G_1$. We will use B-E inequality (Lemma 2.4) to approximate binomial with Gaussian distribution. Since $S_1 \sim \text{Binom}(n, \alpha) \approx \alpha n$, it follows that $\forall v \in S_0^c$,

$$d_{S_0}(v) \sim \left\{ \begin{array}{ll}
\text{Binom}(|S_0|, \frac{1}{2}) & \text{if } v \notin K \\
|S_0 \cap K| + \text{Binom}(|S_0 \setminus K|, \frac{1}{2}) & \text{if } v \in K
\end{array} \right.$$ 

where $|S_0 \cap K| \sim \text{Binom}(k, \alpha) \approx k \alpha = C \sqrt{n} \alpha \text{ w.h.p}$. Thus,

$$P(d_{S_0}(v) \geq \frac{1}{2}|S_0| + \frac{\beta}{2} \sqrt{|S_0|}) \approx \begin{cases} Q(\beta) & \text{if } v \notin K \\
Q(\beta - C \sqrt{\alpha}) & \text{if } v \in K
\end{cases}.$$ 

Finally, in summary,

$$|V_1| = \sum_{v \in V_0} 1 \{v \in S_0^c\} 1 \{d_{S_0}(v) \geq \frac{1}{2}|S_0| + \frac{\beta}{2} \sqrt{|S_0|}\} \approx |V_0| (1 - \alpha)Q(\beta).$$

Similarly,

$$k_1 = |K \cap V_1| = \sum_{v \in K} 1 \{v \in S_0^c\} 1 \{d_{S_0}(v) \geq \frac{1}{2}|S_0| + \frac{\beta}{2} \sqrt{|S_0|}\} \approx k (1 - \alpha)Q(\beta - C \sqrt{\alpha}).$$

Claim 2.3. Let $\tilde{K}$ be the set of the $\frac{k_T'}{2}$ highest-degree vertices in $G_T$. Choose $T = C_0 \log \log n$ (so that w.h.p $\frac{k_T^2}{n_T} \geq \log^2 n$ say), and $n_T \approx n_T' = \rho^T n \geq \frac{n}{\text{polylog}(n)}$ and $k_T \approx k_T' = \tau^T k \geq \frac{k}{\text{polylog}(n)}$. Then $\tilde{K} \subset K$ with high probability.

Proof. By Lemma 2.3, with probability $\geq 1 - n_T e^{-\frac{k_T^2}{8n_T}} \geq 1 - e^{-\text{polylog}(n)}$, the hidden clique in $G_T \sim G(n_T, \frac{1}{2}, k_T)$ have the highest degrees. On the high probability event that $k_T \geq \frac{k_T'}{2}$, we have $\tilde{K} \subset K$. \hfill \square

Now that we have shown $\tilde{K}$ is a subset of the true clique, we still need to expand it to the entire clique. Think of $\tilde{K}$ as a “seed set” and the main point is in this case $s = |\tilde{K}| \geq (1 + \epsilon) \log n$ seeds suffice. However, the caveat is that $\tilde{K}$ obtained from step 1 and 2, and might be dependent on every edge. Fortunately, at this point, this can be addressed by taking a union bound over all $s$-subsets of $K$.
Claim 2.4 (Clean-up). With high probability, the following holds: Let $\tilde{K}$ is an $s$-subset of $K$ (which can be adversarially chosen). Let $K'$ be the union of $\tilde{K}$ and its common neighbors. Let $\hat{K}$ be the $k$ highest-degree vertices on $G' = G[K']$. If $s \geq (1 + \epsilon) \log n$ for any constant $\epsilon$, then $\hat{K} = K$ with high probability.

Proof. Let

$$K' = \tilde{K} \cup \text{common neighbors} = K \cup F,$$

where $F$ denotes the non-clique common neighbors. We first show that $F$ is small: taking a union bound over $\tilde{K}$,

$$\Pr[|F| \geq \ell] \leq \left(\frac{k}{s}\right) \sum_{\ell'=\ell}^{n-k} \binom{n}{\ell} 2^{-s\ell'}$$

$$\leq n \exp(s \log k + \ell(\log n - s))n = \exp((1 + \epsilon) \log n \log k - \epsilon \ell \log n) \to 0$$

if $\ell = \frac{2}{\epsilon} \log k$.

Now in $G'$, for any $v \in K$, we have by definition $d(v) \geq k - 1$; for any $v \notin K$, we have

$$d(v) \leq |F| + d_K(v) \leq |F| + \max_{v \notin L} d_K(v) < k - 1.$$}

This shows that the $k$ highest-degree vertices in $G'$ are precisely the true clique. \qed
§ 3. Spectral Method - basic perturbation theory

Our goal is to use the spectral method to pursue inference. As we will see, in planted clique and many related planted problems, the first few eigenvectors of the population matrix $EX$ contain the information about the planted structures that we are interested in. Since we only have observations $X$ at hand and do not know $EX$, we compute the first few eigenvectors of $X$ instead. Writing $X = EX + (X - EX)$, we expect that the error of estimating the first few eigenvectors of $EX$ can be bounded by the size of the perturbation $X - EX$.

3.1 Review of linear algebra

3.1.1 Eigendecomposition

Suppose that $X$ is a symmetric real valued matrix in $\mathbb{R}^{n \times n}$.

**Definition 3.1.** The pair $(\lambda, v)$ with $\lambda \in \mathbb{R}$ and $v \in \mathbb{R}^n$ is an eigenpair of $X$, consisting of an eigenvalue $\lambda$ and an eigenvector $v$, if

$$Xv = \lambda v.$$ 

We order the eigenvalues of $X$ by their sizes such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. The corresponding eigenvectors $[v_1, \ldots, v_n]$ form an orthonormal basis (ONB) of $\mathbb{R}^n$. Denote $V = [v_1, \ldots, v_n]$, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$. We can write the eigendecomposition of $X$ as

$$X = VA\Lambda V^\top = \sum_{i=1}^n \lambda_i v_i v_i^\top.$$ 

Also note that $\text{rank}(X) = r \iff$ there exist exactly $r$ nonzero $\lambda_i$’s.

3.1.2 Singular value decomposition (SVD)

Now suppose that $X \in \mathbb{R}^{m \times n}$ is a real valued rectangular matrix. The singular value decomposition (SVD) of $X$ is

$$X = U \Sigma V^\top = \sum_{i=1}^r \sigma_i U_i V_i^\top$$

where $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r \times r}$, $\sigma_i \geq 0$, $U = [U_1, \ldots, U_r] \in \mathbb{R}^{m \times r}$ and $V = [V_1, \ldots, V_r] \in \mathbb{R}^{n \times r}$. The columns of $U$ are orthonormal and we call them left singular vectors and likewise the columns of $V$ are orthonormal too and we call them right singular vectors.

We can calculate $\Sigma, U$ and $V$ by taking eigendecompositions of $XX^\top$ and $X^\top X$. Indeed,

$$XX^\top = U \Sigma^2 U^\top \in \mathbb{R}^{m \times m} \quad \text{and} \quad X^\top X = V \Sigma^2 V^\top \in \mathbb{R}^{n \times n},$$

and

$$\sigma_i = \sqrt{\lambda_i(XX^\top)} = \sqrt{\lambda_i(X^\top X)}. $$
3.1.3 Matrix norms

Suppose again that $X \in \mathbb{R}^{m \times n}$. There are multiple ways to define a norm on $X$.

- We view $X$ as a $mn$-dimensional vector with euclidean norm and define the Frobenius norm
  \[ \|X\|_F = \|\text{vec}(X)\|_2 = \sqrt{\sum_{i,j} X_{ij}^2}. \]

- We view $X$ as a linear operator from $(\mathbb{R}^n, \|\cdot\|_p) \to (\mathbb{R}^m, \|\cdot\|_q)$ with operator norm
  \[ \|X\|_{p \to q} = \sup_{\|v\|_p = 1} \|Av\|_q. \]

For this course the most relevant matrix is the case of $p = q = 2$, where we equip $\mathbb{R}^n$ with the euclidean inner product. We denote \[ \|X\|_{2 \to 2} =: \|X\|_{op}, \]
also known as the spectral norm.

We now prove that
\[ \|X\|_{op} = \sigma_{\max}(X) \]
Using the SVD of $X$:
\[ \|X\|_{op}^2 = \sup_{\|v\|_2 = 1} \|Xv\|_2^2 = \sup_{\|v\|_2 = 1} \left\| \sum_i \sigma_i U_i V_i^T v \right\|_2^2 = \sup_{\|v\|_2 = 1} \sum_i \sigma_i^2 (V_i, v)^2 = \sigma_{\max}(X)^2. \]

**Remark 3.1.**

- $\|\cdot\|_{op}$ is a norm and $\|X\|_{op} = \|X^T\|_{op}$.
- If $X = x$ is a vector then $\|X\|_{op} = \|x\|_2$.
- If $\|\cdot\|_{op}$ is orthogonal invariant, i.e. for any $R \in O(n)$, $R' \in O(m)$ we have $\|R'XR\|_{op} = \|X\|_{op}$.
- If $X = [X_1, \ldots, X_n]$ has orthonormal rows (columns), then $\|X\|_{op} = 1$.

**Remark 3.2.** Recall the matrix inner product: $\langle X, Y \rangle = \text{trace}(Y^T X) = \sum_{i,j} X_{ij} Y_{ij}$. Using this we can write
\[ \|X\|_{op} = \sigma_{\max}(X) = \sup_{\|A\|_{p=1}, \text{rank}(A)=1} \|X, A\| = \sup_{\|u\|_2 = 1} \langle X, uu^T \rangle. \]

Likewise, if $X$ is real and symmetric we have that
\[ \lambda_{\max}(X) = \sup_{\|v\|_2 = 1} \langle X, vv^T \rangle, \quad \|X\|_{op} = \sigma_{\max}(X) = \sup_{\|v\|_2 = 1} |\langle X, vv^T \rangle|. \]

Similar relations hold for $\sigma_{\min}$ and $\lambda_{\min}$ if one substitutes the sup’s above for inf’s.

3.2 Pertubation of eigenstructures

In this section we assume that we are given two matrices, $X$ and $Y = X + Z$ where $Z$ is a ‘pertubation’ of $X$. We are interested if eigenvectors and eigenvalues of $X$ and $Y$ are close when $Z$ is ‘small’. Unfortunately, in general this is not the case.
3.2.1 Negative results

Eigenvalues  The eigenvalues \( \lambda_i \) are the roots of the polynomial \( \det(\lambda I - X) = 0 \), which is a polynomial in \( \lambda \) of degree \( n \). Although the roots are continuous in the coefficients of the polynomial, in general the modulus of continuity is not Lipschitz and only \( \frac{1}{\text{degree}} \)-Hölder, and this is tight. Indeed, consider the two matrices

\[
X = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad X_\varepsilon = \begin{bmatrix} 0 & 1 \\ \varepsilon & 0 \end{bmatrix}
\]

Then \( \lambda_1(X) = \lambda_2(X) = 0 \), but \( \lambda_1(X_\varepsilon) = \sqrt{\varepsilon} \) and \( \lambda_2(X_\varepsilon) = -\sqrt{\varepsilon} \). More generally consider

\[
X = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & \ldots & 1 \\ 0 & \ldots & \ldots & \ldots & 0 \end{bmatrix} \quad \text{and} \quad X_\varepsilon = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & \ldots & 1 \\ \varepsilon & \ldots & \ldots & \ldots & 0 \end{bmatrix}
\]

One can show that \( \lambda_i(X) = 0 \) but that \( \lambda_i(X_\varepsilon) = \varepsilon^{1/n} \).

Therefore we need more assumptions on \( X \) to be able to obtain Lipschitz bounds, e.g. that \( X \) is a real and symmetric matrix.

Eigenvectors  But even in the symmetric case eigenvector perturbations may fail dramatically. For \( \varepsilon > 0 \) consider

\[
X = \begin{bmatrix} 1 + \varepsilon & 0 \\ 0 & 1 + \varepsilon \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{bmatrix}.
\]

The eigenvalues of these two matrices are the same

\[
\lambda_1(X) = \lambda_1(Y) = 1 + \varepsilon, \quad \lambda_2(X) = \lambda_2(Y) = 1 - \varepsilon.
\]

However, the eigenvectors are far apart:

\[
v_1(X) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad v_2(X) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \text{but} \quad v_1(Y) = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad v_2(Y) = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}
\]

The lesson from this is that we need separation between the eigenvalues, a spectral (eigen) gap.

3.2.2 Perturbation bound for eigenvalues

Let \( X, Y, Z \) be real symmetric matrices in \( \mathbb{R}^{n \times n} \) and suppose \( Y = X + Z \). We have that

\[
\lambda_1(X) + \lambda_n(Z) = \lambda_1(X) + \inf_{\|v\|_2=1} \langle Z, vv^\top \rangle \\
\leq \sup_{\|v\|_2=1} \langle X + Z, vv^\top \rangle = \lambda_1(Y) \\
\leq \lambda_1(X) + \sup_{\|v\|_2=1} \langle Z, vv^\top \rangle = \lambda_1(X) + \lambda_1(Z)
\]

and therefore

\[
|\lambda_1(X) - \lambda_1(Y)| \leq \max(|\lambda_1(Z)|, |\lambda_n(Z)|) = \|Z\|_{\text{op}}.
\]

More generally we have the following theorem (homework):

**Theorem 3.1** (Weyl’s inequality / Lidski’s inequality).

\[
|\lambda_i(X) - \lambda_i(Y)| \leq \|Z\|_{\text{op}}.
\]
3.2.3 Perturbation bounds for eigenspaces

Let $X, Y, Z$ again be real symmetric matrices in $\mathbb{R}^{n \times n}$ and suppose $Y = X + Z$. Suppose that $X = \sum \lambda_i u_i u_i^\top$ and $Y = \sum \rho_i v_i v_i^\top$. We want to prove a perturbation bound for $u \triangleq u_1$ and $v \triangleq v_1$ and more generally for $U = [u_1, \ldots, u_r]$ and $V = [v_1, \ldots, v_r]$. However, considering $\|u - v\|_2$ makes no sense as $u$ and $v$ are only determined up to their sign, and similarly $U$ and $V$ are only defined up to orthogonal transformation. There are two possible workarounds:

- Consider the distance
  \[
  \min_{s \in \{\pm 1\}} \|u + sv\|_2 = \sqrt{2 - 2|\langle u, v \rangle|} = \sqrt{2 - 2 \cos \theta} = \sqrt{2 \sin^2 \frac{\theta}{2}},
  \]

  and, more generally, $\inf_{R \in O(r)} \|U - VR\|$. 

- Consider the distance between the linear subspaces spanned by $u$ and $v$, defined through their respective projection matrices:
  \[
  \left\| uu^\top - vv^\top \right\|_F^2 = 2(1 - \langle u, v \rangle^2) = 2 \sin^2(\theta),
  \]

  and in the general case $\|UU^\top - VV^\top\|_F$ or $\|UU^\top - VV^\top\|_{op}$.

**Theorem 3.2** (Davis-Kahan). Let $\cos \theta = |\langle u_1, v_1 \rangle|$. Then

\[
\sin \theta \leq \frac{\|Z\|_{op}}{\max(\rho_1 - \lambda_2, \lambda_1 - \rho_2)}.
\]

**Proof.** Assume that $\rho_1 \geq \lambda_2$. Let us start from the eigenvalue equations:

\[
Xu = \lambda_1 u \quad \text{and} \quad Yv = \rho_1 v.
\]

Denote $U_\perp = [u_2, \ldots, u_n] \in \mathbb{R}^{n \times n-1}$. Then

\[
U_\perp^\top X = \begin{bmatrix}
  u_2^\top \\
  \vdots \\
  u_n^\top
\end{bmatrix} X = \begin{bmatrix}
  \lambda_2 u_2^\top \\
  \vdots \\
  \lambda_n u_n^\top
\end{bmatrix} = \begin{bmatrix}
  \lambda_2 \\
  \vdots \\
  \lambda_n
\end{bmatrix} \begin{bmatrix}
  u_2^\top \\
  \vdots \\
  u_n^\top
\end{bmatrix}.
\]

Hence

\[
U_\perp^\top (X + Z)v = \rho_1 U_\perp^\top v \iff \begin{bmatrix}
  \lambda_2 \\
  \vdots \\
  \lambda_n
\end{bmatrix} U_\perp^\top v + U_\perp^\top ZV = \rho_1 U_\perp^\top v
\]

\[
\iff \begin{bmatrix}
  \rho_1 - \lambda_2 \\
  \vdots \\
  \rho_1 - \lambda_n
\end{bmatrix} U_\perp^\top v = U_\perp^\top ZV
\]

\[
\iff U_\perp^\top v = \begin{bmatrix}
  \frac{1}{\rho_1 - \lambda_2} \\
  \vdots \\
  \frac{1}{\rho_1 - \lambda_n}
\end{bmatrix} U_\perp^\top ZV.
\]

25
Taking the $\| \cdot \|_2$-norm on both sides gives
\[
\|U^\top v\|_2 \leq \left\| \begin{bmatrix} \frac{1}{\rho_1 - \lambda_2} & \cdots & \frac{1}{\rho_1 - \lambda_n} \\ \vdots \\ \frac{1}{\rho_1 - \lambda_n} \end{bmatrix} \right\|_{op} \|U^\top\|_{op} \|Z\|_{op} = \frac{\|Z\|_{op}}{|\rho_1 - \lambda_2|}.
\]
Finally, note that
\[
\|U^\top v\|^2 = v^\top U^\top_{\perp} v = v^\top (I - uu^\top) v = 1 - \langle u, v \rangle^2 = \sin^2(\theta).
\]
If $\rho_1 < \lambda_2$, then $\lambda_1 > \rho_2$. Exchanging the roles of $X$ and $Y$ we obtain the other statement. 

More generally, considering the first $r$ eigenvectors we have for $U = [U_1, \ldots, U_r]$ and $V = [V_1, \ldots, V_r]$ that for any unitarily invariant norm $\| \cdot \|$
\[
\|U^\top V\| \leq \frac{\|Z\|}{\max(|\rho_r - \lambda_{r+1}|, |\lambda_r - \rho_{r+1}|)}.
\]
One can generalize this to singular vectors by a technique sometimes called self-adjoint dilation:\footnote{Thanks for Cheng Mao for pointing this out.}
For $X = USV^\top \in \mathbb{R}^{m \times n}$, $Y = \tilde{U}\tilde{S}\tilde{V}^\top$ consider the matrix
\[
\begin{bmatrix} 0 & X \\ X^\top & 0 \end{bmatrix} \in \mathbb{R}^{(m+n) \times (m+n)},
\]
and likewise for $Y$. Observe that
\[
\begin{bmatrix} 0 & X \\ X^\top & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} = \sigma_1 \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & X \\ X^\top & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ -v_1 \end{bmatrix} = -\sigma_1 \begin{bmatrix} u_1 \\ -v_1 \end{bmatrix}.
\]
Now we can apply the Davis-Kahan Theorem (and $\sin \frac{\theta}{2} \leq \sin \theta$) to obtain
\[
\min_{s \in \{\pm 1\}} \left\| \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} + s \begin{bmatrix} \tilde{u}_1 \\ \tilde{v}_1 \end{bmatrix} \right\|_2 \leq \frac{2}{\sigma_1(X) - \sigma_2(Y)} \left\| \begin{bmatrix} 0 & Z \\ Z^\top & 0 \end{bmatrix} \right\|_{op} = \frac{2\|Z\|_{op}}{|\sigma_1(X) - \sigma_2(Y)|}.
\]
§ 4. Basic Random Matrix Theory

Let $Z = (Z_{ij}) \in \mathbb{R}^{n \times n}$ be an i.i.d or a symmetric with upper part i.i.d (i.e. $Z_{ij} = Z_{ji} \overset{i.i.d}{\sim} P$ for $1 \leq i \leq j \leq n$) matrix with $\mathbb{E}Z_{ij} = 0$ and bounded variance and sub-Gaussian tails. We will prove next that w.h.p.

$$\|Z\|_{op} \leq C\sqrt{n}.$$ 

In comparison we have w.h.p.

$$\|Z\|_{F} \asymp n.$$ 

For now we will discuss some intuition why $\sqrt{n}$ is the right order for the operator norm by focusing on $P = \mathcal{N}(0, 1)$. First observe that

$$Z = \begin{bmatrix} Z_{1}^\top \\ \vdots \\ Z_{n}^\top \end{bmatrix}$$

and $\|Z\|_{op} \asymp \sqrt{n}$ by the CLT.

Intuition 1: Observe that $\theta_i \overset{\Delta}{=} \frac{Z_i}{\|Z_i\|_2} \sim \text{Unif}(S^{n-1})$ and that $\langle \theta_i, \theta_j \rangle \asymp \frac{1}{n} \langle Z_i, Z_j \rangle = \frac{\sum_k x_{ik}x_{jk}}{n} \sim \frac{1}{\sqrt{n}}$. Therefore the rows of $Z$ are almost orthogonal and hence the operator norm should roughly equal the $\| \cdot \|_2$-norm of the largest row.

Intuition 2: Since $\|Z\|_{op} = \sup_v \frac{\|Zv\|_2}{\|v\|_2}$, fix a particular $v \in S^{n-1}$. Then

$$Zv = \begin{bmatrix} \langle Z_1, v \rangle \\ \vdots \\ \langle Z_n, v \rangle \end{bmatrix} \sim \mathcal{N}(0, I_n).$$

This shows that $\|Z\|_{op} \geq \sqrt{n}$. A better choice is $v = Z_1$. Indeed, in that case

$$Zv = \begin{bmatrix} \|Z_1\|_2^2 \\ \langle Z_2, Z_1 \rangle \\ \vdots \\ \langle Z_n, Z_1 \rangle \end{bmatrix}.$$ 

As before $\|Z_1\|_2^2 \asymp n$ and $\sum_{i>1}(Z_i, Z_1)^2 \asymp n^2$ which yields that $\|Zv\|_2 \asymp \sqrt{2}n$, which shows that $\|Z\|_{op} \geq \sqrt{2n}$. In fact one can do even better and prove that w.h.p.

$$\|Z\|_{op} = (2 + o(1))\sqrt{n}.$$
4.1 Gaussian Random Matrix

For simplicity, let’s consider $Z$ having independent $N(0, 1)$ off-diagonals and $N(0, 2)$ diagonals. It will become transparent that the variance of the diagonal is immaterial, provided it is small, say, a constant. This model is referred to as Gaussian Orthogonal Ensemble (GOE).

$$\|Z\|_{op} = \sigma_{\text{max}} = \max_{\|v\|_2=1} |\langle Z, vv^\top \rangle|.$$  

For $\forall$ fixed $v \in S^{n-1},$

$$\langle Z, vv^\top \rangle = \sum_i Z_{ii}v_i^2 + 2\sum_{i<j} Z_{ij}v_iv_j \sim N(0, 2\sum_i v_i^4 + 4\sum_{i<j} v_i^2v_j^2) = N(0, 2).$$  

$$\Rightarrow \mathbb{P}(|\langle Z, vv^\top \rangle| > t) \leq 2e^{-\frac{t^2}{4}}, \forall t > 0.$$

**Remark 4.1.** The distributions of the diagonals are not important for the operator norm. To see this, note

$$\|Z\|_{op} \leq \|Z_o\|_{op} + \|\text{diag}(Z)\|_{op}$$

where $Z_o$ is the same as $Z$ except that the diagonals are set to zero, and diag$(Z) = \text{diag}(Z_{ii})$. By union bound, $\|\text{diag}(Z)\|_{op} = \max_{1 \leq i \leq n} |Z_{ii}| = O_p(\sqrt{\log n}) \ll O_p(\sqrt{n})$, and thus negligible.

To bound $\mathbb{P}(\max_{v \in S^{n-1}} |\langle Z, vv^\top \rangle| > t)$, we would like to apply the union bound. However, the sphere here is not a finite set. In order to handle this, we can use the discretization technique — the $\epsilon$-net argument.

**Definition 4.1.** $V \subset S^{n-1}$ is called an $\epsilon$-net (covering), if $\forall u \in S^{n-1}, \exists v \in V$ s.t. $\|u - v\|_2 \leq \epsilon$.

**Lemma 4.1.** For $V$ an $\epsilon$-net,

$$\max_{v \in V} |\langle Z, vv^\top \rangle| \leq \|Z\|_{op} \leq \frac{1}{1 - 2\epsilon} \max_{v \in V} |\langle Z, vv^\top \rangle|.$$  

**Proof.** Choose $u \in S^{n-1}$ such that $|\langle Z, uu^\top \rangle| = \|Z\|_{op}$. Then $\exists v \in V$ such that $\|u - v\|_2 \leq \epsilon$. It follows that

$$\|Z\|_{op} = |\langle Z, uu^\top \rangle| \leq |\langle Z, vv^\top \rangle| + |\langle Z, uu^\top - vv^\top \rangle|$$

$$= |\langle Z, vv^\top \rangle| + |\langle Z, uu^\top - uv^\top + uv^\top - vv^\top \rangle|$$

$$\leq |\langle Z, vv^\top \rangle| + |\langle Z, (u - v)^T \rangle| + |\langle Z, (u - v)v^\top \rangle|$$

$$\leq |\langle Z, vv^\top \rangle| + 2\|Z(u - v)\|_2$$

$$\leq \max_{v \in V} |\langle Z, vv^\top \rangle| + 2\epsilon \|Z\|_{op}. \quad \blacksquare$$

**Definition 4.2.** For $A \subset \mathbb{R}^d$, $V = \{v_1, \ldots, v_m\} \subset A$ is called an $\epsilon$-packing, if $\forall i \neq j$, $\|v_i - v_j\|_2 \geq \epsilon$. An $\epsilon$-packing $V$ is maximal if it cannot be made bigger, i.e., $\forall u \in A \setminus V$, $V \cup \{u\}$ is not an $\epsilon$-packing.

We make two key observation for these concepts:

- Any maximal $\epsilon$-packing is an $\epsilon$-net.
• ∀ ε-packing V of A, |V| ≤ vol(A + \frac{\varepsilon}{2} B)/vol(\frac{\varepsilon}{2} B), where B is the unit norm ball. Here A + B ≜ {x + y : x ∈ A, y ∈ B} is the Minkowski sum of two sets.

The first observation is just by definition. We can construct a maximal ε-packing through greedy search. The second one is because we can put |V| balls of radius \frac{\varepsilon}{2} into A + \frac{\varepsilon}{2}B and keep them disjoint. So the total volume of balls should not exceed that of the A + \frac{\varepsilon}{2} B. Among many measures, we choose volume because it’s location invariant. We can summarize the observations as

covering # ≤ packing # ≤ volume ratio.

Now set A = S^{n-1}. Then A + \frac{\varepsilon}{2} B ⊂ B + \frac{\varepsilon}{2} B = (1 + \frac{\varepsilon}{2}) B.\footnote{The first inclusion does not lose much volume, because the volume of a ball in high dimension is concentrated near the shell anyway.} The volume ratio

\[
\frac{vol(A + \frac{\varepsilon}{2} B)}{vol(\frac{\varepsilon}{2} B)} \leq \frac{vol((1 + \frac{\varepsilon}{2}) B)}{vol(\frac{\varepsilon}{2} B)} = \frac{(1 + \frac{\varepsilon}{2})^n vol(B)}{(\frac{\varepsilon}{2})^n vol(B)} = \left(1 + \frac{2}{\varepsilon}\right)^n.
\]

What we discussed above concludes the following lemma.

Lemma 4.2 (Size of ε-net). There exists an ε-net V for S^{n-1}, of size |V| ≤ (1 + \frac{2}{\varepsilon})^n.

Theorem 4.1. \|Z\|_{op} ≤ C\sqrt{n} w.h.p..

Proof. Set \varepsilon = \frac{1}{4} and choose V as in Lemma 4.2. By Lemma 4.1, \|Z\|_{op} ≤ 2 max_{v \in V} |\langle Z, vv^\top \rangle|.

\[
\mathbb{P}(\max_{v \in V} |\langle Z, vv^\top \rangle| > t) \leq \sum_{v \in V} \mathbb{P}(|\langle Z, vv^\top \rangle| > t) \leq |V| \cdot 2e^{-\frac{t^2}{2}} = 2e^{n\log 9 - \frac{t^2}{2}}.
\]

Choose t = \frac{C}{2}\sqrt{n} with C > 4\sqrt{\log 9} a universal constant. Then we know \|Z\|_{op} ≤ C\sqrt{n} with probability at least 1 − 2e^{-C'n}, where C' = C^2/16 − \log 9 > 0.

\[
4.2 \text{ Sub-Gaussian Random Matrix}
\]

Reviewing the whole proof of Theorem 4.1, we can see there is only one part that the Gaussian assumption is used: the tail bound \mathbb{P}(|\langle Z, vv^\top \rangle| > t) ≤ 2e^{-\frac{t^2}{2}}. Thus the result of Theorem 4.1 can be naturally extended to other r.v.s with such tail bound.

Definition 4.3. A r.v. X is sub-Gaussian (SG) with parameter \sigma^2 if ∀λ, \mathbb{E}e^{λ(X−EX)} ≤ e^{σ^2λ^2/2}.

For a σ^2-SG r.v. X and t > 0, a direct result of Chernoff bound is \mathbb{P}(X−EX > t) ≤ e^{σ^2t^2/2−λt}, ∀λ. Choose λ = t/σ^2, and we have \mathbb{P}(X−EX > t) ≤ e^{-\frac{t^2}{2σ^2}}. The similar result for the other side tail combines to show \mathbb{P}(|X−EX| > t) ≤ 2e^{-\frac{t^2}{2σ^2}}. Thus, a σ^2-SG r.v. do have the same tail bound as N(0, σ^2). We can also view the tail bound as the definition of σ^2-SG. σ^2-SG r.v.s have variance at most σ^2, which can be shown easily through Taylor expansion.

Lemma 4.3 (Hoeffding). Bounded r.v.s are SG. If X ∈ [−a, a] a.s. for some a > 0, then it’s 4a^2-SG.
Proof. First, we prove when $X$ is a Rademacher r.v.,

$$\mathbb{E}e^{\lambda X} = \frac{1}{2}(e^\lambda + e^{-\lambda}) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{2^k k!} = e^{\lambda^2/2}.$$  

Second, when $|X| \leq a$ a.s., we apply so-called symmetrization technique. Let $X' \sim X$ and $\epsilon$ be a Rademacher r.v., and they three are independent. Then $X - X'$ has symmetric distribution and $\epsilon(X - X') = \mathcal{D} X - X'$.

$$
\mathbb{E}e^{\lambda (X - \mathbb{E}X)} = \mathbb{E}e^{\lambda (X - \mathbb{E}X')} = \mathbb{E}e^{\lambda X} e^{-\lambda \mathbb{E}X'} \\
\leq \mathbb{E}e^{\lambda X} \mathbb{E}e^{-\lambda X'} = \mathbb{E}e^{\lambda (X - X')} \\
= \mathbb{E} \left( \mathbb{E}(e^{\lambda(X-X')}|X, X') \right) \\
\leq \mathbb{E} \left( e^{\lambda^2(X-X')^2/2} \right) \leq e^{2\lambda^2 a^2}.
$$

The last inequality is because $|X - X'| \leq 2a$ a.s.. \hfill \Box

Remark 4.2. This bound is good (tight enough) if $\text{Var}(X) \approx a^2$ and is loose when $\text{Var}(X) = o(a^2)$. For example, let’s consider $X \sim \text{Bern}(p)$, which we will encounter in studying sparse random graphs. By Lemma 4.3, it’s 1-SG regardless of the value of $p$. In contrast, $\text{Var}(X) \approx p = o(1)$. We can improve the sub-Gaussian constant to $\sigma^2(p) \lesssim 1/\log^2 p$, and this is the best possible. Thus, dealing with sparse graphs using sub-gaussian technology leads to highly suboptimal result.

Lemma 4.4. 1. If $X$ is $\sigma^2$-SG, then $\alpha X$ is $\alpha^2 \sigma^2$-SG.

2. If $X_1, \ldots, X_n$ are independent $\sigma_i^2$-SG, then $\sum_{i=1}^n X_i$ is $(\sum_{i=1}^n \sigma_i^2)$-SG.

3. If $X$ is $\sigma^2$-SG and $\tau^2 > \sigma^2$, then $X$ is $\tau^2$-SG.

Theorem 4.2. Suppose $Z = (Z_{ij})_{n \times n}$ a real symmetric matrix with $\mathbb{E}Z = 0$. For $i \leq j$, $Z_{ij} \sim \sigma^2$-SG and are independent. Then $\|Z\|_{op} \leq C \sqrt{n \sigma^2}$ with probability at least $1 - 2e^{-C' n}$ for some universal constant $C, C'$.

Proof. For $\forall v \in S^{n-1}$, $\langle Z, vv^\top \rangle = \sum_i Z_{ii}v_i^2 + 2 \sum_{i<j} Z_{ij}v_i v_j$ is SG with parameter

$$
\sigma^2 \left( \sum_i v_i^4 + 4 \sum_{i<j} v_i^2 v_j^2 \right) \leq 2 \sigma^2 \left( \sum_i v_i^2 \right)^2 = 2 \sigma^2.
$$

We have $\mathbb{P}(|X - \mathbb{E}X| > t) \leq 2e^{-t^2/4\sigma^2}$. The rest is identical to the proof of Theorem 4.1. \hfill \Box
§ 5. Spectral method for Planted Clique

We now apply the perturbation result from Lecture 3 and the basic random matrix result from Lecture 4 to the hidden clique problem, following [AKS98].

5.1 Spectral methods for Planted Clique Model

Let $G \sim G(\frac{1}{2}, n, k)$. In other words, $K \subset [n]$ is a hidden $k$-clique, and $G$ has adjacency matrix

$$A_{ij} = \begin{cases} 1 & i, j \in K \\ \text{Bern}(\frac{1}{2}) & \text{o/w} \end{cases}.$$ 

Let $W$ be a real symmetric matrix that

$$W_{ij} = \begin{cases} 2A_{ij} - 1 & i \neq j \\ 0 & i = j \end{cases}.$$ 

The following is a spectral method to find the clique:

1. Find the top eigenvector $u$ of $W$.
2. Let $\tilde{K}$ be the index vector of $k$ largest $|u_i|$.
3. (Clean up) Define $\hat{K}$ as the set of vertices in $G$ having $\geq 3k^4$ neighbors in $\tilde{K}$. In other words,

$$\hat{K} = \{i \in V(G) : d_{\tilde{K}}(i) \geq 3k^4\}.$$ 

**Theorem 5.1 ([AKS98]).** If $k \geq C\sqrt{n}$ for $C$ large enough, then $\mathbb{P}(\hat{K} = K) \to 1$.

**Proof.** First, we show $\tilde{K}$ is approximately correct: $|\tilde{K} \cap K| \geq (1-\epsilon)k$ w.h.p for some $\epsilon = \epsilon(C)$.

Fix some small $\epsilon > 0$ that we will choose later. Let $W^* = \xi\xi^\top$, $\xi = 1_K = (1 \{i \in K\})_{1 \leq i \leq n}$. Now $W^*$ is rank one, so $v = \frac{1}{\sqrt{k}}\xi$ is its top eigenvector. By Davis-Kahan’s sin $\Theta$-theorem (Theorem 3.2), provided that $\lambda_1(W^*) > \lambda_2(W) > 0$,

$$\min_{\pm} \|u \pm v\|_2 \leq \frac{\|W - W^*\|_{\text{op}}}{\lambda_1(W^*) - \lambda_2(W)}.$$  \hspace{1cm} (5.1)

WLOG, assume the LHS is $\|u - v\|_2$. Note that $\|W - W^*\|_{\text{op}} \leq \|\mathbb{E}W - W\|_{\text{op}} + \|\mathbb{E}W - W^*\|_{\text{op}} \leq \|\mathbb{E}W - W\|_{\text{op}} + 1 \leq C_0\sqrt{n} + 1$ w.h.p for some universal $C_0 > 1$ by Theorem 4.2. By Weyl’s inequality (Theorem 3.1), $|\lambda_2(W)| = |\lambda_2(W^*) - \lambda_2(W)| \leq \|W - W^*\|_{\text{op}} \leq C_0\sqrt{n} + 1$ under the event above. And we know $\lambda_1(W^*) = \|\xi\|^2_2 = k$. Plug back into (5.1) and we get that

$$\|u - v\|_2 \leq \frac{C_0\sqrt{n} + 1}{C\sqrt{n} - C_0\sqrt{n} - 1} \leq \epsilon$$  \hspace{1cm} (5.2)
holds w.h.p for $C$ big enough.

Second, we claim that $\|u - v\|_2 \leq \epsilon$ actually implies

$$|\tilde{K} \cap K| \geq (1 - \epsilon')k.$$  \hfill (5.3)

To see this, note that $|K| = |\tilde{K}| = k$, thus $|K \setminus \tilde{K}| = |\tilde{K} \setminus K|$. If $\exists i \in (5.2)$ is guaranteed. Therefore, $\epsilon$ the whole proof, we choose $\|\epsilon \|

By Chernoff bound, for all $\epsilon$, the whole proof, we choose $\|\epsilon \|$

In all, in either case, (5.3) holds with $\epsilon' = 4\epsilon^2$.

Third, we claim that $\tilde{K} = K$ with high probability. Think under the event of $\|u - v\|_2 \leq \epsilon$. If vertex $i \in K$, $d_{\tilde{K}}(i) \geq d_{K \cap \tilde{K}}(i) \geq |\tilde{K} \cap K| - 1 \geq (1 - \epsilon')k - 1$. So vertex $i \in \tilde{K}$ when $\epsilon' < \frac{1}{4}$. If vertex $i \notin K$, $d_{\tilde{K}}(i) \leq d_{K}(i) + |\tilde{K} \setminus K|$. From above, we know $|K \setminus \tilde{K}| \leq \epsilon'k$. And $d_{K}(i) \sim \text{Bin}(k, \frac{1}{2})$. By Chernoff bound, for all $\epsilon' \leq 1/8$,

$$\mathbb{P}(d_{K}(i) \geq (\frac{3}{4} - \epsilon')k) \leq \mathbb{P}(d_{K}(i) \geq \frac{5}{8}k) \leq e^{-\frac{k}{\frac{1}{2}}}.$$

Applying a union bound over all vertices $i \notin K$ gives that

$$\mathbb{P}\left( \forall i \notin K : d_{K}(i) \geq (\frac{3}{4} - \epsilon')k \right) \leq ne^{-\frac{\frac{k}{\frac{1}{2}}}.$$

In all, under events $\|u - v\| \leq \epsilon$ and $d_{K}(i) \leq (\frac{3}{4} - \epsilon')k$ for all $i \notin K$, we have $\tilde{K} = K$. To wrap up the whole proof, we choose $\epsilon = \frac{1}{3}$. Then $\epsilon' = 4\epsilon^2 = \frac{1}{16}$. Choose $C \geq 17C_0$ then the second inequality in (5.2) is guaranteed. Therefore,

$$\mathbb{P}(\tilde{K} \neq K) \leq \mathbb{P}(\|u - v\| > \epsilon) + \mathbb{P}\left( \forall i \notin K : d_{K}(i) > (\frac{3}{4} - \epsilon')k \right)$$

$$\leq \mathbb{P}(\|W - \mathbb{E}W\|_{op} > C_0\sqrt{n}) + ne^{-\frac{k}{\frac{1}{2}}}$$

$$\leq 2e^{-C_0n} + ne^{-\frac{\frac{k}{\frac{1}{2}}} \sqrt{n}} \to 0.$$

\hfill \Box

Remark 5.1. 1. An alternative algorithm can take $u$ as the second leading eigenvector of $A$. The top eigenvector of $A$ is almost deterministic and not informative, since it is almost $\propto 1$.

2. Thresholding technique is widely used in non-parametric estimation. Here, the step 3 (clean up) can be viewed as a version of thresholding.
5.2 Improving the constant

Next we show that the constant $C$ in Theorem 5.1 can be made arbitrarily small, at the pricing of increasing the time complexity (still poly($n$) but with bigger exponent). This part is generic and applies to any algorithm. The idea is as following. Fix a subset of vertices $S \subset V$, $|S| = s$. Define $N_s(S)$ as the common neighbor of $S$, i.e., $N_s(S) = \{v \in V \setminus S : \forall u \in S, v \sim u\} = (\bigcap_{u \in S} N(u)) \setminus S$. Let’s say $s = 2$. Next, look at the induced subgraph $G' = G[N_s(S)]$. If $S \subset K$, then $G' = G(|N_s(S)|, \frac{1}{2}, k - 2)$ and $|N_s(S)| \sim \text{Bin}(n - k, \frac{1}{2}) + k - 2 = (1 + o_P(1))\frac{n}{2}$ when $k = o(n)$. So as we can see, by this subgraph operation, $n$ decays exponentially while $k$ decays linearly. The upgraded algorithm is summarized below:

For any $s$-subset $S \subset V$, run the existing algorithm on $G' = G[N_s(S)]$ and output $Q$. Repeat until $S \cup Q$ is a $k$-clique. And the final output is $S \cup Q$.

When the search over $S$ finds $S \subset K$, the requirement in Theorem 5.1 asks for $k - s \geq C\sqrt{n} \cdot 2^{-s}$ to guarantee consistency of $Q$, thus also guarantees the consistency of $S \cup Q$ in the original graph $G$. Pick $s \approx 2\log_2\frac{C}{\delta}$, then the algorithm above is guaranteed to be consistent with requirement $k \geq \delta\sqrt{n}$. The extra search time is at most $\binom{n}{s} \approx n^s$ that is polynomial in $n$. 
In this lecture we discuss semidefinite programming (SDP) relaxation in the context of the planted clique problem. We discuss two SDPs:

- A standard form of SDP relaxation for the planted clique problem (after [FK00, HWX16])
- A convexified maximum likelihood estimator with nuclear norm constraint, which can also be written as SDP (after [CX16]).

We show that both methods succeed in finding the hidden clique of size $\Omega(\sqrt{n})$ with high probability, using the following two types of proof techniques respectively:

- Dual proof: we construct the needed Lagrangian multipliers (also called dual certificates or dual witnesses) that together with the desired solution $X^*$ fulfill the KKT condition, thereby certifying the optimality of $X^*$.
- Primal proof: we show that no feasible solution other than the desired $X^*$ achieve a higher objective function.

Although the two methods are of distinct nature (one is constructive and one is non-constructive), for analyzing convex programs both methods are ultimately equivalent; nevertheless, the specific execution (e.g. explicit construction of dual certificates) need not be the same.

In addition, we show that the standard SDP relaxation of the hidden clique problem, unlike previously discussed methods like the degree test or the spectral method, is robust with respect to any monotone adversary. Aside from robustness, SDP also possesses the advantages of not requiring any cleanup step (opposed to spectral methods or the iterative degree tests), and that it can be solved in polynomial time. However empirically solving a large SDP can be quite slow.

6.1 The planted clique problem and spectral method revisited

Recall the in the planted clique model, a clique $K$ of size $k$ is planted in the Erdős-Rényi graph with success probability $1/2$. Denote $A$ as the corresponding adjacency matrix as $A$, and $G(n, 1/2, k)$ the distribution $A$ is generated from. Define $W \in \mathbb{R}^{n \times n}$ to be the following transformation of $A$.

$$W_{ij} = \begin{cases} 2A_{ij} - 1, & i \neq j; \\ 0, & i = j. \end{cases}$$

Note that $W$ takes value 1 on edges connecting two members within the clique, and is i.i.d. Rademacher(1/2) for all other off-diagonal entries. It is easy to see that the MLE for the planted
clique problem can be written as
\[
\hat{u}_{\text{MLE}} = \arg \max_u \sum_{i,j} u_i u_j W_{ij}
\]
s.t. \( u \in \{0, 1\}^n \),
\[
\sum u_i = k,
\]
or equivalently in matrix form,
\[
\hat{u}_{\text{MLE}} = \arg \max_u \langle W, uu^\top \rangle
\]
s.t. \( u \in \{0, 1\}^n \),
\[
\|u\|^2 = k.
\]
(6.1)

Notice that the spectral method is a relaxation of (6.1). By relaxation we mean enlarging the constraint set in order to speed up the computation. Indeed, the spectral method takes the top eigenvector of \( W \) by solving
\[
\hat{u}_{\text{spectral}} = \arg \max_u \langle W, uu^\top \rangle
\]
s.t. \( u \in \mathbb{R}^n \),
\[
\|u\|^2 = k.
\]
(6.2)

One would hope that after relaxation, the optimality of the optimizer in the original problem is not lost. Unfortunately that is not the case for the relaxation done in (6.2). We proved previously in Lecture 5 that the spectral method typically only recovers most of the members in the clique, thus requires a cleanup step to recover the entire clique. In other words, we relaxed the constraint set too much. We will develop two tighter relaxations to (6.1), namely our semidefinite programs. The standard form of SDP is derived using the idea of lifting. Let us first illustrate the lifting idea on spectral program (6.2).

**Definition 6.1.** An optimization problem is said to be convex if the objective function is a convex function and the constraint set is a convex set.

Although easy to solve, the optimization (6.2) is not convex. Nevertheless, it can be written as a convex optimization via the lifting idea.

**Definition 6.2.** A symmetric matrix \( X \in \mathbb{R}^{n \times n} \) is called positive semidefinite (PSD), denote by \( X \succeq 0 \), if \( y^\top X y \geq 0 \) for all \( y \in \mathbb{R}^n \).

Consider the following program:
\[
\hat{X}_{\text{spectral}} = \arg \max_X \langle W, X \rangle
\]
s.t. \( X \succeq 0 \),
\[
\text{Tr}(X) = k.
\]
(6.3)

**Proposition 6.1.** The optimization (6.2) is equivalent to (6.3).
To prove Proposition 6.1 we need to introduce the notion of extremal points in a convex set. Let $S$ be a convex set. We say a point $s \in S$ is an extremal point of $S$ if it cannot be written as convex combinations of other points in $S$. The importance of extremal points is that points in a convex set $S$ can be written as convex combinations of extreme points. There are various results of this flavor in convex analysis, the most general being the Krein-Milman theorem. We recall the following result for Euclidean space:

**Theorem 6.1 (Carathéodory’s theorem).** Suppose $S$ is a convex set in $\mathbb{R}^d$. Then for each $s \in S$, it can be written as a convex combination of $d + 1$ points in the set of extreme points of $S$.

**Proof of Proposition 6.1.** Write $X = uu^\top$ to reformulate (6.2) as

$$
\max_X \langle W, X \rangle \\
\text{s.t. } X \succeq 0, \\
\quad \text{Tr}(X) = k, \\
\quad \text{rank}(X) = 1.
$$

The lifting step only drops the rank one constraint on $X$. The proposition can be proved by arguing that dropping the rank constraint does not incur any sub-optimality. Notice that the objective function is linear in $X$ and the constraint set is convex. Therefore optimality occurs at one of the extreme points. To see that simply notice that for each feasible $X$, by Carathéodory’s theorem it can be written as a convex combination of a finite number of extreme points in the feasible set. Write $X = \sum_i X_i \alpha_i$, which gives

$$
\langle W, X \rangle = \sum_i \langle W, X_i \rangle \alpha_i.
$$

Therefore the objective function evaluated at $X$ has to be beaten (or at least match) its value at one of the extreme points. Given that all extreme points of the feasible set $\{ X : X \succeq 0, \text{Tr}(X) = k \}$ are of rank one, the rank constraint is automatically enforced by the optimization (6.3).

**6.2 Standard SDP relaxation**

**6.2.1 Formulation**

Start by rewriting the MLE program (6.1) in a lifted form.

$$
\hat{X}_{\text{MLE}} = \arg \max_X \langle W, X \rangle \\
\text{s.t. } X \succeq 0, \\
\quad 0 \leq X \leq J, \text{(entrywise)} \\
\quad \text{Tr}(X) = k, \\
\quad \langle X, J \rangle = k^2, \\
\quad \text{rank}(X) = 1.
$$

**Proposition 6.2.** The MLE optimization (6.1) is equivalent to (6.4).

**Proof.** It is easy to check that for all $u$ in the feasible set of (6.1), the matrix $uu^\top$ is in the feasible set of (6.4). We only need to check the other direction. In other words, we need to show that every $X$ in the feasible set of (6.4) can be written as $uu^\top$ for some $u$ in the feasible set of (6.1).
The positive semidefinite constraint combined with the rank one constraint imply that $X = uu^T$ for some $u \in \mathbb{R}^n$. The trace constraint gets translated to $\sum_i u_i^2 = k$; the constraint $\langle X, J \rangle = k^2$ is equivalent to $|\sum_i u_i| = k$. What’s more, by looking at the diagonal entries $X_{ii} = u_i^2 \leq 1$ we have $u_i \in [-1, 1]$. Also, $X_{ij} \geq 0$, so all the non-zeros of $u$ are either all positive or all negative. Assume the former. Then $\sum_i u_i^2 = k$ and $\sum_i u_i = k$ force the integrality $u_i \in \{0, 1\}$. Thus $u$ lies in the feasible set of (6.1).

The SDP form of the MLE is obtained by dropping the rank one constraint and the $X \leq J$ constraint in (6.4). Define

$$\hat{X}_{\text{SDP}} = \arg \max_X \langle W, X \rangle$$

s.t. $X \succeq 0$

$$X \succeq 0,$$

$\text{Tr}(X) = k$,

$\langle X, J \rangle = k^2.$ (6.5)

**Remark 6.1.** Recall that a standard form for a linear program is

$$\max_x \langle a, x \rangle$$

s.t. $\langle b_i, x \rangle \leq 0$ for $i = 1, \ldots, m$.

Compare with the standard form for an SDP:

$$\max_{X \text{ symmetric}} \langle W, X \rangle$$

s.t. $\langle B_i, X \rangle \leq 0$ for $i = 1, \ldots, m$,

$X \succeq 0$.

The set of all positive semidefinite matrices form a cone. The PSD constraint is a cone condition. Notice that a matrix $X \in \mathbb{R}^{n \times n}$ is call PSD if $\langle X, uu^T \rangle \geq 0$ for all $u \in \mathbb{R}^n$. Hence the PSD constraint can be viewed as a continuum of linear constraints.

### 6.2.2 Statistical guarantee: dual proof

Denote by $\xi = (1 \{i \in K\})_{1 \leq i \leq n}$ the indicator vector of the true clique $K$. Write $X^* = \xi \xi^T$ for the matrix marking if an edge connects two members in $K$. We show that if the size of the clique is of order $\sqrt{n}$, then the SDP (6.5) consistently recovers $X^*$.

**Theorem 6.2.** There exists a constant $c > 0$ such that if $k \geq c\sqrt{n}$, then the optimization (6.5) has unique maximizer that is $X^*$ with high probability.

**Remark 6.2.** With some work the constant $c$ can be reduced to 1. We will not optimize the choice of $c$ in the proof of the theorem.

**Remark 6.3.** Clearly $X^*$ is an optimizer of (6.5). To prove Theorem 6.2 we will mostly need to establish the uniqueness of $X^*$. To see the optimality of $X^*$, notice that

$$\langle W, X^* \rangle = k^2 - k$$

because $W$ takes value 1 for all off-diagonal entries in $K \times K$. 37
For each $X$ in the feasible set of (6.5),
\[
\langle W, X \rangle = \langle W + I, X \rangle - \text{Tr}(X) \\
\leq \langle J, X \rangle - \text{Tr}(X) = k^2 - k.
\]
The inequality is from $W + I \leq J$ and the entrywise non-negativity of $X$. Therefore $\langle W, X \rangle \leq \langle W, X^* \rangle$ for all feasible $X$.

The proof is via the standard dual approach to optimality: we will construct a set of Lagrangian multipliers for (6.5) (also called “dual certificates” or “dual witnesses”) that certifies the optimality of $X^*$.

Attach to each constraint in (6.5) a Lagrangian multiplier, i.e., $S \succeq 0$, $B \geq 0$, $\eta, \lambda \in \mathbb{R}$. Write out the Lagrangian for the SDP (6.5).

\[
L(X, S, B, \eta, \lambda) = \langle W, X \rangle + \langle S, X \rangle + \langle B, X \rangle + \eta(k - \text{Tr}(X)) + \lambda(k^2 - \langle X, J \rangle).
\]

Since the inner products of two PSD matrices is always nonnegative, one important observation is that
\[
\max_X \langle W, X \rangle \leq \min_{S, B, \eta, \lambda} \max_X L(X, S, B, \eta, \lambda). \quad (6.6)
\]

**Lemma 6.1.** Suppose there exists $S \succeq 0$, $B \geq 0$, $\eta, \lambda \in \mathbb{R}$ such that
\[
W + S + B - \eta I - \lambda J = 0, \quad \text{(first-order condition)}
\]
\[
\langle S, X^* \rangle = 0, \quad \langle B, X^* \rangle = 0, \quad \text{(complementary slackness)}
\]
\[
\lambda_{n-1}(S) > 0, \quad \text{(uniqueness)}
\]
then $X^* = \xi \xi^T$ is the unique global maximizer for (6.5).

**Proof.** At the high level, the reasoning for any duality result is always of the following type:

1. First-order condition ensures $L(X, S, B, \eta, \lambda)$ is the same for any feasible solution $X$;
2. Complementary slackness ensures $L(X^*, S, B, \eta, \lambda) = \langle W, X^* \rangle$.

Then we are done: $\forall$ feasible $X$,
\[
\langle W, X \rangle \leq L(X, S, B, \eta, \lambda) = L(X^*, S, B, \eta, \lambda) = \langle W, X^* \rangle.
\]

Indeed, let’s rewrite the Lagrangian multiplier as
\[
L(X, S, B, \eta, \lambda) = \langle X, W + S + B - \eta I - \lambda J \rangle + k \eta + k^2 \lambda.
\]
By the first order condition, the first term is 0 thanks to the first order condition. Hence $L(X, S, B, \eta, \lambda) = k \eta + k^2 \lambda$ does not depend on $X$. By (6.6), for any feasible $X$ we have
\[
\langle W, X \rangle \leq k \eta + k^2 \lambda.
\]
By the complementary slackness condition $X^*$ achieves the above with equality
\[
\langle W, X^* \rangle = k \eta + k^2 \lambda.
\]
This shows $X^*$ is a maximizer.
To prove $X^*$ is the maximizer (uniqueness), suppose for some feasible $X'$ we have

$$\langle W, X' \rangle = \langle W, X^* \rangle.$$  

Then we must have $\langle S, X' \rangle = 0$. Note that positive semidefiniteness of $S$ implies that $\langle S, X^* \rangle = 0$ is equivalent to $S \xi = 0$, i.e., $\xi$ is an eigenvector for the smallest eigenvalue $\lambda_n = 0$ of $S$. Since $X'$ is positive semidefinite and $S$ has a strictly positive second smallest singular value, $\langle S, X' \rangle = 0$ forces $X' = C \xi \xi^\top = CX^*$ for some constant $C$. The trace constraint ensures that $\text{Tr}(X) = \text{Tr}(X^*)$. Hence $C$ has to be one, meaning $X' = X^*$.

Proof of Theorem 6.2. From Lemma 6.1, it suffices to construct $B \geq 0, \eta \lambda \in \mathbb{R}$ such that

$$S = \eta I + \lambda J - B - W \succeq 0,$$

and $S \xi = 0, \langle B, X^* \rangle = 0, \lambda_{n-1}(S) > 0$.

The condition $S \xi = 0$ is equivalent to

$$\eta \xi + \lambda k 1 = B \xi + W \xi. \quad (6.7)$$

Recall that $X^* = \xi \xi^\top$. The condition $\langle B, X^* \rangle = 0$ is equivalent to $B_{ij} = 0$ for all $(i, j) \in K \times K$. Therefore for all $i \in K$, the $i$'th entry of $B \xi$ is zero. Denote $y = W \xi$. For $i \in K$, pull out the $i$'th place in the vector equality (6.7).

$$\eta + k \lambda = (B \xi)_i + y_i = k - 1.$$  

Deduce that $\eta = k(1 - \lambda) - 1$.

For $i \notin K$, we have

$$k \lambda = (B \xi)_i + y_i. \quad (6.8)$$

Construct $B$ from $B = \xi b^\top + b \xi^\top$ for some $b \in \mathbb{R}^n$ such that $b_i = 0$ for all $i \in K$. Such matrix $B$ is of rank 2 and takes the block form

$$B = \begin{bmatrix} 0 & \text{column-wise constant} \\ \text{row-wise constant} & 0 \end{bmatrix}.$$  

For $B$ defined as such, we have $B \xi = kb$. Hence (6.8) can be rewritten as

$$k \lambda = kb_i + y_i,$$

implying that the choice of $b$ satisfies $b_i = \lambda - y_i/k$ for all $i \notin K$.

We still need to ensure that $B \succeq 0$. Equivalently, we need $b_i \geq 0$ for all $i$. This entails

$$\lambda \geq \frac{1}{k} \max_{i \notin K} y_i,$$

We can choose $\lambda = 1/2$ (in reality $\lambda = o(1)$ works) to ensure this holds with high probability.

It remains to verify $S \succeq 0$ and $\lambda_{n-1}(S) > 0$. In other words, we need

$$x^T S x > 0 \text{ for all } x \in S^{n-1} \text{ s.t. } \langle x, \xi \rangle = 0.$$

39
Use the first-order condition to write
\[ x^T S x = \eta + \lambda x^T J x - x^T B x - x^T W x. \]
The second term in the right-hand side is nonnegative by positive semidefiniteness of the all ones matrix. We have \( x^T B x = 0 \) from \( \langle x, \xi \rangle = 0 \). Write \( W = \mathbb{E} W + (W - \mathbb{E} W) \) to deduce that
\[ x^T S x \geq \frac{k}{2} - 1 - x^T \mathbb{E} W x - \|W - \mathbb{E} W\|_{op}. \]
Note that \( \mathbb{E} W = \xi \xi^T - \text{diag}(\xi) \), therefore \( x^T \mathbb{E} W x \leq 0 \) for all \( x \perp \xi \). We proved that \( \|W - \mathbb{E} W\|_{op} \leq c' \sqrt{n} \) with high probability. Hence \( x^T S x \geq (k - c' \sqrt{n}) > 0 \) as long as \( k \geq c \sqrt{n} \) for some large enough constant \( c \).

### 6.2.3 A negative result

We showed exact recovery (with high probability) of the clique of the SDP (6.5) under the assumption that \( k \) is at least some large multiple of \( \sqrt{n} \). We will show the requirement on the clique size cannot be improved. In fact if \( k \) is smaller than some small constant multiple of \( \sqrt{n} \), there will typically be a spurious maximizer for (6.5) that does not provide any information on the location of the clique.

**Theorem 6.3.** If \( k \leq c \sqrt{n} \) (for example \( c = 1/2 \) works), then with high probability, the SDP (6.5) has a maximizer \( X \in \mathbb{R}^{n \times n} \) that is supported by \( K^C \times K^C \). In other words,

1. \( X \) is in the feasible set of (6.5).
2. \( \langle W, X \rangle = \langle W, X^* \rangle = k^2 - k \).

**Proof.** Define \( Z = W_{K^C, K^C} \in \mathbb{R}^{m \times m} \), where \( m = n - k \). We will construct the feasible alternative solution \( X \) as follows. Let
\[ X = \begin{bmatrix} 0 & 0 \\ 0 & Y \end{bmatrix}, \]
where \( Y \in \mathbb{R}^{m \times m} \) is a matrix we need to specify so that \( X \) satisfies the feasibility conditions. In terms of \( Y \), that means

1. \( Y \succeq 0 \),
2. \( Y \geq 0 \),
3. \( \text{Tr}(Y) = k \),
4. \( \langle Y, J \rangle = k^2 \).

Let
\[ Y = \frac{k}{m} I + \alpha Z + \beta (J - I). \]
In other words, \( Y_{ii} = \frac{k}{m} \) and \( Y_{ij} = \alpha Z_{ij} + \beta \). It is easy to check that conditions 2 is satisfied as long as \( \beta \geq \alpha \geq 0 \). Condition 3 is satisfied by specifying the diagonal entries of \( Y \) to be \( k/m \). Condition 4 translates to
\[ k + \alpha \langle Z, J \rangle + \beta m (m - 1) = k^2. \]

We also need \( \langle X, W \rangle = k(k - 1) \), i.e.
\[ \alpha m (m - 1) + \beta \langle Z, J \rangle = k(k - 1). \]
Solve the linear system (6.9) with (6.10) to obtain the pair
\[
\alpha = \beta = \frac{k^2 - k}{m^2 - m - \langle Z, J \rangle} = \frac{k^2}{n^2}(1 + o_P(1))
\]
for \( k \leq \sqrt{n}/2 \).

It remains to ensure condition 1 holds. To show that \( Y \succeq 0 \) it suffices to show
\[
\frac{k}{n} \geq \alpha \|Z\|_{\text{op}}.
\]
The operator norm of the i.i.d. Radamacher matrix \( Z \) is smaller than \( 2\sqrt{n} \) with high probability. Therefore, the right-hand side can be upper bounded by \( k^2/n^2 \sqrt{n}(2 + o_P(1)) \). This means \( k \leq \sqrt{n}/2 + o_P(1) \).

### 6.2.4 Rounding an approximate optimal solution

Unlike linear programs, there are no known solver for semidefinite programs that outputs the exact optimizer in polynomial time. Via the ellipsoid method, an SDP can only be solved in polynomial time up to some accuracy. To be exact, it is possible to produce a feasible solution, in \( \text{poly}(n, m, 1/\epsilon) \) time, e.g. by the ellipsoid method, whose objective value is at least \( (1 - \epsilon)\text{OPT} \), where \( n \) = number of variables, \( m \) = number of constraints and \( \epsilon \) is the relative accuracy. Thus, in order to obtain a poly-time algorithm, we need to that show, as a sanity check, that the same statistical guarantee can be attained if we only solve the SDP relaxation up to certain relative accuracy \( \epsilon = \frac{1}{\text{poly}(n)} \) followed by some simple post processing (rounding); otherwise, it defeats the purpose.

In the context of the planted clique problem, we showed in Theorem 6.2 that as long as \( k \geq c\sqrt{n} \) for some constant \( c \), then with high probability, \( X^* \) is the unique optimizer of (6.5). Suppose we found a feasible solution \( X \) to (6.5), such that
\[
\langle W, X \rangle \geq (1 - \epsilon)\langle W, X^* \rangle.
\]
Next, we can apply simple rounding scheme to convert \( X \) to \( \hat{X} \in \{0,1\}^{n \times n} \), where
\[
\hat{X}_{ij} = \begin{cases} 
0 & \text{if } X_{ij} \leq \frac{1}{2}, \\
1 & \text{if } X_{ij} > \frac{1}{2}. 
\end{cases}
\]

**Theorem 6.4 (Rounding).** Under the assumptions of Theorem 6.2, if \( \epsilon \leq c_1 \sqrt{n}/k^3 \) for some \( c_1 > 0 \), then \( \hat{X} = X^* \) with high probability.

**Proof.** Suppose, for the sake of contradiction, that \( \hat{X} \neq X^* \). By assumption \( \langle W, X \rangle \geq \langle W, X^* \rangle - \delta \), with \( \delta = k(k - 1)\epsilon \). From the definition of \( \hat{X} \), we know that \( \hat{X} \neq X^* \) means
\[
\exists (i_0, j_0) \in K \times K, \quad \text{s.t. } X^*_{i_0,j_0} = 1, \quad \text{but } X_{i_0,j_0} \leq \frac{1}{2};
\]
or
\[
\exists (i_1, j_1) \notin K \times K, \quad \text{s.t. } X^*_{i_1,j_1} = 0, \quad \text{but } X_{i_1,j_1} > \frac{1}{2}.
\]

\(^1\)This should be contrasted with the positive result \( k \leq \sqrt{n}(2 + o_P(1)) \). In fact, this can be improved to \( k \leq \sqrt{n}(1 + o_P(1)) \) by choosing \( \lambda = o(1) \) in the proof of Theorem 6.2.
As a consequence, we have \( \|X^* - X\|_F^2 > 1/2 \). However, there is no contradiction to the optimality gap yet, because when \( i_2, j_2 \) is an edge, it might balance the loss of the objective value inside the clique.

Next, we use the dual variables to assess the suboptimality. Recall the set of dual certificates constructed in the proof of Theorem 6.2:

\[
S \succeq 0, \quad \text{s.t. } S\xi = 0, \quad B \geq 0, \quad \eta, \lambda \in \mathbb{R}.
\]

Under the assumptions of Theorem 6.2 we showed

\[
\lambda_{n-1}(S) \geq c_2 \sqrt{n}
\]

for some \( c_2 > 0 \) with high probability. Recall that if the first order conditions on the Lagrangian multiplier are satisfied, the Lagrangian \( L(X, S, B, \eta, \lambda) = k\eta + k^2\lambda \) does not depend on \( X \). Deduce that

\[
\langle W, X^* - X \rangle = \langle S, X \rangle + \langle B, X \rangle.
\]

Denote the optimality gap \( \delta = \langle W, X^* - X \rangle \), which satisfies \( \delta \leq k(k-1)\epsilon \), we have from the above

\[
\langle S, X \rangle \leq \delta.
\]

Let \( u = \xi/\sqrt{k} \) be the (unit) eigenvector of \( S \) corresponding to zero eigenvalue. We have

\[
S \succeq \lambda_{n-1}(S) \left( I - uu^\top \right).
\]

Therefore by positive semidefiniteness of \( X \),

\[
\langle S, X \rangle \geq c_2 \sqrt{n} \langle X, I - \frac{1}{k} X^* \rangle.
\]

Together with the upper bound on \( \langle S, X \rangle \) we have

\[
\langle X, I - \frac{1}{k} X^* \rangle \leq \frac{\delta}{c_2 \sqrt{n}} \quad \text{Tr}(X) = k \quad \iff \quad \langle X, X^* \rangle \geq k^2 - \frac{k\delta}{c_2 \sqrt{n}}.
\]

We are now ready to obtain an upper bound on \( \|X^* - X\|_F^2 \).

\[
\|X^* - X\|_F^2 = \|X^*\|_F^2 + \|X\|_F^2 - 2 \langle X, X^* \rangle.
\]

The truth \( X^* = \xi\xi^\top \) is with Frobenius norm \( k^2 \). Again thanks to \( X \succeq 0 \),

\[
\|X\|_F \leq \|X\|_* = \text{Tr}(X) = k.
\]

Therefore

\[
\|X^* - X\|_F^2 \leq k^2 + k^2 - 2 \left( k^2 - \frac{k\delta}{c_2 \sqrt{n}} \right) = \frac{2k\delta}{c_2 \sqrt{n}} \leq 0.2
\]

as long as \( \delta < 0.1c_2 \sqrt{n}/k \), which is satisfied if \( \epsilon < 0.1c_2 \sqrt{n}/k^3 \). We have arrived a contradiction with \( \|X^* - X\|_F \geq 0.5 \).
6.2.5 Robustness against monotone adversary

In this section we consider a semi-random model to address the robustness of the methods we analyzed so far, following [FK00, FK01]. We show that the standard SDP relaxation of the hidden clique problem, unlike previously discussed methods like the degree test or the spectral method, is robust with respect to any monotone adversary. An adversary is a (possibly random) modification on the observed data designed to derail certain estimators. A monotone adversary in the context of the planted clique problem takes the observed adjacency matrix $A \sim G(n, 1/2, k)$ and it is allowed to arbitrarily delete edges while leaving the clique intact.

An important observation is that although monotone adversary seems to be helpful as it only makes the clique more pronounced, it can break the consistency of both the degree test and spectral method.

- We proved in previous lectures that as long as the size of the clique is of order $k \approx \sqrt{n \log n}$, the degree test consistently recovers location of the true clique $K$. To design a monotone adversary that breaks the degree test, simply remove all edges between $i \in K$ and $j \in K^c$. The modified adjacency matrix takes the form

$$\tilde{A} \sim \begin{bmatrix} 1 & 0 \\ 0 & G(n-k, \frac{1}{2}) \end{bmatrix}.$$ 

After modification, the degree of vertex $i$ for $i \in K$ is $k$; while the degree of $i$ for $i \notin K$ is distributed $Bin(n-k, 1/2)$. The degree test will obviously fail at picking out the location of $K$.

- To break the spectral method, design the monotone adversary as follows:

$$\tilde{A} \sim \begin{bmatrix} k\text{-clique} & 0 & 0 \\ 0 & G(\frac{n-k}{2}, \frac{1}{2}) & 0 \\ 0 & 0 & G(\frac{n-k}{2}, \frac{1}{2}) \end{bmatrix}.$$ 

The top two eigenvectors of $\tilde{A}$ would typically be aligned with the two blocks contained in $K^c$ (with $\lambda_1 \approx \lambda_2 \approx \frac{n-k}{4}$), and they are no longer informative about the hidden clique $K$.

Next we argue that $\tilde{X}^{SDP}$ is automatically robust against monotone adversary. Denote the corresponding $W$ after modification as $\tilde{W}$. Then for all feasible $X \neq X^*$, we have

$$\langle \tilde{W}, X \rangle \leq \langle W, X \rangle < \langle W, X^* \rangle = \langle \tilde{W}, X^* \rangle.$$ 

where the strict inequality is because $X^*$ is the unique global maximizer of the original problem, and the last equality is because the adversary kept the planted clique intact. In other words, whenever $X^*$ is the maximizer under $W$, it remains the maximizer under $\tilde{W}$.

6.3 Convexified MLE

Recall the lifted version of the MLE (6.4). The trace constraint $\text{Tr}(X) = k$ is equivalent to $\sum_i \lambda_i(X) = k$. By positive semideniteness of $X$ all eigenvalues of $X$ are nonnegative and are also singular values. Therefore the operator norm $\|X\|_* = \sum_i \lambda(i)(X) = k$. 

43
Following [CX16], the convexified MLE is obtained by relaxing the trace constraint in (6.4) to an inequality constraint and dropping the rank-one constraint, hence making the feasible set convex.

Define
\[
\hat{X}_{\text{convex}} = \arg \max_X \langle W, X \rangle \\
\text{s.t.} \quad \|X\|_* \leq k, \\
0 \leq X \leq J, \\
\langle X, J \rangle = k^2. \tag{6.11}
\]

**Remark 6.4 (Matrix norm revisited).** We continue the discussion of matrix norms in Section 3.1.3. Recall that the nuclear norm \(\|X\|_*\) is the \(\ell_1\) norm of the singular values of \(X\), which when \(X\) is symmetric, equals the summation of the absolute values of the eigenvalues of \(X\).

More generally the Schatten \(p\)-norm of a matrix \(X\) (denoted as \(\|X\|_{S_p}\)) is defined as the \(\ell_p\) norm of the singular values of \(X\). For instance, we have
\[
\|X\|_{S_2} = \|X\|_F, \quad \text{the Frobenius norm of } X;
\]
\[
\|X\|_{S_\infty} = \sigma_{\max}(X) = \|X\|_{\text{op}}, \quad \text{the operator norm of } X;
\]
\[
\|X\|_{S_1} = \|X\|_*, \quad \text{the nuclear norm of } X.
\]

In general, the dual norm of a norm \(\|\cdot\|\) is defined as
\[
\|\cdot\|_\ast = \max_{y: \|y\| \leq 1} \langle \cdot, y \rangle. \tag{6.12}
\]

The subscript \(\ast\) here stands for the dual norm, not to be confused with the nuclear norm.

From duality between the usual \(\ell_p\) vector norms, it is easy to derive that
\[
(\|\cdot\|_{S_p})_\ast = \|\cdot\|_{S_q},
\]
where \(1/p + 1/q = 1\). In particular, we have for that the nuclear norm is dual with the operator norm. Hence for the nuclear norm of \(X\),
\[
\|X\|_* = \max_{\|Y\|_{\text{op}} \leq 1} \langle X, Y \rangle.
\]

### 6.3.1 SDP formulation

The following proposition allows us to rewrite the program (6.11) as an SDP in the standard form:

**Proposition 6.3.** For a matrix \(X \in \mathbb{R}^{m \times n}\), \(\|X\|_* \leq 1\) if and only if there exists \(W_1 \in \mathbb{R}^{m \times m}\), \(W_n \in \mathbb{R}^{n \times n}\), such that
\[
\text{Tr}(W_1) + \text{Tr}(W_2) \leq 2, \quad \text{and}
\]
\[
\begin{bmatrix}
W_1 & X \\
X^\top & W_2
\end{bmatrix} \succeq 0. \tag{6.13}
\]

**Proof.** ("if" part) the PSD assumption (6.13) implies that
\[
\begin{bmatrix}
u^\top, & -v^\top
\end{bmatrix}
\begin{bmatrix}
W_1 & X \\
X^\top & W_2
\end{bmatrix}
\begin{bmatrix}
u \\
v
\end{bmatrix} \succeq 0 \quad \forall u, v.
\]
Choose \( u = u_i \) to be the \( i \)'th left singular vector of \( X \), and \( v = v_i \) to be the \( i \)'th right singular vector of \( X \). We have
\[
2u_i^\top X v_i \leq u_i^\top W_i u_i + v_i^\top W_i V_i.
\]
Note that the left-hand side is equal to twice the \( i \)'th singular value of \( X \). Take summation of the inequality above over \( 1 \leq i \leq \min\{m, n\} \) to deduce that
\[
2\|X\|_* \leq \langle W_1, \sum u_i u_i^\top \rangle + \langle W_2, \sum v_i v_i^\top \rangle.
\]
Note that \( W_1 \) is positive semidefinite (simply take \( v = 0 \)). Similarly we have \( W_2 \succeq 0 \). Combine with the fact that \( \sum u_i u_i^\top \succeq I \), \( \sum v_i v_i^\top \preceq I \) we have that the right-hand side of the display above is bounded by \( \Tr(W_1) + \Tr(W_2) \leq 2 \).

("only if" part) Suppose \( X = U \Sigma V^\top \) is the singular value decomposition of \( X \). Choose
\[
W_1 = U \Sigma U^\top, \quad W_2 = V \Sigma V^\top.
\]
First check the trace condition:
\[
\Tr(W_1) + \Tr(W_2) = 2\Tr(\Sigma) = 2\|X\|_* \leq 2.
\]
Next check the positive semidefinite condition. For all \( u \in \mathbb{R}^m, v \in \mathbb{R}^n \), we have
\[
[u^\top, -v^\top] \begin{bmatrix} W_1 & X \\ X^\top & W_2 \end{bmatrix} \begin{bmatrix} u \\ -v \end{bmatrix} = [u^\top, -v^\top] \begin{bmatrix} U \Sigma U^\top & U \Sigma V^\top \\ V \Sigma U^\top & V \Sigma V^\top \end{bmatrix} \begin{bmatrix} u \\ -v \end{bmatrix} = (U^\top u - V^\top v)^\top \Sigma (U^\top u - V^\top v) \geq 0
\]
since \( \Sigma \) is a diagonal matrix with nonnegative diagonal entries.

### 6.3.2 Subgradient and norms

Suppose a function \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) is convex and differentiable. Then the gradient \( \nabla f(x_0) \) at each \( x_0 \) defines an affine minorant:
\[
f(x) \geq f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle, \quad \forall x.
\]
When \( f \) is not differentiable, we could work with the sub-\emph{gradients} of \( f \). A sub-gradient of \( f \) at \( x_0 \) is defined as any element of
\[
\partial f(x_0) = \left\{ u \in \mathbb{R}^d : f(x) \geq f(x_0) + \langle u, x - x_0 \rangle, \quad \forall x \right\}.
\]
For example, \( \partial | \cdot |(0) = [-1, 1] \). Note that for smooth \( f \), subgradient and gradient coincide, i.e.,
\[
\partial f(x_0) = \{ \nabla f(x_0) \}.
\]
Next we find the sub-gradient of the nuclear norm \( \| \cdot \|_* \). Being the \( \ell_1 \) norm of the singular values of \( X \), it is not differentiable. In fact, it is easy to show that the subgradients of any norm are precisely those vectors for which the duality (6.12) is tight:

**Proposition 6.1** (Subgradient of norms). Let \( \| \cdot \| \) and \( \| \cdot \|_* \) be a pair of dual norms. Then
\[
\partial \| \cdot \|_*(x_0) = \begin{cases} 
\{ y : \langle x_0, y \rangle = \|x_0\|_*, \|y\| = 1 \} & x_0 \neq 0 \\
\{ y : \|y\| \leq 1 \} & x_0 = 0
\end{cases}
\] (6.14)
Proof. “⊂”: For any $y$ in the RHS, and for any $x$, $\|x\|_* + \langle y, x - x_0 \rangle = \langle y, x \rangle \leq \|y\| \|x\|_* \leq \|x\|_*$. 
“⊄”: Suppose $\|x\|_* \geq \|x_0\|_* + \langle y, x - x_0 \rangle$ for all $x$. Then

\[
\langle y, x_0 \rangle - \|x_0\|_* \geq \sup_{x \in \mathbb{R}^d} \langle y, x \rangle - \|x\|_* \overset{(6.12)}{=} \begin{cases} 0 & \|y\| \leq 1 \\ +\infty & \|y\| > 1 \end{cases}
\]

This implies that, necessarily, $\|y\| \leq 1$ and $\langle y, x_0 \rangle = \|x_0\|_*$. If $x_0 \neq 0$, then this further means $\|y\| = 1$.

Specializing to the nuclear norm, we have the following (we only need ⊂ direction for analyzing convexified MLE next):

**Corollary 6.1.** For any $X \neq 0$, denote its SVD as $X = U\Sigma V^\top$. Then

\[
\partial \| \cdot \|_*(X) = \{UV^\top + P^\perp(Y) : \|Y\|_{op} \leq 1\}
\]

where $P^\perp(Y) = (I - UU^\top)Y(I - VV^\top)$ is the projection onto the orthogonal complement of the linear subspace $T$ of matrices whose column span is orthogonal to that of $X$ (i.e., $\text{span}(U)$) and whose row span is orthogonal to that of $X$ (i.e., $\text{span}(V)$), that is, $T = \{UA^\top + BV^\top : A \in \mathbb{R}^{n \times r}, B \in \mathbb{R}^{n \times r}\}$.

Proof. Exercise.

### 6.3.3 Statistical guarantee: primal proof

**Theorem 6.5.** Assume that $k \geq C_0 \sqrt{n}$ for some constant $C_0$. Whp, the unique solution to (6.11) is given by $X = X^*$.

Proof. We will show that whp, the objective function of any feasible $X \neq X^*$ is inferior, i.e., $\langle X, W \rangle < \langle X^*, W \rangle$. To this end we will show that whp, for any feasible $X$,

\[
\langle X^* - X, W \rangle \gtrsim \|X - X^*\|_1 \tag{6.15}
\]

As before, $\xi = 1_{K^*}$ be the indicator vector of the hidden clique and $u = \frac{1}{\sqrt{k}} \xi$. Then $X^* = \xi \xi^\top$, which is almost the same as $\mathbb{E}W = X^* - \text{diag}(\xi)$. Let $E = uu^\top = \frac{1}{k} X^*$ denote the projection matrix onto $\text{span}(\xi)$. As in Corollary 6.1, define the projection operator $P^\perp(Y) = (I - E)Y(I - E)$ and $P(Y) = Y - P^\perp(Y) = EY + YE - EYE$.

Write

\[
\langle X^* - X, W \rangle = \underbrace{\langle X^* - X, X^* \rangle}_{(a)} + \underbrace{\langle X^* - X, P^\perp(W - X^*) \rangle}_{(b)} + \underbrace{\langle X^* - X, P(W - X^*) \rangle}_{(c)}.
\]

Then

(a): This term dominates:

\[
(a) = \sum_{(i,j) \in K^* \times K^*} (1 - X_{ij}) = \frac{1}{2} \|X - X^*\|_1
\]

where the last step is due to $\langle X - X^*, J \rangle = 0$ (from feasibility) so that $\sum_{(i,j) \in K^* \times K^*} (1 - X_{ij}) = \sum_{(i,j) \notin K^* \times K^*} X_{ij}$.
We find a subgradient of $\| \cdot \|_*$ at $X^*$. Note that $X^*$ is rank-one, so by Corollary 6.1, 
$$
\partial \| \cdot \|_*(X^*) = \{ E + P^\perp(Y) : \|Y\|_{op} \leq 1 \}.
$$
Now,
$$
0 \geq \|X\|_* - \|X^*\|_* \ 	ext{by feasibility}
\geq \langle X - X^*, E \rangle + \langle X - X^*, P^\perp(Y) \rangle.
\geq -\frac{1}{2}\|X - X^*\|_{\ell_1}
$$

Taking $Y = \pm \frac{W - X^*}{\|W - X^*\|_{op}}$, we get
$$
|b| \leq \frac{\|W - X^*\|_{op}}{2k} \|X - X^*\|_{\ell_1}
$$

By duality, we have
$$
|c| \leq \|P(W - X^*)\|_{\ell_{\infty}} \|X - X^*\|_{\ell_1}.
$$
Combining (a), (b) and (c), we get
$$
\langle X^* - X, W \rangle \geq \left( \frac{1}{2} - \frac{\|W - X^*\|_{op}}{2k} - \|P(W - X^*)\|_{\ell_{\infty}} \right) \|X - X^*\|_{\ell_1}.
$$
Here $\|W - X^*\|_{op} \leq \|W - E[W]\|_{op} + \|\text{diag}(\xi)\|_{op} \leq C\sqrt{n} + 1$ for some constant $C$ whp, by Theorem 4.2. Since $k = C_0\sqrt{n}$ for some sufficiently large $C_0$, it suffices to show that $\|P(W - X^*)\|_{\ell_{\infty}} = o(1)$ whp.

Note that $W - X^* = W - E[W] - \text{diag}(\xi)$ and $P(\text{diag}(\xi)) = E\text{diag}(\xi) + \text{diag}(\xi)E - E\text{diag}(\xi)E = E$ so
$$
\|P(W - X^*)\|_{\ell_{\infty}} \leq \|P(W - E[W])\|_{\ell_{\infty}} + \|P(\text{diag}(\xi))\|_{\ell_{\infty}} 1/k
$$

Next, for any $Y$, $P(Y) = EY + YE - EYE$, where $\|EYE\|_{\ell_{\infty}} \leq \|EY\|_{\ell_{\infty}} \|E\|_{\ell_{\infty}} \leq \|EY\|_{\ell_{\infty}}$. Thus
$$
\|P(Y)\|_{\ell_{\infty}} \leq 3\|EY\|_{\ell_{\infty}},
$$
where we used the fact that for any symmetric $Y$ (e.g. $W - E[W]$) whose support is disjoint from that of $E$,
$$
\|EY\|_{\ell_{\infty}} = \|YE\|_{\ell_{\infty}} = \frac{1}{k} \max_{i \notin K^*} \sum_{j \in K^*} Y_{ij}.
$$
Furthermore, for each $i$, $P \left[ |\sum_{j \in K^*} (W_{ij} - EW_{ij})| \geq \sqrt{k}t \right] \leq \exp(-ct^2)$, by Hoeffding’s inequality (Lemma 2.2). Therefore whp, $\|P(W - E[W])\|_{\ell_{\infty}} \leq 3\|E(W - E[W])\|_{\ell_{\infty}} \leq \sqrt{C\log n}/k$, and we are done.

**Exercise**: Give a dual-based proof of Theorem 6.5 by identifying the appropriate dual certificates.
Part II

Planted partition model
7. Detection threshold for SBM

7.1 Planted partition model and overview

In the second part of the course, we will study the problem of community detection in a broad sense. Consider the following abstract planted partition model, where a matrix $A = (A_{ij})_{1 \leq i < j \leq n}$ is observed whose distribution depends on the latent labels $\sigma = (\sigma_1, \ldots, \sigma_n) \in \{\pm\}^n$, such that

$$A_{ij} \sim \begin{cases} P & \sigma_i = \sigma_j \\ Q & \sigma_i \neq \sigma_j \end{cases}$$

Given $A$, the goal is to recover the labels $\sigma$ accurately.

Two prominent special cases are the following:

**Stochastic block model (SBM)** Here $P = \text{Bern}(p)$ and $Q = \text{Bern}(q)$. In this case the set of vertices $[n]$ is partitioned into two communities $V_+ = \{ i : \sigma_i = + \}$ and $V_- = \{ i : \sigma_i = + \}$, and $A$ is the adjacency matrix of a random graph, such that two nodes $i$ and $j$ are connected with probability $p$ if they belong to the same community, and with probability $q$ if otherwise. The case of $p > q$ is referred to as “ assortative” and $p < q$ as “disassortative”.

The community structure is determined by the vector $\sigma$, which, depending on the problem formulation, could either be fixed or random. We will frequently consider special cases:

- iid model: Each $\sigma_i$ is equally likely to be $\pm$ (Rademacher) and independently.
- exact bisection: $|V_+| = |V_-| = n/2$ (when $n$ is even) and the partition is chosen uniformly at random from all bisections.

Typically these two models behave very similarly.

**Spiked Wigner model (Rank-one deformation)** Here $P = N(\sqrt{\frac{\lambda}{n}}, 1)$ and $Q = N(-\sqrt{\frac{\lambda}{n}}, 1)$. In matrix notation,

$$A = \sqrt{\frac{\lambda}{n}} \sigma \sigma^\top + Z$$

where $Z$ is such that $\{Z_{ij} : 1 \leq i < j \leq n\}$ are iid $N(0, 1)$. Therefore $A$ can be viewed as a rank-one perturbation of a Gaussian Wigner matrix.

As opposed to the treatment of the planted clique problem in Part I, we will be focusing on

- Sharp threshold, i.e., finding the exact constant in the fundamental limit (and achieving them with fast algorithms).

- “Sparse” graphs, where the edge density tends to zero (at different speed), unlike the hidden clique model $G(n, \frac{1}{2}, k)$

We will focus on the following three formulations (recovery guarantees):
**Detection**  Here there is a null model. For example,

- For spiked Wigner model, the null hypothesis is $A$ is iid Gaussian. The sharp threshold is given by $\lambda = 1$, in the sense that for any fixed $\epsilon$, it is possible to test the hypotheses with vanishing error probability if $\lambda \geq 1 + \epsilon$, and impossible if $\lambda \leq 1 - \epsilon$.

- For SBM with bisection, we want to test against the null hypothesis of no community structure, that is, an Erdős-Rényi graph $G(n, p + q/2)$ with the same average degree. The most interesting regime is bounded average degree $p = \frac{a}{n}, q = \frac{b}{n}$ for constants $a, b$, and the sharp threshold is given by $(\frac{a-b}{2(a+b)})^2 = 1$.

**Correlated (weak) recovery**  Here and below, there is no null model. The goal is to recover the community structure (labels) better than random guessing. Let $\hat{\sigma} = \hat{\sigma}(A)$ be the estimator. Its overlap with the true labels $\sigma$ is $|\langle \hat{\sigma}, \sigma \rangle|$ and the number of misclassification errors (up to a global sign flip) is expressed as

$$\ell(\sigma, \hat{\sigma}) = \min_{\pm} \| \hat{\sigma} \pm \sigma \|_1 = n - |\langle \hat{\sigma}, \sigma \rangle|.$$  

In the iid setting, random guessing would yield, by CLT, $|\langle \hat{\sigma}, \sigma \rangle| = O_P(\sqrt{n})$ and $\mathbb{E}[|\langle \hat{\sigma}, \sigma \rangle|] = o(n)$. The goal of weak recovery is to achieve a positive correlation, namely

$$\mathbb{E}[|\langle \hat{\sigma}, \sigma \rangle|] = \Omega(n)$$

Although in general detection and correlated recovery are two different problems, for both SBM and spiked Wigner the thresholds coincide. In fact, for certain models one can have a generic reduction between the problems (e.g. spiked Wigner, see Homework).

**(Almost) exact recovery**  Almost exact recovery means achieving a vanishing misclassification rate: $\mathbb{E}\ell(\sigma, \hat{\sigma}) = o(n)$. Typically the sharp threshold is expressed in terms of Hellinger distance as $H^2(P, Q) \gg \frac{1}{n}$.

Exact recovery means $\ell(\sigma, \hat{\sigma}) = 0$ with probability tending to 1. Typically the sharp threshold is given by $H^2(P, Q) = \frac{(2+\epsilon) \log n}{n}$.

A more statistical flavored question is to characterize the optimal (in the sense of minimax) misclassification rate $\frac{1}{n} \ell(\sigma, \hat{\sigma})$, which typically behaves as $\exp(-H^2(P, Q)/2)$.

### 7.2 Detection threshold for SBM

We want to test the hypothesis

$$H_0 : G \sim G(n, \frac{p + q}{2}) \quad \text{vs.} \quad H_1 : G \sim SBM(n, p, q).$$

Under the SBM model, we assume the the labels $\sigma = (\sigma_1, \ldots, \sigma_n)$ are either iid $\text{Rad}(\frac{1}{2})$, or drawn uniformly at random from all bisections. The detection problem is non-trivial in the regime of bounded average degree:

$$p = \frac{a}{n}, \quad q = \frac{b}{n}, \quad (7.2)$$

where $a, b$ are constants.
Theorem 7.1. If \( \frac{(a-b)^2}{2(a+b)} > 1 \), detection is possible, in the sense of total variation that
\[
\text{TV}(\text{Law}(G|H_0), \text{Law}(G|H_1)) \to 1
\] (7.3)

If \( \frac{(a-b)^2}{2(a+b)} \leq 1 \), detection is impossible, in the sense that
\[
\text{TV}(\text{Law}(G|H_0), \text{Law}(G|H_1)) \leq 1 - \Omega(1).
\] (7.4)

We start with the impossibility part. For non-detection it is enough to show
\[
\chi^2(\text{Law}(G|H_0)||\text{Law}(G|H_1)) = O(1).
\] (7.5)

Remark 7.1 (Contiguity). Recall the notion of contiguity (of two sequences of probability measures \((P_n)\) and \((Q_n)\)). We say \((P_n)\) is contiguous to \((Q_n)\) if for any sequence of events \(E_n, Q_n(E_n) \to 0 \implies P_n(E_n) \to 0\). Contiguity implies non-detection, because for any sequence of tests
\[
Q_n(\text{failure}) \to 0 \implies P_n(\text{success}) \to 0
\]
which is bad news.

A sufficient condition of contiguity is bounded second moment of likelihood, i.e., \(\chi^2(P_n||Q_n) = O(1)\). Indeed, by Cauchy-Schwarz,
\[
P_n(E_n) = \mathbb{E}_{Q_n} \left[ \frac{P_n}{Q_n} 1 \{E_n\} \right] \leq \sqrt{\mathbb{E}_{Q_n} \left[ \left( \frac{P_n}{Q_n} \right)^2 \right]} Q_n(E_n) \to 0
\]

The following lemma is very useful for computing \(\chi^2(\text{mixture distribution}||\text{simple distribution})\). The introduction of two iid copies of randomness is typical in second moment calculation (cf. Section 1.2.2).

Lemma 7.1 (Second moment trick). Suppose we have a parametric family of distributions \(\{P_\theta : \theta \in \Theta\}\). Given a prior on the parameter space \(\Theta\), define the mixture distribution:
\[
P_\pi \triangleq \int P_\theta \pi(d\theta).
\]
Then we have \(\chi^2(P_\pi||Q) = \mathbb{E}G(\theta, \bar{\theta}) - 1\), where \(\theta, \bar{\theta} \sim \pi\) and \(G(\theta, \bar{\theta})\) is defined by
\[
G(\theta, \bar{\theta}) \triangleq \int \frac{P_\theta P_{\bar{\theta}}}{Q}.
\]

Proof. The proof is just by Fubini:
\[
\int \frac{P_\pi^2}{Q} = \int \frac{(\int P_\theta(x)\pi(d\theta)) (\int P_{\bar{\theta}}(x)\pi(d\bar{\theta}))}{Q(x)} \mu(dx)
= \int \pi(d\theta) \pi(d\bar{\theta}) \left( \frac{P_\theta(x)P_{\bar{\theta}}(x)}{Q(x)} \mu(dx) \right)_G(\theta, \bar{\theta})
\]

\[\square\]
Example 7.1 (Gaussian). Consider $P_\theta = N(\theta, I_d)$ and $Q = N(0, I_d)$, and let $\pi$ be some distribution on $\mathbb{R}^d$. Then $\chi^2(P_\pi \| Q) = \mathbb{E}[(\theta, \tilde{\theta})] - 1$, where $\theta, \tilde{\theta} \overset{i.i.d.}{\sim} \pi$.

The calculation for SBM can be carried out in a very general setting. Consider $P$ and $Q$ in place of Bern$(p)$ and Bern$(q)$. For each label $\sigma \in \{\pm\}^n$, the distribution of the adjacency matrix is

$$P_\sigma = \text{Law}(A | \sigma) = \prod_{i<j} \left( P_{1\{\sigma_i = \sigma_j\}} + Q_{1\{\sigma_i \neq \sigma_j\}} \right) = \prod_{i<j} \left( \frac{P + Q}{2} + \frac{P - Q}{2} \sigma_i \sigma_j \right)$$

and the null distribution is $P_0 = \prod_{i<j} \left( P + Q \right)$. Fix two assignment $\sigma, \hat{\sigma} \in \{\pm 1\}^n$. Then

$$G(\sigma, \hat{\sigma}) = \int \frac{P_\sigma P_{\hat{\sigma}}}{P_0}$$

$$= \prod_{i<j} \left[ \int \frac{P + Q}{2} + \frac{P - Q}{2} \sigma_i \sigma_j \right] \left[ \int \frac{P + Q}{2} + \frac{P - Q}{2} \hat{\sigma}_i \hat{\sigma}_j \right]$$

$$= \prod_{i<j} \left[ 1 + \rho \sigma_i \sigma_j \hat{\sigma}_i \hat{\sigma}_j \right]$$

$$\leq \exp \left( \rho \sum_{i<j} \sigma_i \sigma_j \hat{\sigma}_i \hat{\sigma}_j \right) \leq \exp \left( \frac{\rho}{2} \langle \sigma, \hat{\sigma} \rangle^2 \right)$$

Thus, by Lemma 7.1, we have

$$\chi^2(P_1 \| P_0) + 1 = \mathbb{E}_{\sigma, \hat{\sigma}} \left[ \exp \left( \frac{\rho}{2} \langle \sigma, \hat{\sigma} \rangle^2 \right) \right]$$

where $\hat{\sigma}$ is an iid copy of $\sigma$.

For SBM$(n, p, q)$, under the scaling $\left(7.2\right)$, we have

$$\rho = \frac{\tau + o(1)}{n}, \quad \tau \triangleq \frac{(a - b)^2}{2(a + b)}.$$  

Next we consider two situations:

**Random labels:** $\sigma, \hat{\sigma} \overset{iid}{\sim} \{\pm 1\}^n$. By CLT, $\frac{1}{\sqrt{n}} \langle \sigma, \hat{\sigma} \rangle = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sigma_i \hat{\sigma}_i \overset{D}{\rightarrow} Z \sim N(0,1)$. Assuming convergence of MGF (see Lemma 7.2 next), we have

$$\chi^2(P_1 \| P_0) + 1 = \mathbb{E} \exp \left( \frac{\tau + o(1)}{2n} \langle \sigma, \hat{\sigma} \rangle^2 \right)$$

$$\rightarrow \mathbb{E} \left( \frac{\tau + o(1)}{2} Z^2 \right)$$

$$= \begin{cases} 
\infty & \text{if } \tau \geq 1 \\
\text{constant} & \text{if } \tau < 1.
\end{cases} \quad (7.7)$$

52
Exact bisection: Let us consider the case where $\sigma, \hat{\sigma}$ are drawn iid and uniformly at random from the set $\{\theta \in \{\pm 1\}^n : \sum \theta_i = 0\}$. For simplicity, write

$$\sigma = 2\xi - 1, \quad \hat{\sigma} = 2\hat{\xi} - 1,$$

Then $\langle \sigma, \hat{\sigma} \rangle = 4\langle \xi, \hat{\xi} \rangle - n$. Both $\xi, \hat{\xi}$ are iid uniform random $\frac{n}{2}$-sparse binary vectors. So

$$\langle \xi, \hat{\xi} \rangle \sim \text{Hypergeometric}(n, \frac{n}{2}, \frac{n}{2}),$$

which means (check!)

$$\frac{\langle \xi, \hat{\xi} \rangle - \frac{n}{4}}{\sqrt{\frac{n}{16}}} \rightarrow Z \sim N(0, 1).$$

Thus the dichotomy (7.7) applies to bisection as well.

To pass from weak convergence to convergence of the MGF, the following lemma is useful:

Lemma 7.2 (Convergence of MGF). Assume that $X_n \xrightarrow{D} X$. Let $M_n(t) = \mathbb{E} \exp(tX_n)$ and $M(t) = \mathbb{E} \exp(tX)$. If there exists some constant $\alpha > 0$ such that

$$\sup_n P(|X_n| > x) \leq \exp(-\alpha x)$$

for all $x > 0$, then $M_n(t) \rightarrow M(t)$ for all $|t| < \alpha$.

Remark 7.2. • The critical case of $\frac{(a-b)^2}{2(a+b)} = 1$ also implies non-detection. Proving this is outside the scope of this section as the $\chi^2$ truly blows up.

• The threshold of the spiked Wigner model (7.1) is given by $\lambda = 1$. This can be proved by the same second moment method (homework).

7.3 Test by counting cycles

Below we describe a test for

$$H_0 : G \sim G(n, \frac{p+q}{2}) \quad \text{vs.} \quad H_1 : G \sim SBM(n, p, q).$$

that achieves the sharp threshold in Theorem 7.1, following [MNS15]. We will consider the labels being iid Rad$(\frac{1}{2})$. The test is based on counting “short” cycles – by short we mean much shorter than the longest cycle, but the length still need to be slowing growing. As there is no generic polynomial-time algorithm for counting $k$-cycles ($C_k$) for growing $k$, in the next section we make it polynomial-time relying on the randomness of the graph.

Consider the number of $k$-cycles (not induced cycles) as the test statistic, denoted by $X_k$. As a warmup, consider the behavior of $X_k$ in $G(n, \frac{d}{n})$. Then by union bound,

$$\mathbb{P}(X_k > 0) \leq \mathbb{E}[X_k] = \binom{n}{k}k! \frac{1}{2^k} \left(\frac{d}{n}\right)^k \leq d^k,$$

where the overcounting factor $2k$ is the number of symmetries (automorphisms) of $C_k$, namely, cyclic shift and flip. Thus there are no cycles of growing length if $d < 1$. Of course, this first-moment

\footnote{Note that the variance of $\text{Hypergeometric}(n, \frac{n}{2}, \frac{n}{2})$ is exactly half of its counterpart $\text{Binom}(\frac{n}{2}, \frac{1}{2})$. Why? Think about sampling with and without replacements.}
calculation does not tell us about existence. Nevertheless it is known that if \( d \geq 1 \), the longest cycle is of length \( \Omega(n) \) [Bol01, Chap. 8].

Now let’s get back to the original problem of testing \( G(n, \frac{a+b}{2n}) \) versus \( SBM(n, \frac{a}{n}, \frac{b}{n}) \). Assume that \( a > b \). Define

\[
    s = \frac{a - b}{2}, \quad d = \frac{a + b}{2}.
\]

The threshold is then given by \( s^2 \geq d \). Since \( d > s \), this implies \( s > 1 \) and \( a > 2 \).

**Intuition:** For \( k \) not too big, \( X_k \) has a Poisson limit under both model with different parameters. To prove the success of the test (based on thresholding \( X_k \)), it suffices to compute its mean and variance. We will show

- **Under \( H_0 \):** \( \mathbb{E}X_k \approx d^k \), \( \text{Var}X_k \leq d^k \)
- **Under \( H_1 \):** \( \mathbb{E}X_k \approx d^k + s^k \), \( \text{Var}X_k \leq d^k \)

Under the condition \( s^2 > d \), we have

\[
    \mathbb{E}_1[X_k] - \mathbb{E}_0[X_k] \gg \sqrt{\text{Var}_0(X_k) + \text{Var}_1(X_k)}
\]

as \( k \) growing, and hence the test \( \mathbb{1}\{X_k \leq d^k + \frac{s^k}{2}\} \) succeeds.

### 7.3.1 First moment calculation

**Under \( H_0 \).** First we note that

\[
    X_k = \frac{1}{2k} \sum_{\substack{v_1,\ldots,v_k; \\
                              \text{all ordered } k\text{-tuple from } V(G)\text{;}}} \mathbb{1}\{v_1 \sim v_2, v_2 \sim v_3, \ldots, v_k \sim v_1\},
\]

which implies

\[
    \mathbb{E}X_k = \frac{1}{2k} \binom{n}{k} k! \prod_{i=1}^{k} \mathbb{P}\{v_i \sim v_{i+1}, v_{i+1} \sim v_{i+2}, \ldots, v_k \sim v_1\} \approx \frac{1 + o(1)}{2k} d^k \tag{7.8}
\]

under \( H_0 \), where the last equality holds provided \( k = o(\sqrt{n}) \) (Why? Think about birthday problem).

**Under \( H_1 \).** We just need to recompute the probability in (7.8), which now depends on the labels of the vertices. Consider the adjacency matrix \( A \). Then given any two vertices \( v_i, v_{i+1} \), we have

\[
    A_{v_i, v_{i+1}} \sim \begin{cases} \text{Bern}(p) & \text{if } \sigma_i = \sigma_{i+1} \\ \text{Bern}(q) & \text{if } \sigma_i \neq \sigma_{i+1}. \end{cases}
\]

Given any \( k \)-tuple \( \{v_1, v_2, \ldots, v_k\} \) of vertices, suppose \( N \) denotes the number of disagreements of adjacent labels, given by

\[
    N = \sum_{i=1}^{k} \mathbb{1}\{\sigma(v_i) \neq \sigma(v_{i+1})\}
\]
with \( k + 1 \) understood as 1 circularly. Write

\[
N = \sum_{i=1}^{k-1} \mathbb{1}_{\{\sigma(v_i) \neq \sigma(v_{i+1})\}} + \mathbb{1}_{\{\sigma(v_k) \neq \sigma(v_1)\}}
\]

Then we have \( T \sim \text{Binom}(k - 1, \frac{1}{2}) \) and

\[
S = \begin{cases} 
0 & \text{if } T \text{ is even} \\
1 & \text{if } T \text{ is odd}
\end{cases}
\]

is a parity bit, so that \( N = S + T \) is always even.

It is clear that conditioned on \( N = m \), the probability of \( v_1, \ldots, v_k \) forming a cycle is \( q^mp^{k-m} \).

Note that

\[
P(N = m) = \begin{cases} 
0 & m \text{ odd} \\
P(\text{Binom}(k - 1, \frac{1}{2}) = m - 1 \text{ or } m) = \binom{k}{m}2^{-k+1} & m \text{ even}
\end{cases}
\]

Thus

\[
P(v_1 \sim v_2 \sim \cdots \sim v_k) = \sum_{m=0}^{k} q^mp^{k-m} \cdot P(N = m)
\]

\[
= \sum_{m=0}^{k} q^mp^{k-m} \binom{k}{m}2^{-k+1}
\]

\[
= \sum_{m=0}^{k} \frac{(-q)^mp^{k-m} + q^mp^{k-m}}{2} \binom{k}{m}2^{-k+1}
\]

\[
= \left( \frac{p + q}{2} \right)^k + \left( \frac{p - q}{2} \right)^k = n^{-k}(s^k + d^k).
\]

Thus, under \( H_1 \),

\[
\mathbb{E}(X_k) = \frac{[n]_k}{2k} \left\{ \left( \frac{p + q}{2} \right)^k + \left( \frac{p - q}{2} \right)^k \right\}
\]

\[
\approx \frac{1 + o(1)}{2k} \left( s^k + d^k \right).
\]

### 7.3.2 Variance analysis

We only consider the variance under the null, as the alternative is similar. Given ordered \( k \)-tuple of vertices \( T = (v_1, \ldots, v_k) \), define \( b_T = \mathbb{1}_{\{v_1 \sim v_2, \ldots, v_k \sim v_1\}} \). Then under \( H_0 \), we have

\[
\text{Var}(X_k) = \frac{1}{4k^2} \sum_{T, T'} \text{Cov}(b_T, b_{T'}) = \frac{1}{4k^2} \left( \sum_{T} \text{Var}(b_T) + \sum_{T \neq T' \cap T' \neq \emptyset} \text{Cov}(b_T, b_{T'}) \right).
\]

Consider two distinct \( k \)-cycles \( T \) and \( T' \) that are overlapping. Let

\[
\ell = \text{number of common edges}, \quad v = \text{number of common vertices}.
\]

Note that
• \( \text{Cov}(b_T, b_{T'}) \leq \mathbb{E}[b_T b_{T'}] = p^{2k-\ell} \)

• Crucially, \( v \geq \ell + 1 \).

This is because the intersection of two cycles is a forest (each connected component is a path), so that \( v = \ell + \text{cc} \).

Combining all this, we get

\[
\sum_{T \neq T'} \text{Cov}(b_T, b_{T'}) \leq \sum_{\ell=1}^{k-1} \frac{n}{2k-\ell-1} \binom{d}{\ell}^{2k-\ell} \leq \frac{1}{n} k^{k+1} d^{2k} = o(1), \quad \text{provided that } k = o(\log n/ \log \log n).
\]

So we get

\[
\text{Var}(X_k) = \frac{1}{4k^2} d^k + o(1).
\]

### 7.4 Approximately counting cycles in polynomial time

A caveat: The naive way of counting (exhaustive search) \( k \)-cycles takes \( n^k \) time, which is not polynomial in \( n \) if \( k \to \infty \). From the previous analysis, we see that we need to count \( k \)-cycles with slowly growing \( k \).

Fix: The trick is to use the sparsity of the random graph and approximately count the number of \( k \)-cycles.

**Definition 7.1** (\( \ell \)-tangle free). An \( \ell \)-tangle is a connected subgraph of diameter at most \( 2\ell \) that contains at least two cycles.

A graph \( G \) is called \( \ell \)-tangle free if no subgraph of \( G \) is an \( \ell \)-tangle. In other words, for all \( v \in V(G) \), its \( \ell \)-hop neighborhood \( N_{\ell}(v) \) contains at most one cycle.

**Lemma 7.3.** If \( G \sim G(n, \frac{d}{n}) \) and \( d \) is a constant, then \( G \) is \( \ell \)-tangle free if \( \ell = o(\log n) \) (In general \( \ell \log d = c \log n \) for small constant \( c \) suffices).

**Proof.** Suppose \( G \) contains an \( \ell \)-tangle. Then \( G \) must contain a subgraph of the following form

![Diagram of a cycle graph](image)

with \( m \) edges and \( v \) vertices, such that \( m \leq 4\ell \) and \( m \geq v + 1 \). Then by union bound, such a graph exists with probability

\[
\leq n^v \left( \frac{d}{n} \right)^m \leq \frac{d^{O(\ell)}}{n} \to 0,
\]

when \( \ell \log d \ll \log n \). \( \square \)
Next we discuss the connection between counting and linear algebra. Let’s start with triangles \( (k = 3) \):

**Example 7.2** (Counting triangles). Suppose that \( A \) is the adjacency matrix of \( G \). Given any vertex \( v \) in \( G \),

\[
(A^3)_{vv} = \sum_{a,b} A_{va} A_{ab} A_{bv}
\]

is in fact twice the number of triangles incident to \( v \). Therefore, \( \text{Tr}(A^3) = 6 \times \text{the number of triangles in } G \).

To count find \( k \)-cycles one can consider computing \( \text{Tr}(A^k) \), which can be done in the time of eigenvalue decomposition. But

\[
\text{Tr}(A^k) = \text{number of closed walks of length } k \gg \text{number of } k \text{-cycles}.
\]

The strategy next is use the tangle-free structure and count the number of *non-backtracking* (NB) paths.

**Definition 7.2** (Non-backtracking walk). We say

- \((v_1, v_2, \ldots, v_k)\) is a NB walk if \( v_t \sim v_{t+1} \) and \( v_t \neq v_{t-2} \) for all \( t \).
- \((v_1, v_2, \ldots, v_k)\) is a NB cycle if \( v_t \sim v_{t+1} \) and \( v_t \neq v_{t-2} \) for all \( t \) and \( v_1 = v_k \).

For example,

![Figure 7.1: Examples of backtracking and non-backtracking.](image)

**Consequences**: Conditioned on \( G \) being \( 2k \)-tangle free, any NB cycle of \( k \) steps is either a \( k \)-cycle, or an \( m \) cycle traversed for \( \frac{k}{m} \) times. Otherwise, we have a \( 2k \)-tangle such as two short cycles sharing a vertex (see Fig. 7.1 above). This reduces the problem to counting the number of NB cycles of length \( m \), for all \( m = 1, \ldots, 2k \).

It is easy to count NB walk of length \( k \) recursively: Let \( N^m_{uv} \) be the number of NB walks from \( u \) to \( v \) of length \( k \). Then our goal is expressed as

\[
\sum_{v \in V(G)} N^m_{vv}
\]

So it suffices to compute \( N^m_{uv} \) for all pairs \( u, v \).
It turns out $N_{uv}^m$ is given by the following three-term recursion:

$$N_{uv}^{m+1} = \sum_{w \sim v} N_{uw}^m - (d_v - 1)N_{uv}^{m-1}.$$  \hspace{1cm} (7.9)

In matrix notation: let $N^{(m)} = (N_{uv}^m)$ and $D = \text{diag}(d_v)$. Then we have\footnote{In the special case of $d$-regular graphs, (7.10) becomes $N^{(m+1)} = N^{(m)} \cdot A - (d - 1)N^{(m-1)}$. This means $N^{(m)}$ is a polynomial of $A$, in fact, the Chebyshev polynomial, which satisfies the same three-term recurrence. See [ABLS07] for more.}

$$\begin{cases} N^{(m+1)} = N^{(m)} \cdot A - N^{(m-1)}(D - I), \\
N^{(1)} = A, \quad N^{(2)} = A^2 - D \end{cases} \hspace{1cm} (7.10)$$

which means we can compute all $N_{uv}^m$'s using matrix multiplication.

Finally, to justify (7.9), simply notice that the first term on the RHS counts all NB walks of length $m$ from $u$ to a neighbor $w$ of $v$, which, followed by another step from $w$ to $v$, constitute a walk of length $m + 1$ from $u$ to $v$. But, it can be backtracking. So we need to subtract those out, and that precisely the second term: fix any NB walk from $u$ to $v$ of length $m - 1$, say, $u, \ldots, v', v$, where $v' \in N(v)$. Append this walk by $w \in N(v) \setminus \{v'\}$ constitutes a NB walk from $u$ to $w$ in $m$ steps.
§ 8. Correlated recovery

Recall model

\[ G \sim SBM(n, p, q) \]
\[ \sigma = (\sigma_1, \ldots, \sigma_n) \in \{\pm 1\}^n \]
\[ \mathbb{P}[i \sim j] = \begin{cases} p & \text{if } \sigma_i = \sigma_j \\ q & \text{if } \sigma_i \neq \sigma_j. \end{cases} \]

**Goal:** As described in Section 7.1, an estimator \( \hat{\sigma} = \hat{\sigma}(G) \) achieves correlated recovery if the overlap is strictly better than random guessing, that is,

|\langle \hat{\sigma}, \sigma \rangle| \geq \Omega(n) \text{ as } n \to \infty \iff \min_{\pm} \|\hat{\sigma} \pm \sigma\|_1 \leq \left(\frac{1}{2} - \Omega(1)\right) \cdot n.

### 8.1 Impossibility

We start with an information theoretic characterization of correlated recovery:

**Theorem 8.1** (Mutual information characterization). Correlated recovery is possible \( \iff I(\sigma_1, \sigma_2; G) = \Omega(1) \text{ as } n \to \infty. \)

**Remark 8.1** (Mutual information and probability of error). Note that for all \( x_1, x_2 \in \{\pm\}, \)

\[ \text{Law}(G|\sigma_1 = x_1, \sigma_2 = x_2) = \text{Law}(G|\sigma_1 = -x_1, \sigma_2 = -x_2) \]

This means the product \( \sigma_1 \sigma_2 \) is a sufficient statistic of the pair \( (\sigma_1, \sigma_2) \) for \( G \) and hence

\[ I(\sigma_1, \sigma_2; G) = I(\sigma_1 \sigma_2; G) \]

The condition \( I(\sigma_1 \sigma_2; G) = \Omega(1) \) means that \( G \) offers some nontrivial information so that one can decide whether a (or any) pair of vertices have the same label better than chance. This can be quantified as follows.

Aside: mutual information vs probability of error. Suppose we have two random variables \( X \sim \text{Rad}(\frac{1}{2}) \) and \( Y \). Then

\[ \min_{\hat{X}(\cdot)} \mathbb{P}(X \neq \hat{X}(Y)) = \frac{1}{2} \left[ 1 - \text{TV}(P_+, P_-) \right]. \tag{8.1} \]

where

\[ P_+ \triangleq \mathcal{L}(Y|X = +) = \]
\[ P_- \triangleq \mathcal{L}(Y|X = -) \]
So no better than random guess ⇔ \(\text{TV}(P_+, P_-) = o(1)\). We claim this is equivalent to \(I(X; Y) \to 0\). Indeed,

\[
I(X; Y) = \mathbb{E}_X \left[ D(P_{Y|X} \| P_Y) \right] \\
= \frac{1}{2} \left[ D(P_+ \| \bar{P}) + D(P_- \| \bar{P}) \right] \\
\geq \text{TV}^2(P_+, \bar{P}) + \text{TV}^2(P_-, \bar{P}) \\
= \frac{1}{2} \text{TV}^2(P_+, P_-).
\]

On the other hand, from the inequality \(D \leq \chi^2\) we get

\[
I(X; Y) \leq \frac{1}{2} \left[ \chi^2(P_+ \| \bar{P}) + \chi^2(P_- \| \bar{P}) \right] \\
= \frac{1}{2} \left[ \int \frac{(P_+ - \bar{P})^2}{\bar{P}} + \int \frac{(P_- - \bar{P})^2}{2} \right] \\
= \int \frac{(P_- - P_+)^2}{2(P_+ + P_-)} \leq \frac{1}{2} \int |P_+ - P_-| = \text{TV}(P_+, P_-).
\]

\[\Rightarrow\]

Remark 8.2. Mutual information characterization in Theorem 8.1 holds under much more general conditions, e.g., \(k\)-community SBM. See [WX18, Appendix A].

Proof of Theorem 8.1.

(\(\Rightarrow\)) Suppose that \(I(\sigma_1, \sigma_2; G) \geq \epsilon\). Then by symmetry \(I(\sigma_i, \sigma_j; G) \geq \epsilon\) for all \(i \neq j\). Therefore, by Remark 8.1 and (8.1), for all \(i \neq j\), \(\exists \hat{T}_{ij} = \hat{T}_{ij}; G\), such that

\[
\mathbb{P}\{\hat{T}_{ij} = \sigma_1 \sigma_j\} \geq \frac{1}{2} + \delta.
\]

for some \(\delta = \delta(\epsilon)\). Then we can define an estimator of the labels \(\hat{\sigma} = (\hat{\sigma}_1, \ldots, \hat{\sigma}_n)\) by

\[
\hat{\sigma}_1 = +, \hat{\sigma}_i = \hat{T}_{1i}; \ i = 2, \ldots, n.
\]

Then the expected number of correctly classified nodes is

\[
\max \sum_{i \in [n]} \mathbb{P}[\sigma_i = \pm \hat{\sigma}_i] = \sum_{i \in [n]} \mathbb{P}[T_{1i} = \hat{T}_{1i}] \geq (1/2 + \delta)n.
\]

(\(\Leftarrow\)) Suppose \(I(\sigma_i, \sigma_j; G) = o(1)\). Then \(\forall \hat{T}_{ij}, \mathbb{P}[\hat{T}_{ij} = \sigma_i \sigma_j] = \frac{1}{2} + o(1)\). This means given \(\hat{\sigma} = (\hat{\sigma}_1, \ldots, \hat{\sigma}_n)\), we have

\[
2n^2 - \mathbb{E}||\langle \sigma, \hat{\sigma} \rangle||^2 = \mathbb{E}||\sigma \sigma^T - \hat{\sigma} \hat{\sigma}^T||^2_F \\
= 4 \cdot \sum_{i \neq j} \mathbb{P}(\sigma_i \sigma_j \neq \hat{\sigma}_i \hat{\sigma}_j) \\
= 2n^2 - o(n^2),
\]

which means \(\mathbb{E}||\langle \hat{\sigma}, \sigma \rangle||^2 = o(n^2)\) , or \(||\langle \sigma, \hat{\sigma} \rangle|| = o_P(n)\).
Next we show that
\[ \tau = \frac{(a - b)^2}{2(a + b)} < 1 \implies I(\sigma_1, \sigma_2; G) = o(1) \implies \text{Correlation recovery impossible}. \]

First note the following variational representation of total variation:
\[ \text{TV}(P_+, P_-) = \frac{1}{2} \inf_Q \int \frac{(P_+ - P_-)^2}{Q}. \quad (8.2) \]

**Proof.** By C-S, \[ \int \frac{(P_+ - P_-)^2}{Q} = \int \frac{(P_+ - P_-)^2}{\sqrt{2}} \geq \left( \int |P_+ - P_-| \right)^2 = 4 \text{TV}^2, \] with equality if \[ Q = |P_+ - P_-|/\int |P_+ - P_-|. \]

To apply this variational representation, take \( Q = \text{Law of } G(n, \frac{d}{n}) \). To show \( I(\sigma_1, \sigma_2; G) = o(1) \), it suffices to show \( \int \frac{(P_+ - P_-)^2}{Q} = o(1) \). This is a second-moment calculation similar to what we did in Lecture 7 for detection. The difference is that here there is no null model. Write
\[ \int \frac{(P_+ - P_-)^2}{Q} = \int \frac{P_+^2}{Q} + \int \frac{P_-^2}{Q} - 2 \int \frac{P_+ P_-}{Q}. \]

Next we show that \( \int \frac{P_+ P_-}{Q} = \text{constant} + o(1) \), \( z, \tilde{z} \in \{\pm 1\} \). Consider the case of iid labels. By the same argument in Section 7.2, we have
\[ \int \frac{P_+ P_-}{Q} F_{\text{ubini}} = \mathbb{E} \left[ \exp\left( \frac{\tau + o(1)}{n} \sum_{i \neq j} \sigma_i \tilde{\sigma}_i \sigma_j \tilde{\sigma}_j \right) \sigma_1 \sigma_2 = z, \tilde{\sigma}_1 \tilde{\sigma}_2 = \tilde{z} \right] \]
\[ = (1 + o(1)) \mathbb{E} \left[ \exp\left( \frac{\tau + o(1)}{2} \frac{1}{N(0,1)^2} \sigma \tilde{\sigma} \right) \sigma_1 \sigma_2 = z, \tilde{\sigma}_1 \tilde{\sigma}_2 = \tilde{z} \right] \]
\[ \to \mathbb{E} \exp\left( \frac{\tau + o(1)}{2} N(0,1)^2 \right) = C(\tau). \]

Here the justification of the CLT steps is almost the same as before: \( \frac{1}{\sqrt{n}} \langle \sigma, \tilde{\sigma} \rangle = \frac{1}{\sqrt{n}} \sum_{j=3}^n \sigma_j \tilde{\sigma}_j + \frac{1}{\sqrt{n}} (\sigma_1 \tilde{\sigma}_1 + \sigma_2 \tilde{\sigma}_2) \), where the first term is asymptotically \( N(0,1) \) and independent of \( \sigma_1, \sigma_2, \tilde{\sigma}_1, \tilde{\sigma}_2 \), and the last term is negligible.

More generally,
- For exact bisection the same statement holds true, except one should be more careful with the conditioning.
- For the spiked Wigner model (7.1), the same calculation shows that \( \lambda < 1 \implies \text{correlated recovery is impossible}. \)
- In fact, for the spiked Wigner model, one can directly prove (by a sample splitting reduction) that impossibility of detection \( \implies \text{impossibility of correlated recovery} \) (Homework).

### 8.2 Correlated recovery via spectral methods

Next we explain how to achieve the sharp threshold of correlated recovery via suitable versions of spectral methods. We only provide the main ideas and some proof sketch.
Spiked Wigner model: Let’s rewrite (7.1) as follows:

\[ W = \frac{\mu}{n} \sigma \sigma^\top + Z. \]

where the entries of \( Z \) is \( N(0, \frac{1}{n}) \), so that its eigenvalues are between \([-2, 2]\) with high probability.

Consider the following spectral method for estimation \( \sigma \): take the top eigenvector \( \hat{u} = u_1(W) \) of the matrix \( W \) corresponding to the largest eigenvalue \( \lambda_1 \), and report sign(\( u_1 \)) as the estimate \( \hat{\sigma} \). Let \( u = \frac{1}{\sqrt{n}} \sigma \). This method succeeds in correlated recovery if and only (why?) if \( |\langle u, \hat{u} \rangle| \) is bounded away from 0.

The well-known BBP phase transition [BBAP05] states that

\[ \lambda_1(W) \rightarrow \begin{cases} \mu + \frac{1}{\mu} & \text{if } \mu > 1 \\ \frac{2}{\mu} & \text{if } \mu \leq 1, \end{cases} \]

and correspondingly, \( \hat{u} \) is correlated with \( u \) if and only if \( \lambda_1(W) \) escapes the bulk of the spectrum, namely,

\[ |\langle u, \hat{u} \rangle| \rightarrow \begin{cases} 1 - \frac{1}{\mu^2} & \text{if } \mu > 1 \\ 0 & \text{if } \mu \leq 1, \end{cases} \]

SBM\( (n, p, q) \) model: Suppose that the adjacency matrix of the graph \( G \) is given by \( A \). Mimicking the above Gaussian result, the ”wishful thinking” on our part is to view

\[ A = \mathbb{E}A + A - \mathbb{E}A \]

where

\[ \mathbb{E}A = \begin{pmatrix} p & q \\ q & p \end{pmatrix} = \frac{p+q}{2} \begin{pmatrix} 1 & + \\ + & - \end{pmatrix} + \frac{p-q}{2} \begin{pmatrix} - & - \\ + & + \end{pmatrix} \]

and \( \text{Var}(A_{ij} - \mathbb{E}A_{ij}) = \frac{d + o(1)}{n} \), with \( d = (a + b)/2 \). The first eigenvector of \( \mathbb{E}A \) is uninformative, and the second is exactly the label. So we can consider taking the signs of the second eigenvector of \( A \). If we pretend the entries of the perturbation \( A - \mathbb{E}A \) are iid \( N(0, d n) \), then making analogy to the Gaussian result shows that the sharp thresholding is given by \( s = \frac{a-b}{\sqrt{d}} > \sqrt{d} \), which is the exactly the sharp threshold we want to show.

However, applying spectral method to \( A \) itself does not work, as sparse graphs are plagued by high degree vertices. Indeed, for \( G(n, \frac{d}{n}) \) with constant \( d \), it is known [KS03]

\[ \lambda_1(A) = \|A\| = \sqrt{d_{\max}(1 + o(1))}, \quad d_{\max} = \Theta \left( \frac{\log n}{\log \log n} \right). \quad (8.3) \]

In fact, not only the top eigenvalue, \( \lambda_i(A) = \lambda_1(1 - o(1)) \) for an unbounded many of \( i \) [KS03, Sec. 4]. Suppose that \( d_i = d_{\max}, e_i \) is the \( i \)-th coordinate vector. Then \( \|A\| \geq \frac{\|A_{e_i}\|}{\|e_i\|} = \sqrt{d_{\max}} \). As the matrix has all non-negative entries, by Perron frobenious theorem we can say that \( \|A\| = \lambda_1(A) \), which concludes the proof.
To see the effect of high-degree vertices, let’s look at power iteration: say \(d_i = d_{\max}\). Then

\[
(A^{2k})_{ii} = \sum_{i_2, \ldots, i_{2k}} A_{ii_2} A_{i_2 i_3} \cdots A_{i_{2k} i}
\]
\[
\geq d_{\max}^k,
\]
where the last inequality follows by restricting to those backtracking paths that goes from \(i\) to one of its neighbors and immediately goes back. Thus

\[
\|A\|^{2k} \geq \|Ae_i\|^2 = e_i^T A^{2k} e_i \geq d_{\max}^k
\]

Thus \(\lambda_1(A) = \|A\| \geq \sqrt{d_{\max}}\), where the first inequality follows from Perron-Frobenius theorem applied to the nonnegative matrix \(A\). The other side can be shown by arguing that most of the contribution in the moment calculation comes from those backtracking paths. Thus the top eigenvalue \(\lambda_1(A)\) is not bounded. In fact, correspondingly, the limiting spectral distribution of the bulk has unbounded support.

The fact that \(d_{\max}\) is unbounded even when the average degree \(d\) is bounded is because of the following: for each \(v\),

\[
d_v \sim \text{Binom}(n, \frac{d}{n}) \approx \text{Poi}(d)
\]

Pretending they are independent, the maximum of \(n\) iid Poisson is given by the \(\frac{1}{n}\)-quantile, namely,

\[
e^{-\frac{d}{n} k} \approx \frac{1}{n}, \text{ that is, } k \approx \frac{\log n}{\log \log n}.
\]

In summary: Adjacency matrix of sparse graphs is plagued by high-degree vertices, and the top eigenvector is localized on those vertices and not informative.

Solutions:

1. Regularize, e.g., remove high-degree vertices then apply spectral methods. However, it is unclear whether this achieves the sharp thresholds of \(s^2 \geq d\). In [CO10] a sufficient condition of \(s^2 \geq d \log d\) is shown.

2. Turn to other matrices, e.g., the non-backtracking matrix, which we briefly explain next. The motivation comes from the above moment calculation (8.4), wherein the pathological behavior is due to backtracking in the neighborhood of the high-degree vertices, so we remove those.

### 8.3 Spectrum of non-backtracking matrices

Given a simple undirected graph \(G = (V, E)\). Denote the set of oriented edges (ordered pairs) by \(\vec{E} = \{(u, v) : \{u, v\} \in E\}\). The non-backtracking matrix \(B \in \{0, 1\}^{\vec{E} \times \vec{E}}\) is defined as follows: for \(e = (e_1, e_2), f = (f_1, f_2) \in \vec{E}\),

\[
B_{ef} = 1_{\{e_2 = f_1\}} 1_{\{e_1 \neq f_2\}}
\]

**Properties of NB matrix:** Let \(n = |V|, m = |E|\).

1. \(B\) is a \(2m \times 2m\) matrix, and can be partitioned into four \(m \times m\) blocks:

\[
B = \begin{matrix}
B_{11} & B_{12} \\
B_{21} & B_{22}
\end{matrix}
\]

\[B_{11} = B_{22}^T, \quad B_{12}, B_{21} \text{ symmetric,}\]

63
2. Row sum: \( \forall e = (u,v), \sum_{e' \in \overrightarrow{E}} B_{ee'} = d_v - 1. \)

3. Singular values of \( B \) are \( \{d_v - 1 : v \in V\} \cup \{1\} \) and thus not informative. (Why? Consier \( BB^\top \)).

4. Spectrum (eigenvalues) of \( B \): \( \det(I - \lambda B) = (1 - \lambda^2)^{m-n} \det(I - \lambda A + (D - I)\lambda^2). \)

5. Ihara-Bass identity [Ter10, p. 89]:

\[
\det(I - \lambda B) = (1 - \lambda^2)^{m-n} \det(I - \lambda A + \lambda^2(D - I)),
\]

where \( D = \text{diag}(d_v) \). This means \( B \) has \( 2(m-n) \) useless eigenvalues that are equal to \( \pm 1 \), and the rest of the \( 2n \) eigenvalues are useful.

6. \( B \) is not symmetric, but satisfies the following symmetry: Given \( e = (e_1, e_2) \), let \( e^{-1} = (e_2, e_1) \) denote its reversal. Then

\[(B^\top)_{ef} = B_{e^{-1}f^{-1}}.\] (8.6)

In matrix notation, let \( P = (\mathbb{1} \{ e = f^{-1}\}) \) denote the involution that maps a vector \((x_e : e \in \overrightarrow{E})\) to \((x_{e^{-1}} : e \in \overrightarrow{E})\) such that \( P^\top = P \) and \( P^2 = I \). Then

\[B^\top = PBP\]

(in other words, \( BP \) is a symmetric) and consequently \( B^k = PB^kP \).

For sparse random graphs, the spectrum of the NB matrix looks like the following for \( G(n, \frac{d}{n}) \) and \( SBM(n, \frac{a}{n}, \frac{b}{n}) \): [BLM18]

In addition, the following result gives a spectral method based on \( B \) that achieves the optimal threshold:

**Theorem 8.2** ([BLM18]). Let \( s = \frac{a-b}{2} \), \( d = \frac{a+b}{2} \). Let \( u_2 = u_2(B) \) be the second largest eigenvector of \( B \). Define

\[\hat{\sigma}_u = \text{sign} \left( \sum_{e : e_1 = u} (u_2)_e \right).\]

Then \( \hat{\sigma} \) achieves correlated recovery if \( s^2 > d \).

Proving this result is outside the scope here. We explain some intuitions:
Why is $B$ not hindered by high-degree vertices? This applies to both Erdős-Rényi and SBM. Here we consider the former. In the previous section, we see for $G(n, \frac{d}{n})$, the outlier eigenvalues of $A$ exist due to high-degree vertices. This no longer occurs for $B$. To explain some intuition, we apply the trace method to $B^k(B^T)^k$ for some large $k$. Claim that for each oriented edge $e$,

$$(B^k(B^T)^k)_{ee} = \# \text{NB walks starting with } e \text{ in } k \text{ steps then reversing the last step and returning to } e \text{ in } k \text{ steps such as}$$

Indeed, using the symmetry property,

$$(B^k(B^T)^k)_{ee} = \sum_{e_2 \ldots e_{2k}} B_{e_1e_2}B_{e_2e_3} \ldots B_{e_ke_{k+1}}B_{e_{k+1}e_{k+2}} \ldots B_{e_{2k}e_{2k+1}}$$

To simplify the counting in (8.7), crucially, recall the locally tree-like structure of sparse graphs: with high probability, for each vertex $u$, its $k$-hop neighborhood $N_k(u)$ is a tree, provided that $k$ is not too big, e.g. $k = o(\log n)$. If $N_k(v)$ is a tree, then for each summand in (8.7), the path must reverse itself (otherwise there will be a cycle). Thus, on the event that locally tree-like structure holds, we have

$$(B^k(B^T)^k)_{ee} = k\text{th generation descendents of } u \approx d^k,$$

even if the degree of $u$ is as large as $\frac{\log n}{\log \log n}!$ To justify the last step,

- For $G(n, \frac{d}{n})$, the local neighborhood behaves as (can be coupled to) a Galton Watson tree with offspring distribution $Poi(d)$.
- For $SBM(n, \frac{a}{n}, \frac{b}{n})$, the local neighborhood behaves as a two-type Galton Watson tree, where the total offspring distribution is still $Poi(d)$, and each $+$ has $Poi(\frac{a}{2})$ children of type $+$ and $Poi(\frac{b}{2})$ children of type $-$, and vice versa. This can be encoded into the following matrix:

$$M = \begin{bmatrix} \frac{a}{2} & \frac{b}{2} \\ \frac{b}{2} & \frac{a}{2} \end{bmatrix}.$$  

Basic results in branching process states that the total number of $k$th-gen children grows exponentially as $d^k$.

Finally,

$$\sum_{k=1}^{2m} |\lambda_k(B)|^{2k} = \|B^k\|_F^2 = Tr(B^k(B^k)^T) \approx 2md^m$$

which implies that the bulk of the eigenvalues belong to the disk of radius $\sqrt{d}$. 

65
Why is the eigenvector of $B$ informative? This applies to SBM.

Let $\xi \in \mathbb{R}^E$ denote the 2nd eigenvector of $B$. Let $\xi^* \in \mathbb{R}^E$ be defined by $\xi^*_e = \sigma(e_2)$, where $e = (e_1, e_2)$ as usual. For each node $u$, we estimate its label $\sigma(u)$ by $\hat{\sigma}_u = \text{sign}\left(\sum_{e : e_1 = u} \xi_e\right)$.

To gain some insight, let’s proceed with the following wishful thinking: Suppose we can apply power method to study the behavior of the eigenvectors. Since $\xi^*$ is orthogonal to the all-one vector, the 1st eigenvector of $B$ in the population case, let’s hope we can gain some insight about the 2nd eigenvector $\xi$ by studying $B^k \xi^*$ for some large $k$. Then for each node $u$,

$$\sum_{e : e_1 = u} (B^k \xi^*)_e = \sum_{e : e_1 = u} \sum_f (B^k)_{ef} \xi^*_f$$

$$= \sum_{v : \sigma(v) = +} \sum_{e_1 = u} \sum_{f : f_2 = v} (B^k)_{ef} - \sum_{w : \sigma(v) = -} \sum_{e_1 = u} \sum_{f : f_2 = v} (B^k)_{ef}$$

# of $k$th-gen children of type $+ \frac{Z^+_k}{k^2}$ # of $k$th-gen children of type $- \frac{Z^-_k}{k^2}$

where in the last step follows again from the tree structure of $N_u(k)$.

The celebrated result of Kesten-Stigum [KS66] says that the behavior of this number is governed by the matrix $M$ in (8.8), whose eigenvalues are $\lambda_1 = d$ and $\lambda_2 = s$. If $\lambda_2^2 > \lambda_1$, then $(Z^+_k - Z^-_k)/\lambda_k$ converges to some non-degenerate limit $X$, where $X$ is correlated with the label of the root. This means that $\sum_{e : e_1 = u} (B^k \xi^*)_e$ has non-trivial correlation with $u$, and correlated recovery can be achieved by majority vote.

Nevertheless, the above plan is too simplistic as $B$ is asymmetric so straightforward power method does not work. In reality, to apply the power method properly, one needs to study the SVD of $B$ by considering $B^k(B^k)^\top$. But as opposed to the above calculation for $B^k$ which only involves the number of children at the $k$th generation, the same calculation with $B^k(B^k)^\top$ will involve the number of children of all generations up to $k$. For details, see [BLM18, Sec 8].

---

1The rationale of the power method is that $\frac{1}{\|B^k \xi^*\|} B^k \xi^*$ will converge to $\xi$, but since the matrix $B$ is not symmetric, this does not quite work.

2To see this, note that $B^2(B^2)^\top$ will involve paths like $u, u_1, u_2, u_1, u_2$, where $u_i$ is a $i$th-gen children.
§ 9. Semidefinite programming II: general SBM

We consider a two-community SBM model in a very broad sense (cf. Lecture 7). Let \( \sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_n\} \in \{\pm 1\}^n \) be the community labels of nodes. The weighted adjacency matrix is \( A = (A_{ij}) \), where \( \{A_{ij} : 1 \leq i < j \leq n\} \) are independent conditioned on \( \sigma \), such that \( A_{ij} \sim \begin{cases} P & \sigma_i = \sigma_j \\ Q & \text{o.w.} \end{cases} \).

Let \( d_H(x, y) = \sum_i \mathbb{1}\{x_i \neq y_i\} \) denote the Hamming distance. Consider the loss function

\[
\ell(\sigma, \hat{\sigma}) = \min\{d_H(\sigma, \hat{\sigma}), d_H(\sigma, -\hat{\sigma})\},
\]

which is the number of misclassified vertices (up to a global relabeling). We say

1. \( \hat{\sigma} \) achieves almost exact recovery if \( \mathbb{E}\ell(\sigma, \hat{\sigma}) = o(n) \);
2. \( \hat{\sigma} \) achieves exact recovery if \( \ell(\sigma, \hat{\sigma}) = 0 \) w.h.p.

Under mild assumptions on the distributions \( P, Q \), we will see that the requirements for these two types of recovery are \( H^2(P, Q) \gg \frac{1}{n} \) for almost exact recovery, and \( H^2(P, Q) \geq \frac{(2+\epsilon)\log n}{n} \) for exact recovery. In this lecture, we will focus on the exact recovery.

### 9.1 MLE and SDP relaxation

The log-likelihood is

\[
\log p(A | \sigma) = \sum_{i,j} \log p(A_{ij} | \sigma_i, \sigma_j)
\]

\[
= \sum_{i,j} \log p(A_{ij}) \mathbb{1}\{\sigma_i = \sigma_j\} + \log q(A_{ij}) \mathbb{1}\{\sigma_i \neq \sigma_j\}
\]

\[
= \sum_{i,j} \frac{\log p(A_{ij}) + \log q(A_{ij})}{2} + \frac{\log p(A_{ij}) - \log q(A_{ij})}{2} \sigma_i \sigma_j,
\]

where \( p, q \) are the densities of \( P, Q \) w.r.t some dominating measure. And we assume \( P \ll Q \). Define the likelihood ratio (LLR) matrix \( W \) with \( W_{ij} = \log \frac{p}{q}(A_{ij}) \) for \( i \neq j \) and \( W_{ii} = 0 \) for convenience. Then MLE is equivalently formulated as solving

\[
\max_{\sigma \in \{\pm 1\}^n} \langle W, \sigma \sigma^\top \rangle.
\]
This is closely related with the “min-cut” problem. For a weighted graph \( G = (V, E, W) \), a cut is a partition of \( V \) into two disjoint subsets \( S \) and \( S^c \). The value of a cut \((S, S^c)\) is the total edge weights between the two subsets:

\[
\sum_{i \in S, j \in S^c} W_{ij}.
\] (9.2)

We can define \( \sigma_i = 2\mathbb{1}\{i \in S\} - 1 \in \{ \pm 1 \} \) to be the label of cut, then the cut value is

\[
\sum_{i,j} W_{ij} \frac{1 - \sigma_i \sigma_j}{2}.
\]

So we can see that, solving for MLE of SBM (9.1) is equivalent to searching for the min-cut for the complete graph weighted with the LLR matrix \( W \).

Like what we did in the clique problem (Lecture 6), we can reformulate (9.1) as

\[
\begin{align*}
\max & \quad \langle W, X \rangle \\
\text{s.t.} & \quad X \succeq 0 \\
& \quad \text{diag}(X) = I \\
& \quad \text{rank}(X) = 1.
\end{align*}
\] (9.3)

Dropping the rank-one constraint, we arrive at the following SDP relaxation:

\[
\begin{align*}
\max & \quad \langle W, X \rangle \\
\text{s.t.} & \quad X \succeq 0 \\
& \quad \text{diag}(X) = I.
\end{align*}
\] (9.4)

For bisection (two equally sized communities), we have \( \langle \sigma, 1 \rangle = 0 \) and we can further strengthen the above min-cut SDP (9.3) by adding a hard constraint \( \langle X, J \rangle = 0 \) and get that

\[
\begin{align*}
\max & \quad \langle W, X \rangle \\
\text{s.t.} & \quad X \succeq 0 \\
& \quad \text{diag}(X) = I \\
& \quad \langle X, J \rangle = 0.
\end{align*}
\] (9.5)

**Theorem 9.1 (Min-cut SDP).** Assume that

\[
H^2(P, Q) \geq \frac{2(1 + \epsilon) \log n}{n}, \quad \text{for a fixed constant } \epsilon \in (0, 1)
\] (9.5)

\[
\| W - EW \|_{\text{op}} = o(\log n)
\] (9.6)

\[
\text{Var}(W_{ij}) = o\left(\frac{\log^2 n}{n}\right)
\] (9.7)

\[
\mathbb{E}[\{W_{ij} - \mathbb{E}[W_{ij}]\}^4] = o(\log^3 n)
\] (9.8)

\[
D(P\|Q) - D(Q\|P) = O\left(H^2(P, Q) \log n\right).
\] (9.9)

Then w.h.p., the unique solution to the min-cut SDP (9.3) is \( \sigma \sigma^\top \).

**Theorem 9.2 (Min-bisection SDP).** Consider bisection \( \langle \sigma, 1 \rangle = 0 \). Suppose the conditions (9.5) and (9.6) hold. Then w.h.p., the unique solution to the min-bisection SDP (9.4) is \( \sigma \sigma^\top \).
Remark 9.1. The assumptions of the preceding theorem may be interpreted as follows:

- The assumption (9.5) is fundamental since it is the information-theoretical limit that guarantees the MLE to succeed, and it is also necessary under further regularity conditions on the distributions $P, Q$, e.g., for Bernoulli and Gaussians.

- The assumption (9.6) on the spectral deviation of the weight matrix is the main assumption for SDP relaxation to succeed. To appreciate this assumption, let us compute a lower bound to $\|E[W]\|_{\text{op}}$. Specifically, $E[W]$ is a rank-two matrix, with

$$E[W_{ij}] = \begin{cases} D(P\|Q) & \sigma_i = \sigma_j \\ -D(Q\|P) & \sigma_i \neq \sigma_j. \end{cases}$$  \hfill (9.10)

Thus $E[W]\sigma_i = kD(P\|Q) + (n-k)D(Q\|P)$ if $\sigma_i = +1$ and $-kD(Q\|P) - (n-k)D(P\|Q)$ if $\sigma_i = -1$, where $k$ is the number of $+1$'s in $\sigma$. Recall that the KL divergence dominates the Hellinger distance:

$$\min\{D(P\|Q), D(Q\|P)\} \geq 2H^2(P\|Q).$$ \hfill (9.11)

Thus $\|E[W]\|_{\text{op}} \geq n \min\{D(P\|Q), D(Q\|P)\} \stackrel{(9.5)}{=} \Omega(\log n)$. This means $E[W]$ is on the scale of $\log n$ which dominates $(W - E[W])$, thanks to (9.6).

- The assumptions (9.7), (9.8), and (9.9) are for technical reasons. Note that (9.7) is consistent with (9.6), because for Wigner matrix we expect the spectral norm is $\sqrt{n} \cdot \text{Var}$.

Remark 9.2 (Specialization to SBM). For SBM($n,p,q$), set $p = \alpha \frac{\log n}{n}, q = \beta \frac{\log n}{n}$. Then

$$H^2(P,Q) = (\sqrt{p} - \sqrt{q})^2 + (\sqrt{1-p} - \sqrt{1-q})^2 = (\sqrt{\alpha} - \sqrt{\beta})^2 \log n (1 + o(1)).$$

So the condition for Hellinger distance in Theorem 9.1 is $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$.

For the LLR:

$$W_{ij} = \log \frac{p}{q} \mathbb{1}\{A_{ij} = 1\} + \log \frac{1-p}{1-q} \mathbb{1}\{A_{ij} = 0\}$$

$$= \log \frac{p(1-q)}{q(1-p)} A_{ij} + \log \frac{1-p}{1-q}.$$

Thus

$$W = \log \frac{p(1-q)}{q(1-p)} \cdot A + \log \frac{1-p}{1-q} \cdot J \hspace{1cm} (9.12)$$

where $J$ is the all-ones matrix. Thus the SDP (9.3) is equivalent with the following penalized form:

$$\max \langle A, X \rangle + \tau \langle J, X \rangle$$

s.t. $X \succeq 0$

$$\text{diag}(X) = I.$$

\footnote{Related calculations are carried out in [HWX17] for the single-community model with general distributions.}
where \( \tau = \frac{\log \frac{1-q}{p}}{\log \frac{1-q}{q(1-p)}} \). For bisection, the SDP (9.4) is equivalent to

\[
\max \langle A, X \rangle \\
\text{s.t. } X \succeq 0 \\
\text{diag}(X) = I \\
\langle X, J \rangle = 0.
\]

Finally, from (9.12) assuming \( \log \frac{p(1-q)}{q(1-p)} \) is bounded, we have

\[
\|W - EW\|_{op} = o(\log n) \iff \|A - EA\|_{op} = o(\log n).
\]

Indeed, \( p, q = \Omega(\frac{\log n}{n}) \), it can be shown (cf. e.g. [HWX16, Theorem 5]) that whp,

\[
\|A - EA\|_{op} \lesssim \sqrt{\log n}.
\]

This is consistent with the Gaussian heuristic in Lecture 4 that \( \|A - EA\|_{op} \lesssim \sqrt{n \cdot \text{variance}} \).

### 9.2 Proof of Theorem 9.2

In this section, we give the proof of Theorem 9.2 for min-bisection SDP; Theorem 9.1 for min-cut SDP is proved analogously but with a bit more detailed calculations in the next section.

The following deterministic lemma gives a sufficient condition for (9.4) to achieve the exact recovery.

**Lemma 9.1** (Duality). \( X^* = \sigma \sigma^\top \) is the unique maximizer of (9.4) if \( \exists D = \text{diag}(d_i), S \succeq 0, \lambda \in \mathbb{R}, \text{s.t.} \)

\[
S = D - W + \lambda J, \\
S \sigma = 0, \\
\lambda_{n-1}(S) > 0.
\]

**Proof of Lemma 9.1.** The proof follows the usual route of KKT conditions (cf. Section 6.2). Define the following function as the Lagrangian form of (9.4) with specified parameters

\[
L(X, D, S, \lambda) = \langle W, X \rangle + \langle S, X \rangle + \text{Tr}(D) - \langle D, X \rangle - \lambda \langle J, X \rangle.
\]

For any feasible \( X \) in (9.4), \( \text{Tr}(D) - \langle D, X \rangle - \lambda \langle J, X \rangle = 0, \langle S, X \rangle \geq 0. \) So

\[
\langle W, X \rangle \leq L(X, D, S, \lambda) \overset{(9.13)}{=} L(X^*, D, S, \lambda) \overset{(9.14)}{=} \langle W, X^* \rangle.
\]

Thus \( X^* \) is an optimal solution to (9.4).

Finally, we prove the uniqueness. Note that if \( \langle W, X \rangle = \langle W, X^* \rangle \), then \( \langle S, X \rangle = 0. \) But \( \lambda_{n-1}(S) > 0, \) so the column space of \( X \) is spanned by \( \sigma, \) which means \( X = cX^*. \) Finally, \( c = 1 \) since \( \text{diag}(X) = I. \)

With Lemma 9.1, to prove Theorem 9.2, it suffices to construct \( (D, \lambda) \) and verify that the conditions of Lemma 9.1 are satisfied with high probability.
Proof of Theorem 9.2. From (9.13) and (9.14) and \( \langle \sigma, 1 \rangle = 0 \), we know \( D\sigma = W\sigma \), which means
\[
d_i = \sum_j W_{ij} \sigma_i \sigma_j.
\]
It remains to show \( S \) as defined in (9.13) satisfies w.h.p. \( \lambda_{n-1}(S) > 0 \), that is
\[
\inf_{x \perp \sigma, \|x\|_2 = 1} x^\top (D - W + \lambda J)x > 0.
\] (9.15)
Note that by the assumption (9.6), \( |x^\top (W - EW)x| \leq \|W - EW\|_{op} = o(\log n) \). Thus it suffices to show
\[
\inf_{x \perp \sigma, \|x\|_2 = 1} x^\top (D - E[W] + \lambda J)x = \Omega(\log n).
\] (9.16)
Write \( s = D(P\|Q), t = D(Q\|P) \) as shorthand. By (9.10), we have
\[
\mathbb{E}W = s - \frac{t}{2} J + \frac{s + t}{2} \sigma \sigma^\top - s I.
\]
Thus to show (9.16), it suffices to show
\[
\inf_{x \perp \sigma, \|x\|_2 = 1} x^\top \left( D - \frac{s - t}{2} J + \lambda J \right) x = \Omega(\log n).
\] (9.17)
We finish the proof in two steps. First, we choose \( \lambda \geq \frac{s-t}{2} \), so that \( \left(-\frac{s-t}{2} + \lambda\right) x^\top Jx \geq 0 \). Then, we show that under the information-theoretic condition (9.5), \( \min_{i \in [n]} d_i \geq \epsilon(1 + \epsilon) \log n \) with high probability. This is where the condition (9.5) is from.

**Lemma 9.2.** Assume that (9.5) holds. Then whp, \( \min_{i \in [n]} d_i \geq \frac{1}{2} \epsilon n H^2(P,Q) \geq \epsilon(1 + \epsilon) \log n \).

\[\square\]

9.2.1 Proof of Lemma 9.2: Large deviation

This is a good exercise on the (information-theoretic flavored) large deviation analysis. Assume that \( H^2(P,Q) \geq \frac{2(1+\epsilon)\log n}{n} \). By union bound, it suffices to show
\[
P(d_i \leq c \log n) = o(1/n)
\]
for each \( i \in [n] \) and \( c = \frac{e^{nH^2(P,Q)}}{2\log n} \). Let us focus on \( d_i \) for a node \( i \) with \( \sigma_i = +1 \) and simply call it \( d \); the case for a node \( i \) with \( \sigma_i = -1 \) can be proved analogously.

Define \( X \) and \( Y \) be distributed as the law of \( \log \frac{dP}{dQ} \) under \( P \) and \( Q \), respectively. Then
\[
d = \sum_{i=1}^k X_i - \sum_{i=1}^{n-k} Y_i, \quad \mathbb{E}d = kD(P\|Q) + (n-k)D(Q\|P),
\]
where \( X_i \) are iid copies of \( X \) and \( Y_i \) are iid copies of \( Y \). To apply Chernoff bound, denote the log moment generating function (log MGF) of \( X \) and \( Y \) as
\[
\psi_P(\theta) = \log \mathbb{E}[e^{\theta X}] = \log \int P^{1+\theta} Q^{-\theta}, \quad \psi_Q(\theta) = \log \mathbb{E}[e^{\theta Y}] = \log \int P^{\theta} Q^{1-\theta} = \psi_P(\theta - 1). \] (9.18)
For any $\theta > 0$,
\[
\mathbb{P}(d \leq c \log n) = \mathbb{P}
\left( \sum_{i=1}^{n-k} Y_i - \sum_{i=1}^{k} X_i \geq -c \log n \right)
\leq \mathbb{E} \exp \left( \theta \sum_{i=1}^{n-k} Y_i - \theta \sum_{i=1}^{k} X_i + \theta c \log n \right)
= \exp \left( (n-k) \psi_Q(\theta) + k \psi_P(-\theta) + \theta c \log n \right)
= \exp \left( (n-k) \psi_P(\theta - 1) + k \psi_P(-\theta) + \theta c \log n \right).
\]
Choose $\theta = \frac{1}{2}$, note that $\psi_P(-1/2) = \log \int \sqrt{PQ} = \log(1 - \frac{H^2(P,Q)}{2})$, and recall that $c = \frac{\epsilon n H^2(P,Q)}{2 \log n}$.
Then we have as desired
\[
\mathbb{P}[d \leq 0] \leq \exp \left\{ n \log \left( 1 - \frac{H^2(P,Q)}{2} \right) + \frac{\epsilon n H^2(P,Q)}{4} \right\}
\leq \exp \left\{ \frac{(1-\epsilon/2)nH^2(P,Q)}{2} \right\}
\leq n^{-(1-\epsilon/2)(1+\epsilon)} = o(1/n),
\]
where (a) follows from $\log(1-x) \leq -x$ for $x \in [0,1]$ and (b) follows from the assumption (9.5).

**Remark 9.3 (Optimality).** We explain heuristically the sharp threshold expressed in terms of the Hellinger distance:

- **Exact recovery:** Under further assumptions, the condition (9.5) on $H^2(P,Q)$ is necessary for MLE (and hence any method) to achieve exact recovery. Indeed, consider another solution $\hat{\sigma}$ that differs from $\sigma$ only on a single coordinate. Then the potential increment of the likelihood is $\langle W, \hat{\sigma}\hat{\sigma}^\top - \sigma\sigma^\top \rangle_D = -4d_i$ defined above. We have showed that $H^2(P,Q) = \frac{2(1+\epsilon)}{n} \log n$ ensures that $\min d_i > 0$. If the above large deviation analysis is tight, then in the opposite condition $H^2(P,Q) = \frac{2(1-\epsilon)}{n} \log n$, we will have $\min d_i < 0$, meaning MLE will make a mistake.

- **Almost exact recovery:** Consider the oracle situation when $\sigma_2, \ldots, \sigma_n$ are observed, with roughly half $+$ and half $-$, and the only goal is to estimate $\sigma_1$. This is equivalent to testing the hypothesis of $H_0 : \sigma_1 = +$ versus $H_1 : \sigma_1 = -$. In this case, only the first row of $A$ is useful. Hence $\sigma_1$ cannot be tested with vanishing probability of error, if
\[
TV(P^{\otimes \frac{k}{2}} \otimes Q^{\otimes \frac{k}{2}}, Q^{\otimes \frac{k}{2}} \otimes P^{\otimes \frac{k}{2}}) \leq 1 - c \iff H^2(P \otimes Q, Q \otimes P) = O(1/n) \iff H^2(P, Q) = O(1/n).
\]

### 9.3 Proof of Theorem 9.1

The following deterministic lemma gives a sufficient condition for (9.3) to achieve the exact recovery.

---

2 We chose $\theta = 1/2$ independent of $k$. To see why, consider the special case of bisection where $k = n/2$. Due to the convexity of the log MGF $\psi_P$, for any $\theta$, $\psi_P(\theta - 1) + \psi_P(-\theta) \geq 2\psi_P(-1/2)$, with equality if $\theta = 1/2$, which is the optimal choice.
Lemma 9.3 (Duality). \( X^* = \sigma \sigma^T \) is the unique maximizer of (9.3) if \( \exists D = \text{diag}(d_i) \) s.t.

\[
S = D - W, \quad (9.19)
\]

\[
S \sigma = 0, \quad (9.20)
\]

\[
\lambda_{n-1}(S) > 0.
\]

The proof is identical to the proof of Lemma 9.1 with \( \lambda \) setting to be 0. With Lemma 9.3, to prove Theorem 9.1, it suffices to construct \( D \) and verify that the conditions of Lemma 9.3 are satisfied with high probability.

Proof of Theorem 9.1. The proof is almost identical to the proof of Theorem 9.2 with \( \lambda \) setting to be 0. The only difference is in showing

\[
\inf_{x \perp \sigma, \|x\|_2 = 1} x^T \left( D - \frac{s - t}{2} J \right) x = \Omega(\log n).
\]

(9.21)

To evaluate this minimum, let us define a unit vector \( \xi \) such that

\[
\xi_i = \begin{cases} \sqrt{\frac{n-k}{nk}} & \text{if } \sigma_i = +1, \\ \sqrt{\frac{k}{n(n-k)}} & \text{if } \sigma_i = -1, \end{cases}
\]

(9.22)

Then it is straightforward to verify that \( \text{span}(\sigma, 1) = \text{span}(\sigma, \xi) \), consisting of vectors which takes constant values on each of the two communities. For any feasible \( x \) in (9.21), we have \( x = \cos \theta \xi + \sin \theta z \) for some unit vector \( z \in \text{span}(\sigma, 1)^\perp \).

First consider the easy case where the community size is very unbalanced, that is, \( k \) or \( n - k \) is \( o\left(\frac{n}{\log n}\right) \). Recall that \( k = |\{i : \sigma_i = +\}| \) is the size of the community with label +1. Then by Lemma 9.2, whp,

\[
x^T \left( D - \frac{s - t}{2} J \right) x \geq \min d_i - \frac{s - t}{2} (\langle \xi, 1 \rangle)^2 = \min d_i - \frac{s - t}{2} \frac{4k(n-k)}{n} = \Omega(\log n),
\]

where the last equality uses assumption (9.9) and (9.5).

Next, assume that both \( k = \Omega\left(\frac{n}{\log n}\right) \) and \( n - k = \Omega\left(\frac{n}{\log n}\right) \). Expanding the quadratic form, we have

\[
x^T \left( D - \frac{s - t}{2} J \right) x = \cos^2 \theta \cdot (I) + 2 \cos \theta \sin \theta \cdot (II) + \sin^2 \theta \cdot (III),
\]

where

- \((I) = \xi^T \left( D - \frac{s - t}{2} J \right) \xi \geq \xi^T \left( \mathbb{E} D - \frac{s - t}{2} J \right) \xi - \|D - \mathbb{E} D\|_2 = nt - \|D - \mathbb{E} D\|_2.
\]

Here \( \xi^T J \xi = (\langle \xi, 1 \rangle)^2 = \frac{4k(n-k)}{n}, \quad \xi^T \mathbb{E} D \xi = \sum_i \mathbb{E}[d_i] \xi_i^2 = [(n-k)d_+ + kd_-]/n = \frac{s+t}{2} + \frac{t-s}{2n}(n-2k)^2, \) with \( \mathbb{E}[d_i] = \frac{s+t}{2} n + \sigma_i \frac{s-t}{2} (2k-n), \) so that (by some miracle) \( \xi^T \left( \mathbb{E} D - \frac{s-t}{2} J \right) \xi = nt. \)

- \((II) = \xi^T \left( D - \frac{s - t}{2} J \right) z = \xi^T (D - \mathbb{E} D) z \geq -\|D - \mathbb{E} D\|_2,
\]

where we used \( z \perp 1 \) and \( \xi^T \mathbb{E} D z = \langle z, \mathbb{E} D \xi \rangle = 0, \) since the entries of \( \mathbb{E} D \xi \) are constant in each of the two communities.
\[(\text{III}) = z^\top \left( D - \frac{s-t}{2} J \right) z = z^\top Dz \geq \min_i d_i.\]

Overall, we have

$$\inf_{x \perp \sigma, \|x\|_2 = 1} x^\top \left( D - \frac{s-t}{2} J \right) x \geq \min\{nt, \min_i d_i\} - 3\| (D - \mathbb{E} D) \xi \|_2 = \Omega(\log n),$$

where $nt = n D(Q\|P) \geq 2 n H^2(P, Q) = \Omega(\log n)$, $\min_i d_i = \Omega(\log n)$ whp by Lemma 9.2, and the following

**Lemma 9.4.** Assume that $k = \Omega(\frac{n}{\log n})$ and $n - k = \Omega(\frac{n}{\log n})$. Whp, $\| (D - \mathbb{E} D) \xi \|_2 = o(\log n)$.

This completes the proof of Theorem 9.1. \qed
Ranking from comparisons arises in various applications, including recommender systems, social choice and sports tournament. We consider the following setup. Suppose that there are items 1, …, n associated with unknown ranks $\pi^*(1), \ldots, \pi^*(n)$, where $\pi^* : [n] \to [n]$ is a permutation. Observing a set of pairwise comparisons, each of the form $i \succ j$ meaning that “item $i$ beats item $j$”, we aim to recover the ranking $\pi^*$.

### 10.1 Modeling pairwise comparisons

We first give an overview of common models for ranking from pairwise comparisons.

#### 10.1.1 Models for probabilities of outcomes

Each pairwise comparison is a Bernoulli outcome. Let us denote the probability that the item at rank $k$ beats the item at rank $\ell$ by $M_{k,\ell}$ where $M \in \mathbb{R}^{n \times n}$, so that

$$\mathbb{I}\{i \succ j\} \sim \text{Ber}(M_{\pi^*(i),\pi^*(j)})$$

In the sequel, we present several models on the matrix $M$ of probabilities. It is vacuous to compare an item to itself, so we assume without loss of generality that $M_{i,i} = 1/2$ for $i \in [n]$. Moreover, we consider the case that there is one and only one winner in a pairwise comparison, so it always holds that $M_{k,\ell} + M_{\ell,k} = 1$.

**Parametric models** Parametric models assume that for $i \in [n]$, item $i$ is associated with a strength parameter $\theta_i \in \mathbb{R}$, and

$$M_{\pi^*(i),\pi^*(j)} = F(\theta_i - \theta_j)$$

where $F : \mathbb{R} \to (0,1)$ is a known, increasing link function. Two classical examples are the logistic function $F(x) = \frac{1}{1 + e^{-x}}$ and the Gaussian cumulative density function, which correspond to the Bradley-Terry model and the Thurstone model respectively.

**Noisy sorting** The noisy sorting model [BM08] assumes that

$$M_{k,\ell} = \begin{cases} 
1/2 + \lambda & \text{if } k > \ell, \\
1/2 - \lambda & \text{if } k < \ell.
\end{cases} \quad (10.1)$$

This is the model we focus on later, as it is simple yet captures important concepts and tools.
Strong stochastic transitivity  Strong stochastic transitivity (SST) means that for any triplet $(k, \ell, m) \in [n]^3$ such that $k < \ell < m$, we have

$$M_{k,m} \geq M_{k,\ell} \lor M_{\ell,m}.$$  

In matrix terminology, this is saying that $M$ is bivariate isotonic (bi-isotonic) in addition to the constraint $M + M = 11^\top$. More precisely, all the columns of $M$ are nonincreasing while all the rows of $M$ are nondecreasing. Note that any parametric model, as well as the noisy sorting model, satisfies SST.

10.1.2 Sampling models

We consider uniform sampling. Namely, for $m \in [N]$ where $N$ is the sample size, we observe independent outcomes

$$y_m \sim \text{Ber}(M_{\pi^*(i_m), \pi^*(j_m)}),$$  

where the random pairs $(i_m, j_m)$ are sampled uniformly randomly with replacement from all possible pairs $\{ (i, j) \}_{i \neq j}$. Here $y_m = 1$ means that $i_m \succ j_m$ and $y_m = 0$ means that $j_m \succ i_m$. We collect the outcomes of comparisons in a matrix $A \in \mathbb{R}^{n \times n}$ whose entry $A_{i,j}$ is defined to be the number of times item $i$ beats item $j$.

Note that for parametric models, we have for $m \in [N]$,

$$\mathbb{E}[y_m] = F(\theta_{i_m} - \theta_{j_m}) = F(x_m^\top \theta),$$  

where $x_m = e_{i_m} - e_{j_m}$ is the design point. This is simply the setup of generalized linear regression. Particularly, the Bradley-Terry model is essentially logistic regression with this special design.

10.2 Kendall’s tau and minimax rates for noisy sorting

In general, we would like to estimate both $\pi^*$ and $M$, but let us focus on estimating $\pi^*$ under the noisy sorting model (10.1) for the rest of the notes. Full details of the discussion can be found in the paper [MWR18].

Consider the Kendall tau distance, i.e., the number of inversions between permutations, defined as

$$d_{KT}(\pi, \sigma) = \sum_{i,j \in [n]} \mathbb{I}(\pi(i) > \pi(j), \sigma(i) < \sigma(j)).$$

Note that $d_{KT}(\pi, \sigma) \in [0, \binom{n}{2})$ and it is equal to the minimum number of adjacent transpositions required to change from $\pi$ to $\sigma$ (think of bubble sort). A closely related distance is the $\ell_1$-distance, also known as Spearman’s footrule, defined as

$$\|\pi - \sigma\|_1 = \sum_{i=1}^{n} |\pi(i) - \sigma(i)|.$$  

It is well known [DG77] that

$$d_{KT}(\pi, \sigma) \leq \|\pi - \sigma\|_1 \leq 2d_{KT}(\pi, \sigma).$$  

Theorem 10.1. Consider the noisy sorting model (10.1) with $\lambda \in (0, \frac{1}{2} - c]$ where $c$ is a positive constant. Suppose $N$ independent comparisons are given according to (10.2). Then it holds that

$$\min_{\tilde{\pi}} \max_{\pi^*} \mathbb{E}_{\pi^*} [d_{KT}(\tilde{\pi}, \pi^*)] \geq \frac{n^3}{\lambda^2} \wedge n^2.$$  

76
10.2.1 Inversions and metric entropy

Before proving the theorem, we study the metric entropy of the set of permutations \( \mathfrak{S}_n \) with respect to the Kendall tau distance \( d_{KT} \). Let \( \mathcal{B}(\pi, r) = \{ \sigma \in \mathfrak{S}_n : d_{KT}(\pi, \sigma) \leq r \} \).

The inversion table \( b_1, \ldots, b_n \) of a permutation \( \pi \in \mathfrak{S}_n \) is defined by

\[
b_i = \sum_{j=i}^{n} \mathbb{1}(\pi(i) > \pi(j)).
\]

Note that \( b_i \in \{0, 1, \ldots, n-i\} \) and \( d_{KT}(\pi, \text{id}) = \sum_{i=1}^{n} b_i \). On can reconstruct a permutation using its inversion table \( \{b_i\}_{i=1}^{n} \), so the set of inversion tables is bijective to \( \mathfrak{S}_n \). (Try the permutation (3 5 2 4 1) which has inversion table (4 2 0 1 0).

**Lemma 10.1.** For \( 0 \leq k \leq \binom{n}{2} \), we have that

\[
n \log(k/n) - n \leq \log |\mathcal{B}(\text{id}, k)| \leq n \log(1 + k/n) + n.
\]

**Proof.** According to the discussion above, \( |\mathcal{B}(\text{id}, k)| \) is equal to the number of inversion tables \( b_1, \ldots, b_n \) such that \( \sum_{i=1}^{n} b_i \leq k \) where \( b_i \in \{0, 1, \ldots, n-i\} \). On the one hand, if \( b_i \leq [k/n] \) for all \( i \in [n] \), then \( \sum_{i=1}^{n} b_i \leq k \), so a lower bound is given by

\[
|\mathcal{B}(\text{id}, k)| \geq \prod_{i=1}^{n} ([k/n] + 1) \wedge (n-i+1)
\]

\[
\geq \prod_{i=1}^{n-[k/n]} ([k/n] + 1) \prod_{i=n-[k/n]+1}^{n} (n-i+1)
\]

\[
\geq (k/n)^{n-k/n} [k/n]!.
\]

Using Stirling’s approximation, we see that

\[
\log |\mathcal{B}(\text{id}, k)| \geq n \log(k/n) - (k/n) \log(k/n) + [k/n] \log[k/n] - [k/n]
\]

\[
\geq n \log(k/n) - n.
\]

On the other hand, if \( b_i \) is only required to be a nonnegative integer for each \( i \in [n] \), then we can use a standard “stars and bars” counting argument to get an upper bound

\[
|\mathcal{B}(\text{id}, k)| \leq \binom{n+k}{n} \leq e^n (1 + k/n)^n.
\]

Taking the logarithm finishes the proof. \[\square\]

For any \( \varepsilon > 0 \) and \( S \subseteq \mathfrak{S}_n \), let \( N(S, \varepsilon) \) and \( D(S, \varepsilon) \) denote respectively the \( \varepsilon \)-covering number and the \( \varepsilon \)-packing number of \( S \) with respect to \( d_{KT} \).

**Proposition 10.1.** We have that for \( \varepsilon \in (0, r) \),

\[
n \log \left( \frac{r}{n + \varepsilon} \right) - 2n \leq \log N(\mathcal{B}(\pi, r), \varepsilon) \leq \log D(\mathcal{B}(\pi, r), \varepsilon) \leq n \log \left( \frac{2n + 2r}{\varepsilon} \right) + 2n.
\]

For \( n \lesssim \varepsilon < r \leq \binom{n}{2} \), the \( \varepsilon \)-metric entropy of \( \mathcal{B}(\pi, r) \) scales as \( n \log \frac{r}{\varepsilon} \). In other words, \( \mathfrak{S}_n \) equipped with \( d_{KT} \) is a doubling space\(^1\) with doubling dimension \( \Theta(n) \).

---

\(^1\)A metric space \((X, d)\) is called a doubling space with doubling dimension \( \log_2 M \), if \( M \) is the smallest number such that any ball of radius \( r \) in \((X, d)\) can be covered with \( M \) balls of radius \( r/2 \).
10.2.2 Proof of the minimax upper bound

Proof. The relation between the covering and the packing number is standard. We employ a volume argument for the bounds. Let \( \mathcal{P} \) be a \( 2\varepsilon \)-packing of \( \mathcal{B}(\pi, r) \) so that the balls \( \mathcal{B}(\sigma, \varepsilon) \) are disjoint for \( \sigma \in \mathcal{P} \). By the triangle inequality, \( \mathcal{B}(\sigma, \varepsilon) \subseteq \mathcal{B}(\pi, r + \varepsilon) \) for each \( \sigma \in \mathcal{P} \). By the invariance of the Kendall tau distance under composition, Lemma 10.1 yields

\[
\log D(\mathcal{B}(\pi, r), 2\varepsilon) \leq n \log(1 + r/n) + n - n \log(\varepsilon/n) + n = n \log \left( \frac{n + r}{n} \right) + 2n.
\]

In addition, if \( \mathcal{N} \) is an \( \varepsilon \)-net of \( \mathcal{B}(\pi, r) \), then the set of balls \( \{ \mathcal{B}(\sigma, \varepsilon) \}_{\sigma \in \mathcal{N}} \) covers \( \mathcal{B}(\pi, r) \). By Lemma 10.1, we obtain

\[
\log N(\mathcal{B}(\pi, r), \varepsilon) \geq n \log(\ell / n) - n - n \log(1 + \varepsilon/n) - n = n \log \left( \frac{\varepsilon}{n + \varepsilon} \right) - 2n,
\]

as claimed. \( \square \)

10.2.2 Proof of the minimax upper bound

We only present the proof of the upper bound in Theorem 10.1 with \( \lambda = 1/4 \) for simplicity. The estimator we use is a sieve maximum likelihood estimator (MLE), meaning that it is the MLE over a net (called a sieve). More precisely, define \( \varphi = \frac{1}{n} \binom{n}{2} \). Let \( \mathcal{P} \) be a maximal \( \varphi \)-packing (and thus a \( \varphi \)-net) of \( \mathcal{G}_n \) with respect to \( d_{KT} \). Consider the sieve MLE

\[
\hat{\pi} \in \arg\max_{\pi \in \mathcal{P}} \sum_{\pi(i) < \pi(j)} A_{i,j}.
\]

**Basic setup** Since \( \mathcal{P} \) is a \( \varphi \)-net, there exists \( \sigma \in \mathcal{P} \) such that \( D \triangleq d_{KT}(\sigma, \pi^*) \leq \varphi \). By definition of \( \hat{\pi} \), \( \sum_{\hat{\pi}(i) < \hat{\pi}(j)} A_{i,j} \geq \sum_{\sigma(i) < \sigma(j)} A_{i,j} \). Cancelling concordant pairs \((i, j)\) under \( \hat{\pi} \) and \( \sigma \), we see that

\[
\sum_{\hat{\pi}(i) > \hat{\pi}(j), \sigma(i) > \sigma(j)} A_{i,j} \geq \sum_{\hat{\pi}(i) > \hat{\pi}(j), \sigma(i) < \sigma(j)} A_{i,j}.
\]

Splitting the summands according to \( \pi^* \) yields that

\[
\sum_{\sigma(i) > \sigma(j), \pi^*(i) < \pi^*(j)} A_{i,j} + \sum_{\sigma(i) < \sigma(j), \pi^*(i) > \pi^*(j)} A_{i,j} \geq \sum_{\sigma(i) > \sigma(j), \pi^*(i) < \pi^*(j)} A_{i,j} + \sum_{\sigma(i) < \sigma(j), \pi^*(i) > \pi^*(j)} A_{i,j}.
\]

Since \( A_{i,j} \geq 0 \), we may drop the rightmost term and drop the condition \( \hat{\pi}(i) < \hat{\pi}(j) \) in the leftmost term to obtain that

\[
\sum_{\sigma(i) > \sigma(j), \pi^*(i) < \pi^*(j)} A_{i,j} + \sum_{\sigma(i) < \sigma(j), \pi^*(i) > \pi^*(j)} A_{i,j} \geq \sum_{\sigma(i) > \sigma(j), \sigma(i) < \sigma(j), \pi^*(i) < \pi^*(j)} A_{i,j}.
\]

To set up the rest of the proof, we define, for \( \pi \in \mathcal{P} \),

\[
L_\pi = |\{(i, j) \in [n]^2 : \pi(i) < \pi(j), \sigma(i) > \sigma(j), \pi^*(i) > \pi^*(j)\}| = |\{(i, j) \in [n]^2 : \pi(i) > \pi(j), \sigma(i) < \sigma(j), \pi^*(i) < \pi^*(j)\}|.
\]

78
Moreover, define the random variables

\[ X_\pi = \sum_{\pi(i) > \pi(j), \pi^*(i) < \pi^*(j)} A_{i,j}, \quad Y_\pi = \sum_{\pi(i) < \pi(j), \pi^*(i) > \pi^*(j)} A_{i,j}, \quad \text{and} \quad Z = \sum_{\pi(i) = \pi(j), \pi^*(i) = \pi^*(j)} A_{i,j}. \]

We show that the random process \( X_\pi - Y_\pi - Z \) is positive with high probability if \( d_{KT}(\pi, \sigma) \) is large.

**Binomial tails** For a single pairwise comparison sampled uniformly from the possible \( \binom{n}{2} \) pairs, the probability that

1. the chosen pair \((i, j)\) satisfies \( \pi(i) > \pi(j), \sigma(i) < \sigma(j) \) and \( \pi^*(i) < \pi^*(j) \), and

2. item \( i \) wins the comparison,

is equal to \( \frac{3}{4} L_\pi \binom{n}{2}^{-1} \). By definition, \( X_\pi \) is the number of times the above event happens if \( N \) independent pairwise comparisons take place, so \( X_\pi \sim \text{Bin}(N, \frac{3}{4} L_\pi \binom{n}{2}^{-1}) \). Similarly, we have \( Y_\pi \sim \text{Bin}(N, \frac{1}{4} L_\pi \binom{n}{2}^{-1}) \) and \( Z \sim \text{Bin}(N, \frac{3}{4} D \binom{n}{2}^{-1}) \). The tails of a Binomial random variable can be bounded by the following lemma.

**Lemma 10.2.** For \( 0 < r < p < s < 1 \) and \( X \sim \text{Bin}(N, p) \), we have

\[ \mathbb{P}(X \leq rN) \leq \exp \left( -N \frac{(p - r)^2}{2p(1 - r)} \right) \quad \text{and} \quad \mathbb{P}(X \geq sN) \leq \exp \left( -N \frac{(p - s)^2}{2s(1 - p)} \right). \]

Therefore, we obtain

1. \( \mathbb{P}(X_\pi \leq \frac{5}{8} L_\pi N \binom{n}{2}^{-1}) \leq \exp \left( -L_\pi N \binom{n}{2}^{-1}/128 \right) \),

2. \( \mathbb{P}(Y_\pi \geq \frac{3}{8} L_\pi N \binom{n}{2}^{-1}) \leq \exp \left( -L_\pi N \binom{n}{2}^{-1}/128 \right) \), and

3. \( \mathbb{P}(Z \geq 2\varphi N \binom{n}{2}^{-1}) \leq \exp \left( -\varphi N \binom{n}{2}^{-1}/4 \right) = \exp(-n/4) \).

Then we have that

\[ \mathbb{P}(X_\pi - Y_\pi \leq \frac{1}{4} L_\pi N \binom{n}{2}^{-1}) \leq 2 \exp \left( -L_\pi N \binom{n}{2}^{-1}/128 \right). \]  \quad (10.6)

**Peeling and union bounds** For an integer \( r \in \left[ C\varphi, \binom{n}{2} \right] \) where \( C \) is a sufficiently large constant to be chosen, consider the slice \( S_r = \{ \pi \in \mathcal{P} : L_\pi = r \} \). Note that if \( \pi \in S_r \), then

\[ d_{KT}(\pi, \pi^*) = | \{(i, j) : \hat{\pi}(i) < \hat{\pi}(j), \pi^*(i) > \pi^*(j) \}| \]

\[ \leq 4 \sum_{\pi(i) = \pi(j), \pi^*(i) = \pi^*(j)} A \]

\[ + | \{(i, j) : \pi(i) = \pi(j), \pi^*(i) = \pi^*(j) \}| \]

\[ = L_\pi + d_{KT}(\sigma, \pi^*) \leq r + \varphi, \]  \quad (10.7)

showing that \( S_r \subseteq \mathcal{B}(\pi^*, r + \varphi) \). Therefore, Proposition 10.1 gives

\[ \log |S_r| \leq n \log \frac{2n + 2r + 2\varphi}{\varphi} + 2n \leq n \log \frac{45r}{\varphi}. \]
By (10.6) and a union bound over $S_r$, we have $\min_{\pi \in S_r} (X_\pi - Y_\pi) > \frac{1}{4} rN \left(\frac{n}{2}\right)^{-1}$ with probability

$$1 - \exp \left( n \log \frac{45r}{\varphi} + \log 2 - \frac{rN}{128\left(\frac{n}{2}\right)} \right) \geq 1 - \exp(-2n),$$

where the inequality holds by the definition of $\varphi$ and the range of $r$. Then a union bound over integers $r \in [C\varphi, \left(\frac{n}{2}\right)]$ yields that

$$X_\pi - Y_\pi > C \frac{\varphi N \left(\frac{n}{2}\right)^{-1}}{4}$$

for all $\pi \in P$ such that $L_\pi \geq C\varphi$ with probability at least $1 - e^{-n}$. This is larger than the above high probability upper bound on $Z$, so we conclude that with probability at least $1 - e^{-n/8},$

$$X_\pi - Y_\pi - Z > 0$$

for all $\pi \in P$ with $L_\pi \geq C\varphi$. However, (10.5) says that $X_\hat{\pi} - Y_\hat{\pi} - Z \leq 0$, so $L_\hat{\pi} \leq C\varphi$ on the above event. By (10.7), $d_{KT}(\hat{\pi}, \pi^*) \leq L_\pi + \varphi$ on the same event, which completes the proof.

10.3 An efficient algorithm for noisy sorting

Let us move on to present an efficient algorithm. We continue to assume $\lambda = 1/4$. To recover the underlying order of items, it is equivalent to estimate the row sums $\sum_{j=1}^n M_{\pi^*(i), \pi^*(j)}$ which we call scores of the items. Initially, for each $i \in [n]$, we estimate the score of item $i$ by the number of wins item $i$ has. If item $i$ has a much higher score than item $j$ in the first stage, then we are confident that item $i$ is stronger than item $j$. Hence in the second stage, we know $M_{\pi^*(i), \pi^*(j)} = 3/4$ with high probability. For those pairs that we are not certain about, $M_{\pi^*(i), \pi^*(j)}$ is still estimated by its empirical version. The variance of each score is thus greatly reduced in the second stage, thereby yielding a more accurate order of the items. Then we iterate this process to obtain finer and finer estimates of the scores and the underlying order.

To present the $T$-stage sorting algorithm formally, we split the sample into $T$ subsamples each containing $N/T$ pairwise comparisons. For $t \in [T]$, we define a matrix $A^{(t)} \in \mathbb{R}^{n \times n}$ by setting $A^{(t)}_{i,j}$ to be the number of times item $i$ beats item $j$ in the $t$-th sample. The algorithm proceeds as follows:

1. For $i \in [n]$, define $I^{(0)}(i) = [n]$, $I^{(0)}_+(i) = \emptyset$ and $I^{(0)}_-(i) = \emptyset$. For $0 \leq t \leq T$, we use $I^{(t)}(i)$ to denote the set of items $j$ whose ranking relative to $i$ has not been determined by the algorithm at stage $t$.

2. At stage $t$, compute the score $S^{(t)}_i$ of item $i$:

$$S^{(t)}_i = \frac{T \left(\frac{n}{2}\right)}{N} \sum_{j \in I^{(t-1)}(i)} A^{(t)}_{i,j} + \frac{3}{4} |I^{(t-1)}_-(i)| + \frac{1}{4} |I^{(t-1)}_+(i)|.$$

3. Set the threshold

$$\tau^{(t)}_i \approx n \sqrt{|I^{(t-1)}(i)| TN^{-1} \log(nT)},$$

and define the sets

$I^{(t)}_+(i) = \{ j \in [n] : S^{(t)}_j - S^{(t)}_i < -\tau^{(t)}_i \}$,

$I^{(t)}_-(i) = \{ j \in [n] : S^{(t)}_j - S^{(t)}_i > \tau^{(t)}_i \}$,

and

$I^{(t)}(i) = [n] \setminus (I^{(t)}_-(i) \cup I^{(t)}_+(i))$. 

80
4. Repeat step 2 and 3 for \( t = 1, \ldots, T \). Output a permutation \( \hat{\pi}^{MS} \) by sorting the scores \( S_i^{(T)} \) in nonincreasing order, i.e., \( S_i^{(T)} \geq S_j^{(T)} \) if \( \hat{\pi}^{MS}(i) < \hat{\pi}^{MS}(j) \).

We take \( T = \lceil \log \log n \rceil \) so that the overall time complexity of the algorithm is only \( O(n^2 \log \log n) \).

**Theorem 10.2.** With probability at least \( 1 - n^{-7} \), the algorithm with \( T = \lceil \log \log n \rceil \) stages outputs an estimator \( \hat{\pi}^{MS} \) that satisfies

\[
\| \hat{\pi}^{MS} - \pi^* \|_\infty \lesssim \frac{n^2}{N} (\log n) \log \log n
\]

and

\[
d_{KT}(\hat{\pi}^{MS}, \pi^*) \lesssim \frac{n^3}{N} (\log n) \log \log n.
\]

The second statement follows from the first one together with (10.3).

### 10.3.1 Proof (sketch) of Theorem 10.2

Assume that \( \pi^* = \text{id} \) without loss of generality. We define a score

\[
s_i^* = \sum_{j \in [n] \setminus \{i\}} M_{i,j} = \frac{i}{2} + \frac{n}{4} - \frac{3}{4}
\]

for each \( i \in [n] \), which is simply the \( i \)-th row sum of \( M \) minus 1/2.

**Lemma 10.3.** Fix \( t \in [T] \), \( I \subseteq [n] \) and \( i \in I \). Let us define

\[
S = \frac{T(n)}{N} \sum_{j \in I} A_{i,j}^{(t)} + \frac{3}{4} |\{ j \in [n] \setminus I : j < i \}| + \frac{1}{4} |\{ j \in [n] \setminus I : j > i \}|.
\]

If \( |I| \) is not too small, then it holds with probability at least \( 1 - (nT)^{-9} \) that

\[
|S - s_i^*| \lesssim n \sqrt{|I|TN^{-1} \log(nT)}.
\]

**Proof.** The probability that a uniform pair consists of item \( i \) and an item in \( I \setminus \{i\} \), and that item \( i \) wins the comparison, is equal to \( q \triangleq \left( \sum_{j \in I \setminus \{i\}} M_{i,j} \right) / \binom{n}{2} \). Thus the random variable \( X \triangleq \sum_{j \in I} A_{i,j}^{(t)} \) has distribution \( \text{Bin}(N/T, q) \). In particular, we have \( \mathbb{E}[X] = Nq/T = \frac{N}{T} \sum_{j \in I \setminus \{i\}} M_{i,j} \), so \( S \) is an unbiased estimate of \( s_i^* \). Moreover, we have the tail bound

\[
\mathbb{P}\left( |X - \mathbb{E}[X]| \gtrsim \sqrt{qNT^{-1} \log(nT)} \right) \leq (nT)^{-9},
\]

from which the conclusion follows. \( \square \)

We apply Lemma 10.3 inductively to each stage of the algorithm. By a union bound over all \( i \in [n] \) and \( t \in [T] \), all the events studied below hold with high probability. For \( t \in [T] \), define

\[
\mathcal{E}^{(t-1)} \triangleq \{ j < i \text{ for all } j \in I^{(t-1)}(i) \text{ and } j > i \text{, for all } j \in I^{(t-1)}_+(i) \}.
\]

On the event \( \mathcal{E}^{(t-1)} \), the score \( S_i^{(t)} \) is exactly the quantity \( S \) in Lemma 10.3 with \( I = I^{(t-1)}(i) \), so

\[
|S_i^{(t)} - s_i^*| \lesssim n \sqrt{|I^{(t-1)}(i)|TN^{-1} \log(nT)} = \tau_i^{(t)}/2. \quad (10.8)
\]

81
For any \( j \in I^{(t)}_+(i) \), by definition \( S^{(t)}_j - S^{(t)}_i < -\tau^{(t)}_i \), so we have \( s^*_j < s^*_i \) and thus \( j > i \). Similarly, \( j < i \) for any \( j \in I^{(t)}_-(i) \). Hence \( E^{(t)} \) occurs with high probability. Moreover, if \( |s^*_j - s^*_i| > 2\tau^{(t)}_i \), then \( |S^{(t)}_j - S^{(t)}_i| > \tau^{(t)}_i \), so \( j \notin I^{(t)}(i) \). Hence if \( j \in I^{(t)}(i) \), then \( |j - i| \leq \tau^{(t)}_i \). Consequently,

\[
|I^{(t)}(i)| \lesssim \tau^{(t)}_i \lesssim n\sqrt{|I^{(t-1)}(i)|TN^{-1}\log(nT)}.
\]  

(10.9)

Note that if we have \( \alpha^{(0)} = n \) and the iterative relation \( \alpha^{(t)} \leq \beta \sqrt{\alpha^{(t-1)}} \) where \( \alpha^{(t)} > 0 \) and \( \beta > 0 \), then it is easily seen that \( \alpha^{(t)} \leq \beta^2 n^{2^{-t}} \). Consequently, we obtain that

\[
|I^{(T-1)}(i)| \lesssim \frac{n^2}{N} T \log(nT)n^{2^{-T+1}} \lesssim \frac{n^2}{N}(\log n)(\log \log n)
\]

for \( T = \lfloor \log \log n \rfloor \). Taking \( T \) to be larger does not make \( |I^{(T-1)}(i)| \) smaller, because Lemma 10.3 requires a lower bound on \( |I^{(T-1)}(i)| \). The details are left out. It follows from (10.8) that

\[
|S^{(T)}_i - s^*_i| \lesssim \frac{n^2}{N}(\log n)(\log \log n) =: \delta.
\]

As the permutation \( \hat{\pi}^{MS} \) is defined by sorting the scores \( S^{(T)}_i \) in nonincreasing order, we see that \( \hat{\pi}^{MS}_i < \hat{\pi}^{MS}_j \) for all pairs \((i, j)\) with \( s^*_i - s^*_j > 2\delta \), i.e., \( j - i > \delta \).

Finally, suppose that \( \hat{\pi}^{MS}(i) - i > \delta \) for some \( i \in [n] \). Then there exists \( j > i + \delta \) such that \( \hat{\pi}^{MS}(j) < \hat{\pi}^{MS}(i) \), contradicting the guarantee we have just proved. A similar argument leads to a contradiction if \( \hat{\pi}^{MS}(i) - i < -\delta \). Therefore, we obtain that \( |\hat{\pi}^{MS}(i) - i| \leq \delta \), completing the proof.
Previously in Lecture 9, we discussed the exact recovery of SBM. In this lecture, we turn to the almost exact recovery. Let $\sigma$ be the community labels of nodes. Recall that the loss to evaluate a community estimate $\hat{\sigma}$ is 

$$
\ell(\sigma, \hat{\sigma}) = \frac{1}{n} \min_{\delta} d_H(\sigma, \hat{\sigma}),
$$

where $d_H(x, y)$ is the Hamming distance $\sum_i I\{x_i \neq y_i\}$. We call $\hat{\sigma}$ an almost exact recovery if $E\ell(\sigma, \hat{\sigma}) = o(1)$, and an Exact recovery if $\ell(\sigma, \hat{\sigma}) = 0$ w.h.p. We have seen that the requirement is $H^2(P, Q) \geq (2+\epsilon) \log \frac{n}{\epsilon}$ for all $\epsilon > 0$ for exact recovery. Here we are going to show the necessary and sufficient condition for almost exact recovery is $H^2(P, Q) \gg \frac{1}{n}$, which can be achieved by SDP relaxation.

We first introduce the key technical tool: Grothendieck Inequality (Theorem 11.1). Then we discuss its application to SBM following Guédon-Vershynin [GV16].

### 11.1 $\| \cdot \|_{\infty \to 1}$ norm

Consider $A \in \mathbb{R}^{n \times m}$. We look at the following optimization

$$
\max_{x, y = \pm 1} \sum_{1 \leq i \leq n, 1 \leq j \leq m} a_{ij} x_i y_j = \max_{x \in \{\pm\}^n, y \in \{\pm\}^m} \langle A, xy^\top \rangle.
$$

(11.1)

**Remark 11.1.** The objective above (11.1) is a norm of $A$, denoted as $\|A\|_{\infty \to 1} = \max_{\|x\|_\infty \leq 1} \|Ax\|_1$. This is easily seen by writing $\| \cdot \|_1$ in the dual form.

**Remark 11.2.** $\|A\|_{\infty \to 1}$ is closely related to the cut norm. The cut norm $\|A\|_{\text{cut}}$ is defined as (cf. the min cut in Lecture 9)

$$
\|A\|_{\text{cut}} = \max_{I \subset [n], J \subset [m]} \left| \sum_{i \in I, j \in J} a_{ij} \right|.
$$

The relation of the two norms is

$$
\|A\|_{\text{cut}} \leq \|A\|_{\infty \to 1} \leq 4 \|A\|_{\text{cut}}.
$$

The left side inequality can be seen by

$$
\|A\|_{\text{cut}} = \max_{I \subset [n], J \subset [m]} \left| \sum_{i \in I, j \in J} a_{ij} \right| \leq \max_{I \subset [n], J \subset [m]} \sum_{i \in I} \left| \sum_{j \in J} a_{ij} \right| \leq \max_{J} \|Ax_J\|_1 \leq \|A\|_{\infty \to 1}.
$$

$x_J$ is the indicator vector of $J$. The right side inequality can be shown by writing $x = I\{I\} - I\{I^c\}, y = I\{J\} - I\{J^c\}$ in (11.1).
We have the SDP relaxation of $\|A\|_{\infty \to 1}$: for $r \geq n + m$ (otherwise it is nonconvex),

$$\text{SDP}(A) = \max_{u_i, v_j \in \mathbb{R}, \|u_i\| = \|v_j\| = 1} \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} \langle u_i, v_j \rangle. \quad (11.2)$$

**Remark 11.3.** When $r = 1$, SDP$(A)$ corresponds to $\|A\|_{\infty \to 1}$. Thus it is indeed a “relaxation” of the norm: $\|A\|_{\infty \to 1} \leq \text{SDP}(A)$.

**Remark 11.4** (Dimension-free). SDP$(A)$ is dimension-free in the sense that the value does not depend on $r$ as long as $r \geq n + m$. In particular, if it helps construction, we are free to consider the infinite-dimensional setting, e.g., the decision variables $u_i, v_j$ take values in the Hilbert space of random variables – and we will do so next.

**Remark 11.5** (Standard Form). SDP$(A)$ can be written into a standard SDP form

$$\text{SDP}(A) = \max_{X \succeq 0, X_{ii} = 1} \langle W, X \rangle$$

where $W = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}$. The correspondence is by writing

$$X = \begin{pmatrix} U^T \\ V^T \end{pmatrix} \begin{pmatrix} U & V \end{pmatrix} = \begin{pmatrix} U^T U & U^T V \\ V^T U & V^T V \end{pmatrix}.$$

We can see the role of $r$ in SDP$(A)$ is rank$(X) \leq r$. But $X$ is $n + m$ by $n + m$, so as long as $r \geq n + m$, this constrain disappears.

### 11.2 Grothendieck Inequality

**Theorem 11.1** (Grothendieck Inequality).

$$\|A\|_{\infty \to 1} \leq \text{SDP}(A) \leq k \|A\|_{\infty \to 1}.$$  

*Here the absolute constant $k$ can be chosen as $k = \frac{1}{\pi - 1} \approx 3.66$ (with the world record $\approx 1.78$).*

**Proof:** following Rietz [Rie74]. The left side is obvious and stated in Remark 11.3. We focus on the right side. The main idea is randomized rounding. Let $u_i, v_j \in S^{d-1}$ achieve the maximum in SDP$(A)$ (11.2), $d = n + m$. We hope such $u_i, v_j$ can match (not too far away from in objective) the $x_i, y_j$ in (11.1). If we take some random $x_i, y_j$, then we can have the lower bound

$$\|A\|_{\infty \to 1} \geq \mathbb{E} \sum a_{ij} x_i y_j = \sum a_{ij} E(x_i y_j).$$

But $a_{ij}$ can be positive or negative, so we cannot go further directly.

Consider $x_i = \text{sign}(\langle g, u_i \rangle), y_j = \text{sgn}(\langle g, v_j \rangle), g \sim N(0, I_d)$.

**Fact 11.1.** Note $\frac{g}{\|g\|} \sim \text{unif} (S^{d-1})$, so

$$\mathbb{E} x_i y_j = \frac{2}{\pi} \text{arcsin}(u_i, v_j).$$
Denote by $g_{u_i} = \langle g, u_i \rangle$, $g_{v_j} = \langle g, v_j \rangle$. We consider the generic setting $x_i = f(g_{u_i}), y_j = f(g_{v_j})$ for some $f : \mathbb{R} \to [-1, 1]$. We have the following facts

**Fact 11.2.**

1. $\mathbb{E}g_u g_v = \langle u, v \rangle$.
2. $\mathbb{E}u f(g_v) = \langle u, v \rangle \mathbb{E}Z g(Z) = \langle u, v \rangle K$, $Z \sim N(0, 1)$.
3. $\mathbb{E}(u - f(g_u))^2 = 1 - 2K + L$, $L \triangleq \mathbb{E}f^2(Z)$.

The facts hold noticing each $g_u \sim N(0, 1)$. Then

$$
\|A\|_{\infty \to 1} \geq \sum a_{ij}E \langle x_i, y_j \rangle \\
= \sum a_{ij}E f(g_{u_i})f(g_{v_j}) \\
= \sum a_{ij}E (g_{u_i} - f(g_{u_i}))(g_{v_j} - f(g_{v_j})) - \sum a_{ij}E g_{u_i} g_{v_j} \\
+ \sum a_{ij}E (g_{v_j} f(g_{u_i}) + g_{u_i} f(g_{v_j}))
$$

(def of $u_i, v_j$) $= \sum a_{ij}E (g_{u_i} - f(g_{u_i}))(g_{v_j} - f(g_{v_j})) + (2K - 1)\text{SDP}(A)$.  

The magical next step is observing $(\ast)$ is a feasible representation (after normalizing) in (11.2). Thus

$|\langle \ast \rangle| \leq (1 - 2K + L)\text{SDP}(A)$. \\
$\Rightarrow \|A\|_{\infty \to 1} \geq (4K - L - 2)\text{SDP}(A)$.

Let $f = \text{sgn}$. Then $L = 1, K = \mathbb{E}|Z| = \sqrt{\frac{2}{\pi}}$, $4K - L - 2 = 4\sqrt{\frac{2}{\pi}} - 3 \approx 0.19 < 0.2$. So we proved that we can choose $k = 5$ in the theorem.

Moreover, there is a natural way to improve the constant. If we replace $f$ by $\alpha f$ in the derivation above, then

$$
\alpha^2\|A\|_{\infty \to 1} \geq \sum a_{ij}E \alpha f(g_{u_i}) \cdot \alpha f(g_{v_j}) \\
= \sum a_{ij}E (g_{u_i} - f(g_{u_i}))(g_{v_j} - f(g_{v_j})) - \sum a_{ij}E g_{u_i} g_{v_j} \\
+ \sum a_{ij}E (\alpha g_{v_j} f(g_{u_i}) + \alpha g_{u_i} f(g_{v_j}))
$$

$= |\langle \ast \rangle(\alpha)| + (2\alpha K - 1)\text{SDP}(A)$. 

And

$$
|\langle \ast \rangle(\alpha)| \leq (1 - 2\alpha K + \alpha^2 L)\text{SDP}(A).
$$

Then the statement would be

$$
\|A\|_{\infty \to 1} \geq \left(\frac{4K}{\alpha} - \frac{2}{\alpha^2} - L\right)\text{SDP}(A).
$$

The optimal $\alpha$ is $\frac{1}{K}$, and the bound is

$$
\|A\|_{\infty \to 1} \geq (2K^2 - L)\text{SDP}(A)
$$

In the case $f = \text{sgn}$, $2K^2 - L = \frac{4}{\pi} - 3$, as suggested in the theorem.

Clearly, the best strategy is to let

$$
f = \arg\max_{|f| \leq \mathbb{E}f(Z) > 0} 2K^2 - L.
$$

The solution is given by the bounded linear function, $f(x) = \max(0, \min(x, 1))$.  

85
Remark 11.6 (PSD $A$). When $A$ is psd, the optimal bound is

$$\|A\|_{\infty \to 1} \geq \frac{2}{\pi} \text{SDP}(A).$$

In this case, we can lower bound $(\star)$ by 0 instead, since by design $u_i = v_i$ in this case. Then

$$\|A\|_{\infty \to 1} \geq \text{SDP}(A) \geq \sup_{\alpha} \frac{2\alpha K - 1}{\alpha^2} = K^2 = \frac{2}{\pi}.$$  

To show the sharpness, we construct examples that (asymptotically) achieve the bound. Let $l_i \sim \text{unif}(S^{d-1})$, $A_{ij} = \frac{1}{n^2} \langle l_i, l_j \rangle$. Choose $u_i = v_i = l_i$ in (11.2), we have

$$\text{SDP}(A) \geq \frac{1}{n^2} \sum_{i,j} \langle l_i, l_j \rangle^2 \approx \mathbb{E} \langle l, l' \rangle^2 = \frac{1}{d} (1 + o(1)).$$

But

$$\|A\|_{\infty \to 1} \text{ w.h.p} \leq \frac{2}{d} \left( \frac{2}{\pi} + o(1) \right).$$

This is because

$$\|A\|_{\infty \to 1} = \max_{x \in \{\pm\}^n} \langle A, xx^\top \rangle = \max_{x \in \{\pm\}^n} \frac{1}{n^2} \sum_{i,j} x_i x_j \langle l_i, l_j \rangle = \max_{x \in \{\pm\}^n} \frac{1}{n} \sum_i x_i l_i \|l_i\|_2^2 = \|l\|_{2 \to 1}.$$  

Remark 11.7 (Connection to max-cut). Given a weighted graph with weight matrix $W$, similar to min-cut in (9.2), define:

$$\text{maxcut}(W) \triangleq \max_{I \subset [n]} \sum_{i \in I, j \in \overline{I}} W_{ij}.$$  

Then

$$2 \text{maxcut}(W) = \max_{\sigma \in \{\pm\}^n} \sum_{i,j} W_{ij} (1 - \sigma_i \sigma_j) = \max_{\sigma \in \{\pm\}^n} \langle W, J - \sigma \sigma^\top \rangle \leq \max_{X \succeq 0, \sum_{i=1}^n X_{ii} = 1} \langle W, J - X \rangle \triangleq GW(W).$$  

The same as the proof of Theorem 11.1, note here $W_{ij} \geq 0$,

$$2 \text{maxcut}(W) \geq \sum_{i,j} W_{ij} (1 - \frac{2}{\pi} \arccos(u_i, v_j)) \geq 0.878 \sum_{i,j} W_{ij} (1 - \langle u_i, v_j \rangle) = 0.878GW(W).$$  

### 11.3 Application to SBM

Consider $SBM(n, p, q)$, $p = \frac{a}{n}$, $q = \frac{b}{n}$, and bisection $\langle \sigma, 1 \rangle = 0$. Define $d = \frac{a+b}{2}$, $s = a - b$. Recall in the bisection case (see Remark 10.2), the MLE has the following SDP relaxation

$$\hat{X} = \arg \max \langle A, X \rangle.$$  

\begin{align*}
X &\succeq 0 \\
X_{ii} &= 1 \\
\langle X, J \rangle &= 0
\end{align*}
We claim that the necessary and sufficient condition is
\[
\frac{(a-b)^2}{a+b} \to \infty.
\]
Here \(\frac{(a-b)^2}{a+b}\) can be interpreted as the signal-to-noise ratio (snr). In the more general \(P/Q\) model, the condition is \(H^2(P,Q) \gg \frac{1}{n}\), which recovers the above when \(P = \text{Bern}(p)\) and \(Q = \text{Bern}(q)\).

**Theorem 11.2 ([GV16]).** Let \(\hat{v} = \text{the top eigenvector of } \hat{X}\), and \(\hat{\sigma} = \text{sgn}(\hat{v})\). Then
\[
\mathbb{E} l(\hat{\sigma}, \sigma) \overset{\text{(also w.h.p)}}{\lesssim} \frac{1}{\sqrt{\text{snr}}}.
\]

Note: The above misclassification rate is later sharpened to exponential (optimal) by [FC18]:
\[
\mathbb{E} l(\hat{\sigma}, \sigma) \leq \exp(-\Omega(\text{snr})�.\]

**Proof.** Define the population solution
\[
X^* = \arg \max \langle E A, X \rangle.
\]
\[
X \succeq 0
\]
\[
X_{ii} = 1
\]
\[
\langle X, J \rangle = 0
\]
We can calculate
\[
E A = \frac{p+q}{2} J + \frac{p-q}{2} \sigma \sigma^T - p I
\]
and justify \(\sigma \sigma^T = X^*\).
\[
\langle E A, \hat{X} \rangle = \langle A, \hat{X} \rangle - \langle A - E A, \hat{X} \rangle
\]
\[
\geq \langle A, X^* \rangle - \langle A - E A, \hat{X} \rangle
\]
\[
= \langle E A, X^* \rangle + \underbrace{\langle A - E A, X^* \rangle - \langle A - E A, \hat{X} \rangle}_{\equiv -\delta}.
\]
If we can somehow say \(\delta \leq 0\), in other words \(\langle A - E A, X^* \rangle - \langle A - E A, \hat{X} \rangle \geq 0\), then we can conclude \(\langle E A, \hat{X} \rangle \geq \langle E A, X^* \rangle\), and thus \(\hat{X} = X^*\). Though this is not possible in general, we can show \(\delta\) is not too big to get the conclusion. Let \(\hat{v} = v_1(X), v = v_1(X^*) = \frac{\pi}{\sqrt{n}}\), then by sin\(\theta\) law,
\[
\min \|\hat{v} \pm v\|_2 \overset{\text{w.h.p}}{\lesssim} \frac{\|\hat{X} - X^*\|_F}{\lambda_1(X^*) - \lambda_2(X)} \leq \frac{\|\hat{X} - X^*\|_F}{n - 0} = \frac{\|\hat{X} - X^*\|_F}{n}.
\]
Also note that for every \(\sigma_i \neq \hat{\sigma}_i\), \(\|\hat{v} \pm v\|_2^2\) differs at least \(\frac{1}{n}\) at this \(i\). Thus
\[
l(\hat{\sigma}, \sigma) \leq \frac{1}{n} \cdot n \min \|\hat{v} \pm v\|_2^2 \lesssim \frac{\|\hat{X} - X^*\|_F^2}{n^2}.
\]
Suppose \(n \sqrt{d} \overset{\text{w.h.p}}{\gtrsim} \delta \geq \langle E A, X^* \rangle - \langle E A, \hat{X} \rangle = \frac{p-q}{2}(n^2 - \langle \sigma \sigma^T, \hat{X} \rangle)\). Then
\[
\|\hat{X} - X^*\|_F^2 = \|\hat{X}\|_F^2 + \|X^*\|_F^2 - 2\langle \hat{X}, X^* \rangle
\]
\[
= \|\hat{X}\|_F^2 + n^2 - 2\langle \hat{X}, \sigma \sigma^T \rangle
\]
\[
\leq \text{Tr}(\hat{X})^2 + n^2 - 2\langle \hat{X}, \sigma \sigma^T \rangle
\]
\[
= 2(n^2 - \langle \sigma \sigma^T, \hat{X} \rangle) \lesssim \frac{\delta}{p-q} = \frac{n \delta}{a-b} \leq \frac{n^2}{\sqrt{\text{snr}}}.
\]
This completes the proof. So it remains only to show $\delta \lesssim n \sqrt{d}$. Denote $W = A - \mathbb{E}A$. We want to show

$$
\frac{1}{2} |\delta| \leq \text{SDP}(W) = \max_{X \succeq 0, X_{ii} = 1} \langle W, X \rangle \overset{w.h.p}{\lesssim} n \sqrt{d}.
$$

By Grothendieck Inequality,

$$
\text{SDP}(W) = \max_{\|u\|_1 = 1} \sum_{i,j} W_{ij} \langle u_i, u_j \rangle \\
\leq \max_{\|u\|_1 = \|v\|_1 = 1} \sum_{i,j} W_{ij} \langle u_i, v_j \rangle \\
\overset{G.I.}{\lesssim} \|W\|_{\infty \to 1} \\
= \max_{x, y \in \{\pm\}^n} \langle W, xy^T \rangle.
$$

By Hoeffding’s inequality Lemma 2.2,

$$
P(\|\langle W, xy^T \rangle\| \geq t) \leq \exp(-c \cdot \frac{t^2}{n^2}).
$$

To apply union bound on $x, y$, which in total $4^n$, we need to choose $t \sim n^{3/2}$. We apply Bernstein’s inequality instead,

$$
P(\|\langle W, xy^T \rangle\| \geq t) \leq \exp(-c \cdot \frac{t^2}{t + nd}),
$$

then we can choose $t \sim n \sqrt{d}$. \hfill \Box
Initiated by [BR13], studying the computational limit of statistical problems is a broad topic. Results of varying precision have been obtained for different problems. As a case study, let us consider the problem of submatrix detection (biclustering):

Observe

\[ X = M + Z \]

where \( X \) is a \( N \times N \) matrix, \( M \) is the signal and \( Z \) is the noise. \( Z \) is i.i.d \( N(0, 1) \) and \( M = \mu \mathbf{1}_S \mathbf{1}_S^T \) where \( S \subset [N], |S| = K \), and \( S \) is chosen uniformly at random. The problem is parametrized by \((N, K, \mu)\). We consider the asymptotic regime where \( N \to \infty \) and it is convenient to look at the exponents so let \( K = N^\alpha \) where \( \alpha \in [0, 1] \) and \( \mu = N^{-\beta} \) where \( \beta \in \mathbb{R} \). We consider the problem of detection, i.e., testing between

\[
\begin{cases}
  H_0 : X = Z \quad (\mu = 0) \\
  H_1 : X = M + Z
\end{cases}
\]

The main result can be represented in an “easy-hard-impossible” phase transition diagram shown below.

![Difficulty of the problem in different regimes](image)

Figure 12.1: Difficulty of the problem in different regimes

In "impossible" regime, i.e., when \( \beta > \max\{\alpha/2, 2\alpha - 1\} \), we will show that for any test \( \phi : \mathbb{R}^{N \times N} \to \{0, 1\} \),

\[
\mathbb{P}_0(\phi = 1) + \sup_{S \in \binom{[N]}{K}} \mathbb{P}_S(\phi = 0) \to 1
\]

by showing that \( \text{TV}(\mathbb{P}_0, \frac{1}{(N^\alpha)} \sum_{S \in \binom{[N]}{K}} \mathbb{P}_S) \to 0 \).

In "easy" regime, i.e., when \( \beta < \max\{2\alpha - 1, 0\} \), there exist efficient tests that runs in \( O(N^2) \) times. More specifically, we consider the following two test statistics
1. Linear test statistic

\[ T_{\text{SUM}} = \sum_{i,j} X_{i,j} \sim \begin{cases} N(0, N^2) & \text{under } H_0 \\ N(\mu k^2, N^2) & \text{under } H_1 \end{cases} \]

This test succeeds if \( \mu k^2 \gg N \), i.e., if \( \beta < 2\alpha - 1 \).

2. Max test statistic

\[ T_{\text{MAX}} = \max_{i,j} X_{i,j} \mu \gg \sqrt{\log N} \approx \begin{cases} \Theta(\sqrt{\log N}) & \text{under } H_0 \\ \mu & \text{under } H_1 \end{cases} \]

This test succeeds if \( \mu \gg \sqrt{\log N} \), i.e., if \( \beta < 0 \).

Now we return to the proof of the impossibility regime. Let \( P_0 = \mathcal{L}(Z) \) denote the law under the null and \( P_1 = \mathcal{L}(X) \) denote the law under the alternative if \( M \) is draw uniformly at random. We will prove that \( \text{TV}(P_0, P_1) \to 0 \) by proving that \( \chi^2(P_1||P_0) \to 0 \). After some algebra we get

\[ \chi^2(P_1||P_0) = \mathbb{E} \exp \left( \langle M, \tilde{M} \rangle \right) - 1 \]

where \( M = \mu 1_S 1_S^T \) and \( \tilde{M} = \mu 1_{\tilde{S}} 1_{\tilde{S}}^T \) are i.i.d copies. Observe that

\[ \langle M, \tilde{M} \rangle = \mu^2 |S \cap \tilde{S}|^2 = \mu^2 H^2 \]

where \( H \overset{\Delta}{=} |S \cap \tilde{S}| \) follows \( \text{Hyp}(N, K, K) \) from previous lectures. Putting things together, we get

\[ \chi^2(P_1||P_0) = \mathbb{E} \exp (\mu^2 H^2) - 1 \]

To characterize the behavior of \( \mathbb{E} \exp (\mu^2 H^2) \), we make use of the following lemma.

**Lemma 12.1.** \( \mathbb{E} \exp (\lambda H^2) \leq C_1 \) if \( \lambda \leq C_2 \left( \frac{1}{K} \log \frac{eN}{K} \wedge \frac{N^2}{K^2} \right) \). Moreover, \( C_1 \to 1 \) if \( C_2 \to 0 \).

Therefore, Lemma 12.1 is suggesting \( \mu^2 \ll \frac{1}{K} \log \frac{eN}{K} \wedge \frac{N^2}{K^2} \), which is equivalent to \( \beta > \alpha/2 \) or \( \beta > 2\alpha - 1 \).

Several remarks on this Lemma is in order.

1. By Jensen inequality,

\[ \mathbb{E} \exp (\lambda H^2) \geq \exp(\lambda \frac{N^2}{K^4}) \]

Therefore, \( \lambda = O\left( \frac{N^2}{K^4} \right) \) is necessary.

2. 

\[ \mathbb{E} \exp (\lambda H^2) \geq \exp(\lambda K^2) \mathbb{P}(H = K) \]

Note that \( \mathbb{P}(H = K) = \frac{1}{(K)} \geq \left( \frac{N}{K} \right)^{-K} \). Therefore, \( \lambda = O\left( \frac{1}{K} \log \frac{eN}{K} \right) \) is also necessary.

3. Sharp constant is known.
In "hard" regime, i.e., $2\alpha - 1 < \beta < \alpha/2$, we first consider the scan test which is not computationally efficient, but works.

$$T_{\text{SCAN}} = \max_{|T| = K} \sum_{i \in T, j \in T} X_{ij}.$$  

Under $H_1$, $T_{\text{SCAN}} \geq \sum_{i \in S, j \in S} X_{i,j} \sim N(\mu K^2, K^2)$. Under $H_0$, $T_{\text{SCAN}} \lesssim \sqrt{K^2 \log \left(\frac{N}{K}\right) = K^{3/2} \sqrt{\log \frac{eN}{K}}}$.

Therefore, the test works if $\mu K^2 \gg K^{3/2} \Rightarrow \beta < \alpha/2$.

Now we discuss why in the regime $2\alpha - 1 < \beta < \alpha/2$ the problem is computationally hard. But first we have to discuss what we mean by polynomial-time algorithm. Conventionally, polynomial-time algorithm means that

$$\text{running time} = O\left(\text{poly}(\# \text{ of bits to describe the input})\right).$$

However, gaussian random variables are theoretical objects that are continuous and takes infinite number of values so takes infinite number of bits to describe precisely. So how do we make sense of the "Gaussian" noise?

Idea: find a discrete model that is asymptotically equivalent to Gaussian. To define asymptotically equivalent we first introduce the notion of LeCam deficiency of $P$ w.r.t $Q$. Let $P = \{P_\theta : \theta \in \Theta\}$ on $X$ and $Q = \{Q_\theta : \theta \in \Theta\}$ on $Y$. Let $T : X \rightarrow Y$ be a Markov kernel (conditional distribution) so that $TP(dy) = \int T(dy|x) P(dx)$. Define

$$\delta(P, Q) \triangleq \inf_{\theta \in \Theta} \sup_T \text{TV}(TP_\theta, Q_\theta).$$

Define the LeCam distance

$$\Delta(P, Q) \triangleq \max\{\delta(P, Q), \delta(Q, P)\}.$$  

Then asymptotically equivalent simply means $\Delta(P, Q) \rightarrow 0$. Now let $X = M + Z$ be $P$. We discretize $X$ into $X_t$, denoted $P_t$, by letting $(X_t)_{ij} = \frac{[X_{ij}2^t]}{2^t}$.

**Lemma 12.2.** $\Delta(P^d, P^d_t) \leq c N^2 \exp(-ct) \rightarrow 0$ as long as $t \geq C \log N$.

*Proof sketch.* $\delta(P, P_t) = 0$ by definition since we can let kernel $T$ to be the discretization procedure. $\delta(P_t, P) \lesssim N^2 \exp(-ct)$ because the total variation between $(X_t)_{ij} + \text{Unif}(0, 2^{-t})$ and $(X_t)_{ij}$ is roughly $\exp(-ct)$. \hfill $\square$

Therefore, the above discussion implies that it makes sense to talk about polynomial in the dimension $N$ instead of the number of bits to describe the inputs.

**Reduction**  Informally, Problem $A$ is at least as hard as Problem $B$ if $B$ can be reduced to $A$ in polynomial time.

We will show in the following sequel that the problem in ”hard” regime is at least as hard as the planted clique problem. Recall that the planted clique problem is a testing problem where $H_0 : G \sim G(n, \frac{1}{2})$ and $H_1 : G \sim G(n, \frac{1}{2}, k)$. We will turn a planted clique problem of size $n$ into a submatrix detection problem of size $N = nl$ where $l \rightarrow \infty$. The reduction consists of three steps.

1. We find a kernel $T : 0, 1 \rightarrow \mathbb{R}$ such that $\text{Ber}(\frac{1}{2}) \mapsto N(0, \frac{1}{l^2})$ and $\text{Ber}(1) \mapsto N(\mu, \frac{1}{l^2})$. Let

$$\frac{1}{2}(P_0 + P_1) = N(0, \frac{1}{l^2}) = Q \quad P_1 = N(\mu, \frac{1}{l^2}) = P.$$  

91
Then we get $P_0 = 2Q - P$. Indeed, we can choose $P_0$ and $P_1$ such that
\[ \frac{1}{2}(P_0 + P_1) = N(0, \frac{1}{l^2}) \quad P_1 \approx N(\mu, \frac{1}{l^2}). \]
Here the approximation denotes that the total variation distance is small.

2. Randomize the block.

3. Reduction scheme. For any $s, t \in [n]$, generate a $l \times l$ block $T$ by
\[
T = \begin{cases} 
  P_0 & \text{if } A_{st} = 0 \\
  P_1 & \text{if } A_{st} = 1 
\end{cases}.
\]
This is possible because if $X_1, \ldots, X_n$ i.i.d $N(\mu, 1)$, then $\bar{X}$ is sufficient statistic. In other words, we can simulate $X_1, \ldots, X_n$ given $\bar{X}$.

Observe that under $H_0$, $X \overset{d}{=} Z$. Under $H_1$, $\text{TV}(\mathcal{L}(X), \mathcal{L}(M + Z)) \to 0$. The approximation error comes from two places. a) $P_1 \neq N(\mu, \frac{1}{l^2})$ and b) diagonal $l \times l$ blocks are always $N(0, 1)$.

So the reduction scheme works for $N = nl, K = kl, l \ll \frac{1}{\mu}$. Pick $l - \frac{1}{\mu} \frac{1}{\log N}$, we see that $k \ll \sqrt{n} \iff \mu \ll \frac{N}{K^2} \iff \beta > 2\alpha - 1$. 
Bibliography


